Review: Chapters 3 -8

Chapter 4

1. Random Phenomenon- have long-run regularity
   a. probability as long-term proportion
   b. subjective interpretation

2. Probability Models
   a. Sample Space
      i. Event
   b. probability rules
      i. any probability is a number between 0 and 1, inclusive.
      ii. all possible outcomes together must have probability 1.
      iii. if A is an event then \( P(A^c) = 1 - P(A) \) (Complement Rule)
      iv. if A and B are events with no outcomes in common (disjoint), then
         \[ P(A \text{ or } B) = P(A) + P(B) \]
   c. Finite Sample Space: \( P(A) \) is sum of probabilities of its component outcomes
   d. Independent Events - \( P(A \text{ and } B \text{ occur}) = P(A)P(B) \)

3. Random Variables
   a. Discrete and Continuous
   b. Probability distributions give values of R.V. and corresponding probabilities
      i. Probabilities as areas under the density curve (use geometry or calculus to find them)
      ii. Probabilities from normal distributions

4. Means and Variances of R.V.
   a. How to calculate for discrete R.V., and for continuous R.V. using calculus
      i. Rules for linear functions, \( aX + b \).
      ii. Rules for sum of independent R.V.'s
   b. Law of Large Numbers
      i. proportion of occurrences gets closer to true probability in the LONG-RUN
         \((\hat{p} \to p)\)
      ii. average of a set of outcomes gets closer to the population mean in the LONG-RUN
         \((\overline{x} \to \mu)\)
      iii. beware – the myth of small numbers

Chapter 3

A. Experiments
   1. Recognize whether a study is an observational study or an experiment (an experiment must impose a treatment).
   2. Recognize bias due to confounding of explanatory variables with lurking variables in either an observational study or an experiment.
   3. Identify the factors (explanatory variables), levels, treatments, response variables, and experimental units or subjects in an experiment.
4. Know and explain the basic principles of experimental design.
   a. Understand the importance of comparative experiments.
   b. Explain why a randomized comparative experiment can give good
evidence for cause and effect relationships.
   c. Explain the need for replication.
5. Outline the design of a completely randomized experiment using a diagram. The
diagram should show the sizes of the groups, the specific treatments, and the
response variable.
6. Use Table B to carry out the random assignment of subjects to groups in a
completely randomized experiment.
7. Recognize the placebo effect. Recognize when the double-blind technique should
   be used.
8. Know what a block design is and when it is appropriate to use this design.
   Recognize a matched pairs experiment.
9. Understand to which groups results of an experiment can be generalized and that
   lack of realism can prevent us from generalizing results.

B. Sampling
1. Identify the population in a sampling situation.
2. Identify the sampling frame in a sampling situation.
3. Know terms such a sample, probability sampling, and bias.
4. Distinguish sample from census.
5. Recognize bias due to voluntary response samples and other inferior sampling
   methods.
6. Understand the definition of a simple random sample and be able to recognize if
   a sampling scheme produces an SRS.
7. Use Table B to select a simple random sample from a population.
8. Recognize the presence of undercoverage and nonresponse as sources of error
   in a sample survey. Recognize the effect of the wording of questions on the
   responses.
9. Use random digits to select a stratified random sample from a population when
   the strata are defined.

C. Sampling Distributions
1. Identify parameters and statistics in a sample or experiment
2. Recognize the fact of sampling variability: a statistic will take different values when you
repeat a sample or experiment.
3. Interpret a sampling distribution as describing the values taken by a statistic in all possible
   repetitions of a sample or experiment under the same conditions.
4. Interpret the sampling distribution of a statistic as describing the probabilities of its
   possible values.
5. Distinguish bias and variability in a sampling distribution.
Chapter 5. Sampling Distributions.

A. The Sampling Distribution for Counts and Proportions.
   1. The Binomial Setting
      Two possible outcomes, a fixed number, n, of independent observations, with a fixed
      probability, p, of success for each observation.
   2. Approximate Binomial setting: SRS of less than 5% of population
   3. Exact Binomial Distribution of the count X of successes. Use Table C.
      Know the formulas for mean and standard deviation of X.
   4. Sample proportion. (X/n). Know its mean and standard deviation.
   5. Normal approximation for distribution of counts and proportions. Uses the exact means
      and variances. (\( \hat{p} \) is approximately normal with mean \( p \) and standard
      deviation \( \sqrt{\frac{p(1-p)}{n}} \))

B. The Sampling Distribution of a Sample Mean
   1. For a SRS, \( \bar{x} \) has mean \( \mu \) and standard deviation \( \frac{\sigma}{\sqrt{n}} \)
   2. If the population is normal, then the sampling distribution of \( \bar{x} \) is exactly normal.
      a. A linear combination of independent normal R.V.’s is also normal.
   3. If the population is not normal, then the Central Limit Theorem says that the sampling
      distribution of \( \bar{x} \) is approximately normal for SRS’s of large size (greater than 30)
      with mean \( \mu \) and standard deviation \( \frac{\sigma}{\sqrt{n}} \)

Chapter 6

A. Confidence Intervals
   1. State in nontechnical language what is mean by “95% confidence." Be clear
      about the difference between confidence and probability.( A 95% (or 90% or
      80%, etc...) confidence interval is an interval obtained by a method that, in
      95% (or 90% or 80%, etc...) of all samples, will produce an interval containing
      the true population parameter.)
   2. Calculate a confidence interval for the mean \( \mu \) of a normal population with known
      standard deviation \( \sigma \). Know the formula for the margin of error, and for the
      interval endpoints...( level C confidence interval: \( \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \) where \( z^* \) is
      obtained from Table D)
   3. Recognize when you can safely use this method for computing confidence
      intervals and when the data collection design or a small sample from a
      skewed population makes it inaccurate. Understand that the margin of error
      does not include the effects of undercoverage, nonresponse, or other
      practical difficulties. It just accounts for sampling variability.
   4. Understand how the margin of error of a confidence interval changes with the
      sample size, standard deviation, and the level of confidence C.
5. Understand how the critical value $z^*$ is obtained.
6. Find the sample size required to obtain a confidence interval of specified margin of error $m$ when the confidence level and other information are given.

B. Significance Tests

1. Hypotheses
   a. null hypothesis ($H_0$)
      i. statement we assume is true & assess strength of evidence against
      ii. generally a statement of "no effect" or "no difference"
      iii. general form: population parameter = specific value (e.g. $\mu = 0$ or $p = 0.5$)
   b. alternate hypothesis ($H_a$)
      i. statement which, before we look at data, we hope or suspect is true instead of $H_0$
      ii. general forms.
         - two-sided: population parameter $\neq$ specific value (e.g. $\mu \neq 0$ or $p \neq 0.5$)
         - one-sided: population parameter $>$ specific value (e.g. $\mu > 0$ or $p > 0.5$)
           population parameter $<$ specific value (e.g. $\mu < 0$ or $p < 0.5$)

2. Test statistic ($Z$) – a number calculated from the data that measures how far the data differ from what we expect when the null hypothesis is true. Usually a standardized value of an estimate of the population parameter.

3. (A) Critical value approach. Reject $H_0$ if $Z$ is too far away from its expectation when $H_0$ is true, i.e. when $Z$ is in the "critical region." The critical region is chosen so that the probability of rejecting $H_0$ when it is true is a specified value $\alpha$.

3. (B) P-values
   a. interpretation:
      A probability (computed assuming $H_0$ is true) that a sample would yield results as extreme or more extreme than the results we got. Small P-values provide strong evidence against the null hypothesis. It is computed as the value of $\alpha$ that would result if the observed value of $Z$ were on the border of the critical region.
   b. often computed as some area under a density curve of a normal sampling distribution
      i. for a one-sided alternative hypothesis, P-value involves area in only one tail
      ii. for a two-sided alternative hypothesis, P-value involves total area in both tails
   c. a test is statistically significant at level $\alpha$ (i.e. reject $H_0$) if the P-value of the test is less than or equal to the significance level $\alpha$. This makes the two forms of significance testing completely equivalent.

3. Tests for Population Means
   a. assumptions needed for validity
      i. data are from a SRS
      ii. $n$ is large
      iii. population standard deviation, $\sigma$, is known
   b. state hypotheses
      i. $H_0$: $\mu = \mu_0$ where $\mu_0$ is some known value
      ii. $H_a$: $\mu \neq \mu_0$ where $\neq$ is one of $<, \neq, >$ depending on the particular problem
   c. calculate test statistic (the standard score of $\bar{x}$): $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
d. Compare $z$ to the critical value $z^*$ and reject $H_0$ when $z < z^*$, $z > z^*$, or $|z| > z^*$ depending on the alternative. The value of $z^*$ is chosen to make the probability of rejecting $H_0 = \alpha$ when $H_0$ is true.

**ALTERNATIVELY:**
calculate $P$-value using Table A
  i. $H_a$ determines the appropriate area under the normal curve to calculate

e. Know the equivalence between two sided hypothesis tests and confidence intervals.

4. Some Cautions on Tests of Significance
   a. $P$-values are more informative than significance levels (which have often been used inappropriately)
   b. lack of statistical significance can have practical significance and shouldn't be ignored
   c. beware of searching for significance; a large set of tests will often produce one or more significant results by chance (even though $H_0$ really was true in every case)
   d. Recognize that significance testing does not measure the size or importance of an effect (statistical significance vs. practical significance). Especially with very large sample sizes, a result may be statistically significant without having much practical significance. With small sample sizes, effects that are of practical significance may not be statistically significant.
   e. Recognize when you can use the $z$ test and when the data collection design or a small sample from a skewed population makes it inappropriate. Statistical significance means basically nothing if data was collected through a study with poor design.

**Chapter 8**

Confidence Intervals and Tests for Proportions

1. Know the assumptions for validity: data are from a SRS, $n$ is large, but not more than 5% of the population, $p$ is not too close to 0 or 1.

Confidence intervals for Population Proportions, $p$

1. Use the "large sample" method based on the sample proportion $\hat{p}$

   a level $C$ confidence interval is: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $z^*$ is obtained from Table D.

   This would be the default method provided that the conditions for its use are satisfied. Know what those conditions are.

2. Use the Plus Four estimate $\tilde{p} = \frac{X + 2}{n + 4}$

   a level $C$ confidence interval is: $\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + 4}}$ where $z^*$ is obtained from Table D.

   Know the conditions for using this approach.

3. find the sample size needed to produce a specified margin of error and confidence level using the "large sample" method. Know the conservative approach and the guessed $p^*$ approach.
Tests for Population Proportions

1. state hypotheses
   a. $H_0: p = p_0$ where $p_0$ is some known value
   b. $H_a: p \neq p_0$ where $\neq$ is one of $<, =, >$ depending on the particular problem

2. calculate test statistic (the standard score of $\hat{p}$): $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

3. calculate P-value using Table A
   a. $H_a$ determines the appropriate area under the normal curve to calculate
   b. Know when this approach is justified.

Chapter 7

A. One-Sample t Procedures

1. Recognize when the t procedures are appropriate in practice, in particular that they are moderately robust against lack of normality but are strongly influenced by outliers.

2. Also recognize when the design of the study, outliers, or a small sample from a skewed distribution make the t procedures risky.

3. Use t to obtain a confidence interval at a stated level of confidence for the mean $\mu$ of a population when the population standard deviation $\sigma$ is unknown and the sample standard deviation $s$ is used as an estimate of $\sigma$.

4. Carry out a t test for the hypothesis that a population mean $\mu$ has a specified value against either a one-sided or a two-sided alternative. Use Table D of t distribution critical values to approximately give a range for the P-value or carry out a fixed level $\alpha$ test.

5. Know the equivalence between two sided hypothesis tests and confidence intervals.