Advanced Placement Statistics
Learning Targets
J. Fortney

Text: *Stats Modeling the World, 2nd edition.* Bock, Velleman, and De Veaux.

This document provides a list of key learning objectives for each of the major content units of the Advanced Placement Statistics course. Although the objectives are provided within individual teaching units, the material in the course is taught and assessed in a cumulative manner. Hence, learning objectives in any given unit will be assessed in future units as well.

Unit 1: Exploring Data (*Chapters 1-4)*

1. Recognize and intuitively understand that statistics is a way of reasoning whose goal is to help us understand the world. It is *not* a set of recipes or formulas.
2. Believe and daily apply the concept that data are not "just numbers." They *always* have a context (who, what, when, where, why, and how.)
3. Incorporate into all work (classwork, homework, quizzes, tests, and projects) the guiding principle that no statistical analysis is complete without a connection back to the real-world circumstances from which it arose.
4. Identify the cases and variables in any data set.
5. Classify a variable, in context, as categorical or quantitative, and if quantitative, identify the units by which the variable is being measured.
6. Construct and interpret contingency tables, conditional distribution tables, stem-and-leaf displays, bar charts, histograms, timeplots, and dot plots, and discuss the relative advantages and disadvantages of each as a descriptive tool.
7. Determine when, and in what context, conditional distributions and/or marginal distributions should be used.
8. Compare the distributions of a variable for two or more groups by comparing, in context, their shapes, centers, spreads, and any unusual features.
9. On the TI-84, use list operations fluently, and generate displays and summary statistics.
10. Describe anomalies or extraordinary features revealed by the display of a variable.
11. Understand how bias, measurement error, and display distortion can affect the interpretation of data.
12. Know how to determine when a data value is an *outlier* as opposed to merely the largest or smallest value in the data set.
13. Identify which pictorial display is most appropriate for any given quantitative variable.
14. Identify and fluently use the appropriate "summary measures" for both categorical and quantitative data.
15. Know and understand *Simpson's Paradox,* thus being able to demonstrate that when averages are taken across different groups, they can *appear* to contradict the overall averages.

Unit 2: Describing distributions numerically; Standard deviation and the Normal model. (*Chapters 5-6*)

1. Find and describe, in context, the five-number summary for a given quantitative data set.
2. Determine the mean and the standard deviation and describe such, in context, for a quantitative data set.
3. Select a suitable measure of spread and a suitable measure of center for a variable based on information about its distribution.

4. Construct a boxplot by hand from a five-number summary, and compare groups via the use of side-by-side boxplots.
5. Change the center and/or spread of a variable by adding (or subtracting) a constant or by multiplying (or dividing) by a constant, and recognize that information may be easier to understand after such a shifting or rescaling of the data.
6. Compare values of two different variables (or for two different groups being described by the same variable) by using their *z*-scores.
7. Use normal models and the 68-95-99.7 rule to estimate the percentage of observations falling within one, two, or three standard deviations of the mean.
8. Determine whether a variable satisfies the "nearly normal condition" by constructing and interpreting a normal probability plot or a histogram.
9. Explain why the median and the inter-quartile range *resist* the effects of outliers, while the mean and the standard deviation do not.
10. Understand the *power* of standardizing data and recognize that a *z*-score can identify unusual or surprising values among data.
11. Determine what assumptions/conditions are necessary to justify the use of a Normal model.
12. Demonstrate an understanding of the critical difference between a *distribution* and a *model.*
13. State the effects (if any) on summary statistics when data is shifted or rescaled.
14. Explain the relationship between the median and the mean in skewed distributions.
15. Recognize that the "answer" in statistics is *never* "just a number." Rather, the number must be accompanied by an interpretation in the context of the real-world data.
16. On the TI-84, confidently use the "catalog," and use "1-Var Stats" and "Statplot" to describe numerical data sets.
17. On the TI-84, understand how (and when) to use the commands *normalcdf, normalpdf,* and *invNorm.*

<u>Unit 3: Exploring Relationships Between Variables</u> (*Chapters 7-10)*

1. Construct a scatterplot by hand and via the TI-84.
2. Read and interpret a correlation display produced by a statistics software package.
3. Identify and describe points that deviate from an overall pattern of a bivariate data set.
4. Find a regression equation from the summary statistics for each variable and from the correlation between the variables.
5. Recognize when a well-chosen re-expression may help improve and simplify statistical analysis.
6. Re-express data with powers and find an effective re-expression with the TI-84.
7. Describe the direction, form, and strength of an association between two variables.
8. Understand why shifting, scaling, or standardizing variables has no effect on the numerical value of the correlation.
9. Check the "straight enough condition" before determining a choice of statistical methods, and know why this condition must be applied.
10. Know why a high correlation does *not* necessarily imply a cause-and-effect relationship between two variables being studied.
11. Know, by definition, what residuals are and what they tell us about the underlying data.
12. Explain, in context, the meaning of r-squared in a regression analysis.
13. Explain why outliers can play such havoc in the construction of the best-fit line.
14. Demonstrate an understanding of what the standard deviation of the residuals tells us about the amount of "scatter" around the best-fit line.

15. Explain how the existence of one or more lurking variables can confound the relationship between the two variables whose association is being studied.
16. Explain what is meant by "regression to the mean."
17. Demonstrate a clear understanding of the difference between the "appropriateness of a model" and the "strength of a model."
18. State the assumptions and conditions that are required in order for correlation to be an appropriate descriptive measure.
19. On the TI-84, confidently use and create scatterplots, work with LinReg, LnReg, ExpReg, PwrReg, and the RESID list.

Unit 4: Gathering Data and Conducting Experiments (*Chapters 11-13)*

1. Perform simulations using dice, cards, m&ms, a spinner, or some such other non-technology-based method.
2. Perform simulations with random number tables as well as with a calculator via the use of the randInt command.
3. Discuss the results of a simulation study and draw conclusions about the question being investigated.
4. Recognize when a simulation might usefully model random behavior in a real-world situation.
5. Draw a simple random sample from a master list of a population.
6. Identify the population, the parameter of interest, the sampling frame, the sample, the sampling method, and any potential sources of bias in a statistical study.
7. Identify the factors, the treatments, and the response variable in a description of a designed experiment.
8. Design an experiment in which blocking is used to reduce variation.
9. Know the four basic principles of sound experimental design…control, randomization, replication, and blocking.
10. Describe and understand the distinctions between treatments, levels, and factors.
11. Describe the underlying importance of randomization as a defense against bias.
12. Know and appropriately use the basic concepts and terminology of sampling.
13. State and elaborate upon the differences between and among sampling, observational studies, and experimentation.
14. Explain the difference between *bias* and *error.*
15. Explain the difference between a *statistic* and a *parameter.*
16. Describe the differences between *confounding* and *lurking* variables.
17. Describe the differences between cluster sampling, multi-stage sampling, and systematic sampling, and provide situations for which each of these sampling methods would be most appropriate.
18. Discuss the dangers inherent in any form of convenience sampling or voluntary response sampling.
19. Recognize the difference between a retrospective and a prospective study, and why each must be used in various situations.
20. Understand and discuss the ethics of experimentation and the need for approval panels in many experiments and studies.

Unit 5: Probability Rules! (*Chapters 14-15)*

1. Know and correctly use the basic terminology for probability: trial, outcome, event, sample space, sampling frame, disjoint events, independent events, etc.
2. Understand and apply the addition rule, the multiplication rule, and the complement rule, as well as the rules for conditional probability.
3. Describe how the *general addition rule* is different from the *addition rule for disjoint events.*
4. Describe how the *general multiplication rule* is different from the *multiplication rule for disjoint events.*
5. Use two-way tables to organize the information needed to determine conditional probabilities.
6. Describe, via definition and example, what is meant by "statistically significant results."
7. Define, using technically correct probability terms, what is meant by a *random phenomena.*
8. Understand, at an intuitive level, that short-run anomalies get "drowned out" in the long run. (The basis for the "Law of Large Numbers.")
9. Use the basic facts about probability to determine whether an assignment of probability is legitimate.
10. Construct and use well-labeled tree diagrams and well-labeled Venn diagrams, and understand that each are helpful in understanding conditional probability as well as reverse conditioning.
11. Recognize random outcomes in a real-world situation.
12. Build all "probability learning" around the understanding that random phenomena are unpredictable in the short run but show long-run regularity.
13. Recognize that the relative frequency of an outcome "settles down" as we gather more random outcomes.
14. Recognize when events are disjoint and when they are independent.
15. Determine the sample space for any given event.
16. Recognize and understand the fundamental notion that disjoint events *cannot* be independent.
17. Understand and apply the concepts of independent and dependent events and know how they are related to compound events and conditional probability.

<u>Unit 6: Random Variables and Probability Models</u> (*Chapters 16-17*)

1. Understand and daily apply a sound knowledge of the distinction between *assumptions* and *conditions.*
2. Understand and daily apply a sound knowledge of the difference between a *discrete* and a *continuous* random variable.
3. Know the appropriate and required assumptions and conditions for using a Geometric, Binomial, or Normal model.
4. Know, *by definition*, how Binomial and Geometric models are different.
5. Know that Bernoulli trials are characterized by two outcomes (a fixed probability of success and independence from trial to trial) and thus be able to determine whether a given setting involves Bernoulli trials.
6. Interpret means, standard deviations, and probabilities in the Bernoulli trial context.
7. Understand that random variables *must* be independent in order to determine the variability of their sum or difference by adding variances.
8. Interpret the meaning of the expected value and standard deviation of a random variable in the proper context.
9. Use the TI-84 to find geometric and binomial probabilities via use of the commands *binompdf, binomcdf, geometpdf, and geometcdf.*
10. Calculate binomial probabilities via the Normal Approximation process *(if, of course, the necessary conditions for doing so are met.)*.

11. Recognize that a "probability model" consists of the set of possible outcomes of a random variable and the probability of each.
12. Know that the *expected value* of a random variable is the mean of its probability model.
13. Find the probability model, the expected value, and the variance for a discrete random variable.
14. Recognize, on both an intuitive and practical level, that the *expected value* of a random variable may very well be impossible on any given trial.
15. Believe and intuitively grasp that probability models are still *just models;* they may be useful, but they are *not* reality.
16. Understand that normality of an underlying population can't be *assumed;* thus recognize that using a normal model when it really does not apply will lead to potential wrong answers and misleading conclusions.

<u>Unit 7: Sampling Distribution Models and Confidence Intervals for Proportions</u> (*Chapters 18-19)*

1. Demonstrate understanding of a sampling distribution by simulation, and thus show that sampling variability is both understandable and predictable.
2. Understand that *sampling error* is <u>always</u> present, and that the goal is to distinguish random variability from results that are truly unusual.
3. State the *Central Limit Theorem* and describe, with detail, what makes it so powerful.
4. State the assumptions and conditions that support the use of the Normal model for the sampling distribution of sampling proportions.
5. Explain the imperative difference between the terms "sampling distribution" and "distribution of the sample."
6. Use a sampling distribution model to make statements about the distribution of a proportion or mean under repeated sampling.
7. Interpret a sampling distribution model as describing the values taken by a statistic in all possible realizations of a sample or randomized experiment under the same conditions.
8. Construct and interpret a one-proportion z-interval.
9. Determine when the margin of error is too large to be useful.
10. Understand that the variability of a statistic depends on the size of the sample with statistics based on larger samples being less variable.
11. Find the minimum sample size required to meet given margin of error and confidence level specifications.
12. Understand, both practically and intuitively, that a higher level of confidence results in a wider confidence interval, and that a larger sample size will create a narrower confidence interval.
13. Define and apply the *plausible independence* condition, the *random sampling* condition, the *10%* condition, and the *success/failure* condition.
14. Know how to examine data for violations of conditions that would hence make inference about a population proportion unwise or invalid.
15. Confidently use the 1-PropZint command on the TI-84.

<u>Unit 8: Hypothesis Testing for Proportions</u> (*Chapters 20-22*)

1. Understand that the four-step process for hypothesis testing (hypothesis, model, mechanics, and conclusion) is critical to a sound implementation of inferential statistics.
2. Define the "null and alternative hypotheses" and describe how they are used in inferential statistics.

3. Understand that the *p*-value is the *conditional probability* that the observed results could have happened if the null hypothesis is true.
4. Describe the difference between *statistical significance* and *practical significance* within the context of hypothesis testing conclusions.
5. Know that if the sample data are out of line with the null hypothesis model, the *p*-value will be small, while if the sample data are consitent with the null hypothesis model, the *p*-value will be large.
6. Describe how the critical value for a test is related to the specified alpha level.
7. Carry out a complete four-step hypothesis testing procedure for a one-proportion *z*-test.
8. Carry out a complete multi-step hypothesis testing procedure for a two-proportion *z*-test.
9. Interpret the *p*-value for a given test in the real-world context of the problem.
10. Choose between a one-sided and two-sided hypothesis test and explain your reasoning for doing so.
11. Explain, in both technical and non-technical terms, the difference between a Type I error and a Type II error for a hypothesis test.
12. Explain how the choice of alpha level potentially offers protection against one type of error while possibly increasing the risk of the other type of error.
13. Explain how Type I and Type II errors relate to both sample size and power.
14. Construct a two-proportion *z*-interval.
15. Define "statistical pooling" and describe how it is used when comparing two proportions.
16. Define, in both technical and non-technical terms, what is meant by the *power* of a test.
17. Demonstrate how a confidence interval can be used to *expand* upon a conclusion following a hypothesis test.
18. Confidently use the 2-PropZInt and 2-PropZTest commands on the TI-84.

Unit 9: Inference for Means; Paired Samples and Blocks (*Chapters 23-25)*

1. Distinguish between groups of data that are *matched* and those that are *independent.*
2. Recognize whether a design that compares two groups is "paired" or not.
3. Identify the assumptions and conditions that are necessary for using the single-sample *t*-models.
4. State the *independence assumption* and explain its importance.
5. State the *nearly normal condition* and explain when it can be used.
6. Identify the assumptions and conditions necessary for constructing a two-sample *t*-interval or performing a two-sample *t*-test.
7. Identify the assumptions and conditions necessary for the "paired *t*-test" and the "paired *t*-interval."
8. Understand that the *t*-models form a whole *family* of related distributions and explain how a given curve depends on the *degrees of freedom*.
9. Explain why the null hypothesis for a test of paired data is a statement about the *mean of the differences.*
10. Define what is meant by a *pooled t-test* and describe the conditions under which it is used.
11. Construct a one-sample *t*-interval and provide an interpretation in the context of the problem.
12. Carry out a complete four-step hypothesis testing procedure for a one-sample *t*-test for the mean.
13. Examine outliers carefully to determine if they can be "set aside" so as to meet the *nearly normal* condition.
14. Construct a two-sample *t*-interval and provide an interpretation in the context of the problem.
15. Carry out a multi-step two-sample *t*-test for the difference between two means.
16. Confidently use the following commands on the TI-84: invT, tcdf, tpdf, Tinterval, T-Test, 2-SampTTest, 2-SampTInt, T-Test, and TInterval.

Unit 10: The Chi-Square Tests and Inferences for Regression (*Chapters 26, 27*)

1. Use hypothesis testing to determine whether an *observed distribution* is “statistically significantly different” from what we expected.
2. Know that the Chi-Square Test of Independence tests the significance of associations between categorical variables.
3. Recognize and intuitively understand that the Chi-Square Models form a whole family of related distributions that depend on the number of degrees of freedom.
4. Discuss the key graphical differences between Normal models, *t*-models, and Chi-Square models.
5. Recognize that a failure of independence between two categorical variables does not demonstrate a cause-and-effect relationship between the variables.
6. Understand and apply a sound knowledge of the fact that the null hypothesis for a Goodness-of-Fit test states that the given probability model is correct.
7. Understand and apply a sound knowledge of the fact that the null hypothesis for a test of homogeneity states that the two groups are drawn from populations in which the variable has the same model.
8. Understand and apply a sound knowledge of the fact that the null hypothesis for a test of independence states that there is no association between the variables in question.
9. Recognize and implement an understanding that the degrees of freedom for a Chi-Square test depend on the dimensions of the table and not on the sample size.
10. Determine the appropriate number of degrees of freedom for any table.
11. Determine the expected value for any cell of a table.
12. Calculate the *component* of χ^2 for any cell.
13. Recognize when a test of Goodness-of-Fit, a test of homogeneity, or a test of independence would be appropriate for a table of counts.
14. Display and interpret, *in the context of the given problem,* the counts in a two-way table.
15. State the assumptions and conditions that are required before using regression inference.
16. State the “equal variance” assumption and describe the steps by which it is verified.
17. Understand that under certain assumptions, the sampling distribution for the slope of a regression line can be described by a Student’s *t*-model with *n*-2 degrees of freedom.
18. Describe *why* the sampling distributions of slopes of regression lines are described by the *t*-models.
19. Know how to test the standard hypothesis that the true regression slope is zero.
20. Interpret a confidence interval for the slope of a regression line.
21. Recognize why there is strong evidence of an association between variables if the slope of a regression line is significantly different from zero.
22. Use a residual plot to confirm the appropriateness of a linear model.
23. Perform a complete regression slope *t*-test.
24. Find relevant regression numbers in a standard computer regression output.
25. Confidently use all commands on the TI-84 relevant to Chi-Square testing and Linear Regression *t*-tests.

Unit #11: End-of-Course Project

Use all of your learning from the entire academic year to design and implement a statistical study or experiment; this must include a 20-25 minute presentation to the class along with a substantial written report.

Additional Learning Targets

Statistical History: *(Who are we indebted to?)* Develop a sound appreciation and understanding of the contributions of the many men and women who contributed to the development of Statistics as it is practiced today. *The learning targets that follow are woven into multiple units throughout the course, and are assessed as such.*

1. *John Tukey:* One of the greatest statisticians of the 20th century, he devised the Stem-and-Leaf display, as well as "Tukey's Quick Test" for determining whether or not to reject the null hypothesis in a two-sample *t*-test.

2. *Sir Ronald Fisher:* A great statistician of the 20th century; famously testified in court that a causal relationship just might underlie the correlation of smoking and cancer.

3. *George Gallup:* American 20th century statistician who founded The Gallup Organization.

4. *Sir Frances Galton:* Nineteenth-century Englishman who was the first to speak of "regression" to the mean, and to apply statistical methods to the study of human differences.

5. *Karl Gauss*: German mathematician and scientist who was influential in the development of the Method of Least Squares, a procedure used to minimize the impact of measurement error.

6. *Legendre*: The first mathematician to officially publish the "least squares" solution to the problem of fitting a line to data.

7. *Pierre-Simon LaPlace*: Proved the Central Limit Theorem (upon which most of statistics rests) in 1810.

8. *William of Occam:* The 14th century English philosopher and Theologian to whom "Occam's Razor" is attributed.

9. *John Venn*: Creator of the Venn Diagram and author of the influential book on probability, *The Logic of Chance.*

10. *Abraham DeMoivre:* Made significant contributions to probability theory.

11. *Gustav Fechner:* Nineteenth-century founder of experimental psychology; he did a great deal to advance the study of experimental design.

12. *Daniel and Jacob Bernoulli*: Their work with mathematics and with games of chance resulted in what we know as Bernoulli trials and Bernoulli numbers.

13. *George Box:* Famous 20th century British statistician who coined the well-known phrase "*All models are wrong—but some are useful.*"

14. *William Gossett*: Irish statistician best known by his pen name "Student" and for his work on Student's *t*-distributions.
