Interpret Standard Deviation	Outlier Rule
Linear Transformations	Describe the Distribution OR Compare the Distributions
SOCS	Using Normalcdf and Invnorm (Calculator Tips)
Interpret a z–score	What is an Outlier?
Interpret LSRL Slope “b”	Interpret LSRL y–intercept “a”

Upper Bound = $Q_3 + 1.5(IQR)$

Lower Bound = $Q_1 - 1.5(IQR)$

$IQR = Q_3 - Q_1$

Standard Deviation measures spread by giving the "typical" or "average" distance that the observations **(context)** are away from their **(context)** mean

SOCS!
Shape, Outliers, Center, Spread
Only discuss outliers if there are obviously outliers present. Be sure to address SCS in context!

If it says "Compare"
YOU MUST USE comparison phrases like "is greater than" or "is less than" for Center & Spread

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.

Normalcdf (min, max, mean, standard deviation)

Invnorm (area to the left as a decimal, mean, standard deviation)

Shape – Skewed Left (Mean < Median)
Skewed Right (Mean > Median)
Fairly Symmetric (Mean ≈ Median)
Outliers – Discuss them if there are obvious ones
Center – Mean or Median
Spread – Range, *IQR*, or Standard Deviation

Note: Also be on the lookout for gaps, clusters or other unusual features of the data set. Make Observations!

When given 1 variable data:
An outlier is any value that falls more than $1.5(IQR)$ above Q_3 or below Q_1

Regression Outlier:
Any value that falls outside the pattern of the rest of the data.

$$z = \frac{\text{value - mean}}{\text{standard deviation}}$$

A z-score describes how many standard deviations a value or statistic (x, $\bar{x}$, $\hat{p}$, etc.) 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.

When the x variable **(context)** is zero, the y variable **(context)** is estimated to be put value here.

For every one unit change in the x variable **(context)** the y variable **(context)** is predicted to increase/decrease by ____ units **(context).**

Interpret r^2	Interpret r
Interpret LSRL "SE_b"	Interpret LSRL "s"
Interpret LSRL "$\hat{y}$"	Extrapolation
Interpreting a Residual Plot	What is a Residual?
Sampling Techniques	Experimental Designs

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 = stronger
- Exactly 1 or −1 = perfectly straight line
- Positive r = positive correlation
- Negative r = negative correlation

___% 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)**

SE_b measures the standard deviation of the estimated slope for predicting the y variable **(context)** from the x variable **(context).**

SE_b measures how far the estimated slope will be from the true slope, on average.

Using a LSRL to predict outside the domain of the explanatory variable.

(Can lead to ridiculous conclusions if the current linear trend does not continue)

$\hat{y}$ is the "estimated" or "predicted" y-value **(context)** for a given x-value **(context)**

Residual = $y - \hat{y}$

A residual measures the difference between the actual (observed) y-value in a scatterplot and the y-value that is predicted by the LSRL using its corresponding x value.

In the calculator: $L_3 = L_2 - Y_1(L_1)$

1. **Is there a curved pattern?** If so, a linear model may not be appropriate.
2. **Are the residuals small in size?** If so, predictions using the linear model will be fairly precise.
3. **Is there increasing (or decreasing) spread?** If so, predictions for larger (smaller) values of x will be more variable.

1. **CRD** (Completely Randomized Design) – All experimental units are allocated at random among all treatments
2. **RBD** (Randomized Block Design) – Experimental units are put into homogeneous blocks. The random assignment of the units to the treatments is carried out separately within each block.
3. **Matched Pairs** – A form of blocking in which each subject receives both treatments in a random order or the subjects are matched in pairs as closely as possible and one subject in each pair receives each treatment, determined at random.

1. **SRS**– Number the entire population, draw numbers from a hat (every set of n individuals has equal chance of selection)
2. **Stratified** – Split the population into homogeneous groups, select an SRS from each group.
3. **Cluster** – Split the population into heterogeneous groups called clusters, and randomly select whole clusters for the sample. Ex. Choosing a carton of eggs actually chooses a cluster (group) of 12 eggs.
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.

Goal of Blocking Benefit of Blocking	Advantage of using a Stratified Random Sample Over an SRS
Experiment Or Observational Study?	Does ___ CAUSE ___?
SRS	Why use a control group?
Complementary Events	*P*(at least one)
Two Events are Independent If…	Interpreting Probability

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.

The goal of blocking is to create groups of homogeneous experimental units.

The benefit of blocking is the reduction of the effect of variation within the experimental units. **(context)**

Association is NOT Causation!

An observed association, no matter how strong, is not evidence of causation. Only a well-designed, controlled experiment can lead to conclusions of cause and effect.

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.

A control group gives the researchers a comparison group to be used to evaluate the effectiveness of the treatment(s). **(context)** (gauge the effect of the treatment compared to no treatment at all)

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.

$P(\text{at least one}) = 1 - P(\text{none})$

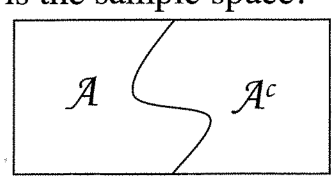
Ex. $P(\text{at least one 6 in three rolls}) =$ ___

$P(\text{Get at least one six}) = 1 - P(\text{No Sixes})$

$= 1 - (5/6)^3$

$= 0.4213$

Two mutually exclusive events whose union is the sample space.



Ex: Rain/Not Rain,
Draw at least one heart / Draw NO hearts

The probability of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions. Probability is a long-term relative frequency.

$P(B) = P(B|A)$

Or

$P(B) = P(B|A^c)$

Meaning: Knowing that Event A has occurred (or not occurred) doesn't change the probability that event B occurs.

Interpreting Expected Value/Mean	Mean and Standard Deviation of a Discrete 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 Distribution (Conditions)	Geometric Distribution (Conditions)
Binomial Distribution (Calculator Usage)	Mean and Standard Deviation Of a Binomial Random Variable
Why Large Samples Give More Trustworthy Results… (When collected appropriately)	The Sampling Distribution of the Sample Mean (Central Limit Theorem)

Also on the formula sheet!

Mean (Expected Value):

$$\mu_x = \sum x_i p_i$$

(Multiply & add across the table)

Standard Deviation:

$$\sigma_x = \sqrt{\sum (x_i - \mu_x) p_i}$$

Square root of the sum of (Each x value – the mean)^2(its probability)

The mean/expected value of a random variable is the long-run average outcome of a random phenomenon carried out a very large number of times.

Mean of a Sum of 2 RV's:

$$\mu_{X+Y} = \mu_X + \mu_Y$$

Stdev of a Sum of 2 Independent RV's:

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Stdev of a Sum 2 Dependent RV's:

Cannot be determined because it depends on how strongly they are correlated.

Mean of a Difference of 2 RV's:

$$\mu_{X-Y} = \mu_X - \mu_Y$$

Stdev of a Difference of 2 Indep RV's:

$$\sigma_{X-Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Stdev of a Difference of 2 Dependent RV's:

Cannot be determined because it depends on how strongly they are correlated.

---

1. **B**inary? Trials can be classified as success/failure
2. **I**ndependent? Trials must be independent.
3. **T**rials? The goal is to count the number of trials until the first success occurs
4. **S**uccess? The probability of success (p) must be the same for each trial.

1. **B**inary? Trials can be classified as success/failure
2. **I**ndependent? Trials must be independent.
3. **N**umber? The number of trials (n) must be fixed in advance
4. **S**uccess? The probability of success (p) must be the same for each trial.

Also on the formula sheet!

Mean: $\mu_x = np$

Standard Deviation: $\sigma_x = \sqrt{np(1-p)}$

Exactly 5: $P(X = 5) =$ Binompdf$(n, p, 5)$

At Most 5: $P(X \leq 5) =$ Binomcdf$(n, p, 5)$

Less Than 5: $P(X < 5) =$ Binomcdf$(n, p, 4)$

At Least 5: $P(X \geq 5) = 1-$Binomcdf$(n, p, 4)$

More Than 5: $P(X > 5) = 1-$Binomcdf$(n, p, 5)$

Remember to define X, n, and p!

---

1. If the population distribution is Normal the sampling distribution will also be Normal with the same mean as the population. Additionally, as n increases the sampling distribution's standard deviation will decrease
2. If the population distribution is not Normal the sampling distribution will become more and more Normal as n increases. The sampling distribution will have the same mean as the population and as n increases the sampling distribution's standard deviation will decrease.

When collected appropriately, large samples yield more precise results than small samples because in a large sample the values of the sample statistic tend to be closer to the true population parameter.

Unbiased Estimator	Bias
Explain a *P*-value	Can we generalize the results to the population of interest?
Finding the Sample Size (For a given margin of error)	Carrying out a Two-Sided Test from a Confidence Interval
<u>4-Step Process</u> Confidence Intervals	<u>4-Step Process</u> Significance Tests
Interpreting a Confidence Interval (Not a Confidence Level)	Interpreting a Confidence Level (The Meaning of 95% Confident)

The systematic favoring of certain outcomes due to flawed sample selection, poor question wording, undercoverage, nonresponse, etc. Bias deals with the **center** of a sampling distribution being "off"!	The data is collected in such a way that there is no systematic tendency to overestimate or underestimate the true value of the population parameter. (The mean of the sampling distribution equals the true value of the parameter being estimated)
Yes, if: A large random sample was taken from the same population we hope to draw conclusions about.	Assuming that the null is true **(context)** the *P*-value measures the chance of observing a statistic (or difference in statistics) **(context)** as large as or larger than the one actually observed.
We do/(do not) have enough evidence to reject H_0: $\mu = ?$ in favor of H_a: $\mu \neq ?$ at the $\alpha = 0.05$ level because <u>?</u> falls outside/(inside) the 95% CI. $\alpha = 1 -$ confidence level	For one mean: $m = z^*\left(\frac{\sigma}{\sqrt{n}}\right)$ For one proportion: $m = z^*\sqrt{\frac{p(1-p)}{n}}$ If an estimation of p is not given, use 0.5 for p. Solve for ***n***.
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: Interpret the result of your test in the context of the problem.	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.
Intervals produced with this <u>method</u> will capture the true population _______ in about 95% of all possible samples of this same size from this same population.	I am ___% confident that the interval from ___ to ___ captures the true ____.

<u>Paired t-test</u> Phrasing Hints, H_0 and H_a, Conclusion	<u>Two Sample t-test</u> Phrasing Hints, H_0 and H_a, Conclusion
Type I Error, Type II Error, & Power	Factors that Affect Power
<u>Inference for Means</u> (Conditions)	<u>Inference for Proportions</u> (Conditions)
Types of Chi-Square Tests	<u>Chi-Square Tests</u> df and Expected Counts
<u>Inference for Counts</u> <u>(Chi-Squared Tests)</u> (Conditions)	<u>Inference for Regression</u> (Conditions)

Key Phrase: DIFFERENCE IN THE MEANS

H_0: $\mu_1 - \mu_2 = 0$ OR $\mu_1 = \mu_2$

H_a: $\mu_1 - \mu_2 < 0, >0, \neq 0$

$\mu_1 - \mu_2$ = The difference between the mean ___ for all ___ and the mean ___ for all ___.

We do/(do not) have enough evidence at the 0.05 level to conclude that the difference between the mean ___ for all __ and the mean ___ for all __ is ___.

Key Phrase: MEAN DIFFERENCE

H_0: $\mu_{\text{Diff}} = 0$

H_a: $\mu_{\text{Diff}} < 0, > 0, \neq 0$

μ_{Diff} = The mean difference in __ for all __.

We do/(do not) have enough evidence at the 0.05 level to conclude that the mean difference in __ for all __ is ___.

1. **Sample Size:** To increase power, increase sample size.
2. **Increase α:** A 5% test of significance will have a greater chance of rejecting the null than a 1% test.
3. **Consider an alternative that is farther away from μ_0:** Values of μ that are in H_a, but lie close to the hypothesized value are harder to detect than values of μ that are far from μ_0.

1. **Type I Error:** Rejecting H_0 when H_0 is actually true. (Ex. Convicting an innocent person)
2. **Type II Error:** Failing to (II) reject H_0 when H_0 should be rejected. (Ex. Letting a guilty person go free)
3. **Power:** Probability of rejecting H_0 when H_0 should be rejected. (Rejecting Correctly)

Random: Data from a random sample(s) or randomized experiment
Normal: At least 10 successes and failures (in both groups, for a two sample problem)
Independent: Independent observations and independent samples/groups; 10% condition if sampling without replacement

Random: Data from a random sample(s) or randomized experiment
Normal: Population distribution is normal or large sample(s) ($n_1 \geq 30$ or $n_1 \geq 30$ and $n_2 \geq 30$)
Independent: Independent observations and independent samples/groups; 10% condition if sampling without replacement

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

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

1. **Goodness of Fit:** Use to test the distribution of one group or sample as compared to a hypothesized distribution.
2. **Homogeniety:** Use when you you have a sample from 2 or more independent populations or 2 or more groups in an experiment. Each individual must be classified based upon a single categorical variable.
3. **Association/Indepencence:** Use when you have a single sample from a single population. Individuals in the sample are classified by two categorical variables.

Linear: True relationship between the variables is linear.
Independent observations, 10% condition if sampling without replacement
Normal: Responses vary normally around the regression line for all x-values
Equal Variance around the regression line for all x-values
Random: Data from a random sample or randomized experiment

Random: Data from a random sample(s) or randomized experiment
Large Sample Size: All expected counts are at least 5.
Independent: Independent observations and independent samples/groups; 10% condition if sampling without replacement