MA 486-25 (Statistics), Dr Chernov Final exam Thu, May 31, 2001

Student's name _____

1. (10 pts) In the following table, 2000 individuals are classified by gender and by whether they answer Yes, No, or Not Sure in a certain poll. Test the null hypothesis that the probabilities of the answers are independent of the gender. Let $\alpha = 0.01$.

Gender	Yes	No	Not Sure
Male	530	280	60
Female	470	570	90

Answers: $\hat{p}_1 = 0.5, \hat{p}_2 = 0.425, \hat{p}_3 = 0.075$,

$$Q = \frac{(530 - 0.5 \cdot 870)^2}{0.5 \cdot 870} + \cdots + \frac{(90 - 0.075 \cdot 1130)^2}{0.075 \cdot 1130} = 76.026$$

Critical region

$$Q > \chi^2_{0.01}(2) = 9.210$$

Accept H_1.

2. (14 pts) In a regression problem, $n = 14$ data points are observed and the following values are found for the Gaussian brackets:

$$[X] = 20,\ [Y] = 6,\ [X^2] = 100,\ [XY] = 8,\ [Y^2] = 3$$

Find $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma^2}$. Find 98% confidence intervals for β, and σ^2. Test the hypothesis $\beta = 0$ against the alternative $\beta \neq 0$ at 98% level.

Find a 95% prediction interval for Y when $x = 2$.

Answers:

$$\hat{\beta} = -0.008, \qquad \hat{\alpha} = 0.44$$
$$\text{RSS} = 0.424, \qquad \hat{\sigma^2} = 0.0303$$

CI for β is

$$-0.008 \pm 2.681\sqrt{\frac{14 \cdot 0.0303}{12(100 - 400/14)}} = [-0.0676, 0.0516]$$

CI for σ^2:

$$\left[\frac{14 \cdot 0.0303}{26.22}, \frac{14 \cdot 0.0303}{3.571}\right] = [0.0162, 0.1187]$$

Test statistic

$$T = -0.3597$$

Critical region $|T| > 2.681$. Accept H_0.

Prediction interval

$$0.424 \pm \sqrt{\frac{14 \cdot 0.0303}{12}} \times \sqrt{1 + \frac{1}{14} + \frac{(2 - 20/14)^2}{100 - 400/14}} \times 2.179 = 0.424 \pm 0.425$$

3. (10 pts) Test the hypothesis $H_0 : F_X = F_Y$ at the 2% level, based on two samples from the X and Y distributions, respectively:

$$X : \ 3,\ 8,\ 4,\ 10,\ 2,\ -3,\ 4,\ 2,\ 4,\ 8,\ 2,\ 9,\ 3,$$

$$Y : \ 0,\ 7,\ -1,\ 12,\ 1,\ 5,\ -1,\ 6,\ 5,\ 0,\ 6$$

Use the run test with normal approximation. Find the p-value (as best as you can).

[Bonus] Compute the exact probability $P(R = 8)$, where R is the number of runs.

Answers: The combined sample

$$xyyyyyxxxxxxxyyyyyxxxxy$$

Hence the number of runs is $R = 6$. Then

$$\mu_R = \frac{2 \cdot 13 \cdot 11}{13 + 11} + 1 = 12.9$$

and, similarly, $\mathrm{Var}(R) = 5.656$,

$$Z = \frac{6 - 12.9}{\sqrt{5.656}} = -2.90$$

Critical region $Z < -2.054$. Accept H_1. The p-value is $\Phi(-2.90) = 0.0019$.

Bonus:

$$P(R = 8) = \frac{2\, C_{12,3}\, C_{10,3}}{C_{24,11}} = 0.02$$

4. (12 pts) (a) Given a random sample

$$-2.4,\ 69.9,\ 0.2,\ 4.3,\ -7.2,\ 11.7,\ 0.8,\ 2.9,\ 5.1,\ -91.6$$

Find an approximate 98% confidence interval for the median, m, of the corresponding distribution. What is the exact confidence level of your interval?

(b) For the previous sample, find the probability

$$P(0.2 < \pi_{0.8} < 4.3)$$

(c) For a random sample of size $n = 432$ from an unknown distribution, find an approximate 98% confidence interval for the first quartile, $\pi_{0.25}$. Use normal approximation. Give the answer in the form $(y_{...}, y_{...})$.

Answers:

(a) CI is $[Y_2, Y_9] = [-7.2, 11.7]$. The exact confidence level is 97.86%.

(b)

$$P(Y_4 < \pi_{0.8} < Y_7) = P(4 \le b(10, 0.8) < 7) = 0.12$$

(c)

$$r_1 = 432 \cdot 0.25 + 0.5 - 2.326\sqrt{432\,(1/4)\,(3/4)} \approx 88$$

$$r_2 = 432 \cdot 0.25 + 0.5 + 2.326\sqrt{432\,(1/4)\,(3/4)} \approx 129$$

So, CI is $[Y_{88}, Y_{129}]$.

5. (10 pts) Let X_1, X_2, X_3 be three independent random variables that have some normal distributions $N(\mu_i, \sigma^2)$. Test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

at the level $\alpha = 5\%$. The observed data are given in the table below:

X_1 :	7	10	9	10		
X_2 :	7	5	7	6	5	
X_3 :	9	7	10	8	10	10

Construct an ANOVA table (without p-value) and state your conclusion.

Answers: sample means are 9, 6, 9, the grand mean is 8.

$$SS(T) = 4(9-8)^2 + 5(6-8)^2 + 6(9-8)^2 = 30$$

$$SS(TO) = 1^2 + 2^2 + 1^2 + 2^2 + \cdots + 0^2 + 2^2 + 2^2 = 48$$

So, $SS(E) = 48 - 30 = 18$. The Anova table:

Treatment	2	30	15
Error	12	18	1.5

Then F-ratio is $15/1.5 = 10$. The critical region

$$\text{F} - \text{ratio} > 3.89$$

Accept H_1.

6. (14 pts) Test the hypothesis $H_0 : m_X = m_Y$ against $H_1 : m_X > m_Y$. The following data were observed:

$$X : \ 4,\ 9,\ 1,\ 6,\ 5,\ 10$$

$$Y : \ 2,\ 5,\ 0,\ -1,\ 8,\ 1$$

(a) Use the median test, compute the p-value (use the exact formula for the probabilities of the V-statistic).
(b) Use the Wilcoxon test (for two samples), again compute the p-value.
(Bonus) Sketch the q-q plot.
Which hypothesis would you accept at the 10% level?

Answers:

(a) by the median test:

$$V = 2$$

the p-value is

$$\frac{C_{6,0}\, C_{6,6}}{C_{12,6}} + \frac{C_{6,1}\, C_{6,5}}{C_{12,6}} + \frac{C_{6,2}\, C_{6,4}}{C_{12,6}} = 0.28$$

(b) by the Wilcoxon test: ranks are

$$\mathbf{1},\ \mathbf{2},\ \mathbf{3.5},\ 3.5,\ \mathbf{5},\ 6,\ \mathbf{7.5},\ 7.5,\ 9,\ \mathbf{10},\ 11,\ 12$$

(the x-ranks are in bold). So,

$$W = 1 + 2 + 3.5 + 5 + 7.5 + 10 = 29$$

$$Z = \frac{29 - 6 \cdot 13/2}{\sqrt{6 \cdot 6 \cdot 13/2}} = -1.60$$

Critical region is $Z < -1.282$. Accept H_1. Tye p-value is $\Phi(-1.60) = 0.0548$.

7. (10 pts) The following numbers were generated by a computer program:

$$-0.4,\ 0.2,\ -0.1,\ 0.8,\ 0.3,\ -0.6,\ 0.1,\ 0.7,\ 1.0,\ -0.5$$

Test the hypothesis that the program generates a uniform random variable on the interval $(-1, 1)$, i.e. $X = U(-1, 1)$. Use the Kolmogorov-Smirnov test with Table VIII at the level $1 - \alpha = 95\%$. Also, sketch an empirical distribution function.

Indicate how you would construct a 95% confidence band around the empirical distribution function.

Answers: The distribution function is $F(x) = (x + 1)/2$. The test statistic is $D_n = 0.2$. Critical region is $D_n > 0.41$. Accept H_0.

8. (10 pts) Five experimental data points are given:

$$(2,0),\ (4,-1),\ (1,-1),\ (-2,5),\ (0,2)$$

Compute the Gaussian brackets $[X]$, $[Y]$, $[X^2]$, $[XY]$, $[Y^2]$. Estimate the parameters α and β of the regression line $y = \alpha + \beta x$. Compute the RSS (Residual Sum of Squares).

Draw a scatter plot, marking the data points and the regression line.

[Bonus] Compute the sample covariance c_{xy} and the sample correlation coefficient r.

Answers:

$$[X] = 5,\ \ [Y] = 5,\ \ [XY] = -15,\ \ [X^2] = 25,\ \ [Y^2] = 31$$

$$\hat{\beta} = -1, \qquad \hat{\alpha} = 2$$

$$\mathrm{RSS} = 6$$

$$c_{xy} = 5, \qquad r = -0.877$$

9. (10 pts) It is claimed that the median price of houses in Alabama is \$150,000. A random sample of 11 houses on the market gives the following prices (in thousands of dollars):

$$315, 122, 84, 169, 219, 99, 175, 280, 63, 100, 146$$

Test the hypothesis $H_0 : m = 150$ against $H_1 : m \neq 150$ at the level 5%. Use the Wilcoxon test.

[Bonus] Find the p-value of the test.

Answers: ranks are

$$-1,\ 2,\ 3,\ -4,\ -5,\ -6,\ -7,\ 8,\ -9,\ 10,\ 11$$

their sum is

$$W = 2$$

Hence

$$Z = \frac{2}{\sqrt{11 \cdot 12 \cdot 23/6}} = 0.89$$

Critical region is

$$|Z| > 1.960$$

Accept H_0.

The p-value is $P(|N(0,1)| > 0.089) = 2(1 - \Phi(0.089)) = 0.9282$.