

# AP STATISTICS

# AP EXAM STUDY GUIDE

## Table of Contents

**Topic 1:** Sampling Techniques and Sources of Bias

**Topic 2:** Experimental Design

**Topic 3:** Analyzing Data

**Topic 4:** Normal Distributions and Z-Scores

**Topic 5:** Probability Rules

**Topic 6:** Probability Distributions

**Topic 7:** Binomial and Geometric Distributions

**Topic 8:** Sampling Distributions

***Addendum:** Procedures for running Confidence Intervals and Significance Tests. **Memorize these pages!!***

**Topic 9:** Confidence Intervals

**Topic 10:** Significance Tests

**Topic 11:** Chi-Squared Tests and Types of Error

**Topic 12:** Bivariate Data

**Topic 13:** Confidence Intervals and Significance Tests with Bivariate Data

## **Additional Topics**

**BONUS Topic:** Advice for the AP Exam (from someone who's passed six of them)

### **You are responsible for...**

- Completing this study guide (5 points per topic)
- Completing the Practice Problems (5 points per topic)
- *Studying hard and doing your best!*

**Topic 1: Sampling Techniques and Sources of Bias (Notes: 1.1 and 1.2)**

1. Know and understand the difference between a *population* and *sample*
  - How is each one measured (what do we use to measure them)?
  
  - Why do we often measure samples instead of populations?
  
2. Know the different types of *bias* and how to spot them in different situations
  - **Bias** is anything that causes a sample to be **not representative of the population of interest**
    - You must be able to articulate what the bias is, why it should be considered bias, and how it distorts the results from what they otherwise might be.
  - What is the difference between *sampling error* and *sampling bias*?
  
  - How can a small sample size affect the validity of the sample? (*this is related to sampling error rather than bias*)

Define the types of <b>sampling bias</b> (a bias in <i>who</i> was in the sample)	Define the types of <b>response bias</b> (a bias in <i>what</i> the sample is saying)
<p>Undercoverage</p> <p>Nonresponse bias</p> <p>Voluntary response bias</p>	<p>Loaded Questions</p> <p>False answers</p>

3. Know the different types of sampling techniques and how to identify which one is being used (as well as the *advantages* and *disadvantages* of each)

Simple Random Sample (SRS)	Stratified Random Sample  <i>*Stratifying will <b>reduce variability</b> of possible sample results!</i>
Systematic Random Sample	Cluster Sample
Multistage Sample	Convenience Sample

4. Know how to design a random sampling procedure

- **Random number generator** will be your friend!
- “Describe a method...” (NOTE: blanks will be filled in with the context of the problem!)
  - **START WITH:** Assign each \_\_\_\_\_ (unit, subject, etc.) a *different* number between \_\_\_\_ and \_\_\_\_
  - Describe how you will implement the sampling method you want to use
  - Randomly select \_\_\_\_\_ numbers, ignoring repeats, and include the \_\_\_\_\_ (unit, subject, etc.) that corresponds with those numbers in your sample.

**Example:** Mr. Frederick wants to create an advisory committee of 20 randomly-selected students out of the 1,950 students at Grant. Describe how he could do so using a...

Simple random sample	Systematic Random Sample
Stratified Random Sample	Cluster Sample
Multistage Sample	Convenience Sample

## Topic 2: Experimental Design (Notes: 1.3)

### 1. Know the vocabulary of experiments and experimental design

- What is the difference between an Experiment and an Observational Study? Which one lets us establish cause-and-effect relationships? **HINT:** *There is one “dead giveaway” keyword when identifying an experiment. It starts with the letter A.*
- Define *Treatment* –
- Define *Confounding* –
- Define *Experimental Units* (*Subjects* when human) –

### 2. Know the four principles of a good experiment

- 
- 
- 
- 

### 3. Know methods for **controlling** an experiment to prevent bias

- Control group (what is it, and what does it allow us to do?)  
(**NOTE:** A control group is **NOT** mandatory; it is just one way to get comparison, which **IS** mandatory)
- Placebo effect –

- Blind study –
- Double-blind study –

4. Know the different types of experimental design and how to identify which one is being used (as well as the *advantages* and *disadvantages* of each)

- Completely Randomized Design
- Randomized Block Design (“Blocking”)
- Matched Pairs Design

5. Be able to discuss *generalizability* – the extent to which the results of a sample (or experimental group) can be applied to a certain population

- You can generalize to the population *from which the sample or experimental group was taken*
- **BIAS** can hurt (or even eliminate) generalizability. You need **RANDOMNESS** to avoid this!
  - For example, a study that consists of **volunteers** should only be generalized to those volunteers! You *might* be able to generalize to “people who are similar to the volunteers,” but absolutely no further, because they weren’t *randomly selected*!
  - **NOTE:** *Even a relatively small sample size (not ridiculously small, but somewhat small) can be valid as long as it’s random!*

**Example:**

*A researcher studied a random sample of 100 teens in Oklahoma. To which populations will the results of this researcher’s findings be generalizable? (Circle ALL that apply)*

- A. The 100 Oklahoma teens in the study
- B. All teens in Oklahoma
- C. All teens
- D. All Oklahomans

### Topic 3: Analyzing Data (Notes: 2.1, 2.2)

1. The 5 things you should discuss when analyzing a **distribution** of data:

*NOTE: If asked to compare data sets, make sure you explicitly compare them (For example, “The first distribution has a greater mean than the second distribution, while the second distribution has a greater spread than the first”)*

#### 2. Center

Measure	How to find it	Resistant to the effects of outliers?
Mean Population: $\mu$ Sample: $\bar{x}$		
Median		

- The best one to use is usually \_\_\_\_\_, unless the data is skewed, at which point \_\_\_\_\_ should be used

#### 3. Shape

Shape	Normal	Skewed Left	Skewed Right	Uniform	Bimodal
Sketch					
Which is greater, mean or median? (or are they = )					

#### 4. Spread

Measure	Paired with... (mean or median)	How to find it	Resistant to the effects of outliers?
Standard Deviation Population: $\sigma$ Sample: $s$		$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$ $s = \sqrt{\frac{\Sigma(x-\mu)^2}{n-1}}$ <b>Or use 1-Var Stats!</b>	
Variance Population: $\sigma^2$ Sample: $s^2$		$\sigma^2 = \frac{\Sigma(x-\mu)^2}{n}$ $s^2 = \frac{\Sigma(x-\mu)^2}{n-1}$ <b>Or use 1-Var Stats!</b>	
Lower Quartile (Q1)		Midpoint of Minimum and Median <b>Or use 1-Var Stats!</b>	
Upper Quartile (Q3)		Midpoint of Median and Maximum <b>Or use 1-Var Stats!</b>	

Range			
Interquartile Range (IQR)			

5. **Outliers** (You may ALSO want to point out gaps, clusters, and any other “interesting” features a data set may have)

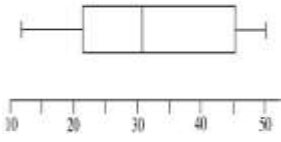
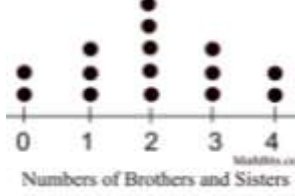
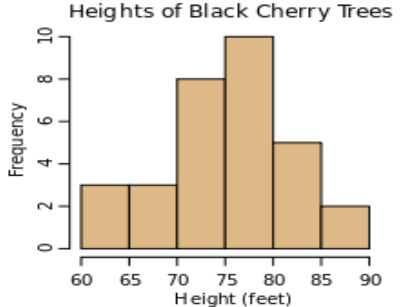
- *What is an outlier?*
- **NOTE:** An outlier **CAN** change the value of the Median, Q1, Q3, etc. if the addition of an outlier causes the *position* of numbers to change. However, this change will *usually* be slight
- *How to identify outliers: IQR TEST* (remember, this is a *general guideline*, not a strict rule!)

How it works:

**Example:** Min = 11, Q1 = 32, Med = 36, Q3 = 44, Max = 51

Any point *below* \_\_\_\_\_ or *above* \_\_\_\_\_ can be considered an outlier. **Outliers in this data set:**

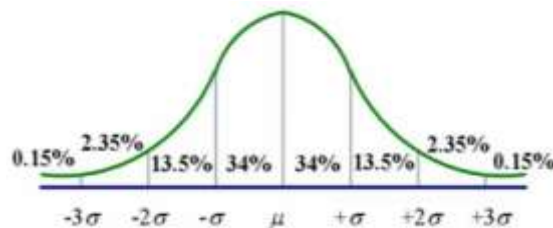
### 6. Graphs

Boxplot	Stemplot	Dotplot	Histogram
	<pre> stem   leaf ----- -----  6     9  7      8     7 8 8 9  9     0 6 7 7 10     0 Key: 6   8 means 68 </pre>		
<p><i>Notes:</i></p> <ul style="list-style-type: none"> <li>• Min, Q1, Med, Q3, Max</li> <li>• <b>Cannot show shape</b> (but <i>can</i> show skews)</li> <li>• <b>Outliers</b> should be marked with a *</li> </ul>	<p><i>Notes:</i></p> <ul style="list-style-type: none"> <li>• Remember to give a <i>key</i> to show what the numbers mean</li> <li>• <b>Do not skip stems</b></li> <li>• If given a <b>back to back</b> stemplot, <i>always</i> read stem first, then leaf</li> </ul>	<p><i>So easy a caveman could do it!</i></p>	<p><i>Notes:</i></p> <ul style="list-style-type: none"> <li>• X-axis shows <i>intervals</i>, y-axis shows the <i>frequency</i> (number of data points that belong in that interval)</li> <li>• <b>Finding the median:</b> Figure out how many data points there are, use <math>\frac{n+1}{2}</math> to find the <i>position</i> of the median, then figure out which interval contains that position!</li> </ul> <p><b>EXAMPLE:</b>  Number of data points: _____   Position of median: _____   Interval containing median: _____</p>
<p>This data point is 24, <b>NOT</b> 42</p>	<pre> Boys   Girls ----- -----  7     0  1     1 146    2 268  458    3 34466889 122289   4 436  3479   5 4 </pre>		

## Topic 4: Normal Distributions and Z-Scores (Notes: 2.3)

1. Know how to analyze a normal distribution

- *THEORETICAL* distribution (in reality, we consider data to be \_\_\_\_\_ normal)
- It's like a **histogram** in which the center is the \_\_\_\_\_ and the intervals are each one \_\_\_\_\_



2. Know how to use the **Empirical Rule**

- About \_\_\_\_\_% of data is within 1 Standard Deviation of the mean
- About \_\_\_\_\_% of data is within 2 Standard Deviations of the mean
- About \_\_\_\_\_% of data is within 3 Standard Deviations of the mean

3. Know how to calculate and interpret **z-score**

- A data point's z-score is the \_\_\_\_\_
- **Formula** (NOT in AP exam):  $z = \frac{x - \mu}{\sigma}$
- Z-scores can help us compare two **unlike** measurements  
*Example: Suppose the weights of apples are normally distributed with a mean of 85 grams and a standard deviation of 8 grams. The weights of oranges are also normally distributed with a mean of 131 grams and a standard deviation of 20 grams. Amy has an apple that weighs 90 grams and an orange that weighs 155 grams.*

1. Calculate **and interpret** the z-score of Amy's apple

2. Which is *relatively* larger, Amy's apple or her orange? **Explain.**

3. How large would Amy's apple have to be in order to be comparable to her orange?

4. Know how to use Z-scores to calculate the percentage of data points above, below, or between certain boundaries

*\*This works **ONLY** for normally-distributed data!! **DO NOT** do these procedures if you do not **know** that your data is normally distributed!*

With Z-table	With Calculator
<ul style="list-style-type: none"> <li>• Z-table gives the percentage of values <b>below</b> a given z-score</li> <li>• You can use the z-table <b>backwards</b> – if you know the percentage, find it on the z-table, and see what z-score it equates to!</li> </ul>	<ul style="list-style-type: none"> <li>• NormalCDF (if <i>looking for</i> percentage/probability)</li> <li>• InvNorm (if <i>given</i> percentage or probability)</li> <li>• To adequately <i>show work</i>, you must write...</li> </ul>



## Topic 5: Probability Rules (Notes: 3.1)

### 1. Understand what probability *is*

- How do you calculate the probability of an outcome?
- What is the Law of Large Numbers?
- What are *mutually exclusive* outcomes?
- What are *independent* events?
- Why can two events that are mutually exclusive *never* be independent?

### 2. Know the basic rules of probability

- When calculating the probability of getting more than one outcome for a given event, what formula should you use? **HINT:** Always account for any overlap between outcomes!
- When analyzing events with multiple outcomes, what visual aide will be the most beneficial?
- When calculating the probability of *multiple* events, what rule or formula should you use?
- When, and *how*, do you use the *combinations* ( $nCr$ ) function in your calculator?

- When analyzing a series of multiple events, each with multiple possible outcomes, what visual aide will be helpful?
- When calculating the probability of multiple independent events, what three things should you account for? **HINT:** *The formula on the formula sheet may help you!*
- How does the above procedure change when the events are dependent?
- What is *conditional* probability, and how do you calculate the conditional probability of a given event?

Situation	Rule	Formula
“At least one”	Opposite of “none”	$1 - P(0)$
Multiple outcomes – mutually exclusive	Add probabilities	$P(A \cup B) = P(A) + P(B)$ <b>NOTE:</b> $P(A \cap B) = 0$ (no overlap for mutually exclusive events)
Multiple outcomes – NOT mutually exclusive	Add probabilities but <b>subtract the overlap</b> *If using a Venn Diagram, just add up the 3 sections in the diagram	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiple events – Independent	Multiply probabilities, and account for COMBINATIONS in which these events can occur (nCr)	$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ $nCr \cdot (p_{\text{success}})^{\# \text{ of successes}} \cdot (p_{\text{fail}})^{\# \text{ of fails}}$
Multiple events – Dependent	Multiply probabilities *Account for the <b>change</b> in probability with each trial *Account for <b>combinations</b> (nCr)	$nCr \cdot p_{\text{event 1}} \cdot p_{\text{event 2}} \cdot p_{\text{event 3}} \dots$ <b>NOTE:</b> Remember these probabilities CHANGE!!
Conditional Probability (A <u>given</u> B)	$\frac{\text{Probability of both events}}{\text{Probability of first event}}$	$P(A B) = \frac{P(A \cap B)}{P(B)}$

**Topic 6: Probability Distributions (Notes: 3.2, 3.3)**

1. Know the different types of random variables and how their distributions work

- What is the difference between a discrete and a continuous random variable?
- For continuous random variables, what is the probability of getting *exactly* one given outcome? \_\_\_\_\_
- How do you calculate the **expected value** of a discrete random variable?
- What is the **definition** of expected value? (It mean something *very specific*)
- What formula can you use to calculate the spread (st. dev.) of a discrete random variable *by hand*?
- How are variance and standard deviation related?

2. Know how transforming and combining a random variable changes that variable’s distribution

Action	Effect on <b>Center</b> (mean)	Effect on <b>Spread</b> (standard deviation)
Adding/Subtracting a <b>CONSTANT</b> (number)		
Multiplying/Dividing by a <b>CONSTANT</b> (number)		
Combining (adding or subtracting two random variables to each other)		

**HINTS:**

- If  $X$  and  $Y$  are normally distributed, so are  $X + Y$  and  $X - Y$ . This means **use normalCDF!**
- $X > Y$  is the same as  $X - Y > 0$  (likewise,  $X < Y$  is the same as  $X - Y < 0$ )

**Topic 7: Binomial and Geometric Distributions (Notes: 3.4)**

1. Know and understand how to use a Binomial Distribution

• **Using the Binomial Distributions**

– Only works in *binomial* settings, which occurs when the following conditions are met (“BINS”)

- **B:** \_\_\_\_\_
- **I:** \_\_\_\_\_
- **N:** \_\_\_\_\_
- **S:** \_\_\_\_\_

– BinomPDF: finds \_\_\_\_\_

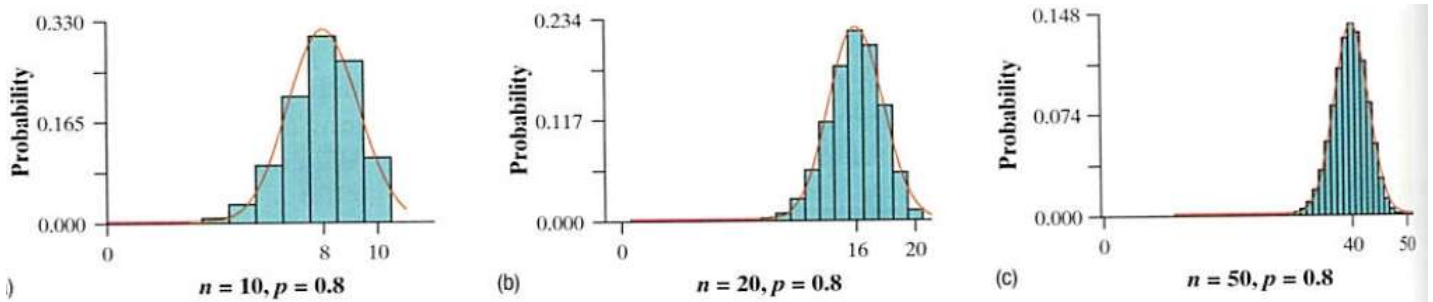
– BinomCDF: finds \_\_\_\_\_

• **Binomial Curve**

– CENTER: \_\_\_\_\_ (number of trials • probability of success = expected # of successes)

– SPREAD: Standard Deviation,  $\sigma =$  \_\_\_\_\_

– SHAPE: Approaches **normality** if you can *expect* at least \_\_\_\_\_ successes and \_\_\_\_\_ failures



**Example:**

Genetics says that children receive genes from each of their parents independently. Each child of a particular set of parents has probability a probability of 0.25 of having Type O blood. Suppose these parents have 6 children. Let  $X =$  the number of children with Type O blood.

a. Calculate the mean and standard deviation of the number of children who will have Type O blood

b. Find the probability of each of the following

$P(X = 4)$ ; exactly 4 children will have Type O blood

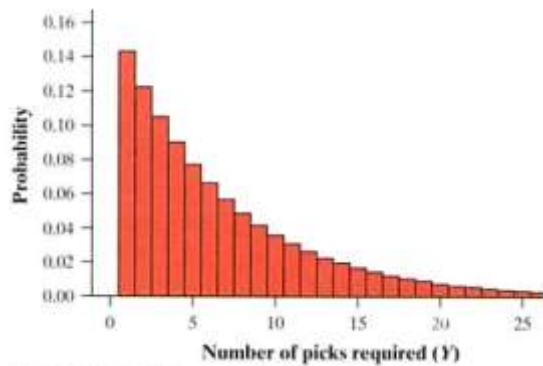
$P(X \leq 3)$ ; 3 or fewer children have Type O blood

$P(X > 1)$ ; More than 1 child will have Type O blood

$P(X \geq 3)$ ; 3 or more children will have Type O blood.

2. Know and understand how to use a Geometric distribution

- Geometric Distribution – a density curve that allows us to determine how many trials it will take to get \_\_\_\_\_ (also think of it as \_\_\_\_\_)
  - Events need to be \_\_\_\_\_ (of course)
- How to calculate it
  - **Calculator**
    - GeometPDF is used for \_\_\_\_\_, the probability that the first success will happen \_\_\_\_\_ the  $K^{\text{th}}$  trial
    - GeometCDF is used for \_\_\_\_\_, the probability that the first success will happen \_\_\_\_\_ the  $K^{\text{th}}$  trial
    - Type in \_\_\_\_\_ and \_\_\_\_\_
- EXPECTED VALUE (mean) of a Geometric Random Variable is \_\_\_\_\_ (If  $n = \frac{1}{p}$ , then  $np = 1$ )
- Shape is always \_\_\_\_\_
  - As you continue, the probability of having \_\_\_\_\_ gets \_\_\_\_\_



**Examples:**

1. A slot machine has a win rate of 8%. A gambler wants to play at this slot machine until they win – then, they will leave.

a. What is the expected number of games the gambler will have to play in order to win? \_\_\_\_\_

b. Find the probability that it will take the gambler...

7 spins to win

10 or fewer spins

More than 20 spins

## Topic 8: Sampling Distributions (Notes: 4.1)

1. Know the basics of *sampling distributions*

- What is the difference between a *parameter* and a *statistic*?
- What is the difference between a *proportion* and a *mean*?
- What is a *sampling* distribution?
- Know the difference between a sample distribution and a sampling distribution
  - Sample distribution – a graph of data taken from one sample
  - Sampling distribution – a graph of statistics taken from multiple samples

2. Know the importance of the **Central Limit Theorem** (define it below)

3. Know how to analyze a **normal distribution**, and use it to find the probability of a sample statistic occurring, *given* an assumed population mean and standard deviation

- What function in the calculator should we use to do this? \_\_\_\_\_

**From the AP Formula Sheet:**

If  $X$  has a binomial distribution with parameters  $n$  and  $p$ , then...

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

If  $\bar{x}$  is the mean of a random sample of size  $n$  from an infinite population with mean  $\mu$  and standard deviation  $\sigma$ , then...

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- **REMEMBER:** These formulas are for **CONVERSION** from the population standard deviation! If you're already given the standard deviation of the sampling distribution, just use it!

4. Know the **CONDITIONS** that must be met for the Central Limit Theorem to apply, and thus for **inference** to occur

Condition	How to meet the condition	Ensures _____ of the <b>sampling distribution</b> is appropriate for inference (center, shape, or spread)
1.	<p><b>For Proportions:</b></p> <p><b>For Means:</b></p>	<p><i>NOTE: If the <u>population</u> has an approximately normal distribution, this condition can be considered “met” regardless of sample size!</i></p>
2.		
3.		

What must we do if the conditions are not met? \_\_\_\_\_

### PROCEDURES FOR CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS

#### 1. State what you're doing

<p>Confidence Intervals</p> <ul style="list-style-type: none"> <li>• Procedure you're using</li> <li>• The <i>parameter</i> (population) of interest!</li> <li>• Confidence level</li> </ul> <p>“We will use a _____ Interval to estimate, with _____% confidence, the <i>true</i> (mean/proportion) of _____ (context) _____”</p>	<p>Significance Tests</p> <ul style="list-style-type: none"> <li>• Procedure you're using</li> <li>• The <i>parameter</i> (population) of interest!</li> <li>• Hypotheses, <math>H_0</math> and <math>H_a</math></li> <li>• Significance Level, <math>\alpha</math> (If none is given, use .05)</li> </ul> <p>“We will use a _____ Test to test the following hypotheses at the <math>\alpha =</math> _____ level”</p>
--	--

*Additional Notes:*

- Remember,  $H_0$  implies “no change” or “no difference”
- If you are doing a 2-Sample or 2-Proportion test, state **both** populations – indicate which one is which!
- For a **Paired** t-test, find the *difference* between the matched pairs, and use these *differences* as your one sample!  **$H_0$ :**  $\mu_{\text{Difference}} = 0$ ,  **$H_a$ :**  $\mu_{\text{Difference}}$  is  $>$ ,  $<$ , or  $\neq 0$

#### 2. Check your conditions

**NOTE:** If a problem says “*assume conditions are met*”, you do not have to go through this process!!

- Sample Size (also known as “Large Counts”)
  - **If met, the SHAPE of the sampling distribution is Normal (or  $\chi^2$  distribution for  $\chi^2$  tests)**
  - *Means* ( $\mu$ ):
    - 30 or more, OR
    - Graph of the sample shows no obvious skews or outliers (**t-test only**), OR
    - Population is *known* to be normal
  - *Proportions* ( $p$ ):
    - At least 10 expected successes and 10 expected failures (find *expected* value of each)

- Randomness
  - Ensures that the **CENTER (the sample statistic) is legitimate**
  - *Samples and Observational Studies*: Randomly selected from the population
  - *Experiments*: Randomly assigned into treatment or control group(s)
  - **Note**: If you are running a 2-sample interval or test, you must check and STATE that both samples are random!
- Independence
  - Ensures that the **SPREAD (the standard deviation) formulas that you're given are reliable**
  - *Samples and Observational Studies*: sample must be *less* than 10% of the population
  - *Experiments*: Groups should be independent of each other (i.e. not matched pairs)
    - If there ARE matched pairs, do a PAIRED t-test; find the *difference* between each pair and use *those* numbers in a 1-sample t-test!

### 3. Do the calculation (create the interval or run the test)

<p><i>Confidence Intervals</i></p> <ul style="list-style-type: none"> <li>• Re-state <i>type</i> and <i>confidence level</i> (just to be safe)</li> <li>• Give interval: (lower, upper)</li> </ul>	<p><i>Significance Tests</i></p> <ul style="list-style-type: none"> <li>• Test Statistic (z, t, or <math>\chi^2</math>)</li> <li>• Degrees of Freedom (t and <math>\chi^2</math> ONLY)</li> <li>• <i>p-value</i></li> </ul>
--	---

### 4. State your conclusion

<p><i>Confidence Intervals</i></p> <ul style="list-style-type: none"> <li>• Give the % confidence</li> <li>• Give the interval <i>in context</i> (including <b>PROPER UNITS</b>)</li> </ul> <p>“I am _____% confident that the <i>true mean</i> (or <i>true proportion</i>) of _____ (<i>context</i>) _____ is between _____ and _____.”</p>	<p><i>Significance Tests</i></p> <ul style="list-style-type: none"> <li>• State whether <math>p &lt; \alpha</math> (reject) or <math>p &gt; \alpha</math> (fail to reject)</li> <li>• Give the consequences <i>in context</i></li> <li>• <b>Chi-Squared</b>: You may be asked to perform a follow-up analysis to see where the biggest gaps between observed and expected values are.</li> </ul> <p>REJECT: “Because <math>p &lt; \alpha</math>, we can reject <math>H_0</math>. There is statistically significant evidence to suggest _____ (<i>whatever <math>H_a</math> was</i>) _____</p> <p>FAIL TO REJECT: “Because <math>p &gt; \alpha</math>, we fail to reject <math>H_0</math>. There is NO statistically significant evidence to suggest _____ (<i>whatever <math>H_a</math> was</i>) _____</p>
--	---

**IMPORTANT:** The p-value is **ALWAYS** between \_\_\_\_\_ and \_\_\_\_\_. If your calculator gives something *other* than this, I *guarantee* there will be an E at the end. This represents *scientific notation* ( $\# \cdot 10^x$ ). This means your p-value is **very small** (in fact, many statisticians just write “ $p < .001$ ” and call it a day). **As far as we're concerned, p-values this low will always be significant!**

**ALSO IMPORTANT:** Know the difference between “interpret the p-value” and “draw conclusions”

- **Interpretation:** IF  $H_0$  is true, the probability that we would get a test statistic as or more extreme as the one we got in our sample (by random chance) is \_\_\_\_\_ (*p-value*)
  - **NOTE:** If there is a *direction* involved ( $<$  or  $>$ ), state that direction (“as high or higher” or “as low or lower”)
- **Draw conclusions:** Rejecting or Failing to Reject  $H_0$  (and associated context)



**Topic 9: Confidence Intervals (Notes: 4.3, 4.4, 4.5)**

1. Understand the purpose of confidence intervals and how they work

- What does a confidence interval allow us to do?
  
- How do we *interpret* a confidence interval? (For instance, to interpret 95% confidence level, what *words* would you say?)
  
- How do we interpret a confidence *level*? (For instance, in a 95% confidence interval, what does the 95% tell us? What does it *mean* to be “95% confident”?)
  
- Know how to use the FORMULA for confidence interval:
  - **Statistic  $\pm$  Critical Value • Standard Deviation of Statistic**
  - **Critical Values** can be found in the *t* table (for *z* distributions, use the \_\_\_\_\_ row)
  - **Standard Deviation:** Use the formula sheet (they are *very* clearly laid out!)
    - *In this context*, St. Dev. of the Sampling Distribution is also called **Standard Error**
- What is the margin of error, and how do we calculate it?

2. Know what type of confidence interval to calculate, and *when* to calculate it

When estimating a <b>population proportion</b>	When estimating the <i>difference</i> between two <b>population proportions</b>
When estimating a <b>population mean</b> and the population standard deviation is <i>known</i> ( <b>RARE</b> )	When estimating the <i>difference</i> between two <b>population means</b> and the population standard deviations are <i>known</i> ( <b>RARE</b> )
<i>Note: The true name of this procedure and the calculator name are slightly different. Know <b>both!</b></i>	

When estimating a **population mean** and the population standard deviation is **NOT known**

When estimating the ***difference between two population means*** and the population standard deviations are **NOT known**

*Note: The true name of this procedure and the calculator name are slightly different. Know **both!***

3. Know the essentials of the *t*-distribution

- When do we use it?
- How do we calculate the *degrees of freedom* of a *t*-distribution?

4. Know the **four-step process** of statistical inference (in this case, creating a confidence interval)

- 1.
- 2.
- 3.
- 4.

5. Know how to *check conditions*

- What conditions must you check, and where in the study guide can you look to find them?
- If dealing with a *t*-distribution and your sample size is not 30 or more, what *other* methods can you use to check for normality? **Be specific!**

## 6. Know how to *manipulate* confidence intervals

- Be able to solve for  $n$  or  $z^*$  (or  $t^*$ ) (**NOTE:** In multiple choice, you can always *plug in* the choices!)
    - If a sample proportion is not given in this case, assume  $p = \rule{1cm}{0.4pt}$  (this gives us the greatest margin of error to work with)
  - Remember that the sample statistic (“point estimate”) is in the  $\rule{1cm}{0.4pt}$  of the confidence interval (and that the distance between the sample statistic and the ends of the confidence interval is the  $\rule{1cm}{0.4pt}$ )
  - Know what happens to the margin of error (and thus *width* of the confidence interval) if we...
    - Increase sample size:
    - Decrease sample size:
    - Increase confidence level:
    - Decrease confidence level:
  - If you adjust sample size, confidence interval changes by the **square root** of that amount (since  $n$  is inside the square root in all standard deviation formulas)
    - **Example:** What will happen to the confidence interval if you multiply the sample size by 4?
- 

## Topic 10: Significance Tests (Notes: 4.2, 4.5)

### 1. Understand what significance tests are for and allow us to do

- What are the two types of hypotheses used in significance tests, and what *symbols* do we use to represent them?
- What is a *null hypothesis*, and what does the null hypothesis *always* assume to be true?  
**NOTE:** The answer is *slightly* different for 1-sample and 2-sample tests – know **both!**
- What is an *alternative hypothesis*? What are the 3 types of alternative hypotheses you could have?  
**NOTE:** The answer is *slightly* different for 1-sample and 2-sample tests – know **both!**

- Significance levels (alpha-levels) determine the p-value below which a test's results should be considered significant. If no alpha level is given, it is a good *general* rule to use \_\_\_\_\_

2. Know what type of significance test to run, and *when* to run it

When testing a claim about a <b>population proportion</b>	When testing a claim about the <b><i>difference between two population proportions</i></b>
When testing a claim about a <b>population mean</b> and the population standard deviation is <u>known</u> ( <b>RARE</b> )	When testing a claim about the <b><i>difference between two population means</i></b> and the population standard deviations are <u>known</u> ( <b>RARE</b> )
When testing a claim about a <b>population mean</b> and the population standard deviation is <u>NOT known</u>	When testing a claim about the <b><i>difference between two population means</i></b> and the population standard deviations are <u>NOT known</u>
When testing a claim about a study or experiment that utilizes <i>matched pairs</i>	<i>In the calculator, which type of test would you select?</i>

3. Know how to *interpret* the results of a significance test

- What two (for t-tests, three) things should you report after running a significance test in your calculator?
- How do you *interpret* a p-value? What does that number *mean*?
- How do you analyze (interpret the results of) a test for which the p-value is *less* than alpha (for instance,  $p < .05$ ). *What would you write?*
- How do you analyze (interpret the results of) a test for which the p-value is *greater* than alpha (for instance,  $p > .05$ ). *What would you write?*

**Topic 11: Chi-Squared Tests and Types of Error (Notes: 4.6 and 4.7)**

1. Know the similarities and differences between Chi-squared and the other types of significance tests

- When do we use Chi-squared tests? In other words, what do Chi-squared tests allow us to measure?
- What are the three types of chi-squared tests, and when do we use each?

Type	Purpose/When to use	Name in Calculator

*NOTE: The biggest difference between the second and third type is context. Other than that, they are essentially the same.*

- What are the null and alternative hypotheses of a Chi-squared test?

2. Know the *conditions* of a Chi-Squared test

- Same conditions as other significance tests
- How is the *sample size* condition different for Chi-Squared tests, and how do we check it?

### 3. Know how to calculate and interpret the Chi-squared statistic

- How can we find *expected counts*?
  - Goodness-of-fit: **READ THE PROBLEM!**
    - Sometimes, you may *expect* certain proportions out of a total (like we did with M&Ms).
    - Sometimes, you may *expect* that the data is *equally distributed* among the categories (in this case, just use simple division!)
  - Homogeneity and Independence: *What formula do we use to calculate each expected value?*
- How do we calculate *degrees of freedom* for a chi-squared test?
  - Goodness-of-Fit:
  - Homogeneity and Independence:
- When running a Chi-Squared test, what three things must you report? **NOTE:** *The interpretation and analysis/drawing conclusions aspects of these are the same as the other significance tests.*

---

4. Know what Type I and Type II error are; be able to spot them in context, *and* discuss what the *consequences* of these types of error would be if they happened in a real-life situation (including possibly evaluating which one would be worse in that situation)



**HINT:** *The chart on your 5.4 notes may be a handy tool to help you understand and remember which is which!*

- What is a Type I error?
- What is a Type II error?
- What variables are used to represent the probability that Type I error and Type II error, respectively, will happen?

5. Know what *power* is, why it's important, and how it can be influenced

- What is the definition of *power*?
- How is power calculated?
- How can power be *increased*? List 3 ways.

6. Understand the relationship between Power, Type I Error, and Type II error

Power	Type I Error ( $\alpha$ )	Type II Error ( $\beta$ )
<i>Increases</i> 		
<i>Decreases</i> 		

Fill in each of the following blanks with either "same" or "opposite"

Type I and Type II error always go the \_\_\_\_\_ direction

Power and Type I error always go the \_\_\_\_\_ direction

Power and Type II error always go the \_\_\_\_\_ direction

Suppose you want to avoid a *Type I* error at all costs. Should you use a significance level of .10, .05, or .01? Explain.

## Topic 11: Bivariate Data (Notes: 5.1, 5.2)

### 1. Know how to analyze a correlation between two variables

- Explanatory and Response variables (which one is x and which one is y?)

- 5 things we should look for in bivariate data:

<i>Characteristic</i>	<i>Possibilities</i>	<i>What the r-value tells us</i>
<b>Shape</b>		R-value assumes that shape is...
<b>Strength</b>		
<b>Direction</b>		
<b>Outliers</b> (especially if they substantially alters the equation of the <i>regression line</i> , or line of best fit)		
<b>Context</b> (as always) – what two variables are we examining?		

- X and Y are correlated. Does this mean that X *causes* Y? If not, what else might be going ON?

### 2. Know how to analyze the least-squares regression line (line of best fit): $\hat{y} = mx + b$

- $\hat{y}$  is the \_\_\_\_\_ value of y for a given value of x

- *Interpretation of Slope:*

- *Interpretation of Y-intercept:*

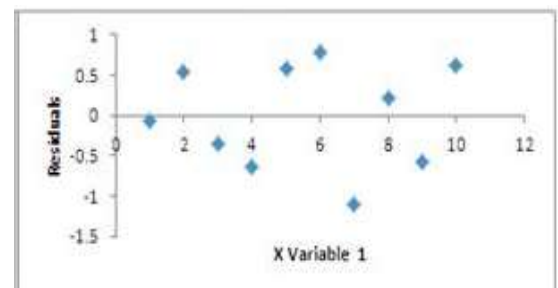
- $r^2$  value (“**coefficient of determination**”)

- Extrapolation

### 3. Know how to analyze *residuals* and *residual plot*

- What *is* a residual?

*Residual Plot*





- How do you calculate a residual?
- What information does a residual plot give you?

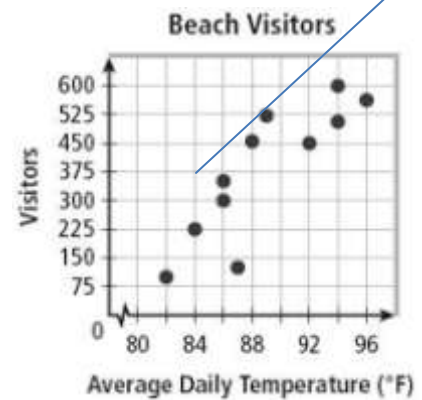
4. Know how to handle *curved* data (linear transformations)

- Make sure that all **interpretations** (see above) take all transformations into account!

Exponential Regression	<b>Equation Format:</b> $\hat{y} =$ _____ <b>Transformations:</b> $x \rightarrow$ _____ $y \rightarrow$ _____
Logarithmic Regression	<b>Equation Format:</b> $\hat{y} =$ _____ <b>Transformations:</b> $x \rightarrow$ _____ $y \rightarrow$ _____
Power Regression	<b>Equation Format:</b> $\hat{y} =$ _____ <b>Transformations:</b> $x \rightarrow$ _____ $y \rightarrow$ _____

**Example**

a. Analyze the correlation shown



b. Give **and interpret** the value of the slope of the regression line

c. Give **and interpret** the value of the y-intercept of the regression line

Predictor	Coef	SE Coef	t	P
Constant	-2486.13	96.84	-2.11	.03
Temp	32.23	15.3	4.76	.000
r = .85		r <sup>2</sup> = .72		

d. Give **and interpret** the r<sup>2</sup> value of the regression line

e. If tomorrow's temperature is going to be 90°, predict how many visitors the beach will have tomorrow. **Show work!**

**Topic 13: Confidence Intervals and Significance Tests with Bivariate Data (Notes: 5.3)**

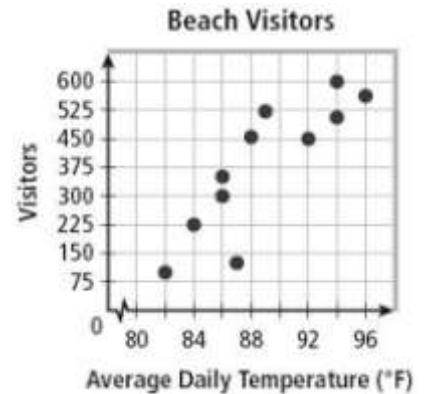
- A regression line is created using a \_\_\_\_\_ of data. Confidence intervals and significance tests allow us to predict and test the \_\_\_\_\_ slope of the relationship between the explanatory and response variables (x and y)
  - You can also do this for y-intercept, but this is not something to worry about for the exam
- The AP exam will most likely ask you to use a \_\_\_\_\_ to make inference
  - Remember, everything dealing with slope is in the row with the \_\_\_\_\_ (“constant” refers to the **y-intercept**)
  - If you need to do them in the calculator...
    - 1. Put all Xs in one list and Ys in another list
    - 2. Go to **LinRegInterval** or **LinRegT-Test**, type in the inputs, and get your results!
- Confidence Interval
  - Confidence interval = Statistic  $\pm$  Critical Value  $\cdot$  Standard Deviation of Statistic
  - For a linear regression, this becomes \_\_\_\_\_
    - “SE Coef” can be found in the \_\_\_\_\_
    - $t^*$  can be found in your \_\_\_\_\_
      - For **degrees of freedom (DF)**, use \_\_\_\_\_
  - Interpretation (assuming 95% confidence)
    - **I am 95% confident that the slope of the *true* regression line of the relationship between     x     and     y     is between \_\_\_\_\_ and \_\_\_\_\_.**
- Significance Test
  - Ho: Assume that there is \_\_\_\_\_ between the variables (this means **slope ( $\beta$ ) = \_\_\_\_\_**)
  - Ha can be \_\_\_\_\_, \_\_\_\_\_, or \_\_\_\_\_ (just like before)
  - t and p can **both** be found in the \_\_\_\_\_. Interpret as usual!
    - The **formula** for the test statistic is \_\_\_\_\_

- Conditions!! (Use the acronym **LINEaR**)

- **L:** \_\_\_\_\_
- **I:** \_\_\_\_\_ (or use \_\_\_\_\_)
- **N:** \_\_\_\_\_
- **E:** \_\_\_\_\_ (can think of this as \_\_\_\_\_)
- **and**
- **R:** \_\_\_\_\_

**Example**

The Florida Tourism Department is studying the habits of beachgoers across the state. They observe a certain beach on 11 randomly-selected days during the peak season (May thru August) and record the Average Daily Temperature and the number of visitors who come to the beach that day. A scatterplot of the data is shown, as is a computer output of the data.



Assume that temperatures and number of visitors are both normally distributed.

Predictor	Coef	SE Coef	<i>T</i>	<i>P</i>
Constant	-2486.13	96.84	-2.11	.03
Temp	32.23	6.76	4.76	.000
<i>r</i> = .85		<i>r</i> <sup>2</sup> = .72		

a. CHECK conditions for inference:

b. Construct and interpret a 95% confidence interval of the slope of this regression line

c. Is there significant evidence at the  $\alpha = .05$  level to suggest that there is a relationship between average daily temperature and number of visitors?

## Additional Topics

1. Know the **SYMBOLS** for parameters and statistics. **Mis-using a symbol WILL cause you to get docked on the exam!!**

<i>Measure</i>	<i>Parameter Symbol (Population)</i>	<i>Statistic Symbol (Sample)</i>
Mean	$\mu$	$\bar{x}$
Standard Deviation (also applies to Variance)	$\sigma$	S
Proportion	p	$\hat{p}$
Number of data points	N	N

2. Know how to work with *percentiles* (“relative frequency”)

- A data point’s percentile tells the percentage of the data that is less than or equal to that data point
  - **Example:** If you’re in the 85<sup>th</sup> percentile, 85% of the population is at or below your level
  - This means **Q1** is the 25<sup>th</sup> percentile, **Median** is the 50<sup>th</sup>, and **Q3** is the 75<sup>th</sup>
- The numbers in the z-table can be considered *percentiles* (for instance, the z-score 0.45 corresponds with .6736 in the z-table, which is the 67<sup>th</sup> percentile)

## AP EXAM ADVICE

*General advice for ALL your exams:*

- **Be prepared**
  - Have your pencils and materials ready to go
  - Get a good night’s sleep! (This will feel strange to some of you)
  - **Be on time.** You WILL NOT be admitted to the testing room if you are late.
  - *Leave the personal drama at the door.* Do not let it bring you down on an exam this important!
- *Don’t try and do too much!* I have seen many students write great answers, only to get docked because they added an incorrect piece of information or tried to make a claim that wasn’t there. *Answer the question as fully yet concisely as possible, and then get out!*
- Read each question **VERY** carefully! AP loves to throw curveballs and you need to be sure of what the question is asking you to do!
- **TIME IS OF THE ESSENCE.** If you are stuck on a question, **OR** you know that question may take a while to figure out, *come back to it.* Knock out the easier ones first.
- **Two minute warning** is the best time to start guessing (*especially* on Multiple Choice).
  - The WORST answer you can possibly have is a blank!

*Specific advice for THIS exam:*

- **TIMING:**
  - 2 minutes and 15 seconds for each multiple choice
  - 13 minutes for Free Response #1 – 5
  - 25 minutes for Free Response #6
  - *Some questions will take more or less than this. That's fine. Just pace yourself!*
- **Calculator Check!**
  - Is it charged and/or have working batteries?
    - If your TI-84 is okay at the start of the test but then says “low battery” in the middle of the test, it will last through the duration of the test. **DO NOT WORRY!**
- **Show work!** You **HAVE** to show enough to prove to the AP Readers that you understand the *process* behind your answers (you **WILL** get docked for not showing enough work.)
  - It doesn't matter *how* simple the calculation is. If it's  $1+1 = 2$ , **write that down.**
- **Formula sheet** is your friend! *Especially* the 2<sup>nd</sup> and 3<sup>rd</sup> pages (as well as the **t-table** because it gives you all the *critical values* you could ever want!). Sometimes the formula sheet gives away an otherwise tricky answer.
  - But be careful: do not, and I repeat, **DO NOT TEAR OUT THE FORMULA SHEET FROM THE TEST BOOKLET. THIS WILL INVALIDATE YOUR EXAM.** This happened to someone I knew on the AP Chem Exam; her score was invalidated and she had to take the test again next year.
- If you need to make a graph, **LABEL YOUR AXES!!**
  - If you're doing it to check the Normality (Sample Size) condition for inference, make sure you **write** whether you see any skews or outliers. **Just showing the graph is not enough** (but don't *forget* to put the graph, either! You need **BOTH** the graph **AND** the analysis of skew/outliers)
    - Remember that *boxplots* are the most efficient (but not the *only*) way of checking for this!
- **Watch your language!** Words like *average*, *range*, *skew*, and *significant* have very specific meanings in statistics, so **DO NOT** use these words unless you are using them in the correct *statistical* context (otherwise, **find synonyms**)
  - Average → Typical
  - “Ranges from” → “Goes from”
  - Skews → Distorts
  - Significant → Substantial
  - **NOTE:** *It is okay to use these words for their statistical definitions. Just use synonyms if you're going to venture outside of that.*
  - **If you aren't sure what a word means, avoid using it!!**
- **Stick to the script!** Know how to phrase your analyses of the following (*they are in your study guide*). **These phrasings help ensure you have covered all important aspects of the analysis in a clear and concise manner!**
  - Confidence intervals
  - Confidence *levels*
  - *Interpreting* p-values
  - *Analyzing* or *drawing conclusions* about p-values
  - Slope of a regression line
  - Interpreting  $r^2$

- **Randomization** and a **large sample size** can solve most of life’s problems – they make for better, more accurate, and more reliable (unbiased) results
- DO NOT mix up the language of *sampling* and the language of *experiments*.
  - For example, subjects of experiments are usually not randomly selected (often times that’s *highly unethical*). They *are*, however, randomly *assigned* to groups (at least they *should* be)
- If you use symbols, **DEFINE** what that symbol means. OR you can weave the context *into* your symbol
  - *Both ways are acceptable* (although one is definitely **quicker!**)

Symbols with definitions	Symbols with context <i>interwoven</i>
$P(A \cap B)$ , where $A$ represents being a girl and $B$ represents being a senior	$P(\text{Girl} \cap \text{Senior})$
$\mu = 23$ , where $\mu$ represents the mean weight of the <i>population</i> of piglets (or <i>true</i> mean weight of piglets)	$\mu_{\text{piglets}} = 23$
$p_1 > p_2$ , where $p_1$ represents the <i>true</i> proportion of adults who like snacks, and $p_2$ represents the <i>true</i> proportion of children who like snacks	$P_{\text{adults}} > P_{\text{children}}$

- For **sampling distributions**, make sure you use  $\mu_{\bar{x}}$  (or  $\mu_{\hat{p}}$ ) for mean and  $\sigma_{\bar{x}}$  (or  $\sigma_{\hat{p}}$ ) for standard deviation
  - **IF YOU DON’T KNOW WHAT SYMBOL TO USE, DON’T USE A SYMBOL AT ALL!!** There’s nothing wrong with writing out an answer in words. An incorrect symbol **WILL** get you docked.
- For inference problems (confidence intervals and significance test), **LOOK** for the statement “assume all conditions are met”. **If it is not there, you had better check those conditions!**
  - Also be on the eye out for *randomness* – is it stated? And for 2-sample problems, is it stated for *both* samples?
- If you’re doing an interval or test, always provide the **name** of the procedure when you do it!
- Remember, **NEVER** claim  $H_0$  or  $H_a$  are “true” or “false”. We “reject” or “fail to reject” based on the *probability* of getting a certain result by chance (that’s what significance tests are all about!) and we *know* that probability is NEVER a guarantee!
- **BREATHE!!** We’ve been working for this all year. *You’ve got this!* One wrong answer won’t kill you. Heck, just getting half of the questions right is *almost guaranteed* to be a 3! Don’t overthink – just do your best.

# GOOD LUCK!!