**BSB123 Data Analysis**

**Practice Quiz**

**Question One**

Scores on a government psychometric test are normally distributed with a mean of 150 and a standard deviation of 25.

1. What is the probability an applicant scores more than 120 on the test?

P(X > 120) = 1 – normdist(120, 150, 25, true) = 0.8849

1. What score would the top 20% of applicants achieve?

Top 20% = norm.inv(0.8,150,25) = 171.041

**Question Two**

Eight percent of the worlds population have blue eyes.

1. What is the probability that in a sample of 200 students more than 10% have blue eyes?

n = 200

p=0.08

np = 16, nq = 184, and since both > 10 can use normal distribution to find probability for a sample proportion

P($\hat{p}>0.1)$

$$σ\_{\hat{p}}=\sqrt{\frac{pq}{n}}=\sqrt{\frac{(0.08)(0.92)}{200}}=0.06$$

P($\hat{p}>0.1)$ = 1 – norm.dist(0.1, 0.08, 0.06, true) = 1 – 0.6306 = 0.3694

1. Answer the same question if the sample was only 100 students.

n = 100

p=0.08

np = 8, nq = 92. Since np < 10 we have to go back to the binomial distribution for this question.

If we want to find the probability that more then 10% have blue eyes out of 100 surveyed that is:

P(x > 10 / n=100, p = 0.08) = 1 – binom.dist(10, 100, 0.08, true) = 1 – 0.0.8243 = 0.1757

**Question Three**

In your excel spreadsheet there is a worksheet on Instagram v Facebook usage for different age groups.

1. Conduct a test at the 5% level of significance if Instagram usage among Under 50’S is more than 60%.

Ho: p ≤ 0.6

H1: p > 0.6

Decision Rule: Reject Ho if p – value < 0.05

From worksheet

Number of U50’s n = 200

p = 0.6

np = 120, nq = 80 it is valid to use z distribution for the test

p – value = 1 – norm.dist($\hat{p}, p, σ\_{\hat{p}, } true)$

x = number of U50’s using Instagram = 130

$$\hat{p}= \frac{x}{n}=\frac{130}{200}=0.65$$

$$σ\_{\hat{p}}=\sqrt{\frac{pq}{n}}=\sqrt{\frac{(0.6)(0.4)}{200}}=0.00346$$

P – value = 1 – norm.dist(0.65, 0.6, 0.0346, true) = 1 – 0.9255 = 0.0745

p-value is not less than α = 0.05, we can not reject the Ho and hence we can say the proportion of U50’s using Instagram is not more than 60%.

1. Conduct a test at the 1% level of significance if the proportion of Under 50’s who use Instagram is more than the proportion of 50 and overs.

Let 1 = U50’s

2 = 50 and over

H0: p1 ≤ p2 ⇒ p1 – p2 ≤ 0

H1: p1 > p2 ⇒ p1 – p2 > 0

Note to conduct this test using the normal distribution we need to check if the sampling distribution of the difference in the sample proportion is normally distributed. For this to be true n1p1 , n1(1 – p1), n2p2, and n2(1 – p2) all have to be > 10. Because we do not have individual values for p1 and p2 we use the sample data which requires:

x1, (n1 – x1), x2 and (n2 – x2) all have to be > 10

n1 = 200, x1 = 130, so (n1 – x1) = 70

n2 = 150, x2 = 90, so (n2 – x2) = 60

From this we can see that we can use the normal distribution as all values are > 10.

Decision Rule: Reject Ho if p – value > α (0.01)

For an upper tail test:

P – value = 1 – norm.dist($\hat{p}\_{1}-\hat{p}\_{2}, p\_{1}-p\_{2}, σ\_{\left(\hat{p}\_{1}-\hat{p}\_{2}\right)}, true)$

$$\hat{p}\_{1}=\frac{130}{200}=0.65$$

$$\hat{p}\_{2}=\frac{90}{150}=0.60$$

$$\hat{p}\_{1}-\hat{p}\_{2}=0.65-0.6=0.05$$

$$σ\_{(\hat{p}\_{1}-\hat{p}\_{2})}=\sqrt{\hat{p}\hat{q}\left(\frac{1}{n\_{1}}+\frac{1}{n\_{2}}\right)}$$

And the pooled proportion is:

$$\hat{p}=\frac{x\_{1}+x\_{2}}{n\_{1}+n\_{2}}$$

$$\hat{p}=\frac{130+90}{200+150}=0.6286$$

$$σ\_{(\hat{p}\_{1}-\hat{p}\_{2})}=\sqrt{0.6286(0.3714)\left(\frac{1}{200}+\frac{1}{150}\right)}=0.0522$$

P – value = 1 – norm.dist(0.05, 0, 0.0522, true) = 1 – 0.831 = 0.169

Hence p-value is not < α and we can not reject the Null Hypothesis. There is insufficient evidence to say the U50’s have a higher proportion of Instagram users than those 50 and over.

**Question Four**

In your excel spreadsheet there is a worksheet called on-line retailer.

An online retailer is looking at the changing trends in sales since pre-covid and 2022.

1. In pre-covid years the standard deviation in sales was $25. While overall spending online has increased, the retailer believes that the variance in sales has also increased. Test at the 5% level of significance if the variance of sales in 2022 is more than the previous $625.

H0: σ2 ≤ 625

H1: σ2 > 625

Note for this test to be valid we need to assume that the distribution of sales is normally distributed.

Decision Rule: Reject Ho if p – value < 0.05 (from question)

This is an upper tail Chi Squared Test.

p – value = chi.dist.rt($χ\_{calc}^{2}, degrees of freedom)$

From data in spreadsheet

n = 40

s2 = 2410.402

σ2 = 625

$$χ\_{calc}^{2}=\frac{(n-1)s^{2}}{σ^{2}}=\frac{(39)(2410.402)}{625}=150.409$$

p – value = chi.dist.rt(150.409, 39) = 0

p – value is < 0.05 and hence we can reject the null hypothesis and say that the variance of retail sales has increased between 2019 and 2022.

1. Conduct a test to see if the average spend of customers in 2022 is more than that in 2019.

H0: µ19 ≥ µ22 OR µ19 - 22 ≥ 0

H1: µ19 < µ22 OR µ19 - µ22 < 0

For the t test to be valid the population must be normally distributed or the sample sizes > 30.

n19 = 35 and n22 = 40 which are both ≥ 30 and hence it is valid to use t test

As this is a test of difference in means we first need to determine if the variances of the two populations are equal or not:

$$H\_{0}: \frac{σ\_{19}^{2}}{σ\_{22}^{2}}=1$$

$$H\_{1}: \frac{σ\_{19}^{2}}{σ\_{22}^{2}}\ne 1$$

This is an F test of Variances which can be done in Excel.

Decision Rule: Reject Ho if p – value < α (5% chosen because no value in question)

Using F.Test(array 1, array2) we get

P – value = 0.00025

This is less than α (0.05) and hence we can conclude variances are not equal.

Based on this result we can now use Excel to conduct a test of means assuming unequal variances

Excel Output

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Unequal Variances |   |   |
|   |   |   |
|  | *2019* | *2022* |
| Mean | 125.9429 | 183.495 |
| Variance | 670.7896 | 2410.401513 |
| Observations | 35 | 40 |
| Hypothesized Mean Difference | 0 |   |
| df | 61 |   |
| t Stat | -6.45776 |   |
| P(T<=t) one-tail | 9.82E-09 |   |
| t Critical one-tail | 1.670219 |   |
| P(T<=t) two-tail | 1.96E-08 |   |
| t Critical two-tail | 1.999624 |   |

Our test of the means is a one tail test and hence p – value = 9.8E-9 = 0

P – value is less than 0 and hence we can conclude that the average value of retail sales has increased in 2022 compared to 2019