ABS1 Exam

Shunyu Wu 21-119-631

2022/1/14

QUESTION 1

1A

- Difference between parameters and estimates
  - **Parameters** are descriptive measures of a whole population. Yet they usually have unknown values because measuring an entire population is impractival. So we can acquire parameter **estimates** by taking a random sample from the population.
  - Estimates can have **standard error**. We can estimate parameters using 2 types of parameter estimates: 1. **point estimates** – usually the value of a parameter, for example the estimate population mean by sample mean. 2. **interval estimate(Confidence intevals)** – a range of values which might contain the population parameter.
- Relationship between hypothesis tests and confidence intervals in the context of a one- sample t-test
  - **1-sample t-test** compares the mean of a random sample from a normal population to the mean proposed in the null hypothesis, hypothesis testing asks how unlikely they differ. Here the **confidence interval** provides a plausible range for a parameter. Given the facts, any values for the parameter inside the interval are reasonable, whilst those outside are implausible.
  - Both confidence intervals and hypothesis testing are inferential techniques that utilize the 1 sample to estimate the mean or assess the strength and validity of a hypothesis.

1B

Null hypothesis is a specific statement(usually simplier than ulternative hypothesis) about a population parameter made for the purposes of argument.

The essence of the null hypothesis is that differences between treatments and controls can be explained by chance. So the **idea** of the null hypothesis is that the difference arises not from a genuine difference between the totals represented by the two, but from a spurious difference manifested by chance random sampling. So the logical premise of the null hypothesis is that there is no difference in the totality behind the sample. It is **useful** if the null hypothesis can be proven wrong. But it would be misleading if people don't really understand null hypothesis and the meaning of p-value.

No I don't think we should abandon classical hypothesis testing and p-values completely. A P-value represents the likelihood of receiving the data, or something equally or more remarkable, assuming the null hypothesis is true. It is confusing because what we really need to know is not the probability of observing the phenomenon when the effect is not present, but the probability of the effect being present. The p-value gives the correct answer, but is for the wrong question. It will be interesting if people understand what it really does.

QUESTION 2

When should I use covariance and when should I use correlation. What's the difference. And how can I measure the correlation between different kinds of variables (continuous, categorical, . . .)?

QUESTION 3

3A Yes. Using linear Regression method is appropriate because we can see from the data set plot that the relationship between X and Y can be described by a line.

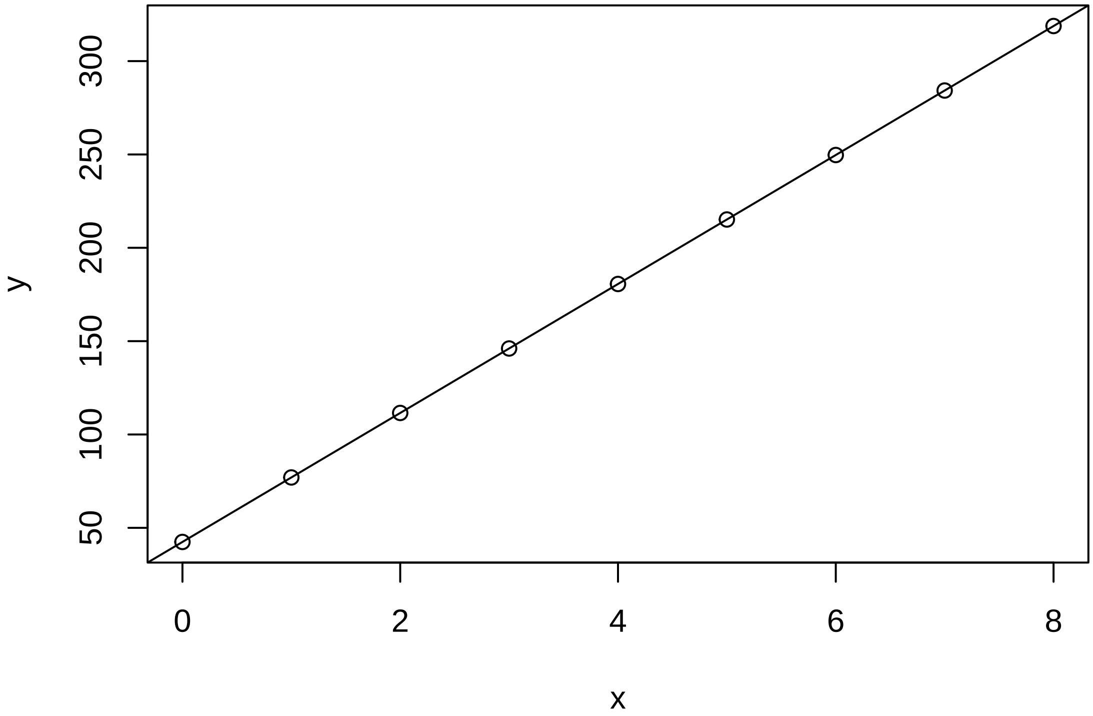
3B Yes. The conclusion is doubtful. Because it is unwise to extrapolate beyond the range of the data. In the example data set we only know the relationship within about $X \in [0, 8]$, $x = 13.2$ is beyond range. To fix this we might need more data for larger X.

3C I would perform data transformation to changes each data point by some simple mathematical formula. And let them fit better in a linear line.

3D

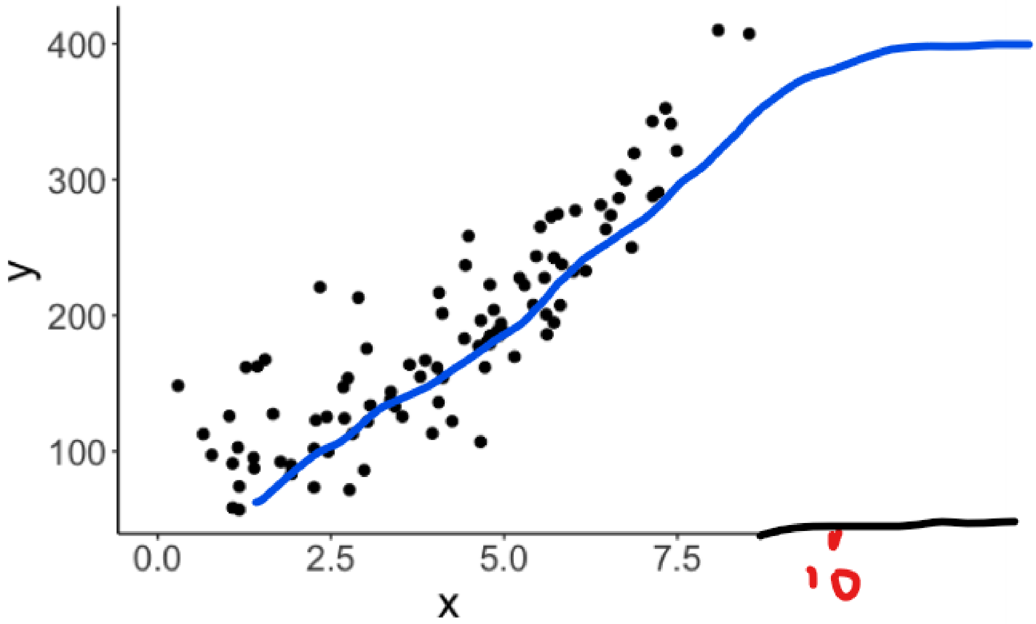
```r
# If I have data in R
# abline(lm(y ~ x, data = df), col = "red")

# Since I only have the output
x <- 0:8
y <-  42.441 + 34.544 * x
par(mar = c(4, 4, 4, 4))
plot(x,y)
abline(lm(y~x))
```



3E

In fact, it is empirically known that after reaching a threshold of learning time, the improvement in grade performance becomes less according to increased study time, making the y-x relationship more of a logarithmic than a linear.



QUESTION 4

4A

I would use 1-sample t test on the transformed data. Because as we can see from these two Q-Q plots that the transformed data is normally distributed. And the sample size is pretty small(n=20). So tests type like permutation test won't really work.

4B

Since we are supposed to test if the mean of the original raw data (= non-transformed) is significantly different from 1. We would use **one-sample t-test**, and would test whether the mean of the log-transformed data is different from 0 (log 1 = 0).

$Y = log(X), \overline{Y} = \frac{y_1 + ... + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$

H_0 : The mean of the log transfromed is 0. $\overline{Y} = 0$

QUESTION 5

5A As the Degree of Freedom $df = (N_1 - 1) + (N_2 - 1)$, the sample size $N_1 + N_2 = 4e + 05 + 2 \approx 4e + 05$

5B Yes, because p < 0.05 (p-value = 0.01457)

5c Yes, because from t-test we know there is significant difference between old and new means of tumor sizes.