[WEEK 3] Inferential Statistics - Part 2

CS5701 - Quantitative Data Analysis

Dr Sarath Dantu

Department of Computer Science
Brunel University London
sarath.dantu AT brunel.ac.uk

5th October 2022

Week 3 - Housekeeping

- I will start at 10:05 prompt
- The Lecture is being recorded ...the recording will be shared in Brightspace but it can take a few hours to make it available
- For the labs - head to TOWER A 407

Week 3 - Learning Outcomes

This week after the lecture, the lab and the independent practice you should be:

- able to express the NULL and alternative hypothesis for testing
- able to interpret the results of hypothesis testing
- familiar with the approaches to run hypothesis tests for different situations
- Use R functions to perform hypothesis testing

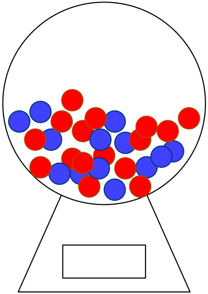
This week's material is also covered in chapter 6 of Crawley

Hypothesis Testing

- So far we have focused on one sample of data, and what can be said about its parameters
- We may want to compare parameters computed from a sample to an underlying population

Example - Gumball

Recall from last week.... what if we wanted to check if indeed 0.5 of the sweets in it are red?



Hypothesis Testing

- The **null hypothesis** or H_0 is the hypothesis that is being tested (and trying to be disproved)
- The **alternative hypothesis** or H_1 represents the alternative value.

The intuition here is: **could these observations really have occurred by chance?**

Hypothesis Testing

- The **null hypothesis** or H_0 is the hypothesis that is being tested (and trying to be disproved)
- The **alternative hypothesis** or H_1 represents the alternative value.

We want to quantify: **"How likely is our sample if what we know about the population is true?** - we will refer to this measure as the **p-value**.

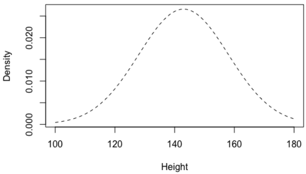
Hypothesis Testing

- The **null hypothesis** or H_0 is the hypothesis that is being tested (and trying to be disproved)
- The **alternative hypothesis** or H_1 represents the alternative value.

The intuition here is: **could these observations really have occurred by chance?**
For example if we believe the mean height of a 12 year old child is 143cm, our sample has mean height of 148cm.... how likely would this mean be, if we sample from a Normal distribution with a mean of 143cm?

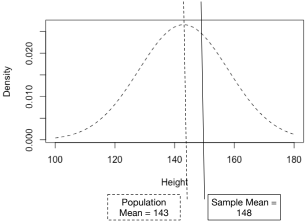
Hypothesis testing the mean - height example

- We assume that the distribution of 12 year old children's heights follows the density $\rightarrow$
- This density has a mean of 143 (population mean)
- How can we use a sample mean and this distribution to answer the question asked?



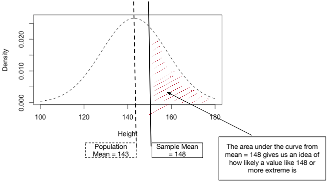
Hypothesis testing the mean - height example

- This density gives us an idea of how frequent or likely the x-axis values are
- When we sum the area under the curve from a certain x value or between two x values we can quantify how likely such values are



Hypothesis testing the mean - height example

- if the population mean is 143 and it has a distribution as $\rightarrow$
- and If we wanted to know how likely is a sample where the mean is 148
- We would want the sum of the area under the curve from 148 and up (shaded area)



Hypothesis testing the mean - height example

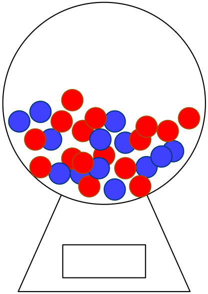
For example if we believe the mean height of a 12 year old child is 143cm, our sample has mean height of 148cm....is such a sample likely given the assumption that the mean height for the population (12 year olds) is 143?

- If this area is small - then it means the result we are seeing from our sample is "rare"
- if this area is large then it means the result we are seeing are likely

The process of hypothesis testing is the structure we use to answer this question.

Example Gumball

- We are told that the gumball machine has 50% red sweets.
- If we take 100 sweets and 42 of them are red?
- How likely is this result (0.42) if the proportion of sweets in the large gumball machine is 0.5?



Hypothesis Testing Steps

1. Formulate the Hypotheses: H_0 and H_1
2. Identify and compute the **test statistic** used to assess the evidence against the **Null Hypothesis** H_0
3. Compute the **p-value** which answers the question: If the Null Hypothesis is true, what is the probability of observing the test statistic at least as extreme as the one computed in (2)
4. Compare the **p-value** to the significance level α that is required (typically 0.05), if p-value $\leq \alpha$ then we rule against H_0

Hypothesis Testing for the Mean for a Normal population - (step1)

This is often referred to as **one-sample t-test**.

- Assuming a sample of observations $x_1, \ldots, x_n$ of a random variable from a *normally* distributed population $N(\mu, \sigma^2)$
- If we want to test a **null hypothesis** H_0 that the mean μ has a specific value μ_0 against an **alternative hypothesis** that $\mu \neq \mu_0$

Our hypotheses are:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Computing the test statistic (step 2)

- The **test statistic** in this case is based on the sample mean $\bar{x}$
- Intuitively if $\bar{x}$ is close to μ_0 then this is evidence in support of H_0
- If H_0 is true then

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

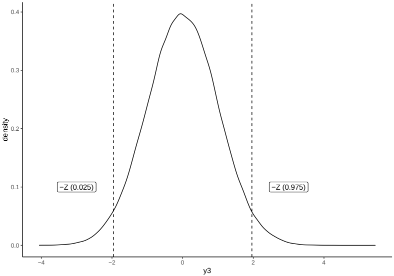
- We compute this for the sample by substituting for the sample mean and standard deviation to obtain a value for z_{obs}

p-value (step 3)

- The **p-value** associated to a test is the probability that the test statistics takes a value equal to, or more extreme than the observed value, assuming that H_0 is true.
- When testing a mean the p-value is given by:

$$P(|Z| \geq |z_{obs}||H_0 is TRUE)$$

The distribution of the test statistic under H_0 is below. The area under the curve outside the dashed lines is the p-value. (recall `pnorm`)



compare p-value (step 4)

- The p-value is the area under the curve for values more extreme that the one in our sample (when it is two sided we look both ways)
- If that p-value is greater than α the we conclude that our sample mean is likely enough not to reject the NULL hypothesis H_0
- If the p-value is smaller than α the we conclude that our sample mean is unlikely enough to accept H_1

example - 12 year old children heights

- Lets go back the example of measuring the heights of 12 year old children
- The population mean (in the UK for 12 year old) is 143 cm
- If we have a sample mean of 148 and a standard deviation of the sample of 4.9
- (step 1) the hypothesis: $H_0 : \mu = 143$, $H_1 : \mu \neq 143$
- (step 2) The test statistic is:

$$z = \frac{148 - 143}{4.9/\sqrt{30}} = 5.6$$

- (step 3) This is equivalent to a p-value of 0.0001032
- So in this case we reject the null hypothesis as it is very very unlikely that we would see such 30 observations if the mean was indeed 143 cm.

In R

```
```{r}
height<-rnorm(30, mean=147, sd=4.9)
t.test(height, mu=143,  alternative="two.sided")
```
One Sample t-test

data: height
t = 4.4942, df = 29, p-value = 0.0001032
alternative hypothesis: true mean is not equal to 143
95 percent confidence interval:
 145.4577 149.5627
sample estimates:
mean of x
 147.5102
```

Hypothesis Test for a proportion

- If we want to test a hypothesis related to a proportion (e.g. proportion of people voting for one political party)
- the test statistic to use is

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

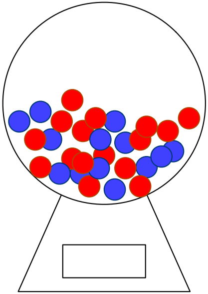
- where p is the sample proportion, π_0 is the H_0 proportion.

in R use `prop.test()` see slide 24.

Brunel University London

Example Gumball

- We are told that the gumball machine has 50% red sweets.
- If we take 100 sweets and 42 of them were red?
- How likely is it that the proportion of sweets in the large gumball machine is 0.5?



Hypothesis testing - Gumball

- (step 1) Formulate the hypothesis: $H_0 : \pi = 0.5, H_1 : \pi < 0.5$
- (step 2) Compute the test statistic using $p = 0.42$ from the sample,
- (step 3) Compute the p-value using `prop.test` in R we obtain p-value$= 0.07$
- (step 4) At the confidence level of 95% (equivalent to $\alpha = .05$) this is not significant.
- Intuitively we dont have enough evidence to reject the NULL hypothesis.

Hypothesis testing - Gumball - R

```
```{r}
prop.test(42,100, p=0.5, alternative = "less")
```
1-sample proportions test with continuity correction

data: 42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.06681
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.5072341
sample estimates:
   p
0.42
```

Recap - Hypothesis Testing

Recap on hypothesis testing and looked at the 4 step process:

1. Formulate all hypotheses
2. Calculate the test statistic
3. Translate it into a p-value
4. Compare the p-value to α and decide

Types of Error - OPTIONAL

- When we do hypothesis tests there are two type of error we can make
- We can reject H_0 when it is in fact true
- We can fail to reject H_0 when it is in fact false

	H_0 is true	H_0 is false
accept H_0	NO ERROR	Type II
reject H_0	Type I	NO ERROR

- $\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$
- $\beta = P(\text{type II error}) = P(\text{fail to reject} H_0 | H_0 \text{ is false})$

Brunel University London

Types of error - contd OPTIONAL

- Ideally we want both α and β to be small
- We can control α by selecting the value
- When it comes to β, $1-\beta$ is known as the **power** of the hypothesis test
- One way of controlling the **power** is to ensure the sample size is in line with the expected effect (we wont dwell on this, but see Crawley Page 9 if you want more details)

Hypothesis Testing

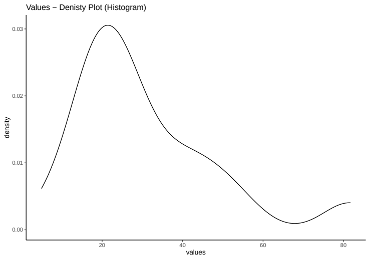
In the next part of the lecture we will cover hypothesis tests to answer the following questions:

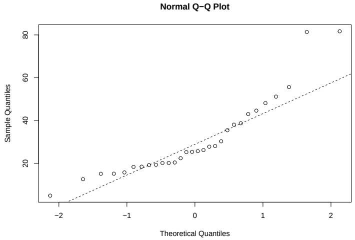
1. Is the data in one sample normally distributed?
2. Is the variance of two samples the same?
3. Do two samples have the same mean?
4. What can be done when the data is not normally distributed?
5. Testing a hypothesis related to a proportion?

1. Testing for normality

- The simplest test for **normality** is the **quantile-quantile** or **Q-Q plot**
- It plots the ranked samples from the sample available against a similar number of of ranked quantiles from the normal distribution
- This plot can be easily produced in R - (see page 79 in Crawley)

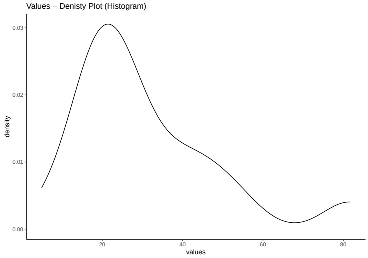
Looking at this data - is this normally distributed?

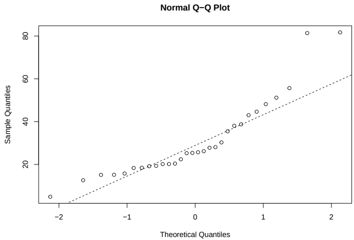




1. Testing for normality - example

Looking at this data - is this normally distributed?





This data does not look normally distributed

1. Testing for normality - numerical methods

It is not necessary to rely on visual inspection there are some tests that can be applied:

- **Kolmogorov-Smirnov (K-S)** test
- **Shapiro-Wilks** test
- The Shapiro-Wilks test is recommended as it provides better power
- These are hypothesis tests where $H_0 : x \ N(\mu, \sigma)$ and $H_1 : x \neq N(\mu, \sigma)$
- Therefore is the p-value obtained from the test (`shapiro.test`) is < 0.05 then we reject H_0 and assume the data is not normally distributed.
- Tor the sample data from the plots the p-value for the shapiro test is 0.0009 - so this data is not normally distributed!

Brunel University London

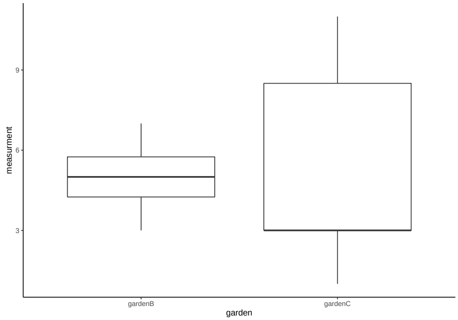
2. Comparing Variances

- Before comparing two samples means, we need to test whether the sample variances are different.
- To do so we use **Fisher's F test** which involves dividing the larger variance by the smaller one
- In order to determine if this is significant or not the **critical value** can be obtained from the **F-distribution**
- The degrees of freedom are $n-1$ for each sample
- `var.test(x,y)` will perform this test for you in R

2. Comparing Variances - example

We are going to use data on the Ozone levels in two Gardens B and C:

	gardenB	gardenC
1	5	3
2	5	3
3	6	2
4	7	1
5	4	10
6	4	4
7	3	3
8	5	11
9	6	3
10	5	10



2. (optional) Comparing Variances - In R

We want to test whether the variance in the Ozone levels in Garden B are different from Garden C? The H_0 hypothesis is that the variances are the same, H_1 that they are different

```
var.test(gardens$gardenB, gardens$gardenC)
F test to compare two variances

data:  gardens$gardenB and gardens$gardenC
F = 0.09375, num df = 9, denom df = 9, p—value = 0.001624
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02328617 0.37743695
sample estimates:
ratio of variances
           0.09375
```

The p-value is small 0,001 therefore in this case we reject the hypothesis that the variances are equal!

3. Two sample Hypothesis testing for means

- In some cases we may want to compare two samples, that may have been collected under different conditions (perhaps two treatments)
- We assume $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$
- We can test the following hypotheses:

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$$

or

$$H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 \neq 0$$

3. Testing sample means of two samples

There are two sample tests for comparing two sample means

- **Student's t-test** when samples are independent and the errors are normally distributed
  - `t.test(x,y)` will perform the test in R - when the variances are not equal R will use the **Welch** t test automatically.
  - a worked example is in Crawley 90 - 95
- (OPTIONAL) **Wilcoxon's Rank Sum test** when the samples are independent, but the errors are not normally distributed

3. Worked example

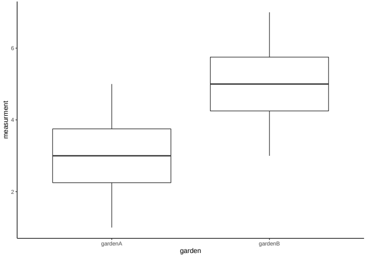
Data is available that measures the ozone levels for two gardens A and B (different from previous). We want to test whether their means are the same ($\alpha = 0.05$).

	gardenA	gardenB
1	3	5
2	4	5
3	4	6
4	3	7
5	2	4
6	3	4
7	1	3
8	3	5
9	5	6
10	2	5

The first step is to test the variances...

3. Worked example

	gardenA	gardenB
1	3	5
2	4	5
3	4	6
4	3	7
5	2	4
6	3	4
7	1	3
8	3	5
9	5	6
10	2	5



Do the variances look similar?

Firstly we should test for equal variances:

```
var.test(two.sample$gardenA, two.sample$gardenB)

F test to compare two variances

data:  two.sample$gardenA and two.sample$gardenB
F = 1, num df = 9, denom df = 9, p-value = 1
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2483859 4.0259942
sample estimates:
ratio of variances
                 1
```

The H_0 that variances are the same holds .....What about testing the means?

3. Worked example - continued

To test the hypothesis of equal means $H_0 : \mu_{GardenA} = \mu_{GardenB}$ vs $H_1 : \mu_{GardenA} \neq \mu_{GardenB}$

```
t.test(two.sample$gardenA, two.sample$gardenB)

Welch Two Sample t-test

data:  two.sample$gardenA and two.sample$gardenB
t = -3.873, df = 18, p-value = 0.001115
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0849115 -0.9150885
sample estimates:
mean of x mean of y
        3         5
```

In summary: equal variances but different means!

4. (optional) Wilcoxon Rank-Sum Test

- This is the **non parametric** alternative to the t-test
- It is used when we cannot assume a normal distribution
- The test works by putting all the measurements into one column (taking note which sample each measurement came from) then assigning a rank to each value
- The sum of the ranks in each group is then computed, if the samples are similar in location then the rank sums will be similar.

  For more see page 96 of Crawley.

Week 3 - Learning Outcomes

This week after the lecture, the lab and the independent practice you should be:

- able to express the NULL and alternative hypothesis for testing
- able to interpret the results of hypothesis testing
- familiar with the approaches to run hypothesis tests for different situations (mean, proportion, normality, two means and two variances)

In the lab you will practice using R to do hypothesis testing

This week's material is also covered in chapter 6 of Crawley