Categorical vs. Quantitative Data	Marginal vs. Conditional Distributions
SOCS	Comparing Distributions
Outlier Rule	Interpret Standard Deviation
How does shape affect measures of center?	Interpret a z–score
Percentiles	Linear Transformations

In a two-way table, marginal distributions consider only one variable and use the total row/column of the table only. Conditional distributions describe the distribution of one variable for a specific value of the other (one row/column inside the table).	Data are categorical if they fall into groups or categories and data are quantitative if they take on numerical values where it makes sense to find an average. -Use bar graphs, pie graphs, or segmented bar charts for categorical variables such as color or gender. -Use dotplots, stemplots, histograms, or boxplots for quantitative variables such as age or weight.
Address: Shape, Outliers, Center, Spread in context! YOU MUST USE comparison phrases like "is greater than" or "is less than" for Center & Spread	**S**hape – Skewed Left, Skewed Right, Symmetric, Uniform, Unimodal, Bimodal **O**utliers – Discuss them if there are obvious ones **C**enter – Mean or Median **S**pread – Range, *IQR*, or Standard Deviation Note: Also be on the lookout for gaps, clusters or other unusual features of the data set.
Standard Deviation measures spread by giving the "typical" distance that the observations **(context)** are away from the mean **(context)**.	Upper Cutoff = $Q_3 + 1.5(IQR)$ Lower Cutoff = $Q_1 - 1.5(IQR)$ $IQR = Q_3 - Q_1$
$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$ A *z*-score describes how many standard deviations a value falls away from the mean of the distribution and in what direction. The further the *z*-score is away from zero the more "surprising" the value of the statistic is.	In general, Skewed Left (Mean < Median) Skewed Right (Mean > Median) Fairly Symmetric (Mean ≈ Median)
Adding "a" to every member of a data set adds "a" to the measures of position, but does not change the measures of spread or the shape. Multiplying every member of a data set by "b" multiplies the measures of position by "b" and multiplies most measures of spread by |b|, but does not change the shape.	The kth percentile of a distribution is the point that has k% of the values less than that point. For example, a student who scores at the 90th percentile got a higher score than 90% of the other test takers.

The Standard Normal Distribution	Using Normalcdf and InvNorm (Calculator Tips)
Describing an association in a scatterplot	Interpret r
Interpret LSRL Slope "b"	Interpret LSRL y–intercept "a"
What is a Residual?	Interpreting a Residual Plot
Interpret LSRL "$\hat{y}$"	Extrapolation

Using boundaries to find area: Normalcdf (min, max, mean, SD) Using area to find boundary: Invnorm (area to the left as a decimal, mean, SD) If used on AP® exam, make sure to label each input!	The standard Normal distribution is the Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The Normal table displays values for the standard Normal distribution.
Correlation measures the **strength** and **direction** of the **linear relationship** between *x* and *y*. • *r* is always between –1 and 1 • Close to zero = very weak • Close to 1 or –1 = strong • Exactly 1 or –1 = perfectly straight line • Positive *r* = positive correlation Negative *r* = negative correlation	Address the following, *in context*: Direction Outliers Form Strength
When the *x* variable **(context)** is zero, the *y* variable **(context)** is *predicted* to be **_____**.	For every one unit change in the *x* variable **(context)** the *y* variable **(context)** is *predicted* to increase/decrease by ____ units **(context).**
If there is a leftover pattern in the residual plot, then the model used does not have the same form as the association (the model is not appropriate). If there is no leftover pattern in the residual plot, then the model is appropriate.	Residual = $y - \hat{y}$ (Actual – Predicted) A residual measures the difference between the actual *y* value and the *y* value that is predicted by the LSRL.
Using a LSRL to predict outside the domain of the explanatory variable. (Can lead to ridiculous conclusions if the observed association does not continue)	$\hat{y}$ is the "estimated" or "predicted" *y*-value **(context)** for a given *x*-value **(context)**

Interpret LSRL "*s*"	Interpret r^2
Outliers and Influential Points in Regression	Reading Computer Output for Regression
SRS	Using a Random Digit Table to Select a Sample
Sampling Techniques	Advantage of using a Stratified Random Sample Over an SRS
Bias	Experiment vs. Observational Study

___% of the variation in *y* **(context)** is accounted for by the LSRL of *y* **(context)** on *x* **(context)**. Or ___% of the variation in *y* **(context)** is accounted for by using the linear regression model with *x* **(context)** as the explanatory variable.	*s* = ___ is the standard deviation of the residuals. It measures the typical distance between the actual *y* values **(context)** and their predicted *y* values **(context)** in a regression setting
Using foot length (*x*) to predict height (*y*): Predictor Coef SE Coef T P Constant 103.41 19.50 5.30 0.000 Foot length 2.7469 0.7833 3.51 0.004 S = 7.95126 R-Sq = 48.6% R-Sq(adj) = 44.7% *Y* intercept = 103.41 and Slope = 2.7469 *s* = 7.95126 and r^2 = 0.486	Any point that falls outside the pattern of the association should be considered an outlier. A point is influential if it has a big effect on a calculation, such as the correlation or equation of the least-squares regression line. Points separated in the *x*-direction are often influential.
Step 1: Label. Give each member of the population a numerical label with the same number of digits. Use as few digits as possible. **Step 2: Randomize.** Read consecutive groups of digits of the appropriate length from left to right across a line in table. Ignore any group of digits that wasn't used as a label or that duplicates a label already in the sample. Stop when you have chosen *n* different labels. Your sample contains the individuals whose labels you find.	An SRS (simple random sample) is a sample taken in such a way that **every set** of *n* individuals has an equal chance to be the sample actually selected.
Stratified random sampling guarantees that each of the strata will be represented. When strata are chosen properly, a stratified random sample will produce better (less variable/more precise) information than an SRS of the same size.	1. **SRS**– Names in a hat 2. **Stratified** – Split the population into homogeneous groups, select an SRS from each group. 3. **Cluster** – Split the population into groups (often based on location) called clusters, and randomly select whole clusters for the sample. 4. **Census** – An attempt to reach the entire population 5. **Convenience**– Selects individuals easiest to reach 6. **Voluntary Response** – People choose themselves by responding to a general appeal.
A study is an experiment ONLY if researchers <u>impose</u> a treatment upon the experimental units. In an observational study researchers make no attempt to influence the results and cannot conclude cause-and-effect.	A sampling method is biased if it consistently produces estimates that are too small or consistently produces estimates that are too large.

Confounding	Why use a control group?
Blinding	Experimental Designs
Benefit of Blocking	Scope of Inference: Generalizing to a Larger Population
Scope of Inference: Cause-and-Effect	Interpreting Probability
Law of Large Numbers	Conducting a simulation

A control group gives the researchers a comparison group to be used to evaluate the effectiveness of the treatment(s). **(context)** It allows the researchers to measure the effect of the treatment (**context**) compared to no treatment at all.	Two variables are confounded if it cannot be determined which variable is causing the change in the response variable. For example, if people who take vitamins on their own have less cancer, we cannot say for sure that the vitamins are causing the reduction in cancer. It could be other characteristics of vitamin takers, such as diet or exercise.
CRD (Completely Randomized Design) – Units are allocated at random among all treatments **RBD** (Randomized Block Design) –Units are put into homogeneous blocks and randomly assigned to treatments within each block. **Matched Pairs** – A form of blocking in which each subject receives both treatments in a random order or subjects are matched in pairs with one subject in each pair receiving each treatment, determined at random.	When the subjects in an experiment don't know which treatment they are receiving, they are blind. If the people interacting with the subjects and measuring the response variable don't know which subjects received which treatments, they are blind. If both groups are blind, the study is double-blind.
We can generalize the results of a study to a larger population if we used a random sample from that population.	Blocking helps account for the variability in the response variable **(context)** that is caused by the blocking variable **(context).** If there really is a difference in the effectiveness of the treatments, using an appropriate blocking variable will increase power (probability of finding convincing evidence that the treatments are not equally effective).
The probability of an event is the proportion of times the event would occur in a very large number of repetitions. Probability is a long-term relative frequency.	We can make a cause-and-effect conclusion if we randomly assign treatments to experimental units in an experiment. Otherwise, Association is NOT Causation!
State: Ask a question about some chance process. **Plan**: Describe how to use a random device to simulate one trial of the process and indicate what will be recorded at the end of each trial. **Do**: Do many trials. **Conclude**: Answer the question of interest.	The Law of Large Numbers says that if we observe many repetitions of a chance process, the observed proportion of times that an event occurs approaches a single value, called the probability of that event.

Complementary Events	Conditional Probability
Two Events are Independent If…	Two Events are Mutually Exclusive If…
Interpreting Expected Value/Mean	Mean and Standard Deviation of a Discrete Random Variable
Mean and SD of a Transformation of a Random Variable	Mean and Standard Deviation of a Difference of Two Random Variables
Mean and Standard Deviation of a Sum of Two Random Variables	Binomial Setting and Random Variable

Probability that one event occurs given that another event is already known to have occurred. $P(\text{A given B}) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ (on formula sheet)	Two mutually exclusive events whose union is the sample space. For example: -Rain / No Rain -Draw at least one heart / Draw NO hearts
$P(A \text{ and } B) = 0$ Events A and B are mutually exclusive if they share no outcomes.	$P(B) = P(B \mid A) = P(B \mid A^C)$ Events A and B are independent if knowing that Event A has occurred (or has not occurred) doesn't change the probability that event B occurs.
Also on the formula sheet! **Mean (Expected Value):** $\mu_x = \sum x_i p_i$ **Standard Deviation:** $\sigma_x = \sqrt{\sum (x_i - \mu_x) p_i}$	If we were to repeat the chance process (**context**) many times, the average value of ____ (**context**) would be about ____.
<u>Mean of a Difference of 2 RVs:</u> $\mu_{X-Y} = \mu_X - \mu_Y$ <u>SD of a Difference of 2 *Independent* RVs:</u> $\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$	If $Y = a + bX$ $\mu_Y = a + b\mu_X$ $\sigma_Y = \lvert b \rvert \sigma_X$
Binary? Each trial can be classified as success/failure **I**ndependent? Trials must be independent. **N**umber? The number of trials (n) must be fixed in advance **S**uccess? The probability of success (p) must be the same for each trial. X = number of successes in n trials	<u>Mean of a Sum of 2 RVs:</u> $\mu_{X+Y} = \mu_X + \mu_Y$ <u>SD of a Sum of 2 *Independent* RVs:</u> $\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$

Binomial Distribution (Calculator Usage)	Mean and Standard Deviation Of a Binomial RV
Geometric Setting and Random Variable	Parameter vs. Statistic
What is a sampling distribution?	What is the sampling distribution of $\hat{p}$?
What is the sampling distribution of $\bar{x}$?	What is the Central Limit Theorem (CLT)?
Unbiased Estimator	4-Step Process Confidence Intervals

Also on the formula sheet! Mean: $\mu_X = np$ Standard Deviation: $\sigma_X = \sqrt{np(1-p)}$	Exactly 5: $P(X = 5)$ = Binom**p**df(n, p, 5) At Most 5: $P(X \le 5)$ = Binom**c**df(n, p, 5) Less Than 5: $P(X < 5)$ = Binom**c**df(n, p, 4) At Least 5: $P(X \ge 5)$ = 1 – Binom**c**df(n, p, 4) More Than 5: $P(X > 5)$ =1 – Binom**c**df(n, p, 5) Remember to label n, p, and X!
A parameter measures a characteristic of a population, such as a population mean μ or population proportion p. A statistic measures a characteristic of a sample, such as a sample mean $\bar{x}$ or sample proportion $\hat{p}$. Statistics are used to estimate parameters.	Arises when we perform independent trials of the same chance process and record the number of trials it takes to get one success. On each trial, the probability p of success must be the same. X = number of trials needed to achieve one success
Also on the formula sheet! Center: $\mu_{\hat{p}} = p$ Spread: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ if $n < N/10$ Shape: Approximately Normal if $np \ge 10$ and $n(1-p) \ge 10$	A sampling distribution is the distribution of a sample statistic in all possible samples of the same size. It describes the possible values of a statistic and how likely these values are. Contrast with the distribution of the population and the distribution of a sample.
If the population distribution is not Normal the sampling distribution of the sample mean $\bar{x}$ will become more and more Normal as n increases.	Also on the formula sheet! Center: $\mu_{\bar{x}} = \mu$ Spread: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ if $n < N/10$ Shape: Normal if population is Normal Or Approximately Normal if $n \ge 30$ (CLT)
STATE: What parameter do you want to estimate, and at what confidence level? PLAN: Choose the appropriate inference method. Check conditions. DO: If the conditions are met, perform calculations. CONCLUDE: Interpret your interval in the context of the problem.	A statistic is an unbiased estimator of a parameter if the mean of its sampling distribution equals the true value of the parameter being estimated. In other words, the sampling distribution of the statistic is centered in the right place.

Interpreting a Confidence Interval	Interpreting a Confidence Level (The Meaning of 95% Confidence)
Standard Error vs. Margin of Error	What factors affect the Margin of Error?
Inference for Means (Conditions)	Inference for Proportions (Conditions)
Finding the Sample Size (For a given margin of error *m*)	4-Step Process Significance Tests
Explain a *P*-value	Carrying out a Two-Sided Test from a Confidence Interval

If many, many samples are selected and many, many confidence intervals are calculated, about __% of them will capture the true ____.	I am ___% confident that the interval from ___ to ___ captures the true ____.
The margin of error decreases when: -The sample size increases -The confidence level decreases	The standard error of a statistic estimates how far the value of the statistic *typically* differs from the true value of the parameter. The margin of error estimates how far we expect the parameter to differ from the statistic, *at most.*
Random: Data from a random sample or randomized experiment **10%:** The sample must be ≤ 10% of population **Large Counts:** At least 10 successes and failures: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (1 sample *z* test for *p*: $np_0 \geq 10$ and $n(1-p_0) \geq 10$)	**Random:** Data from a random sample or randomized experiment **10%:** The sample must be ≤ 10% of population **Normal/Large Sample:** Population distribution is Normal or sample size is large ($n \geq 30$). If $n < 30$, graph sample data and verify no strong skewness or outliers. Include graph!
STATE: What hypotheses do you want to test, and at what significance level? Define any parameters you use. PLAN: Choose the appropriate inference method. Check conditions. DO: If the conditions are met, perform calculations. Compute the test statistic and find the *P*-value. CONCLUDE: Make a decision about the hypotheses in the context of the problem.	For one mean: $m = z^* \frac{\sigma}{\sqrt{n}}$ For one proportion: $m = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ If a value of $\hat{p}$ is not given, use $\hat{p}$ = 0.5.
α = 1 – confidence level If the null hypothesis value is in the interval, then it is a plausible value that should not be rejected. If the null hypothesis value is not in the interval, then it is not a plausible value and should be rejected.	Assuming that the null is true **(context)** there is a ___ probability of observing a statistic (**context)** as large as or larger than the one actually observed by chance alone.

Type I Error & Type II Error	Power
Factors that Affect Power	Paired *t*-test Identification Hints, H_0 and H_a
Two Sample *t*-test Identification Hints, H_0 and H_a	Chi-Square Tests (Conditions)
Types of Chi-Square Tests	Chi-Square Tests df and Expected Counts
Inference for Regression (Conditions)	Inference for Regression with Computer Output

Power: Probability of avoiding a Type II error
= Probability of finding convincing evidence that H_a is true when in reality H_a is true.

Type I Error: Finding convincing evidence that H_a is true when in reality H_a is not true. (Rejecting H_0 when H_0 is actually true).

Type II Error: Not finding convincing evidence that H_a is true when in reality H_a is true. (Failing II reject H_0 when H_a is true).

Key Phrase: MEAN DIFFERENCE

Two lists of numbers are paired (each row could have a unique label).

H_0: $\mu_d = 0$ H_a: $\mu_d < 0, > 0, \neq 0$
μ_{Diff} = The true mean difference in __.

Sample Size: To increase power, increase sample size.
Significance Level α: A larger value of α increases power.
Effect Size: The farther the true value is from the hypothesized value, the larger the power.
Data Collection: Using blocking rather than a completely randomized design can increase power.

Random: Data from a random sample(s) or randomized experiment
10%: The sample must be ≤ 10% of the population.
Large Counts: All *expected* counts are at least 5.

Key Phrase: DIFFERENCE IN MEANS

Two lists of numbers have no association (could be scrambled).

H_0: $\mu_1 - \mu_2 = 0$ H_a: $\mu_1 - \mu_2 < 0, > 0, \neq 0$
$\mu_1 - \mu_2$ = The true difference in mean ___ for ___ and ___.

1. Goodness of Fit:
   df = # of categories – 1
   Expected Counts: Sample size times hypothesized proportion in each category.

2. Homogeneity or Independence:
   df = (# of rows – 1)(# of columns – 1)
   Expected Counts: $\frac{\text{(row total)(column total)}}{\text{table total}}$

Goodness of Fit: Use to compare the distribution of a categorical variable in one population to a hypothesized distribution.
Homogeniety: Use to compare distribution of a categorical variable for 2+ populations or treatments.
Independence: Use to test the association between two categorical variables in one population.

Using foot length (x) to predict height (y) with n = 15

Predictor	Coef	SE Coef	T	P
Constant	103.41	19.50	5.30	0.000
Foot length	2.7469	0.7833	3.51	0.004
S = 7.95126	R-Sq = 48.6%	R-Sq(adj) = 44.7%		

$df = n - 2 = 15 - 2 = 13$
Test for $\beta = 0$: $t = 3.51$ and P-value = 0.004
CI for β: $2.7469 \pm t^*$ (0.7833)

Linear: Association between the variables is linear. Check with residual plot.
Independent observations, 10% condition if sampling without replacement
Normal: Responses vary normally around the regression line for all x-values. Check with graph of residuals.
Equal SD around the regression line for all x-values. Check with residual plot.
Random: Data from a random sample or randomized experiment