AP Statistics Summary

Michel Liao [*]

May 2021

Contents

[*] These notes come from College Board's Daily Videos for AP Statistics.

1 Brief Notes

1.1 Introduction

This summary includes the most important ideas you need to know for the AP exam. I would use this as a supplement to a prep book or other study guides.

In addition, be sure to check out the progress checks on MyAP (they closely mirror the actual prompts on the exam, *especially* hypothesis testing).

1.2 Contact Me

There may be typos. If you notice any or have suggestions on improving the summary, please email me at michel.liao@systemgreen.org.

2 Exploring One-Variable Data

2.1 Definitions

- A **variable** is a characteristic that changes from one individual to another.
- **Categorical variables** take on values that are categories or group labels. **Quantitative variables** take on numerical values for a measured or counted quantity.[1]

2.2 Describe a Distribution

- Shape: Symmetric, skewed left, skewed right, unimodal, bimodal, and uniform.
- Center: Mean, median, and Q1 and Q3 (median of first and second half, respectively).[2]
- Variability (spread): Range, IQR (Q3-Q1), and standard deviation (the typical distance each value is away from the mean), and variance (square of SD).
- Unusual features: Outliers, gaps, and clusters.
  - More than 1.5IQR below Q1 or more than 1.5IQR above Q3
  - Two or more standard deviations above/below the mean.
  - Mean, SD, and range are highly influenced by outliers (nonresistant). Median and IQR aren't greatly affected (resistant).
    - Use median and IQR for skewed distributions. Use mean and SD for symmetric distributions.
    - Skewed right distribution, mean $>$ median. Skewed left distrtibution, mean $<$ median.

2.3 Boxplots

- Use a five-number summary:
  1. Minimum
  2. Q1
  3. Median (Q2)
  4. Q3
  5. Maximum

[1] You can distinguish a quantitative variable because it makes sense to take an average of those values. E.g., you wouldn't average zip codes, so it's a categorical variable (even though it has a number value).

[2] Don't include the median.

6. Outliers

- When there are outliers, plot them as dots and change the minimum and maximum lines to be the lowest and highest values in the sample that aren't outliers.

2.4 Comparing Distributions

- Use shape, center, spread, and outliers.
- Use comparative words like same, similar, greater, less than, and use context.

2.5 The Normal Distribution

- The **percentile** is the percent of data values less than or equal to a given value.
  - The value of [] is at the pth percentile. About p percent of the values are less than or equal to [].
- A **standardized score** is given by

$$\text{standardized score} = \frac{\text{data value} - \text{mean}}{\text{standard deviation}} \implies \text{z-score} = \frac{x_i - \mu}{\sigma}.$$

  - The value of [] is [z-score] standard deviations above/below the mean.
- The **Empirical Rule** says that 68%, 95%, and 99.7% of the data is within 1, 2, and 3 SD of the mean, respectively.
  - Exploit symmetry (divide by 2).
- Normalcdf and invNorm are cool. Set $\mu = 0, \sigma = 1$ to find z-score stuff.

3 Exploring Two-Variable Data

3.1 Relationships

- If distributions aren't the same for each group, then there is an association between two variables.
- **Join relative frequency** is a cell divided by table total. **Marginal relative frequency** is a row/column total divided by table total. **Conditional relative frequency** is for a specific part (row/column) of a two-way table.
- Because the distribution of conditional relative frequencies is different for each age group, the two variables are associated.

- Explanatory variable explains the changes in the response variable.
- Describe a scatter plot using direction (positive/negative), form (linear/non-linear), strength (strong/ weak), and unusual features (clusters, apparent outliers).

3.2 Linear Regression Models

- **Extrapolation** occurs when we made predictions outside the interval of our current data's x-values. Current trends may not continue. Used to answer "is this prediction reliable?"
- Interpreting correlation coefficient: strength, direction, linear, context.
- Residual $= y - \hat{y}$.
- For a good fit, residual plot should show apparent randomness and centered at 0. With a bad fit, there is a pattern, which our model failed to capture. A linear model may not be the best fit.

3.3 Least Squares Regression Line

- A linear model that minimizes the sum of the squared residuals.
- A LSRL contains the point $(\overline{x}, \overline{y})$.
- $$b = r(\frac{s_y}{s_x}).$$
- Interpreting slope: For every 1 [unit] increase in [explanatory variable], our model predicts an average increase/decrease of [slope] in [response variable].
- Interpreting y-intercept: When the [explanatory variable] is zero [units], our model predicts that the response variable would be [y-intercept].
  - The y-intercept is only meaningful in contexts where the explanatory varaible can reasonably take on a value of 0.
- Interpreting the coefficient of determination (r^2): r^2% of the variation in the [response variable] can be explained by the linear relationship/our model with the [explanatory variable].[3]
- **High-leverage points** are points with unusually large/small x-values (away from $\overline{x}$). They usually have a big effect on the slope/y-intercept of the LSRL.

[3] r^2 cannot be negative.

- **Influential points** are points that, if removed, will change the slope, y-intercept, or correlation substantially (high-leverage points (change slope/y-int) and outliers (change correlation)).

- If the data shows non-linear relationships, we can transform the data (log, squaring, exponentiating).

4 Collecting Data

- Generalize when your sample is randomly selected or representative of that population.

- **Observational studies** are taken without imposing treatments on individuals and cannot infer cause and effect. **Experiments** are where different treatments are imposed upon subjects and can infer cause and effect (if well done).

4.1 Data Collection

- **Census** collects data from all individuals in a population.

- **Simple random sample** is easier than a census; every group of a given size has an equal chance of being chosen.

- **Cluster random sample** is where a population is divided into clusters of individuals near one another and an SRS of clusters is taken. All individuals within the chosen clusters are sampled.

  - You want clusters that are heterogeneous and clusters that are similar to each other.

- **Stratified random sample** is where a population is divided into strata, based on a similar characteristic and an SRS is taken within (not the whole) each stratum. The selected individuals combine to the larger sample.

  - You want strata that are homogeneous and not similar to each other.

- **Systematic random sample** is where you choose a start point and sample at a fixed periodic interval.

- **Bias** is a systematic tendency to favor certain responses over others.

  - **Undercoverage bias** is where a part of the population has a reduced chance of being in the sample.
  - **Nonresponse bias** is when individuals chosen for a sample don't respond.

  - **Voluntary response bias** is when invitations are sent to all individuals, but people choose to participate.
  - **Question wording bias** is when survey questions are confusing or leading.
  - **Self-reported response bias** is when individuals inaccurately report their own traits.

- Identifying sources of bias:

  1. Identify the population and sample.
  2. Explain how the sample might differ from the population.
  3. Explain how it leads to overestimate or underestimate.

4.2 Experimental Design

- **Confounding variable** is a variable that is related to the explanatory variable that influences the response variable and may create a false perception of association between the two.

- Well-designed experiment:

  - Compares at least two treatment groups (one can be a control).
  - Random assignment of treatments.
  - Replication (enough experimental units to find individual differences).
  - Control confounding variables.

- **Randomized block design** ensures that units in each block are similar with regard to a blocking variable, which helps separate natural variability from differences due to the blocking variable. (Matched pairs is similar where the blocks are of size 2 or each subject receives both treatments[4]).

- **Placebo** is a fake treatment similar to the treatments being tested.

4.3 Inference and Experiments

- Random assignment allows us to conclude that large observed changes are not by chance (statistically significant). It allows us to contribute that the treatment caused the effect.

- Conclusions from samples can be attributed to the population when the sample is representative of the population (random selection).

[4] Matched pairs can only be applied when there are 2, and only 2, treatments.

5 Probability, Random Variables, and Probability Distributions

- **Random process** is when we are aware of possible outcomes but have no idea what the outcome will be.
- **Law of large numbers** says that the simulated probability gets closer to the theoretical probability with more flips. Also, the variability from the true probability decreases.
- Generate random integers with math $\implies$ PROB $\implies$ randInt(.

5.1 Probability

- For equally likely outcomes,
$$P(A) = \frac{\text{\# of successes}}{\text{of possible outcomes}}.$$
- Valid probability distributions are characterized by the sum of all probabilities being 1.
- Complement of $P(A)$ is $P(A')$ or $P(A^C)$, which is equal to $1 - P(A)$.
- The probability of two mutually exclusive events is 0.
- $$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
- Events A and B are independent if and only if
$$P(B|A) = P(B), P(B|A') = P(B), P(A|B) = P(A), P(A|B') = P(A).$$
In other words, A and B are independent if
$$P(A \cap B) = P(A) \cdot P(B).$$
- $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

5.2 Random Variables & Probability Distributions

- **Discrete** random variables take an integer number of values. **Continuous** random variables can be any real number.[5]
- When multiplying/dividing distributions by a number, the center and spread changes. When adding/subtracting by a number, the center changes.

[5] Kind of any... not really, though.

- Describe a probability distribution using shape, center, and spread.
- Discrete random variable calculations: Use L1 and L2 $\Longrightarrow$ stat $\Longrightarrow$ 1-Var Stats $\Longrightarrow$ List as your x, FreqList as your y
- For independent random variables X and Y and real numbers a and b:
  - The mean of
$$aX + bY = a\mu_x + b\mu_y.$$
  - The standard deviation of
$$aX + bY = \sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}.$$

5.3 Binomial Distribution

- Define your variables, e.g., let X = ...
- The binomial setting involves repeated trials where the following conditions are met:
  1. B - binary; two possible outcomes: success or failure
  2. I - independent trials
  3. N - fixed number of trials, n
  4. S - same probability of success, p
- binomcdf for less than or equal to a certain number of successes and binompdf for exactly a certain number of successes.
- The mean (expected value) of a binomial random variable X is
$$\mu_x = np.$$
- The standard deviation of a binomial random variable X is
$$\sigma_x = \sqrt{np(1-p)}.$$

5.4 Geometric Distribution

- Properties of a geometric setting:
  1. Two possible outcomes: success/failure
  2. Independent trials
  3. Same probability of success
- Find the first success at some x with geometpdf. Find the probability the first success occurs at the first or second or third or fourth... or xth trial with geometcdf.
- The mean (expected value) of a geometric random variable X is $\mu_x = \frac{1}{p}$.
- The standard deviation of a geometric random variable X is $\sigma_x = \frac{\sqrt{1-p}}{p}$.

6 Sampling Distributions

- Different samples from the same population produce different statistics (means/proportions). If you take many samples, the proportion/mean will cluster around the true population proportion/mean.

6.1 Normal Distribution Revisited

- Use normalcdf and invNorm (make sure to label your values).
- Given random variables X and Y, you can find the mean and standard deviation of the distribution $X \pm Y$ with

$$\mu_{X \pm Y} = \mu_X \pm \mu_y, \sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

- The distribution of $X \pm Y$ can be modeled with a normal distribution if the probability distribution of each X and Y is approximately normal and independent.
- Sample size must be large enough for a binomial distribution to be approximately normal.
    1. Unimodal, roughly symmetric, bell-shaped
    2. Empirical rule roughly applies

6.2 Central Limit Theorem

- A **sampling distribution** of a statistic is the distribution of values for the statistic for *all possible samples* of the same size from a given population.
- If the population distribution is normal, the sampling distribution is normal (no matter how small the sample size).
- **Central limit theorem** (CLT) says that when the sample size is sufficiently large (≥ 30), the sampling distribution of the mean of a random variable will be approximately normal.
- A **randomization distribution** is reallocating the response values to treatment groups. We can use it to find the likelihood of an observed outcome happening by chance alone.

6.3 Biased and Unbiased Point Estimates

- A sample statistic is a point estimator of the corresponding population parameter.
- An **unbiased estimator** occurs when, on average, the value of the estimator is equal to the population parameter (e.g. mean is but range isn't).

6.4 Sample Proportions

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

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

  - This assumes that the sample size is less than 10% of the population size, or the **10% condition** (or sampling with replacement).

- **Large counts condition** says the sampling distribution of $\hat{p}$ will be approximately normal when

  $$np \geq 10 \text{ and } n(1-p) \geq 10.$$

- Interpreting $\mu_{\hat{p}}$: For all random samples of size $n =$[] from this population, the sample proportions of [random variable] will have a mean of [$\mu_{\hat{p}}$].

- Interpreting $\sigma_{\hat{p}}$: For all random samples of size $n =$[] from this population, the sample proportions of [random variable] typically vary by about [$\sigma_{\hat{p}}$] from the population proportion of μ.

6.5 Difference in Sample Proportions

-

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be approximately normal when the large counts condition is met for both sampling distributions.

- Interpreting the mean and standard deviation follow a similar format as for one sample proportion.

6.6 Sample Means

- $$\mu_{\overline{x}} = \mu.$$

- $$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}.$$

  - Assumes the 10% condition is met and independence.

- The sampling distribution of $\overline{x}$ is approximately normal when the population distribution is approximately normal <u>or</u> the CLT applies ($n \geq 30$).

- Interpreting $\mu_{\overline{x}}$: For all random samples of size $n =$[] from this population, the sample mean [random variable] will have a mean of [$\mu_{\overline{x}}$].

- Interpreting $\sigma_{\overline{x}}$: For all random samples of size $n =$[] from this population, the sample mean [random variable] will typically vary by about $\sigma_{\overline{x}}$ from the population mean of [$\mu_{\overline{x}}$].

6.7 Difference in Sample Means

- $$\mu_{\overline{x}_1 - \overline{x}_2} = \mu_1 - \mu_2.$$
- $$\sigma_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$
  - Assuming the 10% condition is met and independence.
- The sampling distribution of $\overline{x}_1 - \overline{x}_2$ will be approximately normal when both population distributions are approximately normal <u>or</u> both sampling distributions meet the CLT (≥ 30).
- Interpreting the mean and standard deviation follow a similar format as for one sample mean.

7 Inference for Proportions

7.1 Confidence Intervals

1. Construct a C% confidence interval for [p = population parameter].
2. To estimate the proportion of successes in a single population, use a one-sample z interval for p.
3. Conditions:[6]
   - Random sample from the population
   - 10% condition $10n < N$.
   - Large counts condition $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$.
4. CI = point estimate ± (critical value)(standard error of statistic)[7] or stat $\implies$ TESTS $\implies$ 1-PropZInt.
5. Conclusion: We are C% confident that the interval [] to [] captures the [population parameter].
6. Interpreting the confidence level: In repeated random sampling with the same sample size, approximately C% of the C% confidence intervals would capture the [population parameter].

[6] The first two conditions establish independence and the third establishes normality.

[7] The standard error of a statistic is how far the statistic typically varies from the parameter. It's like standard deviation (σ) except replace p with $\hat{p}$

7. We can decrease the margin of error by increasing the sample size or decrease the confidence level.

7.2 Significance Tests for a Proportion

1. Let $p =$ [proportion of all...]

2. H_0: $p = 0.5$.[8]

   H_a: $p \neq 0.5$ or $p > 0.5$ or $p < 0.5$.

   - When the H_a inequality is $>$ or $<$m it's called one-sided. When it's $\neq$, it's called two-sided.

3. Identify the significance level. If there isn't one, you can use $\alpha = 0.05$.

4. One-sample z test for p.

5. Conditions:

   (a) Random sampling

   (b) 10% condition $10n < N$.

   (c) Large counts condition $np_0 \geq 10, n(1 - p_0) \geq 10$.[9]

6. standardized test statistic $= \frac{\text{sample statistic}-\text{null value of the parameter}}{\text{standard deviation of the statistic}}$.

7. stat $\implies$ TESTS $\implies$ 1-PropZTest.

8. Interpreting p-value: Assuming that [null hypothesis], there is a [p-value] probability we get a sample proportion as extreme or more extreme than [sample proportion] in either direction (or greater than or less than) by chance alone in a random sample of [].

9. Conclusion: Since the p-value of [] $\geq\leq$ α =[], we reject/fail to reject the null hypothesis in favor of the alternative hypothesis. There is/isn't convincing evidence that [H_a in context].

7.3 Potential Errors

- **Type I** error occurs when we find convincing evidence for H_a but H_0 is actually true.

- **Type II** error occurs when we do not find convincing evidence for H_a but H_0 is actually false.

- When H_0 is true, $P(\text{Type I Error}) = \alpha$.

- Decreasing the probability of a Type I error increases the probability of a Type II error.

[8] p won't always be equal to 0.5, it's just a placeholder.

[9] p_0 refers to the proportion in the null hypothesis.

- Power is the probability of avoiding a Type II error, or the probability of correctly rejecting a false null. In other words, $P(\text{Type II error}) = 1 - \text{power}$.

- Power of a test is greater when sample size (n) increases, significance level (α) increases, standard error decreases, or the true parameter value is farther from the null.

7.4 Confidence Interval for Two Proportions

1. Construct a C% confidence interval for the difference in $p_1 =$ [] and $p_2 =$ [].

2. Two-sample z interval for a difference in proportions.

   - Works when two random samples are selected or subjects are randomly assigned to two groups in an experiment.

3. Conditions:

   (a) Two random samples (doesn't apply to experiments).

   (b) 10% condition for both samples if no replacement. If the data is from two groups that are randomly assigned, we just check they were randomly assigned.

   (c) Large counts condition for both samples.

4. stat $\Longrightarrow$ TESTS $\Longrightarrow$ 2-PropZInt.

5. Conclusion: We are C% confident that the interval [] to [] captures [difference in proportions].

7.5 Significance Tests for Two Proportions

1. Let $p_1 =$[] and $p_2 =$[].

2. H_0: $p_1 = p_2$ or $p_1 - p_2 = 0$

   H_a: $p_1 \neq p_2$ or $p_1 < p_2$ or $p_1 > p_2$.

3. Two-sample z test for a difference in proportions.

   - Works with two random samples or subjects are randomly assigned to two groups in an experiment.

4. A pooled proportion is combining the success rates of each of the two samples.[10]

5. Conditions:

[10] This is because we assume $p_1 = p_2$.

   (a) Random samples.
   (b) 10% condition for both samples.
   (c) Large counts condition, but use the *pooled proportion*, $\hat{p}_C$. So, we check $n_1\hat{p}_C \geq 10$, $n_2(1-\hat{p}_C) \geq 10$, $n_2\hat{p}_C \geq 10$, $n_2(1-\hat{p}_C) \geq 10$.

6. The test statistic uses $\hat{p}_C$ instead of the population proportion.

7. Interpreting p-value: Assuming H_0 is true, there is a [p-value] probability of getting a difference in proportions of [observed difference] or [greater/less/more different], by chance alone.

8. Conclusion: Since the p-value of [] $\leq\geq$ $\alpha =$ [] we reject H_0. There is/is not convincing evidence that [H_a in context].

8 Inference for Means

- z^* is only useful when we know σ, but that rarely happens. So, we use s instead of σ and t^* instead of z^*.

8.1 Confidence Intervals for Population Means

1. Construct a C% confidence interval for [$\mu =$].

2. One-sample t interval for population mean

3. Conditions:

   (a) Random sample.
   (b) 10% condition.
   (c) CLT applies or sample is free of strong skewness and outliers (boxplot or dotplot).

4. For the margin of error, since we don't know the population standard deviation, we use standard error instead. The degrees of freedom is $n-1$.

   - We can decrease margin of error by increasing sample size or decreasing the confidence level.

5. stat $\Longrightarrow$ TESTS $\Longrightarrow$ TInterval.

6. Interpreting CI: We are C% confident that the interval from [] to [] captures the [parameter to be estimated].

7. Interpreting confidence level: In repeated random sampling with the same sample size, approximately C% of those C% confidence intervals will capture the population mean.

8.2 Significance Test for Population Means

1. Let $\mu =$ [population mean].
2. H_0: $\mu =$ [value].

   H_a: $\mu \neq$ [value] or $\mu >$ [value] or $\mu <$ [value].
3. One-sample t test for a population mean.
4. Conditions
   (a) Random sample or randomized experiment.
   (b) 10% condition.
   (c) CLT applies or no strong skewness or outliers.
5. stat $\Longrightarrow$ TESTS $\Longrightarrow$ TTest.
6. Interpreting p-value: Assuming H_0 is true, there is a [p-value] probability of getting a sample mean of [observed mean] or [greater/less/more different], by chance alone in the random sample.
7. Conclusion: Because the p-value of [] $\leq>$ $\alpha =$[], we reject/fail to reject H_0. There is/is not convincing evidence that [H_a in context.]

8.3 Confidence Intervals for Difference of Two Means

1. Construct a C% confidence interval for the difference in $\mu_1 =$ [context] $\mu_2 =$ [context].
2. Two-sample t-interval for the difference in population means
3. Conditions
   (a) Two random samples or randomized experiment.
   (b) 10% condition.
   (c) CLT applies for both or no strong skewness/outliers for both.
4. stat $\Longrightarrow$ TESTS $\Longrightarrow$ 2-SampTInt.
5. Interpreting CI: We are C% confident that the interval from [] to [] captures the [parameter to be estimated].
6. Conclusion: Because the p-value of [] $\leq>$ $\alpha =$[], we reject/fail to reject H_0. There is/is not convincing evidence that [H_a in context.]

8.4 Significance Test for Difference in Population Means

1. Let $\mu =$ [population mean].
2. H_0: $\mu =$ [value].

   H_a: $\mu \neq$ [value] or $\mu >$ [value] or $\mu <$ [value].
3. Two-sample t test for a difference in population means.
4. Conditions
   (a) Random sample or randomized experiment.
   (b) 10% condition to both.
   (c) CLT applies or no strong skewness or outliers to both.
5. stat $\implies$ TESTS $\implies$ TTest.
6. Interpreting p-value: Assuming H_0 is true, there is a [p-value] probability of getting a sample mean of [observed mean] or [greater/less/more different], by chance alone in the random sample.
7. Conclusion: Because the p-value of [] $\leq>$ $\alpha =$[], we reject/fail to reject H_0. There is/is not convincing evidence that [H_a in context.]

9 Chi-Square Tests

- Expected value of a two way table is (row total)(column total)/(table total) or use matrices.

9.1 Chi-Square Goodness-of-Fit Test

1. H_0: The proportions for the categories in a single categorical variable are equal to the specified values. (List out each proportion and include ”where p_1 is...”)

   H_a: At least one of the proportions is not as specified in the null hypothesis.
2. Identify the procedure.
3. Conditions:
   (a) Data should come from random sample or randomized experiment.
   (b) Sample should be less than or equal to 10% of the population.
   (c) All expected counts greater than 5 (show using a table/calculated values).
4. Calculate using χ^2GOF-Test. Remember df $=$ categories $- 1$.

5. Interpret p-value: Assuming the proportions of a single categorical variable as stated in the null hypothesis is true, there is a [p-value] probability of getting a chi-square statistic as extreme as the one in the study or more extreme, by chance alone in random sampling/assignment.

6. Conclusion: Since the p-value of [] $\leq>$ $\alpha =$ [], we reject/fail to reject H_0. This is/is not convincing evidence that [H_a in context].

7. Contributions: Identify which one had the largest contribution toward the chi-square test statistic using the contributions list.

9.2 Chi-Square Test for Homogeneity or Independence

1. If 1 categorical variable and 2+ populations, homogeneity. If 2 categorical variables and 1 population, independence.

2. Homogeneity:

   H_0: There is no difference in the distribution of [categorical variable] across populations or treatments.

   H_a: There is a difference in the distribution of [categorical variable] across populations or treatments.

   Independence

   H_0: There is no association between [categorical variable] and [categorical variable] in a given population.

   H_a: There is an association between [categorical variable] and [categorical variable] in a given population.[11]

3. Conditions:

   (a) Random sampling or randomized experiment.

   (b) Sample size is less than 10% of population.

   (c) All expected counts are at least 5 (show using matrix or calculated values).

4. Calculate using χ^2-Test, where df = (rows − 1)(columns − 1).

5. Interpret p-value: Assuming the H_0 is true, there is a [p-value] probability of getting a χ^2 of [calculated chi-square] or greater, by chance alone in the random sample(s) or random assignment.

6. Conclusion: Because the p-value of [] $\leq>$ $\alpha =$, we reject/fail to reject H_0. There is/is not convincing evidence that [H_a in context].

[11] Or you can use [categorical variable] and [categorical variable] are/aren't independent.

10 Regression Models

10.1 Confidence Intervals

1. Construct a C% confidence interval for $\beta = []$.

2. One-sample t interval for β.

3. Conditions:[12]

   - Random sample or randomized experiment.
   - The true relationship between x and y is linear.
   - σ_y does not vary with x/is the same with all values of x.
   - For a particular value of x, the y-values are normally distributed (dotplot or CLT).

4. Describe using shape, center ($\mu_b = \beta$), and spread ($\sigma_b = \frac{\sigma}{\sigma_x \sqrt{n}}$)[13].

5. Calculation:[14] stat $\Longrightarrow$ TESTS $\Longrightarrow$ LinRegTInt

6. Decrease margin of error by increasing sample size or decrease the confidence level.

7. Interpret the CI: We are C% confident that the interval from [] to [] captures the slope of the population regression line [context].

8. Interpreting confidence level: In repeated random sampling with the same sample size, approximately C% of C% confidence intervals created will capture the slope of the population regression line.

10.2 Test for Slope

1. H_0: $\beta = 0$.

   H_a: $\beta \neq 0$ or $\beta > 0$ or $\beta < 0$.

2. T-test for slope

3. Conditions:

   - Random sample or randomized experiment.
   - 10% condition.
   - The true relationship between x and y is linear.
   - σ_y does not vary with x/is the same with all values of x.

[12] Scatter plots, residual plots, and dot plots are helpful for justification.

[13] The standard deviation formula works if the 10% condition applies. Use s_x to estimate σ_x and s to estimate σ.

[14] Remember, $df = n - 2$.

- For a particular value of x, the y-values are normally distributed (dotplot or CLT).

4. Calculation: stat $\Longrightarrow$ TESTS $\Longrightarrow$ LinRegTTest[15]

5. Interpret p-value: Assuming [H_0 in context], there is an approximately [p-value] probability of getting a sample regression line with a slope of [] or greater by chance alone in a random sample of [].

6. Conclusion: Because the p-value of [] $\leq >$ $\alpha =$ [], we reject/fail to reject H_0. There is/isn't convincing evidence that [H_a in context].

[15] $df = n - 2$.