STAT 5: Comprehensive Final Review
TikTok Trends

Antonio Aguirre

Department of Statistics, University of California, Santa Cruz

Instructions

This review covers all key topics for the STAT 5 final exam. Use formulas and reasoning, show your work, and interpret your answers in context where requested.

1. Descriptive Statistics: Mean, Variance, and Standard Deviation

Key Concepts

- The **sample mean** $\overline{x}$ summarizes the central tendency of the data.
- The **sample variance** s^2 and **standard deviation** s measure how spread out the data values are.
- *Always show your work.*

Problem 1. Understanding TikTok Creators' Editing Habits

A researcher records the number of hours spent editing TikTok videos over the last week for a random sample of $n = 8$ creators:

$$\boxed{2,\ 3,\ 3,\ 4,\ 4,\ 5,\ 7,\ 8}$$

(a) **(Central Tendency & Spread)** Calculate the **sample mean** $\overline{x}$ and the **sample variance** s^2 for these data. *Write out the formula and show your calculation steps clearly.*

(b) **(Measure of Dispersion)** Using your result from (a), compute the **sample standard deviation** s. *Explain in one sentence what this number tells you about editing hours among these TikTok creators.*

2. Probability Rules: Unions, Complements, and Intersections

Tips

- Use rules for probability of unions, intersections, and complements.
- Let $A =$ "user follows food creators"; $B =$ "user follows dance creators."
- Visual aids (like a Venn diagram) are encouraged!

Problem 2. TikTok Content Preferences
A city-wide survey of TikTok users reports:

- 60% follow **food** content creators ($P(A) =?$),
- 50% follow **dance** content creators ($P(B) =?$),
- 35% follow **both** food and dance creators ($P(A \cap B) =?$).

Assume all probabilities are out of all TikTok users in the city.

(a) **(At least one category)** What is the probability that a randomly selected user follows **at least one** of these two categories? *Write your answer as $P(A \cup B)$. Show your formula and calculation.*

(b) **(Neither category)** What is the probability that a user follows **neither** food nor dance creators? *Express your answer as $P((A \cup B)^c)$ and show your reasoning.*

(c) **(Food but not dance)** What is the probability that a user follows **food creators but not dance creators**? *Express your answer using set notation (e.g., $A \setminus B$ or $A \cap B^c$) and show your calculation.*

4. Sampling and Study Design

Problem 4. How to Sample TikTok Habits Fairly?

A research team wants to estimate the **average daily TikTok usage** among undergraduates at a large university. To collect data, they:

- Obtain a complete list of all undergraduate majors offered at the university.
- Within each major, randomly select 15% of the students (for example, if Biology has 200 students, they randomly choose 30).
- Combine all selected students into their final sample.

(a) **(Identify the sampling method)** What type of sampling design is being used in this study? *Name the method and explain your reasoning using correct terminology.*

(b) **(Why use this method?)** State **one advantage** of using this sampling approach in the context of this research. *(For example: Why not just take a simple random sample from all undergraduates?)*

5. Point Estimates & Proportions

Key Concepts

- The **sample proportion** $\hat{p}$ estimates the probability or fraction of individuals in the population with a certain characteristic.
- When comparing two groups, always state both sample sizes, sample proportions, and the difference clearly.
- **Always show your formulas and plug in the numbers step by step.**

Problem 5. Estimating TikTok Posting Rates

In the study from Problem 4, $n = 300$ students were surveyed, and 72 reported posting at least one TikTok video in the last week.

(a) **(Single group)**
Calculate the **sample proportion** $\hat{p}$ of students who posted a TikTok last week.
Write the formula, plug in the values, and interpret what this number means in context.

(b) **(Comparing two groups)**
Suppose a second independent group of $n = 400$ students is surveyed and 88 posted a video.
Find the sample proportion for this second group and compute the **difference in sample proportions** $(\hat{p}_1 - \hat{p}_2)$.
Write out each proportion, the subtraction, and briefly interpret what the difference tells you about the two groups.

6. Confidence Intervals for Mean and Proportion

Key Concepts

- A confidence interval gives a plausible range of values for a population parameter based on sample data.
- Always specify the parameter, the formula you're using, and interpret the result in context.
- For large n (like $n = 60$), use the z-interval for means: $\overline{x} \pm z^* \dfrac{s}{\sqrt{n}}$

Problem 6. Estimating TikTok Screen Time

A random sample of $n = 60$ TikTok users had a **mean daily screen time** of $\overline{x} = 100$ minutes, with a sample standard deviation $s = 20$ minutes.

(a) **(Calculation)** Construct a 95% confidence interval for the **mean daily screen time** for all TikTok users at this university. *Use the z-interval formula for means, and use $z^* \approx 2$ for simplicity. Show your work step by step: formula, substitution, interval.*

$$\overline{x} \pm z^* \frac{s}{\sqrt{n}}$$

(b) **(Interpretation)** Interpret your confidence interval in the context of this study. *Be specific about the population and what the numbers mean.*

Problem 7. Estimating Posting Rates

Recall from Problem 5: Out of $n = 300$ surveyed students, 72 reported posting at least one TikTok last week.

(a) Construct a 95% confidence interval for the **proportion** of all students at this university who posted a TikTok last week. *Use the standard normal formula for proportions, $z^* \approx 2$, and show your calculation.*

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(b) (Optional challenge) How would your answer change if the sample size were much smaller (e.g., $n = 15$)? Briefly explain.

7. Hypothesis Testing: Proportion and Mean

Key Concepts

- Hypothesis tests help you decide if a sample provides enough evidence to support a claim about a population.
- Always define the parameter, state hypotheses in both symbols and words, and follow a structured process.
- For large samples, use the z-test for proportions and means; use t-test if the population standard deviation is unknown and n is not large.

Problem 8. Testing a TikTok Posting Rate Claim

A TikTok influencer claims that **more than 30% of college students** have posted at least one video in the past week. Recall from Problem 5: Out of $n = 300$ students surveyed, 72 reported posting a video.

Test this claim at the $\alpha = 0.05$ significance level.

(a) **(State hypotheses)** Define the population parameter and write the null (H_0) and alternative (H_a) hypotheses both in symbols and in plain English.

(b) **(Test statistic)** Calculate the sample proportion $\hat{p}$ and compute the test statistic using the z-test for a single proportion:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Show your substitution and calculation steps.

(c) **(Decision & reasoning)** Without calculating the exact p-value, decide whether to reject H_0 or not. *Justify your answer by comparing your test statistic to the critical value for $\alpha = 0.05$ (one-sided test). Briefly interpret what your result means in this TikTok context.*

Problem 9. Testing for Higher TikTok Screen Time

Suppose the **national average daily TikTok screen time is 90 minutes**. Recall from Problem 6: A sample of $n = 60$ users at this university had $\overline{x} = 100$ minutes, $s = 20$ minutes.

Test if this university's students use TikTok more than the national average at the $\alpha = 0.05$ level.

(a) **(State hypotheses)** Clearly define the population mean and write the null and alternative hypotheses in both symbols and words.

(b) **(Test statistic)** Compute the test statistic using the one-sample z-test for means (use $z^* \approx 2$ for $\alpha = 0.05$):

$$Z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Show your work.

(c) **(Decision & interpretation)** Based on your test statistic and the critical value, state your decision about the null hypothesis. *Interpret your result in the context of TikTok screen time for this university.*

8. Statistical Errors and Significance Level

Key Concepts

- In hypothesis testing, there are two possible types of errors:
  - **Type I error**: Rejecting the null hypothesis when it is actually true.
  - **Type II error**: Failing to reject the null hypothesis when it is actually false.
- The significance level α is the probability of making a Type I error.

Problem 10. Understanding Hypothesis Test Errors (in TikTok Context)
Recall the hypotheses from Problem 8:

- H_0: The proportion of college students who posted a TikTok last week is *at most* 0.30.
- H_a: The proportion is *greater than* 0.30.

(a) **(Type I error in context)** In your own words, describe what a **Type I error** would mean in the context of this TikTok posting study.

(b) **(Type II error in context)** In your own words, describe what a **Type II error** would mean in the context of this TikTok posting study.

(c) **(Significance level)** Which type of error is directly controlled by the significance level α? Briefly explain.

9. Simple Linear Regression

Key Concepts and Tips

- Simple linear regression models the relationship between an explanatory variable x (predictor) and a response variable y (outcome).
- The regression equation is $\hat{y} = b_0 + b_1 x$, where:
  - b_1 (slope): How much y is predicted to change for a one-unit increase in x.
  - b_0 (intercept): The predicted value of y when $x = 0$.
  - R^2: The proportion of variability in y explained by x.
- Use $b_1 = r\dfrac{s_y}{s_x}$ and $b_0 = \overline{y} - b_1\overline{x}$.

Problem 11. Predicting TikTok Likes from Follower Counts
A research team studies $n = 12$ TikTok users. For each user, they record:

- x = number of followers,
- y = average number of likes per video.

The sample statistics are:

- Mean followers: $\overline{x} = 5{,}000$ Standard deviation: $s_x = 2{,}000$
- Mean likes: $\overline{y} = 400$ Standard deviation: $s_y = 100$
- Correlation: $r = 0.8$

(a) **(Calculating the line)** Calculate the **slope** (b_1) and **intercept** (b_0) of the least-squares regression line for predicting likes from followers. *Show the formulas, substitution, and results.*

(b) **(Writing the regression equation)** Write the equation of the regression line in the form $\hat{y} = b_0 + b_1 x$.

(c) **(Interpreting the slope)** Interpret the slope (b_1) in the context of this study. *What does it mean about the relationship between followers and likes?*

(d) **(R^2 calculation and interpretation)** Calculate R^2 and interpret its meaning in the context of TikTok likes and followers.

10. Reading and Interpreting Two-Way Tables

Key Concepts

- Two-way tables help organize counts and allow you to calculate probabilities for combined events.
- Always start by calculating the **total sample size**, row totals, or column totals as needed.
- Express your answers as probabilities (e.g., $\frac{\text{count}}{\text{total}}$) and simplify where possible.

Problem 12. TikTok Posting by Gender: Exploring a Two-Way Table

A survey of college students records their self-identified gender and TikTok posting frequency over the last month. The results are summarized below:

	Never	Sometimes	Often
Female	30	25	15
Male	35	20	10

Note: Begin by calculating the total number of students in the sample.

(a) **(Joint Probability)** What is the probability that a randomly selected student from this sample is **male and posts TikToks often**? *Show your setup and calculation.*

(b) **(Union Probability)** What is the probability that a randomly selected student is **female or posts TikToks often**? *Be careful: Remember to use the general addition rule for "or." Show your setup, intermediate steps, and calculation.*

(c) **(Conditional Probability)** What is the probability that a student **posts sometimes**, *given* that they are female? *Write your answer as* $P(\textit{Sometimes} \mid \textit{Female})$ *and show your calculation.*

Antonio Aguirre
Department of Statistics, UCSC