Math 125—Introductory Statistics
Measurable Outcomes

Mathematics Department, UMass Boston

Reference text: Numbers in brackets refer to sections of Freedman, Pisani, and Purves, *Statistics,* fourth edition.

Note: Outcomes marked **(Optional)** may appear on the final exam with the unanimous consent of all instructors.

1. Design of experiments

1(a) **(Optional)** Design a controlled experiment. Consider using randomization, placebos, and double blinding, and explain the pros and cons of each. [1.1]

1(b) **(Optional)** Set up workarounds to avoid potential pitfalls when using historical controls. [1.2, and 2.1 to 2.3]

1(c) **(Optional)** Work successfully with observational studies: where the researcher does not assign the subjects to treatment or control. [2.1]

1(d) **(Optional)** Avoid thinking that association proves causation. [2.1]

1(e) **(Optional)** Be prepared to control for a confounding factor. [2.5]

2. The histogram

2(a) Define a histogram as a series of blocks, each of which represents a percentage of the distribution by its area. [3.1]

2(b) Determine the height of a block and use that height to draw the block. [3.2]

2(c) Understand the concept of a density scale in general and, in particular, for the height of a block, being careful to use the exact units. [3.3]

2(d) Characterize a variable as: (A) qualitative or (B) quantitative, and decide if a quantitative variable is (B i) discrete or (B ii) continuous. [3.4]

2(e) Find the exact endpoints for a block of continuous data, say for a designated integer value (for example, 63 inches). Give two different answers, depending on whether the data were rounded or truncated. [3.4]

2(f) Find exact endpoints and draw the base of a block, for an interval of whole-numbered values. Such a case would be 63 to 68 inches, where the heights were characterized as 63, 64, 65, 66, 67, or 68 inches. Note the width of the base. Give two different answers, depending on whether the data were rounded or truncated. [3.4]

2(g) Apply the endpoint convention for discrete data—the continuity correction. [3.4]

2(h) **(Optional)** Control for a variable. [3.5]

2(i) **(Optional)** Make a cross-tabulation. [3.6]

3. The average and the standard deviation

3(a) Define and calculate the average and the median. [4.2]

3(b) Recognize the difference between cross-sectional and longitudinal data. [4.2]

3(c) Represent the average and the median of a given histogram. [4.3]

3(d) Define the Standard Deviation (SD), and state the rough percentage of the entries on a list which are within one or two SDs. [4.5]

3(e) Define the root-mean-square (RMS), and calculate the SD by finding the RMS deviation from average.

4. The normal curve and its use to approximate data

4(a) Recognize the normal curve and state some of its properties. [5.1]

4(b) Define standard units; and convert a value to standard units, both verbally and with the use of the formula. [5.1]

4(c) Find the area under the normal curve for a specified interval using the text-book provided table that lists the area from $-z$ to z and the height of the curve at z (for z in increments of 0.05 from 0.00 to 4.45). [5.2]

4(d) Use the normal approximation for data to find the percentage of the population that falls into a given range, once the average and SD have been found for the distribution. [5.3]

4(e) Define a percentile and a percentile rank, and define the interquartile range. [5.4]

4(f) When the data follow the normal curve, find the percentile rank for a given value and determine the value of the distribution at a given percentile. [5.5]

4(g) Define a change of scale as multiplying each member of the list by he same constant and then adding the same constant to the result or vice versa, and observe the common changes of scale between Fahrenheit and Celsius temperatures. [5.6]

4(h) Find the effects of various changes of scale on the average, SD, and the standard units. [5.6]

5. Measurement error

5(a) Estimate the likely size of the chance error in a single measurement based on a series of repeated measurements. [6.1]

5(b) Think about how to deal with outliers in the series of repeated measurements. [6.2]

5(c) Contrast the effects of bias (systematic errors) and of chance errors. [6.3]

6. Plotting points and lines

6(a) Read points off a graph. [7.1]

6(b) Plot points. [7.2]

6(c) Define the slope and the intercept of a line. [7.3]

6(d) Plot a line. [7.4]

6(e) Given an equation in the form $y = mx + b$, plot various points and show that they fall on a line, called the graph of the equation. Find the slope and intercept of the line. [7.5]

6(f) Given a line drawn in the Cartesian coordinate system, find its equation and find the height of this line for a given x. [7.5]

6(g) Plot the line for a given equation in the form $y = mx+b$. [7.5]

7. Correlation

7(a) Define and draw the scatter diagram and use it to estimate the association between the two variables. [8.1]

7(b) Define the correlation coefficient (r) and what it measures, giving the range of possible values. Explain positive vs. negative correlation, and visualize various values by drawing their clouds. [8.2]

7(c) Interpret r graphically by estimating the strength of the association from the look of the scatter diagram. [8.2]

7(d) Define the SD line. [8.3]

7(e) Compute the correlation coefficient for a small data set by converting each variable to standard units and finding the average of these products. [8.4]

7(f) Recognize the correlation coefficient as a pure (unitless) number. [9.1]

7(g) Show that the correlation coefficient is not affected by interchanging the two variables, adding the same number to all the values of one variable, or multiplying all the values of one variable by the same positive number. [9.1]

7(h) Appreciate the limited relevance of the correlation coeficient in the case of non-linear assocition, or when there are outliers. [9.3]

7(i) Be aware that ecological correlations (based on rates or averages) tend to overstate the strength of an association. [9.4]

7(j) Realize that correlation measures association, but association is not the same as causation. [9.5]

8. Regression

8(a) Define regression. [10.1]

8(b) Using the regression method, state the amount of increase for y associated with each increase of one SD in x. [10.1]

8(c) Estimate the average value of y for a given x. [10.1]

8(d) Define the graph of averages and, starting with the graph of averages, explain what the regression line does. [10.2]

8(e) Use the regression method for individuals to predict the value of y, based on that individual's x. [10.3]

8(f) Use the regression method to predict percentile ranks—for football-shaped graphs. [10.3]

8(g) Describe the regression effect ("regression to mediocrity") and avoid thinking that this is anything but the spread around the line. In particular, in a test-retest situation, do not fall for the regression fallacy that when first tests having a particular result (r_1) associate with a specific result (r_2) on the second test, second tests having the result (r_2) will automatically associate with exactly the same aforementioned result (r_1) on the first test. [10.4]

8(h) Draw and explain the two possible regression lines for a given scatter diagram. [10.5]

9. The r.m.s. error for regression

9(a) Derive the r.m.s. error from the all of the vertical distances between points in the scatter diagram and the regression line. [11.1]

9(b) Derive the SD from the errors (vertical distances) of all the points in the scatter diagram from a horizontal line through the average of y. [11.1]

9(c) Apply the formula to compute the r.m.s. error with the correct units. [11.2]

9(d) Plot the residual errors. [11.3]

9(e) Note the similarity of errors in vertical strips of a homoscedastic scatter diagram. [11.4]

9(f) Use the normal curve inside a vertical strip in the case of a football-shaped scatter diagram to find probabilities. [11.5]

10. The regression line

10(a) Find the slope and the intercept of the regression line of y on x. [12.1]

10(b) In an observational study, do not use the regreesion line to predict the results of interventions to change the value of x. [12.1]

10(c) Show that the method of least squares produces the regression line. [12.2]

11. Probability

11(a) Define probability; show several possible ways of expressing chance (as a fraction, decimal, or percent, or as odds); state the range of values (0% to 100%); relate a chance to the chance of the opposite thing; and assign chances in the box model. [13.1]

11(b) Define, find, and use conditional probabilites. Introduce $P(A)$ and $P(A|B)$ notation. [13.2]

11(c) State and apply the multiplication rule. (Please note that "both A and B" is the same as "A and B;" the phrase "all of A, B, and C" is the same as "A and B and C;" and "none of: A, B, C" is the same as "(not A) and (not B) and (not C).") [13.3]

11(d) Define independence and decide if two things are independent. [13.4]

11(e) Assign probabilites by listing the ways. [14.1]

11(f) Define mutually exclusive and decide whether two things are mutually exclusive. [14.2]

11(g) State and apply the addition rule. (Assume that "at least one of A or B" means the same as "A or B." In Math 125, "or" is always inclusive. So "A or B" means the same as "A or B or both" or the sometimes-used "A and/or B.") [14.2]

11(h) Find the logical opposite of a disjunction: "not (A or B)" is the same as "(not A) and (not B)" and both are the same as "neither A nor B."

11(i) Calculate the probability of a disjunction when the events are not mutually exclusive by first calculating the probability of the opposite event. [14.4]

11(j) Show that real-world probabilities are often very closely approximated by the model. [14.5]

12. Binomial Probabilities

12(a) Find the number of ways of arranging k or one thing and $n-k$ of another thing in a row: the binomial coefficient "n choose k." [15.1]

12(b) State and apply the binomial formula, which gives the chance that an event will occur exactly k times out of n. [15.2]

13. The law of averages

13(a) State and explain the law of averages: the difference between the *number* observed and the *number* expected gets larger as the number of trials increases while the difference between the *percentage* observed and the *percentage* expected gets smaller. [16.1]

13(b) Describe the sum of draws for a box model. [16.2 and 16.3]

13(c) Apply box models to gambling situations. [16.4]

14. The expected value and standard error

14(a) Define the expected value and write and apply the formula for the sum of the draws made at random with replacement from a box. [17.1]

14(b) Describe the standard error (SE) and write and apply the square root law for the standard error for the sum of draws randomly drawn with replacement from a box of numbered tickets. [17.2]

14(c) Use the normal curve to find probabilites for the sum of the draws when the number of draws is sufficiently large. [17.3]

14(d) Apply the short-cut to find the SD when the list contains only two different numbers. [17.4]

14(e) Classify and create a new zero-one box for the number of times tickets with a certain number or numbers result when draws are made at random with replacement from a box of numbered tickets. Use the sum of the draws from that box to get the expected value and standard error for the count. [17.5]

14(f) Recognize that almost all gambling games will lead to the player's ruin in the long run. [17.7]

15. The normal approximation for probability histograms

15(a) Define and draw a probability histogram, where chance is represented by area. [18.2]

15(b) Observe that probability histograms of the number of heads when a coin is tossed a large number of times are approximated by the normal curve. Start with a 0-1 counting box. [18.3]

15(c) Apply the continuity correction, where the base of the rectangle over a counting number runs from that number minus 0.5 to that number plus 0.5. Decide whether the correction is needed in a specific example based on the size of the rectangles and the accuracy demanded. [18.4]

15(d) **(Optional)** Do not expect the normal approximation for the sum of the draws to apply as well when the box is lopsided unless the sum of the draws is quite large.

15(e) State the central limit theorem. [18.6]

16. Sampling

16(a) Define the terms used in sampling (population, sample, parameter, and statistic) and justify the procedure. [19.1]

16(b) Designate the preferred method of sampling at random without replacement as simple random sampling. [19.4]

16(c) Find the expected value and standard error for a sample percentage. [20.1]

16(d) State the square root law for the sample error for a percentage. [20.1]

16(e) Apply the normal approximation for chances of the sample percentage, being aware of the requirement to convert to a 0-1 box as the first step. [20.2]

16(f) **(Optional)** State and apply the correction factor, used when the sample is drawn with replacement. [20.3]

16(g) Use the bootstrap to estimate the unknown population percentage from the fractions of 0's and 1's in the sample and, from this, estimate the SD of the box. [21.1]

16(h) Find a confidence interval for a population percentage based on the percentage of a given sample, its size, and a specified confidence level. [21.2]

16(i) Explain what a confidence interval really means. Connect the confidence level with the results and avoid making a probability statement about this particular interval. [21.3]

16(j) Use the half-sample method to calculate a standard error in the Current Population Survey. [22.1 through 22.5]

16(k) State the EV and SE for the average of draws from a box. [23.1]

16(l) Apply the normal approximation to the probability histogram for the average of the draws. [23.1]

16(m) State and apply the square root law for the SE for the average of the draws from a box. [23.1]

16(n) Use the bootstrap to estimate the SD of the box from the SD of the sample and use that estimate to find confidence intervals for the population average. [23.2]

17. A model for measurement error

17(a) Explain and apply the Gauss model for measurement error. State the needed assumptions and discuss the methods for estimating the sample SD. [24.3]

18. Chance models in genetics

18(a) **(Optional)** Explain how Mendel discovered genes and give some examples of a statistical model that might apply. [25.1 through 25.4]

19. Tests of significance

19(a) Explain the null hypothesis (with its associated box model) and the alternative hypothesis. [26.2]

19(b) Discuss the roles of the test statistic and the observed significance level. [26.3]

19(c) For a given test of signficance, define a test statistic, calculate the observed significance level, and make a conclusion. [26.4]

19(d) Run a test of significance involving classifying and counting, with a zero-one box and the test statistic based on the sum of the draws. [26.5]

19(e) **(Optional)** Run student's t-test for a small sample. [26.6]

19(f) **(Optional)** State and apply the formula for the standard error for a difference. [27.1]

19(g) **(Optional)** Compare two sample averages. [27.2]

19(h) **(Optional)** Compare two sample percentages. [27.2]

19(i) **(Optional)** Apply the two-sample methods to certain randomized controlled experiments. [27.3 and 27.4]

19(j) Describe and run a Chi-square test. [28.1 and 28.2]

19(k) **(Optional)** Use a Chi-square test to test independence in an $m \times n$ table. [28.4]

19(l) Ask the question: Was the result significant? [29.1]

19(m) Recognize data snooping. [29.2]

19(n) Ask the question: Was the result important? [29.3]

19(o) Always consider the role of the model before making a test of significance. [29.4]

19(p) Ask the questions: Does the difference prove the point? Was this a good test of what we were trying to prove? [29.5]