Mathematical Statistics

Nikolai Chernov

1 Exploratory Data Analysis

1.1 Example

In a class of 20 students, the teacher records the following test scores:

16	5	13	14	20	15	18	11	16	12
17	10	18	8	19	13	15	9	11	16

This is what we call *raw data*, or unprocessed measurements (facts, observations). Statistics is an *art of data analysis*, this is its first goal. We will see how it does that.

1.2 Ordering

The first thing to do is *order the available measurements*:

5 8 9 10 11 11 12 13 13 14 15 15 16 16 16 17 18 18 19 20

Looking at this row one can see easily that the lowest (worst) score is 5, the highest (best) score is 20, and typical scores (in the middle) are 13–16. This is good enough for a start.

1.3 Terminology

A sequence of raw (unprocessed) observed data is called a *sample* and commonly denoted by

$$x_1,\ x_2,\ x_3,\ \dots,\ x_n \tag{\textbf{sample}}$$

Here n is the number of observed values, called the *size of the sample.*

An ordered sample is denoted by

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$$

so that $x_{(1)}$ is the smallest observed value, $x_{(2)}$ is the second smallest, etc., up to the largest value $x_{(n)}$.

In Example 1.1, $x_1 = 16$, but $x_{(1)} = 5$.

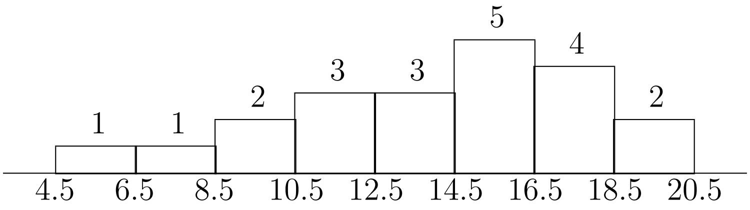
1.4 Frequency table

There are repetitions in our ordered data (for example, 16 appears three times). A more compact way to record the ordered data is a *frequency table*:

$$\begin{array}{r|l}
20 & | \\
19 & | \\
18 & || \\
17 & || \\
16 & ||| \\
\multicolumn{2}{c}{\cdots} \\
8 & | \\
5 & |
\end{array}$$

1.5 Histogram

Another way to visualize data is constructing a *histogram*:



Here the entire range (from 5 to 20) is divided into eight intervals (*bins*), and over each interval a histogram bar is drawn, of size proportional to the number of data points that fall into that bin. The bins are sometimes called *class intervals*, and the midpoint of each interval is its *class mark* (not shown here).

The choice of the number of bins and their positions (locations of the endpoints) is made by statisticians, and it takes experience to construct a histogram that better demonstrates principal features of the sample.

A histogram usually contains less information than the original sample. In the above example, scores 15, 15, 16, 16, 16 are combined into one (the tallest) bar. There is no way to tell how many 15's and 16's are, exactly, in the original sample, if we only see the above histogram. When constructing a histogram one faces a trade-off: shorter bins retain more detailed information about the original sample, but longer bins usually make the histogram more easily readable.

1.6 Numerical characteristics - 1

A sample can be characterized by certain numerical values (‘summaries’), such as the smallest and largest measurements. In our sample

$$\min = 5, \qquad \max = 20, \qquad \text{range} = 15$$

(the *range* is the difference between maximum and minimum).

The *sample median* is the middle point in the ordered sample. If its size is even (like $n = 20$, in our example), then the sample median is the average of the two middle points. In our case it is

$$\tilde{m} = \frac{14 + 15}{2} = 14.5 \qquad \textbf{(median)}$$

Let us further divide the ordered sample into four equal parts:

$$5\ \ 8\ \ 9\ \ 10\ \ 11 \ \Big|\ 11\ \ 12\ \ 13\ \ 13\ \ 14 \ \Big|\ 15\ \ 15\ \ 16\ \ 16\ \ 16 \ \Big|\ 17\ \ 18\ \ 18\ \ 19\ \ 20$$

The average value of the data points around the first and third division bars are called the *first* and *third quartiles*, respectively:

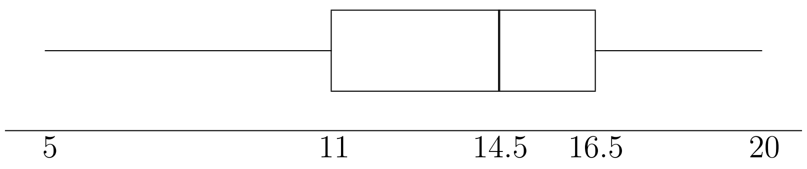
$$\tilde{q}_1 = \frac{11 + 11}{2} = 11, \qquad \tilde{q}_3 = \frac{16 + 17}{2} = 16.5 \qquad \textbf{(quartiles)}$$

(of course, the second quartile is the median). The *interquartile range* is

$$\text{IQR} = 16.5 - 11 = 5.5 \qquad \textbf{(IQR)}$$

1.7 Box-and-whisker diagram

It is common to consider the middle 50% of the data (between the first and third quartiles) as *typical values*, while the lower 25% and the higher 25% ends of it as unusual, extreme values. This is symbolized by a *box-and-whisker diagram* (or, simply, a *box plot*), whose meaning is quite clear:



The two middle boxes are usually short and ‘fat’, representing the bulk of the sample, while the arms (whiskers) are long and narrow.

1.8 Numerical characteristics - 2

Other important numerical characteristics of random samples are

sample mean: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

sample variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

sample standard deviation: $$s = \sqrt{s^2}$$

These are (statistical) analogues of probabilistic notions of *mean value*, *variance* and *standard deviation*, respectively.

The evaluation of the above quantities is quite laborious, it is best done with a computer or an advanced calculator. For our sample of students' scores they are

$$\bar{x} = 14, \qquad s^2 = \frac{307}{19} \approx 16, \qquad s \approx 4$$

One may wonder why the denominator in the formula for s^2 is $n-1$, and not simply n. Occasionally we do use n there (see Section 3.20), and the corresponding characteristic is denoted by

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

There is, however, a good reason to prefer $n-1$ over n (that is, s^2 over V), as it is explained below in Section 3.21.

1.9 Mode

The value that occurs most often in the sample is called the *mode.* In our sample of students' scores, it is 16 (it occurs three times).

1.10 Outliers

Observations that lie far from the rest of the sample are called *outliers.* In our sample of test scores, there are no obvious outliers, but the value 5 can be regarded as an outlier, to some extend.

Outliers may be produced by unusual (uncharacteristic) random events, or simply result from human errors in recording data or computer glitches in transferring data. Outliers are unwanted in common statistical practice, and certain methods are designed to filter them away.

2 Simulation (optional; skipped in 2014)

So, statistics deals with observed data. In practice, the data is collected from observations of real phenomena (students' scores, stock market index, weather records, etc.) or measured in experiments (weights of chocolate bars produced by a candy factory, size of fish caught in a lake, etc.).

Another way of obtaining data, suitable for teaching statistics and for theoretical research, is *computer simulation.* This means that computer generates random numbers, which can be treated and processed as experimentally observed data. Our course includes certain computer projects where the students use MATLAB to generate random numbers. Given a random variable X, MATLAB can produce n values $x_1, \dots, x_n$ of that variable.

2.1 Uniform $U(0,1)$ random variable

Nearly every computer language and software package includes a basic random number generator (RNG) that produces values of the uniform random variable $X = U(0,1)$. In MATLAB, the following commands invoke this basic RNG:

x=rand	returns one random number
x=rand(m,n)	an $m \times n$ matrix of random numbers

In the latter case **x** will be a matrix of m rows and n columns, consisting of random values of $X = U(0,1)$.

2.2 Remark

Theoretically, $X = U(0,1)$ may take any value in the interval $(0,1)$, and every particular value occurs with probability zero. Practically, computers can only handle finitely many special (*binary*) numbers, so the RNG returns only binary numbers between 0 and 1. Depending on the computer arithmetic, there are about 10^{10} to 10^{15} binary numbers between 0 and 1, so each one comes with a positive probability, and sooner or later they will start repeating themselves. (This fact should be kept in mind when doing computer experiments.) By convention, the RNG *is allowed* to return 0, but *not* 1.

2.3 Pseudo-random numbers and resetting RNG

Every computer RNG generates a sequence of random numbers following a specific algorithm. It always starts with the same number x_1, followed by the same number x_2, etc[1]. If you want the RNG to produce a different sequence, you need to change the *state* of the RNG, or 'reset' it. In MATLAB, the command **s=rand('state')** returns a 35-element vector containing the current state. Then you can modify it and reset the generator by using one of the following commands:

rand('state',s)	Resets the state to **s**
rand('state',0)	Resets the RNG to its initial state
rand('state',j)	Resets the RNG to its **j**-th state (here **j** is an integer),
rand('state',sum(100*clock))	Resets it to a different state each time

2.4 Inversion

This method is based on the following fact established in probability theory: if X is a continuous random variable with distribution function $F(x)$, then $Y = F(X)$ is a uniform $U(0, 1)$ variable. Conversely, if $Y = U(0, 1)$, then $X = F^{-1}(Y)$. Thus one can use random values of Y (produced by the basic RNG) and compute X by the formula $X = F^{-1}(Y)$.

This method requires the inverse function $F^{-1}(x)$, which is only available in a few simple cases (see the next three sections). In general, the computation of F^{-1} is prohibitively expensive, so the method is very inefficient.

2.5 Uniform $U(a, b)$ random variable

If $X = U(a, b)$ with arbitrary $a < b$, then its distribution function $F(x) = (x - a)/(b - a)$ has a simple inverse $F^{-1}(y) = a + (b - a)y$. Thus the value of X can be generated by $X = a + (b - a)Y$, where $Y = U(0, 1)$.

2.6 Exponential random variable

If X is an exponential random variable, then its distribution function is $F(x) = 1 - e^{-x/\mu}$ (here $\mu = 1/\lambda$ denotes the mean value of X). It has a simple inverse $F^{-1}(y) = -\mu \ln(1 - y)$. Thus the value of X can be generated by $X = -\mu \ln(1 - Y)$, where $Y = U(0, 1)$.

Note that $1 - Y$ here may take value 1, but not 0, see the last sentence in Section 2.2, thus ensuring the safety in the computation of $\ln(1 - Y)$.

[1] For this reason, the numbers returned by an RNG are not purely random; they are called *pseudo-random numbers*.

With MATLAB Statistical Toolbox, you don't need the above: exponential random variable can be generated directly by special commands

x=exprnd(u)	returns one random value of X
x=exprnd(u,m,n)	an $m \times n$ matrix of random numbers

where **u** denotes μ, the mean value of X.

2.7 Cauchy random variable

Cauchy random variable X has distribution function $F(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x$, hence its inverse is $F^{-1}(y) = \tan(\pi y - \pi/2)$. Thus the Cauchy variable X can be generated by $X = \tan(\pi Y - \pi/2)$, where $Y = U(0,1)$.

2.8 Special purpose: normal random variable

The following algorithm is commonly used to generate values of the standard normal random variable $Z = \mathcal{N}(0,1)$: one generates two values y_1, y_2 of the uniform variable $Y = U(0,1)$ and computes

$$\begin{aligned} z_1 &= \sqrt{-2 \ln y_1} \cdot \sin(2\pi y_2) \\ z_2 &= \sqrt{-2 \ln y_1} \cdot \cos(2\pi y_2) \end{aligned}$$

This gives two *independent* values of Z. If you only need one value, you can use either z_1 or z_2; but in practice we usually need a long sequence of independent values of Z, hence it is good to have two at once.

Having generated Z, the variable $X = \mathcal{N}(\mu, \sigma^2)$ can be obtained by $X = \mu + \sigma Z$.

In MATLAB, you don't need the above formulas: you can generate $Z = \mathcal{N}(0,1)$ by

x=randn	returns one random value of Z
x=randn(m,n)	an $m \times n$ matrix of random numbers

In addition, MATLAB Statistical Toolbox provides special commands to generate any normal random variable $X = \mathcal{N}(\mu, \sigma^2)$

x=normrnd(u,s)	returns one random value of X
x=normrnd(u,s,m,n)	an $m \times n$ matrix of random numbers

here **u** denotes μ and **s** denotes σ (not σ^2).

2.9 Rejection method

This is the most popular general-purpose algorithm. Suppose we want to generate a random variable X with density function $f(x)$. Often we *can* generate another random variable, Y, with density $g(x)$ (for example, Y may be uniform, or exponential, or normal), such that $Cg(x) \geq f(x)$ for some constant $C > 0$ and all x.

Then we generate a random value y of Y and accept it if

$$\frac{f(y)}{Cg(y)} \geq w$$

and reject it otherwise; here w is a (separately generated) random value of $W = U(0,1)$. The accepted values of Y are taken as random values of X.

This method requires two random values (one for Y and one for $W = U(0,1)$) per value of X, and some pairs of Y and W may be rejected. For better efficiency, the fraction of rejected values of Y should be small. It is known that the overall fraction of accepted values of Y is $1/C$. In other words, to generate n values of X one needs, approximately, Cn values of Y (plus Cn values of $W = U(0,1)$). So one wants to make c as small as possible. To achieve this, select Y whose density $g(x)$ is as similar to the density $f(x)$ as possible.

2.10 Bernoulli and binomial random variables

Generating discrete random variables requires special approaches. For example, a Bernoulli random variable X takes two values: 1 with probability p and 0 with probability $q = 1 - p$. So one can generate a random value y of $Y = U(0,1)$ and set

$$X = \begin{cases} 1 & \text{if } y < p \\ 0 & \text{otherwise} \end{cases}$$

To generate a binomial random variable $X = b(n,p)$ one can generate n independent Bernoulli random variables $X_1, \ldots, X_n$ and set $X = X_1 + \cdots + X_n$ (this is rather inefficient, though).

More generally, if a discrete random variable X takes values $x_1, x_2, \ldots$ with corresponding probabilities $p_1, p_2, \ldots$, then one can generate a random value y of $Y = U(0,1)$ and set $X = x_n$ where $n \geq 1$ is the first (smallest) integer such that

$$y < p_1 + \cdots + p_n.$$

This method is quite efficient and only requires one call of the basic RNG.

3 Maximum likelihood estimation (MLE)

3.1 Lottery example

Suppose a person buys 45 lottery tickets in a student fair, and 15 of them win. What is the fraction of winning tickets in this lottery?

Let us describe this example in probabilistic terms. Denote the fraction of winning tickets by p. Then each ticket wins with probability p. If someone buys 45 tickets, then the number of winning tickets is a binomial random variable $X = b(45, p)$. From probability theory, we know that

$$\mathbb{P}(X = 15) = \binom{45}{15} p^{15} (1-p)^{30}$$

Note that the value of the random variable (that is, 15) is known, but the parameter p is unknown.

3.2 Statistics versus probability theory

In probability theory, random variables are usually completely specified and their parameters known; the main goal is to compute probabilities of random values that the variables can take.

In statistics, the situation is opposite. The *values* of random variables are known (observed), but their theoretical characteristics (such as types and/or parameters) are unknown or only partially known. The goal is to determine the unknown theoretical characteristics of random variables by observing and analyzing their values.

In this sense, probability theory and statistics are "opposite" (better to say, complementary) to each other (like derivatives and integrals in calculus).

3.3 Lottery example continued

Intuitively, the fraction of winning tickets appears to be 1/3. Of course, one can never be sure: the person who bought 45 tickets may be very lucky (the real fraction may be much smaller) or very unlucky (the real fraction may be much larger). Nonetheless, 1/3 seems to be the best (most appropriate) estimate of p, see explanations in Section 3.5.

3.4 Unknown parameters versus estimates

The unknown parameter p cannot be precisely determined, unless one buys *all* the lottery tickets. In statistics, unknown parameters can only be *estimated*. The value 1/3 presented in Section 3.3 is just our *guess*. To distinguish the unknown parameter p from its estimate, we denote the latter by $\hat{p}$, so in our example, $\hat{p} = 1/3$ (while the value of p remains unknown).

3.5 Lottery example continued

So why is our estimate $\hat{p} = 1/3$ the best? Is there any argument to support this choice? Yes, here is the argument.

The probability

$$\mathbb{P}(X = 15) = \binom{45}{15} p^{15} (1-p)^{30}$$

gives the likelihood of the value $X = 15$. In this formula, p is an unknown quantity, a variable, so we can treat it as a function of p:

$$L(p) = \binom{45}{15} p^{15} (1-p)^{30}$$

which is called the *likelihood function.* It achieves its maximum (see below) at the point $p = 1/3$. This value of p is the most likely, or most probable. Since we select an estimate of p by maximizing the likelihood function $L(p)$, it is called the *maximum likelihood estimate* (MLE).

3.6 Computation of $\hat{p}$

To find the maximum of $L(p)$ it is convenient to take its logarithm:

$$\ln L(p) = \ln \binom{45}{15} + 15 \ln p + 30 \ln(1-p)$$

(this is called the *log-likelihood function*), and then differentiate it:

$$\frac{d}{dp} \ln L(p) = \frac{15}{p} - \frac{30}{1-p}$$

The maximum is achieved at the point where $\frac{d}{dp} \ln L(p) = 0$, thus we get equation

$$\frac{15}{p} - \frac{30}{1-p} = 0$$

Solving it yields our estimate $\hat{p} = 1/3$.

3.7 MLE for general binomials

More generally, let $X = b(n, p)$ be a binomial random variable with known n but unknown p, and x an observed value of X. Then

$$\mathbb{P}(X = x) = \binom{n}{x} p^x \, (1-p)^{n-x}$$

and we consider this probability as the likelihood function $L(p)$, where x and n are known, but p is unknown. The log-likelihood function is

$$\ln L(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p)$$

and its derivative is

$$\frac{d}{dp} \ln L(p) = \frac{x}{p} - \frac{n-x}{1-p}$$

This derivative equals zero at the point $p = x/n$, hence the MLE is

$$\hat{p} = \frac{x}{n}$$

3.8 Mean value of the MLE

We note that x is a random value of the variable X, hence $\hat{p} = X/n$ is also a random variable. As a random variable, $\hat{p}$ has a distribution and all relevant characteristics: mean value, variance, etc. Its mean value is

$$\mathbb{E}(\hat{p}) = \frac{\mathbb{E}(X)}{n} = \frac{np}{n} = p$$

It is remarkable that the mean value of our estimate $\hat{p}$ coincides with the unknown parameter p that we are estimating. Thus, on average, the estimate $\hat{p}$ is just right – it is precise, there is no systematic error (bias).

3.9 Unbiased estimates

An estimate $\hat{\theta}$ of an unknown parameter θ is called *unbiased* if its mean value coincides with θ:

$$\mathbb{E}(\hat{\theta}) = \theta$$

If the estimate is biased, then the difference

$$\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

is called the *bias* of $\hat{\theta}$.

3.10 Variance of $\hat{p}$

The variance of our estimate $\hat{p}$ is

$$\mathsf{Var}(\hat{p}) = \frac{\mathsf{Var}X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}$$

where $q = 1 - p$, and its standard deviation is

$$\sigma_{\hat{p}} = \frac{\sqrt{pq}}{\sqrt{n}}$$

This means that typical (expected) error $\hat{p} - p$ is about $\sqrt{pq}/\sqrt{n}$, hence

$$\hat{p} \approx p \pm \frac{\sqrt{pq}}{\sqrt{n}}$$

In our example, expected error is

$$\hat{p} - p \approx \pm \frac{\sqrt{1/3 \cdot 2/3}}{\sqrt{45}} = \pm 0.07$$

so the true (unknown) value of p may differ from our estimate $\hat{p} = 1/3$ by about 0.07 (on average).

3.11 Mean Square Error (MSE)

A common measure of accuracy of an estimate $\hat{\theta}$ of an unknown parameter θ is the *mean squared error*

$$\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2$$

For an unbiased estimate, $\theta = \mathbb{E}(\hat{\theta})$, so then $\text{MSE}(\theta) = \mathsf{Var}(\hat{\theta})$. For biased estimates, we have the following simple decomposition:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 + [\mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \mathsf{Var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2\end{aligned}$$

thus the contributions from the variance and from the bias get separated. In practice, the bias is usually zero or negligible, so the main source of errors is $\mathsf{Var}(\hat{\theta})$. The typical (expected) error $|\hat{\theta} - \theta|$ is given by

$$\sqrt{\text{MSE}(\hat{\theta})} = \text{(for unbiased only)} = \sigma_{\hat{\theta}},$$

which is called the *root-mean-squared error*.

3.12 Likelihood function for exponentials

Suppose we want to estimate the average lifetime of light bulbs produced by a factory (a common problem in *quality control*). To this end, one randomly picks n light bulbs, turns them on until every bulb burns down, and records their lifetimes $x_1, x_2, \dots, x_n$. Now how to estimate the average lifetime? Would the sample mean $\bar{x}$ be a reasonable estimate?

In probability theory we learned that the lifetime can be fairly accurately modeled by an exponential random variable, which has density function $f(x) = \lambda e^{-\lambda x}$, and $\lambda > 0$ represents the (unknown) parameter. The average lifetime is $\mathbb{E}(X) = 1/\lambda$. Since we need the value of $\mathbb{E}(X)$, rather than λ, we change parameter: denote $\mu = \mathbb{E}(X) = 1/\lambda$, and accordingly replace λ with $1/\mu$. The formula for the density function becomes $f(x) = \mu^{-1}e^{-x/\mu}$.

In our experiment, we obtained n random values $x_1, \dots, x_n$ of X. Since they are obtained independently, their joint density function is

$$L(\mu) = f(x_1) \cdots f(x_n) = \mu^{-n} e^{-\frac{x_1 + \cdots + x_n}{\mu}}$$

The joint density function gives the probability, or likelihood, of the values $x_1, \dots, x_n$. Since the only unknown quantity here is μ, we get a function of μ and call it the likelihood function.

3.13 MLE for exponentials

To find the MLE estimate of μ, we follow the same steps as in Section 3.7. First, we take the logarithm of the likelihood function

$$\ln L(\mu) = -n \ln \mu - \frac{x_1 + \cdots + x_n}{\mu}$$

Then we differentiate it with respect to the unknown parameter

$$\frac{d}{d\mu} \ln L(\mu) = -\frac{n}{\mu} + \frac{x_1 + \cdots + x_n}{\mu^2}$$

Setting the derivative to zero we arrive at equation

$$\frac{n}{\mu} = \frac{x_1 + \cdots + x_n}{\mu^2}$$

Solving it gives

$$\hat{\mu} = \frac{x_1 + \cdots + x_n}{n} = \bar{x}$$

Hence the MLE for the average lifetime μ is, indeed, the sample mean.

3.14 Mean value and variance of $\hat{\mu}$

The mean value of $\hat{\mu}$ is

$$\mathbb{E}(\hat{\mu}) = \frac{n\,\mathbb{E}(X)}{n} = \mathbb{E}(X) = \mu$$

hence the estimate is unbiased. Its accuracy is characterized by its variance

$$\mathsf{Var}(\hat{\mu}) = \frac{n\,\mathsf{Var}(X)}{n^2} = \frac{1}{\lambda^2 n} = \frac{\mu^2}{n}$$

so the typical (expected) error $|\hat{\mu} - \mu|$ will be $\sigma_{\hat{\mu}} = \mu/\sqrt{n}$.

3.15 Remark

It is common in statistics that typical errors of estimates are proportional to $1/\sqrt{n}$, where n is the size of the sample. A common rule of thumb is that the expected error is $\sim 1/\sqrt{n}$. Hence to get the error ~ 0.1 one needs to collect $n = 100$ data; to get the error ~ 0.01 one needs $n = 10,000$ data, etc.

3.16 Estimating λ for exponentials

The MLE estimate of the parameter $\lambda = 1/\mu$ will be

$$\hat{\lambda} = \frac{1}{\hat{\mu}} = \frac{n}{x_1 + \cdots + x_n} = \bar{x}^{-1}$$

This estimate (unlike $\hat{\mu}$) is biased, but its bias is quite hard to compute.

3.17 General scheme

Summarizing the above examples, we describe a general scheme for evaluating a maximum likelihood estimate (MLE) for a parameter θ.

Suppose a random variable X has density function $f(x;\theta)$ that depends on an unknown parameter θ (if X is a discrete variable, we need to use its probability density function $f(x;\theta) = \mathbb{P}(X = x)$). Let $x_1, \dots, x_n$ be a random sample of n (independently) obtained values of X.

Step 1. Write down the likelihood function

$$L(\theta) = f(x_1;\theta) \cdots f(x_n;\theta)$$

Step 2. Write down the log-likelihood function

$$\ln L(\theta) = \ln f(x_1;\theta) + \cdots + \ln f(x_n;\theta)$$

Step 3. Differentiate the log-likelihood function and set it to zero

$$\frac{d}{d\theta} \ln L(\theta) = 0$$

Step 4. Solve the above equation for θ, the solution will be $\hat{\theta}$.

3.18 MLE for geometric random variable

A match factory wants to determine the quality of their matches. Ideally, a match should light up on the first strike. But in reality, it may fail and require more than one strike. The factory needs to determine the average number of strikes it takes to light up a match.

In an experiment, n matches are chosen randomly and a technician strikes them until they light up and records the numbers of strikes $x_1, \dots, x_n$ it takes. Now how does he determine the average number of strikes? By taking the sample mean?

Striking a match is a sequence of trials till the first success. This is modelled by a geometric random variable X, which has one (unknown) parameter – the probability of success p. Its mean value is $\mathbb{E}(X) = 1/p$. The probability density function is

$$f(x;p) = \mathbb{P}(X = x) = pq^{x-1} \qquad \text{for} \;\; x = 1, 2, \dots$$

where $q = p - 1$. Now the likelihood function is

$$L(p) = pq^{x_1-1} \cdots pq^{x_n-1} = p^n q^{x_1+\cdots+x_n-n}$$

Its logarithm is

$$\ln L(p) = n \ln p + (x_1 + \cdots + x_n - n) \ln(1 - p)$$

and its derivative

$$\frac{d}{dp} \ln L(p) = \frac{n}{p} - \frac{x_1 + \cdots + x_n - n}{1 - p}$$

Solving the equation

$$\frac{n}{p} - \frac{x_1 + \cdots + x_n - n}{1 - p} = 0$$

gives

$$\hat{p} = \frac{n}{x_1 + \cdots + x_n} = \bar{x}^{-1}$$

Thus the MLE for the average $\mathbb{E}(X) = 1/p$ is, indeed, $1/\hat{p} = \bar{x}$.

This result makes good sense: p is the probability of success, the numerator of this fraction is the total number of observed successes (strikes in which the match lights up), and the denominator is the total number of strikes.

This estimate is biased, i.e. $\mathbb{E}(\hat{p}) \neq p$. However, in the limit $n \to \infty$ the fraction of successes $\hat{p}$ approaches the probability of success p, according to the Law of Large Numbers, hence $\hat{p} \to p$ as $n \to \infty$.

3.19 Consistent estimates

An estimate $\hat{\theta}$ of an unknown parameter θ is *consistent* if $\hat{\theta} \to \theta$ as $n \to \infty$ in the probabilistic sense. Precisely, for any small positive number $y > 0$ we must have

$$\mathbb{P}\big(|\hat{\theta} - \theta| > y\big) \to 0 \qquad \text{as} \;\; n \to \infty$$

that is the probability of any deviations of $\hat{\theta}$ from θ vanishes in the limit $n \to \infty$.

Most of the estimates used in practice are consistent, even if they are biased. All MLE estimates are consistent.

3.20 MLE for normals

Suppose biologists want to describe the length of a certain breed of fish (say, salmon). The length of a fish is a random quantity affected by many factors, which are essentially independent. The central limit theorem in probability says that random quantities resulting from many independent factors have approximately normal distribution. Their densities are bell-shaped curves – peaking in the middle (at the most typical value) and decaying symmetrically to the left and right. This principle is almost universal – most random quantities in nature and human society have approximately normal distributions.

A normal random variable $X = \mathcal{N}(\mu, \sigma^2)$ has two parameters μ (the mean value) and σ^2 (the variance). To describe the size of fish completely, the biologists need to determine the values of both μ and σ^2. Suppose they catch n fish randomly and measure their sizes $x_1, \ldots, x_n$. How should they estimate μ and σ^2? By the sample mean $\bar{x}$ and the sample variance s^2?

The probability density function of a normal random variable is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For convenience, we replace σ^2 with θ:

$$f(x;\mu,\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-\frac{(x-\mu)^2}{2\theta}}$$

The likelihood function is

$$L(\mu,\theta) = \prod_{i=1}^n f(x_i;\mu,\theta) = \frac{1}{(2\pi\theta)^{n/2}}\, e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}}$$

Its logarithm is

$$\ln L(\mu,\theta) = -\frac{n}{2}\ln(2\pi\theta) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\theta}$$

There are *two* parameters here, and thus we need to take two *partial derivatives* (with respect to both parameters) and set them to zero:

$$\frac{d}{d\mu}\ln L(\mu,\theta) = \frac{1}{\theta}\sum_{i=1}^n (x_i-\mu) = 0$$

and

$$\frac{d}{d\theta}\ln L(\mu,\theta) = -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^n (x_i-\mu)^2 = 0$$

From the first equation we obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x} \tag{MLE-1}$$

Substituting this into the second equation and solving it for $\theta = \sigma^2$ gives

$$\hat{\theta} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 = V \tag{MLE-2}$$

(the quantity V was introduced in Section 1.8).

Thus, the MLE for the average μ is, indeed, the sample mean $\bar{x}$. But the MLE for the variance σ^2 is (surprisingly) *not* the sample variance s^2. This is one of many strange things that happen in statistics. The way statisticians deal with it is quite instructive, we will see it next.

3.21 Bias of MLE's for normals

The mean value of $\hat{\mu}$ is

$$\mathbb{E}(\hat{\mu}) = \frac{n\,\mathbb{E}(X)}{n} = \mathbb{E}(X) = \mu$$

thus the estimate $\hat{\mu}$ is unbiased. Its variance is

$$\mathsf{Var}(\hat{\mu}) = \frac{n\,\mathsf{Var}(X)}{n^2} = \frac{\sigma^2}{n}$$

thus the typical error in estimating μ is $\sigma/\sqrt{n}$.

Is $\hat{\sigma}^2 = V$ also unbiased? To compute its mean value we first simplify

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= \sum(x_i^2 - 2\bar{x}x_i + \bar{x}^2)\\ &= \sum x_i^2 - 2\bar{x}\bar{x}n + n\bar{x}^2\\ &= \sum x_i^2 - n\bar{x}^2\end{aligned}$$

Now

$$\begin{aligned}\mathbb{E}\Big[\sum(x_i - \bar{x})^2\Big] &= \mathbb{E}\Big[\sum x_i^2\Big] - n\mathbb{E}\big(\bar{x}^2\big)\\ &= n\mathbb{E}(X^2) - n\big(\mathsf{Var}(\bar{x}) + [\mathbb{E}(\bar{x})]^2\big)\\ &= n\big(\mathsf{Var}(X) + [\mathbb{E}(X)]^2\big) - n\big(\tfrac{1}{n}\,\mathsf{Var}(X) + [\mathbb{E}(X)]^2\big)\\ &= (n-1)\,\mathsf{Var}(X)\end{aligned}$$

We used the facts $\mathbb{E}(X^2) = \mathsf{Var}(X) + [\mathbb{E}(X)]^2$ and $\mathbb{E}(\bar{x}) = \mathbb{E}(X)$ and $\mathsf{Var}(\bar{x}) = \frac{1}{n}\mathsf{Var}(X)$ established in probability theory.

So we conclude that

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}(V) = \frac{n-1}{n}\,\mathsf{Var}(X) = \frac{n-1}{n}\,\sigma^2$$

Since $\mathbb{E}(\hat{\sigma}^2) \neq \sigma^2$, the MLE estimate $\hat{\sigma}^2$ is *biased.*

3.22 Unbiased version of $\hat{\sigma}^2$

While the bias of $\hat{\sigma}^2 = V$ is small, $\mathbb{E}(\hat{\sigma}^2) - \sigma^2 = -\sigma^2/n$, it is annoying and many statisticians consider it unacceptable. The bias can be eliminated by multiplying $\hat{\sigma}^2$ with $\frac{n}{n-1}$, and one arrives at a new estimate of $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$$

which is the sample variance introduced in Section 1.8. This estimate of σ^2 is unbiased. For this reason it is commonly used in practice instead of the MLE estimate $\hat{\sigma}^2 = V$ obtained in Section 3.20.

We see that most statisticians are willing to sacrifice the theoretical principle of maximum likelihood estimation (albeit only slightly) in order to achieve the convenience of an unbiased estimate.

3.23 Remark

The estimates $\bar{x}$ and s^2, though obtained from the same sample, are *independent*. This means that the value of one does not imply anything about the value of the other. Suppose s^2 happens to be very small, i.e. the values $x_1, \dots, x_n$ are close to each other. You might think that you are lucky – errors are small, so the sample mean $\bar{x}$ would be very close to the actual mean μ... This isn't the right assumption, there is no relation between the value of s^2 and the error of the estimate $\bar{x}$.

3.24 General estimates of $\mathbb{E}(X)$ and $\mathsf{Var}(X)$

Given a sample $x_1, \dots, x_n$ of values of a random variable X (not necessarily normal) it is often desirable to estimate its mean value $\mathbb{E}(X)$ and variance $\mathsf{Var}(X)$.

One can use the sample mean $\bar{x}$ to estimate $\mathbb{E}(X)$, and this estimate will always be unbiased, since $\mathbb{E}(\bar{x}) = \mathbb{E}(X)$.

Next, one can use V or s^2 (see 1.8) to estimate $\mathsf{Var}(X)$. Both estimates are good, but there is a little difference: the estimate V is biased while s^2 is unbiased. Indeed, our calculations in Section 3.21 are valid for *any* random variable X (not only for normals), thus

$$\mathbb{E}(V) = \frac{n-1}{n}\,\mathsf{Var}(X), \qquad \mathbb{E}(s^2) = \mathsf{Var}(X)$$

For this reason, statisticians always prefer s^2 over V.

4 Percentiles

We learned how to estimate the value of an unknown parameter θ. We even evaluated typical errors of an estimate $\hat{\theta}$. For example, in Section 3.10 we found that our estimate $\hat{p} = 1/3$ differed by about 0.07 from the true (unknown) value of p.

However, in practice it is not enough to just say that "typical errors" are ~ 0.07. Suppose an investment company buys stocks that return high dividends with probability at least 20%. They consider stocks that have a short history of returning high dividends with probability 33%, but this estimate is based on limited data and involves a typical error of 7%. So should the company buy these stocks? The actual error may exceed 7% and be as high as 13% or 14%. What is the chance that the error exceeds 13%? Can it be guaranteed, with some level of confidence, that the error stays below a certain value?

Such questions arise in economical applications, they are essential for insurance and warranty purposes. In fact, no serious application of statistics should ignore such questions. To answer them, we will need percentiles.

4.1 Percentiles in probability

Let X be a random variable with distribution function $F(x)$ and density function $f(x)$. For every $0 < p < 1$ the *quantile* (or *percentile*) π_p is such a number that $F(\pi_p) = p$. In terms of the density function $f(x)$

$$\int_{-\infty}^{\pi_p} f(x)\,dx = p, \qquad \int_{\pi_p}^{\infty} f(x)\,dx = 1 - p$$

i.e. the real line is divided by the point π_p into two parts that capture the probabilities p (to the left of π_p) and $1 - p$ (to the right of π_p).

Note that $\pi_{1/2} = m$ (median), $\pi_{1/4} = q_1$ (first quartile) and $\pi_{3/4} = q_3$ (third quartile).

4.2 Percentiles for normals

Let $Z = \mathcal{N}(0, 1)$ be a standard normal random variable. Denote its distribution function by $\Phi(x)$ and density function by $f(x)$. For every $0 < \alpha < 1$, the quantity

$$z_\alpha = \pi_{1-\alpha}$$

is frequently used in statistics. This means that $\Phi(z_\alpha) = 1 - \alpha$, as well as

$$\int_{-\infty}^{z_\alpha} f(x)\,dx = 1 - \alpha, \qquad \int_{z_\alpha}^{\infty} f(x)\,dx = \alpha$$

in other words, z_α divides the real line into two parts that capture the probabilities $1 - \alpha$ (to the left of z_α) and α (to the right of z_α).

The bottom part of Table Va (on page 584) gives the values of z_α for $\alpha = 0.4$, 0.3, 0.2, 0.1, 0.05, etc. For example, $z_{0.1} = 1.282$.

Due to the symmetry of the standard normal distribution, we have

$$\mathbb{P}(Z < -z_\alpha) = \mathbb{P}(Z > z_\alpha) = \alpha, \qquad \mathbb{P}\big(|Z| > z_{\alpha/2}\big) = \alpha$$

The very bottom line in Table Va gives the values of $z_{\alpha/2}$ for certain α's.

Hence, percentiles z_α and $z_{\alpha/2}$ allow us to "chop-off" tails of the standard normal distribution containing the given amount of probability:

- Right tail of probability α;
- Left tail of probability α;
- Two equal tails (one on each side) of combined probability α

Chopping off correct tails will be necessary for the error analysis in the next section.

5 Confidence intervals for normals: one mean

5.1 Estimating μ with known σ^2

Suppose a doctor is checking a patient's blood pressure by using a monitor that is not very reliable, its readings are known to have typical errors of ±10 points. The doctor measures the pressure n times and gets values $x_1, \dots, x_n$. The best estimate of the patient's blood pressure is, apparently, the sample mean $\bar{x}$, but the doctor needs a certain range (interval) where the unknown blood pressure is guaranteed to be with a high probability (say, 99%). Then a statistical analysis is necessary.

The readings of the blood pressure monitor are affected by many independent factors, so by the central limit theorem they are approximately normal, $X = \mathcal{N}(\mu, \sigma^2)$. Here the average μ represents the unknown blood pressure, and the standard deviation $\sigma = 10$ is the known typical error.

The MLE for μ is $\mu = \bar{x}$, but now we want to find an interval (a, b) such that the unknown value of μ will be guaranteed to be in (a, b) with probability $\geq 1 - \alpha$, where the probability is specified, for example, $1 - \alpha = 0.9$ or 0.95 or 0.99, etc.

The estimate $\hat{\mu} = \bar{x}$ has normal distribution with mean $\mathbb{E}(\bar{x}) = \mathbb{E}(X) = \mu$ and variance $\mathsf{Var}(\bar{x}) = \frac{1}{n}\mathsf{Var}(X) = \sigma^2/n$. Therefore, the variable

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is standard normal (has zero mean and variance equal to one). Since

$$\mathbb{P}(|Z| \leq z_{\alpha/2}) = 1 - \alpha$$

the following inequalities will be guaranteed with probability $1 - \alpha$:

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

Solving these inequalities for μ we obtain

$$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$$

This gives an interval where μ is guaranteed to belong with probability $1 - \alpha$. It is called *confidence interval*, and $1 - \alpha$ is called *confidence level*, or *confidence coefficient*.

Note that the interval is symmetric about the estimate $\hat{\mu} = \bar{x}$. For brevity, symmetric intervals may be denoted by

$$\text{CI } = \bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

where one only indicates the center and half-length of the interval.

5.2 Example

Given observed values $x_1, \dots, x_{15}$ of a normal random variable $\mathcal{N}(\mu, 9)$, such that $\bar{x} = 20$, construct a confidence interval for μ with 95% confidence level.

Solution: Note that the value $\sigma^2 = 9$ is given, so $\sigma = 3$. Here $\alpha = 0.05$, hence $\alpha/2 = 0.025$. The percentile $z_{0.025} = 1.960$ is taken from Table Va. The confidence interval is

$$\text{CI } = 20 \pm 1.960 \cdot 3/\sqrt{15} = 20 \pm 1.5182$$

It can also be presented as CI$= [18.4818, 21.5182]$.

5.3 General scheme

Here is a general scheme for constructing confidence intervals. Let $\hat{\theta}$ be an estimate of an unknown parameter θ. Every estimate is a random variable, so it has a certain distribution. That distribution surely depends on θ, on the sample size n, and possibly on some other factors (other parameters). In the above case, $\hat{\mu} = \mathcal{N}(\mu, \sigma^2/n)$.

We need to transform the random variable $\hat{\mu}$ into some other random variable whose distribution is independent of θ and other parameters, and preferably of n as well. Denote new random variable by $Y(\hat{\theta}, \theta, n, \dots)$, where $\dots$ stand for some other parameters. In the above case, $Y = \sqrt{n}(\hat{\mu} - \mu)/\sigma$, and its distribution is standard normal.

Now since the distribution of Y is independent of anything, its percentiles y_α can be pre-computed and tabulated. Then the following inequalities will be guaranteed with probability $1 - \alpha$:

$$-y_{\alpha/2} \le Y(\hat{\theta}, \theta, n, \dots) \le y_{\alpha/2}$$

All we do now is solve these inequalities for θ to obtain a confidence interval with level $1 - \alpha$.

Note: it often happens that the distribution of Y depends on the sample size n (there may be no way to get rid of that dependence), then the corresponding percentiles should be tabulated for *every* n. In practice, this is done only for small $n \le 30$, while for $n > 30$ one resorts to various approximations.

5.4 Estimating μ with unknown σ^2

Suppose a doctor is checking a patient's blood pressure by using an unreliable monitor whose readings involve totally unknown errors. Despite this new complication, the doctor still wants to find a certain range (interval) where the unknown blood pressure is guaranteed to be with a high probability (say, 99%). Is it even possible, without knowing typical errors of the monitor? Yes, one simply should use the observed values $x_1, \dots, x_n$ to *estimate* the unknown reliability of the monitor. Of course, the analysis will be more complicated than before.

First of all, the previous method would not work, because the value of σ is not available. We can try to replace it with its estimate $\hat{\sigma} = s$ (sample standard deviation). For large n (precisely, for $n > 30$), this approximation is considered to be accurate enough, and we obtain the confidence interval

$$\text{CI } = \bar{x} \pm z_{\alpha/2} s/\sqrt{n}$$

For small n (that is, for $n \le 30$), we need to be more accurate. The crucial quantity here is

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

It has a distribution independent of μ and σ, but dependent on n. This distribution is called *t distribution* (or *Student's t distribution*), and the quantity T itself is called a *t random variable* (this is why we denote it by T). Its distribution depends on n and is characterized by $r = n - 1$, which is called *the number of degrees of freedom.*

The t random variable with $r \ge 1$ degrees of freedom has density

$$f(x) = \frac{\text{const}}{\bigl(1 + x^2/r\bigr)^{(r+1)/2}}$$

which is an even function, and its graph is a bell-shaped curve (generally looking like the density of $Z = \mathcal{N}(0,1)$, but it has heavier tails and a lower peak). Its percentiles are denoted by $t_\alpha(r)$ (analogously to z_α), where r stands for the number of degrees of freedom. Due to the symmetry we have

$$\mathbb{P}\bigl(T < -t_\alpha(r)\bigr) = \mathbb{P}\bigl(T > t_\alpha(r)\bigr) = \alpha, \qquad \mathbb{P}\bigl(|T| > t_{\alpha/2}(r)\bigr) = \alpha$$

Thus, the following inequalities will be guaranteed with probability $1 - \alpha$:

$$-t_{\alpha/2}(r) \le \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2}(r)$$

Solving these inequalities for μ we obtain

$$\text{CI } = \bar{x} \pm t_{\alpha/2}(r)\, s/\sqrt{n}$$

Remember that $r = n - 1$. The values of $t_\alpha(r)$ for various α (and all $r \le 30$) are given in Table VI on page 586.

Note: the t-random variable X with r degrees of freedom has mean $\mathbb{E}(X) = 0$ and variance $\mathsf{Var}(X) = r/(r-2)$ for $r \ge 3$.

5.5 Example

Given observed values $x_1, \dots, x_{15}$ of a normal random variable $\mathcal{N}(\mu, \sigma^2)$, such that $\bar{x} = 20$ and $s^2 = 9$, construct a confidence interval for μ with 95% confidence level.

Solution: Since σ^2 is not given, we assume it is unknown. Since $n = 15$ is small ($\le$ 30), we use the t percentile $t_{0.025}(14) = 2.145$. The confidence interval is

$$\text{CI } = 20 \pm 2.145 \cdot 3/\sqrt{15} = 20 \pm 1.6615$$

This is a longer interval than the one in Section 5.2, even though the same numbers were used. Why? Here we only have an estimate of σ^2, instead of its exact value, thus we have less information than we had in Section 5.2, so our errors are larger.

5.6 Remark

The percentiles in Table VI decrease as $r = n - 1$ grows. As a result, confidence intervals get shorter as n (the sample size) increases. For $n = \infty$ (the bottom row), Table VI gives the same percentiles as Table Va for normals, i.e. $t_\alpha(\infty) = z_\alpha$ for every α.

5.7 Summary

We have covered three distinct cases:

- σ is known; then CI$= \bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$;
- σ is unknown and $n > 30$; then CI$= \bar{x} \pm z_{\alpha/2}s/\sqrt{n}$;
- σ is unknown and $n \le 30$; then CI$= \bar{x} \pm t_{\alpha/2}(n-1)\, s/\sqrt{n}$.

5.8 One-sided confidence intervals

It is sometimes desirable in practice to give a one-sided bound on μ: for instance, guarantee that $\mu > a$ with a certain probability. In this case, instead of chopping off two tails, each of probability $\alpha/2$, one chops off one tail of probability α. For instance, the lower bound on μ with a given confidence level $1-\alpha$ will be

- σ is known; then $\mu \geq \bar{x} - z_\alpha \sigma/\sqrt{n}$;
- σ is unknown and $n > 30$; then $\mu \geq \bar{x} - z_\alpha s/\sqrt{n}$;
- σ is unknown and $n \leq 30$; then $\mu \geq \bar{x} - t_\alpha(n-1)\, s/\sqrt{n}$.

The upper bound is obtained similarly.

5.9 Example

Candy bars produced by a factory must weigh at least 50 grams. A random sample of $n = 100$ candy bars yielded $\bar{x} = 51$ and $s^2 = 0.5$. Estimate μ from below with probability 99%.

Solution: Here $\alpha = 0.01$, σ^2 is unknown, and $n > 30$. Therefore we use $z_{0.01} = 2.326$ and obtain

$$\mu \geq 51 - 2.326 \cdot \sqrt{0.5}/\sqrt{100} = 51 - 0.164 = 50.836$$

Thus, the average weight of a candy bar is guaranteed to be at least 50.836 grams with probability 99%.

5.10 Generating t-random variable in MATLAB

With MATLAB Statistical Toolbox, you can generate values of a t random variable by special commands

x=trnd(v)	returns one random value of t
x=trnd(v,m,n)	an $m \times n$ matrix of random numbers

where **v** denotes the number of degrees of freedom.

6 Confidence intervals for normals: two means

Suppose two classes take the same test (say, in Calculus-I). They were taught by different professors, and/or used different textbooks. The university officials want to determine how significant the difference in their scores is. Again, a student's score is a random quantity affected by many factors, so it is approximately a normal random variable. For the first class, it is $X = \mathcal{N}(\mu_X, \sigma_X^2)$ and for the second class $Y = \mathcal{N}(\mu_Y, \sigma_Y^2)$. The university officials want estimate the difference between μ_X and μ_Y. The actual students' scores $x_1, \dots, x_n$ and $y_1, \dots, y_m$ are random values of these two normal variables. Note that the class sizes (n and m) may be different.

6.1 Point estimate We estimate μ_X by the sample mean $\bar{x}$ of the x-values and μ_Y by the sample mean $\bar{y}$ of the y-values. Thus we can estimate $\mu_X - \mu_Y$ by $\bar{x} - \bar{y}$. Next we want to construct a confidence interval for $\mu_X - \mu_Y$.

6.2 Both sigmas are known

First we consider the (rather unrealistic) case where both variances σ_X^2 and σ_Y^2 are known. Recall from probability theory that $\bar{x}$ is a normal random variable $\mathcal{N}(\mu_X, \sigma_X^2/n)$. Similarly, $\bar{y}$ is a normal random variable $\mathcal{N}(\mu_Y, \sigma_Y^2/m)$. Therefore

$$\bar{x} - \bar{y} = \mathcal{N}(\mu_X - \mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$$

(note: when subtracting two independent normal random variables, we add their variances). Hence

$$\mu_X - \mu_Y - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \bar{x} - \bar{y} < \mu_X - \mu_Y + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

with probability $1 - \alpha$. Solving these inequalities for $\mu_X - \mu_Y$ gives

$$\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} < \mu_X - \mu_Y < \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

Suppose we know the variances σ_X^2 and σ_Y^2. Then the above formula gives the two-sided confidence interval for $\mu_X - \mu_Y$:

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

6.3 Both sigmas are unknown, and the sizes are large

If the variances σ_X^2 and σ_Y^2 are unknown, then the above formula cannot be used. We may replace the unknown variances with their best estimates, sample variances s_x^2 and s_y^2, respectively. This will be good enough, if both m and n are large (greater than 30). Then the confidence interval will be

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}.$$

6.4 Example

The scores of 200 students in the final exam in Calculus-I in class A yielded $\bar{x} = 81$ and $s_x^2 = 19$. In the same test, the scores of 1000 students in all the other classes yielded $\bar{y} = 79$ and $s_y^2 = 16$. Construct a 98% confidence interval for $\mu_X - \mu_Y$.

Solution: since both samples are large enough (200 and 1000 values), we use the formula from the previous section:

$$2 \pm 2.326\sqrt{\frac{19}{200} + \frac{16}{1000}} = 2 \pm 0.745.$$

So it is safe to say that the scores in class A are at least 1.255 points above the average in all the other classes (and at most 2.745 points).

6.5 Both sigmas are unknown, and the sizes are small

If the variances σ_X^2 and σ_Y^2 are unknown and *at least* one size (m or n) is small (30 or less), then the simple method of Section 6.3 will not be acceptable.

In that case the construction of confidence intervals is complicated. There are two cases. First, sometimes it is know that $\sigma_X^2 = \sigma_Y^2$ (but its value is unknown). For example, x's and y's may be experimental measurements of different objects made by the same tool (gauge), whose accuracy is the same in both measurements (but we don't know that accuracy).

In this case the confidence interval for $\mu_X - \mu_Y$ is

$$\bar{x} - \bar{y} \pm t_{\alpha/2}(r)\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where $r = n + m - 2$.

6.6 Welch's formula

In the most general case, where the variances σ_X^2 and σ_Y^2 are unknown and there is no reason to assume that they are equal, we have to use Welch's formula:

$$\bar{x} - \bar{y} \pm t_{\alpha/2}(r)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

where

$$r = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2}$$

Of course, r is the number of degrees of freedom, so it must be an integer. If it happens to be a fractional number (such as $r = 5.68$), its fractional part must be dropped (so we should use $r = 5$).

6.7 Example

The math test scores in two small classes produced the following results: $n_x = 22$, $\bar{x} = 75$ and $s_x = 4.6$ for the first class and $n_y = 26$, $\bar{y} = 72$ and $s_y = 5.1$ for the second class. Construct a 98% confidence interval for the difference $\mu_x - \mu_y$.

Solution. Since σ_X and σ_Y are unknown and may be different, and the sizes of the samples are small, we use Welch's formula. First, the number of degrees of freedom is

$$r = \frac{(4.6^2/22 + 5.1^2/26)^2}{(4.6^2/22)^2/21 + (5.1^2/26)^2/25} = 45.8$$

So we use $r = 45$. Now, the CI is

$$\begin{aligned} \text{CI} &= 75 - 72 \pm t_{.01}(45)\sqrt{4.6^2/22 + 5.1^2/26} \\ &= 3 \pm 2.326 \cdot 1.401 \\ &= 3 \pm 3.259 \end{aligned}$$

Finally, $-0.259 < \mu_X - \mu_Y < 6.259$. Note that the interval is large, i.e. the accuracy of our estimate is low. This happens because our samples are quite small (22 and 26 values). In statistics, accurate estimates usually require large data samples. One cannot derive much inference from too small data.

6.8 Special case: matching measurements

Suppose n people participate in a diet program. Their weights before the program starts are $x_1, \dots, x_n$, and after the program is completed they are $y_1, \dots, y_n$. The program manager wants to determine the average weight drop (for advertisement).

Again, based on the central limit theorem, we may assume that $x_1, \dots, x_n$ are values of a normal random variable $\mathcal{N}(\mu_X, \sigma_X^2)$, and $y_1, \dots, y_n$ are values of a normal random variable $\mathcal{N}(\mu_Y, \sigma_Y^2)$. The manager wants to estimate $\mu_X - \mu_Y$. But here x_i's are *not* independent from y_i's, because these are measurements taken on the same n people.

In this case we need to use the individual differences $d_i = x_i - y_i$ (weight drops) and treat $d_1, \dots, d_n$ as values of a normal random variable $\mathcal{N}(\mu_D, \sigma_D^2)$, where $\mu_D = \mu_X - \mu_Y$. Since σ_D^2 is unknown, we will estimate it by the sample variance s_d^2. Then the confidence interval for $\mu_D = \mu_X - \mu_Y$ is constructed as in Section 5.4

$$\bar{d} \pm t_{\alpha/2}(n-1)\, s_d/\sqrt{n}$$

6.9 Example

Twelve participants in a health-fitness program recorded the following drops in their weights during the program:

+2.0	-0.5	+1.4	-2.2	+0.3	-0.8
+3.7	-0.1	+0.6	+0.2	+0.9	-0.1

Construct a 95% confidence interval for the average weight drop.

Solution: Here the sample mean is $\bar{d} = 0.45$ and the sample variance is $s_d^2 = 2.207$, hence the confidence interval is

$$\bar{d} \pm t_{0.025}(11)\sqrt{s_d^2/n} = 0.45 \pm 2.201\,\sqrt{2.207/12}$$

which is $[-0.494, 1.394]$.

Not much for an advertisement... Again, for such a small sample ($n = 12$) statistical conclusions cannot be too accurate.

7 Confidence intervals for normals: one variance

Suppose a doctor wants to determine the accuracy of the readings of his blood pressure monitor. He measures the blood pressure of the same person repeatedly n times and obtains values $x_1, \dots, x_n$. As before, we assume that these are values of a normal random variable $\mathcal{N}(\mu, \sigma^2)$, where μ is the actual blood pressure of the guy and σ is a typical error. The doctor wants to estimate σ^2. Note that the mean value μ (the actual blood pressure of the guy) is of no concern, we are testing the monitor here.

7.1 Point estimate

The unbiased estimate for σ^2 is the sample variance s^2, i.e. $\mathbb{E}(s^2) = \sigma^2$. The estimate s^2 has a distribution depending on σ^2, even its average depends on σ^2. We want to transform s^2 into a random variable whose distribution is independent of σ^2 (and μ). Let us try s^2/σ^2. Now, at least, $\mathbb{E}(s^2/\sigma^2) = 1$, a constant.

Let us find the distribution of s^2/σ^2, and start with the simplest case $n = 2$:

$$\frac{s^2}{\sigma^2} = \frac{x_1^2 + x_2^2 - 2\left(\frac{x_1+x_2}{2}\right)^2}{\sigma^2} = \frac{\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 - x_1x_2}{\sigma^2} = \frac{(x_1 - x_2)^2}{2\sigma^2} = \left(\frac{x_1 - x_2}{\sqrt{2}\,\sigma}\right)^2$$

Since x_1 and x_2 are independent normal, we have $x_1 - x_2 = \mathcal{N}(0, 2\sigma^2)$, hence

$$\frac{x_1 - x_2}{\sqrt{2}\,\sigma} = \mathcal{N}(0,1)$$

So s^2/σ^2 is the square of a standard normal random variable. Indeed, its distribution is independent of σ^2 and μ.

The case $n > 2$ is more complicated and we omit the calculations. As it happens, s^2/σ^2 is the average of squares of $n-1$ independent standard normal random variables:

$$\frac{s^2}{\sigma^2} = \frac{Z_1^2 + \cdots + Z_{n-1}^2}{n-1}$$

where $Z_i = \mathcal{N}(0,1)$ for each i, and $Z_1, \dots, Z_{n-1}$ are independent.

7.2 χ^2 random variable

In probability, the sum of squares of $r \geq 1$ independent standard normal random variables is called a χ^2 *random variable with r degrees of freedom.*

$$\chi^2(r) = Z_1^2 + \cdots + Z_r^2.$$

This type of random variables plays a particularly important role in statistics.

It is known in probability theory that $\mathbb{E}(Z^2) = 1$ and $\mathbb{E}(Z^4) = 3$, hence

$$\mathsf{Var}(Z^2) = \mathbb{E}(Z^4) - \big[\mathbb{E}(Z^2)\big]^2 = 2.$$

Thus,

$$\mathbb{E}\big(\chi^2(r)\big) = r, \qquad \text{and} \qquad \mathsf{Var}\big(\chi^2(r)\big) = 2r.$$

By the central limit theorem, when r is large ($r > 30$), we can approximate $\chi^2(r)$ by a normal random variable

$$\chi^2(r) \approx \mathcal{N}(r, 2r).$$

Percentiles for a χ^2 random variable are denoted by $\chi^2_\alpha(r)$, that is if $X = \chi^2(r)$, then

$$\mathbb{P}\big(X > \chi^2_\alpha(r)\big) = \alpha \qquad \text{and} \qquad \mathbb{P}\big(X < \chi^2_{1-\alpha}(r)\big) = \alpha.$$

The values of percentiles are given in Table IV. Note that the χ^2 distribution is not symmetric, unlike normal and t.

7.3 Confidence interval for σ^2

Since we established that

$$\frac{(n-1)s^2}{\sigma^2} = \chi^2(n-1),$$

we have

$$\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)$$

with probability $1 - \alpha$. Solving the above inequality for σ^2 we obtain

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}$$

which is a confidence interval with confidence level $1-\alpha$. Note that $(n-1)$ in the numerators is a multiplier, while in the denominators it just indicates the number of degrees of freedom.

To obtain a confidence interval for σ, if this is desired, we simply take the square root:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}}$$

7.4 Shortest confidence interval

The above confidence interval of level $1-\alpha$ was constructed by chopping off two tails, each capturing probability $\alpha/2$ (so that the probability $1-\alpha$ is left in between). This principle produces the shortest possible interval for symmetric distributions, such as normal and t, but not for asymmetric distributions, such as χ^2.

Table X gives numerical values that must be used, instead of percentiles in the previous section, to obtain the shortest confidence interval for the given level $1-\alpha$, see the following example.

7.5 Example

A sample $x_1 \ldots, x_{20}$ from $\mathcal{N}(\mu, \sigma^2)$ yielded $s^2 = 8$. Find a 95% confidence interval for σ^2.

Solution. Here $1-\alpha = 0.95$, hence $\alpha/2 = 0.025$. The size $n = 20$, hence the number of degrees of freedom is 19.

The symmetric (but not shortest) confidence interval is

$$\frac{19 \cdot 8}{32.85} < \sigma^2 < \frac{19 \cdot 8}{8.907}$$

hence $4.63 < \sigma^2 < 17.06$.

The shortest confidence interval is, with values from Table X,

$$\frac{19 \cdot 8}{35.927} < \sigma^2 < \frac{19 \cdot 8}{9.663}$$

hence $4.23 < \sigma^2 < 15.73$.

Which one should we use in practice? This is, basically, a matter of taste. Each one is just as good as the other one.

7.6 Remark

The intervals obtained in the previous example are very large. What kind of estimation of σ^2 this is if all we can say is that "it is somewhere between 4 and 17"? Well, again, this is typical in statistics: estimates based on small samples (here only 20 data values) are rarely precise. Furthermore, estimates of variances are usually far less precise than estimates of means.

7.7 Large samples

Table IV provides percentiles for the $\chi^2(r)$ random up to $r = 30$ degrees of freedom, and in addition includes a few higher values ($r = 40, 50, 60, 80$). Table X stops at $r = 30$. So what do we do for $r > 30$, especially for $r > 80$?

We can use normal approximation, see Section 7.2. Accordingly, the random variable

$$\frac{(n-1)s^2/\sigma^2 - (n-1)}{\sqrt{2(n-1)}} = \frac{\sqrt{n-1}\, s^2/\sigma^2 - \sqrt{n-1}}{\sqrt{2}}$$

is approximately standard normal, $Z = \mathcal{N}(0,1)$, hence

$$-z_{\alpha/2} < \frac{\sqrt{n-1}\, s^2/\sigma^2 - \sqrt{n-1}}{\sqrt{2}} < z_{\alpha/2}$$

with probability $1-\alpha$. Solving this inequality for σ^2 gives a $(1-\alpha)$ confidence interval:

$$\frac{\sqrt{n-1}\, s^2}{\sqrt{n-1} + z_{\alpha/2}\sqrt{2}} < \sigma^2 < \frac{\sqrt{n-1}\, s^2}{\sqrt{n-1} - z_{\alpha/2}\sqrt{2}}$$

This formula is good for large n.

7.8 Generating the χ^2 random variable in MATLAB

With MATLAB Statistical Toolbox, you can generate values of a χ^2 random variable by special commands

x=chi2rnd(v)	returns one random value of t
x=chi2rnd(v,m,n)	an $m \times n$ matrix of random numbers

where **v** denotes the number of degrees of freedom.

8 Confidence intervals for normals: two variances

Suppose a doctor gets a better blood pressure monitor and wants to determine how much more accurate it is compared to the old one. The doctor measures the blood pressure of a patient by the new monitor n times and records values $x_1, \dots, x_n$. His records also contain values of the blood pressure (of another patient) obtained by the old monitor: $y_1, \dots, y_m$.

Again we assume that $x_1, \dots, x_n$ are values of a normal random variable $X = \mathcal{N}(\mu_X, \sigma_X^2)$ and $y_1, \dots, y_m$ are values of another normal random variable $Y = \mathcal{N}(\mu_Y, \sigma_Y^2)$. Here μ_X and μ_Y are the actual blood pressures of these two patients (which are of no concern here). The doctor wants to estimate the ratio σ_X^2/σ_Y^2 to determine how much more accurate the new monitor is.

8.1 Point estimate

The unbiased estimates for σ_X^2 and σ_Y^2 are the sample variances s_x^2 and s_y^2, respectively. Hence the best point estimate for the ratio σ_X^2/σ_Y^2 is s_x^2/s_y^2. To construct a confidence interval, we need to know the corresponding distribution.

8.2 F distribution

It is a fact in probability theory (details are beyond the scope of this course) that the quantity

$$\frac{s_y^2/\sigma_Y^2}{s_x^2/\sigma_X^2}$$

has a special distribution called *F distribution* (and the quantity itself is called an *F random variable*). This distribution has two parameters, $r_1 = m - 1$ and $r_2 = n - 1$, which are called (not surprisingly..) the *numbers of degrees of freedom.* The F random variable is denoted by $F(r_1, r_2)$. The first number r_1 is called the *number of numerator degrees of freedom*, and the second number r_2 is called the *number of denominator degrees of freedom.* The reason is quite evident from the above fraction.

The percentiles of the F random variable are denoted by $F_\alpha(r_1, r_2)$, that is if $X = F(r_1, r_2)$, then

$$\mathbb{P}\big(X > F_\alpha(r_1, r_2)\big) = \alpha \qquad \text{and} \qquad \mathbb{P}\big(X < F_{1-\alpha}(r_1, r_2)\big) = \alpha.$$

There is a little symmetry of the F distribution that will be helpful:

$$F_{1-\alpha}(r_1, r_2) = \frac{1}{F_\alpha(r_2, r_1)}$$

(note that we have to switch the numbers of degrees of freedom).

The values of the percentiles of the F random variable are given in Table VII. This is a fairly large (and confusing) table, so you need to practice with it. Its complexity has a good reason, though: it has to cover two variable parameters (r_1 and r_2) and a variable confidence parameter α. So, essentially, it is a three dimensional table.

8.3 Confidence intervals

We conclude that the inequality

$$\frac{1}{F_{\alpha/2}(n-1,m-1)} < \frac{s_y^2/\sigma_Y^2}{s_x^2/\sigma_X^2} < F_{\alpha/2}(m-1,n-1)$$

holds with probability $1-\alpha$. Solving it for the ratio σ_X^2/σ_Y^2 (this ratio we want to estimate) gives

$$\frac{s_x^2}{s_y^2}\cdot\frac{1}{F_{\alpha/2}(n-1,m-1)} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{s_x^2}{s_y^2}\cdot F_{\alpha/2}(m-1,n-1),$$

which is a $(1-\alpha)$ confidence interval.

8.4 Remark

If an interval for the ratio σ_X/σ_Y of the standard deviations is of interest, we simply take the square root:

$$\frac{s_x}{s_y}\cdot\frac{1}{\sqrt{F_{\alpha/2}(n-1,m-1)}} < \frac{\sigma_X}{\sigma_Y} < \frac{s_x}{s_y}\cdot\sqrt{F_{\alpha/2}(m-1,n-1)}.$$

8.5 Example

A sample $x_1\dots,x_{16}$ from $\mathcal{N}(\mu_X,\sigma_X^2)$ yielded $s_x^2=6$ and a sample $y_1\dots,y_{25}$ from $\mathcal{N}(\mu_Y,\sigma_Y^2)$ yielded $s_y^2=4$. Find a 98% confidence interval for σ_X^2/σ_Y^2.

Solution. Here $\alpha=0.02$, hence $\alpha/2=0.01$. We use values $F_{0.01}(15,24)=2.89$ and $F_{0.01}(24,15)=3.29$ from Table VII and construct the interval by

$$\frac{6}{4}\cdot\frac{1}{2.89} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{6}{4}\cdot 3.29,$$

or, finally, it is $[0.519, 4.935]$.

8.6 Remark

Note again that the confidence interval is fairly large: on the one hand, σ_X^2 may be twice as small as σ_Y^2, and on the other hand, it may be five times bigger!

9 Confidence intervals for proportions

In several previous sections, we treated normal random variables. Those are most important in statistics, since they approximate almost everything (thanks to the central limit theorem).

There is, however, one special type of random variables that must be treated separately – it is binomial. Recall our lottery example in Chapter 3: we observed the value x of a binomial random variable $X = b(n,p)$ and estimated the probability of success (the proportion of winning tickets) by $\hat{p} = x/n$. The unknown parameter p is often called a *proportion*. Here we construct confidence intervals for p.

9.1 CI for p

When $n \geq 30$, we can use a normal approximation to X (recall de Moivre-Laplace theorem in probability theory):

$$X = b(n,p) \approx \mathcal{N}(np, npq)$$

(as usual, $q = 1-p$), then

$$\hat{p} = \frac{x}{n} \approx \mathcal{N}\Big(p, \frac{pq}{n}\Big)$$

so that

$$\frac{\hat{p}-p}{\sqrt{pq/n}} \approx \mathcal{N}(0,1)$$

Using the percentiles of a standard normal random variable $\mathcal{N}(0,1)$ we conclude that the inequality

$$-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{pq/n}} < z_{\alpha/2} \qquad (*)$$

hold with probability $1-\alpha$. Solving it for p in the numerator gives the inequality

$$\hat{p} - z_{\alpha/2}\sqrt{pq/n} < p < \hat{p} + z_{\alpha/2}\sqrt{pq/n}$$

which also hold with probability $1-\alpha$. It looks like a confidence interval for p, but it is not good – it contains the unknown p and $q = 1-p$ on both sides.

In practice, for large n, we can safely replace p with its estimate $\hat{p} = x/n$ and, respectively, q with $1-\hat{p} = 1-x/n$ and obtain a confidence interval with level $1-\alpha$

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}.$$

9.2 A slightly more accurate CI for p

In the previous section we cheated slightly: we "solved" (*) for p in the numerator only, ignoring the presence of p in the denominator. Interestingly, the inequality (*) can be solved for p completely to give the following confidence interval for p:

$$\frac{\hat{p} + z^2/2n}{1 + z^2/n} \pm \frac{z\sqrt{\hat{p}(1-\hat{p})/n + z^2/4n^2}}{1 + z^2/n}$$

where $z = z_{\alpha/2}$. This is a messy formula, which we will never use. By using Taylor expansion one can show that the above confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} + \mathcal{O}(1/n)$$

which differs from the one obtained in the previous Section by a small term $\mathcal{O}(1/n)$, which is insignificant.

9.3 Example

Before an election, $n = 400$ people were polled on their preference between two candidates. Suppose 160 preferred candidate A and 240 preferred candidate B. Construct a 90% confidence interval for the proportion of the population preferring the candidate A.

Solution: the point estimate is easy: $160/400 = 0.4$, but what are possible margins of error? Using the above formulas, we construct the required confidence interval:

$$0.4 \pm 1.645\sqrt{0.4 \times 0.6/400} = 0.4 \pm 0.04$$

(note: we used percentile $z_{0.05} = 1.645$). Thus, the unknown proportion is expected to be within the interval $(0.36, 0.44)$, not a bad accuracy.

9.4 Remark

It may be desirable to estimate the unknown proportion from one side only (from above or from below). In the previous example we may want to find an upper bound only to guarantee that p does not exceed a certain amount. Then we construct a one-sided confidence interval

$$\hat{p} + z_\alpha\sqrt{\hat{p}(1-\hat{p})/n} = 0.4 + 1.282\sqrt{0.4 \times 0.6/400} = 0.43.$$

(note: we used percentile $z_{0.1} = 1.282$ instead of $z_{0.05} = 1.645$). The new upper bound 0.43 is a little smaller (i.e., better) than the previous one 0.44. In other words, we relaxed the lower estimate but tighten the upper estimate.

10 Differences between proportions

Suppose a large hotel chain is buying a detergent for their washing machines. There are two brands available, A and B, and the hotel technicians are trying to determine which brand is better and by how much. In an experiment, the technicians tried the detergent A on n_1 stains and observed that it successfully removed x_1 of them. Then they tried the detergent B on n_2 stains and observed that it successfully removed x_2 of them.

It is reasonable to assume that x_1 and x_2 are values of binomial random variables $X_1 = b(n_1, p_1)$ and $X_2 = b(n_2, p_2)$. Here p_1 and p_2 represent the probabilities of successful removal of stains, i.e. the efficiency of each detergent. The technicians want to estimate the difference $p_1 - p_2$.

10.1 Point estimate

The point estimates for p_1 and p_2 are $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$, respectively. So the point estimate for the difference $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. Its distribution, due to normal approximation, is normal:

$$\hat{p}_1 - \hat{p}_2 \approx \mathcal{N}\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big)$$

Hence,

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx \mathcal{N}(0,1),$$

so the above ratio is guaranteed to stay in the interval $[-z_{\alpha/2}, z_{\alpha/2}]$ with probability $1-\alpha$.

Now using the same arguments and tricks as in the previous chapter, we obtain a confidence interval for $p_1 - p_2$ at level $1-\alpha$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

10.2 Example

A technician is testing two brands of detergents. She tried detergent A on $n_1 = 200$ stains and observed that it successfully removed 144 of them. Then she tried detergent B on $n_2 = 100$ stains and observed that it successfully removed 81 of them. By how much is the detergent B more reliable than A? Construct a 90% confidence interval for the difference between their efficiencies (by the efficiency we mean the proportion of successfully removed stains).

Solution. The point estimates of their efficiencies are $\hat{p}_1 = 144/200 = 0.72$ and $\hat{p}_2 = 81/100 = 0.81$.

Since $\alpha = 0.1$, we will use $z_{0.05} = 1.645$. The confidence interval is

$$0.72 - 0.81 \pm 1.645\sqrt{\frac{0.72 \times 0.28}{200} + \frac{0.81 \times 0.19}{100}} = -0.09 \pm 0.08$$

That is, the difference is guaranteed to be in the interval $[-0.17, -0.01]$.

11 Sample size: experimental design

We have seen that some statistical estimates are pretty accurate (the confidence intervals are small), but others are not (the confidence intervals are too large, sometimes ridiculous). The accuracy of a statistical estimate depends on many factors, in particular on the size of the sample n. Suppose we want to increase the accuracy of an estimate, and moreover, to guarantee that the error will not exceed some small quantity $\varepsilon > 0$. This means that we want the length of the confidence interval of level $1 - \alpha$ be at most 2ε.

While many factors (such as the values of unknown parameters) cannot be adjusted to improve the accuracy, the sample size usually can be increased by collecting more observations (more data). Moreover, given the desired accuracy ε we can compute the minimal size n for which this accuracy will be achieved. This, of course, must be done *before* collecting experimental data, so this job is called *experimental design*.

11.1 Normal random variables

Suppose we are estimating the mean value μ of a normal random variable $\mathcal{N}(\mu, \sigma^2)$ with a known variance σ^2. The half-length of the confidence interval is $z_{\alpha/2}\sigma/\sqrt{n}$, see Section 5.1. Thus, if it must not to exceed ε, we need

$$z_{\alpha/2}\sigma/\sqrt{n} \leq \varepsilon$$

Solving this inequality for n gives

$$n \geq \frac{z_{\alpha/2}^2 \sigma^2}{\varepsilon^2}$$

This give the minimal size of the experimental sample n.

What if σ^2 is unknown (as it normally is)? We cannot use its estimate s^2, because our calculations must be done *before* the experiment, so no data are available yet!

In this case, we may use some reasonable guess of the value of σ^2. Alternatively, we may run a preliminary smaller experiment (collecting just a few values of the random variable) with the sole purpose of (roughly) estimating σ^2. Then we compute n and run the real experiment collecting n values of the variable and estimating μ.

11.2 Proportions

Suppose we are estimating the proportion p of a binomial random variable $X = b(n,p)$. The half-length of the confidence interval is $z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. Thus, if it must not to exceed ε, we need

$$z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \le \varepsilon$$

Solving this inequality for n gives

$$n \ge \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{\varepsilon^2}$$

Here we run into a seemingly unsolvable problem: to find the minimal value of n we already have to know the estimate $\hat{p}$, which will only be available *after* the experiment.

There are two ways to resolve this problem. First, we may use a reasonable guess about the value of p (certain expected value, which does not have to be precise, a rough approximation would suffice).

Second, if there are absolutely no expectations or guesses available, we can note that

$$\hat{p}(1-\hat{p}) \le 1/4$$

for *all* values $0 < \hat{p} < 1$. Then it will be always enough to have

$$n \ge \frac{z^2_{\alpha/2}}{4\varepsilon^2} \qquad (*)$$

This will give us a very accurate estimate for n when $\hat{p} \approx 0.5$, but it may be a significant "overshot" if $\hat{p}$ is close to 0 or 1 (in that case a much smaller value of n may be sufficient for our purposes).

11.3 Example

How many people do we need to poll (see Example 9.3) so that the margin of errors in the 95% confidence interval be less than 0.03?

Solution: here $\alpha = 0.05$, hence we use percentile $z_{0.025} = 1.96$. Since no expected (or guessed) value of p is given, we have to use the universal bound (*), which will be quite accurate anyway because in elections usually p is close to 0.5:

$$n \ge \frac{(1.96)^2}{4 \times (0.03)^2} = 1067.11$$

Hence we need to poll at least 1068 people.

12 Advanced topics in estimation

In many examples, we only used the sample mean $\bar{x}$ and/or the sample variance s^2 to estimate unknown parameters, construct confidence intervals, etc. So we only needed two numbers 'summarizing' the entire sample $x_1, \dots, x_n$. Does it mean that the sample itself can be discarded once a few crucial 'summaries', like $\bar{x}$ and s^2, are computed? Wouldn't the individual values of x_i's help in any way if we knew them? Couldn't they improve our conclusions?

12.1 Sufficient statistics

Recall that the likelihood function is

$$L(\theta) = f(x_1;\theta) \cdots f(x_n;\theta)$$

This function involves both unknown parameters θ and observed data $x_1, \dots, x_n$. Suppose that we can 'separate' them so that

$$L(\theta) = g\big(u(x_1, \dots, x_n), \theta\big) \cdot h(x_1, \dots, x_n)$$

where g, h, u and some functions. Then $u(x_1, \dots, x_n)$ is called *sufficient statistic*. The factor $h(x_1, \dots, x_n)$ that also depends on the data is not included in sufficient statistics.

12.2 Sufficient statistics for normals

The likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{[2\pi\sigma^2]^{n/2}}\, e^{-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2}}$$

Here the data and parameter are "tangled together". But the expression in the exponent can be modified as

$$\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} = \frac{\sum_i x_i^2 - 2\mu \sum_i x_i + n\mu^2}{2\sigma^2}$$

So we have separated parameters from the data: the values $x_1, \dots, x_n$ only appear in two expressions:

$$u_1 = \sum_i x_i \qquad \text{and} \qquad u_2 = \sum_i x_i^2$$

These are sufficient statistics for normal samples.

12.3 Remark

Note that u_1 and u_2 are equivalent to $\bar{x}$ and s^2 in the following sense: knowing u_1 and u_2 one can compute $\bar{x} = u_1/n$ and $s^2 = (u_2 - u_1^2/n)/(n-1)$, see Section 3.21, and vice versa: $u_1 = n\bar{x}$ and $u_2 = (n-1)s^2 + n\bar{x}^2$. So we can also say that $\bar{x}$ and s^2 are sufficient statistics for normals.

12.4 Meaning of sufficient statistics

The theory says that for all statistical purposes (estimation, construction of confidence intervals, etc.) it is enough to have the values of sufficient statistics. The values of individual observations $x_1, \dots, x_n$ cannot improve statistical inferences, so they can be discarded.

This is very convenient in practice: instead of recording and storing all n observed values of a normal random variable, we only need to record and store two values of sufficient statistics: u_1 and u_2, see above!

Moreover, the values such as u_1 and u_2 can be easily computed 'on-line', if the data $x_1, \dots, x_n$ arrive sequentially, one by one. Indeed, every x_i must be added to u_1, its square must be added to u_2, then x_i can be discarded.

12.5 Sufficient statistics for Poisson

The likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}\, e^{-\lambda} = \frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}\, e^{-\lambda n}$$

We see here two expressions involving the data: $x_1 + \cdots + x_n$ and $x_1! \cdots x_n!$, however, the latter is just a factor (denoted by $h(x_1, \dots, x_n)$ in the general formula), hence it can be ignored. The only sufficient statistic is

$$u = x_1 + \cdots + x_n.$$

As a rule, the number of sufficient statistics corresponds to the number of unknown parameters, but there are exceptions...

12.6 Sufficient statistics for uniforms

This is a tricky problem. Let $x_1, \dots, x_n$ be random values of a uniform random variable $X = U(a, b)$ with unknown parameters $a < b$. Its density function is $f(x) = 1/(b-a)$ for $a < x < b$ (and zero elsewhere). Hence the likelihood function is

$$L(a,b) = \begin{cases} \frac{1}{(b-a)^n} & \text{if} \quad a < x_i < b \quad \text{for all} \quad i \\ 0 & \text{elsewhere} \end{cases}$$

Here we do not see any data $x_1, \dots, x_n$ in the expression for $L(a,b)$. However, they are involved in the condition "$a < x_i < b$ for all i". So this condition gives us sufficient statistics. It can be rewritten as

$$a < \min\{x_1 \dots, x_n\} = x_{(1)} \qquad b > \max\{x_1 \dots, x_n\} = x_{(n)}.$$

Thus, the extreme values of the sample, $x_{(1)}$ and $x_{(n)}$, are two sufficient statistics.

12.7 MLE for uniforms

We can also compute the maximum likelihood estimate of the parameters a and b of a uniform random variable $X = U(a,b)$. To find the MLE, we need to maximize the likelihood function $L(a,b)$. Clearly, the fraction $1/(b-a)^n$ takes larger values when $b-a$ gets smaller, i.e. when a and b get closer together. However, we cannot make them arbitrarily close because of the restrictions in the previous section. To make them as close to each other as possible we need to set

$$\hat{a} = \min\{x_1 \dots, x_n\} = x_{(1)} \qquad \hat{b} = \max\{x_1 \dots, x_n\} = x_{(n)}.$$

These are the MLE for a and b.

12.8 Asymptotic distribution of MLE

Suppose $\hat{\theta}_n$ is the maximum likelihood estimate of an unknown parameter θ based on a sample $x_1, \dots, x_n$ of size n. In Chapter 3 we learned that the MLE usually have errors of order $1/\sqrt{n}$, that is their typical values are

$$\hat{\theta}_n = \theta \pm \mathcal{O}(1/\sqrt{n}).$$

Here we describe the distribution of MLE much more precisely. The estimate $\hat{\theta}_n$ is, approximately, a normal random variable with mean θ (which is the actual value of the parameter) and variance σ_n^2, i.e.

$$\hat{\theta} \approx \mathcal{N}(\theta, \sigma_n^2)$$

and the variance satisfies a general formula

$$\begin{aligned}\sigma_n^2 &= \frac{1}{-n\,\mathbb{E}\big(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\big)} \\ &= \frac{1}{n\,\mathbb{E}\big(\frac{\partial}{\partial\theta}\ln f(x;\theta)\big)^2}\end{aligned}$$

where $f(x;\theta)$ denotes the probability density function of the random variable. The above two formulas are equivalent, and in practice you can use either one. We will demonstrate how they work below.

As we can see, the variance is $\sim 1/n$, hence the standard deviation is $\sim 1/\sqrt{n}$, which makes typical errors $\sim 1/\sqrt{n}$, as we know already.

12.9 Rao-Cramer lower bound

The previous section gives a precise formula for the variance (i.e., for typical errors) of the MLE. A natural question is – are there better estimates than MLE? That is, can some other estimates have a smaller variance (i.e., smaller typical errors)? The answer is NO.

First of all, the accuracy of an estimate $\hat{\theta}$ is measured by the mean squared error (MSE), see Section 3.11, and we have the decomposition

$$\mathrm{MSE}(\hat{\theta}) = \mathsf{Var}(\hat{\theta}) + [\mathrm{bias}(\hat{\theta})]^2$$

When the bias is significant, the estimate cannot be accurate, it is not good. In our theoretical analysis, we usually restrict ourselves to unbiased estimates (where the bias is zero) or almost unbiased estimates (by this we mean that the bias is of order less than $1/\sqrt{n}$). Then the accuracy of the estimate $\hat{\theta}$ is characterized by its variance $\mathsf{Var}(\hat{\theta})$ only.

A general theorem called *Rao-Cramer lower bound* says that for any unbiased (or almost unbiased, in the above sense) estimate the variance has the lower bound:

$$\begin{aligned}\mathsf{Var}(\hat{\theta}) &\geq \frac{1}{-n\,\mathbb{E}\big(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\big)} \\ &= \frac{1}{n\,\mathbb{E}\big(\frac{\partial}{\partial\theta}\ln f(x;\theta)\big)^2}\end{aligned}$$

These are the exact same (!) formulas we had for σ_n^2 in the previous section. Thus, no estimate can be better (i.e. more accurate) than the MLE.

This theoretical fact explains the overwhelming popularity of maximum likelihood estimates in practical statistics.

12.10 Efficiency

An estimate $\hat{\theta}$ is said to be *efficient* (or 100% efficient) if its variance coincides with the expression given by the Rao-Cramer lower bound (i.e. its variance takes the smallest possible value, so the estimate is optimal).

If the variance $\mathsf{Var}(\hat{\theta})$ is greater than its lower bound

$$\mathsf{Var}_{\min} = \frac{1}{-n\,\mathbb{E}\big(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\big)}$$

then the ratio

$$\text{Eff} = \frac{\mathsf{Var}_{\min}}{\mathsf{Var}(\hat{\theta})}$$

is called the *efficiency* of the estimate $\hat{\theta}$. It never exceeds 1, it can only be smaller than 1.

In practical terms, the value of the efficiency means the following: if the estimate $\hat{\theta}$ has, say, efficiency 1/2 (or 50%), then in order to achieve the same accuracy as the MLE does, our 'poor' estimate $\hat{\theta}$ would require twice as many data points.

12.11 Example: exponentials

Let X be an exponential random variable, then its density is

$$f(x;\mu) = \frac{1}{\mu}\,e^{-\frac{x}{\mu}}, \qquad x > 0$$

where $\mu > 0$ is the parameter. We will computet the Rao-Cramer lower bound. Taking logarithm we get

$$\ln f(x;\mu) = -\ln\mu - \frac{x}{\mu}$$

Differentiating with respect to μ gives

$$\frac{\partial}{\partial\mu}\ln f(x;\mu) = -\frac{1}{\mu} + \frac{x}{\mu^2}$$

Now we have two options. First, we can square this expression and take its mean value:

$$\mathbb{E}\Big(\frac{1}{\mu^2} - \frac{2x}{\mu^3} + \frac{x^2}{\mu^4}\Big) = \frac{1}{\mu^2} - \frac{2\mathbb{E}(X)}{\mu^3} + \frac{\mathbb{E}(X^2)}{\mu^4}$$

For the exponential random variable, $\mathbb{E}(X) = \mu$ and $\mathbb{E}(X^2) = \mathsf{Var}(X) + [\mathbb{E}(X)]^2 = 2\mu^2$. Hence we obtain

$$\frac{1}{\mu^2} - \frac{2\mu}{\mu^3} + \frac{2\mu^2}{\mu^4} = \frac{1}{\mu^2}$$

Thus

$$\mathsf{Var}_{\min} = \frac{\mu^2}{n}.$$

Alternatively, we can take the second order derivative

$$\frac{\partial^2}{\partial\mu^2} \ln f(x;\mu) = \frac{1}{\mu^2} - \frac{2x}{\mu^3}$$

and then the mean value

$$\mathbb{E}\Big(\frac{1}{\mu^2} - \frac{2x}{\mu^3}\Big) = \frac{1}{\mu^2} - \frac{2\mu}{\mu^3} = -\frac{1}{\mu^2}$$

and then

$$\mathsf{Var}_{\min} = \frac{\mu^2}{n}$$

(note that the two minuses cancel out). We have seen in Section 3.14 that the variance of the MLE $\hat{\mu}$ is exactly μ^2/n.

12.12 Example

Let X be a random variable with density function

$$f(x;\theta) = \theta\, x^{\theta-1}, \qquad 0 < x < 1$$

where $\theta > 0$ is a parameter. We will compute the Rao-Cramer lower bound. Taking logarithm we get

$$\ln f(x;\theta) = \ln\theta + (\theta-1)\ln x$$

Differentiating with respect to θ gives

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) = \frac{1}{\theta} + \ln x$$

Differentiating with respect to θ again gives

$$\frac{\partial^2}{\partial\theta^2} \ln f(x;\theta) = -\frac{1}{\theta^2}.$$

Note that the variable x is gone! Now the mean value is

$$\mathbb{E}\Big(-\frac{1}{\theta^2}\Big) = -\frac{1}{\theta^2}$$

and then

$$\mathsf{Var}_{\min} = \frac{\theta^2}{n}$$

(note again that the two minuses cancel out).

13 Hypotheses testing: introduction

In the previous sections, the results of our work was numerical: point estimates and confidence intervals were some numbers. In the real life, numbers are used to make certain practical decisions. In Example 10.2, the results of the test may be used by the management of a large hotel chain to select one detergent over the other.

The life of statisticians would have been easier if all they had to provide were numbers, and then some other 'responsible people' made final decisions. But our life is not that simple. Statisticians have to make (or at least suggest) final decisions, keeping numbers and other technical details to ourselves.

Making a decision usually requires choosing between two (or more) available options (such as recommending to buy detergent A versus detergent B). We call these options *hypotheses*, because they remain hypothetical (potential) until statisticians select one of them and reject the others.

13.1 Example

A manufacturer of detergents knows that the efficiency of its product (the probability of successful removal of stains) is $p_0 = 0.72$. Engineers propose a new technological method claiming that the efficiency of the detergent would increase. A small quantity of detergent was obtained in a lab by using the new technology and tested on $n = 100$ stains. It successfully removed $x = 81$ of them.

Should the manufacturer adopt the new technology? This is a serious question, because updating technology is a costly process. The manufacturer must be confident that the new detergent is indeed better than the old one. Statisticians must determine if the efficiency of the new detergent, let us call it p, is greater (or not greater) than p_0. They must tell the manufacturer: "observed data make us, say, 95% confident that $p > p_0$". Or else, "there is no sufficient evidence to claim, with 95% confidence, that $p > p_0$".

Note: the level of confidence (here 95%) has to be prescribed. One can never be 100% sure of anything when dealing with random events – even a fair coin flipped 100 times may land on Heads all 100 times (this is extremely unlikely, but possible). In practical work, we can set the level of confidence in advance to something high enough, such as 95%, and then feel fairly comfortable with our conclusions. That's what testing hypotheses is all about.

13.2 Null and alternative hypotheses

In the above example, the base hypothesis is that there is no improvement in the efficiency of detergent, i.e. p takes its default value $p = p_0$. This is called the *null hypothesis*. The other hypothesis is that there is an improvement, i.e. p takes an alternative value $p > p_0$. This is called the *alternative hypothesis*:

$$\begin{aligned} H_0 :& \quad p = p_0 \qquad \text{(null hypothesis)} \\ H_1 :& \quad p > p_0 \qquad \text{(alternative hypothesis)} \end{aligned}$$

We note that the null hypothesis only includes *one* value of the parameter p. Such hypotheses are said to be *simple*. On the contrary, the alternative hypothesis covers a whole interval of values of the parameter p. Such hypotheses are said to be *composite*.

13.3 Errors of type I and II

Statisticians need to decide which hypothesis is true. They may accept H_0 and reject H_1, or accept H_1 and reject H_0. At the same time, one of these two hypotheses is actually true, either H_0 or H_1. We have four possible combinations:

		accepted:	
		H_0	H_1
true:	H_0	OK	Error-I
	H_1	Error-II	OK

The diagonal combinations $H_0 - H_0$ and $H_1 - H_1$ are OK, in these cases the statisticians make the right decision. The other two combinations mean statistical errors:

Type I error: accepting H_1 when H_0 is actually true.

Type II error: accepting H_0 when H_1 is actually true.

Statisticians want to reduce the probabilities of errors:

$$\begin{aligned} \mathbb{P}(\text{Error I}) &= \alpha \qquad \text{(significance level)} \\ \mathbb{P}(\text{Error II}) &= \beta \end{aligned}$$

These two values are crucial in hypothesis testing.

13.4 Actual test

How do we actually choose one hypothesis over the other, in the above example? Denote by $\hat{p} = 81/100 = 0.81$ the estimate of the unknown parameter p obtained from the observations. Obviously, if $\hat{p}$ is sufficiently high, i.e. $\hat{p} \geq c$, then we should accept the hypothesis H_1, otherwise we should reject it and accept H_0. Here c is a certain *critical value*. The interval $\hat{p} \geq c$ is the *critical region* (the region of the values of the estimate where we make the 'critical' decision to accept H_1).

How do we choose c? Here is a common practice: set a certain (small) value for the probability of type I error, such as $\alpha = 0.1$ or $\alpha = 0.05$ or $\alpha = 0.01$, and choose c that makes α equal to that value, see below.

13.5 The choice of c

Suppose the value of α is prescribed. The probability of type I error can be computed as

$$\alpha = \mathbb{P}(\hat{p} > c;\ \ H_0 \ \text{ is true}) = \mathbb{P}\big(\tfrac{x}{n} > c;\ \ p = p_0\big)$$

By using normal approximation (de Moivre-Laplace theorem), $x \approx \mathcal{N}(np, npq)$, hence $\frac{x}{n} \approx \mathcal{N}(p, \frac{pq}{n})$, where $p = p_0$ (since we assume the null hypothesis is true) and $q = 1 - p = 1 - p_0$. So the above probability is

$$\alpha = \mathbb{P}\big(\tfrac{x}{n} > c\big) \approx 1 - \Phi\Big(\frac{c - p_0}{\sqrt{p_0(1-p_0)/n}}\Big)$$

thus

$$\frac{c - p_0}{\sqrt{p_0(1-p_0)/n}} = z_\alpha$$

hence

$$c = p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}$$

For example, let $\alpha = 0.05$. Then $z_{0.05} = 1.645$ and

$$c = 0.72 + 1.645\sqrt{0.72 \cdot 0.28/100} = 0.794$$

If the estimate $\hat{p}$ exceeds the critical value $c = 0.794$ (which it does, since we have $\hat{p} = 0.81$), then statisticians accept H_1 with the 95% confidence level.

The condition $\hat{p} > c = p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}$ is equivalent to

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha$$

The new quantity Z here has approximately standard normal distribution $\mathcal{N}(0,1)$. We call it the *test statistic*, it is used to test the given hypotheses. The critical region can be described as $Z > z_\alpha$.

In our example,

$$Z = \frac{0.81 - 0.72}{\sqrt{0.72 \cdot 0.28/100}} = 2.00 > z_{0.05} = 1.645$$

thus we are in the critical region, i.e. accept H_1.

Another popular name for the test statistic Z is *Z-score*. Any test statistics that has a standard normal distribution $\mathcal{N}(0,1)$ is called *Z-score*.

13.6 Power of the test

We designed the test to make α small enough (=0.05), but what about β? To compute the type II error we again use normal approximation:

$$\begin{aligned}\beta &= \mathbb{P}(\hat{p} < c; \quad H_1 \text{ is true}) \\ &= \mathbb{P}(x/n < c; \quad p > p_0) \approx \Phi\Big(\frac{c-p}{\sqrt{p(1-p)/n}}\Big)\end{aligned}$$

We see that β is a function of the unknown parameter p. This is why it could not be used to design the test: if we set β to a small value (such as 0.05), we still would not be able to determine c, because the unknown parameter p is involved in the formula here.

Since we want β to be small, we call the value $K = 1 - \beta$ the *power of the test*. The smaller β, the more powerful the test is.

For example, if the actual efficiency of the new detergent is $p = 0.85$, then

$$\beta = \Phi\Big(\frac{0.794 - 0.85}{\sqrt{0.85 \cdot 0.15/100}}\Big) = \Phi(-1.57) = 0.0582$$

so the power of the test is $K = 0.9418$ (very high). That is, while Type I error would occur with probability 5% (by design), Type II error would occur with probability 5.82% (by calculation).

13.7 Trade-off between α and β

Suppose we try to decrease the probability of type I error and set $\alpha = 0.01$. Then $z_{0.01} = 2.326$ and

$$c = 0.72 + 2.326\sqrt{0.72 \cdot 0.28/100} = 0.8244$$

Assuming again that the the actual efficiency of the new detergent is $p = 0.85$, we obtain

$$\begin{aligned}\beta &= \Phi\Big(\frac{0.8244 - 0.85}{\sqrt{0.85 \cdot 0.15/100}}\Big)\\ &= \Phi(-0.717) = 0.2367\end{aligned}$$

So the probability of type II error increased from 5.82% to 23.67%. The power of the test dropped accordingly from 0.9418 to 0.7633. This means, in practical terms, that the alternative hypothesis is now likely to be rejected even if it is true. By the way, in our example $\hat{p} = 0.81 < c = 0.8244$, so we do reject H_1 when $\alpha = 0.01$ (recall that we have accepted H_1 when $\alpha = 0.05$).

Here is a classical trade-off in statistical hypothesis testing: making α smaller automatically increases β and vice versa. In practice one should look for a reasonable compromise. Here are some guidelines:

Choosing α smaller reduces the chances of accepting the alternative hypothesis H_1, whether it is true or not. This is a 'safe play', reflecting the tendency to stick to default, shying away from alternatives and risks they involve.

Choosing α larger increases the chances of accepting the alternative hypothesis H_1, whether it is true or not. This is a 'risky play', an aggressive strategy oriented to finding alternatives, innovations and profit.

13.8 p-value

In the above example, we first set $\alpha = 0.05$ and accepted H_1. Then we reset α to 0.01 and rejected H_1. Obviously, there is a borderline between acceptance and rejection, i.e. there is a value (called the *probability value*, or for brevity the *p-value*) such that

$$\begin{aligned}&\text{if} \quad \alpha > \text{p-value}, \quad \text{we accept} \quad H_1 \quad (\text{reject } H_0)\\ &\text{if} \quad \alpha < \text{p-value}, \quad \text{we reject} \quad H_1 \quad (\text{accept } H_0)\end{aligned}$$

Many people find these rules hard to memorize. A popular chant can help:

P is low – the null must go; P is high – the null will fly

To compute the p-value, we must use the formula for α, but substitute the estimated value of $\hat{p}$ for the critical value c, that is

$$\begin{aligned}\text{p-value } &= 1 - \Phi\Big(\frac{0.81 - 0.72}{\sqrt{0.72 \cdot 0.28/100}}\Big)\\ &= 1 - \Phi(2.00) = 0.0228\end{aligned}$$

So, for any $\alpha > 0.0228$ we accept H_1, for any $\alpha < 0.0228$ we reject H_1.

Recall that the significance level α (the risk level) must be specified *a priori*, rather than computed from the data. On the contrary, the p-value is computed from the data, there is no need to specify α. This is convenient: now the statisticians can report the p-value and leave the burden of making the final decision to other responsible people.

14 Hypotheses testing: proportions

14.1 Summary of the previous chapter

Suppose $X = b(n,p)$ is a binomial random variable with an unknown proportion p. We have tested the null hypothesis $H_0\colon p = p_0$ against the alternative hypothesis $H_1\colon p > p_0$. In an experiment, we have observed a value x of the variable X. Then we used the test statistic (Z-score)

$$Z = \frac{x/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

and accepted the alternative H_1 whenever $Z > z_\alpha$, where α was the preset level of significance. The p-value was computed by the rule: p-value $= 1 - \Phi(Z)$.

14.2 Two other alternatives

In our example, engineers were interested in increasing the proportion p (the efficiency of the detergent). In other applications it may be desirable to decrease the proportion p (e.g., if it represents the failure rate). Then the alternative hypothesis would be $H_1\colon p < p_0$. In that case we use the same test statistic Z and accept the hypothesis H_1 if Z is too small, precisely if $Z < -z_\alpha$. The p-value is computed by the rule: p-value $= \Phi(Z)$.

In yet other applications it may be desirable to change p either way, or just verify whether or not the value of p equals its default value p_0. Then the alternative hypothesis is $H_1\colon p \neq p_0$ (the *two-sided* hypothesis). In that case we use the same test statistic Z and accept the hypothesis H_1 if Z differs from zero either way, precisely if $|Z| > z_{\alpha/2}$. The p-value is computed by the rule: p-value $= 2\,[1 - \Phi(|Z|)]$. Summary:

H_0	H_1	Critical region	p-value	Test statistic
$p = p_0$	$p > p_0$	$Z > z_\alpha$	$1 - \Phi(Z)$	$Z = \frac{x/n-p_0}{\sqrt{p_0(1-p_0)/n}}$
	$p < p_0$	$Z < -z_\alpha$	$\Phi(Z)$	
	$p \neq p_0$	$\lvert Z\rvert > z_{\alpha/2}$	$2\,[1 - \Phi(\lvert Z\rvert)]$	

14.3 Two proportions

In the previous example, we assumed that the efficiency of the 'old' detergent was known ($p_0 = 0.72$). What if it isn't? Then it should be determined experimentally, just as the efficiency of the new detergent. Then we will have to compare two proportions, as in the original Example 10.2.

Suppose $X_1 = b(n_1, p_1)$ and $X_2 = b(n_2, p_2)$ are two binomial random variables. We want to compare p_1 and p_2. Our null (base) hypothesis is $H_0\colon p_1 = p_2$. The alternative H_1, depending on the particular goals of the test, may one of three forms: $p_1 > p_2$ or $p_1 < p_2$ or $p_1 \neq p_2$. In an experiment, we observe the values x_1 and x_2 of these variables.

As it follows from the analysis in Section 10.1, the following statistic has a standard normal distribution $\mathcal{N}(0, 1)$:

$$\frac{x_1/n_1 - x_2/n_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

But this expression must be adapted to hypotheses testing. First of all, we only need to know its distribution under the null hypothesis, i.e. when $p_1 = p_2$. Then the term $p_1 - p_2$ in the numerator vanishes. In the denominator, the unknown value $p_1 = p_2$ can be replaced with its best estimated (the combined, or *pooled* estimate from the two experiments):

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

(the total number of successes over the total number of trials). Thus we get the test statistic (Z-score)

$$Z = \frac{x_1/n_1 - x_2/n_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The actual test is completed as before, according to the table:

H_0	H_1	Critical region	p-value	Test statistic
	$p_1 > p_2$	$Z > z_\alpha$	$1 - \Phi(Z)$	
$p_1 = p_2$	$p_1 < p_2$	$Z < -z_\alpha$	$\Phi(Z)$	$Z = \frac{x_1/n_1 - x_2/n_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
	$p_1 \neq p_2$	$\lvert Z\rvert > z_{\alpha/2}$	$2\,[1 - \Phi(\lvert Z\rvert)]$	$\hat{p} = \frac{x_1+x_2}{n_1+n_2}$

14.4 Example

A UAB doctor notices that many patients having Menier's disease also have hearing trouble. Menier's disease is very hard to diagnose. So the doctor proposes to include the hearing test into the diagnostic procedure. To determine if his theory is correct the doctor uses hearing test on 96 patients (known to have Menier's disease) and on randomly selected 119 healthy people. The results of his experiment are presented in the table below:

	Fail	Pass	Total
Ill	71	25	96
Healthy	76	43	119

(these are real data obtained at the UAB Medical School in about 1997).

The doctor expects that ill patients would fail the test more frequently than healthy people do. It looks like this indeed happens in his experiment. But is there a sufficient evidence to claim the existence of such a tendency with a high confidence level?

We proceed with the test. Let p_1 denote the probability of failure of the hearing test for ill patients and p_2 that for healthy people. We test the null hypothesis $H_0\colon p_1 = p_2$ against the alternative $H_1\colon p_1 > p_2$. We start with

$$\hat{p} = \frac{71+76}{96+119} = 0.6837$$

(note that we treat 'failures of the hearing test' as 'successes'), then

$$Z = \frac{71/96 - 76/119}{\sqrt{0.6837 \cdot 0.3163 \cdot \left(\frac{1}{96} + \frac{1}{119}\right)}} = 1.5825$$

If we set $\alpha = 0.1$, then $Z > z_{0.1} = 1.282$, so we accept the alternative H_1 (thus validating the doctor's theory).

But if we set $\alpha = 0.05$, then $Z < z_{0.05} = 1.645$, so we reject the alternative H_1 (thus invalidating the doctor's theory).

The p-value here is $1 - \Phi(1.5825) = 0.0571$. Thus if the medical community is willing to take a risk higher than 5.71% in this experiment, then the doctor's theory should be welcomed. Otherwise it should be rejected (until further tests, perhaps). Whatever the actual standards in the medical community, we can say (from statisticians viewpoint) that the data mildly support the doctor's theory, but do not give an overwhelming evidence.

14.5 Remark

In the above example, the doctor expects that ill patients would fail the test more frequently than healthy people do. What if the doctor does not have any specific expectations either way, he only believes that Menier's disease *somehow affects* the results of the hearing test, whether increasing failures or decreasing failures?

Then we have to test the null hypothesis $H_0\colon p_1 = p_2$ against the two-sided alternative $H_1\colon p_1 \neq p_2$. In that case the critical region is $|Z| > z_{\alpha/2}$.

Now if we set $\alpha = 0.1$, then $|Z| < z_{0.05} = 1.645$, we reject the alternative (thus invalidating the doctor's theory).

Only if we set $\alpha = 0.2$, then $|Z| > z_{0.1} = 1.282$, then we accept the alternative (thus validating the doctor's theory).

The p-value now is $2\,[1 - \Phi(1.5825)] = 0.1142$, i.e. the critical risk level is 11.42%.

We see that the acceptance of the two-sided alternative hypothesis is twice as more risky than that of the one-sided alternative. It is common in statistics: better formulated, well-focused hypotheses have higher chances of acceptance.

14.6 Warning

The form of the alternative hypothesis must be specified by the doctor based on general medical considerations (or previous experience). This must be done before the experiment is performed! Neither the doctor, nor statisticians should look into the experimental data for clues of how to specify the alternative hypothesis. This would constitute a 'statistical cheating'. Anyone doing such inappropriate things is making a serious mistake and may arrive at totally wrong conclusions.

15 Hypotheses testing: one normal distribution

15.1 Example

Suppose the average score in calculus tests at a particular university is used to be 60 (out of 100). The university considers adopting a new textbook hoping the students would learn better and get higher scores. In an experimental class of $n = 25$, calculus was taught by the new book, and the average score in this class was 64. Is there a sufficient evidence to claim that the test scores would increase if the new book is adopted?

We agreed to consider the test scores as values of a normal random variable $\mathcal{N}(\mu, \sigma^2)$. Our primary interest here is the average μ: is it greater than 60 or not? The secondary parameter σ is not relevant to the question posed and for simplicity we assume that it is known: $\sigma = 10$.

Thus we need to test the null (base) hypothesis $H_0\colon \mu = 60$ against the alternative $H_1\colon \mu > 60$.

15.2 General method

Suppose we are testing the null hypothesis $H_0\colon \mu = \mu_0$ against the alternative $H_1\colon \mu > \mu_0$.

The point estimate of the unknown parameter μ is $\bar{x}$. It has a normal distribution $\mathcal{N}(\mu, \sigma^2/n)$. Under the null hypothesis $\mu = \mu_0$ its distribution is $\mathcal{N}(\mu_0, \sigma^2/n)$. The following statistics (Z-score)

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

has a standard normal distribution $\mathcal{N}(0, 1)$.

If the alternative hypothesis is true, then μ is large and we expect $\bar{x}$ to be large, and so Z is expected to be large as well. Hence the critical region must be of the form $Z > c$. Given the significance level α, the probability of type I error must be

$$\mathbb{P}(Z > c;\quad H_0 \text{ is true}) = \alpha$$

hence $c = z_\alpha$. The critical region is $Z > z_\alpha$. Note the similarity with Section 14.2. As in that section, now the p-value is $1 - \Phi(Z)$.

15.3 Example (continued)

In our example $Z = \frac{64-60}{10/5} = 2$. So if we set $\alpha = 0.05$, then $Z > z_{0.05} = 1.645$, so we accept H_1 (adopt the new textbook). Even if we set $\alpha = 0.025$, then $Z > z_{0.025} = 1.96$, so we again accept H_1. But if we set $\alpha = 0.01$, then $Z > z_{0.01} = 2.326$, so we accept H_0 (and reject the new textbook). The p-value is $1 - \Phi(2) = 0.0228$, so the 'borderline' risk level is 2.28%.

15.4 Summary

The alternative hypothesis may be of two other forms: $\mu < \mu_0$ and $\mu \neq \mu_0$. Accordingly, the test proceeds as follows:

H_0	H_1	Critical region	p-value	Test statistic
$\mu = \mu_0$	$\mu > \mu_0$	$Z > z_\alpha$	$1 - \Phi(Z)$	$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
	$\mu < \mu_0$	$Z < -z_\alpha$	$\Phi(Z)$	
	$\mu \neq \mu_0$	$\|Z\| > z_{\alpha/2}$	$2\,[1 - \Phi(\|Z\|)]$	

15.5 A general principle

Constructing confidence intervals and testing hypothesis we use the same statistics. There is a general principle: the null hypothesis $H_0\colon \theta = \theta_0$ about the unknown parameter θ is accepted if the value θ_0 is in the $1-\alpha$ confidence interval for θ. Applying this principle we need to remember that when the alternative is two-sided $H_1\colon \theta \neq \theta_0$, then the CI must be two-sided, but if the alternative is one-sided, then the CI also must be two-sided.

15.6 Power function

Suppose the critical region is specified by $\bar{x} > 62$. Then we can compute the power function

$$K(\mu) = 1 - \beta = \mathbb{P}(\bar{x} > 62) = 1 - \Phi\Big(\frac{62 - \mu}{2}\Big)$$

Here are a few of its values:

62	63	64	65	66	67	68
0.5000	0.6915	0.8413	0.9332	0.9772	0.9938	0.9987

We see that the power function (which is the probability of accepting H_1 if it is correct) starts with moderate values (50% to 69%) but quickly gets over 90% and then over 99%. If the actual average of calculus test scores with the new textbook is 68, this fact will be recognized by our test, and the textbook will be adopted with probability 99.87%.

But what if the actual average with the new textbook is only 63? Then the chance of adoption is mere 69.15%, even though the new score of 63 is higher than the old one of 60. Our test is simply not powerful enough to recognize such a relatively small difference ($63 - 60 = 3$) between the old and new averages, the chance of its failure (of making the wrong decision) exceeds 30%.

How to increase the power of the test? Obviously, by testing more students, by enlarging the experimental class. The larger the sample the more accurate statistical conclusions.

15.7 Sample size: experimental design

Here we again come to the experimental design. Suppose we are testing the hypothesis $\mu = \mu_0$ against $\mu > \mu_0$. We want the significance level (which is the probability of type I error) to be equal to a small value, α. In addition, we want the probability of type II error be equal to another small value, β for a particular value of $\mu = \mu_1$. How can we design such a test?

The probability of type I error is

$$\alpha = \mathbb{P}(\bar{x} > c;\ \mu = \mu_0) = 1 - \Phi\Big(\frac{c - \mu_0}{\sigma/\sqrt{n}}\Big)$$

hence

$$\frac{c - \mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

The probability of type II error is

$$\beta = \mathbb{P}(\bar{x} < c;\ \mu = \mu_1) = \Phi\Big(\frac{c - \mu_1}{\sigma/\sqrt{n}}\Big)$$

hence

$$\frac{c - \mu_1}{\sigma/\sqrt{n}} = z_{1-\beta} = -z_\beta$$

We arrive at a system of equations:

$$\begin{aligned} c - \mu_0 &= z_\alpha \sigma/\sqrt{n} \\ c - \mu_1 &= -z_\beta \sigma/\sqrt{n} \end{aligned}$$

Solving it for n and c gives

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2}$$

and

$$c = \frac{\mu_0 z_\beta + \mu_1 z_\alpha}{z_\alpha + z_\beta}$$

We emphasize that if $\mu_1 - \mu_0$ is small, then there isn't much difference between the null hypothesis and the alternative hypothesis, they are hard to distinguish. This is why n must be large to make that distinction possible.

15.8 Example (continued)

Suppose in our calculus test example, we want the significance level to be $\alpha = 0.025$. Also suppose that if the average test score with the new book is 63 (versus the old average of 60), we want this fact to be recognized by our test with probability 95%, i.e. we want $\beta = 0.05$. Then we need

$$n = \frac{(1.96 + 1.645)^2 \cdot 100}{(63 - 60)^2} = 144.4$$

i.e. we need at least 145 students for the experimental class(es). The critical value should be

$$c = \frac{1.645 \cdot 60 + 1.96 \cdot 63}{1.645 + 1.96} = 61.63$$

So if the average score in the experimental class(es) exceeds 61.63, we will adopt the new textbook, otherwise we reject it.

15.9 Remark

In the above analysis we assumed that $\mu_1 > \mu_0$. But the same formulas for n and c apply in the case $\mu_1 < \mu_0$.

15.10 Unknown variance

So far we greatly simplified our life by assuming that σ was known. In practice we rarely have such a luxury. If σ^2 is unknown, then it should be replaced with its best estimate s^2, and the normal distribution – with the t-distribution (the latter has $n-1$ degrees of freedom). The test statistic is

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and the test procedure goes as follows:

H_0	H_1	Critical region	p-value	Test statistic
	$\mu > \mu_0$	$T > t_\alpha(n-1)$	$1 - F(T)$	
$\mu = \mu_0$	$\mu < \mu_0$	$T < -t_\alpha(n-1)$	$F(T)$	$T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
	$\mu \neq \mu_0$	$\lvert T\rvert > t_{\alpha/2}(n-1)$	$2\,[1 - F(\lvert T\rvert)]$	

where $F(x)$ denotes the distribution function of the t random variable with $n-1$ degrees of freedom.

The textbook explains how to compute an "approximate" p-value by using Table VI, see the next example. Exact p-value can be found with the help of the on-line calculator on the instructor's web page.

15.11 Example

Suppose $n = 24$ random values of a normal random variable yielded $\bar{x} = 0.079$ and $s = 0.255$. Test the hypothesis $H_0\colon \mu = 0$ against $H_1\colon \mu > 0$ at the 5% significance level.

Solution: Since σ^2 is unknown we use the T-statistic

$$T = \frac{0.079 - 0}{0.255/\sqrt{24}} = 1.518$$

Since $T = 1.518 < t_{0.05}(23) = 1.714$, we accept the null hypothesis H_0.

To determine the "approximate" p-value, we find in Table VI two percentiles that are closest to the value of the T-statistic on both sides of it: $t_{0.1}(23) = 1.319$ and $t_{0.05}(23) = 1.714$ (note that we must use the same number of degrees of freedom, 23). Then the p-value is between the two corresponding percentages, i.e. the p-value is between 5% and 10%, or in the interval $(0.05, 0.1)$.

The on-line calculator on the instructor's web page gives a precise answer: p-value=0.0713.

15.12 Test for the variance

So far we have tested the mean value μ of a normal random variable $X = \mathcal{N}(\mu, \sigma^2)$. In practice it is sometimes necessary to test the variance σ^2. For example, let σ represent a typical error in readings of a blood pressure monitor. A new brand of monitor is considered by a doctor, who might want to test if its typical readings error σ exceeds a certain threshold σ_0.

The null hypothesis says that $\sigma^2 = \sigma_0^2$. The alternative may be of one of the three forms: $\sigma^2 > \sigma_0^2$ or $\sigma^2 < \sigma_0^2$ or $\sigma^2 \neq \sigma_0^2$. To test these hypotheses we use the statistic

$$\boldsymbol{\chi}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

which has a χ^2 distribution with $n-1$ degrees of freedom. The test procedure is summarized below:

H_0	H_1	Critical region	Test statistic
	$\sigma^2 > \sigma_0^2$	$\boldsymbol{\chi}^2 > \chi^2_\alpha(n-1)$	
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\boldsymbol{\chi}^2 < \chi^2_{1-\alpha}(n-1)$	$\boldsymbol{\chi}^2 = \frac{(n-1)s^2}{\sigma_0^2}$
	$\sigma^2 \neq \sigma_0^2$	$\boldsymbol{\chi}^2 > \chi^2_{\alpha/2}(n-1)$ or $\boldsymbol{\chi}^2 < \chi^2_{1-\alpha/2}(n-1)$	

Note that in the third case $\sigma^2 \neq \sigma_0^2$ the critical region consists of two intervals: $\boldsymbol{\chi}^2 > \chi^2_{\alpha/2}(n-1)$ and $\boldsymbol{\chi}^2 < \chi^2_{1-\alpha/2}(n-1)$. This means that if the test statistic $\boldsymbol{\chi}^2$ falls into either of these intervals, we accept H_1.

The formulas for the p-value are rather complicated, we omit them. The textbook explains how to compute an "approximate" p-value by using Table IV, see the next example.

15.13 Example

Suppose $n = 16$ random values of a normal random variable yielded $\bar{x} = 10$ and $s^2 = 8$. Test the hypothesis $H_0\colon \sigma^2 = 16$ against $H_1\colon \sigma^2 < 16$ at the 1% significance level.

Solution: the test statistic is $\boldsymbol{\chi}^2 = \frac{15\cdot 8}{16} = 7.5$. Since $\boldsymbol{\chi}^2 = 7.5 > \chi^2_{0.99}(15) = 5.229$, we are not in the critical region, so we accept H_0.

To determine the "approximate" p-value, we find in Table IV two percentiles that are closest to the value of our test statistic: $\chi^2_{0.95}(15) = 7.261$ and $\chi^2_{0.90}(15) = 8.547$ (note that we must use the same number of degrees of freedom, 15). Note that the subscripts 0.95 and 0.9 are the values of $1-\alpha$, so the corresponding values of α are 0.05 and 0.1. Then the p-value is between the two corresponding α-percentages, i.e. the p-value is between 5% and 10%, or in the interval $(0.05, 0.1)$.

The on-line calculator on the instructor's web page gives a precise answer: p-value=0.0577.

16 Hypotheses testing: two normals

Let $x_1, \dots, x_n$ be calculus test scores in one class and $y_1, \dots, y_m$ calculus test scores in another class in the same university. The classes were taught from the same textbook but by two different professors. The university wants to combine these two sets of scores for some larger analysis assuming that they can be treated as values of the same normal random variable. But is it a correct assumption? Maybe the professors teach and/or grade too differently?

Just looking up the scores for similarities or differences is not enough, only a formal test can determine whether or not these two samples can be treated as values of the same normal random variable.

16.1 Formal test

Let $x_1, \dots, x_n$ be random values of a normal variable $X = \mathcal{N}(\mu_X, \sigma_X^2)$ and $y_1, \dots, y_m$ random values of a normal variable $Y = \mathcal{N}(\mu_Y, \sigma_Y^2)$. We want to test the hypothesis $H_0\colon X = Y$ against the alternative $H_1\colon X \neq Y$.

Since a normal random variable is completely determined by its mean μ and variance σ^2, the null hypothesis really says that $\mu_X = \mu_Y$ and $\sigma_X^2 = \sigma_Y^2$. We will test these two identities separately, in two steps.

Which identity should we test first? Recall that the information about variances is essential for the analysis of means (Chapter 6) but the analysis of variances does not require any knowledge about means (Chapter 8). So we start with the variances.

16.2 Step 1: variances

We test the null hypothesis $H_0\colon \sigma_X^2 = \sigma_Y^2$ against the alternative $H_1\colon \sigma_X^2 \neq \sigma_Y^2$. We use the F-statistic

$$\mathbf{F} = \frac{s_y^2/\sigma_Y^2}{s_x^2/\sigma_X^2}$$

which has $F(m-1, n-1)$ distribution, see Chapter 8. Under the null hypothesis $\sigma_X^2 = \sigma_Y^2$, thus the F-statistic becomes simple: $\mathbf{F} = s_y^2/s_x^2$. Accordingly, the test proceeds as follows

H_0	H_1	Critical region	Test statistic
	$\sigma_Y^2 > \sigma_X^2$	$\mathbf{F} > F_\alpha(m-1, n-1)$	
$\sigma_Y^2 = \sigma_X^2$	$\sigma_Y^2 < \sigma_X^2$	$\mathbf{F} < 1/F_\alpha(n-1, m-1)$	$\mathbf{F} = s_y^2/s_x^2$
	$\sigma_Y^2 \neq \sigma_X^2$	$\mathbf{F} > F_{\alpha/2}(m-1, n-1)$ or $\mathbf{F} < 1/F_{\alpha/2}(n-1, m-1)$	

We include the cases $H_1\colon \sigma_Y^2 > \sigma_X^2$ and $H_1\colon \sigma_Y^2 < \sigma_X^2$ for completeness. For our test of equality of two normal distributions we only need to know if $\sigma_Y^2 \neq \sigma_X^2$. Note that the critical region consists of two intervals: $\mathbf{F} > F_{\alpha/2}(m-1, n-1)$ and $\mathbf{F} < 1/F_{\alpha/2}(n-1, m-1)$. This means that if the test statistic $\mathbf{F}$ falls into either of these intervals, we accept H_1. An approximate p-value can be found as in the previous chapters.

If the test determines that the variances are distinct (i.e. we accept H_1), then the entire test stops. The normal distributions are different, there is no need to test their means.

If the variances are found to be equal (i.e. we accept H_0), then the test continues.

16.3 Step 2: means

We test the null hypothesis $H_0\colon \mu_X = \mu_Y$ against the alternative $H_1\colon \mu_X \neq \mu_Y$. Since we already determined (at Step 1) that the variances were equal, we can use the facts from Section 6.5. We use the test statistic

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

which has a T-distribution $t(r)$ with $r = n + m - 2$ degrees of freedom.

Accordingly, the test proceeds as follows:

H_0	H_1	Critical region	Test statistic
	$\mu_X > \mu_Y$	$T > t_\alpha(r)$	
$\mu_X = \mu_Y$	$\mu_X < \mu_Y$	$T < -t_\alpha(r)$	$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$
	$\mu_X \neq \mu_Y$	$\lvert T\rvert > t_{\alpha/2}(r)$	

We include the cases $H_1\colon \mu_X > \mu_Y$ and $H_1\colon \mu_X < \mu_Y$ for completeness. For our test of equality of two normal distributions we only need to know if $\mu_X \neq \mu_Y$. An approximate p-value can be found as in the previous chapters.

Now we make the final conclusion. If the test determines that the means are distinct (i.e. we accept H_1), then the normal distributions are different. If the means are found to be equal (i.e. we accept H_0), then the normal distributions are identical.

16.4 Example

Suppose the calculus test scores of 10 students in one class yielded $\bar{x} = 73$ and $s_x = 25$ and the calculus test scores of 13 students in another class yielded $\bar{y} = 82$ and $s_y = 28$. Can we treat all these scores as values of the same normal random variable? Test this hypothesis at a 10% significance level.

Step 1. Testing variances. The test statistic is

$$\mathbf{F} = s_y^2/s_x^2 = 28^2/25^2 = 1.2544$$

Since

$$1.2544 < F_{0.05}(12,9) = 3.07$$

and

$$1.2544 > \frac{1}{F_{0.05}(9,12)} = \frac{1}{2.80} = 0.357$$

we are not in the critical region, so we accept H_0 (the variances are equal). We proceed to Step 2.

Step 2. Testing means. The test statistic is

$$T = \frac{73 - 82}{\sqrt{\frac{9\cdot 25^2 + 12\cdot 28^2}{10+13-2}}\sqrt{\frac{1}{10} + \frac{1}{13}}} = -0.7997$$

Since

$$|T| = 0.7997 < t_{0.05}(21) = 1.721$$

we are not in the critical region, so we accept H_0 (the means are equal).

The final conclusion: the calculus test scores from both classes can be treated as values of the same normal random variable.

16.5 Remark

Step 1 and step 2 can be used as separate tests, if we only want to test the identity between the means or that between the variances of the two normal distributions. For example, if we need to test the hypothesis $H_0\colon \mu_X = \mu_Y$ against the alternative $H_1\colon \mu_X > \mu_Y$ or $H_1\colon \mu_X > \mu_Y$ or $H_1\colon \mu_X \neq \mu_Y$ (and assume that the variances are equal), then we use the table in Step 2.

16.6 Remark

We will not try to compute the p-value of the combined test.

Actually, the test consists of two steps, and each step has its own p-value, so the p-value of the entire test should be the smaller of the two p-values.

16.7 Remark

In the previous example our test statistics were quite far from the corresponding critical values, so the acceptance of the null hypotheses was quite certain (it left no doubts). This demonstrates that the test is not very powerful, it easily misses (ignores) relatively small differences between the samples. The differences have to be large in order for the test to reject the null hypotheses.

17 χ^2 goodness-of-fit test

17.1 Proportions revisited

Recall the test on proportions from Section 14.2: suppose $Y = b(n,p)$ is a binomial random variable with an unknown 'proportion' p. If we test the null hypothesis $H_0\colon p = p_0$ against the two-sided alternative $H_1\colon p \neq p_0$, we use the Z-statistic

$$Z = \frac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}}$$

and the critical region is $|Z| > z_{\alpha/2}$. This critical region can also be expressed by $Z^2 > z^2_{\alpha/2}$. We denote

$$Q = Z^2 = \frac{(y/n - p_0)^2}{p_0(1-p_0)/n}$$

Now recall that in Section 7.2 we introduced a χ^2 random variable. We see now that the statistic $Q = Z^2$ is a χ^2 random variable with one degree of freedom. Since

$$\alpha = \mathbb{P}(|Z| > z_{\alpha/2}) = \mathbb{P}(Z^2 > z^2_{\alpha/2}) = \mathbb{P}(Q > z^2_{\alpha/2})$$

and

$$\alpha = \mathbb{P}\big(Q > \chi^2_\alpha(1)\big),$$

we conclude that $z^2_{\alpha/2} = \chi^2_\alpha(1)$. The critical region is now expressed as $Q > \chi^2_\alpha(1)$.

Let us also modify the expression for Q as follows. Denote by $p_1 = p_0$ the probability of success and $p_2 = 1 - p_0$ the probability of failure. Also let $y_1 = y$ denote the number of successes and $y_2 = n - y_1$ the number of failures. Then

$$\begin{aligned} Q &= \frac{(y_1/n - p_1)^2}{p_1p_2/n} = \frac{(y_1 - np_1)^2}{np_1p_2} \\ &= \frac{(y_1 - np_1)^2}{np_1} + \frac{(y_1 - np_1)^2}{np_2} \\ &= \frac{(y_1 - np_1)^2}{np_1} + \frac{(y_2 - np_2)^2}{np_2} \end{aligned}$$

Here we used two facts. First,

$$\frac{1}{p_1p_2} = \frac{1}{p_1} + \frac{1}{p_2}$$

which can be verified directly by using the obvious relation $p_1 + p_2 = 1$. Second, since $y_2 = n - y_1$ and $np_2 = n - np_1$, we have $y_2 - np_2 = -(y_1 - np_1)$, hence $(y_2 - np_2)^2 = (y_1 - np_1)^2$.

The formula for Q has a remarkable symmetry between successes and failures. In the first term we square the difference between the observed number of successes y_1 and the theoretically expected number of successes np_1 and divide it by the latter. In the second term we square the difference between the observed number of failures y_2 and the theoretically expected number of failures np_2 and divide it by the latter. Recall that for binomial random variables we only have two possible outcomes in every trial: success and failure.

17.2 Pearson's test

In 1900 K. Pearson extended the above scheme to trials with more that two possible outcomes. For example, when we roll a die, we observe one of six possible outcomes.

Suppose we perform n trials in which we observe one of k possible outcomes. In the end we count the number of times every outcome was observed: the first outcome was observed y_1 times, the second outcome y_2 times, etc. Of course,

$$y_1 + y_2 + \cdots + y_k = n \tag{link-1}$$

(the total number of trials). Suppose we expect the first outcome to come up with some probability p_1, the second outcome with some probability p_2, etc. Of course, $p_1 + \cdots + p_k = 1$. Then the (theoretically) expected number of times the outcomes should come up are $np_1,\ np_2, \dots, np_k$. Of course,

$$np_1 + np_2 + \cdots + np_k = n \tag{link-2}$$

By analogy with our 'binomial' statistic Q we compute

$$Q = \frac{(y_1 - np_1)^2}{np_1} + \cdots + \frac{(y_k - np_k)^2}{np_k}$$

Pearson proved that this Q-statistic has a χ^2 distribution with $k - 1$ degrees of freedom.

Now suppose we are testing the null hypothesis that our values $p_1, \dots, p_k$ are correct values of the probabilities of the outcomes in our trials. The alternative is 'everything else', i.e. H_1 simply says that these values of the probabilities (at least some of them) are incorrect.

To test this hypothesis with a significance level α, we compute the Q-statistic as above and check the critical region $Q > \chi^2_\alpha(k-1)$.

17.3 Example

Suppose we roll a die $n = 60$ times and observe 8 ones, 13 twos, 9 threes, 6 fours, 15 fives, and 9 sixes. Our goal is to test the hypothesis that the die is fair, i.e. that the probability of every outcome is 1/6. We use the 10% significance level.

To compute the Q-statistic, we list all the observed frequencies, then all the theoretically expected frequencies, then the corresponding differences (ignoring the signs):

8	13	9	6	15	9
10	10	10	10	10	10
2	3	1	4	5	1

The Q-statistic is

$$Q = \frac{2^2}{10} + \frac{3^2}{10} + \frac{1^2}{10} + \frac{4^2}{10} + \frac{5^2}{10} + \frac{1^2}{10} = \frac{56}{10} = 5.6$$

Sice $5.6 < \chi^2_{0.1}(5) = 9.236$, we accept the null hypothesis: the die appears to be fair (or ‘fair enough’).

17.4 Remarks

Note that the greatest contribution $5^2 = 25$ to the value of the final numerator (56) comes from a single outcome with the maximal discrepancy between ‘theory’ and ‘experiment’ (10 versus 15). This is quite typical for the χ^2 test: one ‘bad apple’ may ‘spoil’ the whole picture.

Also note that the Q-value 5.6 is quite far from its critical value 9.236 (even though we chose a pretty big risk level $\alpha = 0.1$). Thus, the χ^2 test does not appear to be very powerful – the chances of accepting the alternative hypothesis are not high, even if it is true. This is because the χ^2 test has a ‘universal’ alternative – it tests the null hypotheses against ‘everything else’. We already remarked in Section 14.5 that the alternative hypothesis should be well focused in order to increase its chances of acceptance. The nature of the χ^2 test does not allow any focusing.

17.5 Why 'degrees of freedom'?

It is time to explain the mysterious term 'degrees of freedom'. The frequencies $y_1, \ldots, y_k$ are, generally, independent from each other ('free variables') except that they must satisfy one constraint (link1). The probabilities $p_1, \ldots, p_k$ are independent variables ('free parameters') except for the constraint $p_1 + \ldots + p_k = 1$, see the corresponding equation (link2). In both cases, the constraint 'cancels' one degree of freedom – if we know $k-1$ frequencies (or probabilities), we can immediately compute the last one. This explains why there are $k-1$ degrees of freedom.

17.6 Example

When a person is trying to make up a random sequence of digits, he/she usually is avoiding repetitions or putting two numbers that differ by one next to each other (thinking that it would not look 'random'). This is a basis to detect whether the sequence is truly random or was made up.

Suppose in a sequence of 51 digits there are no repetitions and only 8 neighboring pairs differ by one. Is this a truly random sequence? Assume the 5% significance level.

Solution: the probability of a repetition is 1/10 and the probability of a pair of numbers differing by one is (approximately) 2/10. All the other pairs appear with probability $1 - 0.1 - 0.2 = 0.7$. This way we classify pairs into three categories, thus $k = 3$. We record the observed frequencies, then the theoretically expected frequencies, then the corresponding differences (ignoring the signs):

0	8	42
5	10	35
5	2	7

The Q-statistic is

$$Q = \frac{5^2}{5} + \frac{2^2}{10} + \frac{7^2}{35} = 6.8$$

Sice $6.8 > \chi^2_{0.05}(2) = 5.991$, we are in the critical region, thus we accept the alternative hypothesis: the sequence is not truly random, it is made up.

Note: this is one of the tests used by the IRS to detect tax fraud in tax return forms.

17.7 Remarks

The χ^2 test is the most popular in statistics, it has a long history and a solid reputation. Its advantage is the universality, it can be applied to almost any problem. The disadvantage is ... also the universality of the alternative hypothesis – the test cannot be focused to any specific alternative, it has to run against 'everything else' – as a result, its power is low.

The theoretical foundation of the χ^2 test makes use of some normal approximations to binomials. Those are considered to be accurate enough if all the theoretical frequencies satisfy $np_i \geq 5$. We need to verify this condition in our examples.

17.8 Example

Two friends, A and B, play a game by flipping a coin three times. If three heads come up, A pays B three dollars, if two heads and one tail, then A pays B one dollar, etc. They played 80 rounds of that game and recorded the results: three heads appeared 7 times, two heads 21 times, one head 36 times, and zero heads (three tails) 16 times. The friends suspected that the coin may not be really fair, as it lands on tails too often. They decided to verify their guess by using the χ^2 test with a 10% significance level.

Assuming that the coin is fair, the probabilities of all possible outcomes in this game are 1/8 for three heads or three tails and 3/8 for two heads (and one tail) or one head (and two tails). Below is the record of the observed frequencies, the theoretically expected frequencies, and the corresponding differences (ignoring the signs):

7	21	36	16
10	30	30	10
3	9	6	6

The Q-statistic is

$$Q = \frac{3^2}{10} + \frac{9^2}{30} + \frac{6^2}{30} + \frac{6^2}{10} = 8.4$$

Sice $8.4 > \chi^2_{0.1}(3) = 6.25$, we are in the critical region, thus we accept the alternative hypothesis: the coin is not fair, it is 'loaded'!

17.9 Example continued

Now the friends want to estimate the probability that the coin lands heads and redo the test. The compute the total number of tosses $80 \times 3 = 240$ and the total number of times the coin landed on heads: $7\times 3+21\times 2+36=99$. Then they come up with an estimate $99/240=0.4125$.

Thus the probabilities of the above three outcomes are, according to the binomial distribution:

$$\mathbb{P}(3\text{ heads})=0.4125^3=0.07,\quad \mathbb{P}(2\text{ heads})=3\times 0.4125^2\times 0.5875=0.3$$

$$\mathbb{P}(1\text{ head})=3\times 0.4125\times 0.5875^2=0.427,\quad \mathbb{P}(0\text{ heads})=0.5875^3=0.203$$

Now the frequency table looks like

7	21	36	16
5.6	24	34.16	16.24
1.4	3	1.84	0.24

The Q-statistic is

$$Q=\frac{1.4^2}{5.6}+\frac{3^2}{24}+\frac{1.84^2}{34.16}+\frac{0.24^2}{16.24}=0.827$$

Its value is very small, so the null hypothesis will be surely accepted. But! The critical region changes: it is now $Q>\chi^2_{0.1}(2)=4.605$. The number of degrees of freedom has changed: it is 2 instead of 3. Why? Because we have estimated one parameter in the model: the probability of landing on heads. Each estimate of an unknown parameter used in the χ^2 test creates a new link between the frequencies, and it reduces the number of degrees of freedom by one. Generally, if r parameters are estimated, then the number of degrees of freedom is $k-1-r$.

The rest of this Chapter is optional...

17.10 Example

A small company recorded the number of orders it received every day during a period of 50 days. Suppose there were no days without orders, one day with one order, two days with two orders, 10 days with three orders, 9 days with four orders, 6 days with five orders, 5 days with six orders, 7 days with seven orders, 3 days with eight orders, 5 days with nine orders, one day with ten orders, and one days with eleven orders.

The company wants to treat the number of orders per day as a Poisson random variable for some important analysis. Before doing this it wants to test if this assumption is correct, i.e. if the number of orders per day can be treated as values of a Poisson random variable.

First of all, a Poisson random variable has a parameter λ, whose value is unknown and needs to be estimated. Its best estimate is

$$\hat{\lambda} = \bar{x} = \frac{1 \times 1 + 2 \times 2 + 3 \times 10 + 4 \times 9 + \cdots + 11 \times 1}{50} = 5.4$$

Now the probability that x orders are received on a day can be computed by the Poisson formula

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

The following table represents all recorded outcomes, their observed frequencies x_i and their expected theoretical frequencies np_i for $i = 1, \dots, 11$:

outcomes	0	1	2	3	4	5	6	7	8	9	10	11
observation	0	1	2	10	9	6	5	7	3	5	1	1
theory	0.2	1.2	3.3	5.9	8	8.7	7.8	6	4	2.4	1.3	0.6

But here we have two problems. First of all, there are many more outcomes that have not been observed: 0, 12, 13,.... There are infinitely many of them! Second, some of the theoretical frequencies are less than 5 (failing to satisfy the important condition of the applicability of the test).

To solve these problems we *group* outcomes. Outcomes whose theoretical frequencies are low can be combined into one or more group so that the combined theoretical frequencies will satisfy the condition $np_i \geq 5$.

We combine small outcomes 0, 1, 2, 3 into one group with the combined theoretical frequency 10.65. And we combine large outcomes 8, 9, 10, 11, ... into another group with the combined theoretical frequency 8.9. Now the table looks like this:

outcomes	≤ 3	4	5	6	7	≥ 8
observation	13	9	6	5	7	10
theory	10.65	8	8.65	7.8	6	8.9
difference	1.35	1	2.65	2.8	1	1.1

The Q-statistic is

$$Q = \frac{1.35^2}{10.65} + \frac{1^2}{8} + \frac{2.65^2}{8.65} + \frac{2.8^2}{7.8} + \frac{1^2}{6} + \frac{1.1^2}{8.9} = 2.763$$

Since we made six groups of outcomes (there are six terms in the Q-formula), we have $k = 6$. Also, we have estimated one parameter, so the number of degrees of freedom is $6 - 1 - 1 = 4$. The critical region is $Q > \chi^2_\alpha(4)$.

To complete the test, suppose $\alpha = 0.05$. Then since $2.763 < \chi^2_{0.05}(4) = 9.488$, we accept the null hypothesis: the number of orders per day is, indeed, a Poisson random variable.

What is the p-value of the test? Based on Table IV, we only can say that it is greater than 0.1. The one-line calculator on the instructor's web page gives a precise value 0.598. This means the alternative hypothesis can only be accepted if one takes an unrealistically high risk of 59.8%. Nobody takes such a risk in statistics, so the null hypothesis is accepted beyond doubts.

17.11 Remarks

Combining outcomes with small probabilities has advantages and disadvantages. On the one hand, it allows us to form groups that have high enough probabilities and then run the χ^2 test. On the other hand, combining outcomes leads to a certain loss of details of information. This issue is similar to the trade-off between larger bins and smaller bins when constructing a histogram, recall Section 1.5.

In the previous example, we have verified that the data can be treated as values of a Poisson random variable, i.e. that the Poisson distribution *fits* out data well. This explains the name *goodness-of-fit test.*

17.12 Fitting continuous random variables

In the previous example we dealt with a discrete random variable (Poisson). Suppose now that we observe values $x_1, \dots, x_n$ of a continuous random variable. Then we may want to check if that random variable belongs to a particular type, such as exponential or normal. This also can be done by the χ^2 test.

First, we estimate the unknown parameters. Then we need to divide the entire range of possible values into several intervals. For each interval we count the number of observed points x_i in it (these numbers will be treated as frequencies), as well as compute the probability of being in each interval. Then we form the Q-statistic. If there are k intervals and we have estimated r parameters, then the critical region will be $Q > \chi^2_\alpha(k - 1 - r)$.

This approach is common in many applications. It requires a careful selection of intervals (not too big and not too small), just like in the construction of histograms in Section 1.5.

18 Contingency tables

In this and next chapters we discuss several variations of the χ^2 test.

18.1 Example

In a university, officials want to compare the grades received by male students versus those of female students. They pick at random 50 male students and 50 female students and record their calculus grades:

	A	B	C	D	F	Total
Female	8	13	16	10	3	50
Male	4	9	14	16	7	50
Total	12	22	30	26	10	100

Is there a sufficient evidence to claim that distributions of grades for male and female students are different, or should we conclude that they are comparable (homogeneous)? What we do here is the *homogeneity test.*

18.2 Test of equality of two distributions: setting

More generally, suppose trials are performed that have k possible outcomes. In one experiment, n_1 such trials are performed, and the recorded frequencies are $y_{11}, \dots, y_{k1}$. In another experiment, n_2 such trials are performed, and the recorded frequencies are $y_{12}, \dots, y_{k2}$. Note that the first index refers to the outcome, and the second to the experiment. The data can be presented by a 'contingency table':

	1	2	...	k
Exp-I	y_{11}	y_{21}	...	y_{k1}
Exp-II	y_{12}	y_{22}	...	y_{k2}

The probabilities of the k outcomes in the first experiment $p_{11}, \dots, p_{k1}$ are unknown, and so are the probabilities of these k outcomes in the second experiment $p_{12}, \dots, p_{k2}$. We want to test the null hypothesis

$$H_0\colon p_{11} = p_{12}, \dots, p_{k1} = p_{k2}$$

against the alternative that is again sort of 'everything else', i.e. H_1 simply says that H_0 is false.

18.3 Test of equality of two distributions: procedure

Since p_{ij} are unknown, they must be estimated first. Under the null hypothesis, $p_{i1} = p_{i2}$ is just one unknown parameter for each $i = 1, \dots, k$. Its best estimate is obtained by combining data from both experiments:

$$\hat{p}_{i1} = \hat{p}_{i2} = \frac{y_{i1} + y_{i2}}{n_1 + n_2}$$

(the total number of occurrences of the ith outcome over the total number of trials).

Then we compute the Q-statistic by the same general formula as in the previous chapter:

$$Q = \sum_{i=1}^{k} \frac{(y_{i1} - n_1\hat{p}_{i1})^2}{n_1\hat{p}_{i1}} + \sum_{i=1}^{k} \frac{(y_{i2} - n_2\hat{p}_{i2})^2}{n_2\hat{p}_{i2}}$$

The critical region is then $Q > \chi^2_\alpha(r)$, where α is the significance level and r is the number of degrees of freedom.

What is r here? Originally, we have $2k$ random variables y_{ij}. There are two obvious links $y_{11} + \cdots + y_{k1} = n_1$ and $y_{12} + \cdots + y_{k2} = n_2$, which eliminate 2 degrees of freedom. And we have estimated $k-1$ parameters $\hat{p}_{i1} = \hat{p}_{i2}$ for $i = 1, \dots, k-1$ (the last one, $\hat{p}_{k1} = \hat{p}_{k2}$, need not be estimated since the probabilities must add up to one). Thus the total number of degrees of freedom is

$$r = 2k - 2 - (k-1) = k - 1$$

Note that the number of degrees of freedom equals $(k-1) \cdot (2-1)$. Later we will see that it is a general formula:

$$r = (\text{number of rows} - 1) \cdot (\text{number of columns} - 1)$$

18.4 Example 18.1 finished

In our example with male and female students, we have $\hat{p}_{11} = 12/100 = 0.12$, $\hat{p}_{21} = 0.22$, $\hat{p}_{31} = 0.3$, $\hat{p}_{41} = 0.26$, and $\hat{p}_{51} = 0.10$. Then

$$Q = \frac{(8-6)^2}{6} + \frac{(13-11)^2}{11} + \cdots + \frac{(7-5)^2}{5} = 5.18$$

Since $5.18 < \chi^2_{0.05}(4) = 9.488$ (here we assume $\alpha = 0.05$), we accept H_0. That is, both groups of students have similar distributions of grades.

The p-value (according to the on-line calculator on the instructor's web page) is 0.2693.

18.5 Remark

A quick look at the table in Section 18.1 suggests that male students do get lower grades: fewer A's and B's, but more D's and F's. But the χ^2 test was not able to recognize this. We see again that the test is not very powerful, and the reason is its universality: it is unable to 'focus' on any specific alternative hypothesis, it simply checks H_0 against 'everything else'.

18.6 Independence test

We may look at Example 18.1 differently: do the grades depend on the gender of the students? Or are these two attributes independent? So the test described above applies whenever we want to test the independence of two attributes.

In such experiments, every observation comes with two attributes (like every student has a certain gender and gets a certain grade). We count the observed frequency for every pair of values of these two attributes, make a contingency table, and proceed with the χ^2 test as above.

18.7 Example

Let us revisit Example 14.4. There, every patient has two attributes: his/her condition (ill or healthy) and the result of the hearing test (pass or fail). The doctor's theory is that these two attributes are related (correlated). So the doctor can use the χ^2 independence test to check his theory. We recall the experimental results (contingency table):

	Fail	Pass	Total
Ill	71	25	96
Healthy	76	43	119
Total	147	68	215

We first estimate the probabilities:

$$\hat{p}_{11} = \frac{147}{215} = 0.6837, \qquad \hat{p}_{21} = \frac{68}{215} = 0.3163$$

and compute the theoretically expected frequencies:

$$96 \cdot 0.6837 = 65.6, \qquad 96 \cdot 0.3163 = 30.4$$

$$119 \cdot 0.6837 = 81.4, \qquad 119 \cdot 0.3163 = 37.6$$

Then compute the Q-statistic

$$Q = \frac{(71-65.6)^2}{65.6} + \frac{(25-30.4)^2}{30.4} + \frac{(76-81.4)^2}{81.4} + \frac{(43-37.6)^2}{37.6} = 2.54$$

The p-value (according to the on-line calculator on the instructor's web page) is 0.111.

We see that this p-value is almost identical to the one we got using the binomial test in Section 14.5 against the two-sided alternative. (There the p-value was 0.114, the small difference is entirely due to round-off errors.)

Again we see that the χ^2 test can only deal with a 'universal' alternative (which covers 'everything else'). On the contrary, the binomial test used in Section 14.4 could be made more focused (one-sided), and thus it was able to reduce the p-value to 0.057.

18.8 Final remark

Nonetheless, our doctor chose to use the χ^2 test, rather than the binomial test, despite the lower power of the former. The doctor's rationale was that the χ^2 had a very high reputation, and its conclusion would be accepted by the medical community. The binomial test, on the other hand, is less known and looks as something 'special', 'hand-made', and 'unreliable', thus its results might be doubtful.

19 Test about several means

19.1 Example

In a university, several professors teach different sections of Calculus-I, after which the students take a common final exam. The officials want to compare the average performance of students from different sections. Do students taught by different professors receive significantly different average scores, or are all the averages comparable (so that there is no significant difference)?

The officials assume that the scores in each section have a normal distribution with the same variance (in all sections), but possibly different means. They want to determine if the means are significantly different or not.

19.2 Test of equality of several means: settings

Suppose we have several samples from different normal distributions:

$$\begin{array}{ll} x_{11},\dots,x_{1n_1} & \text{from } \mathcal{N}(\mu_1,\sigma^2) \\ x_{21},\dots,x_{2n_2} & \text{from } \mathcal{N}(\mu_2,\sigma^2) \\ \cdots & \cdots \\ x_{m1},\dots,x_{mn_m} & \text{from } \mathcal{N}(\mu_m,\sigma^2) \end{array}$$

The mean values $\mu_1,\dots,\mu_m$ and the (common) variance σ^2 are unknown. We are testing the hypothesis

$$H_0\colon \mu_1=\mu_2=\cdots=\mu_m$$

against a universal alternative (which says that at least some of the means are different).

The unknown parameters $\mu_1,\dots,\mu_m$ are estimated by the sample means

$$\bar{x}_{1\cdot}=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{1i},\quad \dots,\quad \bar{x}_{m\cdot}=\frac{1}{n_m}\sum_{i=1}^{n_m}x_{mi}$$

Note that $\bar{x}_{i\cdot}$ denotes the sample mean within the ith sample; the dot indicates that the second index is eliminated by summation. We also denote

$$\bar{x}_{\cdot\cdot}=\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j}x_{ji}$$

the *grand mean* (here $n = n_1 + \cdots + n_m$).

The individual observations x_{ji} within each sample vary about the corresponding sample mean $\bar{x}_{j\cdot}$, and sample means $\bar{x}_{j\cdot}$ vary about the grand mean $\bar{x}_{\cdot\cdot}$. While the former reflect natural statistical data variations, the latter may reflect possible differences between means μ_j. Our strategy is to 'separate' these two variations and compare their values. This procedure is known as *analysis of variances* (ANOVA).

19.3 Analysis of variances

We compute the total (TO) sum of squares (SS) of variations:

$$\begin{aligned}
\mathrm{SS(TO)} &= \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_{\cdot\cdot})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_{j\cdot} + \bar{x}_{j\cdot} - \bar{x}_{\cdot\cdot})^2 \\
&= \underbrace{\sum_{j}\sum_{i}(x_{ji} - \bar{x}_{j\cdot})^2}_{\mathrm{SS(E)}} + \underbrace{\sum_{j} n_j(\bar{x}_{j\cdot} - \bar{x}_{\cdot\cdot})^2}_{\mathrm{SS(T)}} \\
&\quad + \underbrace{2\sum_{j}\sum_{i}(x_{ji} - \bar{x}_{j\cdot})(\bar{x}_{j\cdot} - \bar{x}_{\cdot\cdot})}_{=0} \\
&= \mathrm{SS(E)} + \mathrm{SS(T)}
\end{aligned}$$

Fortunately, the double sum of cross-products vanishes (all its terms cancel out). The first sum of squares reflects statistical *errors* within samples, the second sum of squares reflects differences between samples (*treatments*).

The terminology here is borrowed from medical sciences, where patients are given different treatments in order to test various methods or types of medicines.

19.4 Test procedure

The following facts are established in (advanced) probability theory:

$$\frac{\mathrm{SS(TO)}}{\sigma^2} \quad \text{is } \chi^2(n-1)$$

$$\frac{\mathrm{SS(E)}}{\sigma^2} \quad \text{is } \chi^2(n-m)$$

$$\frac{\mathrm{SS(T)}}{\sigma^2} \quad \text{is } \chi^2(m-1)$$

(note that $(n-m)+(m-1)=n-1$). Also, the statistics SS(E) and SS(T) are independent.

We cannot use the χ^2 values since σ^2 is unknown. But the ratio

$$\mathbf{F} = \frac{\frac{\mathrm{SS(T)}}{\sigma^2(m-1)}}{\frac{\mathrm{SS(E)}}{\sigma^2(n-m)}} = \frac{\mathrm{SS(T)}/(m-1)}{\mathrm{SS(E)}/(n-m)}$$

does not contain σ^2, which cancels out, and it has an F-distribution with $m-1$ and $n-m$ degrees of freedom. Hence the critical region is

$$\mathbf{F} > F_\alpha(m-1, n-m)$$

where α is the significance level.

The test is usually summarized in the so called *ANOVA table*:

	SS	DoF	MS	F	p-value
Treatment	SS(T)	$m-1$	SS(T)/$(m-1)$		
Error	SS(E)	$n-m$	SS(E)/$(n-m)$	**F**	$\cdots$
Total	SS(TO)	$n-1$	SS(TO)/$(n-1)$		

Here DoF stands for *degrees of freedom* and MS for *mean squares.*

19.5 Example

Four samples, each with three observations, yield the following results:

				$\bar{x}$
X_1 :	13	8	9	10
X_2 :	15	11	13	13
X_3 :	8	12	7	9
X_4 :	11	15	10	12
	grand mean:			11

Note that $m = 4$ and $n_1 = n_2 = n_3 = n_4 = 3$.

We compute the test statistics:

$$\mathrm{SS(E)} = 3^2 + 2^2 + 1^2 + 2^2 + 2^2 + 0^2 + 1^2 + 3^2 + 2^2 + 1^2 + 3^2 + 2^2 = 50$$

$$\mathrm{SS(T)} = 3\,(10 - 11)^2 + 3\,(13 - 11)^2 + 3\,(9 - 11)^2 + 3\,(12 - 11)^2 = 30$$

$$\mathbf{F} = \frac{30/3}{50/8} = 1.6$$

Assume that $\alpha = 0.05$. Since $1.6 < F_{0.05}(3, 8) = 4.07$, then we accept the null-hypothesis H_0: there is no significant differences between the means of these four random variables. The p-value (obtained by the on-line calculator) is 0.2642.

The ANOVA table looks like this:

	SS	DoF	MS	F	p-value
Treatment	30	3	30/3		
Error	50	8	50/8	1.6	0.2642
Total	80	11	80/11		

20 Two-factor analysis of variances

20.1 Example

Suppose the auto industry engineers want to determine which factors affect the gas mileage of cars. In a simplified experiment, they test several types of cars with several brands of gasoline and record the observed gas mileage for each pair of “car + brand of gas”.

Here is a table describing such an experiment with 3 types of cars and 4 brands of gas:

		gasoline				
		1	2	3	4	mean
car	1	16	18	21	21	19
	2	14	15	18	17	16
	3	15	15	18	16	16
	mean	15	16	19	18	17

The last column contains the row means, and the last row contains the column means. The bottom right value 17 is the *grand mean*.

We see that there are some variations within the table, some variations between columns and some – between rows. The question is whether those variations are significant to conclude that the type of car or the brand of gasoline (or both) affect the gas mileage.

20.2 General setting

Suppose in an experiment, the observed result depends on two factors. The first factor has a levels (values) and the second factor has b levels (values). For each combination of values of these two factors, an experimental observation is recorded. Thus we get an $a \times b$ table of observations:

		2nd factor				
		1	2	$\cdots$	b	mean
1st factor	1	x_{11}	x_{12}	$\cdots$	x_{1b}	$\bar{x}_{1\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	a	x_{a1}	x_{a2}	$\cdots$	x_{ab}	$\bar{x}_{a\cdot}$
	mean	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\cdots$	$\bar{x}_{\cdot b}$	$\bar{x}_{\cdot\cdot}$

We use the same convention for denoting means with dots, as in the previous chapter.

Here we are testing two separate hypothesis: H_A : there is no significant variations between rows (i.e., the 1st factor does not affect the observed results), and H_B : there is no significant variations between columns (i.e., the 2nd factor does not affect the observed results). Each hypothesis has its own alternative, which says that there is a significant variation (i.e. the corresponding factors affects the outcome).

20.3 Test procedure

We analyze variances in a way similar to the previous section. Again we compute the total (TO) sum of squares (SS) of variations:

$$\begin{aligned}
\mathrm{SS(TO)} &= \sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\bar{x}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\big[(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot})+(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})+(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot})\big]^2 \\
&= \underbrace{\sum_{i}\sum_{j}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x}_{\cdot\cdot})^2}_{\mathrm{SS(E)}} \\
&\quad + \underbrace{b\sum_{i=1}^{a}(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2}_{\mathrm{SS(A)}} + \underbrace{a\sum_{j=1}^{b}(\bar{x}_{\cdot j}-\bar{x}_{\cdot\cdot})^2}_{\mathrm{SS(B)}} + 2\underbrace{\sum_{i}\sum_{j}\cdots}_{=0} \\
&= \mathrm{SS(E)}+\mathrm{SS(A)}+\mathrm{SS(B)}
\end{aligned}$$

Fortunately again, all the double sums of cross-products vanish (all their terms cancel out), and we do not even include them explicitly.

The first sum of squares reflects statistical *errors* within samples, the second and third sums of squares reflect variations between rows and columns, respectively.

The following facts are established in (advanced) probability theory:

$$\frac{\mathrm{SS(E)}}{\sigma^2} \quad \text{is } \chi^2\big((a-1)(b-1)\big)$$

$$\frac{\mathrm{SS(A)}}{\sigma^2} \quad \text{is } \chi^2(a-1)$$

$$\frac{\mathrm{SS(B)}}{\sigma^2} \quad \text{is } \chi^2(b-1)$$

Here σ^2 denotes the unknown variance of statistical errors. Also, the statistic SS(E) is independent from SS(A) and SS(B).

We cannot use the χ^2 values since σ^2 is unknown. But the ratio

$$F_A = \frac{\mathrm{SS(A)}/(a-1)}{\mathrm{SS(E)}/[(a-1)(b-1)]}$$

has an F-distribution with $a-1$ and $(a-1)(b-1)$ degrees of freedom. Hence the critical region for testing the hypothesis H_A is

$$F_A > F_\alpha\big(a-1, (a-1)(b-1)\big)$$

where α is the significance level.

Similarly, the ratio

$$F_B = \frac{\mathrm{SS(B)}/(b-1)}{\mathrm{SS(E)}/[(a-1)(b-1)]}$$

has an F-distribution with $b-1$ and $(a-1)(b-1)$ degrees of freedom. Hence the critical region for testing the hypothesis H_B is

$$F_B > F_\alpha\big(b-1, (a-1)(b-1)\big)$$

where α is the significance level.

20.4 Example (continued)

We return to our example with 3 cars and 4 brands of gasoline. The statistic SS(E) is the hardest to compute:

$$\mathrm{SS(E)} = (16 - 15 - 19 + 17)^2 + \cdots = 4$$

Now

$$\mathrm{SS(A)} = 4\,(2^2 + 1^2 + 1^2) = 24$$

$$\mathrm{SS(B)} = 3\,(2^2 + 1^2 + 2^2 + 1^2) = 30$$

Thus

$$F_A = \frac{24/2}{4/6} = 18$$

and

$$F_B = \frac{30/3}{4/6} = 15$$

Let us pick $\alpha = 0.01$. Since $18 > F_{0.01}(2,6) = 10.92$, we reject H_A. Since $15 > F_{0.01}(3,6) = 9.78$, we reject H_B as well. Note that the critical values for the two hypotheses are different. The p-values (obtained by the on-line calculator) are 0.0029 for the hypothesis H_A and 0.0034 for the hypothesis H_B. Both p-values are well below 1%, so the rejection of both hypotheses is very safe.

Our final conclusion is that there is a significant variations both between rows and between columns (hence, the type of a car and the brand of gasoline both affect the gas mileage).

20.5 Extension

In the previous example, we had one observation for each combination of values of the two factors (one data per cell in the table). Suppose now we have $c > 1$ observations in each cell, we denote them by x_{ijk}, where i and j are the levels of the two factors and $k = 1, \dots, c$ is the index of individual observations for the given pair i, j of values of the factors.

Now, with more data available, we can test an extra hypothesis in this experiment.

The analysis of variances now is more complicated:

$$\begin{aligned}
\mathrm{SS(TO)} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(x_{ijk}-\bar{x}_{\dots})^2 \\
&= \underbrace{bc\sum_{i=1}^{a}(\bar{x}_{i\cdot\cdot}-\bar{x}_{\dots})^2}_{\mathrm{SS(A)}} + \underbrace{ac\sum_{j=1}^{b}(\bar{x}_{\cdot j\cdot}-\bar{x}_{\dots})^2}_{\mathrm{SS(B)}} \\
&\quad \underbrace{+\,c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{ij\cdot}-\bar{x}_{i\cdot\cdot}-\bar{x}_{\cdot j\cdot}+\bar{x}_{\dots})^2}_{\mathrm{SS(AB)}} + \underbrace{\sum_{i}\sum_{j}\sum_{k}(x_{ijk}-\bar{x}_{ij\cdot})^2}_{\mathrm{SS(E)}} \\
&= \mathrm{SS(A)}+\mathrm{SS(B)}+\mathrm{SS(AB)}+\mathrm{SS(E)}
\end{aligned}$$

where we used the same general rule for denoting sample means. For example,

$$x_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c}x_{ijk}, \quad x_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c}x_{ijk}$$

and so on.

Now we can test three hypotheses: H_A and H_B, as before, and H_{AB} – about *interactions* between the factors A and B. It might happen that not only the factors A and B have direct affect on the observations, but also certain pairs of values of A and B have an extra effect. For example, the car 1 may have the best gas mileage, the gasoline brand 2 may have the best gas mileage, but the car 1 and gasoline 2 may “not mix too well”, so that this pair may have a poor gas mileage. In that case the interaction between the factors has a significant effect and must be included in the analysis.

We test these three hypotheses as follows. The factor A is significant if

$$F_A = \frac{\mathrm{SS(A)}/(a-1)}{\mathrm{SS(E)}/[ab(c-1)]} > F_\alpha\big(a-1, ab(c-1)\big)$$

The factor B is significant if

$$F_B = \frac{\mathrm{SS(B)}/(b-1)}{\mathrm{SS(E)}/[ab(c-1)]} > F_\alpha\big(b-1, ab(c-1)\big)$$

The interaction between the factors A and B is significant if

$$F_{AB} = \frac{\mathrm{SS(AB)}/[(a-1)(b-1)]}{\mathrm{SS(E)}/[ab(c-1)]} > F_\alpha\big((a-1)(b-1), ab(c-1)\big)$$

Here α is the significance level of the test.

In practice, one usually starts by testing the hypotheses A and B. If both factors are determined to be significant, then one tests the interactions. If one of the factors is not significant, then there is no need to test interactions.

20.6 Three factors

It is interesting to discuss a model which involves *three* factors: A, B, and C. For simplicity, assume that each factor has two levels (say, *low* and *high*). We denote the low value by $-$ and the high value by $+$.

For example, it is common in industry to analyze various factors that may affect the quality of the product. In a preliminary test, several potentially significant factors are selected, and for each factor two values (a lower and a higher) are set. Then for each combination of values of all the selected factors an experimental product is manufactured and its quality measured.

All possible combinations of three factors are represented by sequences of pluses and minuses of length three, from $---$ to $+++$. For example, $---$ corresponds to the lower values of all the factors, etc. There are $2 \times 2 \times 2 = 8$ such sequences. For each sequence (combination of values of the factors), a single experimental value is observed, we denoted them by $x_1, \dots, x_8$:

run	A	B	C	observ.	AB	AC	BC	ABC
1	$-$	$-$	$-$	x_1	$+$	$+$	$+$	$-$
2	$+$	$-$	$-$	x_2	$-$	$-$	$+$	$+$
3	$-$	$+$	$-$	x_3	$-$	$+$	$-$	$+$
4	$+$	$+$	$-$	x_4	$+$	$-$	$-$	$-$
5	$-$	$-$	$+$	x_5	$+$	$-$	$-$	$+$
6	$+$	$-$	$+$	x_6	$-$	$+$	$-$	$-$
7	$-$	$+$	$+$	x_7	$-$	$-$	$+$	$-$
8	$+$	$+$	$+$	x_8	$+$	$+$	$+$	$+$

This table is extended by four columns corresponding to interactions: three pairwise interactions AB, AC, and BC, and the triple interaction ABC (between all the factors). Those columns are obtained by 'multiplying' the corresponding columns A, B, and C. The 'multiplication rules' are these: we treat + as +1 and − as −1. Therefore, + times − is −, for example. Also, − times − is +, etc.

Now we can test 7 hypotheses: about the significance of individual factors A, B, and C, about the significance of the pairwise interactions AB, AC, and BC, and about the significance of the triple interaction ABC. We compute seven test statistics:

$$\begin{aligned}
[A] &= (-x_1 + x_2 - x_3 + x_4 - x_5 + x_6 - x_7 + x_8)/8\\
[B] &= (-x_1 - x_2 + x_3 + x_4 - x_5 - x_6 + x_7 + x_8)/8\\
[C] &= (-x_1 - x_2 - x_3 - x_4 + x_5 + x_6 + x_7 + x_8)/8\\
[AB] &= (+x_1 - x_2 - x_3 + x_4 + x_5 - x_6 - x_7 + x_8)/8\\
[AC] &= (+x_1 - x_2 + x_3 - x_4 - x_5 + x_6 - x_7 + x_8)/8\\
[BC] &= (+x_1 + x_2 - x_3 - x_4 - x_5 - x_6 + x_7 + x_8)/8\\
[ABC] &= (-x_1 + x_2 + x_3 - x_4 + x_5 - x_6 - x_7 + x_8)/8
\end{aligned}$$

Note that the sequence of signs in each line is taken from the corresponding column of the table.

20.7 Example

Suppose the following values are observed:

$$\begin{array}{llll}
x_1 = 41.0 & x_2 = 30.5 & x_3 = 47.7 & x_4 = 27.0\\
x_5 = 39.5 & x_6 = 26.5 & x_7 = 48.0 & x_8 = 27.5
\end{array}$$

The test statistics are computed as follows:

[A]	[B]	[C]	[AB]	[AC]	[BC]	[ABC]
−8.06	1.56	0.56	−2.19	−0.31	0.81	0.31

Now we construct a plot consisting of seven points. Their x-coordinates are the ordered statistics:

$$-8.06,\ -2.19,\ -0.31,\ 0.31,\ 0.56,\ 0.81,\ 1.56$$

The y-coordinates of our seven points are the "equally spaced" percentiles of the standard normal distribution $Z = \mathcal{N}(0,1)$, i.e. $z_{1/8}$, $z_{2/8}$, $z_{3/8}$, $z_{4/8}$, $z_{5/8}$, $z_{6/8}$, $z_{7/8}$; the values of these percentiles are

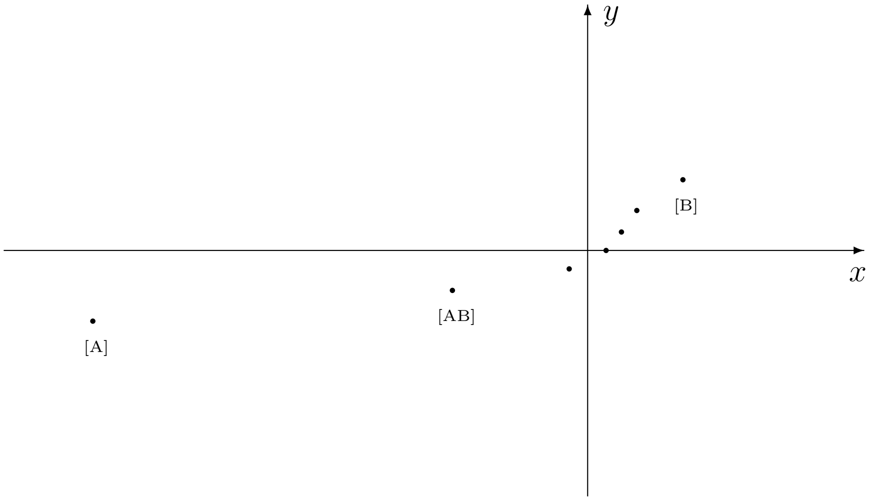
$$-1.15,\ -0.67,\ -0.32,\ 0.00,\ 0.32,\ 0.67,\ 1.15$$

Thus our seven points have the following coordinates:

$$(-8.06, -1.15),\ (-2.19, -0.67), \ldots, (0.81, 0.67),\ (1.56, 1.15).$$

Assuming that none of the factors or their combinations are significant, these seven points should lie approximately on a straight line on the xy plane. However, if some of these points appear 'out of line' (outliers), they indicate the factors or interactions that are significant!

The seven points in our example are plotted below. We see a linear pattern in the middle (near the origin), but some points are 'out of line'. The point that is the farthest from the line is $[A] = (-8.06, -1.15)$, it indicates that the factor A is the most significant. Two more points $[B] = (1.56, 1.15)$ and $[AB] = (-2.19, -0.67)$ appear to be somewhat out-of-line, so the factor B and the interaction AB are of some significance. Other factors and interactions appear to be insignificant.



21 Regression

In the previous section certain points (x_i, y_i) were expected to approximately lie on a straight line. Such things happen in many applications.

For example, let $x_1, \dots, x_n$ be SAT scores of n students and $y_1, \dots, y_n$ their scores in a calculus test in college. We expect that those who got higher SAT scores would get higher calculus scores, so these n points should lie on a certain curve or line with a positive slope.

Or let $s_1, \dots, s_n$ be the values of a stock market index recorded on n consecutive days of trading. Then we may expect that the points $(1, s_1), \dots, (n, s_n)$, where the first coordinate represents time (the day counter), lie on a certain curve describing the evolution of the stock market. If we knew that curve, we would be able to predict the behavior of the stock index in the future!

21.1 Regression

Suppose we want to determine a curve $y = g(x)$ that is best to describe a set of observed pairs of values $(x_1, y_1), \dots, (x_n, y_n)$. That is, we assume that these points approximately lie on a curve and want to determine the equation of that curve.

It is commonly assumed that $x_1, \dots, x_n$ are not random but $y_1, \dots, y_n$ are random (say, x_i represent the time moments when the values y_i are observed and recorded, like in the stock market example above). Then the function $y = g(x)$ is called the *regression* of y on x, and determining that function is called the *regression problem.*

In practical applications, regression is used to estimate (predict) the value of $y = g(x)$ for various values of x. We often call x the *explanatory* or *predictor* variable, and y the *response* variable.

21.2 Maximum likelihood method

Assume that each random value y_i is given by

$$y_i = g(x_i) + \varepsilon_i$$

where $g(x_i)$ is the actual value of the (unknown) function at the point x_i and ε_i is the statistical error (measurement error). It is usually assumed that $\varepsilon_1, \dots, \varepsilon_n$ are independent normal random variables with zero mean and a common variance σ^2. That is, $\varepsilon_i = \mathcal{N}\big(0, \sigma^2\big)$. The variance σ^2 is unknown.

Then $y_i = \mathcal{N}(g(x_i), \sigma^2)$, and its density function is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y_i - g(x_i))^2}{2\sigma^2}}$$

The joint density function (or likelihood function) is

$$L = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum (y_i - g(x_i))^2}{2\sigma^2}}$$

Its logarithm is

$$\ln L = -n \ln \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - g(x_i))^2$$

Now suppose we want to find the best possible curve $y = g(x)$ by the maximum likelihood method, i.e. by maximizing $\ln L$. Clearly, to make $\ln L$ bigger we need to make

$$H = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

smaller, so we want to minimize the value of H, i.e. find a curve such that the sum of squares of the distances from our points to the curve is as small as possible. This is called the *least squares method* or the *least squares fit* (LSF).

Suppose we found such a curve $y = g(x)$, then we can estimate σ^2 by using the maximum likelihood method again. Solving the equation

$$0 = \frac{d}{d\sigma^2} \ln L = -\frac{n}{2\sigma^2} + \frac{\sum (y_i - g(x_i))^2}{2\sigma^4}$$

gives the MLE for σ^2

$$\hat{\sigma}^2 = \frac{\sum (y_i - g(x_i))^2}{n} = \frac{H}{n}$$

The values $r_i = y_i - g(x_i)$ are often called *errors* (of observations) or *residuals*, and H is called the *residual sum of squares* (RSS). Now

$$\hat{\sigma} = \sqrt{\frac{\sum (y_i - g(x_i))^2}{n}}$$

can be called the *root mean squared error*.

21.3 Linear regression

In the simples case, the unknown curve is a line given by equation

$$y = \alpha_1 + \beta x$$

where α_1 is the intercept and β is the slope. These are the parameters to be estimated. For convenience, we change parameter $\alpha_1 = \alpha - \beta\bar{x}$, where $\bar{x}$ is the sample mean (the average) of $x_1, \dots, x_n$. Then the equation of the unknown line is

$$y = \alpha + \beta(x - \bar{x})$$

The least squares method requires the minimization of the function

$$H(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - \alpha - \beta(x_i - \bar{x})\big)^2$$

Setting partial derivatives to zero gives two equations:

$$\begin{aligned} 0 &= -\frac{1}{2}\frac{\partial H}{\partial \alpha} = \sum_{i=1}^{n} y_i - n\alpha \\ 0 &= -\frac{1}{2}\frac{\partial H}{\partial \beta} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) - \beta \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

Thus the MLE for α and β are

$$\hat{\alpha} = \bar{y} \qquad \text{and} \qquad \hat{\beta} = \frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

21.4 Basic statistics for two samples

We have two sets of values: $x_1, \dots, x_n$ and $y_1, \dots, y_n$. Accordingly, we have two sample means

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and two sample variances

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big)$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1}\Bigl(\sum_{i=1}^n y_i^2 - n\bar{y}^2\Bigr)$$

In addition, to measure the dependence (correlation) between these samples we use *sample covariance*

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\Bigl(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\Bigr)$$

and *sample correlation coefficient*

$$r = \frac{c_{xy}}{s_x s_y}$$

Just as in probability theory, the sample correlation coefficient takes values $-1 \le r \le 1$. The values close to 1 indicate strong positive correlation, the values close to -1 indicate strong negative correlation, and if $r = 0$ then the samples are uncorrelated.

In these terms, the least squares estimates of the linear regression parameters are

$$\hat{\alpha} = \bar{y} \qquad \text{and} \qquad \hat{\beta} = \frac{c_{xy}}{s_x^2} = r \cdot \frac{s_y}{s_x}$$

We note that positive slope $\beta > 0$ corresponds to positive correlation $r > 0$, negative slope $\beta < 0$ corresponds to negative correlation $r > 0$. The zero slope $\beta = 0$ corresponds to uncorrelated x and y variables.

21.5 Residuals

The value $\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$ is the estimated value of the unknown function $y = \alpha + \beta(x_i - \bar{x})$ at the point x_i. Recall that the difference $y_i - \hat{y}_i$ is called the residual and

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \bigl(y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\bigr)^2$$

is the residual sum of squares (RSS). There is a shortcut formula for the RSS:

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= (n-1)\Bigl(s_y^2 - \frac{c_{xy}^2}{s_x^2}\Bigr) = (n-1)\frac{s_x^2 s_y^2 - c_{xy}^2}{s_x^2} = (n-1)(1-r^2)s_y^2\end{aligned}$$

Note: if RSS=0, then the points (x_i, y_i) lie exactly on a straight line. This happens precisely when $r = \pm 1$, i.e. when the correlation between x and y takes one of its extreme values.

21.6 Example

Approximate five points

$$(-1, 1), \quad (0, 2), \quad (1, 2), \quad (2, 3), \quad (3, 4)$$

by a line $y = \alpha_1 + \beta x$. In other words, fit a line to these points.

Solution. First, we compute five 'accumulators'

$$\sum x_i = 5, \quad \sum y_i = 12, \quad \sum x_i^2 = 15, \quad \sum y_i^2 = 34, \quad \sum x_i y_i = 19$$

Then we compute basic statistics:

$$\bar{x} = 1, \quad \bar{y} = 2.4, \quad s_x^2 = \tfrac{15-5}{4} = 2.5, \quad s_y^2 = \tfrac{34-5\cdot 2.4^2}{4} = 1.35, \quad c_{xy} = \tfrac{19-5\cdot 2.4}{4} = 1.75$$

Now we get

$$\hat{\alpha} = 2.4 \qquad \text{and} \qquad \hat{\beta} = 0.7$$

The equation of the least squares line is

$$y = 2.4 + 0.7(x - 1) = 1.7 + 0.7x$$

The residual sum of squares is

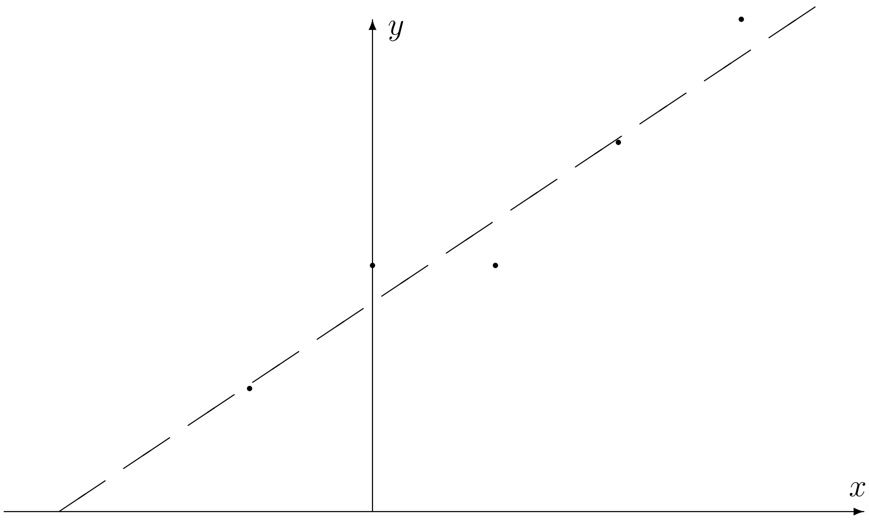
$$\text{RSS} = 4 \cdot \frac{\frac{10}{4} \cdot \frac{5.2}{4} - \frac{49}{16}}{\frac{10}{4}} = \frac{52 - 49}{10} = 0.3$$

Note that RSS is small, which indicates the line is pretty close to the given points.

Lastly, the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{0.3}{5} = 0.06$$

A plot showing the given points and the best line is called *scatter plot.*



21.7 Distributions of estimates

It is known in (advanced) probability theory that the estimates $\hat{\alpha}$ and $\hat{\beta}$ have normal distributions and $\hat{\sigma^2}$ is related to a χ^2 distribution:

$$\hat{\alpha} = \mathcal{N}\Big(\alpha, \frac{\sigma^2}{n}\Big), \qquad \hat{\beta} = \mathcal{N}\Big(\beta, \frac{\sigma^2}{(n-1)s_x^2}\Big), \qquad \hat{\sigma}^2 = \frac{\sigma^2}{n} \cdot \chi^2(n-2)$$

In addition, these three estimates are independent.

21.8 Remark

The independence of the above estimates may help to prevent some misinterpretation of the data. For example, if the observed points happen to lie exactly on a line (so that the RSS is zero or nearly zero), one may feel 'lucky' and assume that one has found the actual line, i.e. its parameters $\hat{\alpha}$ and $\hat{\beta}$ are close to the actual values of α and β. This need not be the case at all: there is no correlation between the RSS and the accuracy of the fit (the accuracy of the estimates of the parameters α and β).

Likewise, if the points are badly misaligned, so that the RSS is large, it would not mean at all that the estimates of α and β are poor: the least squares line may be actually very close to the theoretical line.

21.9 Unbiasedness and efficiency

We note that $\mathbb{E}(\hat{\alpha}) = \alpha$ and $\mathbb{E}(\hat{\beta}) = \beta$, so that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are unbiased. They are also 100% efficient (this fact requires computation of the Rao-Cramer lower bound, which is beyond the scope of this course).

21.10 Variances and their adjustment

The variances of the estimates are

$$\mathsf{Var}(\hat{\alpha}) = \frac{\sigma^2}{n} \qquad \mathsf{Var}(\hat{\beta}) = \frac{\sigma^2}{(n-1)s_x^2}$$

These formulas are useful, too. Suppose we want to increase the accuracy of our estimates, i.e. reduce their variances. Of course, increasing n would help, as always. But we can improve the estimate $\hat{\beta}$ even without increasing n, by increasing s_x^2. This can be achieved by positioning the points x_i 'wisely' – as far from their center $\bar{x}$ as possible. For example, if we are to select n points from an interval (A, B), then we should put $n/2$ points near A and the other $n/2$ points near the other end B.

21.11 Confidence intervals

Knowing the distributions of our estimates of the unknown parameters we can construct confidence intervals.

The confidence interval for α is

$$\hat{\alpha} \pm t_{\gamma/2}(n-2)\sqrt{\frac{\hat{\sigma^2}}{n-2}}$$

The confidence interval for β is

$$\hat{\beta} \pm t_{\gamma/2}(n-2)\sqrt{\frac{n\hat{\sigma^2}}{(n-2)(n-1)s_x^2}}$$

The confidence interval for σ^2 is

$$\left[\frac{n\hat{\sigma^2}}{\chi^2_{\gamma/2}(n-2)}, \frac{n\hat{\sigma^2}}{\chi^2_{1-\gamma/2}(n-2)}\right]$$

We denoted the confidence coefficient by $1 - \gamma$. Note that the number of degrees of freedom is $n - 2$, because we have estimated two parameters in the model (α and β).

21.12 Testing hypotheses

We can also test hypotheses about the unknown parameters. The most common hypothesis is $H_0\colon \beta = 0$. Its meaning is that the unknown line is flat (has zero slope), i.e. there is no correlation between the x and y variables.

The test about β with significance level γ can be summarized in the table:

H_0	H_1	Critical region	Test statistic
$\beta = \beta_0$	$\beta > \beta_0$	$T > t_\gamma(n-2)$	$T = \frac{\hat{\beta}-\beta_0}{\sqrt{\frac{n\hat{\sigma}^2}{(n-2)(n-1)s_x^2}}}$
	$\beta < \beta_0$	$T < -t_\gamma(n-2)$	
	$\beta \neq \beta_0$	$\lvert T\rvert > t_{\gamma/2}(n-2)$	

21.13 Example (continued)

The confidence interval for α is

$$2.4 \pm t_{\gamma/2}(3)\sqrt{\frac{0.06}{3}}$$

The confidence interval for β is

$$0.7 \pm t_{\gamma/2}(3)\sqrt{\frac{0.3}{30}}$$

The confidence interval for σ^2 is

$$\left[\frac{0.3}{\chi^2_{\gamma/2}(3)}, \frac{0.3}{\chi^2_{1-\gamma/2}(3)}\right]$$

If we want to test the hypothesis is $H_0\colon \beta = 0$ against the two-sided alternative $H_1\colon \beta \neq 0$, then we use the T statistic

$$T = \frac{0.7}{\sqrt{0.3/30}} = 7.0$$

The critical region is $T = 7 > t_{\gamma/2}(3)$. It is quite clear that we will accept H_1 for all reasonable values of γ.

21.14 Prediction

The main purpose of approximating the observed points (x_i, y_i) with a function $y = g(x)$, in particular with a line $y = \alpha + \beta(x - \bar{x})$, is to be able to predict the value y at some other points x. For example, if x is the time variable, we may want to predict the value of our function $y = g(x)$ in the future.

Of course, the point estimate of $y = g(x)$ would be $\hat{y} = \hat{\alpha} + \hat{\beta}(x - \bar{x})$. Next we want to construct a confidence interval for y. It is given by the formula

$$\hat{y} \pm c\, t_{\gamma/2}(n-2)$$

where $1 - \gamma$ denotes the confidence level, and

$$c = \sqrt{\frac{n\hat{\sigma^2}}{n-2}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

We note that the half length of this confidence interval is $c\, t_{\gamma/2}(n-2)$, and it is a function of x. It takes the smallest value at $x = \bar{x}$, hence the prediction is the most accurate at the center $\bar{x}$ of the x sample. Then the interval gets larger on both sides of $\bar{x}$, and it grows approximately linearly with $x - \bar{x}$.

The above interval was constructed for the actual value $y = g(x)$ of the unknown function, i.e. for the *model value* of y. Suppose now we want to estimate the *experimentally observed* value y_{exp} at the point x. Our assumptions say that $y_{\text{exp}} = y + \varepsilon$, where ε is a statistical error represented by a normal random variable $\mathcal{N}(0, \sigma^2)$. We see that y_{exp} contains an additional error (the measurement error), so it is 'more random'. The confidence interval for y_{exp} should be larger than the one for y, and it is given by the formula

$$\hat{y} \pm c_{\text{exp}}\, t_{\gamma/2}(n-2)$$

where $1 - \gamma$ denotes the confidence level, and

$$c_{\text{exp}} = \sqrt{\frac{n\hat{\sigma^2}}{n-2}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}$$

Note that c_{exp} is slightly greater than c, hence the second confidence interval is slightly larger than the first.

The rest of this Chapter is optional...

21.15 Polynomial fit

In many practical problems the regression $y = g(x)$ is nonlinear, then one needs to fit some nonlinear functions to the observed points (x_i, y_i). Here we discuss fitting polynomials

$$g(x) = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

here $k \geq 1$ is the degree of the polynomial and $\beta_0, \dots, \beta_k$ are unknown coefficients to be estimated. The degree k is supposed to be chosen. The least squares method is based on the minimization of

$$H(\beta_0, \dots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_k x_i^k)^2$$

Setting partial derivatives to zero gives us a system of equations

$$\begin{aligned} \beta_0 \cdot n + \beta_1 \sum x_i + \cdots + \beta_k \sum x_i^k &= \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \cdots + \beta_k \sum x_i^{k+1} &= \sum x_i y_i \\ \cdots \qquad\qquad & \cdots \\ \beta_0 \sum x_i^k n + \beta_1 \sum x_i^{k+1} + \cdots + \beta_k \sum x_i^{2k} &= \sum x_i^k y_i \end{aligned}$$

This is a system of $k + 1$ linear equations with $k + 1$ unknowns. Solving such systems in practice is difficult for large k, but there are many computer programs that do that quickly and accurately.

In MATLAB, one can fit a polynomial of degree k by using the procedure **polyfit**:

$$\mathbf{p} = \mathbf{polyfit}(\mathbf{x}, \mathbf{y}, \mathbf{k})$$

where **k** is the degree of the polynomial, **x** is the vector of x-values and **y** is the vector of y-values. The procedure returns **p**, the vector of estimated coefficients.

The rest of this chapter is optional...

21.16 Choice of the degree k

If the data points cannot be well approximated by a straight line, one may try parabolas ($k = 2$), cubic polynomials ($k = 3$), etc., until the fit is satisfactory. But how should we decide if the fit is satisfactory or not?

The residual sum of squares (RSS) measures the overall discrepancy between the best polynomial fit and the data points. It steadily gets smaller as k increases. To measure 'how well' the polynomial describes the points one uses the quantity

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_k x_i^k$$

is the estimated (predicted) value of the function. The quantity R^2 show how much of the original variation in the y sample is explained by the fit. The value $1 - R^2$ tells us how much of the variation remains unaccounted for by the fit.

The value of R^2 steadily growth from 0 to 1 as the degree k increases. One usually consider $R^2 = 0.8$ or $R^2 = 0.9$ to be a good fit.

But one should not attempt to go 'too far' and choose the degree k too high. Theoretically, it is possible to reduce the RSS to zero (and then R^2 will reach its maximal value $R^2 = 1$), but we definitely do not want that: the corresponding polynomial curve will be ridiculous – it will wiggle up and down trying to adapt itself to every single data point. This phenomenon is known as *overfit*.

So which degree k is optimal? When do we stop increasing the degree of the polynomial? Some researchers examine the residuals $d_i = y_i - \hat{y}_i$, plot them as a function of x_i. If the residuals have some pattern or follow some trend (for example, go up or down for many consecutive x_i's), then one may try a higher degree to account for that trend. If the residuals look 'chaotic' (without any clear pattern), then the fit is assumed to be good enough. This method is quite subjective, though.

21.17 Cross-validation

A more objective method to find an optimal degree k is the *cross-validation*. One divides the data points into two groups: a *training set* (a larger group) to which one fits a polynomial, and a *test set* (a smaller group) of points on which the polynomial is 'tested'. Precisely, if (x_j, y_j) is the 'test set' ($1 \leq j \leq m$ with some $m < n$), then one computes the predicted values $\hat{y}_j$ and the overall residual sum of squares for the test set

$$\mathrm{RSS}_{\mathrm{test}} = \sum_{j=1}^{m} (y_i - \hat{y}_i)^2$$

which is treated as the discrepancy of the fit. Note that the test points (x_j, y_j) were *not* used in the construction of the polynomial (they were not part of the training set), hence the polynomial need not adapt to them.

The value of $\mathrm{RSS}_{\mathrm{test}}$ usually decreases as k growth, indicating that the fit becomes better, but when the degree k gets too high, the $\mathrm{RSS}_{\mathrm{test}}$ starts growing again. The degree k for which the $\mathrm{RSS}_{\mathrm{test}}$ takes its minimal value is optimal.

21.18 Leave-one-out

The above method requires an arbitrary partitioning of the data set into two groups. For different partitions, one may find different values of the optimal degree k, thus making the results ambiguous and confusing.

To eliminate the dependence on the partition one should combine many different partitions. In a popular algorithm, one partitions the data set of n points into two groups: a training set of $n-1$ points and a single-point 'test' set. The polynomial is constructed by using the $n-1$ training points, and then tested on the remaining point (x_j, y_j) giving a single residual squared $(y_j - \hat{y}_j)^2$.

Then one repeats this procedure for every point of the sample: take a point (x_j, y_j), leave it out, fit a polynomial to the remaining $n-1$ points, evaluate its value $\hat{y}_j$ for $x = x_j$, then compute $(y_j - \hat{y}_j)^2$. The overall sum of squares

$$\mathrm{RSS}_{\mathrm{all}} = \sum_{j=1}^{n} (y_i - \hat{y}_i)^2$$

is then treated as the discrepancy of the fit by polynomials of degree k. The degree k for which the $\mathrm{RSS}_{\mathrm{all}}$ takes its minimal value is optimal.

22 Nonparametric methods

22.1 General description

So far we discussed statistical problems where the unknown random variables were only partially unknown, that is their type (normal, binomial, or other) was known, and the parameter(s) were to be determined. Now we turn to problems where the random variable X is *completely unknown*, i.e. we know nothing about its type or distribution.

How can we characterize a totally unknown random variable, if there are no parameters to test or estimate? Most random variables can be characterized by their mean values and variances, so that $\mathbb{E}(X)$ and $\mathsf{Var}(X)$ can be regarded as important parameters to determine.

But we must remember that not all random variables have mean value or variance (example: the Cauchy random variable has neither). Trying to determine a nonexistent quantity may not be rewarding.

On the other hand, every random variable has *median, quartiles*, and more generally *percentiles.* These are characteristics that can be determined statistically. Note: if we accurately determine sufficiently many percentiles, then we effectively can reconstruct the distribution function of X.

22.2 Order statistics

Recall that a percentile π_p is a number that divides the probability distribution according to the ratio $p : (1-p)$, i.e. satisfying $F(\pi_p) = p$, where F is the distribution function. If we have a sample $x_1, \dots, x_n$ and order it as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, then dividing this ordered sample according to the ratio $p : (1-p)$ seems to be a good way to estimate the percentile π_p.

For brevity, we denote $y_i = x_{(i)}$, i.e. $y_1 \le y_2 \le \cdots \le y_n$ will be the ordered sample. The capital letters Y_r will denote the random variables associated with y_r.

22.3 Estimates for percentiles

To estimate π_p, we compute $r = p(n+1)$. If r is an integer, then y_r is the estimate of π_p. Otherwise we take the two integers r and $r+1$ closest to $p(n+1)$, i.e. $r < p(n+1) < r+1$ and estimate π_p by $(y_r + y_{r+1})/2$:

$$\hat{\pi}_p = \begin{cases} y_r & \text{if } r = p(n+1) \text{ is an integer} \\ \frac{1}{2}[y_r + y_{(r+1)}] & \text{if } r < p(n+1) < r+1 \end{cases}$$

In particular, to estimate the median $m = \pi_{1/2}$, we use the rule

$$\hat{m} = \begin{cases} y_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}[y_{n/2} + y_{(n+2)/2}] & \text{if } n \text{ is even} \end{cases}$$

The order statistics $y_1, \dots, y_n$ will play an instrumental role in the nonparametric statistics.

22.4 Distribution of order statistics

In probability, we learned that the random variable Y_r had distribution function

$$\begin{aligned} G_r(y) = \mathbb{P}(Y_r \leq y) &= \mathbb{P}\big(b(n, F(y)) \geq r\big) \\ &= \sum_{k=r}^{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k} \end{aligned}$$

or, alternatively,

$$\begin{aligned} G_r(y) &= \mathbb{P}\big(b(n, 1 - F(y)) \leq n - r\big) \\ &= \sum_{k=0}^{n-r} \binom{n}{k} [1 - F(y)]^k [F(y)]^{n-k} \end{aligned}$$

Here $b(n, p)$ denotes a binomial random variable with probability of success p. Note that "at least r successes" is the same as "at most $n - k$ failures".

22.5 Practical calculation: Table II

Probabilities $\mathbb{P}(b(n, p) \leq r)$ related to a binomial random variable $X = b(n, p)$ can be found in Table II. It covers values $n = 2, \dots, 25$ (for larger n's, we use normal approximation) and $p = 0.05, 0.1, 0.15, 0.2, \dots, 0.5$.

What do we do if $p > 0.5$? In that case we switch "successes" and "failures", which replaces p with $1 - p$:

$$\mathbb{P}(b(n, p) \leq r) = \mathbb{P}(b(n, 1 - p) \geq n - r) = 1 - \mathbb{P}(b(n, 1 - p) \leq n - r - 1).$$

22.6 Examples

(a) Let $n = 9$ and $F(0.1) = 0.1$. Determine $G_1(0.1)$.

Solution:

$$\begin{aligned} G_1(0.1) &= \mathbb{P}(Y_1 \leq 0.1) = \mathbb{P}(b(9, 0.1) \geq 1) \\ &= 1 - \mathbb{P}(b(9, 0.1) \leq 0) = 1 - 0.3874 = 0.6126 \end{aligned}$$

The value 0.3874 was taken from Table II.

(b) Let $n = 9$ and $F(0.7) = 0.7$. Determine $G_8(0.7)$.

Solution:

$$\begin{aligned} G_8(0.7) &= \mathbb{P}(Y_8 \le 0.7) = \mathbb{P}(b(9, 0.7) \ge 8) = 1 - \mathbb{P}(b(9, 0.7) \le 7) \\ &= 1 - \mathbb{P}(b(9, 0.3) \ge 2) = \mathbb{P}(b(9, 0.3) \le 1) = 0.1960 \end{aligned}$$

The value 0.1960 was taken from Table II. Note that in the second line we used the trick of switching successes and failures.

22.7 Three important formulas

We continue the analysis of Section 22.4. Substitute $y = \pi_p$, then $F(y) = F(\pi_p) = p$, hence

$$\mathbb{P}(Y_r \le \pi_p) = \mathbb{P}\big(b(n, p) \ge r\big) = \sum_{k=r}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

Similarly,

$$\mathbb{P}(Y_r \ge \pi_p) = \mathbb{P}\big(b(n, p) < r\big) = \sum_{k=0}^{r-1} \binom{n}{k} p^k (1-p)^{n-k}$$

and

$$\mathbb{P}(Y_{r_1} \le \pi_p \le Y_{r_2}) = \mathbb{P}\big(r_1 \le b(n, p) < r_2\big) = \sum_{k=r_1}^{r_2-1} \binom{n}{k} p^k (1-p)^{n-k}$$

(In the inequalities on the left hand side, we can replace $\le$ with $<$ and $\ge$ with $>$, because Y_r is a continuous random variable.)

22.8 Example

Let $y_1 \le \cdots \le y_{13}$ be an ordered sample of size $n = 13$ from an unknown random variable. For its median m we have

$$\mathbb{P}(y_4 \le m \le y_{10}) = \mathbb{P}(4 \le b(13, 0.5) \le 9) = 0.9539 - 0.0461 = 0.9078$$

the numerical values are taken from Table II.

Thus, the interval (y_4, y_{10}) can be regarded as a confidence interval for the median m with confidence level 90%.

22.9 Table 8.4-1

Table 8.4-1 on page 438 in the book gives confidence intervals for the median m for small samples of size $5 \le n \le 20$. It also gives the corresponding confidence levels. These are *recommended* confidence intervals, one can change them in practice. But remember: making the CI shorter reduces the confidence level, making the CI wider increases confidence level.

For example, for $n = 13$, the recommended interval is (Y_3, Y_{11}) with level 97.76%. We can use a shorter interval (Y_4, Y_{10}), but it has a lower confidence level of 90.78% (see the previous section).

22.10 Example

Suppose a sample of size $n = 10$ is

$$3.8,\ 4.1,\ 2.5,\ 4.2,\ 3.4,\ 2.8,\ 4.6,\ 3.3,\ 2.8,\ 3.7,$$

Find a confidence interval for the percentile $\pi_{0.4}$.

Solution: let us try $(y_1, y_5) = (2.5, 3.4)$:

$$\mathbb{P}(y_1 \le \pi_{0.4} \le y_5) = \mathbb{P}(2.5 \le \pi_{0.4} \le 3.4) = 0.6331 - 0.0060 = 0.6271$$

This is a very short interval, but the confidence level is rather low, 62.71%.

22.11 Normal approximation for large n

When n is large ($n \ge 20$), we can use normal approximation $b(n,p) \approx \mathcal{N}\big(np, np(1-p)\big)$, then

$$\begin{aligned} G_r(y) &\approx \mathbb{P}\Big(\mathcal{N}\big(nF(y), nF(y)(1-F(y))\big) \ge r - \tfrac{1}{2}\Big) \\ &= 1 - \Phi\bigg(\frac{r - \frac{1}{2} - nF(y)}{\sqrt{nF(y)(1-F(y))}}\bigg) \end{aligned}$$

(we applied histogram correction).

22.12 Example

Find a confidence interval for the median m if the sample size is $n = 100$.

Solution. Let us try (y_{40}, y_{60}):

$$\begin{aligned} \mathbb{P}(y_{40} \le m \le y_{60}) &= \mathbb{P}(40 \le b(100, 0.5) < 60) \\ &\approx \mathbb{P}\big(39.5 \le \mathcal{N}(50, 25) \le 59.5\big) \\ &= \Phi\Big(\frac{59.5 - 50}{5}\Big) - \Phi\Big(\frac{39.5 - 50}{5}\Big) \\ &= \Phi(1.9) - \Phi(-2.1) = 0.9534 \end{aligned}$$

We got a CI with level 95.34%.

22.13 CI for percentiles for large n

Suppose we want to construct a CI for the unknown percentile π_p with confidence level $1-\alpha$. Then we need to find the orders r_1 and r_2 such that

$$\begin{aligned}
1-\alpha &= \mathbb{P}\big(Y_{r_1} \le \pi_p \le Y_{r_2}\big) \\
&= \mathbb{P}\big(r_1 \le b(n,p) \le r_2 - 1\big) \\
&\approx \mathbb{P}\Big(r_1 - \tfrac{1}{2} \le \mathcal{N}\big(np, np(1-p)\big) \le r_2 - \tfrac{1}{2}\Big) \\
&= \Phi\bigg(\frac{r_2 - \frac{1}{2} - np}{\sqrt{np(1-p)}}\bigg) - \Phi\bigg(\frac{r_1 - \frac{1}{2} - np}{\sqrt{np(1-p)}}\bigg)
\end{aligned}$$

Assigning the probabilities $\alpha/2$ to each tail gives us the following formulas:

$$\frac{r_2 - \frac{1}{2} - np}{\sqrt{np(1-p)}} = z_{\alpha/2}$$

$$\frac{r_1 - \frac{1}{2} - np}{\sqrt{np(1-p)}} = -z_{\alpha/2}$$

Thus

$$\begin{aligned}
r_2 &= \tfrac{1}{2} + np + z_{\alpha/2}\sqrt{np(1-p)} \\
r_1 &= \tfrac{1}{2} + np - z_{\alpha/2}\sqrt{np(1-p)}
\end{aligned}$$

Of course we need to round off these values to the nearest integer (to be safe, it is better to round r_1 to the nearest smaller integer and r_2 to the nearest greater integer).

22.14 Example

Find an 80% confidence interval for the first quartile $\pi_{1/4}$ for a sample of size $n = 27$.

Solution: here $\alpha/2 = 0.1$, so we use $z_{0.1} = 1.282$.

$$r_2 = 0.5 + 6.75 + 1.282\sqrt{27 \cdot 0.25 \cdot 0.75} = 10.2$$

$$r_1 = 0.5 + 6.75 - 1.282\sqrt{27 \cdot 0.25 \cdot 0.75} = 4.3$$

So the confidence interval is (y_4, y_{10}). This is not entirely safe: we reduced r_1 (which is safe) but also reduced r_2 (which may be dangerous). To verify our result, we can find the actual confidence level of this interval:

$$\mathbb{P}(y_4 \leq \pi_{1/4} \leq y_{10}) = \mathbb{P}\big(4 \leq b(27, 0.25) \leq 9\big) = 0.8201$$

(the numerical value is obtained by using the on-line calculator on the instructor's web page). Thus the actual confidence level is even higher than the required 80%, so we are OK.

23 Wilcoxon tests

Here we test hypotheses about the median m of an unknown random variable.

23.1 Example

In a polluted lake, the median of length of fish is known to be $m_0 = 3.7$. Authorities clean up the lake expecting the fish to get better and grow. To check up the results, they have $n = 10$ fish caught randomly and measure their lengths:

$$5.0,\ 3.9,\ 5.2,\ 5.5,\ 2.8,\ 6.1,\ 6.4,\ 2.6,\ 1.7,\ 4.3$$

Is there sufficient evidence that the length of fish has increased?

Let m denote the unknown median of the fish length after the clean-up. We are testing the hypothesis $H_0\colon m = m_0$ against the alternative $H_1\colon m > m_0$.

23.2 'Old' sign test

One computes the differences between the sample lengths and the median $m_0 = 3.7$:

$$1.3,\ 0.2,\ 1.5,\ 1.8,\ -0.9,\ 2.4,\ 2.7,\ -1.1,\ -2.0,\ 0.6$$

Here 7 values are positive and 3 values are negative, so it appears that most fish grew. But seven and three are too small numbers, so the test will not be able to substantiate the desired conclusion. A smarter test by Wilcoxon (see below) uses not only the signs but also magnitudes of the observations, hence its conclusions are sharper.

23.3 Wilcoxon test - I

Given a sample $x_1, \dots, x_n$ of values of an unknown random variable we compute the differences $d_i = x_i - m_0$. Then we arrange their absolute values $|d_1|, \dots, |d_n|$ in the increasing order and assign ranks (from 1 to the smallest to n to the biggest). Then we add the signs of d_i's to the ranks; i.e. if $d_i < 0$ then we negate its rank. Lastly we sum up the signed ranks and obtain the Wilcoxon statistic W.

23.4 Example continued

In our example the differences $x_i - m_0$ are

$$1.3,\ 0.2,\ 1.5,\ 1.8,\ -0.9,\ 2.4,\ 2.7,\ -1.1,\ -2.0,\ 0.6$$

Their magnitudes arrange in the increasing order are

$$0.2 < 0.6 < 0.9 < 1.1 < 1.3 < 1.5 < 1.8 < 2.0 < 2.4 < 2.7$$

So the ranks (in the original order) are

$$5,\ 1,\ 6,\ 7,\ 3,\ 9,\ 10,\ 4,\ 8,\ 2$$

The signed ranks (corresponding to the signed differences) are

$$5,\ 1,\ 6,\ 7,\ -3,\ 9,\ 10,\ -4,\ -8,\ 2$$

Their sum is

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25$$

23.5 Distribution of Wilcoxon statistic

The statistic W has approximately normal distribution $W \approx \mathcal{N}(\mu, \sigma^2)$ with

$$\mu = 0 \qquad \text{and} \qquad \sigma^2 = \frac{n(n+1)(2n+1)}{6}$$

So we can compute the Z statistic

$$Z = \frac{W - \mu}{\sigma} = \frac{W}{\sqrt{n(n+1)(2n+1)/6}}$$

Then the test is completed as follows:

H_0	H_1	Critical region	p-value
$m = m_0$	$m > m_0$	$Z > z_\alpha$	$1 - \Phi(Z)$
	$m < m_0$	$Z < -z_\alpha$	$\Phi(Z)$
	$m \neq m_0$	$\lvert Z\rvert > z_{\alpha/2}$	$2\,[1 - \Phi(\lvert Z\rvert)]$

23.6 Example continued

In our example

$$Z = \frac{25}{\sqrt{10 \cdot 11 \cdot 21/6}} = \frac{25}{\sqrt{385}} = 1.274$$

If the significance level $\alpha = 0.1$, then the critical region is $Z > z_{0.1} = 1.282$. We accept H_0. The p-value of the test is 0.1013.

23.7 Remark

It is easy to see why

$$\mathbb{E}(W) = 0 \qquad \text{and} \qquad \mathsf{Var}(W) = \frac{n(n+1)(2n+1)}{6}$$

Under the null hypothesis, $\mathbb{P}(x_i < m_0) = \mathbb{P}(x_i > m_0) = 0.5$. So, every rank has equal chance to be positive or negative, hence its average value is zero.

Next, the variance of rank k is

$$\mathsf{Var}(k) = \mathbb{E}(k^2) = k^2$$

and

$$\mathsf{Var}(W) = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

23.8 Tie breakers

If two differences have equal magnitudes, $|x_i - m_0| = |x_j - m_0|$, then we average the corresponding ranks. For example, if the ordered sequence is $0.2 < 0.6 \leq 0.6 < 1.1 \cdots$, then we assign ranks 1, $2\frac{1}{2}, 2\frac{1}{2}, 4$, etc.

23.9 Median test for two samples

Suppose we have two samples: $x_1, \dots, x_{n_1}$ are values of a random variable X, and $y_1, \dots, y_{n_2}$ are values of another random variable Y. We want to compare their medians m_x and m_y, i.e. test $H_0\colon m_x = m_y$ against the alternative $H_1\colon m_x > m_y$ or $m_x < m_y$.

The traditional 'median test' goes as follows: we combine the two samples, arrange all the $n_1 + n_2$ values in the increasing order, and count the number of x's in the lower half of the combined sample. Let V be that number. If H_0 is true, we expect $V \approx n_1/2$, if $m_x > m_y$ we expect $V < n_1/2$, and if $m_x < m_y$ we expect $V > n_1/2$.

The statistic V has the following distribution (assuming the H_0 hypothesis is true and $n_1 + n_2 = 2k$ is even):

$$\mathbb{P}(V = v) = \frac{\binom{n_1}{v}\binom{n_2}{k-v}}{\binom{n_1+n_2}{k}}$$

Then we can compute the p-value of the test as follows. If we are testing the hypothesis $H_1\colon m_x > m_y$ and v_{exp} denotes the experimental (computed) value of the V statistic, then

$$\text{p-value} = \mathbb{P}(V \le v_{\text{exp}}) = \sum_{v \le v_{\text{exp}}} \frac{\binom{n_1}{v}\binom{n_2}{k-v}}{\binom{n_1+n_2}{k}}$$

If we are testing the hypothesis $H_1\colon m_x < m_y$, then

$$\text{p-value} = \mathbb{P}(V \ge v_{\text{exp}}) = \sum_{v \ge v_{\text{exp}}} \frac{\binom{n_1}{v}\binom{n_2}{k-v}}{\binom{n_1+n_2}{k}}$$

23.10 Example

Let the x sample be 6, 3, 2, 4, 9, and the y sample be 7, 7, 5, 10, 15. Test the hypothesis $H_0\colon m_x = m_y$ against the alternative $H_1\colon m_x < m_y$.

Solution. Here $n_1 = n_2 = 5$. The combined sample is

$$2,\ 3,\ 4,\ 5,\ 6 \;\Big|\; 7,\ 7,\ 9,\ 10,\ 15$$

and there are four x's in the lower half (2, 3, 4, 6), so $v_{\text{exp}} = 4$. Then the p-value is

$$\begin{aligned}\text{p-value} &= \mathbb{P}(V \ge 4) = \mathbb{P}(V = 4) + \mathbb{P}(V = 5)\\ &= \frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} + \frac{\binom{5}{5}\binom{5}{0}}{\binom{10}{5}} = \frac{\binom{5}{4}\binom{5}{1} + \binom{5}{5}\binom{5}{0}}{\binom{10}{5}} = \frac{5 \cdot 5 + 1 \cdot 1}{252} = \frac{26}{252}\end{aligned}$$

So the p-value is about 10%. This result is not very compelling, the test does not clearly demonstrate the validity of either hypothesis H_0 or H_1.

23.11 Remark

The above median test is weak, because it only relies on the number of x's in the lower half of the combined sample and does not use the magnitude of x's. The following smarter test by Wilcoxon improves the median test.

23.12 Wilcoxon test - II

We combine the two samples, arrange all the $n_1 + n_2$ values in the increasing order, and assign ranks (from 1 to the smallest to n to the biggest). The we compute the Wilcoxon statistics

$$W = \text{sum of ranks of } y'\text{s}$$

In our example, y's have ranks 4, $6\frac{1}{2}$, $6\frac{1}{2}$, 9 and 10 (note that we used the tie breaker rule to average the ranks of equal values), so

$$W = 4 + 6\tfrac{1}{2} + 6\tfrac{1}{2} + 9 + 10 = 36$$

23.13 Distribution of the second Wilcoxon statistic

The statistic W has approximately normal distribution $W \approx \mathcal{N}(\mu, \sigma^2)$ with

$$\mu = \frac{n_2(n_1 + n_2 + 1)}{2} \qquad \text{and} \qquad \sigma^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

So we can compute the Z statistic

$$Z = \frac{W - \mu}{\sigma} = \frac{W - n_2(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}}$$

Then the test is completed as follows:

H_0	H_1	Critical region	p-value
	$m_x > m_y$	$Z < -z_\alpha$	$\Phi(Z)$
$m_x = m_y$	$m_x < m_y$	$Z > z_\alpha$	$1 - \Phi(Z)$
	$m_x \neq m_y$	$\lvert Z\rvert > z_{\alpha/2}$	$2\,[1 - \Phi(\lvert Z\rvert)]$

23.14 Example finished

In our example

$$Z = \frac{36 - 5 \cdot 11/2}{\sqrt{5 \cdot 5 \cdot 11/12}} = 1.78$$

Hence

$$\text{p-value} = 1 - \Phi(1.78) = 0.0375$$

The p-value of 3.75% is a much more definite indication in favor of the alternative hypothesis than 10% in 23.10. The Wilcoxon test is thus sharper than the median test.

24 Run tests

In the previous chapter we tested the hypothesis that two random variables X and Y had the same median $m_x = m_y$. If they do, then still X and Y may be different random variables (having different distribution functions). For example, the random variables $X = \mathcal{N}(0,1)$ and $Y = \mathcal{N}(0,10000)$ have zero median but very different distributions (typical values of X are within the interval $[-3,3]$, and typical values of Y are or order ± 100).

Here we test the hypothesis that X and Y have identical distributions, i.e. the same distribution function $H_0\colon F_X = F_Y$. The alternative hypothesis will be $H_1\colon F_X \neq F_Y$ (the denial of the null hypothesis).

24.1 Run test

Let $x_1, \dots, x_{n_1}$ observed values of X and $y_1, \dots, y_{n_2}$ observed values of Y. We combine the two samples, arrange all the $n_1 + n_2$ values in the increasing order, and underline consecutive x's and consecutive y's in the combined sample to get something like this:

$$\underline{x}\,\underline{yyy}\,\underline{xx}\,\underline{yyy}\,\underline{x}\,\underline{y}\,\underline{xx}\,\underline{y}\cdots$$

Every string of consecutive x's or y's (including singletons) is called a *run*. We count runs in the entire combined sample. Let R denote the number of runs.

If the null hypothesis is true, i.e. X and Y have identical distributions, then x's and y's should be mixed up evenly in the combined sample, thus the runs should be short and the number of runs large.

On the other hand, if $F_X \neq F_Y$, then there must be intervals where x's appear more frequently than y's, and vice versa. Then we expect longer runs and their number smaller.

Thus the critical region of the test is $R < C$, where C is a critical value. Let r_{exp} denote the experimental (computed) value of the R statistic. Then the p-value of the test can be computed by

$$\text{p-value} = \mathbb{P}(R \le r_{\text{exp}}) = \sum_{r \le r_{\text{exp}}} \mathbb{P}(R = r).$$

24.2 Distribution of R for small samples

When n is small, we use exact formulas for the probabilities $\mathbb{P}(R = r)$. Here they are.

If $r = 2k$ is an even number, then

$$\mathbb{P}(R = 2k) = \frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_2}}$$

If $r = 2k+1$ is an odd number, then

$$\mathbb{P}(R = 2k+1) = \frac{\binom{n_1-1}{k}\binom{n_2-1}{k-1} + \binom{n_1-1}{k-1}\binom{n_2-1}{k}}{\binom{n_1+n_2}{n_2}}$$

24.3 Example

Consider the data from Example 23.10. Here is the combined sample with underlined runs:

$$\underline{2\,3\,4}\;\underline{5}\;\underline{6}\;\underline{7\,7}\;\underline{9}\;\underline{10\,15}$$

The total number of runs is $R = 6$.

The p-value of the test is $\sum_{r\le 6}\mathbb{P}(R = r)$. Here we go:

$$\mathbb{P}(R = 2) = \frac{2\binom{4}{0}\binom{4}{0}}{\binom{10}{5}} = \frac{2}{252}$$

$$\mathbb{P}(R = 3) = \frac{\binom{4}{1}\binom{4}{0} + \binom{4}{0}\binom{4}{1}}{\binom{10}{5}} = \frac{8}{252}$$

$$\mathbb{P}(R = 4) = \frac{2\binom{4}{1}\binom{4}{1}}{\binom{10}{5}} = \frac{32}{252}$$

$$\mathbb{P}(R = 5) = \frac{\binom{4}{2}\binom{4}{1} + \binom{4}{1}\binom{4}{2}}{\binom{10}{5}} = \frac{48}{252}$$

$$\mathbb{P}(R = 6) = \frac{2\binom{4}{2}\binom{4}{2}}{\binom{10}{5}} = \frac{72}{252}$$

The total is

$$\text{p-value} = \sum_{r=2}^{6}\mathbb{P}(R = r) = \frac{162}{252} \approx 0.64$$

The p-value of 64% is an overwhelming evidence in favor of H_0. Thus, we conclude that the two random variables have identical distributions.

24.4 Controversy?

But stop! In the previous Chapter, we analyzed the same example by the second Wilcoxon test and concluded that $m_x < m_y$, i.e. X and Y had different medians! How can they have the same distribution?

Well, it is not uncommon in statistical practice that different methods applied to the same data sets lead to different (often logically inconsistent and even opposite) conclusions. Every statistical conclusion may or may not be correct, and there is always a chance that it is wrong.

In hypotheses testing, when we accept H_0 (as in 24.3), we simply conclude that there is not enough evidence to accept H_1. The run test was not able to recognize the difference between the two samples, thus it had to 'give up' and stick to the null hypothesis. The second Wilcoxon test was 'smarter' and caught the difference between the two samples, thus arriving at the alternative hypothesis.

24.5 Distribution of R for large samples

When the samples are large ($n_1 \geq 10$ and $n_2 \geq 10$), then R is approximately a normal random variable $R \approx \mathcal{N}(\mu, \sigma^2)$, where

$$\mu = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

and

$$\sigma^2 = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

So we can compute the Z statistic

$$Z = \frac{R - \mu}{\sigma}$$

Now the critical region will be $Z < -z_\alpha$ and the p-value$= \Phi(Z)$.

24.6 Example

In our example $n_1 = n_2 = 5$, so

$$\mu = \frac{2 \cdot 5 \cdot 5}{10} + 1 = 6$$

$$\sigma^2 = \frac{50(50 - 10)}{10^2 \cdot 9} = \frac{20}{9}$$

so $Z = 0$ and the p-value is $\Phi(0) = 0.5$. This is different from the exact p-value=0.62 obtained earlier, but in the same 'ballpark'.

The rest of this chapter is optional...

24.7 Tests for randomness

The run test can be modified to test something quite unusual – the “randomness” of experimental data. We have always assumed that the observations $x_1, \dots, x_n$ were statistically independent, i.e. arrived randomly. However, in practice the values $x_1, \dots, x_n$ may follow a certain pattern.

For example, consider the stock market index recorded at the end of every trading day during a month. It often happens that the index tends to decrease toward the end of the month. Or it often follows periodic cycles – it drops in the beginning every week and recovers by the end of the week. Or it may change alternatively - drops followed by rises, followed by drops, etc.

In such cases the sequence $x_1, \dots, x_n$ is not completely random but follows certain patterns (trends up or down, or cycles). To detect these patterns we can use an adapted run test.

24.8 Run test for randomness

We divide the sample $x_1, \dots, x_n$ into upper and lower halves. Then we replace every x_i with a U (if it belongs to the upper half) or an L (if it belongs to the lower half). Then the sequence $x_1, \dots, x_n$ becomes a sequence of U's and L's. We underline strings of consecutive U's and consecutive L's (‘runs’) and get something like this:

$$\underline{LL}\,\underline{UUU}\,\underline{L}\,\underline{U}\,\underline{LLLL}\,\underline{U}\,\underline{LL}\,\underline{UUU}\cdots$$

Let R be the number of runs.

24.9 Tie breaking

If $n = 2k$ is even, then there is equal number $n_1 = n_2 = k$ of U's and L's. If $n = 2k + 1$ is odd, then we make the number of U's larger ($n_1 = k + 1$) and the number of L's smaller ($n_2 = k$).

24.10 Distribution of R

If the sample is purely random (without patterns), then the R statistic has distribution, which is described in Section 24.2 for small n and in Section 24.5 for large n. So we can use all those formulas.

24.11 Test procedure for trends

If we are trying to detect a trend (up or down), then we expect the sample to start with L's and end with U's or vice versa, thus runs are long and their number is small. Then the critical region is $R < C$, where C is some critical value, and the p-value can be computed by

$$\text{p-value} = \mathbb{P}(R \le r_{\text{exp}}) = \sum_{r \le r_{\text{exp}}} \mathbb{P}(R = r)$$

where r_{exp} is the experimental value of the R statistic.

24.12 Test procedure for cycles

If we are trying to detect cycles, then we expect that L's and U's alternate, thus runs are short (mostly singletons) and their number is big. Then the critical region is $R > C$ and the p-value can be computed by

$$\text{p-value} = \mathbb{P}(R \ge r_{\text{exp}}) = \sum_{r \ge r_{\text{exp}}} \mathbb{P}(R = r)$$

where r_{exp} is the experimental value of the R statistic.

24.13 Example

Consider a sample $5, 2, 3, 6, 8, 4, 10, 7$. It looks like the numbers tend to increase. Can we conclude that these numbers are not completely random?

Solution: We are testing the sample for a trend (up or down). Here $n = 8$ is even, so $n_1 = n_2 = 4$. The upper half is $6, 8, 10, 7$ and the lower half is $5, 2, 3, 4$. The sequence is

$$\underline{LLL}\,\underline{UU}\,\underline{L}\,\underline{UU}$$

so there are $R = 4$ runs. The p-value can be computed as

$$\begin{aligned}\text{p-value} &= \mathbb{P}(R = 2) + \mathbb{P}(R = 3) + \mathbb{P}(R = 4)\\ &= \frac{2\binom{3}{0}\binom{3}{0}}{\binom{8}{4}} + \frac{\binom{3}{1}\binom{3}{0} + \binom{3}{0}\binom{3}{1}}{\binom{8}{4}} + \frac{2\binom{3}{1}\binom{3}{1}}{\binom{8}{4}}\\ &= \frac{2}{70} + \frac{6}{70} + \frac{18}{70} = \frac{26}{70} \approx 0.37\end{aligned}$$

The p-value of 37% is a strong evidence in favor of H_0. Hence the test failed to detect a trend (no surprise, we only have 8 observations!..).

24.14 Another example

Suppose in the previous example we are testing the hypothesis for cycles. Then the p-value would be

$$\begin{aligned}
\text{p-value} &= \mathbb{P}(R \geq 4) \\
&= \mathbb{P}(R = 4) + \mathbb{P}(R = 5) + \mathbb{P}(R = 6) + \mathbb{P}(R = 7) + \mathbb{P}(R = 8) \\
&= 1 - \mathbb{P}(R = 2) - \mathbb{P}(R = 3) \\
&= 1 - \frac{2\binom{3}{0}\binom{3}{0}}{\binom{8}{4}} - \frac{\binom{3}{1}\binom{3}{0} + \binom{3}{0}\binom{3}{1}}{\binom{8}{4}} \\
&= 1 - \frac{2}{70} - \frac{6}{70} = \frac{62}{70} \approx 0.89
\end{aligned}$$

The p-value of 89% is an 'absolute' evidence in favor of H_0, i.e. there is not a trace of evidence of cycles (which is obvious even when you just look at the sample).

24.15 Remark

For large samples ($n \geq 20$) we should use normal approximation with all the formulas of Section 24.5.

Let us apply normal approximation to the above example (even though n is too small, so the approximation would not be accurate): we have $R \approx \mathcal{N}(\mu, \sigma^2)$, where

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1 = 5$$

and

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} = \frac{12}{7}$$

So the Z statistic is

$$Z = \frac{4 - 5}{\sqrt{12/7}} = -0.76$$

Now for the trend test the p-value is $\Phi(-0.76) = 0.2236$. For the cycle test the p-value is $1 - \Phi(-0.76) = 0.7764$.

25 Kolmogorov-Smirnov test

This is our last nonparametric test.

25.1 Test about the distribution function

Suppose we have a sample $x_1, \dots, x_n$ of observed values of an unknown random variable X and want to check whether X has a hypothetical distribution function $F(x)$. That is, let $F_X(x)$ denote the unknown distribution function of X, then we want to test the hypothesis $H_0\colon F_X(x) = F(x)$ against the alternative $H_1\colon F_X(x) \neq F(x)$.

25.2 Empirical distribution function

First of all, we need to estimate the unknown object, in this case the distribution function $F_X(x)$. Its value at any point x is equal to $\mathbb{P}(X \leq x)$. This probability can be estimates by using methods for binomials, see Section 3.7.

Consider the event $\{X \leq x\}$, think of it as 'success', then $p = \mathbb{P}(X \leq x)$ is the probability of success. In our sample, the empirical number of successes is $\#\{i\colon x_i \leq x\}$, hence the estimate of p is

$$\hat{p} = \frac{\#\{i\colon x_i \leq x\}}{n}$$

This is our estimate for $F_X(x)$, it is denoted by $F_n(x)$ and called the *empirical distribution function* (EDF).

25.3 Construction of EDF

Let $y_1, \dots, y_n$ denote the ordered sample $x_1, \dots, x_n$, that is $y_i = x_{(i)}$. Then

$$F_n(x) = \begin{cases} 0 & \text{for } \; x < y_1 \\ i/n & \text{for } \; y_i \leq x < y_{i+1} \\ 1 & \text{for } \; x \geq y_n \end{cases}$$

That is, $F_n(x)$ is a step function that jumps up by $1/n$ at every sample point, see illustration later.

25.4 Distribution of $F_n(x)$

The number of successes $\#\{i\colon x_i \leq x\}$ has distribution $b(n, p)$, where $p = F(x)$. Therefore the value $F_n(x)$ has distribution $\frac{1}{n} b(n, F(x))$. In particular, its mean value and variance are

$$\mathbb{E}\big(F_n(x)\big) = F(x), \qquad \mathsf{Var}\big(F_n(x)\big) = \frac{F(x)(1-F(x))}{n}$$

A typical error (standard deviation) is

$$\big|F_n(x) - F(x)\big| \sim \frac{\sqrt{F(x)(1-F(x))}}{\sqrt{n}}$$

25.5 Test statistic

Our test statistic will be

$$D_n = \sup_x \big|F_n(x) - F(x)\big|$$

which is the maximal distance between the graph of the the empirical distribution function and that of the hypothetical distribution function. Of course, whenever this distance is too large, we should reject H_0. Thus the critical region is $D_n > d$, where d is the critical value.

The value d is given in Table VIII in the book, it depends on the sample size n and the significance level α.

For large n, the table gives an asymptotic formula in the form $a/\sqrt{n}$, where a is some constant. This makes sense because the distance between $F_n(x)$ and $F(x)$ is of order $1/\sqrt{n}$, see the previous section.

25.6 Practical computation of D_n

Even though the formula for D_n involves finding the maximum difference $F_n(x) - F(x)$ over all real x's, in practice it is enough to compute that difference only at the sample points, i.e. we should compute $F_n(x_i) - F(x_i)$ for all $i = 1, \dots, n$ and take the largest one (in absolute value).

Note however that the empirical distribution function F_n is discontinuous at every sample point x_i, i.e. it takes two different values – one on the left and one on the right. We need to try both, i.e. we actually need to compute $2n$ differences, rather than n.

25.7 Example

Test the hypothesis that the sample

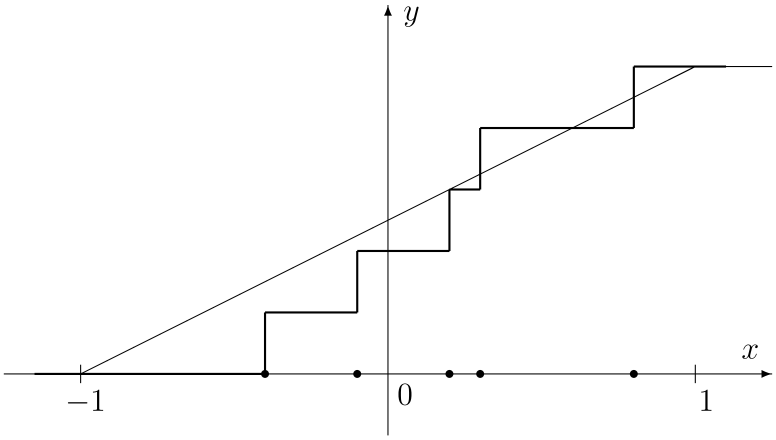
$$-0.4, \quad 0.2, \quad -0.1, \quad 0.8, \quad 0.3$$

came from the random variable $X = U(-1, 1)$. Let $\alpha = 0.1$.

Solution. The hypothetical distribution function is $F(x) = (x + 1)/2$. The table below records the values of $F(x)$ and $F_n(x)$ at the sample points, as well as the differences:

x	-0.4	-0.1	0.2	0.3	0.8
$F(x)$	0.3	0.45	0.6	0.65	0.9
left $F_n(x)$	0	0.2	0.4	0.6	0.8
right $F_n(x)$	0.2	0.4	0.6	0.8	1.0
left diff.	**0.3**	0.25	0.2	0.05	0.1
right diff.	0.1	0.05	0.0	0.15	0.1

The largest difference is $D_n = 0.3$ (in bold, see the table). Since $D_n < d = 0.51$ (from Table VIII), we accept H_0.



25.8 Confidence band for F_X

The empirical distribution function $F_n(x)$ provides the best estimate for the unknown distribution function $F_X(x)$. If we want a confidence interval with level $1 - \alpha$, then we move $F_n(x)$ up and down by the distance d, where d is taken from Table VIII and corresponds to the given sample size n and α.

This gives us two bounds - an upper bound for $F(x)$ and a lower bound for $F(x)$; the area in between is called the *confidence band.* That band is where we expect $F(x)$ to be, with probability $1 - \alpha$.

Note: if the confidence band sticks above the line $y = 1$ or below the line $y = 0$, it should be trimmed accordingly, since all distribution functions must be between 0 and 1.

25.9 Variant: two sample

Kolmogorov-Smirnov (KS) test can be adjusted to the problem discussed in the preamble to Chapter 24: given two samples, $x_1, \dots, x_{n_1}$ from a random variable X and $y_1, \dots, y_{n_2}$ from a random variable Y, we want to test the hypothesis $H_0\colon F_X = F_Y$ against the alternative $H_1\colon F_X \neq F_Y$.

In this case we construct the two empirical distribution functions: $F_{n_1}(x)$ for X and $F_{n_2}(x)$ for Y, and compute

$$D_n = \sup_x \bigl|F_{n_1}(x) - F_{n_2}(x)\bigr|$$

The critical region is again $D_n > d$, where d is given in Table VIII.

25.10 The p-value

There is a formula for the p-value of the Kolmogorov-Smirnov test:

$$\text{p-value} \approx Q\bigl(\sqrt{n}\, D_n\bigr)$$

where Q is a function defined by infinite series

$$Q(t) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$$

Here n is the size of the sample for the standard KS test (one sample), and

$$n = \frac{n_1 n_2}{n_1 + n_2}$$

for the adapted variant that deals with two samples of sizes n_1 and n_2.

The infinite series in the formula for $Q(t)$ converges very fast, so it can be easily computed numerically, by taking just a few terms.

25.11 Improvement: Anderson-Darling test

The formula for D_n has a certain drawback. Since the difference $F_n(x) - F(x)$ has variance

$$\mathsf{Var}\big(F_n(x) - F(x)\big) = \frac{F(x)(1 - F(x))}{n}$$

(see section 25.4), it is smaller when $F(x) \approx 0$ and $F(x) \approx 1$ and larger when $F(x) \approx 0.5$. Thus the formula

$$D_n = \sup_x \big|F_n(x) - F(x)\big|$$

is not well balanced – it is likely to overlook statistically significant differences in the 'extreme' ranges, where $F(x) \approx 0$ and $F(x) \approx 1$.

Anderson and Darling proposed a more balanced formula for D:

$$D_n^* = \sup_x \frac{\big|F_n(x) - F(x)\big|}{\sqrt{F(x)(1 - F(x))}}$$

It is harder to compute, but it makes the test more efficient.