# Stat 50 Elementary Stastitics

## Study Guide for Final Exam: Chapters 1-12

Note: you may bring a GoodCheatSheet with you and your calculator. **I will not provide any formulas for you.** Please come to the final prepared!

Start by reviewing each of your exams, quizzes, and the exam reviews. The lecture notes are a great resource as well, since they contain the information and examples that I find the most important. Please take a look at old homework problems as well and make sure you know them.

**Chapter 1**

- Given a research objective, be able to identify the target population(s) and variable(s) that need to be studied to answer that question.
- Be able to explain why we select *samples* to study, rather than study the whole *population* (census).
- Given a study, be able to identify whether random sampling, convenience sampling, or volunteer sampling was used to select the sample. Be able to identify random vs. non-random sampling.
- Quantitative vs. Qualitative and Continuous vs. Discrete Variables
- Observational study vs. randomized experiment and the benefit of the latter. Know what a double-blind experiment is.
- Understand the basic idea behind a biased vs. unbiased study. Knowing the meaning of voluntary bias, self-interest bias, social acceptability bias, and leading question bias.
- Know what a confounder is, also called lurking variable. Understand that a correlation between two variables doesn't automatically mean that one variable cause an increase or decrease of the other, but that there could be another, or several other, variables that are the actual cause.(Ch. 12)
- Given an ungrouped quantitative data set, be able to construct a frequency distribution table with classes of equal width, containing frequency, relative frequency, percent, and midpoint columns.

**Chapter 2**

- Given a frequency, relative frequency, or percentage table for a quantitative variable, be able to construct a histogram, bar graph, and polygon. Conversely, given a graph, be able to reconstruct the frequency, relative frequency, or percentage distribution table. (Ch. 1 and 2)
- Be able to construct and interpret a stem-and-leaf plot and dotplot graph.
- Understand and be able to calculate the mean, median, mode, standard deviation, variance, and range for an ungrouped data set using correct notation (s vs. $\sigma$ ) and correct units.(STATS →CALC→1-VarStats)
- Skewness and outliers. Know which parameters are more sensitive to outliers.

- Be able to calculate the mean, standard deviation, and variance for grouped data (from a frequency, relative frequency, or percentage distribution table) by using the midpoints and frequencies of the various classes.

- Be able to calculate the quartiles, the inner quartile range, the $k^{th}$ percentile, or the percentile rank of a particular data value for a data set using correct notation and units.

- Calculate and interpret z-scores.

- Given the mean, the standard deviation, and an interval of data values $\bar{x} - ks$ to $\bar{x} + ks$, be able to determine $k$ (the number of standard deviations above and below the mean) and use Chebyshev's Theorem to calculate the minimum percentage $\left(1-\frac{1}{k^2}\right)\cdot 100\%$ of data values within $k$ standard deviations from the mean.

- Conversely, given a minimum percentage of data values, be able to determine $k$ using Chebyshev's Thm. Then find the interval of data values using *k*, the mean, and the standard deviation: $\bar{x} - ks$ to $\bar{x} + ks$.

- Be able to do calculations similar to the previous two items for approximately bell-shaped data using approximate percentages and the Empirical Rule.

- Be able to construct a boxplot using your calculator and interpret it (is it skewed, what is the median, etc.)

**Chapters 3**

- Be able to draw a tree diagram with probabilities for an experiment.

- Be able to do some basic counting, such as how many possible outcomes are there when tossing a dice five times, or how many different ATM codes with four digits exists.

- Given a two-way classification table, or a probability tree, be able to calculate probabilities including "and", "or" and conditional probabilities either directly from the table or using the appropriate general formulas:

  $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

  $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$ $\qquad$ $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

- Be able to explain what the complement of an event is, and be able to use its formula: $P(A^C) = 1 - P(A)$

- Be able to explain what mutually exclusive events are.

- For *mutually exclusive* events $A$ and $B$, $\quad P(A \text{ and } B) = 0$ so $\quad P(A \text{ or } B) = P(A) + P(B)$

- Be able to explain what independent events are, how to test if two events are independent, and how this relates to the probability of the intersection two events:

  If $P(A) = P(A \mid B)$, then $A$ and $B$ are independent.
  If $P(A) \neq P(A \mid B)$, then $A$ and $B$ are dependent.

  For *independent* events $A$ and $B$, $\quad P(A \text{ and } B) = P(A) \cdot P(B)$

**Chapter 4**

- Given a random variable, be able to explain whether it is qualitative or quantitative, and be able to classify a quantitative random variable as discrete or continuous.

- Given a small population and an experiment, be able to find the probability distribution table for a discrete random variable using the given information or by making a tree diagram with probabilities, using sampling with or without replacement.

- Be able to identify when a table is the probability distribution for a discrete random variable.

- Using a probability distribution function table to calculate the probability that a discrete random variable is a single value or within an interval of values.

- Be able to find the mean and standard deviation of a discrete probability distribution. Be able to interpret the mean as the expected value of the variable.

- Be able to explain what the criteria of a binomial experiment are, and apply these criteria to specific situations to determine if an experiment is binomial.

- Be able to determine if an experiment is a binomial experiment by looking at its tree diagram.

- Be able to recognize a binomial problem. It is usually a probability problem where we are given a certain percentage or probability $p$ of selecting an element having a certain characteristic (which does not change after each selection), and we need to calculate the probability of selecting $x$ out of $n$ elements having that characteristic.

- Calculate the probability using the appropriate formula or program [PRGM -> BINOML83].

- Be able to calculate the mean and standard deviation of a binomial distribution using the appropriate formulas. Be able to interpret the mean as the expected number of successes out of $n$ trials.

**Chapter 5-7**

- Be able to explain what the mean and standard deviation tell us about the center and spread of a normal distribution curve.

- Be able to explain what the standard normal distribution is ( $\mu = 0, \sigma = 1$ )

- Given $x$, calculate its corresponding $z$ -score, or vice versa.
  Remember, $z$ tells us the number of standard deviations $\sigma$ that $x$ is from the mean $\mu$:

- $z = \dfrac{x - \mu}{\sigma}$ $\quad x = \mu + z\sigma$ or on calculator using PRGM -> NORMAL83 or INVNOR83

- Be able to calculate the probability that a normally distributed variable *x* is over a certain interval. Draw a picture with correctly labeled areas and axis.

- Be able to calculate the appropriate $z$ or $x$ value given the probability or percentage of data values
  Be able to explain the difference between a population distribution and a sampling distribution.

- Be able to obtain the sampling distribution of the sample means $\bar{x}$, and understand what the axes in such a sampling distribution represent.

- The sampling distribution of $\bar{x}$ will be approximately normally distributed when…
  1. The population distribution is already normally distributed (regardless of sample size).
  2. The sample sizes taken are large $n \geq 30$, regardless of the shape of the population distribution. (this is the Central Limit Theorem for Means).

  In either case, $\mu_{\bar{x}} = \mu,\ \sigma_{\bar{x}} = \sigma/\sqrt{n}$

- Be able to recognize a probability problem that uses the sampling distribution of all sample means: you are asked to calculate the probability of selecting a simple random sample of a certain size that has a *sample mean* over a certain interval, or you are asked to calculate the percentage of simple random samples of a certain size that have means over a certain interval.

- Be able to calculate the probability that the sample mean $\bar{x}$ is over a certain interval.

- The sampling distribution of $\hat{p}$ will be normally distributed when $np > 10$ and $nq > 10$:

  then $\mu_{\hat{p}} = p,\ \sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

- Be able to recognize a probability problem that uses the sampling distribution of all sample proportions: it looks similar to a binomial problem, but rather than finding the probability of a certain number of successes as we do in a binomial problem, we are asked the find the probability of selecting a *sample* in which a certain interval of proportions or percentages of them have a certain characteristic.

- Calculate the probability that the sample proportion $\hat{p}$ is over a certain interval.

- Assess normality by using dotplots, boxplots, histograms, stem-and-leaf plots, and normal quantile plots.

**Chapter 8 and additional topics in 10 - Confidence Intervals**

- A sample mean $\bar{x}$ is a point estimate of a population mean $\mu$.
  A sample proportion $\hat{p}$ is a point estimate of a population proportion $p$.
  A difference of sample means $\bar{x}_1 - \bar{x}_2$ is a point estimate of a difference of population means $\mu_1 - \mu_2$.
  A sample mean of differences $\bar{d}$ is a point estimate of a population mean difference $\mu_d$.
  A difference of sample prop. $\hat{p}_1 - \hat{p}_2$ is a point estimate of a difference of population prop. $p_1 - p_2$.

- We construct a confidence interval using a sample statistic (point estimate), together with some error, whenever we want to estimate an unknown population parameter: point estimate $\pm$ margin of error

- Be able to explain what the confidence level tells us: the percentage of samples of the same size $n$ that will make confidence intervals that actually contain the true population parameter; thus, a certain percentage of confidence intervals will not contain the population parameter, and we usually never know if our sample's interval contains the population parameter, or not.

- Be able to use the appropriate formula to estimate the sample size needed to construct a confidence interval for a population mean $\mu$ with the desired error and confidence level.

- Be able to use the appropriate formula to estimate the sample size needed to construct a confidence interval for a population proportion $p$ with the desired error and confidence level. For the most conservative estimate (or when we don't have any $\hat{p}$ available), use $\hat{p} = 0.5$.

- To decrease the error in a confidence interval estimate:
  1. Increase the sample size (preferred, but not always economical or possible).
  2. Decrease the confidence level.

- Be able to calculate confidence intervals for one population mean $\mu$, one population proportion $p$, the difference between two population means $\mu_1 - \mu_2$, the difference between two population proportions $p_1 - p_2$, or the population mean difference $\mu_d$, using the formulas and/or the calculator, and write your answer in the form of a detailed sentence as we did in-class (ex. *"we are 95% confident that the true population mean of ... is between ... and ...*). Remember, when writing conclusions for the confidence intervals for two populations make a comparison between the population parameters of the first and second populations instead of using the words "different" or "difference. You should also avoid using negative numbers in your final conclusion.

- Use the appropriate inverse program to get the $z$ or the $t$ when using the formulas:

| **Desired Estimate:** | **Assumptions:** | **Distribution:** | **Formula:** | **Program:** |
|---|---|---|---|---|
| Conf. Int. for $\mu$ | 1. SRS<br>2. $n > 30$ or population is normal<br>3. $\sigma$ known | $z$ distribution | $\overline{x} \pm z\frac{\sigma}{\sqrt{n}}$ | ZInterval |
| Conf. Int. for $\mu$ | 1. SRS<br>2. $n > 30$ or population is normal<br>3. $\sigma$ unknown, $s$ | $t$ distribution<br>$df = n - 1$ | $\overline{x} \pm t\frac{s}{\sqrt{n}}$ | TInterval |
| Conf. Int. for $p$ | 1. SRS<br>2. $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$ | $z$ distribution | $\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | 1-PropZInt |
| Conf. Int. for $\mu_1 - \mu_2$ | 1. Independent SRSs<br>2. $n_1, n_2 > 30$ or pops. normal<br>3. $\sigma_1, \sigma_2$ unknown, and $s_1, s_2$ known | $t$ distribution<br>$df = n_{smallest}$ -1 | $(\overline{x}_1 - \overline{x}_2) \pm t\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$ | 2-SampTInt (not pooled) |
| Conf. Int. for $\mu_d$ | 1. Paired SRSs<br>2. $n \geq 30$ or pop. of diff. normal<br>3. $\sigma_d$ unknown, $s_d$ known | $t$ distribution<br>$df = n - 1$ | $\overline{d} \pm t\frac{s_d}{\sqrt{n}}$ | TInterval |
| Conf. Int. for $p_1 - p_2$ | 1. Independent SRSs<br>2. Each pop. size$\geq 20 \cdot$ n<br>3. Two categories with at least 10 in each. | $z$ distribution | $(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ | 2-PropZInt |

Note: SRS = Simple Random Sample

**Chapter 9, 10, 11 Hypothesis Tests**

- A hypothesis test is a procedure that helps us make a decision regarding statements made about the characteristics of a <u>population</u>.

- Be able to perform hypothesis tests about a single population mean $\mu$, a single population proportion $p$, the difference between two population means $\mu_1 - \mu_2$, the difference between two population proportions $p_1 - p_2$, the population mean difference $\mu_d$, a goodness of fit test, test of independence, and an analysis of variance test (ANOVA), using the six-step procedure:

**6-Step Procedure for Performing a Hypothesis Test:**

1) State the null and alternative hypotheses of the test.
2) Compute the test statistic. On the final exam you are allowed to use your calculator to do this (if so state calc. program).
3) Draw a picture of the standardized sampling distribution you are using. Label the axis, the test statistic, and the area of the p-value.
4) Calculate the *P*-value.
5) Interpret the *P*-value and make a decision.
   If *P*-value < $\alpha$ then we reject the null hypothesis, and we have sufficient evidence for the alternative hypothesis.
   If *P*-value > $\alpha$ then we do NOT reject the null hypothesis, and we do NOT have sufficient evidence for the alternative hypothesis.
6) State a conclusion in the form of a detailed sentence that addresses the alternative hypothesis.
   When we Reject $H_0$, we say "there is sufficient evidence to show that $H_1$", where $H_1$ is stated in words.
   When we Fail to Reject $H_0$, we say "there is not sufficient evidence to show that $H_1$", where $H_1$ is stated in words.

| **Type of Test:** | **Assumptions:** | **Distribution:** | **Details:** |
|---|---|---|---|
| Test about $\mu$:<br><br>$H_0$: $\mu = value$<br><br>$H_1$: $\mu \neq value$<br>$\mu < value$<br>$\mu > value$ | 1. SRS<br>2. $n > 30$ or pop. is normal<br>3. $\sigma$ known | $z$ distribution | Test Stat: $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$<br><br>P-Value: Normal83, or ZTest |
| Test about $\mu$:<br><br>$H_0$: $\mu = value$<br><br>$H_1$: $\mu \neq value$<br>$\mu < value$<br>$\mu > value$ | 1. SRS<br>2. $n > 30$ or pop. is normal<br>3. $\sigma$ <u>not</u> known, $s$ known | $t$ distribution<br><br>$df = n - 1$ | Test Stat: $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$<br><br>P-Value: T83, or TTest |
| Test about $p$:<br><br>$H_0$: $p = value$<br><br>$H_1$: $p \neq value$<br>$p < value$<br>$p > value$ | 1. SRS<br>2. $np > 10$ and $n(1-p) > 10$ | $z$ distribution | Test Stat: $z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$<br><br>P-Value: Normal83, or 1-PropZTest |
| Test about $\mu_1 - \mu_2$:<br><br>$H_0$: $\mu_1 - \mu_2 = 0$<br><br>$H_1$: $\mu_1 - \mu_2 \neq 0$<br>$\mu_1 - \mu_2 < 0$<br>$\mu_1 - \mu_2 > 0$ | 1. Independent SRSs<br>2. $n_1, n_2 > 30$ or pops. Normal<br>3. $\sigma_1, \sigma_2$ <u>not</u> known, $s_1, s_2$ known | $t$ distribution<br><br>df= $n$-1 for smallest $n$ or use calculator to find df (it will be different) | Test Stat: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$<br><br>P-Value: T83, or 2-SampTTest (not pooled) |

**Chapter 9, 10, 11 (continued)**

| **Type of Test:** | **Assumptions:** | **Distribution:** | **Details:** |
|---|---|---|---|
| Test about $\mu_d$:<br><br>$H_0$: $\mu_d = 0$<br><br>$H_1$: $\mu_d \neq 0$<br>$\mu_d < 0$ | 1. Paired SRSs<br>2. $n > 30$ or pop. of diff. normal<br>3. $\sigma_d$ unknown, $s_d$ known | $t$ distribution<br>$df = n - 1$ | Critical Value(s): TINVRS83<br>Test Stat: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$<br>P-Value: TTest |
| Test about $p_1 - p_2$:<br><br>$H_0$: $p_1 - p_2 = 0$<br><br>$H_1$: $p_1 - p_2 \neq 0$<br>$p_1 - p_2 < 0$<br>$p_1 - p_2 > 0$ | 1. Independent SRSs<br>2. $n_1\hat{p}_1 > 5$,<br>$n_1(1-\hat{p}_1) > 5$<br>$n_2\hat{p}_2 > 5$,<br>$n_2(1-\hat{p}_2) > 5$ | $z$ distribution | Critical Value(s): INVNOR83<br>Test Stat:<br>$z = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}}$,<br>$p = \frac{x_1 + x_2}{n_1 + n_2}$<br>P-Value: INVNOR83, or INVNOR83, or 2-PropZTest |
| **Goodness of Fit Test**<br>$H_0$: Pop. fits expected distr.<br>$H_1$: Pop. does not fit expected distr. | 1. SRS<br>2. All expected frequencies $\geq 5$ | $\chi^2$ distribution<br>$df = k - 1$ | Test Stat: $\sum \dfrac{(O-E)^2}{E}$<br>where each $E = np$<br>P-Value: CHI83, or GOODFT83, or $\chi^2$ GOF-Test |
| **Test of Independence**<br>$H_0$: Two characteristics of a population are independent<br>$H_1$: Two characteristics of a population are dependent. | 1. SRS<br>2. All expected frequencies $\geq 5$ | $\chi^2$ distribution<br>$df = (R-1)(C-1)$ | Test Stat: $\sum \dfrac{(O-E)^2}{E}$<br>where each<br>$E = \dfrac{(Row\ Total)(Column\ Total)}{Sample\ Size}$<br>P-Value: CHI83, or CHITST83 or $\chi^2$-Test |
| **Analysis of Variance (ANOVA)**<br><br>$H_0$: 3+ Pop. means are all equal<br><br>$H_1$: 3+ Pop. means are not all equal | 1. Independent SRSs<br>2. Pops. normal<br>3. $\sigma$'s are all equal<br><br>For this class, we will always assume this to be true, even if not stated. | $F$-distribution | P-Value and Test Stat: ANOVA |

**Chapter 9, 10, 11 (continued)**

- Be able to recognize when two samples are selected independently, and when paired samples are selected dependently.

- Notice that many of the test statistics for means and proportions measure the number of standard deviations in the sampling distribution that the sample statistic is from the null hypothesis value, which measures the evidence against $H_0$.

$$\text{test stat} = \frac{\text{(sample stat)} - \text{(null hypothesis value)}}{\text{(stdev of sampling distribution)}}$$

- Since we use a random sample in a hypothesis test, there is always a chance that we make the wrong the decision in any hypothesis test we perform:

  Type I error: Deciding to reject $H_0$ when $H_0$ is actually true.
  Type II error: Deciding to fail-to-reject $H_0$ when $H_0$ is actually false (when $H_1$ is actually true).

- Based on your conclusion to a hypothesis test, be able to identify whether a type I or type II error could have been made.

- Be able to explain what the significance level ($\alpha$) measures. Remember, it is the probability of making a type I error. In other words, assuming the null hypothesis is true, it is the percentage of all simple random samples that could have been selected that would have lead us to making the type I error of rejecting the null hypothesis when it is true.

**Chapter 12**

- Given paired data between two variables, be able to determine which is the independent vs. the dependent variable and construct a scatter plot for the data.

- Given paired data between two variables, be able to find the equation of the least squares regression line (a.k.a. the line of best fit) for the paired data using correct notation in your equation. Be able to do this on the calculator [STAT→CALC→LinReg(a+bx) or STAT→TESTS→LinRegTTest ]

- Be able to identify the slope and y-intercept in the equation of the regression line and be able to explain, in detail what they mean in a particular situation.

- Be able to graph the regression line on the same graph as the scatter diagram. Be able to do this by-hand.

- Be able to calculate the linear correlation coefficient on the calculator [LinReg(a+bx) or LinRegTTest]. and tell if it is significant (look at p-value in LineRegTTest). Determine whether the linear correlation (based on a situation or from looking at a scattergram) is positive or negative, and whether it indicates a weak, medium, or strong linear relationship between the dependent and independent variables.

- Be able to use the regression equation to make predictions.

- Understand the dangers of using linear regression for making predictions outside of the domain, proving causality (that one variable causes a certain behavior of the other variable), or modeling nonlinear data.