Confidence Intervals, T-tests, and T/Z Distributions - Practice Questions

Confidence Intervals

1. What is a confidence interval, and how should it be interpreted in the context of data analysis?

A **confidence interval (CI)** is a range of values, calculated from sample data, that is likely to contain the true population parameter (like mean or proportion) with a certain level of confidence (e.g., 95%).

Interpretation in data analysis:
If a 95% CI for the accuracy of a model is **(0.82, 0.88)**, it means we are 95% confident that the true accuracy of the model lies between **82% and 88%**. It does **not** mean there's a 95% chance the true value is in the interval; rather, if we repeated the experiment many times, 95% of the intervals constructed would capture the true parameter.

2. If you have a 95% confidence interval for a population mean of (50, 60), what does this tell you about the population mean?

A 95% confidence interval (50, 60) for a population mean means:

We are 95% confident that the true population mean lies between 50 and 60.

◆ This does not mean there's a 95% probability that the mean is in this interval (the mean is fixed; the interval varies from sample to sample).
◆ Instead, it means that if we were to take many random samples and compute confidence intervals in the same way, about 95% of those intervals would contain the true population mean.

Example in Model Evaluation:

Suppose you are evaluating the average F1-score of a machine learning model across multiple test sets.

- If you get a 95% CI = (50, 60), you can say:
  "Based on our data, we are 95% confident the true mean F1-score of the model lies between 50% and 60%."

This interval gives a range of plausible values for the model's true performance, rather than just a single point estimate.

Would you like me to also show you how to compute such a confidence interval in Python (step by step) for a model metric?

3. **How does increasing the sample size affect the width of a confidence interval, and why?**
   Increasing the sample size makes the confidence interval **narrower**, because a larger sample reduces the standard error, giving a more precise estimate of the population mean.
4. **Explain the difference between a 95% and a 99% confidence interval in terms of accuracy and width.**

A 95% confidence interval means we are 95% confident the true population parameter lies within the interval, while a 99% confidence interval means we are 99% confident.

- Accuracy (Confidence): A 99% CI gives greater confidence (higher probability of containing the true value).
- Width: To gain that higher confidence, the interval must be wider than a 95% CI.

5. **When constructing a confidence interval for a mean, why might we use the t-distribution instead of the z-distribution?**
   We use the **t-distribution** instead of the **z-distribution** when the sample size is small (n<30n < 30) or the population standard deviation is unknown, because the t-distribution accounts for extra uncertainty and has heavier tails.

T-tests

1. **What is the purpose of a one-sample t-test, and when would you use it?**

A **one-sample t-test** is used to check whether the **mean of a sample** is significantly different from a known or hypothesized **population mean**.

You use it when:

- The population standard deviation is **unknown**.
- The sample size is **small** (n<30n < 30).
- You want to test if the sample mean differs from a specific value.

2. **Compare the use cases for independent two-sample t-tests vs paired t-tests**.

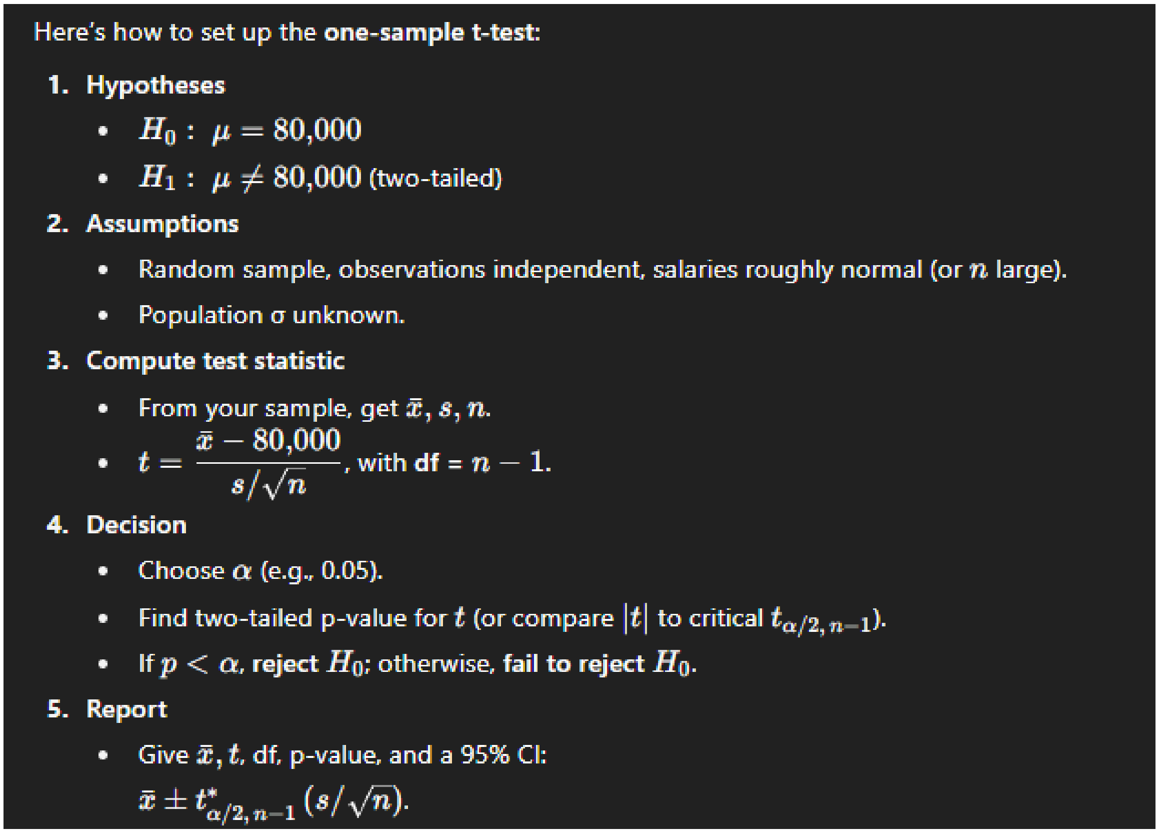
Independent two-sample t-test:

- Used to compare the means of **two independent groups** (no relation between samples).
- Example: Comparing exam scores of students from two different schools.

Paired t-test:

- Used to compare the means of **two related groups** (same subjects measured twice or matched pairs).
- Example: Comparing blood pressure of patients **before and after** taking a drug.

3. A researcher claims that the average salary of data scientists in City A is $80,000. How would you set up a t-test to verify this claim?



Here's how to set up the **one-sample t-test:**

1. **Hypotheses**
   - $H_0: \ \mu = 80{,}000$
   - $H_1: \ \mu \neq 80{,}000$ (two-tailed)
2. **Assumptions**
   - Random sample, observations independent, salaries roughly normal (or n large).
   - Population σ unknown.
3. **Compute test statistic**
   - From your sample, get $\bar{x}, s, n$.
   - $t = \dfrac{\bar{x} - 80{,}000}{s/\sqrt{n}}$, with **df** = $n - 1$.
4. **Decision**
   - Choose α (e.g., 0.05).
   - Find two-tailed p-value for t (or compare $|t|$ to critical $t_{\alpha/2,\,n-1}$).
   - If $p < \alpha$, **reject** H_0; otherwise, **fail to reject** H_0.
5. **Report**
   - Give $\bar{x}, t$, df, p-value, and a 95% CI:
     $\bar{x} \pm t^*_{\alpha/2,\,n-1}\,(s/\sqrt{n})$.

4. Why is it important to check for assumptions such as normality before running a t-test?

The t-test assumes data (or sample means) are normally distributed. If this is violated, results may be invalid—leading to wrong p-values and errors (false positives/negatives). Normality matters most with **small samples**, since the CLT helps with larger ones. If the assumption fails, use data transformation or switch to **non-parametric tests** like Mann-Whitney

5. **In an independent two-sample t-test, what does the p-value tell us about the difference between the two group means?**

The **p-value** in an independent two-sample t-test tells us the probability of observing the difference (or a more extreme one) between the two group means **if the null hypothesis (no difference) is true**.

- **Small p-value (≤ 0.05):** Strong evidence that the group means are significantly different.
- **Large p-value (> 0.05):** Not enough evidence to say the means are different.

T and Z Distributions

1. **Explain the difference between a t-distribution and a z-distribution in terms of shape, sample size, and variability.**

Shape

- *Z-distribution*: Always standard normal, bell-shaped, and fixed.
- *T-distribution*: Also bell-shaped but has heavier tails (more spread), especially for small samples.

Sample Size

- *Z-distribution*: Used when sample size is **large (n > 30)** or population variance (σ^2) is known.
- *T-distribution*: Used when sample size is **small (n < 30)** and population variance is **unknown**.

Variability

- *Z-distribution*: Variability is less since it assumes known population variance.
- *T-distribution*: More variable because it estimates variance from the sample; as sample size increases, it approaches the z-distribution.

2. **Why does the t-distribution approach the z-distribution as the sample size increases?**

The t-distribution approaches the z-distribution as sample size increases because the estimate of variability becomes more accurate, degrees of freedom increase, and by the Central Limit Theorem, the sampling distribution of the mean becomes nearly normal.

3. **When comparing a sample mean to a population mean with a known population standard deviation, which distribution should you use, and why?**

You should use the Z-distribution (standard normal distribution).

Reason:

- The population standard deviation (σ) is known.
- Z-tests are designed for situations where σ is available.
- The sampling distribution of the mean follows a normal distribution (or approximately normal if $n \geq 30$ by the Central Limit Theorem).

In contrast, the t-distribution is used when the population standard deviation is unknown and you estimate it using the sample standard deviation (s).

4. How do degrees of freedom affect the shape of a t-distribution?

Degrees of freedom (df) play a key role in shaping the t-distribution:

- Low degrees of freedom (small sample size):
  - The t-distribution is wider and has heavier tails compared to the normal (z) distribution.
  - This reflects greater uncertainty because small samples provide less reliable estimates.
- High degrees of freedom (large sample size):
  - The t-distribution becomes narrower and more closely resembles the standard normal (z) distribution.
  - With more data, the sample standard deviation is a better estimate of the population standard deviation, so less adjustment is needed.

5. Give a real-world example in data science where you would choose the t-distribution over the z-distribution.

Here's a real-world data science example where the t-distribution is preferred over the z-distribution:

Example:
Suppose you are a data scientist analyzing the average time users spend on a mobile app.

- You randomly select a sample of 30 users and record their daily usage times (in minutes).
- Since the sample size is small ($n < 30$ or around it) and the population standard deviation (σ) is unknown, you cannot use the z-distribution.
- Instead, you estimate the standard deviation using the sample standard deviation (s) and apply the t-distribution to construct a confidence interval or perform a hypothesis test about the mean usage time.

Why t-distribution here?

- Small sample size.
- Population standard deviation is unknown.
- The t-distribution accounts for the extra uncertainty introduced by estimating σ with s.

In contrast, if you had data from 10,000 users and knew σ, you would use the z-distribution.