Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

5.3 P-values, effect size and sample size influences

To p, or not to p? Answer: p! [1]

We introduce p-values, which are our principal tool for deciding whether or not to reject H_0.

A **p-value** is the probability of the event that the 'test statistic' takes the observed value or more extreme (i.e. more unlikely) values under H_0. It is a measure of the discrepancy between the hypothesis H_0 and the data evidence.

- A 'small' p-value indicates that H_0 is not supported by the data.
- A 'large' p-value indicates that H_0 is not inconsistent with the data.

So p-values may be seen as a **risk measure** of rejecting H_0.

Example

Suppose one is interested in evaluating the mean income (in £000s) of a community. Suppose income in the population is modelled as $N(\mu, 25)$[2] and a random sample of $n = 25$ observations is taken, yielding the sample mean $\bar{x} = 17$.

Independently of the data, three expert economists give their own opinions as follows.

- Dr A claims the mean income is $\mu = 16$.
- Ms B claims the mean income is $\mu = 15$.
- Mr C claims the mean income is $\mu = 14$.

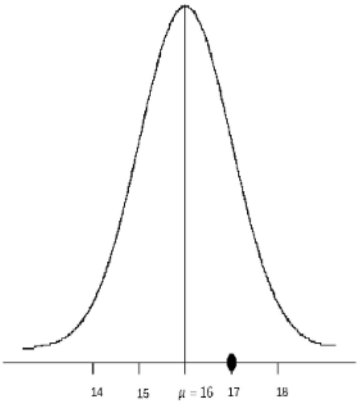
How would you assess these experts' statements?

[1] Available at `http://etheses.lse.ac.uk/31` 😉

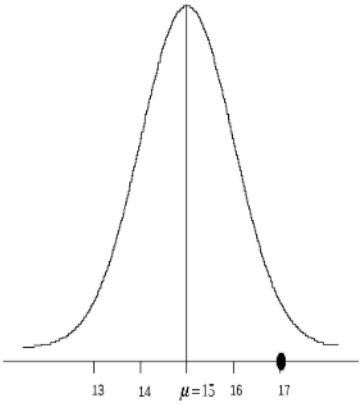
[2] Income is not normally distributed, but heavily positively skewed. We ignore this fact here for simplicity.

Here, $\bar{X} \sim N(\mu, \sigma^2/n) = N(\mu, 1)$. We assess the statements based on this distribution.

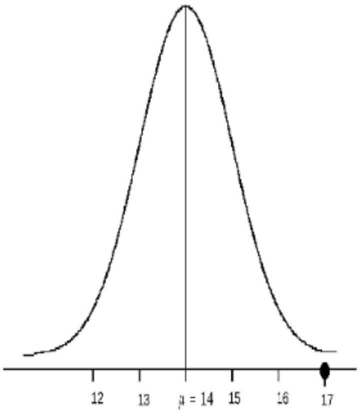
If Dr A's claim is correct, $\bar{X} \sim N(16, 1)$. The observed value $\bar{x} = 17$ is **one standard deviation away from μ**, and may be regarded as a typical observation from the distribution. Hence there is **little inconsistency** between the claim and the data evidence. This is shown below:



If Ms B's claim is correct, $\bar{X} \sim N(15, 1)$. The observed value $\bar{x} = 17$ begins to look a bit 'extreme', as it is **two standard deviations away from μ**. Hence there is **some inconsistency** between the claim and the data evidence. This is shown below:



If Mr C's claim is correct, $\bar{X} \sim N(14, 1)$. The observed value $\bar{x} = 17$ is very extreme, as it is **three standard deviations away from μ**. Hence there is **strong inconsistency** between the claim and the data evidence. This is shown below:



It follows that:

- under $\mathrm{H}_0 : \mu = 16$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 15) = P(|\bar{X} - 16| \geq 1) = 0.3173$
- under $\mathrm{H}_0 : \mu = 15$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 13) = P(|\bar{X} - 15| \geq 2) = 0.0455$
- under $\mathrm{H}_0 : \mu = 14$, $P(\bar{X} \geq 17) + P(\bar{X} \leq 11) = P(|\bar{X} - 14| \geq 3) = 0.0027.$

In summary, we **reject** the hypothesis $\mu = 15$ or $\mu = 14$, as, for example, if the hypothesis $\mu = 14$ is true, the probability of observing $\bar{x} = 17$, or more extreme values, would be as small as 0.003. We are comfortable with this decision, as **a small probability event would be very unlikely to occur in a single experiment**.

On the other hand, we cannot reject the hypothesis $\mu = 16$. However, this does not imply that this hypothesis is necessarily true, as, for example, $\mu = 17$ or 18 are at least as likely as $\mu = 16$. Remember:

not reject $\neq$ accept.

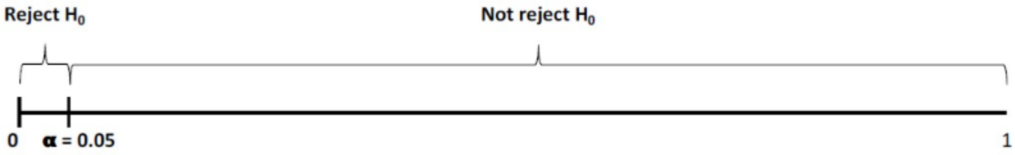
A statistical test is incapable of 'accepting' a hypothesis.

Interpretation of *p*-values

In practice the statistical analysis of data is performed by computers using statistical or econometric software packages. Regardless of the specific hypothesis being tested, the execution of a hypothesis test by a computer returns a p-value. Fortunately, there is a universal decision rule for p-values.

Section 5.2 explained that we control for the probability of a Type I error through our choice of significance level, α, where $\alpha \in [0,\, 1]$. Since p-values are also probabilities, as defined above, we simply compare p-values with our chosen benchmark significance level, α.

The **p-value decision rule** is shown below for $\alpha = 0.05$:



Our decision is to **reject H_0 if the p-value is $\leq \alpha$**. Otherwise, H_0 is not rejected.

Clearly, the magnitude of the p-value (compared with α) determines whether or not H_0 is rejected. Therefore, it is important to consider **two key influences** on the magnitude of the p-value: the effect size and the sample size.

Effect size influence

The **effect size** reflects the difference between what you would *expect* to observe if the null hypothesis is true and what is *actually* observed in a random experiment. Equality between our expectation and observation would equate to a zero effect size, which (while not proof that H_0 is true) provides the most convincing evidence in favour of H_0. As the difference between our expectation and observation increases, the data evidence becomes increasingly inconsistent with H_0 making us more likely to reject H_0. Hence **as the effect size gets larger, the *p*-value gets smaller** (and so is more likely to be below α).

To illustrate this idea, consider the experiment of tossing a coin 100 times and observing the number of heads. Quite rightly, you would not doubt the coin is fair (i.e. unbiased) if you observed *exactly* 50 heads as this is what you would expect from a fair coin (50% of tosses would be expected to be heads, and the other 50% tails). However, it is possible that you are:

- somewhat sceptical that the coin is fair if you observe 40 or 60 heads, say
- even more sceptical that the coin is fair if you observe 35 or 65 heads, say
- highly sceptical that the coin is fair if you observe 30 or 70 heads, say.

In this situation, the greater the difference between the number of heads and tails, the more evidence you have that the coin is not fair.

In fact, if we test:

$$\mathrm{H}_0 : \pi = 0.5 \quad \text{vs.} \quad \mathrm{H}_1 : \pi \neq 0.5$$

where $\pi = P(\text{heads})$, for $n = 100$ tosses of the coin we would expect 50 heads and 50 tails. It can be shown that *for this fixed sample size* the *p*-value is sensitive to the effect size (the difference between the observed sample proportion of heads and the expected proportion of 0.5) as follows:

Observation	Expectation	Effect size	*p*-value	Decision if $\alpha = 0.05$
50 heads & 50 tails	50 heads & 50 tails	$0.5 - 0.5 = 0$	1	H_0 is not rejected
55 heads & 45 tails	50 heads & 50 tails	$0.55 - 0.5 = 0.05$	0.3682	H_0 is not rejected
60 heads & 40 tails	50 heads & 50 tails	$0.6 - 0.5 = 0.1$	0.0569	H_0 is not rejected
70 heads & 30 tails	50 heads & 50 tails	$0.7 - 0.5 = 0.2$	< 0.0001	H_0 is rejected
80 heads & 20 tails	50 heads & 50 tails	$0.8 - 0.5 = 0.3$	≈ 0.0000	H_0 is rejected

So we clearly see the **inverse relationship between the effect size and the *p*-value**.

The above is an example of a **sensitivity analysis** where we consider the pure influence of the effect size on the *p*-value while controlling for (fixing) the sample size. We now proceed to control the effect size to examine the sample size influence.

Sample size influence

Other things equal, **a larger sample size should lead to a more representative random sample** and the characteristics of the sample should more closely resemble those of the population distribution from which the sample is drawn.

In the context of the coin toss, this would mean the observed sample proportion of heads should converge to the true probability of heads, π, as $n \to \infty$.

As such, we consider the **sample size influence** on the p-value. For a non-zero effect size[3] **the p-value decreases as the sample size increases**.

Continuing the coin toss example, let us fix the (absolute) effect size at 0.1, i.e. in each of the following examples the observed sample proportion of heads differs by a fixed proportion of 0.1 (= 10%).

Observation	Expectation	Sample size	p-value	Decision if $\alpha = 0.05$
6 heads & 4 tails	5 heads & 5 tails	$n = 10$	0.7539	H_0 is not rejected
12 heads & 8 tails	10 heads & 10 tails	$n = 20$	0.5034	H_0 is not rejected
18 heads & 12 tails	15 heads & 15 tails	$n = 30$	0.3616	H_0 is not rejected
60 heads & 40 tails	50 heads & 50 tails	$n = 100$	0.0569	H_0 is not rejected
150 heads & 100 tails	125 heads & 125 tails	$n = 250$	0.0019	H_0 is rejected

So we clearly see the **inverse relationship between the sample size and the p-value**.

In Section 5.2, we defined the power of the test as the probability that the test will reject a false null hypothesis. In order to reject the null hypothesis it is necessary to have a sufficiently small p-value (less than α), hence we see that we can unilaterally increase the power of a test by increasing the sample size. Of course, the trade-off would be the increase in data collection costs!

Example

Let $\{X_1, \ldots, X_{20}\}$, taking values either 1 or 0, be the outcomes of an experiment of tossing a coin 20 times, where:

$$P(X_i = 1) = \pi = 1 - P(X_i = 0) \quad \text{for } \pi \in (0,\, 1).$$

We are interested in testing:

$$\text{H}_0 : \pi = 0.5 \quad \text{vs.} \quad \text{H}_1 : \pi \neq 0.5.$$

Suppose there are 17 X_is taking the value 1, and 3 X_is taking the value 0. Will you reject the null hypothesis at the 1% significance level?

[3] A zero effect size would result in non-rejection of H_0, regardless of n.

Let $T = X_1 + \cdots + X_{20}$. Therefore, $T \sim \mathrm{Bin}(20, \pi)$. We use T as the test statistic. With the given sample, we observe $t = 17$. What are the more extreme values of T if H_0 is true?

Under H_0, $\mathrm{E}(T) = n\,\pi = 20 \times 0.5 = 10$. Hence 3 is as extreme as 17, and the more extreme values are:

$$0, \quad 1, \quad 2, \quad 18, \quad 19 \quad \text{and} \quad 20.$$

Therefore, the p-value is:

$$\begin{aligned}
\left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) P_{\mathrm{H}_0}(T = i) &= \left(\sum_{i=0}^{3} + \sum_{i=17}^{20}\right) \frac{20!}{(20-i)!\,i!}\,(0.5)^i\,(1-0.5)^{20-i} \\
&= 2 \times (0.5)^{20} \sum_{i=0}^{3} \frac{20!}{(20-i)!\,i!} \\
&= 2 \times (0.5)^{20} \times \left(1 + 20 + 20 \times \frac{19}{2!} + \frac{20 \times 19 \times 18}{3!}\right) \\
&= 0.0026.
\end{aligned}$$

So we reject the null hypothesis of a fair coin at the 1% significance level, since $0.0026 < 0.01$.