Contents

1 Hypothesis Testing and Power **3**
- 1.1 Formulating a hypothesis test . . . 4
  - 1.1.1 Components of Hypothesis test . . . 4
  - 1.1.2 Examples . . . 6
- 1.2 Interpreting a hypothesis test . . . 7
  - 1.2.1 p-values . . . 7
  - 1.2.2 Errors . . . 7
  - 1.2.3 Inappropriate Hypothesis tests . . . 8
- 1.3 Common types of hypothesis test . . . 8
  - 1.3.1 One-sample t-test . . . 8
  - 1.3.2 Two-sample t-test . . . 9
  - 1.3.3 Comparing proportions . . . 9
- 1.4 Power calculations . . . 10
  - 1.4.1 How power calculations work . . . 10
  - 1.4.2 Power calculations in stata . . . 10
- 1.5 Hypothesis tests and confidence intervals . . . 10

2 Hypothesis Testing and Power Practical **11**
- 2.1 Inference about a proportion . . . 12
- 2.2 More inference about a proportion . . . 12
- 2.3 Inference about a mean . . . 13
- 2.4 Two-sample t-test . . . 14
- 2.5 One sample t-test . . . 14
- 2.6 Power Calculations . . . 15

1 Hypothesis Testing and Power

1.1 Formulating a hypothesis test

The two main forms of inference in frequentist statistics are confidence intervals and hypothesis tests. As we saw in the previous chapter, confidence intervals give a range in which the true population value is likely to lie. A hypothesis test can be thought of as doing the opposite: it presents the strength of evidence the the true population value is not one particular value.

1.1.1 Components of Hypothesis test

To perform a hypothesis test, we begin by creating the *Null Hypothesis*. This proposes a possible value for the true population value. We then calculate a *test statistic* from our sample data, can calculate the probability of seeing that value or one more extreme (further from the the null hypothesis value) if the null hypothesis is true. This probability is called the p-value.

As originally devised by Fisher, the p-value was intended to provide an informal assessment of the strength of the evidence against a particular null hypothesis. However, some researchers were not happy with this. They wanted a decision-making tool: for example, a p-value may say "there is good evidence that this drug is effective", but is the evidence good enough to justify introducing the drug, or not. Neyman and Pearson extended the idea of hypothesis testing by introducing the notion of rejecting the null hypothesis if the p-value is sufficiently small. This introduced the concept of statistical significance: if the p-value is less than 0.05, we say that the test is significant at the 5% level.

Null Hypothesis

The null hypothesis is the hypothesis that we wish to test, and is generally that there is no association between two variables. For example, it may be that the prevalence of a disease is the same in two different groups, or that there is no difference in outcome between the two arms of a drug trial.

The alternative hypothesis is simply "The null hypothesis is untrue", and hence can cover a wide range of possibilities. For example, if the null hypothesis were "there is no difference in prevalence between these two groups", the alternative hypothesis would be "there is a difference in prevalence between these two groups"

Sometimes, people will present a one-sided alternative hypothesis. For example, a drug company way define their null hypothesis as "our drug is no better than our competitor's" and their alternative hypothesis is as "our drug is better than our competitor's". It is rarely justified to use a one-sided test: we would be interested in knowing if their drug was worse than the competitor's, even if they would not. Generally, a one-sided test will have a smaller p-value than a two-sided test, which is often the motivation for doing it, so seeing the words "one-sided test" should ring alarm bells.

However, there are times when a one-sided test is entirely appropriate. For example, we will meet the χ^2-test for testing for an association between two categorical variables in Chapter **??**. This test measures the total difference between the observed numbers in each cell of a table, and the expected numbers in each cell. If the observed numbers are unusually close to the expected numbers if the null hypothesis of no association is true, that does not provide evidence against the null hypothesis. Only if the observed numbers are unusually far from the expected values is there evidence against the null hypothesis, so a one-sided test is entirely appropriate there.

Test statistics

In order to be able to test the null hypothesis, we need to have a statistic whose sampling distribution if the null hypothesis were true is known. For example, it may be a difference in prevalence between two groups, in which case the population value under the null hypothesis would be 0. This value would therefore be the expected value of the sampling distribution.

The T Distribution Suppose that the sampling distribution of our test statistic S has a mean μ and standard deviation σ if the null hypothesis is true. Then the statistic $T = \frac{S-\mu}{\sigma}$ will have a mean of 0 and a standard deviation of 1. This is known as the standard normal distribution. We very often work with "standardised" test statistics.

Figure 1.1 shows a standard normal distribution. The shaded area consists of the lowest 2.5% of the distribution and the highest 2.5%. There is a 5% chance that the test statistic will lie in the shaded area if the null hypothesis is true. This is considered sufficiently unlikely that it provides evidence against the null hypothesis.

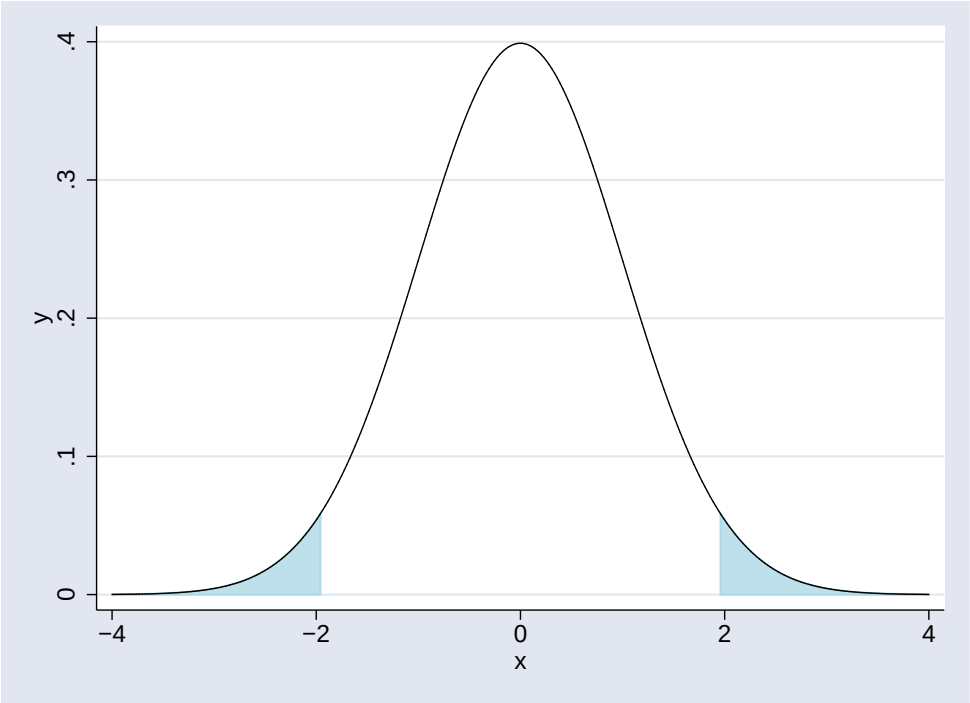

Figure 1.1: A Standard Normal Distribution

However, we again run into the problem that we don't know σ, and are working with a sample-based estimate of it s. So the statistic that we calculate is $T = \frac{S-\mu}{s}$, and this does not follow a standard normal distribution: the distribution it follows is called a t-distribution on $n-1$ degrees of freedom (where n is the sample size). This is where the expression "t-test" comes from.

For small values of n, the tails of the t-distribution are larger than the tails of the normal distribution, so extreme values of the statistic are less uncommon (see Figure 1.2. However, as the sample size increases, the t-distribution gets closer and closer to the normal distribution, and they are practically indistinguishable once the sample size increases beyond 100.

Some test statistics that we may be interested in do not follow a normal distribution, such

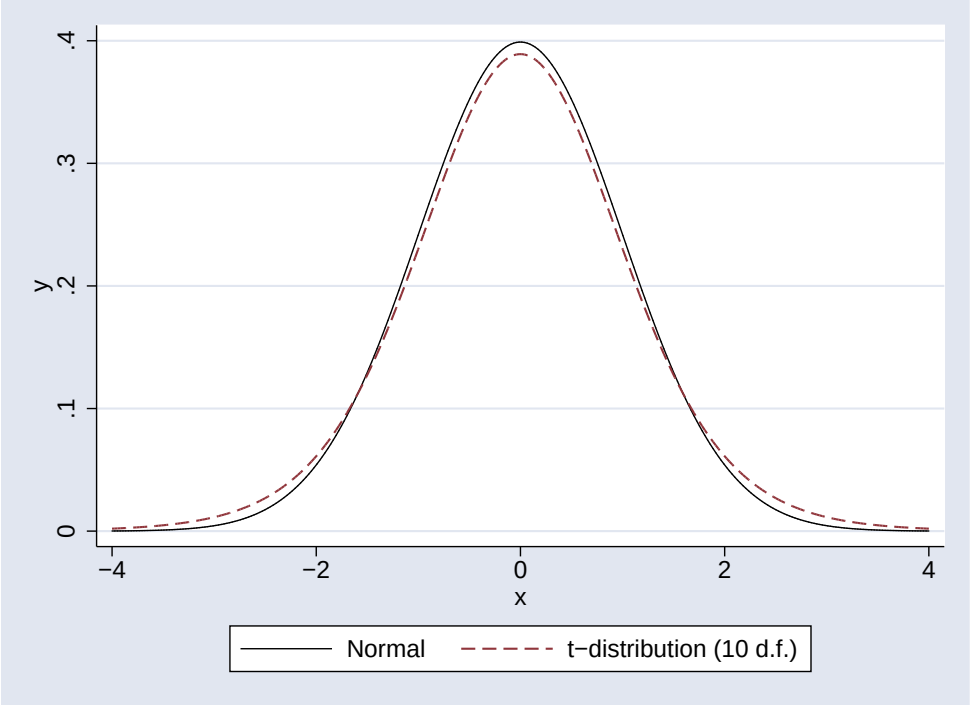

Figure 1.2: T-Distribution compared to a Normal Distribution

as the χ^2 statistic or the Mann-Whitney U statistic. However, as long as the centiles for the distribution of these statistics are known, we can still calculate a p-value based on that statistic.

1.1.2 Examples

Here are a couple of examples of the various components of a hypothesis test

Does height differ between men and women ?

Null hypothesis On average, men and women are the same height

Alternative Hypothesis One gender tends to be taller than the other.

Test Statistic Difference in mean height between men and women.

One-Sided Hypotheses

- Men are taller than women
- Women are taller than men

Which is more popular, Coke or Pepsi ?

Null hypothesis Equal numbers of people prefer Coke and Pepsi

Alternative Hypothesis Most people prefer one drink to the other

Test Statistic Several possibilities:

- Difference in proportions preferring each drink
- Ratio of proportions preferring each drink

One-Sided Hypotheses

- More people prefer Coke
- More people prefer Pepsi

1.2 Interpreting a hypothesis test

Interpreting a hypothsesis test is not easy. The American Statistical Association did a survey of the 25 most common errors in statistics, and 20 of them involve hypothesis testing. In particular, it is not easy to grasp what the p-value means (and more importantly, what it does not mean).

1.2.1 p-values

The p-value that is produced by a hypothesis test is the probability of seeing a test statistic at least as far from the null value as was seen in our sample if the null hypothesis is true. If the p-value is small, it was unlikely that our sample data would have been produced if the null hypothesis were true. But our data *was* produced, and so it provided evidence that the null hypothesis is not true.

The p-value must lie between 0 and 1. The smaller it is, the less likely it is that the data could have been produced if the null hypothesis were true. Conventionally, we say that if the p-value is less than 0.05, there is some evidence against the null hypothesis: the effect is "statistically significant at the 5% level". However, 0.05 is a totally arbitrary value: if the p-value is 0.055, the evidence against the null hypothesis is very nearly as strong as if the p-value were 0.045. For this reason, always present *exact* p-values (to 1 or 2 significant figures).

Large p-values cannot be thought of evidence that the null hypothesis is true. This is because the p-value is affected by both the size of the effect and the size of the study. A large p-value can happen if the null hypothesis it true, or if the study is too small to detect the actual difference between the truth and the null hypothesis.

It may help to think of a hypothesis test as a politician. You ask it a simple question, "Is the null hypothesis true ?", which only has two possible answers, "yes" or "no". And yet you never get the answer "yes" or "no". If $p < 0.05$, the answer is "probably not", and if $p > 0.05$, the answer is "no comment". Interpreting $p > 0.05$ as "no" leads to all sorts of problems

1.2.2 Errors

There are two ways that we can get a hypothesis test wrong. Either the null hypothesis is true, and we conclude that it isn't (Type 1 error), or the null hypothesis is not true, but we fail to find any evidence against it (Type II error).

Type I Error

One time in every 20 that the null hypothesis is true, we will conclude that it isn't (at the 5% significance level). The smaller the p-value, the less likely it is that we are making a type I error. If we test lots of null hypothesis at the same time, there is a good chance that at least one will produce a p-value < 0.05. This can be corrected for in a number of ways: Bonferroni's

correction is the most well-known. There is some debate about when correction for multiple testing is helpful, but it essential that you inform the reader of your study how many hypothesis tests were performed, not just how many were significant.

Type II Error

It may be that the null hypothesis is not true, but that we fail to find evidence against it. This is more likely if the study is small, which it is why it is so important to ensure that a study is sifficiently large to be worth carrying out.

1.2.3 Inappropriate Hypothesis tests

Hypothesis tests are commonly performed in situations in which they are totally inappropriate. For example, in a randomised clinical trial, the null hypothesis (both arms of the trial are random samples from the same population) is true. There is no need to test it: any differences between the arms arose by chance. Even if $p < 0.05$, and you are unlikely to see such a large difference by chance, the entire process of designing a clinical trial is aimed at ensuring that the null hypothesis is true.

Even in observational studies, p-values for the differences in potential confounders between groups being compared are unhelpful. The confounding effect depends solely on the magnitude of the difference in the confounder between the two groups. However, the p-value depends on the sample size as well. You could have two studies with exactly the same confounding effect but different p-values, or very different confounding effects with the same p-value, if the sample sizes of the two studies are different. That will not stop journals asking for p-values for these differences, unfortunately.

1.3 Common types of hypothesis test

1.3.1 One-sample t-test

The simplest type of hypothesis test is comparing a sample mean to a null hypothesis value. This is often referred to as a "one-sample t-test". The test statistic in this case is

$$T = \frac{\bar{x} - \mu}{S.E.(x)}$$

and T can be compared to a t-distribution on $n - 1$ degrees of freedom.

For example, consider the following data consisting of uterine weights (in mg) for a sample of 20 rats. Previous work suggests that the mean uterine weight for the stock from which the sample was drawn was 24mg. Does this sample confirm that suggestion ?

Weights 9, 14, 15, 15, 16, 18, 18, 19, 19, 20, 21, 22, 22, 24, 24, 26, 27, 29, 30, 32

$\bar{x} = 21.0$

S.D.(x) $= 5.912$

In this case, the standard error of $\bar{x}$ is $\frac{5.912}{\sqrt{20}} = 1.322$, so

$$T = \frac{\bar{x} - 24.0}{S.E.(x)}$$

$$\begin{aligned} &= \frac{21.0 - 24.0}{1.322} \\ &= -2.27 \end{aligned}$$

Comparing -2.27 to a t-distribution on 19 degrees of freedom gives a p-value of 0.035 I.e if the stock had a mean uterine weight of 24mg, and we took repeated random samples, less than 4 times in 100 would a sample have such a low mean weight.

This test can be performed in stata with the command `ttest`. Assuming the data were stored in a variable called `x`, the necessary syntax and resulting output would be:

```
. ttest x = 24

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      20          21    1.321881     5.91163    18.23327    23.76673
------------------------------------------------------------------------------
Degrees of freedom: 19

                            Ho: mean(x) = 24

     Ha: mean < 24               Ha: mean != 24                 Ha: mean > 24
       t =  -2.2695                 t =  -2.2695                   t =  -2.2695
   P < t =   0.0175           P > |t| =   0.0351           P > t =   0.9825
```

Note that the one sided alternative test that the weight was lower than 24mg has a p-value equal to half of the two-sided test p-value.

1.3.2 Two-sample t-test

The two-sample t-test is used for comparing the means in two groups.

```
. ttest nurseht, by(sex)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  female |     227     159.774    .4247034    6.398803    158.9371    160.6109
    male |     175    172.9571    .5224808    6.911771    171.9259    173.9884
---------+--------------------------------------------------------------------
combined |     402    165.5129    .4642267    9.307717    164.6003    166.4256
---------+--------------------------------------------------------------------
    diff |           -13.18313    .6666327               -14.49368   -11.87259
------------------------------------------------------------------------------
Degrees of freedom: 400

                Ho: mean(female) - mean(male) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t = -19.7757                 t = -19.7757                   t = -19.7757
   P < t =   0.0000           P > |t| =   0.0000           P > t =   1.0000
```

1.3.3 Comparing proportions

We can also compare proportions between two groups. The stata command to do this is `cs`, and the output from the command is shown below. There are are a number of different ways to test

for a difference in proportions: the null hypothesis could be that the risk difference is 0, or that the risk ratio is 1.

```
. cs back_p sex

                | sex                   |
                |   Exposed   Unexposed |     Total
----------------+-----------------------+----------
          Cases |       637         445 |      1082
       Noncases |      1694        1739 |      3433
----------------+-----------------------+----------
          Total |      2331        2184 |      4515
                |                       |
           Risk |  .2732733    .2037546 |  .2396456
                |                       |
                |      Point estimate   |  [95% Conf. Interval]
                |-----------------------+----------------------
 Risk difference |         .0695187      |   .044767    .0942704
      Risk ratio |         1.341188      |  1.206183    1.491304
 Attr. frac. ex. |         .2543926      |  .1709386     .329446
 Attr. frac. pop |         .1497672      |
                +-------------------------------------------
                           chi2(1) =    29.91  Pr>chi2 = 0.0000
```

1.4 Power calculations

1.4.1 How power calculations work

1.4.2 Power calculations in stata

1.5 Hypothesis tests and confidence intervals

Where possible, confidence intervals should be preferred to hypothesis tests. The confidence interval conveys more information than the hypothesis test (including whether the hypothesis test would be significant at the 5% level in many cases). It is far more useful to know a range of values within which the population value is likely to lie, than a single value that the population value is unlikely to be. There is a movement to replace p-values with confidence intervals both in the epidemiology literature and amongst statisticians.

2 Hypothesis Testing and Power Practical

2.1 Inference about a proportion

Out of 80 women in a random sample of women in Manchester, 13 were asthmatic; this could be used to calculate a 95% confidence interval for the proportion of women in Manchester with asthma. This confidence interval could be compared to the suggested prevalence of 20% in Northern England. An alternative approach would be to test the hypothesis that the true proportion, π, is 0.20.

1.1 What is the expected proportion of women with asthma under the null hypothesis ?

.....

1.2 What is the observed proportion of women with asthma ? .....

1.3 What is the standard error of the expected proportion (remember from last week that the standard error of a proportion p is given by

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

.....

1.4 The appropriate test statistic, T, is given by the formula:

$$\frac{\text{observed proportion - expected proportion}}{\text{standard error of proportion}}$$

Calculate T. .....

1.5 T should be compared to a t-distribution with how many degrees of freedom ? .....

1.6 From tables for the appropriate t-distribution, the corresponding p-value is 0.4. Is is reasonable to suppose that these women are a random sample from a population in which the prevalence of asthma is 20% ? .....

2.2 More inference about a proportion

In the sample heights and weights we have looked at, there were 412 individuals of whom 234 were women. We wish to test that there are equal numbers of men and women in our population.

2.1 What is the null hypothesis proportion of women ? .....

2.2 What is the observed proportion of women ? .....

2.3 What is the null hypothesis standard error for the proportion of women ? .....

2.4 What is an appropriate statistic for testing the null hypothesis ? .....

2.3 Inference about a mean

Load `htwt.dta` into stata with the commands (each command needs to be entered on a separate line).

```
 global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/htwt.dta
```

We wish to test whether the mean height is the same in men and women.

3.1 What is the null hypothesis difference in height between men and women ?

.....

3.2 Use the command `ttest nurseht, by(sex)` to test whether the mean height differs between men and women.

3.3 What is the mean height in men ? .....

3.4 What is the mean height in women ? .....

3.5 What is the mean difference in height between men and women, with its 95% confidence interval ?

.....

3.6 Which of the three hypothesis tests is the appropriate one in this instance ?

.....

3.7 What is the p-value from the t-test ? .....

3.8 What would you conclude ?

.....

2.4 Two-sample t-test

Compare BMI (based on the measured values, i.e. `bmi`) between men and women in `htwt.dta`, using the command `ttest bmi, by(sex).`

4.1 Is there a difference in BMI between men and women ? .....

4.2 What is the mean difference in BMI between men and women and its 95% confidence interval.

.....

4.3 Is there a difference in the standard deviation of BMI between men and women ? (This can be tested with the command `sdtest bmi, by(sex)`

.....

4.4 If there is, repeat the t-test you performed above, using the `unequal` option. Are your conclusions any different ?

.....

2.5 One sample t-test

Load the `bpwide` dataset into stata with the command `sysuse bpwide`. This consists of fiction blood pressure data, taken before and after an intervention. We wish to determine whether the intervention had affected the blood pressure.

5.1 Use the `summarize` command to calculate the mean blood pressure before and after the intervention. Has the blood pressure increased or decreased ?

.....

5.2 Generate a variable containing the change in blood pressure using the command `gen bp_diff = bp_after - bp_before`

5.3 Use the command `ttest bp_diff = 0` to test whether the change in blood pressure is statistically significant. Is it ?

.....

5.4 Give a 95% confidence interval for the change in blood pressure.

.....

2.6 Power Calculations

The following questions can all be answered using the `sampsi` command.

6.1 How many subjects would need to be recruited to have 90% power to detect a difference between unexposed and exposed subjects if the prevalence of the condition is 25% in the unexposed and 40% in the exposed, assuming equal numbers of exposed and unexposed subjects ?

.....

6.2 If the exposure was rare, so it was decided to recruit twice as many unexposed subjects as exposed subjects, how many subjects would need to be recruited ?

.....

6.3 Suppose it were only possible to recruit 100 subjects in each group. What power would the study then have ? .....

6.4 Suppose that we expect a variable to have a mean of 15 and an SD of 5 in group 1, and a mean of 17 and an SD of 6 in group 2. How large would two equal sized groups need to be to have 90% power to detect a difference between the groups ?

.....

6.5 If we wanted 95% power, how large would the groups have to be ?

.....

6.6 Suppose we could only recruit 100 subjects in group 1. How many subjects would we have to recruit from group 2 to have 90% power ?

.....

Hint: the last question can only be answered by trying different numbers for the size of group 2 and seeing what power is achieved. Sensible choice of numbers will give a result fairly quickly. The `PageUp` key is your friend.