Hypothesis Testing – Quiz

Statistical Data Analysis

Question 1

In a hypothesis test, $H_0 : \mu = 50$ and $H_1 : \mu \neq 50$. A sample of $n = 25$ yields $\bar{x} = 53$ with $s = 10$. What is the test statistic t?

A. $t = 0.30$
B. $t = 1.50$
C. $t = 3.00$
D. $t = 0.75$

Question 1

In a hypothesis test, $H_0 : \mu = 50$ and $H_1 : \mu \neq 50$. A sample of $n = 25$ yields $\bar{x} = 53$ with $s = 10$. What is the test statistic t?

A. $t = 0.30$
B. $t = 1.50$
C. $t = 3.00$
D. $t = 0.75$

Answer: B
The one-sample t-statistic is $t = (\bar{x} - \mu_0)/(s/\sqrt{n}) = (53 - 50)/(10/\sqrt{25}) = 3/2 = 1.50$. The denominator $s/\sqrt{n}$ is the standard error of the mean.

Question 2

A researcher sets $\alpha = 0.05$ for a two-tailed test and obtains $p = 0.03$. The 95% confidence interval for the mean is $[51.2, 58.8]$ while $H_0 : \mu = 50$. Which statement is correct?

A. The p-value proves that H_0 is false

B. We need a one-tailed test to draw any conclusion

C. The CI contains 50, so we cannot reject H_0

D. The CI excludes 50, which is consistent with rejecting H_0 at $\alpha = 0.05$

Question 2

A researcher sets $\alpha = 0.05$ for a two-tailed test and obtains $p = 0.03$. The 95% confidence interval for the mean is $[51.2, 58.8]$ while $H_0 : \mu = 50$. Which statement is correct?

A. The p-value proves that H_0 is false

B. We need a one-tailed test to draw any conclusion

C. The CI contains 50, so we cannot reject H_0

D. The CI excludes 50, which is consistent with rejecting H_0 at $\alpha = 0.05$

Answer: D

Since $p = 0.03 < 0.05$, we reject H_0. Correspondingly, the 95% CI $[51.2, 58.8]$ does not contain $\mu_0 = 50$. A confidence interval that excludes the null value is always consistent with rejecting H_0 at the matching significance level.

Question 3

A drug company tests whether a new medication lowers blood pressure. They conclude the drug is effective when in reality it has no effect. What type of error did they commit?

A. Type I error (false positive)
B. Type II error (false negative)
C. No error was committed
D. A power error

Question 3

A drug company tests whether a new medication lowers blood pressure. They conclude the drug is effective when in reality it has no effect. What type of error did they commit?

A. Type I error (false positive)
B. Type II error (false negative)
C. No error was committed
D. A power error

Answer: A
They rejected a true H_0 (the drug truly has no effect), which is a Type I error or false positive. The probability of this error is bounded by the chosen significance level α.

Question 4

A clinical trial fails to detect a real treatment effect with $p = 0.12$ at $\alpha = 0.05$. The true effect size is Cohen's $d = 0.3$ and the sample had only 20 subjects per group. What is the most likely explanation?

A. The significance level was set too high

B. The test statistic was computed incorrectly

C. The study was underpowered to detect a small effect with this sample size

D. Type I error inflated the p-value

Question 4

A clinical trial fails to detect a real treatment effect with $p = 0.12$ at $\alpha = 0.05$. The true effect size is Cohen's $d = 0.3$ and the sample had only 20 subjects per group. What is the most likely explanation?

A. The significance level was set too high

B. The test statistic was computed incorrectly

C. The study was underpowered to detect a small effect with this sample size

D. Type I error inflated the p-value

Answer: C

With a small effect ($d = 0.3$) and only 20 subjects per group, the study likely had insufficient statistical power. Detecting small effects requires much larger samples; this is a classic Type II error scenario caused by inadequate power.

Question 5

A test yields $p = 0.001$. Which interpretation is correct?

A. There is a 0.1% probability that H_0 is true
B. The effect size must be very large
C. We have proven that H_1 is true beyond doubt
D. If H_0 were true, the probability of observing data this extreme or more is 0.001

Question 5

A test yields $p = 0.001$. Which interpretation is correct?

A. There is a 0.1% probability that H_0 is true

B. The effect size must be very large

C. We have proven that H_1 is true beyond doubt

D. If H_0 were true, the probability of observing data this extreme or more is 0.001

Answer: D

The p-value is the probability of obtaining results as extreme as or more extreme than the observed data, assuming H_0 is true. It is explicitly not the probability that H_0 is true, nor does it directly measure effect size.

Question 6

A researcher conducts 20 independent hypothesis tests, each at $\alpha = 0.05$, and finds 3 significant results. All null hypotheses are actually true. Approximately how many false positives would we expect?

A. 1
B. 0
C. 5
D. 20

Question 6

A researcher conducts 20 independent hypothesis tests, each at $\alpha = 0.05$, and finds 3 significant results. All null hypotheses are actually true. Approximately how many false positives would we expect?

A. 1
B. 0
C. 5
D. 20

Answer: A
When all null hypotheses are true, we expect $20 \times 0.05 = 1$ false positive on average. Finding 3 significant results exceeds this expectation, suggesting some may be genuine or the researcher was somewhat unlucky.

Question 7

A study reports a two-sample t-test result: $t(48) = 2.05$, $p = 0.046$, Cohen's $d = 0.12$, with $n = 25$ per group. What is the best assessment?

A. The result is both statistically and practically significant

B. The result is statistically significant but the effect is trivially small, so practical significance is questionable

C. The result is not statistically significant because d is small

D. The test should be repeated with a larger sample before drawing conclusions

Question 7

A study reports a two-sample t-test result: $t(48) = 2.05$, $p = 0.046$, Cohen's $d = 0.12$, with $n = 25$ per group. What is the best assessment?

A. The result is both statistically and practically significant

B. The result is statistically significant but the effect is trivially small, so practical significance is questionable

C. The result is not statistically significant because d is small

D. The test should be repeated with a larger sample before drawing conclusions

Answer: B

With $p = 0.046 < 0.05$ the result is statistically significant, but Cohen's $d = 0.12$ is a negligible effect size (well below the $d = 0.2$ threshold for 'small'). Large samples can make tiny, practically unimportant differences statistically significant, highlighting why effect sizes must accompany p-values.

Question 8

A one-sample t-test of heights against $\mu_0 = 170$ cm with $n = 10$ observations gives $t = 0.949$ and $p = 0.367$. What is the correct conclusion at $\alpha = 0.05$?

A. Reject H_0; the mean height is significantly different from 170 cm

B. Reject H_0; the mean height equals 170 cm

C. Fail to reject H_0; there is no significant evidence that the mean differs from 170 cm

D. Accept H_0; we have proven the population mean is exactly 170 cm

Question 8

A one-sample t-test of heights against $\mu_0 = 170$ cm with $n = 10$ observations gives $t = 0.949$ and $p = 0.367$. What is the correct conclusion at $\alpha = 0.05$?

A. Reject H_0; the mean height is significantly different from 170 cm

B. Reject H_0; the mean height equals 170 cm

C. Fail to reject H_0; there is no significant evidence that the mean differs from 170 cm

D. Accept H_0; we have proven the population mean is exactly 170 cm

Answer: C

Since $p = 0.367 > 0.05$, we fail to reject H_0. We cannot conclude the mean differs from 170 cm. Importantly, failing to reject H_0 does not prove it is true; it only means we lack sufficient evidence against it.

Question 9

For the one-sample t-test formula $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, what are the degrees of freedom?

A. $df = n - 1$
B. $df = n + 1$
C. $df = n$
D. $df = n - 2$

Question 9

For the one-sample t-test formula $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, what are the degrees of freedom?

A. $df = n - 1$
B. $df = n + 1$
C. $df = n$
D. $df = n - 2$

Answer: A
The one-sample t-test has $df = n - 1$ degrees of freedom because one degree of freedom is lost when estimating the sample mean $\bar{x}$ from the data. For example, with $n = 10$ observations, $df = 9$.

Question 10

In a weight-loss study, 6 subjects are measured before and after an intervention. The mean difference is 5.17 kg with $t = 5.77$, $df = 5$, and $p = 0.002$. Which test was used?

A. One-sample t-test
B. Independent two-sample t-test
C. One-way ANOVA
D. Paired t-test

Question 10

In a weight-loss study, 6 subjects are measured before and after an intervention. The mean difference is 5.17 kg with $t = 5.77$, $df = 5$, and $p = 0.002$. Which test was used?

A. One-sample t-test
B. Independent two-sample t-test
C. One-way ANOVA
D. Paired t-test

Answer: D
Before-and-after measurements on the same subjects require a paired t-test, which analyzes the within-subject differences. The degrees of freedom $df = n - 1 = 5$ confirm 6 pairs, consistent with a paired design.

Question 11

Two independent groups are compared: control ($n_1 = 6$, $\bar{x}_1 = 22.8$) and treatment ($n_2 = 6$, $\bar{x}_2 = 29.5$). A pooled two-sample t-test gives $t = 5.99$ with $p < 0.001$. What are the degrees of freedom?

A. $df = 5$
B. $df = 10$
C. $df = 11$
D. $df = 12$

Question 11

Two independent groups are compared: control ($n_1 = 6$, $\bar{x}_1 = 22.8$) and treatment ($n_2 = 6$, $\bar{x}_2 = 29.5$). A pooled two-sample t-test gives $t = 5.99$ with $p < 0.001$. What are the degrees of freedom?

A. $df = 5$
B. $df = 10$
C. $df = 11$
D. $df = 12$

Answer: B
For a pooled (equal-variance) two-sample t-test, $df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$. The two degrees of freedom are lost because we estimate a mean from each of the two samples.

Question 12

Welch's t-test is preferred over the pooled two-sample t-test when:

A. Both groups have the same variance and equal sample sizes
B. The data is paired rather than independent
C. The group variances are unequal or sample sizes differ substantially
D. The sample sizes are both greater than 100

Question 12

Welch's t-test is preferred over the pooled two-sample t-test when:

A. Both groups have the same variance and equal sample sizes

B. The data is paired rather than independent

C. The group variances are unequal or sample sizes differ substantially

D. The sample sizes are both greater than 100

Answer: C

Welch's t-test does not assume equal variances and uses a more complex degrees-of-freedom formula (Welch-Satterthwaite). It is preferred when a variance ratio test (e.g., F-test) suggests unequal variances or when group sizes differ substantially, making the equal-variance assumption risky.

Question 13

An ANOVA of three teaching methods yields $F(2, 12) = 31.7$ with $p < 0.001$. What can we conclude?

A. All three methods produce equal mean scores
B. Method B is significantly better than Methods A and C
C. The within-group variance exceeds the between-group variance
D. At least one teaching method mean differs significantly from the others

Question 13

An ANOVA of three teaching methods yields $F(2, 12) = 31.7$ with $p < 0.001$. What can we conclude?

A. All three methods produce equal mean scores

B. Method B is significantly better than Methods A and C

C. The within-group variance exceeds the between-group variance

D. At least one teaching method mean differs significantly from the others

Answer: D

A significant ANOVA F-test only tells us that at least one group mean differs from the others. It does not identify which specific pairs differ; that requires follow-up post-hoc tests such as Tukey HSD.

Question 14

In an ANOVA table, $SSB = 338.5$, $SSW = 64.0$, $df_{between} = 2$, and $df_{within} = 12$. What is MSW (mean square within)?

A. 5.3
B. 32.0
C. 169.3
D. 64.0

Question 14

In an ANOVA table, $SSB = 338.5$, $SSW = 64.0$, $df_{between} = 2$, and $df_{within} = 12$. What is MSW (mean square within)?

A. 5.3
B. 32.0
C. 169.3
D. 64.0

Answer: A
Mean square within is computed as $MSW = SSW/df_{within} = 64.0/12 \approx 5.3$. This represents the average variability within groups and serves as the denominator of the F-ratio.

Question 15

Levene's test for the ANOVA on three teaching methods yields $p = 0.82$. What does this imply?

A. The group means are equal
B. Welch's ANOVA should be used instead
C. The normality assumption is violated
D. There is no significant evidence of unequal variances, so standard ANOVA is appropriate

Question 15

Levene's test for the ANOVA on three teaching methods yields $p = 0.82$. What does this imply?

A. The group means are equal

B. Welch's ANOVA should be used instead

C. The normality assumption is violated

D. There is no significant evidence of unequal variances, so standard ANOVA is appropriate

Answer: D

Levene's test assesses homogeneity of variances. A large p-value ($p = 0.82$) means we fail to reject the null of equal variances, so the equal-variance assumption of standard ANOVA appears reasonable. If Levene's test had been significant, Welch's ANOVA would be the safer alternative.

Question 16

Tukey HSD applied after a significant ANOVA on three groups (A, B, C) gives adjusted p-values: B−A: $p = 0.0003$, C−A: $p = 0.0008$, C−B: $p = 0.0001$. Which pairs are significantly different at $\alpha = 0.05$?

A. Only B−A and C−B

B. Only C−B

C. All three pairs: B−A, C−A, and C−B

D. None, because the adjusted p-values are too conservative

Question 16

Tukey HSD applied after a significant ANOVA on three groups (A, B, C) gives adjusted p-values: B−A: $p = 0.0003$, C−A: $p = 0.0008$, C−B: $p = 0.0001$. Which pairs are significantly different at $\alpha = 0.05$?

A. Only B−A and C−B

B. Only C−B

C. All three pairs: B−A, C−A, and C−B

D. None, because the adjusted p-values are too conservative

Answer: C

All three adjusted p-values are well below 0.05, so every pairwise comparison is statistically significant. Tukey HSD already controls the family-wise error rate, so no further correction is needed when interpreting these adjusted p-values.

Question 17

With 4 groups and $\alpha = 0.05$ per comparison, the family-wise error rate without correction is $1 - (1 - 0.05)^6 \approx 0.265$. If we apply the Bonferroni correction, what adjusted significance threshold should each individual test use?

A. $\alpha_{adj} = 0.05/4 = 0.0125$
B. $\alpha_{adj} = 0.05/6 \approx 0.0083$
C. $\alpha_{adj} = 0.05/24 \approx 0.0021$
D. $\alpha_{adj} = 0.265/6 \approx 0.044$

Question 17

With 4 groups and $\alpha = 0.05$ per comparison, the family-wise error rate without correction is $1 - (1 - 0.05)^6 \approx 0.265$. If we apply the Bonferroni correction, what adjusted significance threshold should each individual test use?

A. $\alpha_{adj} = 0.05/4 = 0.0125$
B. $\alpha_{adj} = 0.05/6 \approx 0.0083$
C. $\alpha_{adj} = 0.05/24 \approx 0.0021$
D. $\alpha_{adj} = 0.265/6 \approx 0.044$

Answer: B

Bonferroni divides α by the number of tests m, not the number of groups. With 4 groups there are $\binom{4}{2} = 6$ pairwise comparisons, so $\alpha_{adj} = 0.05/6 \approx 0.0083$. This is more conservative than Tukey HSD but straightforward to apply.

Question 18

Which of the following is NOT a factor that increases statistical power?

A. Decreasing the significance level from $\alpha = 0.05$ to $\alpha = 0.01$

B. Increasing the sample size

C. Using a paired design instead of independent samples when subjects can be matched

D. Reducing measurement variability through better instruments

Question 18

Which of the following is NOT a factor that increases statistical power?

A. Decreasing the significance level from $\alpha = 0.05$ to $\alpha = 0.01$

B. Increasing the sample size

C. Using a paired design instead of independent samples when subjects can be matched

D. Reducing measurement variability through better instruments

Answer: A

Decreasing α makes it harder to reject H_0, which reduces power while lowering the Type I error rate. Increasing sample size, using paired designs, and reducing variability all increase power by making the signal easier to detect relative to the noise.

Question 19

A power analysis for a two-sample t-test with Cohen's $d = 0.5$, $\alpha = 0.05$, and target power $= 0.80$ yields a required sample size of $n = 64$ per group. If the researcher can only recruit 30 per group, what happens to the achieved power?

A. Power increases above 0.80 because smaller samples are more sensitive

B. Power remains at 0.80 regardless of sample size

C. Power depends only on effect size, not on sample size

D. Power drops well below 0.80, increasing the risk of a Type II error

Question 19

A power analysis for a two-sample t-test with Cohen's $d = 0.5$, $\alpha = 0.05$, and target power $= 0.80$ yields a required sample size of $n = 64$ per group. If the researcher can only recruit 30 per group, what happens to the achieved power?

A. Power increases above 0.80 because smaller samples are more sensitive

B. Power remains at 0.80 regardless of sample size

C. Power depends only on effect size, not on sample size

D. Power drops well below 0.80, increasing the risk of a Type II error

Answer: D

Power is directly related to sample size. With only 30 per group instead of the required 64, the study is substantially underpowered. This increases the probability of a Type II error, meaning a real medium effect ($d = 0.5$) is more likely to go undetected.

Question 20

A clinical trial comparing 3 treatments ($n = 20$ per group) reports $F(2, 57) = 8.42$, $p = 0.0006$, and $\eta^2 = 0.228$. Post-hoc power is 0.84. What is the best interpretation of $\eta^2 = 0.228$?

A. The p-value is 0.228 after adjustment

B. 22.8% of the total variance in pain scores is explained by treatment group differences

C. The probability of a Type II error is 22.8%

D. Each group mean differs by 22.8% from the grand mean

Question 20

A clinical trial comparing 3 treatments ($n = 20$ per group) reports $F(2, 57) = 8.42$, $p = 0.0006$, and $\eta^2 = 0.228$. Post-hoc power is 0.84. What is the best interpretation of $\eta^2 = 0.228$?

A. The p-value is 0.228 after adjustment

B. 22.8% of the total variance in pain scores is explained by treatment group differences

C. The probability of a Type II error is 22.8%

D. Each group mean differs by 22.8% from the grand mean

Answer: B

Eta-squared (η^2) is the ratio SSB/SST, representing the proportion of total variance explained by the grouping factor. Here, 22.8% of the variability in pain scores is attributable to treatment differences, which is considered a large effect by Cohen's guidelines ($\eta^2 > 0.14$).