MAST7866 — Foundations of Data Science

Computing Session 8 — Hypothesis Testing 1

Within this worksheet we discuss the topic of hypothesis testing. In a test of a hypothesis, we test a certain given theory or belief about a population parameter. We may want to find out, using some sample information, whether or not a given claim (or statement) about a population parameter is true.

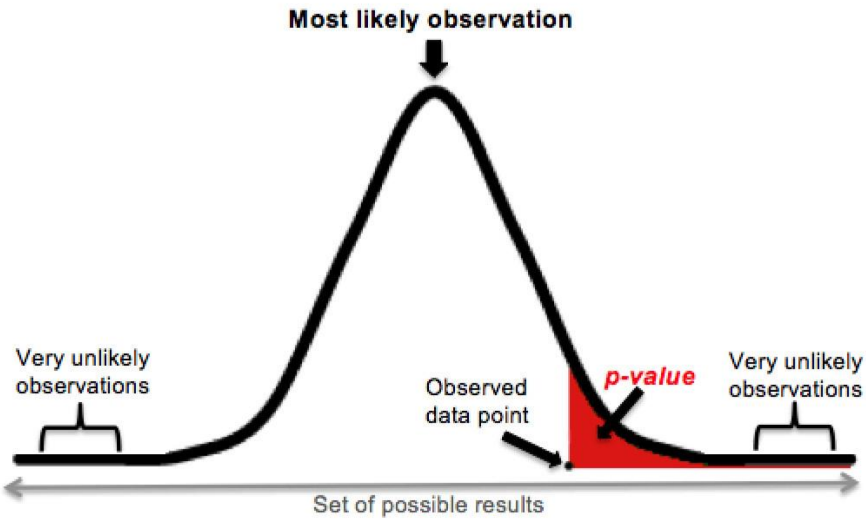
Task 1 – Introduction

Watch the video "Hypothesis Testing" on moodle.

Task 2 – Understanding p-values

When conducting a hypothesis test we can calculate a p-value, which is defined as the smallest level of significance at which the given null hypothesis is rejected. Using this p-value, we state the decision. If we have a predetermined value of α, then we compare the value of p with α and make a decision. We reject the null hypothesis if p-value $\leq \alpha$ and we do not reject the null hypothesis if p-value $> \alpha$.

For a one-tailed test the p-value is given by the shaded red area in the diagram below:



A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

For a two-tailed test, the p-value is twice the area in the tail of the sampling distribution.

This week's seminar focussed on p-values. If you were not at that session please **watch the recording of the session** on moodle.

We can never be certain about whether H_0 is true. The best we can say is that it is very unlikely that we would have observed the data that we did observe if H_0 is true.

Conventionally, we often interpret significance levels as follows:

Significant at	Interpretation
10%	No real evidence against H_0, but may be worth collecting more data
5%	Some evidence against H_0
1%	Strong evidence against H_0
0.1%	Very strong evidence against H_0

Task 3 – Conducting a Hypothesis Test for population mean

Example 1

Suppose we are interested in the mean starting salary (in £'000s) for the students of the School of Mathematics, Statistics and Actuarial Science of the University of Kent.

35	23	34	56	29	25	28	45	29	34	32	31
30	35	45	27	29	40	39	32	29	23	24	29
29	30	28	31	25	26	28	25	26	24	21	

Two experts on labour economics claim the following:

- Economist 1: The mean starting salary is $\mu = 29$.
- Economist 2: The mean starting salary is $\mu = 45$.

Assess the claims of the two economists.

Let us think about the claim of Economist 1 and test

H_0: $\mu = 29$

H_1: $\mu \neq 29$

In order to do this we need to calculate a two-sided t-test.

```
> t.test(data_salaries,mu=29)
```

The output is:

```
        One Sample t-test

data:  data1
t = 1.4312, df = 34, p-value = 0.1615
alternative hypothesis: true mean is not equal to 29
95 percent confidence interval:
 28.26808 33.21763
sample estimates:
mean of x
 30.74286
```

You will see that the output tells us what the alternative hypothesis is that we have tested – in this case it is the required two-sided alternative that the true mean is not equal to 29. You will note that this is the same R command (except for the mu=29 argument) that we used to construct 95% confidence interval for a population mean.

The output tells us that the p-value is 0.1615, since this is greater than our significance level of 0.05, this indicates that we have insufficient evidence to reject the null hypothesis, therefore the mean starting salary of graduates could be £29,000.

Challenge 1

Construct an appropriate hypothesis test for the 2nd economists claim. What do you conclude?

Note that if we want to test a one-sided hypothesis test this can be done by including the argument: alternative = “less” or alternative = “greater” in the t.test function.

Task 4 – Conducting a Hypothesis Test for population proportion

For a population proportion π, we wish to test the hypotheses:

H_0: $\pi = \pi_0$

H_1: $\pi \neq \pi_0$

Example 2

Suppose we want to know whether the proportion of the population who believe that the government's policies will boost the economy has changed over the last 6 months. A previous survey has shown that $\pi = 0.52$ of the people agree with the government's policies. A new survey takes place now which gives us a random sample of size n = 1000. The new survey shows that 567 people are positive that the economy will grow following the proposed economic policies. Test at the 1% significance level whether the current proportion of people who agree with the government's policies is different from 0.52.

The hypothesis test here is:

H_0: $\pi = 0.52$

H_1: $\pi \neq 0.52$

```
> prop.test(x=567, n=1000,p=0.52)

     1-sample proportions test with continuity
     correction

data:  567 out of 1000, null probability 0.52
X-squared = 8.6629, df = 1, p-value = 0.003248
alternative hypothesis: true p is not equal to 0.52
95 percent confidence interval:
 0.5355888 0.5978900
sample estimates:
    p
0.567
```

The p-value in this case is 0.003, thus we can conclude, because this is less than the chosen significance level of 0.01 (1%) that there is strong evidence that the proportion of people who are positive the economy will grow is different from 0.52. As the point estimate is greater than 0.52, we can conclude that the proportion is greater than 0.52.

Challenge 2

Suppose that a survey taken of a random sample of 40 students gives that the sample proportion is 0.55. The same survey is repeated and a random sample of 400 students gives the same values for p. Historically we know that the student proportion is 0.42. For each survey test whether there has been a change in the student proportion. Compare the results you got from the two tests

Task 5 – Understanding the link between confidence intervals and hypothesis testing

Watch the video Confidence Intervals and Hypothesis Testing on moodle.

Challenge 3

Have a look at the output from the R code you have run for the above challenges and compare the confidence intervals and conclusions from the hypothesis tests you have run.