Confidence intervals for a difference of two proportions

Suppose we have two independent simple random samples from two populations. We have m samples $(X_1, \ldots, X_m)$ from the first population which has a Bernoulli(p_X) distribution and we have n samples $(Y_1, \ldots, Y_n)$ from the second population which has a Bernoulli(p_Y) distribution. How can we construct a confidence interval for $p_X - p_Y$, the difference of two proportions?

$\widehat{P}_X = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \widehat{P}_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i$

We know that $\widehat{P}_X$ and $\widehat{P}_Y$ are independent with

$$\widehat{P}_X \sim_{approx} \text{Normal}\left(p_X, \frac{p_X(1-p_X)}{m}\right) \quad \text{and} \quad \widehat{P}_Y \sim_{approx} \text{Normal}\left(p_Y, \frac{p_Y(1-p_Y)}{n}\right)$$

so

$\mathbb{E}[\hat{p}_X - \hat{p}_Y] = \mathbb{E}[\hat{p}_X] - \mathbb{E}[\hat{p}_Y] = p_X - p_Y$

$$\widehat{P}_X - \widehat{P}_Y \sim_{approx} \text{Normal}\left(p_X - p_Y, \frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}\right)$$

This leads to the $1-\alpha$ stochastic confidence interval

$$\underbrace{\widehat{P}_X - \widehat{P}_Y}_{\text{estimator}} \pm \underbrace{z_{1-\alpha/2}}_{\text{critical value}} \underbrace{\sqrt{\frac{\widehat{P}_X(1-\widehat{P}_X)}{m} + \frac{\widehat{P}_Y(1-\widehat{P}_Y)}{n}}}_{\text{standard error of estimator}}$$

Exercise: A 2014 study aimed to assess the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage. A total of 2,729 women who had positive pregnancy tests at clinics were identified for participation but only 1,061 were ultimately interviewed. description study

Of the 208 women who were aged 36 years or above (36+), 52 had a miscarriage, while for the remaining 853 women who were under 36 (< 36), 120 had a miscarriage. Question

Give a 95% confidence interval for the true difference in the rates of miscarriage between women 36+ and women < 36. What does the interval say about the relationship between age and the rate of miscarriage?

p_1 = proportion of miscarriage in 36+ age group.
p_2 = " " " in <36 age group

$\hat{p}_1 = \frac{52}{208} = 0.25 \qquad \hat{p}_2 = \frac{120}{853} \simeq 0.1407. \qquad \hat{p}_1 - \hat{p}_2 = 0.1093$

$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.25 \times 0.75}{208} + \frac{0.1407 \times 0.8593}{853}}$

$= 0.0323$

$0.95 = 1-\alpha \Rightarrow \alpha = 0.05 \qquad z_{0.975} = 1.96$

Our 95% CI for $p_1 - p_2$ is $0.1093 \pm 1.96 \times 0.0323$

0.1093 ± 0.0633

A total of 80 city council buses were randomly selected. The arrival times of each bus at their last stop were compared to the published bus timetable to determine if they were late. Sixteen buses were observed to be late. Construct a 90% confidence interval for the true proportion for council buses that arrived late at their last stop.

let p be the proportion of buses that are late.

general form of CI estimate $\pm$ (critical value) x s.e. (estimate)

estimate $\hat{p} = \frac{16}{80} = 0.2$ $\quad$ s.e. $(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.2 \times 0.8}{80}} = 0.04472\ldots$

$0.90 = 1 - \alpha \Rightarrow \alpha = 0.1$ $\quad$ $z_{0.95} = 1.6449$

$$0.2 \pm 1.6449 \times 0.04472 \Rightarrow 0.2 \pm 0.0736$$

we are 90% confident the true proportion of late buses is in $(0.126, 0.274)$

Having received many complaints from upset passengers that buses were always late, the council decides to implement an electronic ticketing system. Three months after introducing the electronic ticketing system, another random sample of 80 city council buses was selected and 10 were observed to be late at their last stop. Construct a 99% confidence interval for the change in the proportion of buses that arrive late at their last stop.

7

Hypothesis Testing

By the end of this chapter you should:

- Know how to specify null and alternative hypotheses.
- Be able to apply basic statistical tests.
- Know how to interpret a test statistic and a p-value.
- Be able to understand the types of errors that occur in hypothesis testing.
- Know what factors controls the probability of these errors in hypothesis testing.

In the previous chapter we saw how to estimate basic quantities such as a mean or proportion and how to quantify our uncertainty about those estimates. Another problem that arises in the analysis of data is how to make a decision about our model. This arises naturally in a number of settings:

- Do video games increase aggressive behaviour in children?
- Does the new website design get more hits than the old?
-

not dealt with in STAT2203.

- Do difficult subjects cause students to sleep less?

Null and Alternative Hypotheses

In statistics, this problem is called a *hypothesis test*. In hypothesis testing, given data, we wish to determine which of two competing hypotheses: the **null hypothesis** (H_0) and the **alternative hypothesis** (H_1).

We begin with a model for the process generating our data. For example, suppose our data is a realisation of a simple random sample (that is, a realisation of a collection of independent random variables all having the same distribution) from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. In general, we will denote the parameter(s) of the model by θ and the set of all possible parameter values by Θ. We can now specify the null and alternative hypotheses in terms of the parameter θ. capital theta

Theta

Let Θ_0 and Θ_1 form a partition of the parameter space Θ. That is, $\boxed{\Theta_0 \cup \Theta_1 = \Theta}$ and $\boxed{\Theta_0 \cap \Theta_1 = \emptyset}$. The null and alternative hypotheses are then specified as

$$H_0 : \theta \in \Theta_0, \qquad\qquad H_1 : \theta \in \Theta_1.$$

Example: The currently accepted value for the mean density of the Earth is $5.517g/cm^3$. In 1798 Henry Cavendish presented some observations for the mean density of the Earth. Suppose Cavendish's apparatus produced measurements from a $\mathsf{Normal}(\mu, \sigma^2)$ distribution. Potential hypotheses to test would be

- Test $H_0 : \mu = 5.517g/cm^3$ versus $H_1 : \mu \neq 5.517g/cm^3$ (two-sided alternative)

  H_0: measurements from the apparatus are unbiased.

  H_1: measurements from the apparatus are biased.
- Test $H_0 : \mu = 5.517g/cm^3$ versus $H_1 : \mu > 5.517g/cm^3$ (one-sided alternative)

  H_1: measurements from the apparatus overestimate the density of Earth.
- Test $H_0 : \mu = 5.517g/cm^3$ versus $H_1 : \mu < 5.517g/cm^3$ (one-sided alternative)

  H_1: measurements from apparatus under estimates the density of earth.

These two hypotheses are not treated symmetrically. The null hypothesis H_0 is taken as a statement of the "status quo" and we examine the data looking for evidence against H_0.

- If no evidence against H_0 is found, then we accept H_0.
- On the other hand, if evidence against H_0 is found (in the direction of H_1), then we will reject H_0 in favour of the alternative hypothesis H_1.

Test statistics and p-values

So before we can decided whether or not to accept the null hypothesis, we need to be able to quantify the evidence against the null hypothesis. We do this using be constructing a **test statistic** and a **p-value**.

> A test statistic is a function of the data whose distribution under the null hypothesis is known.

Example: Suppose $X_1, \ldots, X_n$ be a simple random sample from $\mathsf{Normal}(\mu, \sigma^2)$ with $\bar{X}$ and S^2 be the usual estimators of μ and σ^2 constructed from the $X_1, \ldots, X_n$. Under the null hypothesis $H_0 : \mu = 5.517 g/cm^3$, the test statistic

$$T(\mathbf{X}) = \frac{\bar{X} - 5.517}{S/\sqrt{n}}$$

has a t_{n-1}-distribution.

When our test statistic computed from the sample data $T(\mathbf{x})$ is 'large' in an appropriate sense, this will indicate evidence against the null hypothesis. This evidence against the null hypothesis is summarised more clearly through the use of a p-value.

p-value : probability of observing data "more extreme" than what we observed if the null hypothesis is true.

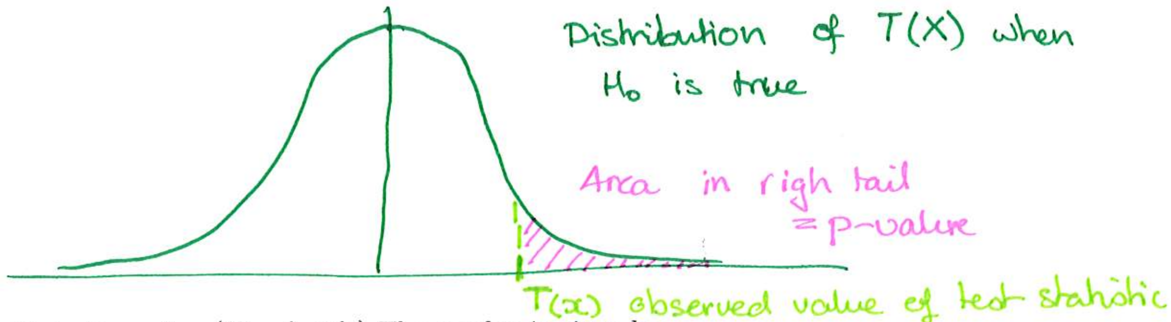
- One sided alternative ($H_1 : \theta > \theta_0$) The p-value is given by

$H_0 : \theta = \theta_0$

$$\mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})),$$

computed from data

where the probability is evaluated under the null hypothesis.

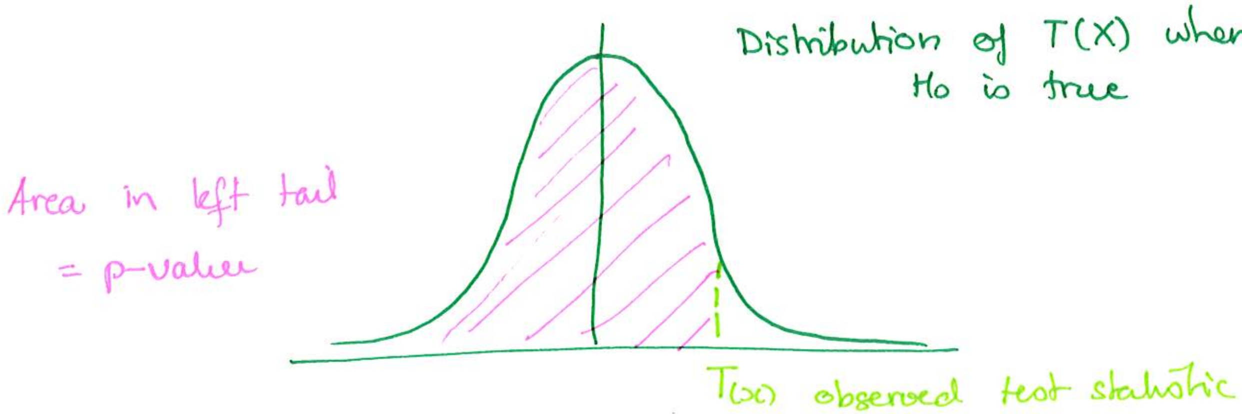


- One sided alternative ($H_1 : \theta < \theta_0$) The p-value is given by

$H_0 : \theta = \theta_0$

$$\mathbb{P}(T(\mathbf{X}) < T(\mathbf{x})),$$

where the probability is evaluated under the null hypothesis.

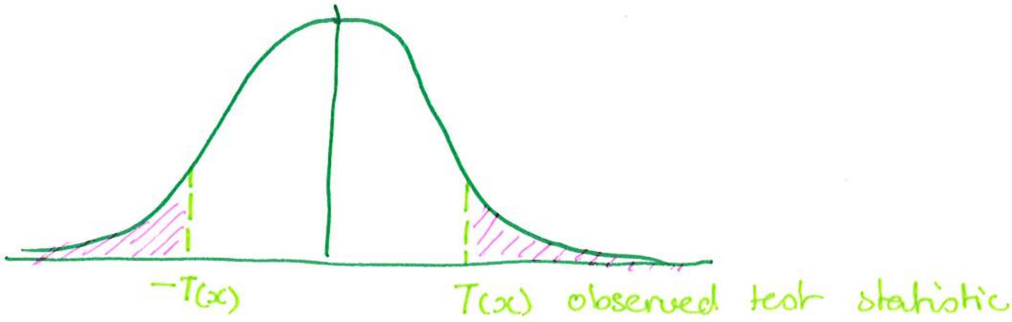


- Two sided alternative ($H_1 : \theta \neq \theta_0$) The p-value is given by

$H_0 : \theta = \theta_0$

$$2 \min \left[\mathbb{P}(T(\mathbf{X}) > T(\mathbf{x})), \mathbb{P}(T(\mathbf{X}) < T(\mathbf{x})) \right],$$

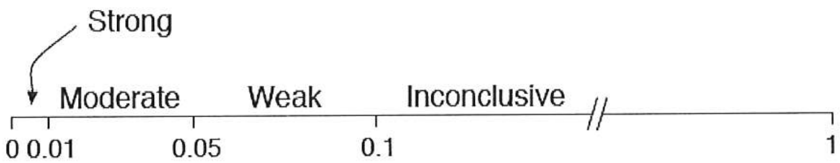
where the probability is evaluated under the null hypothesis.



Like the test statistic, the p-value is a function data and so it also has a distribution. Under the null hypothesis

$$p - \text{value} \sim \mathsf{Uniform}(0, 1).$$

The strength of evidence against the null hypothesis provided by the p-value is summarised in the figure below.



We must decide how small the p-value must be before we reject the null hypothesis. This cut-off point is called the **significance level** and is often denoted by α. The significance level determines the probability that we reject the null hypothesis when it is in fact true.

Question: Suppose you were to toss a coin that you believed was fair several times. How many consecutive heads would need to appear before you begin to doubt that it is really a fair coin?

It is common to use significance levels of 5% or 1%, though sometimes much smaller significance levels are needed.

Example: Assume that the measurements from Cavendish's apparatus are a realisation of a simple random sample from $\mathsf{Normal}(\mu, \sigma^2)$. We wish to test whether or not Cavendish's apparatus gave unbiased measurements of the density of the earth, that is we are testing

$$H_0 : \mu = 5.517 g/cm^3 \quad \text{against} \quad H_1 : \mu \neq 5.517 g/cm^3.$$

Cavendish made 23 measurements of the earth's density, with $\bar{x} = 5.4835 g/cm^3$ and $s = 0.1904 g/cm^3$. The test statistic is

$$T(\mathbf{x}) = \frac{\bar{x} - 5.517}{s/\sqrt{n}} = \frac{5.4835 - 5.517}{0.1904/\sqrt{23}} = -0.8438$$