Analysis of Scientific Data – Summer Semester, 2022

Week 4 Statistics Half of the Tutorial Solutions

Part A

Alice recruited nine pairs of identical twins for a study of two cholesterol-reducing drugs, A and B, with the aim of showing that drug A gave higher reductions in cholesterol than drug B. One of the twins in each pair was given drug A and the other was given drug B, where the choice was made at random. The amount by which cholesterol was reduced in each subject (mg/dL) is given in the following table:

Pair	1	2	3	4	5	6	7	8	9
Drug A	74	55	61	47	53	74	52	40	50
Drug B	63	58	49	41	50	69	59	31	44
Difference	11	-3	12	6	3	5	-7	9	6

The sample mean difference in cholesterol reduction between drug A and drug B was 4.67 mg/dL with a sample standard deviation of 6.265 mg/dL.

a) State the null and alternative hypotheses of interest in terms of μ, the mean difference in cholesterol reduction between drug A and drug B in twins in this population.

$H_0: \mu = 0$

$H_1: \mu > 0$

b) Calculate the t statistic to test this null hypothesis.

$$t_{stat} = \frac{4.67 - 0}{6.265/\sqrt{9}}$$

$$t_{stat} = 2.236233$$

c) What is the corresponding p-value? What do you conclude?

$$p - value = P(t_{stat} \geq 2.236233)$$

$p - value = 0.0279$ [In RStudio: `1-pt(2.236233, df=8)`]

There is moderate evidence to conclude that drug A does give a higher reduction in cholesterol.

d) Based on this data, calculate a 95% confidence interval for the mean difference in cholesterol reduction between drug A and drug B.

95% CI for μ is:

$\overline{X} \pm t^*_{(n-1)} \frac{s}{\sqrt{n}}$ where $t^*_{(n-1)} = 2.306004$ [In RStudio: `qt(0.975, df=8)`]

$4.67 \pm 2.306004 \times \frac{6.265}{\sqrt{9}}$

4.67 ± 4.815704

(-0.1457,9.4857)

e) Suppose we wanted to carry out a new study that could estimate the mean difference in cholesterol reduction with a margin of error of 2 mg/dl. What sample size should the new study use if we keep the 95% confidence level? (Assume that population standard deviation, σ , is 6.265)

Margin of error (MOE)

$MOE = Z^* \frac{\sigma}{\sqrt{n}}$

Rearranging; $n = \left(\frac{Z^* \times \sigma}{MOE}\right)^2$ where $Z^* = 1.959964$ [In RStudio: qnorm(0.975)]

$$n = \left(\frac{1.959964 \times 6.265}{2}\right)^2$$

$$n = 37.69453$$

Thus, the required sample size = 38

Part B

A simple alternative to the *t* test in Part A is to count the number of twins where there was a positive differencebetween drug A and drug B.

a) State the null and alternative hypotheses of interest in terms of *p*, the probability that, in a random pair of identical twins, the one with drug A will have a greater reduction in cholesterol than the one with drug B.

$H_0: p = 0.5$

$H_1: p > 0.5$

b) Assuming H_0 is true, what is the distribution of *X*, the number of positive differences in 9 trials?

$X \sim Binomial(9, 0.5)$

c) Based on the observed number of positive differences, what is the *p*-value for this test? What do you conclude?

$p - value = P(X \geq 7)$

$p - value = P(X = 7) + P(X = 8) + P(X = 9)$

$p - value = 0.0898$ [In R: `sum(dbinom(x=7:9, size=9, prob=0.5))`

Only week evidence to conclude that drug A does give a higher reduction in cholesterol.

Part C

The health officials carryout a survey for women aged 20-60 years and recorded the time in hours that pain lasted at the injection side after the second covid jab. Based on an earlier survey with a similar population of women, the pain duration is assumed to follow a normal distribution with a mean of 18.3 hours and a standard deviation of 6.9 hours.

a) What is the probability that average pain duration time of a randomly selected 5 women in the survey had pains less than 12 hours?

Define X as the pain duration.
Then $X \sim Normal(18.3, 6.9)$

Then the distribution of the sample mean of 5 women is:

$$\bar{X} \sim Normal(18.3, \frac{6.9}{\sqrt{5}})$$

We need to find P($\bar{X}$<12).
$P(\bar{X} < 12) = P(Z < \frac{12 - 18.3}{6.9/\sqrt{5}})$
$P(\bar{X} < 12) = P(Z <$-2.041627)

Using pnorm(-2.041627)
[1] 0.02059427

The probability that average pain duration time less than 12 hours is 0.0206.

b) What is the shortest time of experiencing pain that would be in the upper 10% of pain durations?

In this question we want to find the value q (quantile) such that

$0.90 = P(X \leq q)$

Again, we standardize X
$0.9 = P\left(Z \leq \frac{q - 18.3}{6.9}\right)$
Using the qnorm() in R

qnorm(0.9)
[1] 1.281552

That is, $P(Z \leq 1.2816) = 0.9$.

So we need to solve the equation

$1.281552 = \frac{q - 18.3}{6.9}$

$1.281552 \times 6.9 = q - 18.3$

$q = 18.3 + 1.281552 \times 6.9 = 27.14271$

So, the shortest time of experiencing pain that would be in the upper 10% of pain durations is 27.14 hours.

c) The health officials carried out a similar survey of 20 men aged between 20 and 60 years and recorded the sample mean of 15.5 hours with a standard deviation of 5.6 hours. They believe that mean pain duration for men is less than that of women. Suppose μ is the mean pain duration for men aged between 20 and 60 years who had the second jab. Carry out a hypothesis test to test the health officials' belief.

The null and alternative hypotheses are

H_0: μ =18.3 Vs H_1: μ <18.3

We need to do a one-sample t-test to test the null hypothesis.

The t-test statistic: $t_{test} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$

$t_{test} = \frac{15.5-18.3}{5.6/\sqrt{20}}$= -2.395787

This t test statistic follows a t-distribution with (n-1) degrees of freedom.

Using pt() in R

pt(-2.395787, df=19)

[1] 0.01352045

Thus, p-value = 0.01352045

There is moderate evidence to conclude that mean pain duration for men is less than that of women after the second covid jab.