

# Review for final Exam

Note Title

## Chapter 1 (1.1-1.3)

- Just know vocabulary.

## Chapter 2 (2.1-2.5)

- Know the difference between:
  - ① Categorical and Quantitative
  - ② Discrete and continuous
  - ③ Skewed to the left and skewed to the right.
- Be able to draw and interpret a histogram for quantitative data
- Be able to calculate the mean, standard deviation, and five number summary of data from a random sample.
- Know what an outlier is and how to find one (there are a couple of ways)
- Know, understand, and be able to use the Empirical Rule.
- Know what and how to find quartiles. Once you have them, be able to calculate an IQR and use the IQR to determine outliers.
- Be able to draw a fully labeled box plot.

## Chapter 3 (3.1-3.4)

- Be able to identify a response and an explanatory variable.
- Once identified, be able to appropriately graph and label a scatterplot of the two variables and then perform a full regression analysis. The quiz on regression will be perfect for studying this.
- Know that correlation  $\neq$  causation! Correlation is merely a measure of linear association.
- Know what a lurking variable is.

## Chapter 4 (4.1-4.4)

- Know vocabulary
- Know how to get a random sample using a random digit table.

## Chapter 5 (5.1-5.2) focus on 5.2 but know vocals. from 5.1

- Know vocabulary
- Be able to draw a tree diagram
- Know the rules for the probabilities for a sample space.
- Be able to find the probability of an event

- Know, understand, and be able to use the complement to make working problems easier
- Understand the difference between disjoint and independent events.
- Know and understand how to find the probability of the union or intersection of two events. (dependent + independent events give different results.)

### Chapter 6 (6.1-6.5) and Chapter 7 (7.1-7.3)

- Be able to find the mean of a probability distribution for a discrete random variable

#### Normal Distribution (a continuous distribution)

(6.2) If  $X_1, \dots, X_n$  are normally distributed, in order to find the probability that any  $X$  is  $\geq$  or  $\leq$  some particular value, we must standardize.

Using the z-score. That is, for an arbitrary constant  $c$ ,

$$P(X < c) = P\left(\frac{X - \mu}{\sigma} < \frac{c - \mu}{\sigma}\right) = P\left(Z < \frac{c - \mu}{\sigma}\right).$$

Then, you use the z-score table to find this probability

(6.5) You need to know that  $\bar{X}$ , the sample mean, also has a probability distribution with mean  $= \mu$  and standard error  $= \sigma/\sqrt{n}$

If  $n > 30$ , then the CLT says  $\bar{X}$  is approximately normally distributed.

If we want to find  $P(\bar{X} < c)$  we standardize using

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{c - \mu}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{c - \mu}{\sigma/\sqrt{n}}\right)$$

(7.3)  $\bar{X}$  is a point estimate for  $\mu$  when  $X_1, \dots, X_n$  are normally distributed. If we want to construct a confidence interval to estimate a population mean, we need to keep in mind that since we don't know  $\mu$ , we probably won't know  $\sigma$ ; we'll just be given  $s$ . So, we'll have to use the t-chart rather than the z-chart. The confidence interval endpoints will be, for 95% confidence level,

$$\bar{X} \pm t_{0.025}(s.e.), \text{ where } s.e. = \frac{s}{\sqrt{n}} \text{ and } d.f. = n - 1$$

Should be able to find t-scores for 90% and 99% C.I.'s too.

#### Binomial Distribution (a discrete distribution)

(6.3) Know the conditions for a binomial distribution

If  $X_1, \dots, X_n$  are binomial random variables, then the probability that

$X$  takes on some value, say 2 is

$$P(X=2) = \frac{n!}{2!(n-2)!} p^2 (1-p)^{n-2} \quad \text{where } n = \# \text{ of trials, } p = \text{proportion of successes}$$

The mean and standard deviation of a binomial random variable are

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$

(6.4) The proportion of successes,  $p$ , also has a sampling distribution with mean  $= p$  and s.d.  $= \sqrt{p(1-p)/n}$  = standard error of sample proportion (or s.e. for short)

(7.2)  $\hat{p}$  is a point estimate for  $p$ , the population proportion. In order to construct a confidence interval to estimate a population proportion, we choose a confidence level, say 95%, (usually given) and then a 95% confidence interval has endpoints ( $Z = 1.96$ )

$$\hat{p} \pm 1.96(\text{s.e.}), \quad \text{where} \quad \text{s.e.} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \left( \begin{array}{l} \text{we have to use} \\ \hat{p} \text{ b/c we don't} \\ \text{know } p \end{array} \right).$$

We also need that  $n\hat{p} \geq 15$  and  $n(1-\hat{p}) \geq 15$  for a valid C.I.  
You should be able to find Z-scores for 90% and 99% C.I.'s too.

(8.2) Hypothesis testing for  $p$ , the population proportion.

You must be able to perform all 5 steps of a hypothesis test

① Assumptions + parameters

- Categorical data
- Random sample
- $np \geq 15$  and  $n(1-p) \geq 15$

[ note: In class I wrote  $n\hat{p}$  and  $n(1-\hat{p})$ .  
This was a mistake. I made this mistake because with Confidence Intervals you do use  $n\hat{p}$  and  $n(1-\hat{p})$ . ]

② Hypothesis (usually has to be inferred from story problem)

$$H_0: p = p_0 \quad \text{vs.} \quad H_A: p > p_0, \quad H_A: p < p_0, \quad \text{or} \quad H_A: p \neq p_0$$

③ Test Statistic

$$z = \frac{\hat{p} - p_0}{\text{s.e.}}, \quad \text{where} \quad \text{s.e.} = \sqrt{\frac{p_0(1-p_0)}{n}} \quad \leftarrow \text{we use } p_0 \text{ here and not } \hat{p} \text{ b/c we want to know the standard error when we assume } H_0 \text{ is true.}$$

④ P-value

$$p\text{-value} = P(Z > z) \quad \text{if} \quad H_A: p > p_0$$

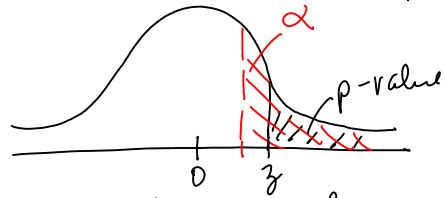
$$p\text{-value} = P(Z < z) \quad \text{if} \quad H_A: p < p_0$$

$$p\text{-value} = 2P(Z \leq -|z|) \quad \text{if} \quad H_A: p \neq p_0$$

## ⑤ Conclusion

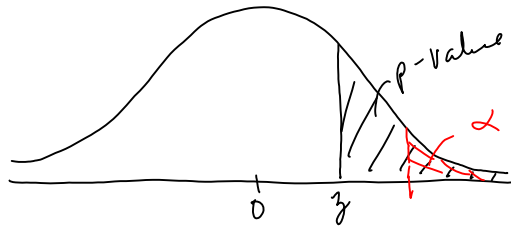
Be prepared to look at significance levels,  $\alpha = 0.1, 0.05,$  and  $0.01$

If  $p\text{-value} < \alpha$ , then the  $p\text{-value}$  area contains extreme values



test statistics that are further away from the mean than the cut off value,  $\alpha$ , allows. So, we would reject  $H_0$ .

If  $p\text{-value} > \alpha$ , then the  $p\text{-value}$  area contains less extreme values



test statistics that are closer to the mean than the cut off value,  $\alpha$ , allows. Since it is desirable to have values closer to the mean, we would fail to reject  $H_0$ .