# AP/Dual Enrollment Statistics

Open Response Final

Review Packet

Madison High School

Mrs. Palmer

Updated 2017

Do not write on this packet.

Please return to Mrs. Palmer on the day of the final.

Packet #

***Please Do Not Write On This Packet***

**Regression Formulas:**

$\hat{y}=b\_{0}+b\_{1}x$ where $b\_{0}=\overbar{y}-b\_{1}\overbar{x}$ and $b\_{1}=r\frac{s\_{y}}{s\_{x}}$

**Sampling Distribution Formulas:**

If random sample of size *n* from an infinite population with mean µ and standard deviation σ, then:

 

**Inference Formulas:**

 Test Statistics: $\frac{statistic-parameter}{SD of statistic}$

 Confidence Interval: $statistic\pm \left(critical \#\right)\*\left(SD of statistic\right)$

|  |  |  |
| --- | --- | --- |
|  | Statistic | Standard Deviation Of Statistic |
| **Proportions**(Normal Model) | Sample Proportion |  |
| Difference of sample proportions |  where $p= \frac{p\_{1}+p\_{2}}{n\_{1}+n\_{2}}$ |
| **Means**(T Model) | Sample Mean |  |
| Difference of sample means |  |

$$χ^{2}=\frac{\left(observed-expected\right)^{2}}{expected}$$

**Test Tips:**

* Write all answers in complete sentences.
* When using a calculator to solve a problem, you must include calculator talk.
* ALWAYS draw a graph or distribution.

**Normal Distributions:** Chapter 2

**What to Study:** To be successful on this section of the final, the student will

* **Z-scores**
	+ Define in context
	+ How to calculate
	+ Using z-scores to compare two or more items.
* **Percentiles**
	+ How to find a percentile
	+ Define in context
* **Ogive Graphs**
	+ Identify a percentile
	+ Estimate the IQR
* **Normal model**
	+ Draw a model - N(mean, standard devation)
	+ 68-95-99.7 Rule
	+ Find any area under the curve (Normalcdf)
	+ Find a value given an area (InvNormal)
	+ Show all work – “Calculator Talk”

**Vocabulary:**

* + percentiles
	+ cumulative relative frequency graphs
	+ z-scores
	+ transforming data
	+ density curves
	+ median of density curve
	+ transform data
	+ mean of density curve
	+ standard deviation of density curve
	+ Normal curves
	+ Normal distributions
	+ 68-95-99.7 rule
	+ 
	+ standard Normal distribution
	+ standard Normal table
	+ Normal probability plot
	+ μ (mu)
	+ σ (sigma)

**Problems to Review:**

1. For each problem below draw a picture of the normal curve and shade the area you have to find. Let Z represent a variable following a standard normal distribution.



1. Find the proportion that is less than z=2.00.
2. Find the proportion that is between z = −.13 and

z = 1.75.

1. Find the proportion that is greater than z=1.86.
2. Find the z−score for the 64th percentile.
3. Find the z−scores that bound the middle 50% of all data
4. Find the z−score for the 24th percentile.
5. Former ISU basketball player Kelvin Cato is 83 inches tall. Assuming that heights follow approximately a normal distribution with mean 70 and standard deviation σ = 3,



1. what is his corresponding z-score?
2. what proportion of men are taller than him?



1. Since the length of a downhill ski is related to the height

of the individuals renting them, it is fair to assume that a

normal distribution would describe the length of women’s

skis at rental outlets in Colorado. The mean of the

distribution is 150 cm and the standard deviation is 10 cm.

* 1. What is the proportion of women’s ski lengths that are less than 130 cm?
	2. What is the proportion of women’s ski lengths that are greater than 125 cm?
	3. What is the proportion of women’s ski lengths that are between 125 and 155?
	4. Very long skies are expensive and there are not many people who rent them. What is the longest women’s ski a rental shop should carry so that only 2 percent of the costumers will ask to rent a longer ski?



1. The BMI for males age 20 to 74 is follows approximately

a normal distribution with mean µ = 27.9 and standard

deviation σ = 7.8. Use the 68-95-99.7 rule to find

* 1. the percentage of males with BMI less than 20.1.
	2. the percentages of males with BMI greater than 12.3.
	3. the BMI values that correspond to the middle 99.7% of the distribution.
	4. the value such that 0.15% of males have BMI’s greater than the value.

**Regression:** Chapter 3

**What to Study:** To be successful on this section of the final, the student will

* Make sure everything is in context!!!!
* Know all vocabulary and be able to use it in context.
* NOT need to create a scatterplot; the graph will be given to you.
* Know how to find r given R2 and R2 give r.
* Know how to find the slope and intercept given means, standard deviation and r.
* Know how to interpret the slope and intercept in context
* Know how to interpret R2 in context “ % of the variation in can be explain using this model with as the explanatory variable”
* Understand what a residual is and how to calculate a residual for a given point.
* Know when we have underestimated or overestimated a residual
* Know when a linear model is appropriate
* Be cautious of extrapolation and lurking variables
* Know the difference between a leverage point and an influential point.
* Read through the Unit II Summary on page 244 (both books) – ignoring any points about re-expressing data (they are from Chapter 10 – which is not on the final.)

**Vocabulary:**

* Scatterplot
* Association
* Explanatory Variable
* Response Variable
* Correlation Coefficient (r)
* Parameter
* Linear Model
* Predicted Value
* Residual
* Least Square
* Line of Best Fit
* Slope
* $\hat{y}$
* Mean-Mean point
* R2
* Extrapolation
* Outlier
* Leverage
* Influential Point

**Problems to Review:**

1. Use the given data to find the equation of the regression line. Round to 3 significant digits, if necessary. Managers rate employees according to job performance and attitude. The results for several randomly selected employees are given below.



1. A random sample of records of electricity usage of homes gives the amount of electricity used in July and size (in square feet) of 135 homes. A regression was done to predict the amount of electricity used (in kilowatt-hours) from size. The residuals plot indicated that a linear model is appropriate. What units does the slope have?
2. Using advertised prices for used Ford Escorts a linear model for the relationship between a car's age and its price is found. The regression has an R2 = 89.4%. Why doesn't the model explain 100% of the variation in the price of an Escort?
3. The relationship between the price of yachts (y) and their length (x) is analyzed. The mean length was 47 feet with a standard deviation of 10. The mean price was $87,000 with a standard deviation of 14,000. The correlation between the price and the length was 0.39. Use the given data to find the equation of the regression line. Round to 3 significant digits, if necessary.



1. Use the graph to the right this problem.
	1. Which point(s) are outliers?
	2. Which point(s) are leverage points?
	3. Which point(s) are influential?
2. The sales price (in thousands of dollars, 150 = $150,000) and size (in square feet) of 25 houses in Albuquerque, New Mexico shown in the plot below are from a random sample of records of resales of homes from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors. This type of data is collected by multiple listing agencies in many cities and is used by realtors as an information base.



**Regression Analysis: Price versus square feet**

The regression equation is

Price = - 27.0 + 0.0838 square feet

Predictor Coef SE Coef T P

Constant -27.03 12.76 -2.12 0.045

square feet 0.083834 0.006196 13.53 0.000

S = 14.4788 R-Sq = 88.8% R-Sq(adj) = 88.4%

1. Interpret (in context) what the slope of the equation means.
2. Interpret (in context) what the intercept of the equation means.
3. Interpret (in context) what means.
4. What do you predict a house of 2000 sq. ft. will sell for?
5. What do you predict a house of 5000 sq. ft. will sell for?
6. What problems, if any, are there with the predictions in parts d and e?
7. Given a house of 1600 sq. ft. sells for $109,000, what is the residual? Interpret the residual in this context.

**Experimental Design:** Chapter 4

**What to Study:** To be successful on this section of the final, the student will

1. **Sample Survey:**
* Selects units randomly from the population of interest for inclusion in the study
* Certain outcomes are measured.
* Makes no random assignment of them to treatments.
* The results of a sample survey can be used to identify associations among variables for the populations from which the units were randomly selected-not just for the people in the study, as in an observational study.
1. **Observational Study:**
* A type of study in which individuals are observed or certain outcomes are measured. No attempt is made to affect the outcome (for example, no treatment is given).
* No treatments
* No random selection of units from the population
* No random assignments of those units to treatments.
* Observes the characteristics of a group of units from one or more existing populations.
1. **Experiment**
* Deliberately imposes some treatments on individuals in order to observe their responses.
* Assigns all units randomly to treatments or treatments to trials.
* May or may not select units at random from the population
* Allows researchers to draw cause-and –effect conclusions for either the study’s units (if the experimental units were not randomly selected) or the population (if the experimental units were randomly selected).
* Three Principals of Experimental Design: control, randomize, and replicate

**Vocabulary:**

* outcome
* trial
* population
* sample
* sample survey
* biased
* randomization
* census
* parameter
* statistic
* simple random sample (SRS)
* sampling frame
* sampling variability
* homogeneous groups
* heterogeneous groups
* stratified random sample
* cluster sampling
* multistage sampling
* systematic sampling
* respondents
* voluntary response sample
* convince sampling
* undercoverage
* nonresponse bias
* response bias
* observational study
* retrospective study
* prospective study
* experiment
* subjects
* participants
* experimental units
* factors
* levels
* treatment
* block
* completely randomized experiments
* statistically significant
* control
* control group
* single-blind
* double-blind
* placebo
* placebo effect
* matching
* confounding

**Problems to Review:**

1. Which of the following is not required in an experimental design?
2. control
3. blocking
4. randomization
5. replication
6. All are required in an experimental design.
7. Which is not a critical part of designing a good experiment?
8. Random selection of subjects.
9. Random assignment of subjects to treatments.
10. Control of known sources of variability.
11. Replication of the experiment on a sufficient number of subjects.
12. All of these are important.
13. Hoping to get information that would allow them to negotiate new rates with their advertisers, Natural Health magazine phoned a random sample of 600 subscribers. 64% of those polled said they use nutritional supplements. Which is true?
14. The population of interest is the people who read this magazine.
15. "64%" is not a statistic; it's the parameter of interest.
16. This sampling design should provide the company with a reasonably accurate estimate of the percentage of all subscribers who use supplements.
17. I and III only
18. I only
19. I and II only
20. I, II, and III
21. II and III only
22. Jennifer is a quality control inspector for a well-known computer modem manufacturer. Jennifer oversees five assembly lines, each assembly line produces the same number of modems per day. Jennifer randomly selects one assembly line each morning and performs further sampling and quality control procedures on that assembly line's modems for the rest of the day. This morning, Jennifer randomly selected the modems of assembly line #5 for quality control inspection. Furthermore, Jennifer will then perform systematic sampling on that assembly line's modems. What has been excluded from being in the sampling frame today?
23. Some of the modems of assembly line #5
24. The modems of assembly lines #1, 2, 3, and 4
25. An unbiased sample for assembly line #5
26. A sampling method for assembly line #5
27. Nothing has been excluded from the sampling frame.
28. Management at a post office is curious about its services in one particular zip code area. Every address within that zip code, including business and residential, is mailed a survey inquiring about the post office's quality of service. The post office does receive a respectable 19% response rate to the survey. What, if any, is the most noticeable bias for this survey?
29. Nonresponse bias
30. Voluntary response bias
31. Undercoverage of the population
32. Response bias
33. There does not seem to be any bias.

**A member of the City Council has proposed a resolution opposing construction of a new state prison there. The council members decide they want to assess public opinion before they vote on this resolution. Identify the method that is proposed to sample local residents to determine the level of public support for the resolution.**

1. Place an announcement in the newspaper asking people to call their council representatives to register their opinions. Council members will tally the calls they receive.
2. convenience
3. systematic
4. cluster
5. voluntary response
6. simple (SRS)
7. Randomly select several city blocks; interview all the adults living on each block.
8. stratified
9. cluster
10. systematic
11. simple (SRS)
12. judgment
13. Call every 500th person in the phone book.
14. convenience
15. cluster
16. simple (SRS)
17. systematic
18. stratified
19. An education researcher was interested in examining the effect of the teaching method and the effect of the particular teacher on students' scores on a reading test. In a study, there are four different teachers (Juliana, Felix, Sonia, and Helen) and three different teaching methods (A, B, and C). The number of students participating in the study is 258. Students are randomly assigned to a teaching method and teacher. After a four-month period, students take a reading test and are given a score out of 10. Identify the levels of the factor "teaching method".
20. Method A, method B, method C
21. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
22. Teaching method and teacher
23. Score on reading test
24. Juliana, Felix, Sonia, and Helen
25. Listening to popular music while driving is dangerous?

*In a study of 7824 drivers from Chico, California, all who renewed their driver’s license in 2009, it was found that of the 2113 who listen to popular music while driving, 4.0% had been in an accident or gotten a ticket in the last two years. On the other hand, of the 5711 who listened to talk radio or nothing at all, just 2.3% had been in an accident or gotten a ticket in the last two years.*

* 1. Is this an observational study or an experiment?

***If you said it is an observational study, answer questions b-h. If you said it is an experiment, answer questions***

***i-o. Answer only one set of questions, leave the other set blank.***

***Questions for an observational study:***

* 1. Is this a retrospective or prospective study?
	2. Explain your answer to part b).
	3. Who were the subjects studied, and how were they selected?
	4. What is the parameter(s) of interest?
	5. State a potential lurking variable or bias.
	6. Can cause –effect be determined based on this study?
	7. Explain your answer to part g).

***Questions for an experiment:***

* 1. What is the factor? What are the levels of the factor?
	2. What is the response?
	3. Who were the subjects studied, and how were they selected?
	4. What is the parameter(s) of interest?
	5. Is this study blind (or double blind)?
	6. Can cause –effect be determined based on this study?
	7. Explain your answer to part n).

**Sampling Distributions:** Chapter 7

**What to Study:** To be successful on this section of the final, the student will

* Know that no sample fully and exactly describes a population; sample proportions and means will vary from sample to sample. That’s sampling error (or, better, sampling variability).
* Know how the Central Limit Theorem describes the behavior of a sample proportions and means, as long as certain assumptions and conditions are met.
* Understand that the variability of a statistic (as measured b the standard deviation of its sampling distribution) depends on the size of the sample. Statistics based on larger samples are less variable.
* Be able to use a sampling distribution model to make simple statements about the distribution of a proportion or mean under repeated sampling.
* Be able to interpret a sampling distribution model as describing the values taken by a statistic in all possible realizations of a sample or randomized experiment und the same conditions.

**Vocabulary:**

* Sampling distribution model
* Sampling variability (sampling error)
* Central Limit Theorem
* Sampling distribution model for a proportion
* Sampling distribution model for a mean

**Problems to Review:**

1. Assume that 25% of students at a university wear contact lenses. We randomly pick 200 students.
2. What is the mean and standard deviation of the proportion of students in this group who may wear contact lenses? Draw the appropriate graph.



1. What is the probability that the proportion for this group of 200 students that wears contact lenses is between 27% and 32%?
2. What is the probability that under 20% of the 200 students wears contact lenses?
3. The lengths of adult yellow-bellied sapsuckers are normally distributed, with mean of 8.5 inches and standard deviation of 0.2 inches.
4. Construct a model for this information.
5. For a randomly selected adult sapsucker, what is the probability that it is between 8.3 and 9.0 inches long?
6. How long is an adult sapsucker in the 80th percentile? Include Z score and graph.
7. A random sample of 20 adult sapsuckers is measured. Find the probability that their mean length is greater than 8.6 inches. Include graph.



1. You pay $10 and roll a die. If you get a five or six, you win $30. If not, you get to roll again. If you get a 5 or 6 on the second roll, you get your $10 back. Suppose you play this game 30 times. Describe the sampling distribution of your mean winnings. In particular, state whether the distribution of the sample mean is normal or approximately normal, and give its mean and standard deviation.
2. Researchers believe that 7% of children have a gene that may be linked to a certain childhood disease. In an effort to track 50 of these children, researchers test 950 newborns for the presence of this gene. What is the probability that they find enough subjects for their study?

**Inference: Confidence Intervals & Hypothesis Tests**

Chapters 8 - 10

**What to Study:** To be successful on this section of the final, the student will

* Correctly identify the different procedures for dealing with categorical and quantitative data.
	+ **Categorical Data**
		- Deals with proportions
		- Deals with the Normal Model
			* 1 – Prop Z
			* 2 – Prop Z
		- Know the assumptions and conditions for inference with categorical data
			* Random
			* 10% Condition
			* Large sample 🡪success/failure (np ≥ 10 and nq ≥ 10)
			* If comparing 2 categorical groups 🡪 Groups must be independent
	+ **Quantitative Data**
		- Deals with means
		- Deals with the Student’s T model
			* T Interval
			* T Test
			* Matched-Paired t-test
			* 2 – Sample Means
		- Know the assumptions and conditions for inference with quantitative data
			* Random
			* 10% Condition
			* Population has a Normal Model 🡪 Histogram or Normal Probability Plot
			* If comparing 2 categorical groups
				+ Groups independent 🡪 2 – Sample Mean
				+ Groups dependent 🡪 Matched Pair
* Understand the following that deals with Confidence Interval:
	+ Correctly calculate a critical value
	+ Correctly calculate a Standard Error and Margin of Error
	+ Correctly calculate any Confidence Interval
	+ Interpret a Confidence Interval.
		- I am \_\_\_\_\_% confident that the true **parameter of interest** lies between **lower value** and **upper value.**
* Understand the following that deals with Hypothesis Tests:
	+ 4 steps to writing up a Hypothesis Test
		- State
		- Plan
		- Do
		- Conclude
	+ How to write hypothesis for the different tests
	+ How to correctly construct a model for the test.
	+ Understand and correctly interpret the P-value.
	+ Understand and identify Type I and Type II Errors.

**Vocabulary:**

* Standard error
* Confidence interval
* One-proportion z-interval
* Margin of error
* Critical value
* Null hypothesis
* Alternative hypothesis
* Two-sided (tailed) alternative
* One-sided (tailed) alternative
* P-value
* One-proportion z-test
* Alpha level
* Statistically significant
* Significance level
* Type I error
* Type II error
* Power
* Effect size
* Variances of independent random variables
* Sampling distribution of the difference between two proportions
* Two-proportion z-interval
* Pooling
* Two-proportion z-test
* t-distribution
* t-table
* degrees of freedom
* one-sample t-interval
* one-sample t test
* two-sample t test
* pooled t-test
* matched-paired t-test
* matched-paired t-interval

**Problems to Review:**

1. The federal guideline for smog is 12% pollutants per 10,000 volume of air. A metropolitan city is trying to bring its smog level into federal guidelines. The city comes up with a new policy where city employees are to use city transportation to and from work. A local environmental group does not think the city is doing enough, and that the pollution percentage is greater than 12%. An independent agency hired by the city to run a test used a random sample size of 250 and found the pollutants at 15%. Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
2. As part of the Pew Internet and American Life Project, researchers surveyed a random sample of 800 teens and a separate random sample of 400 young adults For the teens, 560 said that they owned an iPod or MP3 player. For the young adults, this figure was 268. (According to Erik Erikson's stages of human development, a teen is a person aging from 13 – 19 and a young adult is a person aging 20 – 39.)

1. Is there a significant difference between the population proportions? Test an appropriate hypothesis and state your conclusion. Use a α = 0.05 as your significant level.
2. Regardless of your conclusion, state what a Type I Error would be for this situation.
3. Explain what the P-value means in this context.
4. Every year, each student in a nationally representative sample is given test in various subjects. This test is to determine how the US ranks in the world on education. Recently, a random sample of 9,600 twelfth-grade students from the United States were administered a multiple-choice US history exam. One of the multiple –choice question is below. (The correct answer is C)

*In 1935 and 1936 the Supreme Court declared that important parts of the New Deal were unconstitutional. President Roosevelt responded by threatening to*

1. *Impeach several Supreme Court justices*
2. *Eliminate the Supreme Court*
3. *Appoint additional Supreme Court justices who shared his views*
4. *Override the Supreme Court’s decision by gaining three-fourth majorities in both houses of Congress*

It is believed that 28% of all twelfth-grade students can get this question correct; however on this particular exam 2,772 of the 9,600 students got it correct.

1. Is this evidence of an increase on the correct responses on this question? Test an appropriate hypothesis and state your conclusion. Use a α = 0.05 as your significant level.
2. Regardless of your conclusion, state what a Type II Error would be for this situation and list a possible consequence.
3. How would your conclusion change is α = 0.01? or α = 0.10?
4. A potato chip manufacturer buys potatoes from two different suppliers, Riderwood Farms and Camberley, Inc. The weights of potatoes from Riderwood Farms are approximately Normally distributed with a mean of 175 grams and a standard deviation of 25 grams. The weights of potatoes from Camberley are approximately Normally distributed with a mean of 180 grams and a standard deviation of 30 grams. When shipments arrive at the factory, inspectors randomly select a sample of 20 potatoes from each shipment and weigh them. Let be the difference in the sample mean weight of potatoes from the two suppliers.
5. What is the shape of the sampling distribution of ? Why?
6. Find the mean of the sampling distribution. Show your work.
7. Find the standard deviation of the sampling distribution. Show your work.
8. Ashtyn and Olivia wanted to know if generic chocolate chip cookies have as many chocolate chips as name-brand chocolate chip cookies, on average. To investigate, they randomly selected 10 bags of Chips Ahoy cookies and 10 bags of Great Value cookies and randomly selected 1 cookie from each bag. Then, they carefully broke apart each cookie and counted the number of chocolate chips in each. Here are their results:

 Chips Ahoy: 17, 19, 21, 16, 17, 18, 20, 21, 17, 18

 Great Value: 22, 20, 14, 17, 21, 22, 15, 19, 26, 18

1. Construct and interpret a 99% confidence interval for the difference in the mean number of chocolate chips in Chips Ahoy and Great Value cookies.
2. Does your interval provide convincing evidence that there is a difference in the mean number of chocolate chips?
3. Researchers studying the acquisition of pronunciation often compare measurements made on the recorded speech of adults and children. One variable of interest is called "voice onset time" (VOT), the length of time between the release of a consonant sound (such as "p") and the beginning of an immediately following vowel (such as the "a" in "pat"). For speakers of English, this short time lag can be heard as a period of breathiness between the consonant and the vowel. Here are the results for some randomly-selected 4-year-old children and adults asked to pronounce the word “pat.” VOT is measured in milliseconds and can be either positive or negative.

|  |  |  |  |
| --- | --- | --- | --- |
|   | *n* | $$\overbar{x}$$ | *s* |
| Children | 10 | 60.67 | 39.89 |
| Adults | 20 | 88.17 | 24.74 |

You are interested in whether there is a difference in the VOT of adults and children, so you plan to test $H\_{0}: μ\_{A}-μ\_{C}=0$ against $H\_{0}: μ\_{A}-μ\_{C}\ne 0$, where $μ\_{A}$ and $μ\_{C}$ are the population mean VOT for adults and children, respectively.

1. What additional information would you need to confirm that the conditions for this test have been met?
2. Assuming the conditions have been met, calculate the test statistic (t-score) and P-value for this test. Interpret the P-value in the context of this study, and draw the appropriate conclusion at the α = 0.05 level.
3. Given your conclusion in part (b), which type of error, Type I or Type II, is it possible to make? Describe that error in the context of this study.