Computer projects for Mathematical Statistics, MA 486.

Some practical hints for doing computer projects with MATLAB:

You can save your project to a text file (on a floppy disk or CD or on your web page), so you can come back to the computer lab and resume an unfinished project at a later date/time.

Some simple commands can be just typed and executed. For example, **X=randn(10,1)** produces a vector of 10 random normal numbers, and then **m=mean(X)** computes the sample mean. However, you'll need to do a larger project, so you'll need to type all your commands in an m-file. It should be a text file containing MATLAB commands and (preferably) comments that will help the instructor read it. If you have an m-file (for example, **myfile.m**), then you simply open the MATLAB window on the desktop and type the file name (i.e., **myfile**), then your project will appear on the screen and will be executed.

A sample MATLAB file **means.m** is posted on the instructor's web page, you can download it and try it out. The file contains detailed comments explaining what it does. Your project will have to be larger and more complicated than that file, but you can use that file to start your project.

Note: projects that are already taken are marked here by ⊗

1. ⊗ Simulate n values of a normal random variable $X = N(\mu, \sigma^2)$ (choose μ and $\sigma > 0$ as you like), and compute the sample mean $\bar{x}$, sample median m, sample standard deviation s. Plot these quantities as functions of n on three separate plots (see a general remark in the end). Check that $\bar{x}$ and m converge to μ, as $n \to \infty$, and s converges to σ. Which one converges to μ faster, the sample mean or the sample median? To be sure, estimate the variance of both $\bar{x}$ and m for a particular value of n, such as $n = 100$ (by generating, say, 10000 different random samples of size n and computing the sample variance of the resulting estimates $\bar{x}$ and m. The estimate with the smaller variance is better).

2. ⊗ Simulate n values of an exponential random variable X with parameter $\lambda > 0$ (of your choice), and compute the sample mean $\bar{x}$, sample median m, sample standard deviation s. Plot these quantities as functions of n on three separate plots (see a general remark in the end). Do $\bar{x}$, m, and s converge to any limit values, as $n \to \infty$? What are those values and how are they related? (To describe the relation, you need to recall the properties of exponential random variables.) Estimate the variance of both $\bar{x}$ and m for a particular value of n, such as $n = 100$ (by generating, say, 10000 different random samples of size n and computing the sample variance of the resulting estimates $\bar{x}$ and m). The estimate with the smaller variance is better.

3. Simulate $n = 10000$ values of a variable X with a triangular density function is
$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$
Use the rejection method. For the 'auxiliary' random variable Y use the uniform $U(0, 1)$. Plot the histogram of the resulting values. Does it look like a triangle? Compute the theoretical mean and variance of the above distribution and compare them to the sample mean and variance.

   Also count the number of times m you had to call the random number generator (it is of course greater than $n = 10000$) and compute the ratio n/m. It is consistent with its theoretical value of $1/c$ given in Section 2.9 of classnotes?

4. Simulate n values of the Cauchy random variable X, whose distribution function is

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x$$

Use the inversion method. Note that X has no mean value or variance, but its median is zero. Then compute the sample mean $\bar{x}$, sample median m, sample standard deviation s, and the interquartile range IQR. Plot these quantities as functions of n on four separate plots (see a general remark in the end). Determine if $\bar{x}$ and m converge to anything, as $n \to \infty$. Explain. Do s and IQR seem to converges to anything? Explain. Estimate the variance of both $\bar{x}$ and m for a particular value of n, such as $n = 100$ (by generating, say, 10000 random samples of size n and computing the sample variance of the resulting estimates $\bar{x}$ and m).

5. "Simulate the number π". Simulate n uniformly distributed random points in the square

$$K = \{-1 < x < 1,\ -1 < y < 1\}$$

Determine the number of points, m, that fall into the unit disk $x^2+y^2 < 1$. Note that the probability for a random point to be in the unit disk is $\pi/4$. By the law of large numbers we expect m/n converges to $\pi/4$ as $n \to \infty$. Then use $4m/n$ as an estimate of π. Plot $4m/n$ as a function of n and check that it converges to π as $n \to \infty$. How big n should be (give a "ball park" figure) in order for $4m/n$ to get within $\varepsilon = 0.001$ from π? (Use the central limit theorem.)

6. Simulate n values of a normal random variable $X = N(\mu, \sigma^2)$ (choose μ and $\sigma > 0$ as you like), and compute the sample mean $\bar{x}$ and sample standard deviation s. The theory claims that these estimates are independent. Check this claim experimentally. Repeat this experiment M times and compute the sample correlation coefficient between $\bar{x}$ and σ. Plot this correlation coefficient as functions of M (see a general remark in the end). Check that it converges to zero, as $M \to \infty$ (going up to $M = 10000$ would be enough). The value of n should be small, such as $n = 10$ or $n = 20$.

7. Simulate a sample of $n = 100$ random numbers in which 80 are drawn from a normal distribution $N(5, 1)$ and the other 20 are drawn from

a uniform distribution $U(-100, 100)$ (the latter represent "contamination", or "background noise"). For this sample calculate (a) the sample mean, (b) the trimmed sample means discarding 10%, 20%, and 30% of the data, (c) the sample median. Which estimate appears to be the most accurate? Repeat this experiment 10000 times and compute the standard deviation for each of these five estimates. The estimate with the smallest variance is best.

8. Simulate a sample of $n = 100$ random numbers in which 50 are drawn from a normal distribution $N(5, 1)$ and the other 50 are drawn from a uniform distribution $U(-100, 100)$ (the latter represent a "contamination", or a "background noise"). For this sample calculate (a) the sample mean, (b) the trimmed sample means discarding 10%, 30%, and 60% of the data, (c) the sample median. Which estimate appears to be the most accurate? Repeat this experiment 10000 times and compute the standard deviation for each of these five estimates. The estimate with the smallest variance is best.

9. Simulate a sample of $n = 100$ random numbers drawn from the Cauchy distribution, whose distribution function is

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} x$$

Use the inversion method. Note that X has no mean value or variance, but its median is zero. Then calculate (a) the sample mean, (b) the trimmed sample means discarding 10%, 30%, and 60% of the data, (c) the sample median. Plot these estimates as functions of n (see a general remark in the end). Which estimate appears to be closest to zero? Repeat this experiment with a particular value of n (say $n = 100$) 10000 times and compute the standard deviation for each of these five estimates. The estimate with the smallest standard deviation is best.

10. ⊗ "Simulate the t-distribution". Simulate $n = 8$ values of a normal random variable $N(\mu, \sigma^2)$ (choose μ and $\sigma > 0$ as you wish). Then compute

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

Repeat the experiment 10000 times and plot the histogram of the resulting T-values. Compare it to the density function of the t-random

variable with $n - 1 = 7$ degrees of freedom. Superimpose the two plots to demonstrate their similarity (ideally, they should coincide). Compute the sample mean and variance and compare them to the theoretical mean and variance of the t random variable (see Exercise 5.5-4 in the book).

11. (a) The theory says that if X is a t random variable with one degree of freedom, then so is $1/X$. Verify this claim experimentally: simulate 10000 values of a t random variable with one degree of freedom, plot the histogram of the sample and the histogram of their reciprocal values. Compare. (b) The theory also says that if X and Y are two independent $N(0, 1)$ random variables, then X/Y is a t random variable with one degree of freedom. Verify this claim experimentally: simulate two samples from the $N(0, 1)$ distribution, with 10000 values each. Plot the histogram of the corresponding ratios x_i/y_i, $1 \leq i \leq 10000$, and compare it with the histograms obtained in step (a).

12. (a) The theory says that if Z is a standard normal random variable and U is a χ^2 random variable with r degrees of freedom, and Z and U are independent, then $T = \frac{Z}{\sqrt{U/r}}$ is a t random variable with r degrees of freedom. Verify this claim experimentally: set $r = 10$, simulate 10000 values of pairs (Z, U), compute 10000 values of T and plot the histogram of the T-values. Superimpose it with the density of the t random variable with r degrees of freedom provided by MATLAB to demonstrate their similarity (ideally, they should coincide). Also compute the sample mean and sample variance. Do they match the theoretical values of the mean value and variance of $t(r)$? (see Exercise 5.5-4 in the book). (b) The theory also says that if X is a t random variable with r degrees of freedom, then X^2 is an F random variable with 1 and r degrees of freedom, i.e. $X^2 = F(1, r)$. Verify this claim experimentally: set $r = 10$, simulate 10000 values of $t(r)$, compute their squares, and plot the histograms of the resulting 10000 values. Superimpose it with the density of the t random variable with r degrees of freedom provided by MATLAB to demonstrate their similarity (ideally, they should coincide). Also compute the sample mean and sample variance. Do they match the theoretical values of the mean value and variance of $F(1, r)$?

13. ⊗ "Simulate the χ^2-distribution". Simulate $n = 8$ values of a normal random variable $N(\mu, \sigma^2)$ (choose μ and $\sigma > 0$ as you wish). Then compute

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

Repeat the experiment 10000 times and draw a histogram of the resulting K-values. Compare it with the density function of the χ^2 random variable with $n - 1 = 7$ degrees of freedom provided by MATLAB. Superimpose the two plots to demonstrate their similarity (ideally, they should coincide). Compute the sample mean and sample variance and compare them to the theoretical mean and variance of the χ^2 random variable.

14. ⊗ "Simulate the χ^2-distribution". Simulate 50 values of a discrete random variable that takes values $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$ (set $k = 5$ and choose arbitrary positive probabilities). Then compute the Q statistics. Repeat the experiment 10000 times and draw a histogram of the resulting Q-values. Compare it to the density function of the χ^2 random variable with $k - 1 = 4$ degrees of freedom. Superimpose the two plots to demonstrate their similarity (ideally, they should coincide). Compute the sample mean and sample variance and compare them to the theoretical mean and variance of the χ^2 random variable.

15. ⊗ "Simulate the F-distribution". Simulate $x_1, \ldots, x_{15}$ values of a normal random variable $N(\mu_x, \sigma_x^2)$ and $y_1, \ldots, y_8$ values of another normal random variable $N(\mu_y, \sigma_y^2)$ (choose the parameters as you wish). Then compute

$$F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$$

Repeat the experiment 10000 times and draw a histogram of the resulting F-values. Compare it to the density function of the F random variable with the corresponding degrees of freedom, provided by MATLAB. Superimpose the two plots to demonstrate their similarity (ideally, they should coincide). Compute the sample mean and sample variance and compare them to the theoretical mean and variance of the F random variable.

16. ⊗ Simulate n values of a Poisson random variable $X = \text{poisson}(\lambda)$ (choose $\lambda > 0$ as you like), and compute the sample mean $\bar{x}$, sample

median m, sample standard deviation s. Plot these quantities as functions of n on three separate plots (see a general remark in the end). Do these statistics converge to any limit values, as $n \to \infty$? What are those limits? Do your conclusions agree with the theory? Estimate the variance of $\bar{x}$ and m for a particular value of n, such as $n = 100$ (by generating 10000 random samples of size n and computing the sample variance of the resulting estimates $\bar{x}$ and m). Which of these two estimates is better?

17. Explore the accuracy of the linear least squares fit. Position $n = 10$ points $x_1, \dots, x_n$ on the interval $[-10, 10]$ arbitrarily (however, make sure that $\sum x_i = 0$), then generate y-values by the formula

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

(choose α and $\beta \neq 0$ as you wish), where ε_i are normal random variables $N(0, \sigma^2)$ with $\sigma = 0.2$. Then fit a line $y = \hat{\alpha} + \hat{\beta} x$ to the data. Determine the accuracy of the estimates $\hat{\alpha}$ and $\hat{\beta}$ as follows. Repeat the experiment 1000 times and estimate the mean square error of $\hat{\alpha}$ and $\hat{\beta}$. Then try to rearrange the points $x_1, \dots, x_n$ to increase the accuracy of the estimates (i.e. minimize the mean square errors). What arrangement yields the maximum accuracy?

18. Explore the accuracy of the quadratic least squares fit. Position $n = 10$ points $x_1, \dots, x_n$ on the interval $[-10, 10]$ arbitrarily (however, make sure that $\sum x_i = 0$), then generate y-values by the formula

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$$

(choose α, $\beta \neq 0$, and $\gamma \neq 0$ as you wish), where ε_i are normal random variables $N(0, \sigma^2)$ with $\sigma = 0.1$. Then fit a parabola $y = \hat{\alpha} + \hat{\beta} x + \hat{\gamma} x^2$ to the data. Determine the accuracy of the estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ as follows. Repeat the experiment 1000 times and estimate the mean square error of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$. Then try to rearrange the points $x_1, \dots, x_n$ to increase the accuracy of the estimates (i.e. minimize the mean square errors). What arrangement yields the maximum accuracy?

19. Explore the properties of the linear least squares fit. Position $n = 10$ points $x_1, \dots, x_n$ on the interval $[-10, 10]$ arbitrarily, for example, put

them equally spaced (however, make sure that $\sum x_i = 0$), then generate y-values by the formula

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

(choose α and $\beta \neq 0$ as you wish), where ε_i are normal random variables $N(0, \sigma^2)$ with $\sigma = 0.2$. Then fit a line $y = \hat{\alpha} + \hat{\beta} x$ to the data. The theory says that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are independent. Verify this claim experimentally. Repeat this experiment N times and estimate the covariance between $\hat{\alpha}$ and $\hat{\beta}$. Plot this average as functions of N (see a general remark in the end). Check that it converges to zero, as $N \to \infty$ (going up to $N = 10000$).

20. This project is larger than others and can be done for some extra credit. Explore the polynomial fit with various degrees. First, position $n = 7$ points $x_1, \ldots, x_n$ on the interval $[-9, 9]$ equally spaced, then generate y-values by the formula

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + \varepsilon_i$$

(choose α, β, and γ as you wish, but not close to zero), where ε_i are normal random variables $N(0, \sigma^2)$ with $\sigma = 0.1$. Plot the data, along with the original parabola $y = \alpha + \beta x + \gamma x^2$ used in the simulation. Then fit the polynomial of degree $k = 1, 2, 3, 4, 5, 6$ to the data. Plot the polynomials and see how well they approximate the data points and the original parabola. Compute the RSS (residual sum of squares) and see how it decreases as the degree k grows. But do the higher degree polynomials fit the original parabola better or not? Next, use the "leave-one-out" scheme for cross-validation and recompute the average RSS for each degree $k = 1, 2, 3, 4, 5, 6$. Does this one also decrease as k grows? What degree k yields the smallest the average RSS now? (some MATLAB code for cross-validation scheme may be obtained from the instructor.)

21. Let X be a Poisson random variable with parameter λ. Theory claims that the variable $Y = \sqrt{X}$ has an almost constant variance as $\lambda \to \infty$. Verify this claim experimentally and determine the value of that constant as follows. Pick a large λ, simulate $n = 10000$ values of X, transform them to the values of Y, and compute the sample variance of

the latter. Repeat this experiment for a few large values of λ to make sure that the results does not change much. Plot the obtained values, as a function of λ.

22. Let X be a normal random variable $N(\lambda, \lambda)$. Theory claims that the variable $Y = \sqrt{X}$ has an almost constant variance as $\lambda \to \infty$. Verify this claim experimentally and determine the value of that constant. Pick a large λ, simulate $n = 10000$ values of X, transform them to the values of Y, and compute the sample variance of the latter. Repeat this experiment for a few large values of λ to make sure that the results does not change much. Plot the obtained values, as a function of λ.

23. $\otimes$ Verify that the random number generator **rand** produces independent values of a uniform random variable. Simulate n values, divide them into two groups – odd-numbered values and even-numbered values. Compute the sample correlation coefficient between these two groups. Plot its value as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it converge to zero? Repeat this experiment by dividing the n generated values into two groups differently: putting numbers with indices $1, 2, 5, 6, 9, 10, \ldots$ into one group and numbers with indices $3, 4, 7, 8, 11, 12, \ldots$ into the second group. Do you still observe convergence to zero?

24. $\otimes$ Verify that the random number generator **randn** produces independent values of a normal random variable. Simulate n values, divide them into two groups – odd-numbered values and even-numbered values. Compute the sample correlation coefficient between these two groups. Plot its value as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it converge to zero? Repeat this experiment by dividing the n generated values into two groups differently: putting numbers with indices $1, 2, 3, 7, 8, 9, \ldots$ into one group and numbers with indices $4, 5, 6, 10, 11, 12, \ldots$ into the second group. Do you still observe convergence to zero?

25. Let U and V be independent uniform $U(0, 1)$ random variables. Theory claims that $X = \sqrt{-2 \ln U} \cos(2\pi V)$ and $Y = \sqrt{-2 \ln U} \sin(2\pi V)$ are independent normal random variables. Verify this claim experimentally. Generate n pairs of values of uniform random variables, convert them to n pairs of X and Y. Put together X values in a vector of

length n, and put together Y values in a vector of length n. Compute the sample correlation coefficient between the X vector and the Y vector. Plot its value as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it converge to zero? Redo this experiment by changing the formulas to $X = (\ln U)\cos(2\pi V)$ and $Y = (\ln U)\sin(2\pi V)$. Do you observe independence now?

26. Verify that the random number generator **rand** produces values of a uniform random variable $U(0, 1)$. Simulate n values and plot the empirical distribution function. Superimpose it on the actual distribution function. Do they look similar? Apply Kolmogorov-Smirnov test to check their similarity. Plot the value $d_n\sqrt{n}$ as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it seem to (a) increase, or (b) decrease, or (c) fluctuate?

27. Verify that the random number generator **randn** produces values of a normal random variable $N(0, 1)$. Simulate n values and plot the empirical distribution function. Superimpose it on the actual distribution function. Do they look similar? Apply Kolmogorov-Smirnov test to check their similarity. Plot the value $d_n\sqrt{n}$ as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it seem to (a) increase, or (b) decrease, or (c) fluctuates?

28. Verify that the random number generator **poissrnd** produces values of a Poisson random variable. Choose a value for the parameter (1 or 2 will do). Simulate n values and plot the empirical distribution function. Superimpose it on the actual distribution function. Do they look similar? Apply Kolmogorov-Smirnov test to check their similarity. Plot the value $d_n\sqrt{n}$ as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does it seem to (a) increase, or (b) decrease, or (c) fluctuate?

29. Let U and V be independent uniform $U(0, 1)$ random variables. Theory claims that $X = \sqrt{-2\ln U}\cos(2\pi V)$ and $Y = \sqrt{-2\ln U}\sin(2\pi V)$ are standard normal random variables. Verify this claim experimentally. Generate n pairs of values of uniform random variables, convert them to n pairs of X and Y, then combine the X and Y vectors into one vector Z of length $2n$. Do that for $n = 5000$ and draw a histogram of the resulting 10000 Z-values. Compare it to the density function of

the standard normal random variable $N(0,1)$, provided by MATLAB. Superimpose the two plots to demonstrate their similarity (ideally, they should coincide). Compute the sample mean and sample variance of Z and compare them to the theoretical mean and variance of the normal random variable.

30. ⊗ "Integrate by Monte-Carlo". Suppose you need to compute a definite integral $I = \int_a^b f(x)\,dx$. This can be done by the Monte-Carlo method that involves random numbers. Generate n (independent) values of a uniform random variable $U(a,b)$, call them $x_1, \ldots, x_n$, and compute $S_n = f(x_1) + \cdots + f(x_n)$. Then $\frac{b-a}{n}\, S_n$ will approximate I (the higher n the better). You need to try this method on two integrals: $\int_1^3 \frac{1}{x}\,dx$ and $\int_0^1 \frac{1}{\sqrt{x}}\,dx$. In both cases, do the computation and plot $\frac{b-a}{n}\, S_n$ as a function of n, starting at $n = 100$ and up to $n = 10000$ (see a general remark in the end). Does your estimate converge to the value of the integral?

General remark. In many projects, you are supposed to generate a sample of n values of a certain random variable, compute some statistics and then plot their values as functions of n. Do this for certain values of n such as n = 100, 200, 300,...,10000. This gives you 100 different values of each statistic, well enough for a plot. Important: when increasing n from 100 to 200, then from 200 to 300, etc., **do not** generate a new sample for every new value of n. Instead, add 100 new values to the old sample (that would make your plots much smoother and nicer). How to do that in MATLAB? Keep the vector of the old sample, generate a new vector of 100 random values, and append it to the old vector. The MATLAB file *means.m* available from the instructor's web page shows how to do this.