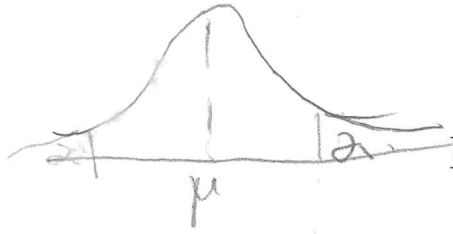
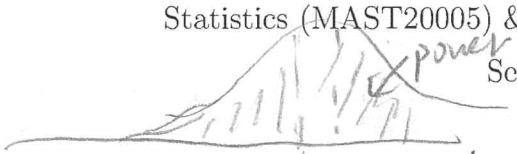


Hypothesis testing

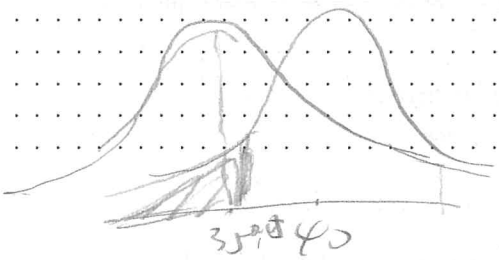
(Module 6)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019



Contents



Aims of this module

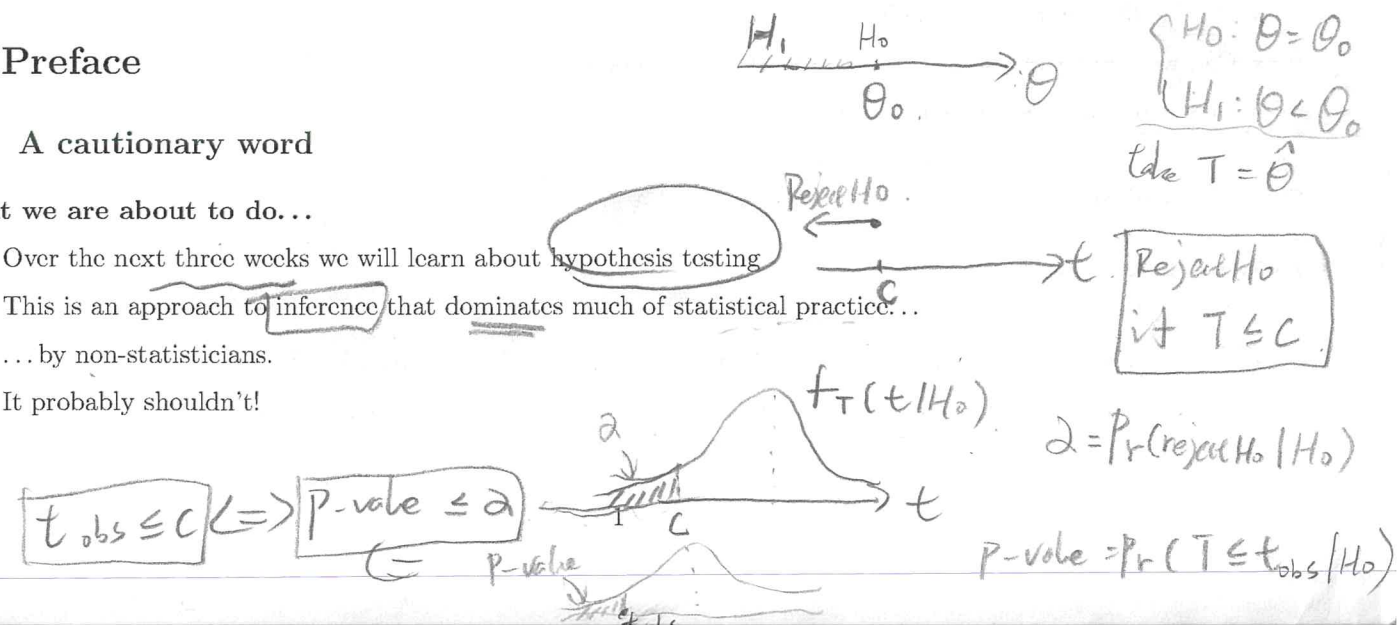
- Introduce the concepts behind **statistical hypothesis testing**
- Explain the connections between estimation and testing
- Work through a number of common testing scenarios
- Emphasise the shortcomings of hypothesis testing

1 Preface

1.1 A cautionary word

What we are about to do...

- Over the next three weeks we will learn about hypothesis testing
- This is an approach to inference that dominates much of statistical practice...
- ... by non-statisticians.
- It probably shouldn't!



Warning!

- The approaches described here are largely considered **NOT best practice** by professional statisticians
- More appropriate procedures usually exist
- ... and we have already learnt some of them!
- But we need to learn these anyway because:
  - Hypothesis testing is ubiquitous
  - Need to understand its weaknesses
  - Sometimes it's useful, or at least convenient

1.2 A motivating example

Factory example

- You run a factory that produces electronic devices
- Currently, about 6% of the devices are faulty
- You want to try a new manufacturing process to reduce this
- How do you know if it is better? Should you switch or keep the old one?
- Run an experiment: make $n = 200$ devices with the new process and summarise this by the number, Y, that are faulty
- You decide that if $Y \leqslant 7$ (i.e. $Y/n \leqslant 0.035$, or 3.5%) then you will switch to the new process
- Is this a sensible procedure?
- We can formulate this as a **statistical hypothesis test**

2 Classical hypothesis testing (Neyman-Pearson)

Research questions as hypotheses

- Research questions / studies are often often framed in terms of hypotheses
- Run an experiment / collect data and then ask:
- Do the data support/contradict the hypothesis?
- Can we frame statistical inference around this paradigm?
- Classical hypothesis testing (due to Neyman & Pearson) aims to do this

2.1 Hypotheses

Describing hypotheses

- A *hypothesis* is a statement about the population distribution
- A *parametric hypothesis* is a statement about the parameters of the population distribution
- A *null hypothesis* is a hypothesis that specifies 'no effect' or 'no change', usually denoted H_0

- An *alternative hypothesis* is a hypothesis that specifies the effect of interest, usually denoted H_1

Null hypotheses

- Special importance is placed on the null hypothesis.
- When the aim of the study/experiment is to demonstrate an effect (as it often is), the 'onus of proof' is to show there is sufficient evidence against the null hypothesis.
- I.e. we assume the null unless proven otherwise.
- Note: what is taken as the null hypothesis (i.e. the actual meaning of 'no change') will depend on the context of the study and where the onus of proof is deemed to lie.

Example

For our factory example:

- We hypothesise that the new process will lead to fewer faulty devices
- Experiment gives: $Y \sim \text{Bi}(200, p)$, where p is the proportion of faulty devices
- Null hypothesis:

$$H_0 : p = 0.06$$

- Alternative hypothesis:

$$H_1 : p < 0.06$$

Types of parametric hypotheses

- A *simple hypothesis*, also called a *sharp hypothesis*, specifies only one value for the parameter(s)
- A *composite hypothesis* specifies many possible values
- Null hypotheses are almost always simple
- Alternative hypotheses are typically composite

simple : H_0
composite : H_1

Specification of hypotheses

- Usually, the null hypothesis is on the boundary of the alternative hypothesis (here, $p = 0.06$ versus $p < 0.06$)
- It is the 'least favourable' element for the alternative hypothesis: it is harder differentiate between $p = 0.06$ and $p = 0.05$ (close to the boundary) than it is between $p = 0.06$ and $p = 0.001$ (far away from the boundary).
- For single parameters, the null is typically of the form $\theta = \theta_0$ and the alternative is either *one-sided* and takes the form $\theta < \theta_0$ or $\theta > \theta_0$, or it is *two-sided* and written as $\theta \neq \theta_0$.

2.2 Tests & statistics

Describing tests

- A *statistical test* (or *hypothesis test* or *statistical hypothesis test*, or simply a *test*) is a decision rule for deciding between H_0 and H_1.
- A *test statistic*, T, is a statistic on which the test is based
- The decision rule usually takes the form:

$$\text{reject } H_0 \text{ if } T \in A$$

value
rejection region
critical value

- The set A is called the *critical region*, or sometimes the *rejection region*.[1] If it is an interval, the boundary value is called the *critical value*.

若 reject H_0 的話

[1] Extra notes (not discussed in the lecture, for your reference only): Some authors are specific with their terminology, referring to A only as the 'rejection region' and reserving the term 'critical region' to refer to the set of values of the data (rather than the set of values of the statistic) that give rise to a rejection, i.e. $\{x : T(x) \in A\}$. This is not very common. Other authors may use the same term(s) to refer to both of these sets.

- For our example, the test statistic is Y, the decision rule is to reject H_0 if $Y \leqslant 7$, the critical region is $(-\infty, 7)$ and the critical value is 7.

Describing test outcomes

Only two possible outcomes:

1. Reject H_0
2. Fail to reject H_0

We never say that we 'accept H_0'. Rather, we conclude that there is **not enough evidence to reject it.**

Often we don't actually believe the null hypothesis. Rather, it serves as the default position of a skeptical judge, whom we must convince otherwise.

Similar to a court case: innocent until proven guilty (H_0 until proven not H_0)

2.3 Errors (Type I, Type II)

Type I error

- What could go wrong with our decision rule for the factory example?
- The new process might produce the same number of faulty devices on average, but by chance we observe at most 7 failures
- Then we would switch to the new process despite not getting any benefit
- We have rejected H_0 when H_0 is actually true; this is called a *Type I error*
- This could be quite costly – changing a production line without reducing faults would be expensive
- (Controlling the probability of a Type I error will help to mitigate against this; see later...)

Type II error

- Could anything else could go wrong if Type I error is managed?
- The new process might reduce faults, but by chance we observe more than 7 failures
- Then we would give up on the new process, forgoing its benefits
- We have failed to reject H_0 when H_0 is false; this is called a *Type II error*
- In this case, the error would be less costly in the short term but might be much more costly long-term
- (So, whilst Type I error is often the one that is specifically controlled, Type II error remains important)

Summary of outcomes

	Do not reject H_0	Reject H_0
H_0 is true	Correct!	**Type I error**
H_0 is false	**Type II error**	Correct!

2.4 Significance level & power

Significance level

$$\alpha = \Pr(\text{Type I error}) = \Pr(\text{reject } H_0 \mid H_0 \text{ true})$$

- This is called the *significance level*, or sometimes the *size*, of the test.
- In our example, under H_0 we have $p = 0.06$ and therefore $Y \sim \text{Bi}(200, 0.06)$, giving:

$$\alpha = \Pr(Y \leqslant 7 \mid p = 0.06) = 0.0829$$

- Calculate in R using: `pbinom(7, 200, 0.06)`

Probability of type II error

$$\beta = \Pr(\text{Type II error}) = \Pr(\text{do not reject } H_0 \mid H_0 \text{ false})$$

...but need to actually condition on a simple hypothesis (an actual value of p) in order for β to be well-defined.

In our example, suppose the new process actually works better and produces only 3% faulty devices on average. Then we have $Y \sim \text{Bi}(200, 0.03)$, giving $\beta = \Pr(Y > 7 \mid p = 0.03) = 0.254$.

We have halved the rate of faulty devices but still have a 25% chance of not adopting the new process!

Power

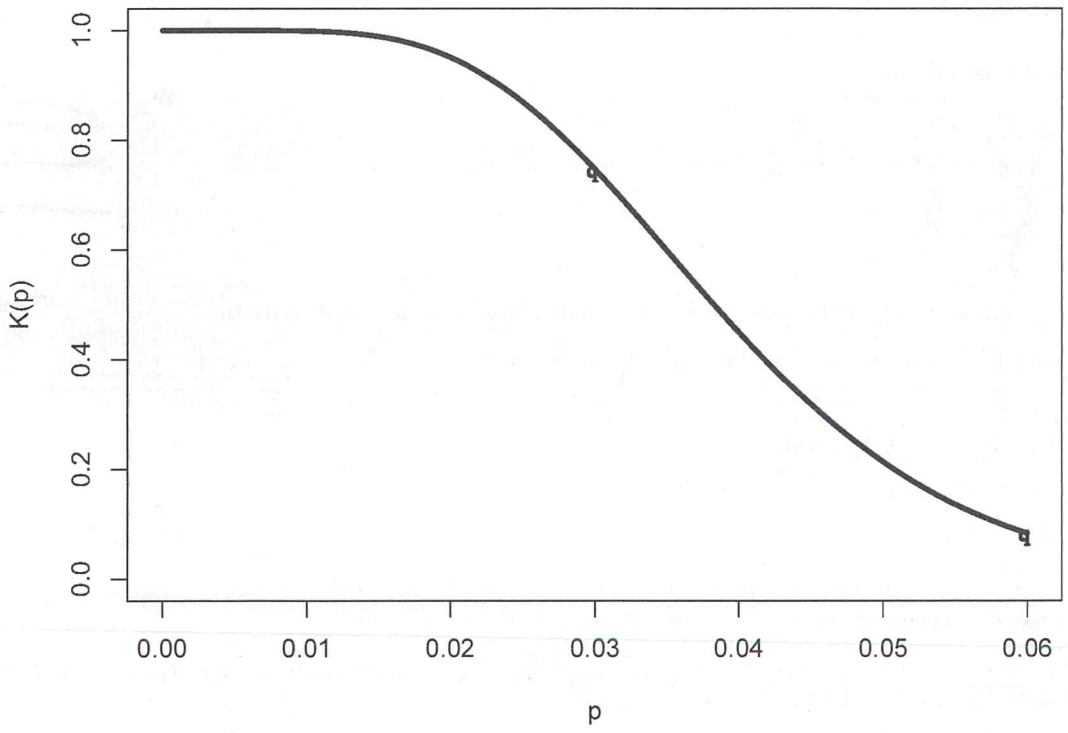
More commonly, we would report the *power* of the test, which is defined as:

$$\text{Power} = 1 - \beta = \Pr(\text{reject } H_0 \mid H_0 \text{ false})$$

$K(\theta)$

Typically, we would present this as a function of the true parameter value, e.g. $K(\theta)$

For our example, we have shown that $K(0.03) = 1 - 0.254 = 0.746$



有了P，才能calculate power.

Remarks about power

- Power is a function, not a single value: need to assume a value of p in order to calculate it
- This point is often forgotten because people talk about 'the' power of a study
- As might be expected, the test is good at detecting values of p that are close to zero but not so good when p is close to $p_0 = 0.06$.
- $K(p_0) = \alpha$, the type I error rate

Controlling errors

- Typically, we construct a test so that it has a specified significance level, α, and then maximise power while respecting that constraint
- In other words, we set the probability of a type I error to be some value (we 'control' it) and then try to minimise the probability of a type II error.
- A widespread convention is to set $\alpha = 0.05$
- I.e. we will incorrectly reject the null hypothesis about 1 time in 20
- Since $K(p_0) = \alpha$, how can we increase power while α is fixed?
- Can do this by:
  - Choosing good/optimal test statistics (see later...)
  - Increasing the sample size

2.5 Alternative formulations

Different ways to present a test

- There a other ways to present the result of a test
- These are all mathematically equivalent
- However, some are more popular than others, because they provide, or seem to provide, more information

Alternative formulation 1: based on a CI

- Instead of comparing a test statistic against a critical region...
- Calculate a $100 \cdot (1 - \alpha)\%$ confidence interval for the parameter of interest
- Reject H_0 if p_0 is not in the interval
- This gives a test with significance level α
- If the CI is constructed from a statistic T, this test is equivalent to using T as a test statistic.
- The convention of using 95% CIs is related to the convention of setting $\alpha = 0.05$

Alternative formulation 2: based on a p-value

- Instead of comparing a test statistic against a critical region...
- Calculate a p-value for the data
- The *p-value* is the probability of observing data (in a hypothetical repetition of the experiment) that is as or more extreme than what was actually observed, under the assumption that H_0 is true.
- It is typically a tail probability of the test statistic, taking the tail(s) that are more likely under H_1 as compared to H_0. (So, the exact details of this will vary between scenarios.)
- Reject H_0 if the p-value is less than the significance level
- **Note:** p-values are, strictly speaking, not part of classical hypothesis testing, but have been adopted as part of modern practice (more info later)

P-values

- P-values are like a 'short cut' to avoid calculating a critical value.
- If the test statistic is T and the decision rule is to reject H_0 if $T < c$, then the p-value is calculated as $p = \Pr(T < t_{\text{obs}})$.
- In this case, values of T that are smaller are 'more extreme', in the sense of being more compatible with H_1 rather than H_0.

- If $t_{\text{obs}} = c$, the p-value is the same as the significance level, α.
- If $t_{\text{obs}} < c$, the p-value is less than α.
- By calculating the p-value, we avoid calculating c, but the decision procedure is mathematically equivalent.
- Many different ways that people refer to p-values: P, p, p, P-value, p-value, p-value, P value, p value, p value

P-values for two-sided alternatives

- When we have a two-sided alternative hypothesis, typically the decision rule is of the form: reject H_0 if $|T| > c$
- Then the p-value is $p = \Pr(|T| > |t_{\text{obs}}|)$
- This is a **two-tailed** probability
- The easy way to calculate this is to simply double the probability of one tail:

$$p = \Pr(|T| > |t_{\text{obs}}|) = 2 \times \Pr(T > |t_{\text{obs}}|)$$

- For more general two-sided rejection regions, we also always double the relevant tail probability. This gives an implicit definition for what it means to be 'more extreme' when the two tails are not symmetric to each other. (See the examples of testing variances later on, for which the distribution of the test statistic is χ^2)

Example

- We run our factory experiment. We obtain $y = 6$ faulty devices out of a total $n = 200$.
- According to our original decision rule ($Y \leqslant 7$), we reject H_0 and decide to adopt the new process.
- Let's try it using a CI...
- Recall that $\alpha = 0.083$. Calculate a one-sided 91.7% confidence interval that gives an upper bound for p. The upper bound is: **5.4%**. This is less than $p_0 = 6\%$, so therefore reject H_0.
- Let's try it using a p-value...
- The p-value is a binomial probability, $\Pr(Y \leqslant 6 \mid p = p_0) = 0.04$. This is less than α, so therefore reject H_0.

3 Significance testing (Fisher)

Significance testing

Pre-dating the classical theory of hypothesis testing was *'significance testing'*, developed by Fisher.

The main differences to the classical theory are:

- Only use a null hypothesis, no reference to an alternative
- Use the p-value to assess the level of significance
- If the p-value is low, use as **informal** evidence that the null hypothesis is unlikely to be true.
- Otherwise, suspend judgement and collect more data.
- Use this procedure only if not much is yet known about the problem, to draw provisional conclusions only.
- This is not a decision procedure; do not talk about accepting or rejecting hypotheses.

Disputes & disagreements

- Bitter clashes between proponents!
- Fisher vs Neyman & Pearson
- In particular, Fisher thought the classical approach was ill-suited for scientific research
- Disputes never resolved (by the proponents)

4 Modern hypothesis testing

Modern practice

- The two approaches have merged in current practice
- It has led to an inconsistent/illogical hybrid
- Largely use the terminology and formulation of the classical theory (Neyman & Pearson) but commonly report the results using a p-value and talk about 'not rejecting' rather than 'accepting' the null (both of which are ideas from Fisher)
- This has given rise to many problems
- Will come back to discuss these at the end...

5 Common scenarios

Common scenarios: overview

Proportions:

- Single proportion
- Two proportions

Normal distribution:

- Single mean
- Single variance
- Two means
- Two variances

5.1 Single proportion

prop.test(....) binom.test( )

Single proportion

- Observe n Bernoulli trials with unknown probability p
- Summarise by $Y \sim \text{Bi}(n, p)$
- Test $H_0 : p = p_0$ versus $H_1 : p > p_0$, and take $\alpha = 0.05$
- Reject H_0 if observed value of Y is too large. That is, if $Y \geqslant c$ for some c.
- Choosing c: need $\Pr(Y \geqslant c \mid p = p_0) = \alpha$
- For large n, when H_0 is true

$$Z = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \approx \text{N}(0,1)$$

- This implies,

$$c = np_0 + \Phi^{-1}(1-\alpha)\sqrt{np_0(1-p_0)}$$

Example (single proportion)

- We buy some dice and suspect they are not properly weighted, meaning that the probability, p, of rolling a six is higher than usual.
- Want to conduct the test $H_0 : p = 1/6$ versus $H_1 : p > 1/6$
- Roll the dice $n = 8000$ times and observe Y sixes.
- The critical value is

$$c = 8000/6 + 1.645\sqrt{8000(1/6)(5/6)} = 1388.162$$

- We observe $y = 1389$ so we reject H_0 at the 5% level of significance and conclude that the die comes up with 6 too often

Single proportion, cont'd

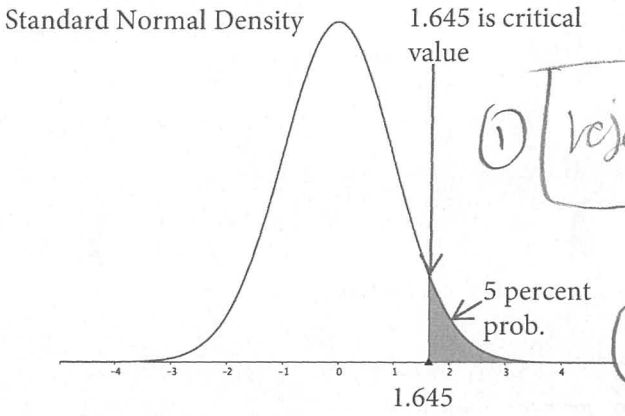
- It is more common to use standardised test statistics
- Here, report Z instead of Y and compare to $\Phi^{-1}(1-\alpha)$ instead of c
- Express Z as the standardised proportion of 6's,

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \approx \mathrm{N}(0,1)$$

- Decision rule: reject H_0 if $Z > \Phi^{-1}(1-\alpha)$
- In the previous example,

$$z = \frac{1389/8000 - 1/6}{\sqrt{(1/6)(5/6)/8000}} = 1.67$$

and since $z > \Phi^{-1}(0.95) = 1.645$ we reject H_0.



- Suppose we used a two-sided alternative, $H_1 : p \neq 1/6$
- This would me we want to be able detect deviations in either direction: whether rolling a six is either lower **or** higher than usual.
- We still compute the same test statistic,

$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathrm{N}(0,1)$$

- but the critical region has changed: we reject H_0 at level α if $|Z| > \Phi^{-1}(1-\alpha/2)$
- In the previous example, we would use $\Phi^{-1}(1-\alpha/2) = 1.96$. Since $z = 1.67$, we would **not** reject H_0.

Summary of tests for a single proportion

H_0	H_1	Critical region
$p = p_0$	$p > p_0$	$z = \frac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} > \Phi^{-1}(1-\alpha)$
$p = p_0$	$p < p_0$	$z = \frac{y/n - p_0}{\sqrt{p_0(1-p_0)/n}} < \Phi^{-1}(\alpha)$
$p = p_0$	$p \neq p_0$	$\lvert z\rvert = \frac{\lvert y/n - p_0\rvert}{\sqrt{p_0(1-p_0)/n}} > \Phi^{-1}(1-\alpha/2)$

Example 2 (single proportion)

- A woman claims she can tell whether the tea or milk was added first to a cup of tea
- Given 40 cups of tea and for each cup the order was determined by tossing a coin
- The woman gave the correct answer 29 times out of 40
- Is this evidence (at the 5% level of significance) that her claim is valid?
- Let p be the probability the woman gets the correct order for a single cup of tea

$$H_0 : p = 0.5 \quad \text{versus} \quad H_1 : p > 0.5$$

- We need evidence against the hypothesis that she is simply guessing, the one-sided alternative is appropriate here.
- Data: $y/n = 29/40 = 0.725$

$$z = \frac{0.725 - 0.5}{\sqrt{0.5 \times 0.5/40}} = 2.84$$

- Critical value $\Phi^{-1}(0.95) = 1.645$, therefore reject H_0 and conclude that the data supports the woman's claim
- Alternatively, we could do this via a p-value:

$$\text{p-value} = \Pr(Z > 2.84) = \Phi(-2.84) = 0.00226$$

- Since $0.00226 < 0.05$, we reject H_0.

R code examples

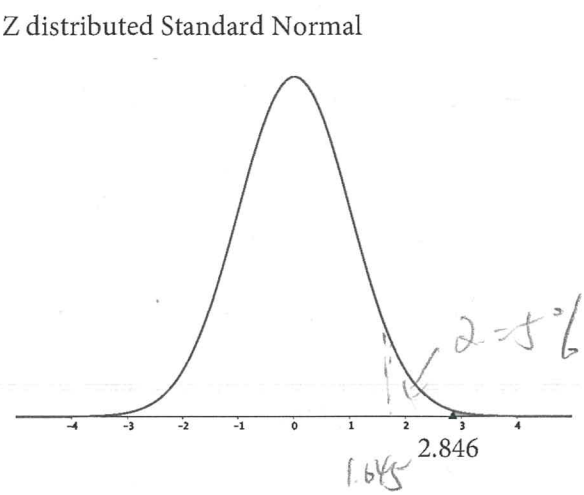
```
> p1 = prop.test(29, 40, p = 0.5,
+                alternative = "greater", correct = FALSE)
> p1

1-sample proportions test without continuity correction

data:  29 out of 40, null probability 0.5
X-squared = 8.1, df = 1, p-value = 0.002213
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.597457 1.000000
sample estimates:
    p
0.725

> sqrt(p1$statistic)
X-squared
  2.84605

> 1 - pnorm(2.846)
[1] 0.002213610
```

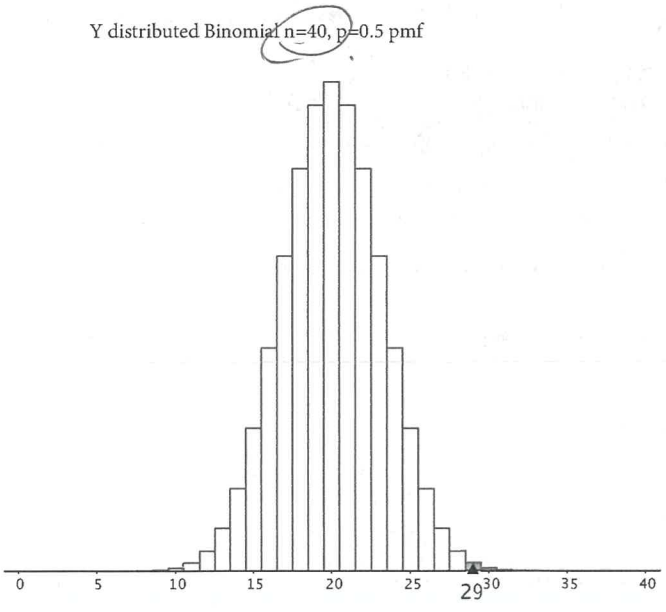


There is also an exact test based on the binomial probabilities:

```
> binom.test(29, 40, p = 0.5, alternative = "greater")

Exact binomial test

data:  29 and 40
number of successes = 29, number of trials = 40,
                                    p-value = 0.003213
alternative hypothesis: true probability of success
                                    is greater than 0.5
95 percent confidence interval:
 0.5861226 1.0000000
sample estimates:
probability of success
                 0.725
```



5.2 Two proportions

Two proportions

- Comparing two proportions: p_1 and p_2 are the probabilities of success in two different populations.
- Wish to test:

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 > p_2$$

based on independent samples (from the two populations) of size n_1 and n_2 with Y_1 and Y_2 successes.

- Know

$$Z = \frac{Y_1/n_1 - Y_2/n_2 - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n2}} \approx \mathrm{N}(0,1)$$

- Under H_0 can assume that $p_1 = p_2 = p$,

$$Z = \frac{Y_1/n_1 - Y_2/n_2}{\sqrt{p(1-p)(1/n_1 + 1/n2)}} \approx \mathrm{N}(0,1)$$

- Let $\hat{p}_1 = y_1/n_1$, $\hat{p}_2 = y_2/n_2$, $\hat{p} = (y_1 + y_2)/(n_1 + n_2)$,
- Reject H_0 at level α if

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n2)}} > \Phi^{-1}(1-\alpha)$$

Example (two proportions)

We run a trial of two insecticides. The standard one kills 425 out of 500 mosquitoes, while the experimental one kills 459 out of 500. Is the experimental insecticide more effective?

Let p_1 and p_2 be the proportion of all mosquitoes killed by experimental and standard spray, respectively.

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 > p_2$$

```
> x <- c(459, 425)
> n <- c(500, 500)
> p.hat <- (x[1] + x[2]) / (n[1] + n[2])
> p1 <- x[1] / n[1]
> p2 <- x[2] / n[2]
> z <- (p1 - p2) / sqrt(p.hat * (1 - p.hat) *
+                       (1 / n[1] + 1 / n[2]))
> pvalue <- 1 - pnorm(z)
> print(c(p1, p2, z, pvalue), digits = 3)
[1] 0.918000 0.850000 3.357560 0.000393
```

Alternatively, can use the R function `prop.test()` which calculates the statistic $\chi^2 = Z^2$ and compares against a χ^2_1 distribution.

```
> prop.test(x, n, alternative = "greater", correct = FALSE)

2-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 11.273, df = 1, p-value = 0.0003932
alternative hypothesis: greater
95 percent confidence interval:
 0.03487541 1.00000000
sample estimates:
prop 1 prop 2
 0.918  0.850
```

Summary of tests for two proportions

H_0	H_1	Critical region
$H_0 : p_1 = p_2$	$H_1 : p_1 > p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1+1/n2)}} > \Phi^{-1}(1-\alpha)$
$H_0 : p_1 = p_2$	$H_1 : p_1 \leq p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1+1/n2)}} < \Phi^{-1}(\alpha)$
$H_0 : p_1 = p_2$	$H_1 : p_1 \neq p_2$	$\lvert z \rvert = \frac{\lvert\hat{p}_1 - \hat{p}_2\rvert}{\sqrt{\hat{p}(1-\hat{p})(1/n_1+1/n2)}} > \Phi^{-1}(1-\alpha/2)$

5.3 Single mean

Example (normal, single mean, known σ)

- A tyre manufacturer claims that a new tyre will last 48,000 km on average. A consumer group tests a sample of 50 tyres and finds the mean is 45,286 km and the standard deviation is known to be $\sigma = 6012.60$ km. Is this evidence against the manufacturer's claim?
- Let μ be the mean tyre lifetime.

$$H_0 : \mu = 48,000 \quad \text{versus} \quad H_1 : \mu < 48,000$$

(Need evidence against the manufacturer to query claims)

- Recall that,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- We reject H_0 in favour of H_1 at level α if

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{50}} < \Phi^{-1}(\alpha)$$

- We have $\Phi^{-1}(0.05) = -1.645$ and,

$$z = \frac{45286 - 48000}{6021.6/\sqrt{50}} = -3.187$$

so we reject H_0 at the 5% level of significance and conclude the tyre life is lower than the claimed 48,000 km

Summary of tests for single mean, σ known

H_0	H_1	Critical region	
$\mu = \mu_0$	$\mu > \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \geqslant \Phi^{-1}(1-\alpha)$	or $\bar{x} \geqslant \mu_0 + \Phi^{-1}(1-\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$	$\mu < \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \leqslant \Phi^{-1}(\alpha)$	or $\bar{x} \leqslant \mu_0 + \Phi^{-1}(\alpha)\frac{\sigma}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert z\rvert = \frac{\lvert\bar{x}-\mu_0\rvert}{\sigma/\sqrt{n}} \geqslant \Phi^{-1}(1-\alpha/2)$	or $\lvert\bar{x}-\mu_0\rvert \geqslant \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$

The critical regions are equivalent to the respective confidence intervals containing μ_0.

Normal, single mean, unknown σ

- Often the variance is not known and the sample size is small.
- Recall if the sample is from $\mathrm{N}(\mu, \sigma^2)$ then

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

and we may base our tests on T

- This is known as the *t-test*

Example (normal, single mean, unknown σ)

- Let $X \sim \mathrm{N}(\mu, \sigma^2)$ model the growth (in mm) of a tumor in a mouse.

$$H_0 : \mu = 4.0 \quad \text{versus} \quad H_1 : \mu \neq 4.0$$

We have $n = 9$, and want a test with significance level $\alpha = 0.1$.

- Reject H_0 if

$$|t| = \frac{|\bar{x} - 4|}{s/\sqrt{9}} > c$$

where c is the 0.95 quantile of t_8

$$\Phi_8^{-1}\left(1 - \frac{0.1}{2}\right) = 13\,\Phi_8^{-1}(0.95)$$

- Conduct experiment with results: $\bar{x} = 4.3$, $s = 1.2$. Also, we can look up / calculate that $c = 1.86$. Therefore our test comparison is,

$$t = \frac{|4.3 - 4.0|}{1.2/\sqrt{9}} = 0.75 < 1.86$$

- At the 10% level of significance we cannot reject H_0 and conclude there is not enough evidence that the tumour mean departs from 4 mm
- The p-value is

$$\Pr(|T| \geqslant 0.75) = 2\Pr(T \geqslant 0.75) = 0.475 > 0.1$$

In R, you can calculate this with the command: `2 * (1 - pt(0.75, 8))` which gives `0.4747312`

Summary of tests for single mean, σ unknown

H_0	H_1	Critical region	
$\mu = \mu_0$	$\mu > \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \geqslant F^{-1}(1-\alpha)$	or $\bar{x} \geqslant \mu_0 + F^{-1}(1-\alpha)\frac{s}{\sqrt{n}}$
$\mu = \mu_0$	$\mu < \mu_0$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \leqslant F^{-1}(\alpha)$	or $\bar{x} \leqslant \mu_0 + F^{-1}(\alpha)\frac{s}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t\rvert = \frac{\lvert\bar{x}-\mu_0\rvert}{s/\sqrt{n}} \geqslant F^{-1}(1-\alpha/2)$	or $\lvert\bar{x}-\mu_0\rvert \geqslant F^{-1}(1-\alpha/2)\frac{s}{\sqrt{n}}$

F^{-1} is the inverse cdf of t_{n-1}.

The critical regions are equivalent to the respective confidence intervals containing μ_0.

Paired-sample t-test

As with confidence intervals, if we observe pairs of numbers (X_i, Y_i) from two different populations, we can take their differences and apply methods for a single sample (in this case, a t-test).

5.4 Single variance

Example (normal, single variance)

- A test about the variance

$$H_0: \sigma^2 = 100 \quad \text{versus} \quad H_1: \sigma^2 \neq 100$$

- $n = 23$, $\alpha = 0.05$, $s^2 = 147.82$.
- Recall

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- So we reject H_0 if

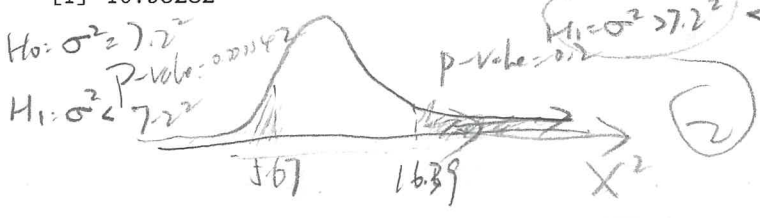
$$\chi^2 < F^{-1}(\alpha/2) = 10.98 \quad \text{or} \quad \chi^2 > F^{-1}(1-\alpha/2) = 36.78$$

Where F^{-1} is the inverse cdf of χ^2_{22}.

R code for quantiles

```
> qchisq(0.975, 22)
[1] 36.78071

> qchisq(0.025, 22)
[1] 10.98232
```



Back to the example

- We actually observe,

$$\chi^2 = \frac{22 \times 147.82}{100} = 32.52$$

- and

$$10.98 < 32.52 < 36.78$$

so we cannot reject H_0.

5.5 Two means

Example (normal, two means, pooled variance)

- A botanist wants to compare the effect of two different hormone concentrations on plant growth.
- Data: X and Y are the growth in the first 26 hours after treatment with hormone 1 & 2, respectively
- We hypothesise less growth with hormone 1
- Suppose $X \sim \mathrm{N}(\mu_X, \sigma^2)$ and $Y \sim n(\mu_Y, \sigma^2)$,

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X < \mu_Y$$

- Samples of sizes n and m. We use the two-sample pivot but assuming H_0 (which makes the $\mu_X - \mu_Y$ term disappear),

$$T = \frac{\bar{X} - \bar{Y}}{S_P\sqrt{1/n + 1/m}} \sim t_{n+m-2}$$

where S_P^2 is the pooled variance estimate:

$$S_P = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}}$$

- Reject H_0 if $t < c$, where c is the α quantile of t_{n+m-2}.
- Here $n = 11$, $m = 13$, $\bar{x} = 1.03$, $s_X^2 = 0.24$, $\bar{y} = 1.66$, $s_Y^2 = 0.35$,

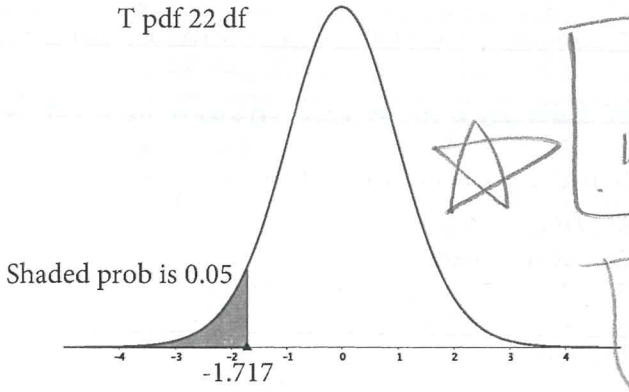
$$S_P^2 = \frac{10 \times 0.24 + 12 \times 0.35}{11 + 13 - 2} = 0.3 = 0.548^2$$

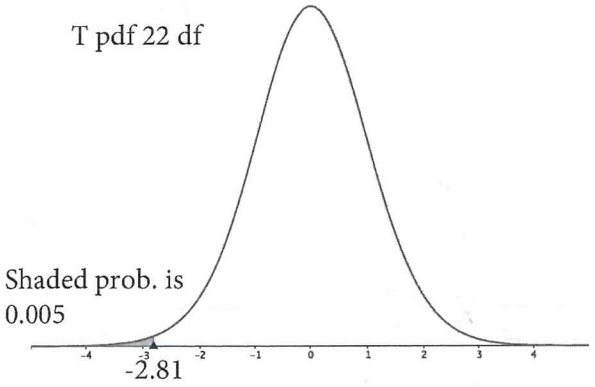
and thus

$$t = \frac{1.03 - 1.06}{\sqrt{0.3 \times (1/11 + 1/13)}} = -2.81$$

- The critical value is $c = -1.717$ (corresponding to $\alpha = 0.05$ and $n + m - 2 = 22$) so we reject H_0 and conclude that there is statistically significant evidence of less growth with hormone 1
- The p-value is

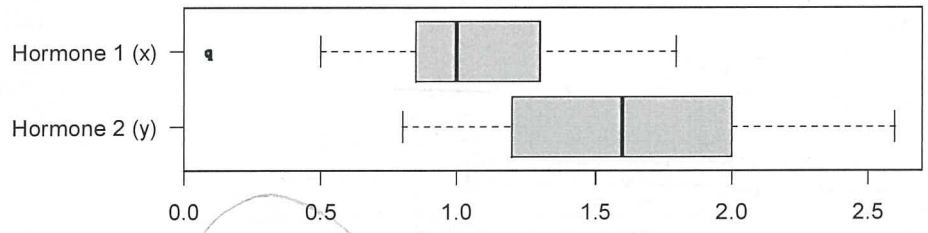
$$\Pr(T < -2.81) = 0.0051 < 0.05$$






```
> x = c(0.8, 1.8, 1.0, 0.1, 0.9, 1.7,
+       1.0, 1.4, 0.9, 1.2, 0.5)

> y = c(1, 0.8, 1.6, 2.6, 1.3, 1.1, 2.4,
+       1.8, 2.5, 1.4, 1.9, 2, 1.2)
```




```
> t.test(x, y, alternative = "less", var.equal = TRUE)

Two Sample t-test

data:  x and y
t = -2.8112, df = 22, p-value = 0.005086
alternative hypothesis:
  true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2468474
sample estimates:
mean of x mean of y
 1.027273  1.661538
```

Example 2 (normal, two means, pooled variance)

The weights (in grams) of packages filled by two methods are $X \sim \mathrm{N}(\mu_X, \sigma^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma^2)$.

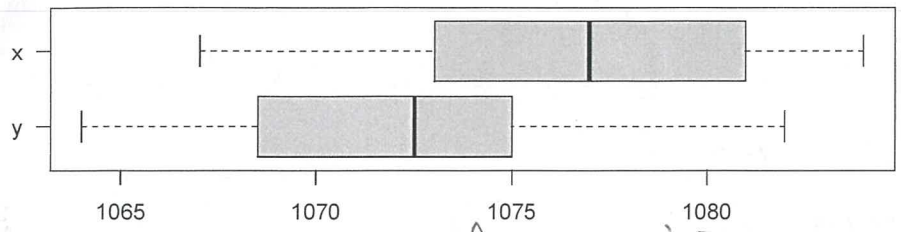
Interested in testing:

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

Similar to before but two-sided alternative, so use a two-sided critical region.

```
> x = c(1071, 1076, 1070, 1083, 1082, 1067,
+       1078, 1080, 1075, 1084, 1075, 1080)

> y = c(1074, 1069, 1075, 1067, 1068, 1079,
+       1082, 1064, 1070, 1073, 1072, 1075)
```




```
> t.test(x, y, var.equal = TRUE)

Two Sample t-test

data:  x and y
t = 2.053, df = 22, p-value = 0.05215
alternative hypothesis:
  true difference in means is not equal to 0
95 percent confidence interval:
 -0.04488773  8.87822107
sample estimates:
mean of x mean of y
 1076.750  1072.333
```

The p-value is 0.052.

Therefore, at the 5% level of significance we do not have enough evidence to reject the null hypothesis.

Given the closeness of result, it would be worth trying to collect more data.

Example (normal, two means, different variances)

$X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ correspond to the thickness of regular gum and bubble gum, with samples of size $n = 40$ and $m = 50$ respectively.

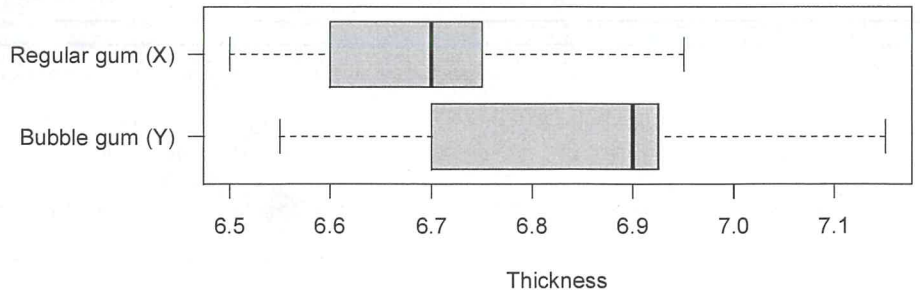
$$H_0 \colon \mu_X = \mu_Y \quad \text{versus} \quad H_1 \colon \mu_X \neq \mu_Y$$

We use the Welch approximation (see Module 3),

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \approx t_r$$

```
> head(gum, 4)
  Thickness Group
1      6.85     X
2      6.60     X
3      6.70     X
4      6.75     X
> table(gum$Group)

 X  Y
50 40
```



```
> t.test(Thickness ~ Group, data = gum)

        Welch Two Sample t-test

data:  Thickness by Group
t = -4.8604, df = 67.219, p-value = 7.357e-06
alternative hypothesis:
  true difference in means is not equal to 0
95 percent confidence interval:
 -0.19784277 -0.08265723
sample estimates:
mean in group X mean in group Y
        6.70100         6.84125

> t.test(Thickness ~ Group, data = gum, var.equal = TRUE)

        Two Sample t-test

data:  Thickness by Group
t = -5.0524, df = 88, p-value = 2.345e-06
alternative hypothesis:
  true difference in means is not equal to 0
95 percent confidence interval:
 -0.19541537 -0.08508463
sample estimates:
mean in group X mean in group Y
        6.70100         6.84125
```

5.6 Two variances

Normal, two variances

- Independent random samples: $X_1, \ldots, X_n \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$.
- Recall

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2_{n-1} \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$$

  and since the samples are independent, these statistics are also independent.

- Want to test,

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

- When H_0 is true, we have $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ and therefore can use the statistic,

$$F = \frac{\frac{(n-1)S_X^2}{\sigma^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma^2}/(m-1)} = \frac{S_X^2}{S_Y^2} \sim \mathrm{F}_{n-1,m-1}$$

Example (normal, two variances)

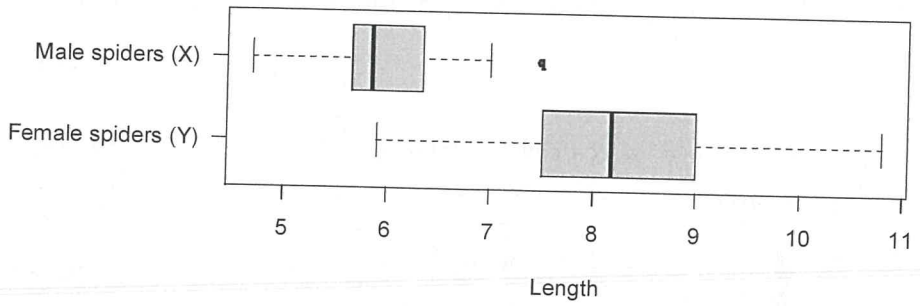
Measure the lengths of male spiders, $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and also female spiders, $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$.

$$H_0 : \sigma^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

```
> head(spiders, 4)
  Length Group
1   5.20     X
2   4.70     X
3   5.75     X
4   7.50     X
```

```
> table(spiders$Group)

 X  Y
30 30
```



The data give: $s_X^2 = 0.4399$, $s_Y^2 = 1.41$,

$$\frac{s_Y^2}{s_X^2} = 3.2055 > 2.67 \quad (0.995 \text{ quantile of } F_{29,29})$$

so we reject H_0 at 1% level of significance.

```
> var.test(Length ~ Group, data = spiders)

        F test to compare two variances

data:  Length by Group
F = 3.2054, num df = 29, denom df = 29, p-value = 0.002458
alternative hypothesis:
  true ratio of variances is not equal to 1
95 percent confidence interval:
 1.525637 6.734441
sample estimates:
ratio of variances
          3.205357
```

6 Usage & (mis)interpretation

Choice of significance level

- Somewhat arbitrary.
- A balance between type I error and type II error. The appropriate balance is likely to depend on your problem.
- Whatever you choose, always remember that you are never guaranteed to be error-free.
- $\alpha = 0.05$ is a very common convention (c.f. 95% confidence level). If you don't have a good basis for choosing a specific α for your problem, then following this convention will usually be acceptable.
- Specific fields of application can have their own conventions which are very different. For example:
  - Genome-wide association studies require p-values of around 10^{-8}
  - High-energy physics (particle physics) requires p-values under 0.003 ('3 sigma') for reporting 'evidence of a particle' and p-values under 0.0000003 (3×10^{-7}; '5 sigma') for reporting a 'discovery'.

(See this blog post for more info.)

Misinterpretations of p-values

- Many misconceptions about p-values:
  - The p-value is the probability that the null hypothesis is true

  - The p-value is the probability that the alternative hypothesis is false
  - A 'significant' p-value implies that the null hypothesis is false
  - A 'significant' p-value implies that the alternative hypothesis is true
  - A 'significant' p-value implies that the effect detected is of large magnitude or of practical importance
- **None of these are true**
- These are just the tip of the iceberg!
- Similar issues arise with oversimplified interpretations of confidence intervals
- Can read much more about this in various articles...

For example, see this article.

'Absence of evidence' versus 'evidence of absence'

- An inability to reject the null could be either because the null is (approximately) true **OR** simply due to insufficient data.
- In this case, absence of evidence is not evidence of absence...
- ...but it could be, if only we quantified our evidence better!

Decisions versus inference

- Hypothesis testing is a decision procedure
- Therefore, it is not actually inference proper
- Decisions are about **behaviour**, inference is about **knowledge**
- Knowledge can drive behaviour, but they are not the same thing
- Decisions are clear, knowledge is ambiguous
- Decisions are black & white, knowledge is a shade of grey

Following the scientific process

- Does hypothesis testing parallel the scientific process?
- I.e. set up a testable hypothesis and then run an experiment to try to disprove it?
- Perhaps... but a binary decision doesn't carry much informative content
- At best, this a very cartoonish view of science
- Better to think of science as a process of cumulative evidence gathering
- Talk about **degrees of evidence** rather than black & white truth claims

Why is hypothesis testing so popular?

- People want 'objective' procedures which lead to conclusive statements of truth.
- Hypothesis testing, esp. when used with p-values, seemed to offer this, especially since it seems to have been 'blessed' by statisticians.
- In reality, it is too good to be true. P-values are too prone to misinterpretation. The ability to draw strong conclusions is a misconception about the nature of inference.
- But the genie is out of the bottle... and has been rampant for more than half a century!
- Statistical education hasn't helped.
- A circular problem: we teach the use of $p = 0.05$ because it's 'in demand', but that only perpetuates it's use.
- But the call for reform is getting stronger now.

- You are lucky, we are teaching you to set the right foot forward from day one!

What's an alternative?

- Think about the actual question at hand. What are you trying to find out?
- Usually, it will be best formulated in terms of estimation or prediction.
- 'How much does my risk of lung cancer increase if I smoke 10 extra cigarettes per day?', instead of simply 'Does smoking cause lung cancer?'
- Interval estimation techniques are a better way to answer such questions

It's all about uncertainty

- Statistics is not magic
- We cannot make uncertainty disappear
- If anything, we do the opposite: we quantify it so that it is plainly visible
- This can sometimes be confronting!
- Always keep your critical thinking hat on: do the results look plausible in light of previous knowledge?
- And be conscious of how you describe your results: which shade of grey are you after this time?

When should we actually use hypothesis testing?

- Follow in Fisher's footsteps
- Use it as an exploratory tool
- Use it when convenient, to help inform further analyses
- If reporting the results, then set them in context and avoid pure black & white conclusions
- It's helpful in designing studies, especially the concept of error probabilities (including power)
- Sometimes we actually require decisions, e.g. quality control applications (such as our factory example)
- Pure hypothesis testing is adequate for such settings, although more sophisticated procedures exist (statistical decision theory)

Why are we learning hypothesis testing?

- Why teach this stuff if it is 'wrong'?
- To understand current practice, and its strengths and weaknesses
- To understand the concepts and language used by others
- Sometimes it is useful and convenient
- Sometimes it is simpler or more practical than alternative procedures, even if we believe the latter are more 'correct'

	Sample size	Sample mean	Sample sed
Model	30	8.63	7.1
Control	30	10.97	8.9

both are Normal.

a). difference exists in between, assume same variance, $\alpha = 0.05$.

b). p-value.

c). Variance differ? give CI.

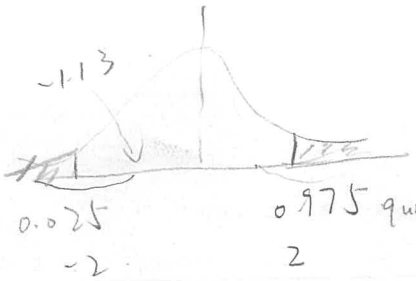
a) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

$$= \sqrt{\frac{29 \times 7.1^2 + 29 \times 8.9^2}{30+30-2}} = 8.05$$

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{8.63 - 10.97}{8.05\sqrt{\frac{1}{30} + \frac{1}{30}}} = -1.13$$



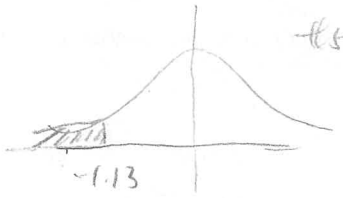
$|t| = |-1.13| < 2$. Do not reject H_0.

b). p-value $= 2 \times Pr(T < -1.13 | H_0)$

$Pr(T < -1.67) < Pr(T < -1.13)$

0.05

p-value > 0.01



$qt(0.05, 58) = -1.67$ given

c) $0.95 = Pr\left(c < \frac{s_y^2/\sigma_y^2}{s_x^2/\sigma_x^2} < d\right)$

c: 0.025, 0.48; d: 0.975, 2.10

$F_{29,29}$

$$= Pr\left(0.48\frac{s_x^2}{s_y^2} < \frac{\sigma_x^2}{\sigma_y^2} < 2.1\frac{s_x^2}{s_y^2}\right)$$

$\frac{s_x^2}{s_y^2} = \frac{7.1^2}{8.9^2} = 0.636$

95% CI: (0.31, 1.34)

don't have evidence to prove that differ.

Distribution-free methods

(Module 7)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

Contents

Aims of this module

- Introduce inference methods that do not make strong distributional assumptions
- Explain the highly used Pearson's chi-squared test

1 Introduction

Distribution-free methods

- So far, have only considered tests that assume a specified form for the population distribution.
- We don't always want to make such assumptions.
- Instead, we can use *distribution-free* methods.
- Here, we will learn about various distribution-free hypothesis tests.

An aside: *distribution-free* versus *non-parametric*

- The term *non-parametric* is also often used to describe methods that do not assume a specific distributional form.
- It is usually a misnomer: the methods typically **do** make use of parameters, but there are usually a large number of them and they adapt to the data.
- Thus, a better term might be **super-parameteric**
- (Note: we won't be covering any advanced methods of this form in this subject.)
- In any case, the convention has stuck, so you will see either of the labels 'distribution-free' or 'non-parameteric' being used.

Distribution-free tests

- Even without making distributional assumptions, it is possible to obtain exact or asymptotic sampling distributions for various statistics.
- Can use these as a basis for hypothesis tests.
- Often the distribution-free test statistic is approximately normally distributed
- ... the Central Limit Theorem strikes again!

2 Testing for a difference in location

Extracting information with fewer assumptions

- How can we assess the information in a sample without assuming a distribution?
- Specifying a distribution is somewhat analogous to specifying a scale of measurement, so...
- How do we compare numbers without a scale?
- Two strategies:
  1. **(Sign)** Only record whether a number is smaller or greater than a reference number, i.e. replace them by binary indicator variables.
  2. **(Rank)** Only retain information about the order of the numbers, i.e. replace them by their rank order.
- Each of these throws away some information, but hopefully retains enough to be useful.
- We now look at a few methods that use these strategies.

Aim: test for the median

- Let X have median m
- We have an iid sample of size n from X
- Can we test $H_0: m = m_0$ with very few assumptions?
- (Want to find distribution-free alternatives to tests about the mean, such as the t-test)
- (Typically consider medians rather than means when distribution-free)

2.1 Sign test

Sign test

- We assume X is continuous — only one assumption
- (No further assumptions!)
- Compute, Y, the number of positive numbers amongst $X_1 - m_0, \ldots, X_n - m_0$
- In other words, replace X_i with $\text{sgn}(X_i - m_0)$
- Under H_0, we have $Y \sim \text{Bi}(n, 0.5)$
- Tests proceed as usual...

Example (sign test)

The time between calls to a switchboard is represented by X.

$$H_0: m = 6.2 \quad \text{versus} \quad H_1: m < 6.2$$

i	x_i	$x_i - 6.2$	Sign	i	x_i	$x_i - 6.2$	Sign
1	6.80	0.60	+1	11	18.90	12.70	+1
2	5.70	−0.50	−1	12	16.90	10.70	+1
3	6.90	0.70	+1	13	10.40	4.20	+1
4	5.30	−0.90	−1	14	44.10	37.90	+1
5	4.10	−2.10	−1	15	2.90	−3.30	−1
6	9.80	3.60	+1	16	2.40	−3.80	−1
7	1.70	−4.50	−1	17	4.80	−1.40	−1
8	7.00	0.80	+1	18	18.90	12.70	+1
9	2.10	−4.10	−1	19	4.80	−1.40	−1
10	19.00	12.80	+1	20	7.90	1.70	+1

- Y is the number of positive signs. Reject H_0 if Y too small. (If median < 6.2 then expect fewer than 1/2 of the observations to be greater than 6.2.)
- Since $\Pr(Y \leqslant 6) = 0.0577 \approx 0.05$, an appropriate rejection rule is to reject H_0 if $Y \leqslant 6$. (In R: `pbinom(6, 20, 0.5)`)
- We observed $y = 11$, so cannot reject H_0.
- The p-value is $\Pr(Y \leqslant 11) = 0.75 > 0.05$ so cannot reject H_0. (In R: `pbinom(11, 20, 0.5)`)

R code

```
> binom.test(11, 20, alternative = "less")

    Exact binomial test

data:  11 and 20
number of successes = 11, number of trials = 20,
 p-value = 0.7483
alternative hypothesis: true probability of
                        success is less than 0.5
95 percent confidence interval:
 0.0000000 0.7413494
sample estimates:
probability of success
                  0.55
```

Sign test for paired samples

Can also use the sign test for paired samples: simply replace (x_i, y_i) with $\text{sgn}(x_i - y_i)$.

For example:

i	x_i	y_i	Sign
1	8.9	10.3	−1
2	26.7	11.7	+1
3	12.4	5.2	+1
4	34.3	36.9	−1

Use of the sign test

- The sign test requires few assumptions
- But it doesn't use information on the size of the differences, so it can be insensitive to departures from H_0
- In other words, large type II error or small power
- Tends to only be used when the data are not numerical but for which comparisons between values are meaningful (e.g. ordinal data)

2.2 Wilcoxon signed-rank test (one-sample)

Wilcoxon one-sample test

- Now, assume the underlying distribution is also symmetrical (as well as continuous)
- Same null hypothesis ($H_0 : m = m_0$) against a one-sided or two-sided alternative
- Determine the ranks of: $|X_1 - m_0|, \ldots, |X_n - m_0|$
- Replace the data by signed ranks, X_i becomes $\text{sgn}(X_i - m_0) \cdot \text{rank}(|X_i - m_0|)$
- The *Wilcoxon signed-rank statistic*, W, is the sum of these signed ranks
- Using this as a basis for a test gives the *Wilcoxon signed-rank test*, also known as the *Wilcoxon one-sample test.*

Alternative definitions

- Textbooks and software packages vary in the statistic they use
- We just defined: W is the sum of the signed ranks
- A popular alternative: V is the sum of the positive ranks only
- V is a bit easier to calculate, esp. by hand
- R uses V
- V and W are deterministically related (can you derive the formula?)
- V and W have different (but related) sampling distributions
- Using either statistic leads to equivalent test procedures

Example (Wilcoxon one-sample test)

- The lengths of 10 fish are:

  5.0, 3.9, 5.2, 5.5, 2.8, 6.1, 6.4, 2.6, 1.7, 4.3

- Interested in testing: $H_0 : m = 3.7$ versus $H_1 : m > 3.7$

i	x_i	$x_i - 3.7$	$\lvert x_i - 3.7 \rvert$	Rank	Signed rank
1	5.0	1.3	1.3	5	5
2	3.9	0.2	0.2	1	1
3	5.2	1.5	1.5	6	6
4	5.5	1.8	1.8	7	7
5	2.8	−0.9	0.9	3	−3
6	6.1	2.4	2.4	9	9
7	6.4	2.7	2.7	10	10
8	2.6	−1.1	1.1	4	−4
9	1.7	−2.0	2.0	8	−8
10	4.3	0.6	0.6	2	2

- The sum of signed ranks is:

$$W = 5 + 1 + 6 + 7 - 3 + 9 + 10 - 4 - 8 + 2 = 25$$

- Alternatively, the sum of positive ranks is:

$$V = 5 + 1 + 6 + 7 + 9 + 10 + 2 = 40$$

Decision rule

- What is an appropriate critical region?
- If $H_1 : m > 3.7$ is true, we expect more positive signs. Then W should be large, so the critical region should be $W \geqslant c$ for a suitable c.
- (For other alternative hypotheses, e.g. two-sided, need to modify this accordingly.)

- If H_0 is true then $\Pr(X_i < m_0) = \Pr(X_i > m_0) = \frac{1}{2}$.
- Assignment of the n signs to the ranks are mutually independent (due to symmetry assumption)
- W is the sum of the integers $1, \ldots, n$, each with a positive or negative sign
- Under H_0, $W = \sum_{i=1}^{n} W_i$ where

$$\Pr(W_i = i) = \Pr(W_i = -i) = \frac{1}{2}, \quad i = 1, \ldots, n$$

- The mean under H_0 is $\mathbb{E}(W_i) = -i \cdot \frac{1}{2} + i \cdot \frac{1}{2} = 0$, so $\mathbb{E}(W) = 0$
- Similarly, $\text{var}(W_i) = \mathbb{E}(W_i^2) = i^2$ and

$$\text{var}(W) = \sum_{i=1}^{n} \text{var}(W_i) = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

- A more advanced argument shows that for large n this statistic approximately follows a normal distribution when H_0 is true. In other words,

$$Z = \frac{W - 0}{\sqrt{n(n+1)(2n+1)/6}} \approx \text{N}(0, 1)$$

- $\Pr(W \geqslant c \mid H_0) \approx \Pr(Z \geqslant z \mid H_0)$, which allows us to determine c.
- In this case, for $n = 10$ and $\alpha = 0.05$, we reject H_0 if

$$Z = \frac{W}{\sqrt{10 \cdot 11 \cdot 21/6}} \geqslant 1.645$$

(because $\Phi^{-1}(0.95) = 1.645$) which is equivalent to

$$W \geqslant 1.645 \times \sqrt{\frac{10 \cdot 11 \cdot 21}{6}} = 32.27$$

- For the example data we have $w = 25$, so we do not reject H_0

Using R

- R uses V rather than W
- For small sample sizes R will use the exact sampling distribution (which we haven't explored) rather than the normal approximation.
- To carry out the test, use: `wilcox.test`
- To work with the sampling distribution of V, use: `psignrank`
- Note: $\mathbb{E}(V) = n(n+1)/4$ and $\text{var}(V) = n(n+1)(2n+1)/24$. You can derive these in a similar way to W.

```
> wilcox.test(x, mu = 3.7, alternative = "greater",
              exact = TRUE)

Wilcoxon signed rank test

data:  x
V = 40, p-value = 0.1162
alternative hypothesis: true location is greater than 3.7

# Calculate exact p-value manually.
> 1 - psignrank(39, 10)
[1] 0.1162109

# Calculate approximate p-value, based on W.
> z <- 25 / sqrt(10 * 11 * 21 / 6)
> 1 - pnorm(z)
[1] 0.1013108
```

⇒ Close agreement between exact and approximate p-values

Paired samples

- Like other tests, we can use the Wilxcon signed-rank test for paired samples by first taking differences and treating these as a sample from a single distribution.
- The assumption of symmetry is quite reasonable in this setting, since under H_0 we would typically assume X and Y have the same distribution and therefore $X - Y \sim Y - X$.
- Indeed, this test is most often used in such a setting, due to the plausibility of this assumption.

Tied ranks

- We assumed a continuous population distribution
- Thus, all observations will differ (with probablity 1)
- In practice, the data are reported to finite precision (e.g. due to rounding), so we could have exactly equal values
- This will lead to ties when ranking our data
- If this happens, the 'rank' assigned for the tied values should be equal to the average of the ranks they span
- Example:

Value:	2.1	4.3	4.3	5.2	5.7	5.7	5.7	5.9
Rank:	1	2.5	2.5	4	6	6	6	8

- The presence of ties complicates the derivation of the sampling distribution, but R knows how to do the right thing

2.3 Wilcoxon rank-sum test (two-sample)

Wilcoxon two-sample test

- We can create a two-sample version of the Wilcoxon test.
- Independent random samples $X_1, \ldots, X_{n_X}$ and $Y_1, \ldots, Y_{n_Y}$ from two different populations with medians m_X and m_Y respectively.
- Want to test H_0: $m_X = m_Y$ against a one-sided or two-sided alternative
- Order the **combined** sample and let W be the sum of the ranks of $Y_1, \ldots, Y_{n_Y}$. This is the *Wilcoxon rank-sum statistic.*
- Note: this captures information on X as well as Y! (Why?)
- The test based on this statistic is called the *Wilcoxon rank-sum test*, also known as the *Wilcoxon two-sample test* and the *Mann-Whitney U test.*

Rejection region

- Suppose our alternative hypothesis is $H_1 : m_X > m_Y$
- If $m_X > m_Y$ then we expect W to be small, since the Y values will tend to be smaller than X and thus have smaller ranks
- Therefore, the critical region should be of the form $W \leqslant c$ for a suitable c.
- Properties of W (derivation not shown):

$$\mathbb{E}(W) = \frac{n_Y(n_X + n_Y + 1)}{2}$$

$$\text{var}(W) = \frac{n_X n_Y (n_X + n_Y + 1)}{12}$$

- W is approximately normally distributed when n_X and n_Y are large

Alternative definitions

- Like for the one-sample version, the definition of the statistic varies
- We just defined: W is the sum of the ranks in the Y sample
- A popular alternative: U is the number of all pairs (X_i, Y_j) such that $Y_j \leqslant X_i$ (the number of 'wins' out of all possible pairwise 'contests')
- U and W are deterministically related (can you derive the formula?)
- U and W have different (but related) sampling distributions
- Using either statistic leads to equivalent test procedures
- Note: $\mathbb{E}(U) = n_X n_Y/2$ and $\text{var}(U) = \text{var}(W)$

Example (Wilcoxon two-sample test)

Two companies package cinnamon. Samples of size eight from each company yield the following weights:

X	117.1	121.3	127.8	121.9	117.4	124.5	119.5	115.1
Y	123.5	125.3	126.5	127.9	122.1	125.6	129.8	117.2

Want to test H_0: $m_X = m_Y$ versus H_1: $m_X \neq m_Y$

Use a significance level of 5%

Using R

- R uses U... but calls it W!
- For small sample sizes R will use the exact sampling distribution, otherwise it will use a normal approximation
- To carry out the test, use: `wilcox.test`
- To work with the sampling distribution of U, use: `pwilcox`

```
> wilcox.test(x, y)

    Wilcoxon rank sum test

data:  x and y
W = 13, p-value = 0.04988
alternative hypothesis:
    true location shift is not equal to 0

# Calculate exact p-value manually.
> 2 * pwilcox(13, 8, 8)
[1] 0.04988345
```

We reject H_0 and conclude that we have sufficient evidence to show that the median weights differ between the two companies.

3 Goodness-of-fit tests (χ^2)

3.1 Introduction

Goodness-of-fit tests

- How well does a given model fit a set of data?
- E.g. if we assume a Poisson model for a set of data, is it reasonable?
- We can assess this with a 'goodness-of-fit' test
- The most commonly used is *Pearson's chi-squared test*

- Unlike most of the other tests we've seen, this operates on categorical (discrete) data
- Can also apply it on continuous data by first partitioning the data into separate classes

3.2 Two classes

Binomial model

- Start with a binomial model $Y_1 \sim \text{Bi}(n, p_1)$
- Our usual test statistic for this is

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \approx \text{N}(0,1)$$

- Therefore,

$$Q_1 = Z^2 \approx \chi^2_1$$

- To test $H_0 : p = p_1$ versus $H_1 : p \neq p_1$, we would reject H_0 if $|Z|$ (and, hence, Q_1) is too large.
- Next, notice that

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}$$

- and

$$(Y_1 - np_1)^2 = (n - Y_1 - n(1-p_1))^2 = (Y_2 - np_2)^2$$

where $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$.

- Therefore,

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

- Y_1 is the observed number of successes, np_1 is the expected number of successes
- Y_2 is the observed number of failures, np_2 is the expected number of failures
- So

$$Q_1 = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_1$$

where O_i is the observed number and E_i is the expected number.

- Even though there are two classes, we have only **one** degree of freedom. This is due to the constraint $Y_1 + Y_2 = n$.

3.3 More than two classes

Multinomial model

- Generalize to k possible outcomes (a multinomial model)
- p_i = probability of the ith class ($\sum_{i=1}^{k} p_i = 1$)
- Suppose we have n trials, with Y_i being the number of outcomes in class i
- $\mathbb{E}(Y_i) = np_i$
- Now we get,

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{k-1}$$

- $k-1$ degrees of freedom because $Y_1 + \cdots + Y_k = n$

Setting up the test

- Specify a categorical distribution: $p_1, p_2, \ldots, p_k$
- We use the Q_{k-1} statistic to test whether are data are consistent with this distribution
- The null hypothesis is that they do (i.e. the p_i define the distribution)
- The alternative is that they do not (i.e. a different set of probabilities define the distribution)
- Under the null, the test statistic will tend to be small (it measures 'badness-of-fit')
- Therefore, reject the null if $Q_{k-1} > c$ where c is the $1-\alpha$ quantile from χ^2_{k-1}.

Remarks

- We are approximating a binomial with a normal
- Good approximation if n is large and the p_i are not too small
- Rule of thumb: need to have all $E_i = np_i \geqslant 5$
- The larger the k (i.e. more classes), the more powerful the test. However, we need the classes to be large enough
- If any of the E_i are too small, can combine some of the classes until they are large enough
- If Q_{k-1} is very small, this indicates that the fit is 'too good'. This can be used as a test for rigging of experiments / fake data. Typically need very large n to do this.
- Often refer to the test statistic as χ^2

Example (completely specified distribution)

- Proportions of commuters using various modes of transport, based on past records:

Bus	Train	Car	Other
0.25	0.15	0.50	0.1

- After a 3-month campaign, a random sample ($n = 80$) found:

Bus	Train	Car	Other
26	15	32	7

- Did the campaign alter commuters behaviour?
- The expected frequencies are:

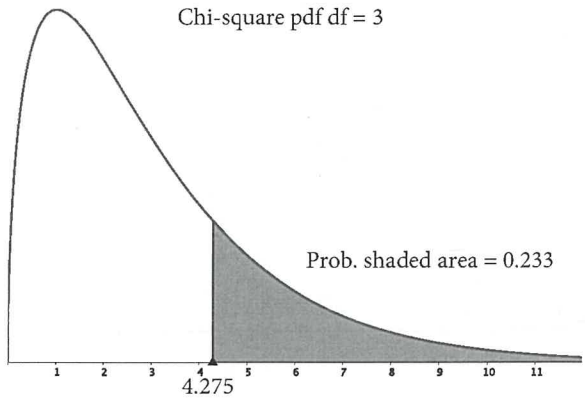
Bus	Train	Car	Other
20	12	40	8

- The value of the test statistic is:

$$\chi^2 = \frac{(26-20)^2}{20} + \frac{(15-12)^2}{12} + \frac{(32-40)^2}{40} + \frac{(7-8)^2}{8} = 4.275$$

- H_0: proportions have not changed, H_1: proportions have changed
- We have 4 classes, so the test statistic here has a χ^2_3 distribution.
- The 0.95 quantile is 7.81, which is greater than $\chi^2 = 4.275$
- Therefore, there is **insufficient** evidence that the proportions have changed
- The p-value is

$$p = \Pr(\chi^2_3 > 4.275) = 0.233 > 0.05$$



Using R

```
> x <- c(  26,   15,   32,   7)
> p <- c(0.25, 0.15, 0.5, 0.1)
> t1 <- chisq.test(x, p = p)
> t1
```

counts probability

```
Chi-squared test for given probabilities

data:  x
X-squared = 4.275, df = 3, p-value = 0.2333
```

```
> rbind(t1$observed, t1$expected)
     [,1] [,2] [,3] [,4]
[1,]   26   15   32    7
[2,]   20   12   40    8

> t1$residuals
[1]  1.3416408  0.8660254 -1.2649111 -0.3535534

> sum(t1$residuals^2)
[1] 4.275

> 1 - pchisq(4.275, 3)
[1] 0.2332594
```

3.4 Estimating parameters

Fitting distributions

- We don't always have an exact model to compare against
- We might specify a family of distributions but still need to estimate some of the parameters
- For example, $\text{Pn}(\lambda)$ or $\text{N}(\mu, \sigma^2)$
- We would need to estimate the parameters using the sample, and use these to specify H_0
- We need to adjust the test to take into account that we've used the data to define H_0 (by design, it will be 'closer' to the data than if it we didn't need to do this)
- The 'cost' of this estimation is 1 degree of freedom for each parameter that is estimated
- The final degrees of freedom is $k - p - 1$, where p is the number of estimated parameters

Example (Poisson distribution)

- X is number of alpha particles emitted in 0.1 sec by a radioactive source
- Fifty observations:

7, 4, 3, 6, 4, 4, 5, 3, 5, 3, 5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3, 9, 11, 6, 7, 4, 5, 4, 7, 3, 2, 8, 6, 7, 4, 1, 9, 8, 4, 8, 9, 3, 9, 7, 7, 9, 3, 10

- Is a Poisson distribution an adequate model for the data?
- H_0: Poisson, H_1: something else
- We have only specified the family of the distribution, not the parameters
- Estimate the Poisson rate parameter λ by the MLE, $\hat{\lambda} = \bar{x} = 5.4$
- Now we ask: does the Pn(5.4) model give a good fit?

First, find an appropriate partition of the value (collapse the data):

```
> X1 <- cut(X, breaks = c(0, 3.5, 4.5, 5.5, 6.5, 7.5, Inf))
> T1 <- table(X1)
> T1
X1
  (0,3.5] (3.5,4.5] (4.5,5.5] (5.5,6.5] (6.5,7.5] (7.5,Inf]
       13         9         6         5         7        10
```

Then, prepare the data for the test:

```
> x <- as.numeric(T1)
> x
[1] 13  9  6  5  7 10

> n <- sum(x)
> p1 <- sum(dpois(0:3, 5.4));
> p2 <- dpois(4, 5.4)
> p3 <- dpois(5, 5.4)
> p4 <- dpois(6, 5.4)
> p5 <- dpois(7, 5.4)
> p6 <- 1 - (p1 + p2 + p3 + p4 + p5)
> p <- c(p1, p2, p3, p4, p5, p6)
```

Then, run the test:

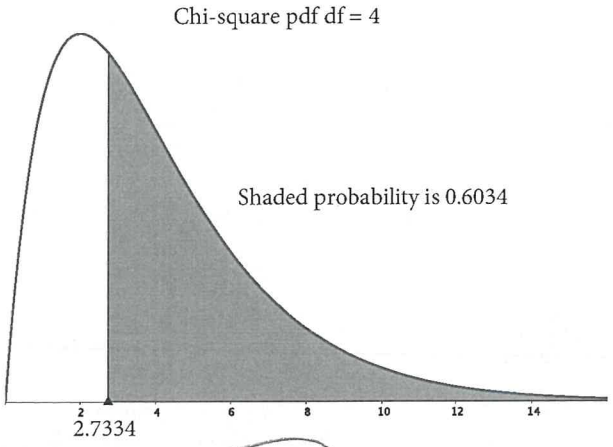
```
> chisq.test(x, p = p)

        Chi-squared test for given probabilities

data:  x
X-squared = 2.7334, df = 5, p-value = 0.741
```

But this is the wrong df! Need to adjust manually:

```
> 1 - pchisq(2.7334, 4)
[1] 0.6033828
```



- Needed to adjust p-values as we have estimated the mean
- The critical value is the 0.95 quantile from χ^2_4, which is 9.488, so we cannot reject H_0
- Not enough evidence against the Poisson model
- Therefore, this is an adequate fit (at least, until further data proves otherwise)

	0 3	4	5	6	7	8+
Observed	13.0	9.0	6.0	5.0	7.0	10.0
Expected	10.7	8.0	8.6	7.8	6.0	8.9

4 Tests of independence (contingency tables)

Contingency tables

- Suppose we have multiple categorical variables (which could be continuous variables partitioned into classes)
- A *contingency table* records the number of observations for each possible cross-classification of these variables
- We are often interested in whether two categorical variables are related to each other
- For example, height and weight
- Define height classes $A_1, \ldots, A_r$, and weight classes $B_1, \ldots, B_c$
- Each person is assigned to a single combination (A_i, B_j)
- A sample of people can be summarised with a $r \times c$ table of counts (a contingency table)

Independence model

- A general model for these data is:

$$p_{ij} = \Pr(A_i \cap B_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, c$$

- Are the two variables independent?
- We can set this up as a hypothesis test:

$$H_0 : p_{ij} = \Pr(A_i)\Pr(B_j) \quad \text{versus} \quad H_1 : p_{ij} \neq \Pr(A_i)\Pr(B_j)$$

- This has the same structure as a goodness-of-fit test, can use Pearson's chi-squared statistic
- Show how this works through an example...

Example (contingency table)

150 executives were classified by sex, A, and whether or not they were firstborn, B:

	Firstborn	Not firstborn	Total
Male	34	74	108
Female	20	22	42
Total	54	96	150

Let's test whether these two variables are independent.

Estimating the marginals

- Recall discrete bivariate distributions:

	Firstborn	Not firstborn	Total
Male	p_{11}	p_{12}	$p_{1\cdot}$
Female	p_{21}	p_{22}	$p_{2\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	1

- The marginals are:

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = \Pr(A_i)$$

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = \Pr(B_j)$$

- The null hypothesis of independence is just, $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$
- Data:

	Firstborn	Not firstborn	Total
Male	y_{11}	y_{12}	$y_{1\cdot}$
Female	y_{21}	y_{22}	$y_{2\cdot}$
Total	$y_{\cdot 1}$	$y_{\cdot 2}$	n

- Estimates:

$$\hat{p}_{i\cdot} = \frac{y_{i\cdot}}{n}$$

$$\hat{p}_{\cdot j} = \frac{y_{\cdot j}}{n}$$

where

$$y_{i\cdot} = \sum_{j=1}^{c} y_{ij}$$

$$y_{\cdot j} = \sum_{i=1}^{r} y_{ij}$$

- Pearson's χ^2 statistic for given p_{ij} is

$$Q = \sum_i \sum_j \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

- Under H_0, an estimator of p_{ij} is

$$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{Y_{i\cdot}Y_{\cdot j}}{n^2}$$

- This gives the following,

$$Q = \sum_i \sum_j \frac{(Y_{ij} - Y_{i\cdot}Y_{\cdot j}/n)^2}{Y_{i\cdot}Y_{\cdot j}/n} \approx \chi^2_{(r-1)(c-1)}$$

Explanation for degrees of freedom

- Recall that we should have $k - p - 1$ degrees of freedom
- Here, $k = rc$, the total number of cells in the table
- We estimated $r - 1$ marginal probabilities for the rows and $c - 1$ for the columns, which makes $p = (r-1)+(c-1)$
- Therefore, the number of degrees of freedom remaining is:

$$\text{df} = rc - (r-1) - (c-1) - 1 = (r-1)(c-1)$$

Using R: set up the data

```
> x <- rbind( male = c(first = 34, later = 74),
+             female = c(first = 20, later = 22))
> x
       first later
male      34    74
female    20    22
```

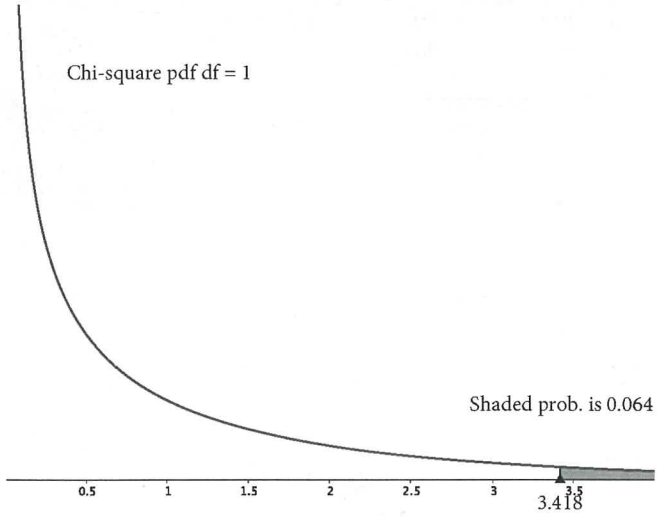
Using R: run the test

```
> c1 <- chisq.test(x, correct = FALSE)
> c1

        Pearson's Chi-squared test

data:  x
X-squared = 3.418, df = 1, p-value = 0.06449
```

We do not have enough evidence to reject H_0 at a 5% significance level.



Using R: more output

```
> c1$observed
       first later
male      34    74
female    20    22

> c1$expected
       first later
male   38.88 69.12
female 15.12 26.88
```