DA-AG-006: Applied Statistics and Inference - Assignment

Question 1: What are Type I and Type II errors in hypothesis testing, and how do they impact decision-making?

Answer:
- Type I error (α): Rejecting the null hypothesis H0 when it is actually true. This is a false positive. The probability of a Type I error is controlled by the significance level α (commonly 0.05).
- Type II error (β): Failing to reject H0 when the alternative hypothesis H1 is actually true. This is a false negative. The probability of correctly rejecting a false null is the power ($1-\beta$).
Impact on decision-making:
- Trade-off: Decreasing α (making tests more conservative) typically increases β (lower power), and vice versa. Choice of α depends on consequences: if false positives are costly (e.g., claiming a drug works when it does not), choose smaller α; if false negatives are more harmful (missing a real effect), prioritize power and increase sample size to reduce β.
- Practical response: Compute required sample size to achieve desired power for a chosen effect size and α; consider consequences and costs of both error types when setting α and designing experiments.

Question 2: What is the P-value in hypothesis testing, and how should it be interpreted in the context of the null hypothesis?

Answer:
- Definition: The p-value is the probability of observing data at least as extreme as what was observed, assuming the null hypothesis H0 is true.
- Interpretation: A small p-value (below the chosen α) indicates that the observed data are unlikely under H0, and provides evidence to reject H0. It is not the probability that H0 is true.
- Cautions: Do not equate p-value with effect size—report confidence intervals and effect sizes alongside p-values. Multiple testing inflates false positive rate; adjust p-values or use corrections (e.g., Bonferroni) when performing many tests.

Question 3: Explain the difference between a Z-test and a T-test, including when to use each.

Answer:
- Z-test: Used when the population standard deviation σ is known and/or sample size is large (commonly $n \geq 30$ so CLT applies). The test statistic uses the normal (Z) distribution.
- T-test: Used when σ is unknown and estimated by the sample standard deviation s, especially for small samples ($n<30$). The test statistic follows Student's t-distribution with $n-1$ degrees of freedom.
- Practical guidance: For means, use a t-test by default when σ is unknown. For large samples the t-distribution approaches normal, so z- and t-tests give similar results. Also consider one-sample vs two-sample and paired designs when choosing the specific t-test variant.

Question 4: What is a confidence interval, and how does the margin of error influence its width and interpretation?

Answer:
- Confidence interval (CI): A range computed from sample data that—under repeated sampling—would contain the true population parameter with a specified confidence level (e.g., 95%). For a sample mean: CI = sample_mean ± margin_of_error.
- Margin of error (ME): ME = critical_value × standard_error. It controls the half-width of the CI. Larger ME $\rightarrow$ wider CI (less precision); smaller ME $\rightarrow$ narrower CI (more precision).
- Factors affecting ME and width: sample size (larger n $\rightarrow$ smaller SE $\rightarrow$ smaller ME), variability in data (larger $\sigma \rightarrow$ larger ME), and confidence level (higher confidence level $\rightarrow$ larger critical value $\rightarrow$ larger ME).

- Interpretation: A 95% CI means that the procedure used to compute the interval will produce intervals that include the true parameter 95% of the time across repeated samples. It does not say there's a 95% probability the specific interval contains the parameter (frequentist interpretation).

Question 5: Describe the purpose and assumptions of an ANOVA test. How does it extend hypothesis testing to more than two groups?

Answer:
- Purpose: ANOVA (Analysis of Variance) tests whether the means of three or more groups are equal (H0: all group means are the same) by comparing between-group variability to within-group variability.
- Key assumptions:
1. Independence of observations.
2. Normality: residuals within each group are approximately normally distributed.
3. Homogeneity of variances: group variances are approximately equal.
- How it extends hypothesis testing: Instead of multiple pairwise t-tests (which inflate Type I error), ANOVA provides a single global test. If ANOVA rejects H0, follow-up post-hoc tests (e.g., Tukey HSD) identify which pairs of groups differ while controlling family-wise error.

Question 6: Write a Python program to perform a one-sample Z-test and interpret the result for a given dataset.

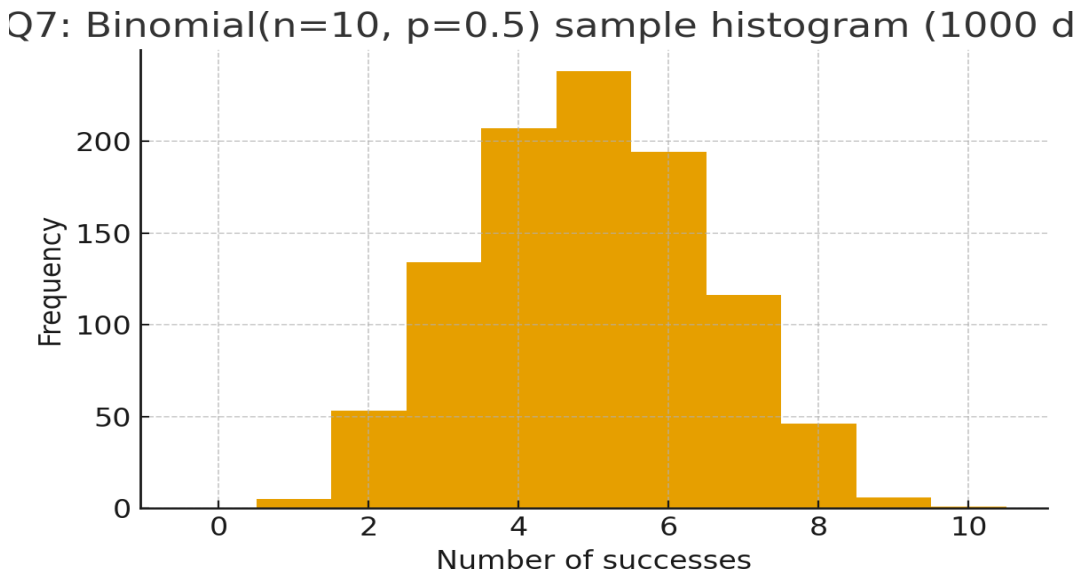
Answer:
Below is a one-sample Z-test implementation and an example. The test compares the sample mean to a known population mean $\mu 0$. We assume population standard deviation σ is known.
Example result: sample mean = 56.4286, z-statistic = 3.5211, two-sided p-value = 0.0004.

```
import numpy as np
from math import sqrt
# Example data
np.random.seed(0)
sample = np.random.normal(loc=52, scale=10, size=30)  # sample drawn from population
n = len(sample)
sample_mean = sample.mean()
sigma = 10.0  # assumed known population std dev
mu0 = 50.0
z_stat = (sample_mean - mu0) / (sigma / sqrt(n))
# two-sided p-value
p_value = 2 * (1 - stats.norm.cdf(abs(z_stat)))
sample_mean, z_stat, p_value
```

Question 7: Simulate a dataset from a binomial distribution (n = 10, p = 0.5) using NumPy and plot the histogram.

Answer: Code and histogram (1000 draws) are provided below.

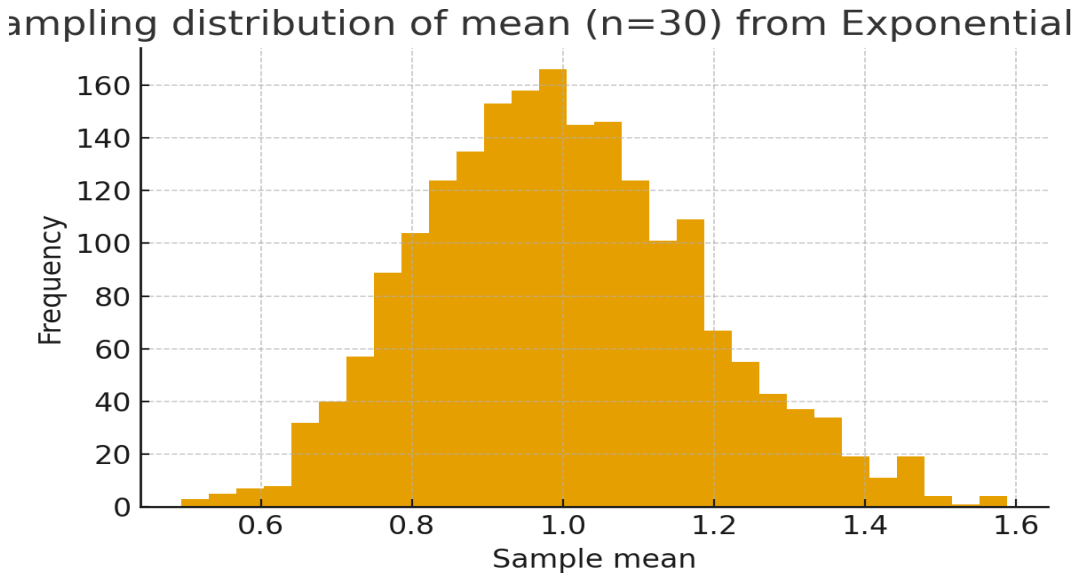
```
import numpy as np
samples = np.random.binomial(10, 0.5, size=1000)
# plt.hist(samples)  # histogram saved as q7_binomial_hist.png
```



Question 8: Generate multiple samples from a non-normal distribution and implement the Central Limit Theorem using Python.

Answer: We used an exponential population, drew 2000 samples of size 30 and plotted the sampling distribution of the mean. The histogram shows approximate normality of sample means.

```
import numpy as np
sample_means = [np.mean(np.random.exponential(scale=1.0, size=30)) for _ in range(2000)]
# histogram saved as q8_clt_hist.png
```



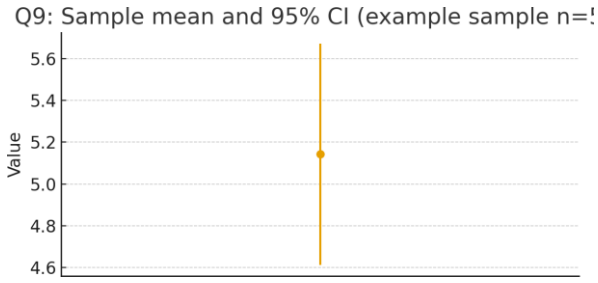
Question 9: Write a Python function to calculate and visualize the confidence interval for a sample mean.

Answer: Example with a sample of n=50 drawn from N(5, 2^2).
Sample mean = 5.1423; 95% CI ≈ (4.6121, 5.6725).
Code (computes t-based 95% CI):

```
import numpy as np
from scipy import stats

def mean_confidence_interval(data, confidence=0.95):
```

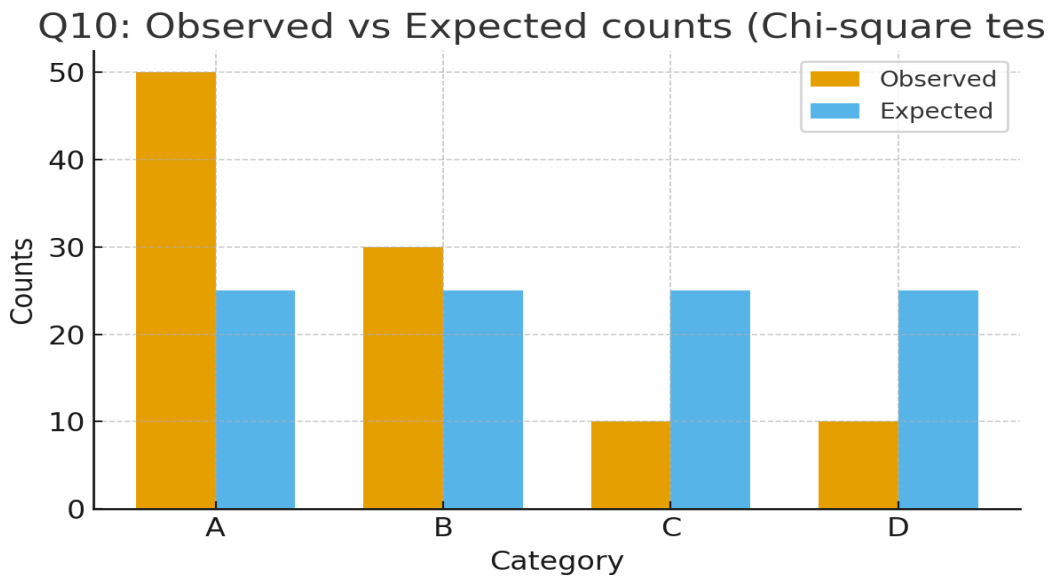
```
n = len(data)
mean = np.mean(data)
se = np.std(data, ddof=1)/np.sqrt(n)
h = se * stats.t.ppf((1+confidence)/2., n-1)
return mean, mean-h, mean+h
```



Question 10: Perform a Chi-square goodness-of-fit test using Python to compare observed and expected distributions, and explain the outcome.

Answer:
Observed counts = [50, 30, 10, 10]
Expected counts (uniform) = [25.0, 25.0, 25.0, 25.0]
Chi-square statistic = 44.0000; p-value = 0.0000.
Interpretation: With p-value = 0.0000, at α = 0.05 we reject the null hypothesis that observed frequencies come from the expected (uniform) distribution.

```
from scipy import stats
observed = np.array([50,30,10,10])
expected = np.array([0.25,0.25,0.25,0.25]) * observed.sum()
chi2_stat, p_val = stats.chisquare(f_obs=observed, f_exp=expected)
# chi2_stat, p_val
```



.