# fiveable 📊 AP STATISTICS CRAM CHART // @thinkfiveable // http://fiveable.me

## Exploring One-Variable Data
### Unit 1 ↓

- **Categorical data** (not numerical) is shown in two-way tables & bar graphs, analyzing proportions
- **Quantitative data** is displayed in histograms, dotplots, box plots, stem and leaf plots, and scatterplots.
- **Mean:** non-resistant (affected by outliers)
- **Median:** resistant (affected by outliers)
- Unimodal = one clear peak, Bimodal = two clear peaks, Uniform = no clear peaks, flat
- Use comparison words when comparing distributions
- **For a histogram** -> make sure you approximate the mean (500-750 units) and use words like "no more" / "approximately" when describing range
- When analyzing distributions, **always CUSS in context -** Center, Unusual features, Shape, Spread (remember skew pulls mean)
- Normal distribution: mound-shaped and symmetric. Its parameters are mu (μ) for mean and sigma σ for standard deviation.
  - Calculate z-score (value-mean / SD), measuring how many SD a value is from a mean
  - The Standard Normal Distribution has a Mean of 0 and a SD of 1
- Empirical Rule: 68% of observations within 1 SD of mean, 95% within 2 SD, 99.7% within 3 SD.

## Exploring Two-Variable Data
### Unit 2 ↓

- For categorical data to be independent, conditional frequency = marginal frequency
- For quantitative data, always describe associations with direction, strength, form.
  - DIrection - **positive / negative (slope)**
  - Form - **linear / non-linear**
  - **r (correlation coefficient)** measures strength & direction, NOT FORM
- **Least Squares regression line (LSRL)** predicts values of response variable (y) given explanatory variable (x)
- LSRL written as **ŷ = a + bx**
  - **ŷ** = predicted value of rsp variable
  - **a** = y-int, **b = slope**
- Residual = predicted - actual
- Look for random scatter on residual plot!
- Using LSRL to make predictions outside the interval of values of x used to make the equation of the line = **extrapolation**
- S & R-sq influenced by outliers (s ↑, r-sq ↓)

Key Interpretations:
**For slope/b:** As the [exp var.] increases by 1 [unit], the [rsp var.] is predicted to increase by **b [units].**
**For y-intercept:** When there are zero [exp var], the predicted [rsp var.] is **y-int.**
**For s:** When using LSRL to predict [rsp var] from [exp var] we are typically off by [value of s].
**For $r^2$ (in %):** About [r-sq]% of variation in [rsp var] is explained by the LSRL using [exp var].
**For a residual:** The actual (rsp var) is about [residual] more/less than the predicted (rsp var).

## Collecting Data
### Unit 3 ↓

- Simple Random Sample (SRS) = every group of a certain size has an equal chance of being selected
- Cluster Sample = Divide pop. into **heterogeneous** groups [**all from some**]
- Stratified Random Sample = Divide pop. into strata of **homogeneous** groups **[some from all]**
- Why stratify by **X -** Explain why indivs in those strata would have different rsps as opposed to some other variable
  - Stratifying -- ↓ variability, ↑ precision
- Bias = **undercoverage**, **nonresponse**, **response bias** (inaccurate) - **Always say if it leads to over/underestimate of a rsp**
- EXPERIMENTS ASSIGN TREATMENTS
- Confounding - When a variable and the exp. variables are associated in a way that their effects on a rsp. Variable can't be distinguished from one another
- Experiments have **comparison, random assignment** (creates roughly equiv. groups of exp. units by balancing the effects of other variables among treatment groups), **control** (helps avoid confounding & ↓ variability in rsp var.), **& replication** (any diffs in effects of treatments can be distinguished from chance differences b/w groups)
- Randomized block design: random assignment of treatments is carried out separately in each block
  - Blocks share a var that may impact rsp
  - ↓ variability in rsp var, allows for easier comparison of treatments
- Matched pairs = compare 2 treatments in block size 2

## Probability, Random Variables, and Probability Distributions
### Unit 4 ↓

- **Probability =** the chance of an event occurring, expressed in a decimal (0-1)
- P(event) = successful/total
- Complement of an event P(not event) is equal to 1 - P(event) (at least, at most, greater/less than)
- P(A and B = P(A∩B) = probability that BOTH events A and B occur
  - Intersection in a 2-way Table
  - Using condt'l probability = P(A) * P(B|A)
- P(A or B) = P(A U B) = probability that either events A or B occur = P(A) + P(B) - P(A and B)
- Conditional Probability = P(A|B) = P(A given B) = probability that event A occurs given B already happened = P(A|B) = P(A and B) / P(B)
- Events are mutually exclusive if P(A or B) = P(A) + P(B)
- **Events are indep if P(A|B) = P(A) OR if P(A and B) = P(A) * P(B)**
- Random variables are quantitative and take numerical values determined by the outcome of a chance event.
- Discrete = only "whole" values possible
- **Expected Value of discrete random variable** = $x_1p_1 + x_2p_2 + ... + x_np_n$
- Continuous random var can be all values in an interval
- Binomial Random Variables: multiple trials of the same event (**B**inary, **I**ndependent (10%), **N**umber of fixed trials, **S**ame probability of success p)
- Geometric Random Variables are STILL independent with fixed probability of success without trials set previously
  - Check Binary, Indep, same prob of success p

## Sampling Distributions
### Unit 5 ↓

**S**ample **s**tatistics help estimate population parameters (x-bar for means and p-hat for proportions)
**A sampling distribution is a distribution of values taken by a statistic in all possible samples of the same size from the same pop.** It shows how a statistic varies in many samples in a pop. Larger samples are less variable and more accurate
When describing a sampling distrib. of p-hat, **it's approx normal if np >= 10 AND n(1-p) >= 10** (**Large Counts Condition) -> Large Counts ensures that sampling distribution is approx. normal** & helps us find z stat for other purposes (like finding a p-value).
Center & spread are found on the formula chart! Make sure to check that the sampling/assignment was random & the 10% condition for indep when it isn't an exp.
When describing a sampling distrib of x-bar:
**Check 10%/random condition, and n > 30.**

## Inference for Categorical Data: Proportions
### Unit 6 ↓

3 Major Conditions: 10 of each, random, 10%
A Conf. Interval = Point Estimate +/- Margin of Error **Larger sample size decreases margin of error!
1-sample CI- state as a **one-sample C% Z-Interval for p,** hypothesis test = **one sample z-test for p**
2-sample CI - state as a **two-sample C% Z-Interval for $p_1$ - $p_2$,** hypothesis test = **two-sample z-test for $p_1$ - $p_2$**
NO **paired data for props. State parameter(s) & hypotheses!**
**For a 2-sided $H_a$ (means or proportions), we can use a confidence interval to make a decision about H0.** Reject if $H_0$ not in interval **[Conf level = opp of signif level)**
Interpreting p-value:
Assuming the **[H0 in context]** is true, the probability that the [observed statistic - x-bar or p-hat] will take a value as or more extreme than it does is **[p-value].** {MAKE SURE TO USE CONTEXT}
Interpreting confidence level:
If we were to select many random samples of the same size # **[from the problem]** from the same population of **[problem]** and construct a C% confidence interval using each sample, about C% of the intervals would capture the **[parameter in context]**.
Interpreting confidence intervals:
We are C% confident that the interval from __ to ___ captures the [true parameter in context]

## Inference for Quantitative Data: Means
### Unit 7 ↓

3 Major Conditions - n > 30, random (no experiments), independent
For a confidence interval for one sample - state as a **one-sample C% t-interval for mu**, hypothesis test = **one-sample t-test for μ**
For a confidence interval of two samples - state as a **2-sample C% t-interval for $\mu_1$- $\mu_2$,** hyp. test = **two-sample t-test for $\mu_1$- $\mu_2$**
Note: there **IS** paired data for mean an (experimental unit received 2 treatments). For paired data, label the parameter mu_diff as the **true mean difference of [context - context]**
**CI = One-sample t C% CI for $\mu_{diff}$**
**Hyp test = One-sample t test for $\mu_{diff}$**
Type I Error - Rejecting Null hypothesis ($H_0$) & finding convincing evidence for the alt. hyp. ($H_a$) **when we should've failed to reject H0 and not found convincing evidence for $H_a$**
Type II error - Failing to reject the Null hypothesis ($H_0$) & not finding convincing evidence for the alt. hyp. ($H_a$) when **we should've rejected H0 and found convincing evidence for $H_a$**
When interpreting Type I/II - replace $H_0$/$H_a$ with the actual null/alternative hypothesis!
**Power = 1 - P(Type II error)** - probability we correctly reject $H_0$ when the reality is that H0 is false [correctly detecting a false $H_0$]
To increase power, you can increase sample size [decreases standard error & dec. p-values], increase significance level [more likely to reject H0], and increase the diff b/w $H_0$ and true $H_a$.

## Inference for Quantitative Data: Slopes + Qualitative Data: Chi-Square
### Units 8 + 9 ↓

**Unit 8 - Chi-Square Inference:**
1 Major Formula:

Chi-square statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

3 Tests: GoF, Independence, and Homogeneity
1 Type of Statistical Inference: Hypothesis Test
- Chi-Squared is a **non-parametric** test meaning we do not make assumptions
- Remember to calculate your df (n-1)

3 Major Conditions: **random, indep., at least 5 success/fail**
Remember to name the correct type of test
**Unit 9 - Inference for Slopes:**
LSRL Equation for Inference for Slopes: μ = a + β x
5 Major Conditions: Linear, Indep, Normal, equal SD, Random

For a **confidence interval -** use this equation: b ± $t^*(SE_b)$
- For unit 9, df=n-2

For a **hypothesis test -** use this equation: $(B-\beta_0)/SE_b$