Learning Objectives

Advanced High School Statistics

Second Edition

David Diez
Data Scientist
OpenIntro

Mine Çetinkaya-Rundel
Associate Professor of the Practice, Duke University
Professional Educator, RStudio

Leah Dorazio
Statistics and Computer Science Teacher
San Francisco University High School

Christopher D Barr
Investment Analyst
Varadero Capital

June 4, 2019

Contents

Chapter 1

Data collection

1.1 Case study: using stents to prevent strokes

1. Understand the four steps of a statistical investigation (identify a question, collect data, analyze data, form a conclusion) in the context of a real-world example.

2. Consider the concept of statistical significance.

1.2 Data basics

1. Identify the individuals and the variables of a study.

2. Identify variables as categorical or numerical. Identify numerical variables as discrete or continuous.

3. Understand what it means for two variables to be associated.

1.3 Overview of data collection principles

1. Distinguish between the population and a sample and between the parameter and a statistic.

2. Know when to summarize a data set using a mean versus a proportion.

3. Understand why anecdotal evidence is unreliable.

4. Identify the four main types of data collection: census, sample survey, experiment, and observation study.

5. Classify a study as observational or experimental, and determine when a study's results can be generalized to the population and when a causal relationship can be drawn.

1.4 Observational studies and sampling strategies

1. Identify possible confounding factors in a study and explain, in context, how they could confound.
2. Distinguish among and describe a convenience sample, a volunteer sample, and a random sample.
3. Identify and describe the effects of different types of bias in sample surveys, including selection bias, non-response, and response bias.
4. Identify and describe how to implement different random sampling methods, including simple, stratified, and cluster.
5. Recognize the benefits and drawbacks of choosing one sampling method over another.
6. Understand when it valid to draw an inference and to what population that inference can be drawn.

1.5 Experiments

1. Identify the subjects/experimental units, treatments, and response variable in an experiment.
2. Identify the three main principles of experiment design and explain their purpose: direct control, randomization, and replication.
3. Explain placebo effect and describe when and how to implement a single-blind and a double-blind experiment.
4. Identify and describe how to implement the following three experimental designs: completely randomized design, blocked design, and matched pairs design.
5. Explain the purpose of random assignment or randomization in each of the three experimental designs.
6. Explain how to randomize treatments in a completely randomized design using technology or a table of random digits (make sure this is explained).
7. Explain when it is reasonable to draw a causal conclusion about the effect of a treatment.
8. Identify the number of factors in experiment, the number of levels for each factor and the total number of treatments.

Chapter 2

Summarizing data

2.1 Examining numerical data

1. Use scatterplots to see the relationship between two numerical variables. Describe the direction, form, and strength of the relationship, as well as any unusual observations.

2. Understand what the term distribution means and how to summarize it in a table or a graph.

3. Create stem-and-leaf plots, dot plots, and histograms to visualize the distribution of a numerical variable. Be able to read off specific information and summary information from these graphs.

4. Identify the shape of a distribution as approximately symmetric, right skewed, or left skewed. Also, identify whether a distribution is unimodal, bimodal, multimodal, or uniform.

5. Read and interpret a cumulative frequency or cumulative relative frequency histogram.

2.2 Numerical summaries and box plots

1. Calculate, interpret, and compare the two measures of center (mean and median) and the three measures of spread (standard deviation, interquartile range, and range).

2. Understand how the shape of a distribution affects the relationship between the mean and the median.

3. Identify and apply the two rules of thumb for identify outliers (one involving standard deviation and mean and the other involving Q_1 and Q_3).

4. Describe the distribution a numerical variable with respect to center, spread, and shape, noting the presence of outliers.

5. Find the 5 number summary and IQR, and draw a box plot with outliers shown.

6. Understand the effect changing units has on each of the summary quantities.

7. Use the empirical rule to summarize approximately symmetric data sets.

8. Use quartiles, percentiles, and Z-scores to measure the relative position of a data point within the data set.

9. Compare the distribution of a numerical variable using dot plots / histograms with the same scale, back-to-back stem-and-leaf plots, or parallel box plots. Compare the distributions with respect to center, spread, shape, and outliers.

2.3 Considering categorical data

1. Use a one-way table or a bar graph to summarize a categorical variable. Use counts (frequency) or proportions (relative frequency).

2. Compare distributions of a categorical variable using a two-way table or a side-by-side or segmented bar chart.

3. Calculate marginal and joint frequencies for two-way tables.

2.4 Case study: gender discrimination (special topic)

1. Recognize that an observed difference in sample statistics may be due to random chance and that we use hypothesis testing to determine if this difference statistically significant (i.e. too large to be attributed to random chance).

2. Set up competing hypotheses and use the results of a simulation to evaluate the degree of support the data provide against the null hypothesis and for the alternative hypothesis.

Chapter 3

Probability

3.1 Defining probability

1. Describe the long-run relative frequency interpretation of probability and understand its relationship to the "Law of Large Numbers".

2. Use Venn diagrams to represent events and their probabilities and to visualize complement, union, and intersection of events.

3. Use the General Addition Rule to find the probability that at least one of several events occurs.

4. Understand when events are disjoint (mutually exclusive) and how that simplifies the General Addition Rule.

5. Apply the Multiplication Rule for finding the joint probability of independent events.

3.2 Conditional probability

1. Understand conditional probability and how to calculate it.

2. Calculate joint and conditional probabilities based on a two-way table.

3. Use the General Multiplication Rule to find the probability of joint events.

4. Determine whether two events are independent and whether they are mutually exclusive, based on the definitions of those terms.

5. Draw a tree diagram with at least two branches to organize possible outcomes and their probabilities. Understand that the second branch represents conditional probabilities.

6. Use the tree diagram or Bayes' Theorem to solve "inverted" conditional probabilities.

3.3 The binomial formula

1. Calculate the number of possible scenarios for obtaining k successes in n trials.
2. Determine whether a scenario is binomial or not.
3. Calculate the probability of obtaining exactly k successes in n independent trials.
4. Recognize that the binomial formula uses the special Addition Rule for mutually exclusive events.
5. Find probabilities of the form "at least" or "at most" by applying the binomial formula multiple times.

3.4 Simulations

1. Understand the purpose of a simulation and recognize the application of the long-run relative frequency interpretation of probability.
2. Understand how random digit tables work and how to assign digits to outcomes.
3. Be able to repeat a simulation a set number of trials or until a condition is true, and use the results to estimate the probability of interest.

3.5 Random variables

1. Define a probability distribution and what makes a distribution a valid probability distribution.
2. Summarize a discrete probability distribution graphically using a histogram and verbally with respect to center, spread, and shape.
3. Calculate and interpret the mean (expected value) and standard deviation of a random variable.
4. Calculate the mean and standard deviation of a transformed random variable.
5. Calculate the mean of the sum or difference of random variables.
6. Calculate the standard deviation of the sum or difference of random variables when those variables are independent.

3.6 Continuous distributions

1. Understand the difference between a discrete random variable and a continuous random variable.
2. Recognize that when working with continuous probability distributions area represents probability and the total area under the curve must equal 1.

Chapter 4

Distribution of random variables

4.1 Normal distribution

1. Calculate and interpret a Z-score.

2. Understand that Z-scores are unitless (standard units) and are not affected by change of units.

3. Use the normal model to approximate a distribution where appropriate.

4. Find probabilities and percentiles using the normal approximation.

5. Find the value that corresponds to a given percentile when the distribution is approximately normal.

4.2 Sampling distribution of a sample mean

1. Understand the concept of a sampling distribution.

2. Describe the center, spread, and shape of the sampling distribution of a sample mean.

3. Distinguish between the standard deviation of a population and the standard deviation of a sampling distribution.

4. Explain the content and importance of the Central Limit Theorem.

5. Identify and explain the conditions for using normal approximation involving a sample mean.

6. Verify that the conditions for normal approximation are met and carry out normal approximation involving a sample mean or sample sum.

4.3 Geometric distribution

1. Determine if a scenario is geometric.
2. Calculate the probabilities of the possible values of a geometric random variable.
3. Find and interpret the mean (expected value) of a geometric distribution.
4. Understand the shape of the geometric distribution.

4.4 Binomial distribution

1. Determine if a scenario is binomial.
2. Calculate the probabilities of the possible values of a binomial random variable.
3. Calculate and interpret the mean (expected value) and standard deviation of the number of successes in n binomial trials.
4. Determine whether a binomial distribution can be modeled as approximately normal. If so, use normal approximation to estimate cumulative binomial probabilities.

4.5 Sampling distribution of a sample proportion

The binomial distribution shows the distribution of the number of successes in n trials. Often, we are interested in the *proportion* of successes rather than the *number* of successes.

1. Describe the center, spread, and shape of the sampling distribution of a sample proportion.
2. Recognize the relationship between the distribution of a sample proportion and the corresponding binomial distribution.
3. Identify and explain the conditions for using normal approximation involving a sample proportion. Recognize that the Central Limit Theorem applies in the case of proportions/counts as well as means/sums.
4. Verify that the conditions for normal approximation are met and carry out normal approximation involving a sample proportion or sample count.

Chapter 5

Foundations for inference

5.1 Estimating unknown parameters

1. Explain the difference between probability and inference and identify when to use which one.

2. Understand the purpose and use of a point estimate.

3. Understand how to measure the variability/error in a point estimate.

4. Recognize the relationship between the standard error of a point estimate and the standard deviation of a sample statistic.

5. Understand how changing the sample size affects the variability/error in a point estimate.

5.2 Confidence intervals

1. Explain the purpose and use of confidence intervals.

2. Construct 95% confidence intervals assuming the point estimate follows a normal distribution.

3. Calculate the critical value for a C% confidence interval when the point estimate follows a normal distribution.

4. Describe how sample size and confidence level affect the width of a confidence interval.

5. Interpret a confidence interval and the confidence level in context.

6. Draw conclusions with a specified confidence level about the values of unknown parameters.

7. Calculate and interpret the margin of error for a C% confidence interval. Distinguish between margin of error and standard error.

5.3 Introducing hypothesis testing

1. Explain the logic of hypothesis testing, including setting up hypotheses and drawing a conclusion based on the set significance level and the calculated p-value.

2. Set up the null and alternative hypothesis in words and in terms of population parameters.

3. Interpret a p-value in context and recognize how the calculation of the p-value depends upon the direction of the alternative hypothesis.

4. Define and interpret the concept statistically significant.

5. Interpret Type I, Type II error, and power in the context of hypothesis testing.

5.4 Does it make sense?

1. Understand the two general conditions for when the confidence interval and hypothesis testing procedures apply. Explain why these conditions are necessary.

2. Distinguish between statistically significant and practically significant. What role does sample size play here?

3. Recognize that not all statistically significant results correspond to real differences, due to Type I errors. What role does the significance level α play here?

Chapter 6

Inference for categorical data

6.1 Inference for a single proportion

1. State and verify whether or not the conditions for inference on a proportion using a normal distribution are met.

2. Recognize that the success-failure condition and the standard error calculation are different for the test and for the confidence interval and explain why this is the case.

3. Carry out a complete hypothesis test and confidence interval procedure for a single proportion.

4. Find the minimum sample size needed to estimate a proportion with C% confidence and a margin of error no greater than a certain value.

5. Recognize that margin of error calculations only measure sampling error, and that other types of errors may be present.

6.2 Difference of two proportions

1. State and verify whether or not the conditions for inference on the difference of two proportions using a normal distribution are met.

2. Recognize that the standard error calculation is different for the test and for the interval and explain why that is the case.

3. Know how to calculate the pooled proportion and when to use it.

4. Carry out a complete confidence interval procedure for the difference of two proportions.

5. Carry out a complete hypothesis test for the difference of two proportions.

6.3 Testing for goodness of fit using chi-square

1. Calculate the expected counts and degrees of freedom for a one-way table.

2. Calculate and interpret the test statistic χ^2.

3. State and verify whether or not the conditions for the chi-square goodness of fit are met.

4. Carry out a complete hypothesis test to evaluate if the distribution of a categorical variable follows a hypothesized distribution.

5. Understand how the degrees of freedom affect the shape of the chi-square curve.

6.4 Homogeneity and independence in two-way tables

1. Calculate the expected counts and degrees of freedom for a chi-square test involving a two-way table.

2. State and verify whether or not the conditions for a chi-square test for a two-way table are met.

3. Explain the difference between the chi-square test of homogeneity and chi-square test of independence.

4. Carry out a complete hypothesis test for homogeneity and for independence.

Chapter 7

Inference for numerical data

7.1 Inference for a single mean with the t-distribution

1. Understand the relationship between a t-distribution and a normal distribution, and explain why we use a t-distribution for inference on a mean.
2. State and verify whether or not the conditions for inference for a mean based on the t-distribution are met. Understand when it is necessary to look at the distribution of the sample data.
3. Know the degrees of freedom associated with a one sample t-procedure.
4. Carry out a complete hypothesis test for a single mean.
5. Carry out a complete confidence interval procedure for a single mean.
6. Find the minimum sample size needed to estimate a mean with C% confidence and a margin of error no greater than a certain value.

7.2 Inference for paired data

1. Distinguish between paired and unpaired data.
2. Recognize that inference procedures for paired data use the same one-sample t-procedures as in the previous section, and that these procedures are applied to the *differences* of the paired observations.
3. Carry out a complete hypothesis test for paired differences.
4. Carry out a complete confidence interval procedure for paired differences.

7.3 Difference of two means using the t-distribution

1. Determine when it is appropriate to use a paired t-procedure versus a two-sample t-procedure.

2. State and verify whether or not the conditions for inference on the difference of two means using the t-distribution are met.

3. Be able to use a calculator or other software to find the degrees of freedom associated with a two-sample t-procedure.

4. Carry out a complete confidence interval procedure for the difference of two means.

5. Carry out a complete hypothesis test for the difference of two means.

Chapter 8

Introduction to linear regression

8.1 Line fitting, residuals, and correlation

1. Distinguish between the data point y and the predicted value $\hat{y}$ based on a model.
2. Calculate a residual and draw a residual plot.
3. Interpret the standard deviation of the residuals.
4. Interpret the correlation coefficient and estimate it from a scatterplot.
5. Know and apply the properties of the correlation coefficient.

8.2 Fitting a line by least squares regression

1. Calculate the slope and y-intercept of the least squares regression line using the relevant summary statistics. Interpret these quantities in context.
2. Understand why the least squares regression line is called the least squares regression line.
3. Interpret the explained variance R^2.
4. Understand the concept of extrapolation and why it is dangerous.
5. Identify outliers and influential points in a scatterplot.

8.3 Inference for the slope of a regression line

1. Recognize that the slope of the sample regression line is a point estimate and has an associated standard error.
2. Be able to read the results of computer regression output and identify the quantities needed for inference for the slope of the regression line, specifically the slope of the sample regression line, the SE of the slope, and the degrees of freedom.

3. State and verify whether or not the conditions are met for inference on the slope of the regression line based using the t-distribution.

4. Carry out a complete confidence interval procedure for the slope of the regression line.

5. Carry out a complete hypothesis test for the slope of the regression line.

6. Distinguish between when to use the linear regression t-test and when to use the matched pairs t-test.

8.4 Transformations for nonlinear data

1. See how a log transformation can bring symmetry to an extremely skewed variable.

2. Recognize that data can often be transformed to produce a linear relationship, and that this transformation often involves log of the y-values and sometimes log of the x-values.

3. Use residual plots to assess whether a linear model for transformed data is reasonable.