Lab 1: Summary from Chapter 1

- **Random sampling** means
    a. All individuals in the population have an equal chance of being included,
    b. The sampling events are independent draws from the population
- Most estimation methods and hypothesis tests assume that the data are a random sample from the population.
- **Biased sampling** means some individuals are more likely than others to be selected.
- **Estimates** are calculated from data. The goal of an estimator is to give values close to the true **parameter value**.
- Estimates are biased if there is a systematic tendency for them to be too high or too low.
- Estimates are **imprecise** if they vary a lot when based on different samples.
- We can detect measurement error by repeatedly measuring the same individual.
- **Measurement bias** is a systematic difference between the true value and the measured value of each individual.

Lab 2: Summary from Chapter 1-3

- Variables in a data set can be **categorical** (nominal or ordinal) or **numerical** (continuous or discrete).
- Frequency distributions show the number or proportion of occurrences of each category in a data set. Use **bar graphs** (for categorical variables) and **histograms**, **cumulative frequency distributions**, and **box plots** (for numerical variables).
- Descriptive statistics (summary statistics) summarize information in the data. We use these **statistics** to get **estimates** of population parameters.
- The **mean** and **median** are the most commonly used "location statistics"
- The **standard deviation** (*s*), **variance** (s^2), and **interquartile range** (*IQR*) describe the spread of the data.
- The "empirical rule": If the data are unimodal and symmetric (roughly normally distributed), 68% of the data are within $\overline{Y}$ ± *s*, and 95% of the data are within $\overline{Y}$ ± 2*s*.
- The median and IQR are often used to summarize skewed data sets.
- For categorical variables, the sample proportion ($\widehat{p}$) is the most common summary.

Lab 3: Summary from Chapter 4

- From a random sample we can calculate a confidence interval – a range of plausible values that are likely to bracket the value of the population parameter.
- The sampling distribution of an estimate: if we were to repeatedly draw samples from a population, calculate an estimate for each sample, and examine the relative frequency distribution of these estimates (for example, as a histogram), then we would be examining the sampling distribution of that estimate.
- The standard error of an estimator is the standard deviation of the sampling distribution; we can estimate the standard error from a single sample using the formula: $SE_{\overline{Y}} = \frac{s}{\sqrt{n}}$ where *s* is the sample standard deviation, and *n* is the sample size.

- The interval $\overline{Y} - 2SE$, $\overline{Y} + 2SE$ is an approximation for the 95% confidence interval for the mean.

Lab 5: Summary from Chapter 6

- Research questions can be converted into hypotheses. Hypotheses focus on population parameters, but we test hypotheses with sample data. The null hypothesis is often a specific statement of "no difference".
- We obtain a test statistic from our sample data and compare it to a null distribution (the sampling distribution of the test statistic if the null hypothesis was true). The *P*-value describes the compatibility of the data with the null hypothesis. If the *P*-value is very small, it suggests that obtaining this particular sample is not very likely given the null hypothesis.
- The *P*-value is compared to an alpha value, which is a threshold value that is called the "significance level" of the test. Alpha values are chosen prior to the start of the study and are often set to be equal to 0.05. If the P-value is less than or equal to alpha, the null hypothesis is rejected. If the P-value is greater than alpha, the null hypothesis is not rejected.
- After doing a statistical test, it is important to return to the original question and make sure one can relate the statistical results to the scientific question. A scientific report about a statistical analysis should include descriptive statistics, the sample size, the test statistic, and the *P*-value.
- Because it is impossible to study the entire population, errors can be made in statistical analyses. A Type I error is the rejection of a correct null hypothesis while a Type II error is the failing to reject an incorrect null hypothesis.

Lab 6: Summary from Chapter 7

- The binomial distribution is useful for studying variables that have two possible outcomes; these outcomes are arbitrarily defined as a "success" or "failure". Specifically, the distribution describes the probability of getting different numbers of "successes" in a fixed number of independent trials (*n*). The probability of success (*p*) in each trial does not change from trial to trial.
- The binomial formula calculates the probability of obtaining X successes and is defined as: $Pr[X] = \binom{n}{X} p^X (1-p)^{n-X}$ where $\binom{n}{X} = \frac{n!}{X!(n-X)!}$.
- The binomial test uses the binomial distribution for hypothesis-testing. The null hypothesis of the binomial test states that the population proportion (*p*) matches a particular value (called the null expectation, p_0). We test this hypothesis by collecting a random sample from the population and calculating a test statistic and *P*-value. Essentially, we are exploring whether the sample data we obtain are consistent with this hypothesized population proportion (p_0); if so, we should not reject the null hypothesis. However, if there is a very low probability of obtaining our sample data from a population with a population proportion of p_0, then we should reject the null hypothesis.
- For the binomial test, the test statistic is simply the observed number of successes in the sample. The test statistic is compared to a null distribution, which is the sampling

distribution of the statistic assuming that the null hypothesis is true. For the binomial test, the null distribution is the binomial distribution. Using the binomial distribution, one can calculate exact *P*-values.

- The binomial test can be applied to any variable with only two outcomes. However, in practice, the binomial test is often used for relatively small sample sizes, with other alternatives used when *n* is large.

Lab 7: Summary from Chapter 8

- The chi-square test compares an observed frequency distribution (for example the number of individuals in different categories) to the frequency distribution that is "expected" under the null hypothesis. The null hypothesis of a chi-square goodness-of-fit test specifies population probabilities or proportions that are used to calculate the expected numbers.
- The test statistic is defined as $\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$ where *O* refers to an observed number and *E* is an expected number for a particular category. The summation is done over all the different categories in the problem. (Note that the use of "category" in this quick summary is meant to be broad: chi-square tests can be used with both true categorical variables as well as discrete numerical variables.)
- The null distribution for the test statistic is the χ^2 distribution with the appropriate degrees of freedom, where *df* = (number of categories) – 1 – (the number of parameters estimated from the data).
- The *P*-value is, as always, the probability of getting the observed test statistic or an even more extreme value, given the null hypothesis is true. This probability is defined as the area under the null distribution to the right of the observed chi- square test statistic.
- There are 3 key assumptions for the chi-square test: (1) that a random sample of the population was chosen, (2) that the expected counts in any category should not be < 1, and (3) that 20% of the categories should not be < 5.

Lab 8: Summary from Chapter 9

- Contingency tables are used to examine patterns of association between two categorical variables. The chi-square test contingency test (also called chi- square test of association, test of independence, or contingency table test) states in the null hypothesis that the two variables are independent of each other. The alternative hypothesis is that the variables are not independent of each other, or in other words, that there is some association between them.
- The test statistic is defined as $\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$ where *O* refers to an observed number and *E* is an expected number for a particular cell in the table. The summation is done over all the different cells in the contingency table. The expected numbers are calculated under the null hypothesis of independence. To calculate the expected number for a cell

in row *i* and column *j* of a contingency table, one multiplies the row *i* total and the column *j* total and divides this product by the grand total (= total number of observations in the data set).

- The null distribution for the test statistic is the χ^2 distribution with the appropriate degrees of freedom, where *df* = (*r* – 1)(*k* – 1) and *r* = number of rows, *k* = number of columns. We determine the P-value using the same approach as in the goodness-of-fit test. Similarly, the assumptions for the test are the same as those of the goodness-of-fit test.
- The estimate of the odds ratio is defined by the following equation: $\widehat{OR} = \frac{\widehat{O}_1}{\widehat{O}_2}$. In this equation, both the numerator and denominator are odds, defined as $\widehat{O} = \frac{\widehat{p}}{1-\widehat{p}}$ (where $\widehat{p}$ is the sample proportion for "successes"). If the odds ratio = 1, the condition or event is equally likely to occur in both groups 1 and 2. If the odds ratio is > 1, then the condition or event is more likely to occur in the first group. If the odds ratio is < 1, then the condition or event is less likely to occur in the first group.
- The odds ratio provides information about the magnitude of an effect. Recall that the *P*-value of the test tells how likely it is to get the observed data given that the null hypothesis is true, but it doesn't provide information on how deviant the data are from the null hypothesis. For example, it is plausible that a contingency test analysis in a very large study (large *n*) could result in a very small *P*-value but that the odds ratio might be close to 1. Thus there is a significant association, but the magnitude of the effect may not be large.
- In estimating a confidence interval for the odds ratio, we use $SE[ln(\widehat{OR})] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$, where *a, b, c,* and *d* are the observed frequencies in the cells of the 2×2 contingency table.
- The 95% CI for $ln(OR)$ is given by: $ln(\widehat{OR})$ ± 1.96 SE[$ln(\widehat{OR})$]. To convert this to the 95% CI for *OR*, we take the antilog of the limits of the interval (raising *e* to the power of each number).

Lab 9: Summary from Chapter 11 and 12

The one-sample *t*-test:

- evaluates whether there is evidence that a random sample was drawn from a population with a mean that differs from μ_0.
- used if the null hypothesis does not specify a variance (or standard deviation).
- assumes that the sample is random and that the variable is normally distributed.
- The $\bar{Y} \pm 2SE_{\bar{Y}}$ provides us with a quick approximation to the 95% confidence interval for the mean.
- An improved confidence interval for the mean can be produced by changing the "2" to the appropriate critical value from the *t*-table.

Lab 10: Summary from Chapter 12

A paired t-test:

- Tests for a difference in the means of two populations when the data consists of a related pair of measurements (each pair consists of measurements from each population, and the two measurements are not independent of each other).
- Is equivalent to calculating the difference between each member of the pair of data points, and using a one-sample *t*-test to evaluate whether these differences are drawn from a population with a mean of zero (μ_d=0).

The two-sample t-test:

- Used to compare the population means based on random samples from each of two populations. The typical null hypothesis being tested is that $\mu_1 = \mu_2$. It is also possible to use a two-sample *t*-test to test if two means differ by a specified amount such as the hypothesis test that $\mu_1 - \mu_2 = d0$, where d0 is the hypothesized difference in means (for example, a possible null hypothesis is that $\mu_1 - \mu_2 = 3.1$).
- Assumes that: (1) the samples are random and (2) the variable is normally distributed.
- The t-test that we will use further assumes that (3) the populations have equal variance. We estimate this variance with the pooled sample variance: $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Lab 11: Summary from Chapter 15

- Analysis of variance (ANOVA) is an extension of the two-sample t-test, and compares the means of multiple groups. If there are only two groups, then the results of the two-sample t-test and the ANOVA will be identical.
- A significant ANOVA result implies that at least one group has a different mean from one other group.
- ANOVA assumes: (1) equal variance of each group, (2) each group has a normal distribution of the variable, and (3) each population is randomly sampled.

Lab 12: Summary from Chapter 16 and 17

Linear regression:

- Estimate the slope of the linear regression equation $Y = \alpha + \beta X$ between an explanatory variable *X* and a response variable *Y*.
- Assumes that: (1) the relationship between *X* and *Y* is linear, (2) each *Y*-measurement at a given *X* is a random sample from a population of Y-measurement, (3) the distribution of *Y*-values at a given *X* is normal, and (4) the variance of the *Y*-values is the same at all values of *X*.
- Hypothesis tests for the regression slope (β) can be carried out by:
    - A t-test of the regression slope, using the test statistic $t = \frac{b - \beta_0}{SE_b}$, where *b* is the sample regression slope, $SE_b = \sqrt{\frac{MS_{residual}}{\sum_i (X_i - \overline{X})^2}}$ and (*df* = *n* – 2).
    - An ANOVA partitioning the sources of variation into regression and residual (see pp. 554 – 555 of our text).
    - R^2 measures the fraction of variation in *Y* explained by linear regression on *X*.

Correlation analysis:

- The covariance measures the strength of an association between two numerical variables.
- The correlation coefficient measures the strength of a linear association between two numerical variables.
- A hypothesis test of correlation coefficient can be carried out to test the null hypothesis that the population correlation coefficient (ρ) is zero (no linear correlation), using the test statistic $t = \frac{r-0}{SE_r}$,where *r* is the sample correlation coefficient, *n* = number of (*X,Y*) pairs, and $SE_r = \sqrt{\frac{1-r^2}{n-2}}$ (df = n – 2).
- Assumes that: (1) the measurements have a bivariate normal distribution in the population, and (2) random sampling of both *X* and *Y*.