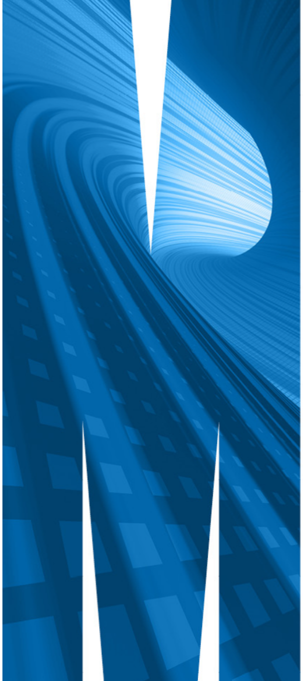
Hypothesis testing



Statistical Thinking (ETC2420 / ETC5242)

Week 4, Semester 2, 2025

Outline

1. Overview
2. Hypothesis testing: key concepts
3. Common simple scenarios
4. Using simulations
5. Usage & (mis)interpretation
6. Wrap-up

Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

Learning goals for Week 4

- Introduce the concepts behind statistical hypothesis testing.
- Carry out parametric tests for one and two proportions using the Central Limit Theorem.
- Introduce the idea of a permutation test.
- Construct a permutation test for independence of two binary variables.
- Explain the disadvantages and appropriate use of hypothesis testing.

Warning

- This week we will learn about hypothesis testing.
- This is an approach to inference that dominates much of statistical practice, especially by non-statisticians.
- It is largely considered **NOT best practice** by professional statisticians
- Better procedures usually exist, and we have already learnt some!
- But we will cover hypothesis testing anyway because:
    - It is ubiquitous
    - Need to understand its weaknesses
    - Sometimes it is useful, or at least convenient.

Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

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 would you know if it is better?
- How would you decide whether to switch or keep the old one?

Factory example

- Run an experiment: make $n = 200$ devices with the new process
- Outcome: let Y be the number of faulty devices (out of 200)
- You decide that if $Y \leqslant 7$ then you will switch to the new process
- (Note: $Y \leqslant 7 \iff Y/n \leqslant 0.035 = 3.5\%$)
- Is this a sensible procedure?
- We can formulate this as a **statistical hypothesis test**.

Research questions as hypotheses

- Research questions/studies are often often framed in terms of hypotheses
- Run an experiment or collect data and then ask:
  - Do the data support/contradict the hypothesis?
- Can we frame statistical inference around this paradigm?

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.

Role of 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.
- In other words, we assume the null unless proven otherwise. Similar to ‘innocent until proven guilty’.
- Note: what is taken as the null hypothesis (the actual meaning of ‘no change’) will depend on the context of the study and where the onus of proof is deemed to lie.

Factory example (cont'd)

For our factory example:

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

$$H_0 \colon p = 0.06$$

- Alternative hypothesis:

$$H_1 \colon p < 0.06$$

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 typically simple
- Alternative hypotheses are typically composite.

Specification of hypotheses

- Usually, the null hypothesis is on the boundary of the alternative hypothesis (here, $p = 0.06$ versus $p < 0.06$).
- For single parameters, the null hypothesis is typically of the form $\theta = \theta_0$ and the alternative hypothesis is one of the following:
  - **one-sided** and takes the form $\theta < \theta_0$ or $\theta > \theta_0$
  - **two-sided** and written as $\theta \neq \theta_0$.

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:

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

- The set A is typically an interval.
- For example, if $A = [3, \infty)$, the decision rule is simply: reject H_0 if $T \geqslant 3$.

Describing tests

- Decision rule:

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

- The set A is called the **critical region**, or sometimes the **rejection region**. If it is an interval, the boundary value is called the **critical value**.
- For our factory example:
  - The test statistic is Y.
  - The decision rule is to reject H_0 if $Y \leqslant 7$.
  - The critical region is $(-\infty, 7]$.
  - 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'.
Instead, 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 otherwise).

Errors in test outcomes

Test outcomes are not guaranteed to be accurate.

What could go wrong with our decision rule for the factory example?

Reminder of the details:

- Null hypothesis (H_0): the new manufacturing process produces faulty products as often as the old process.
- Experiment: make 200 devices with the new manufacturing process.
- Test statistic: Y is the number of faulty devices (out of 200).
- Decision rule: reject H_0 if $Y \leqslant 7$.

Type I error

- The new manufacturing process might produce faulty devices at the same rate as the old process (H_0 is true), but by chance we observe only 7 or fewer failures (reject H_0).
- 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 a **Type I error**, also known as a **false positive**.
- 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 (H_0 is false), but by chance we observe more than 7 failures (fail to reject H_0).
- 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 a **Type II error**, also known as a **false negative**.
- In this case, the error would be less costly in the short term but might be much more costly long-term.
- (While 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

How often do errors occur?

Need to use the **sampling distribution** of the test statistic.

Typically need to make some assumptions, such as assuming the null.

Significance level

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

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

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

- Calculate in R using:

```r
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 $Y \sim \mathrm{Bi}(200, 0.03)$, 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:

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

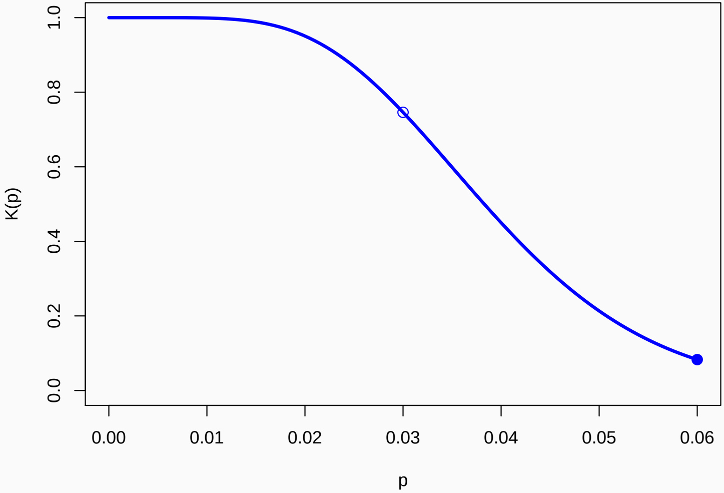
Typically, we would present this as a function of the true parameter value.

For example,

$$K(p') = \Pr(\text{reject } H_0 \mid p = p') = \Pr(Y \leqslant 7 \mid p = p')$$

Example

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



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 type I error

- Typically, we construct a test so that it has a **specified** significance level, α.
- In other words, we set the probability of a type I error to be some value (we ‘control’ it).
- A widespread convention is to set $\alpha = 0.05$
- That is, our test will incorrectly reject the null hypothesis about 1 time in 20.

Maximising power

- While maintaining control of type I error, we try to maximise power (minimise the probability of a type II error) as much as we can.
- Since $K(p_0) = \alpha$, how can we increase power while α is fixed?
- Some possibilities:
  - Choosing good/optimal test statistics
  - Increasing the sample size.

Alternative formulation of a test: via 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.)
- Decision rule: reject H_0 if the p-value is less than the significance level.

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.

P-value nomenclature

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:

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

Factory example (cont'd)

- 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 p-value…
- The p-value is a binomial probability, $\Pr(Y \leqslant 6 \mid p = p_0)$.

```r
pbinom(6, 200, 0.06)
```

```
[1] 0.0412542
```

- This is less than $\alpha = 0.0829$, so therefore reject H_0.

Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

Common scenarios: overview

Scenarios:

- Single mean (vs a null value)
- Two means (comparison)
- Single proportion (vs a null value)
- Two proportions (comparison)

Using the following functions in R:

- `t.test()`
- `prop.test()`

Example (single mean)

- A tyre manufacturer claims that a new tyre will last 60,000 km on average.
- A consumer group tests a sample of 50 tyres and finds the mean is 55,212 km, with a sample standard deviation of 5082 km.
- Is this strong evidence against the manufacturer's claim?
- Let μ be the mean tyre lifetime.

$$H_0 : \mu = 60,000 \quad \text{versus} \quad H_1 : \mu < 60,000$$

- Note the choice of null: we seek strong evidence against the manufacturer to query the claims.

The t-test

- Recall if the sample is from $\mathrm{N}(\mu, \sigma^2)$ then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}$$

- We can use T as the test statistic
- Doing so gives us us the **t-test**
- If the sample is not normally distributed, the CLT still allows us to use the t-test as a good approximation.

Example (single mean, continued)

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

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < c = t_{n-1,\alpha}$$

- Let's use $\alpha = 0.05$. We have $n = 50$, so the critical value is:

```r
qt(0.05, 50 - 1)
```

```
[1] -1.67655
```

- From the data: $\bar{x} = 55212$, $s = 5082$. Also, $\mu_0 = 60000$. The test statistic is:

$$t = \frac{55212 - 60000}{5082/\sqrt{50}} = -6.66$$

- Since $t < c$, we reject H_0 at the 5% level of significance.
- We have enough evidence to suspect that the average tyre lifetime is lower than the claimed 60,000 km.

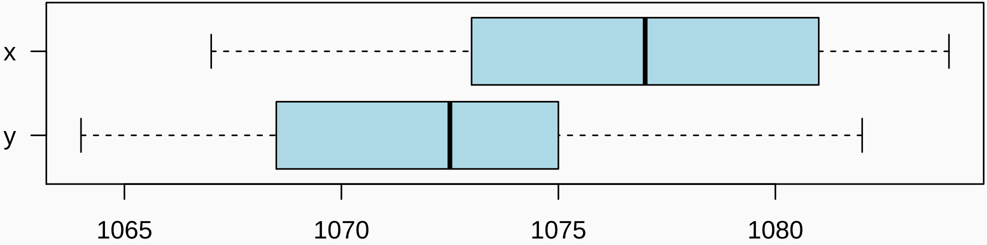
Example (two means)

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:

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

Example (two means)



```
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)
```

Example (two means)

We use the Welch approximation,

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \overset{\text{approx}}{\sim} \mathrm{t}_k$$

where k is given by a specific formula (not shown here).

Reject H_0 if $|T| > c = t_{k,1-\alpha/2}$

Example (two means)

```r
t.test(x, y)
```

```

    Welch Two Sample t-test

data:  x and y
t = 2.053, df = 21.93, p-value = 0.0522
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0456696  8.8790029
sample estimates:
mean of x mean of y 
  1076.75   1072.33 
```

Example (two means)

The p-value is 0.052.

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

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).

Single proportion

- Observe n Bernoulli trials with unknown probability p
- Summarise by $Y \sim \mathrm{Bi}(n, p)$
- 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 $\Pr(Y \geqslant c \mid p = p_0) = \alpha$

Using a normal approximation to determine a critical value:

- For large n, when H_0 is true

$$Y \approx \mathrm{N}(np_0, np_0(1-p_0))$$

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

- This implies,

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

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution.

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 $z_{1-\alpha}$ instead of c
- Express Z as the standardised proportion of 6's,

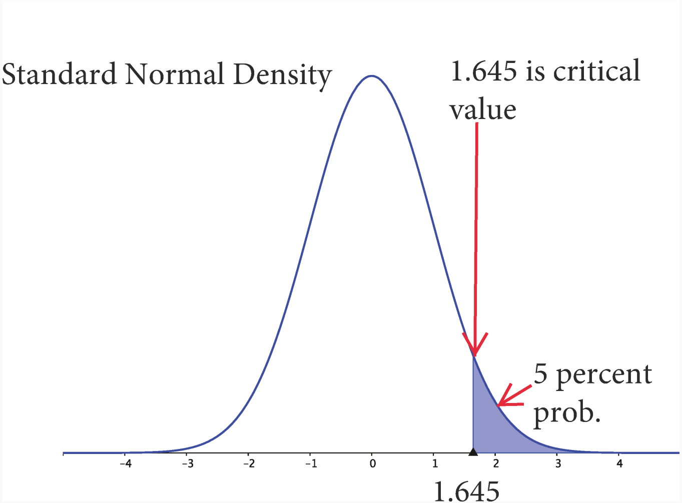
$$Z = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} \approx \mathrm{N}(0,1)$$

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

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

and since $z > z_{0.95} = 1.645$ we reject H_0.

Single proportion, cont'd



Single proportion, cont'd

- 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 $|Z| > z_{1-\alpha/2}$
- In the previous example, we would use $z_{1-\alpha/2} = 1.96$. Since $z = 1.67$, we would **not** reject H_0.

Example (single proportion, continued)

```r
prop.test(x = 1389, n = 8000, p = 1/6, alternative = "greater")
```

```

    1-sample proportions test with continuity correction

data:  1389 out of 8000, null probability 1/6
X-squared = 2.739, df = 1, p-value = 0.049
alternative hypothesis: true p is greater than 0.166667
95 percent confidence interval:
 0.166708 1.000000
sample estimates:
       p 
0.173625 
```

Example (single proportion, continued)

```r
prop.test(x = 1389, n = 8000, p = 1/6, alternative = "two.sided")
```

```

    1-sample proportions test with continuity correction

data:  1389 out of 8000, null probability 1/6
X-squared = 2.739, df = 1, p-value = 0.0979
alternative hypothesis: true p is not equal to 0.166667
95 percent confidence interval:
 0.165420 0.182145
sample estimates:
       p
0.173625
```

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.

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

```r
x <- c(459, 425)
n <- c(500, 500)
```

Example (two proportions)

```r
prop.test(x, n, alternative = "greater")
```

```
    2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 10.62, df = 1, p-value = 0.000559
alternative hypothesis: greater
95 percent confidence interval:
 0.0328754 1.0000000
sample estimates:
prop 1 prop 2 
 0.918  0.850 
```

Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

Going beyond simple scenarios

- What if we don’t know the sampling distribution of the test statistic?
- And we can’t or don’t wish to rely on the CLT
- We can try to simulate the sampling distribution
- Discuss two approaches today:
    - Simulate from an assumed model
    - Simulate by permuting the data

Example (exponential distribution)

- Observe the times (in seconds) between 100 trams at a tram stop.
- Is the population average time more than one minute?

$$H_0\colon \mu = 60 \quad \text{versus} \quad H_1\colon \mu > 60$$

- Assume an exponential distribution for the data

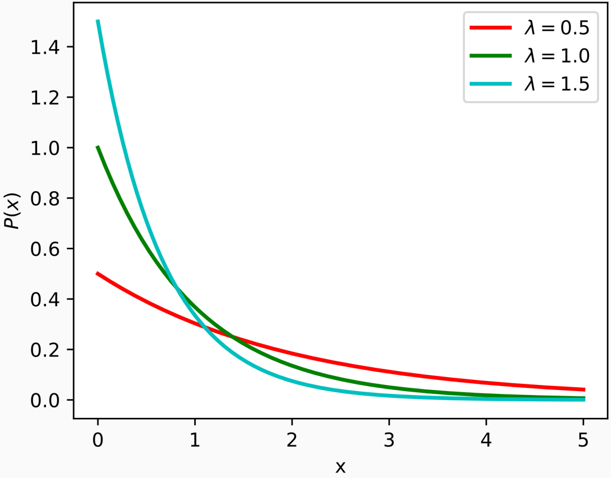
$$X_1, X_2, \dots, X_{100} \sim \mathrm{Exp}(\lambda)$$

- Use test statistic:

$$T = \bar{X}$$

- The observations do not follow a normal distribution...what about *T*?

Exponential distribution



Not symmetric

Not bell-shaped

Simulate the null distribution

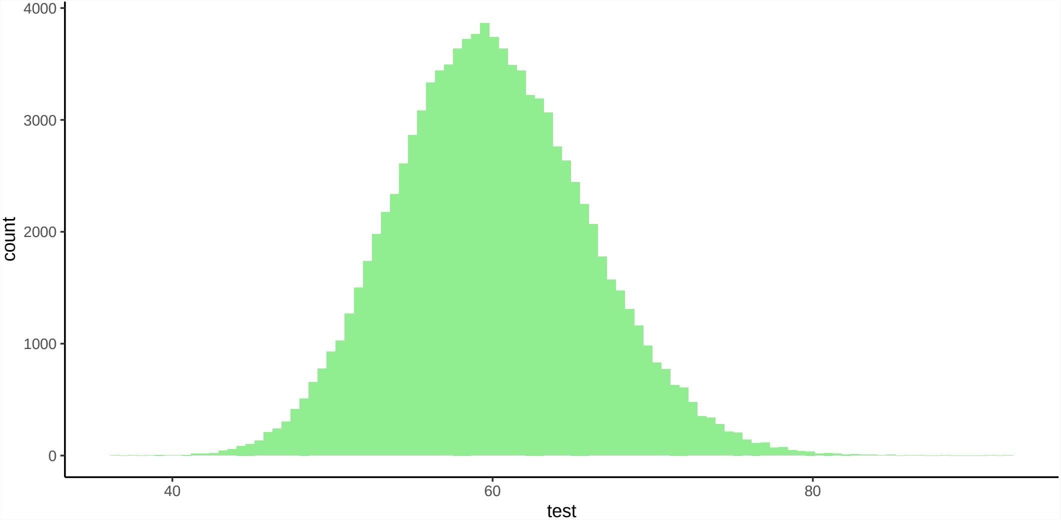
- Want the sampling distribution of T assuming H_0
- That is, when we have a random sample from $\mathrm{Exp}(1/60)$
- We can simulate this, as follows:

1. Sample 100 iid exponentials with mean 60:

   ```
   y <- rexp(100, 1/60)
   ```

2. Compute `mean(y)`
3. Repeat this a large number of times.

A plot of simulated repeated samples



Using the simulations

- We can use the simulated sampling distribution to carry out the test.

$$H_0\colon \mu = 60 \quad \text{versus} \quad H_1\colon \mu > 60$$

- For our observed data we have $t = 71$ (average time between trams was 71 seconds).
- Is this surprisingly large if H_0 were true?
- Can calculate a p-value as the proportion of simulated values (of T) that exceed t.
- In this case, it is about 0.04.

A useful general technique

- Note that the sampling distribution of T looks close to a normal.
- Indeed, this is the CLT in action!
- However, this simulation technique works even when the CLT does not.
- For example, when using statistics other than the sample mean.

Example (gender discrimination)

Do male and female employees have the same chance of being promoted?

- Population proportion of **males** promoted: p_M
- Population proportion of **females** promoted: p_F

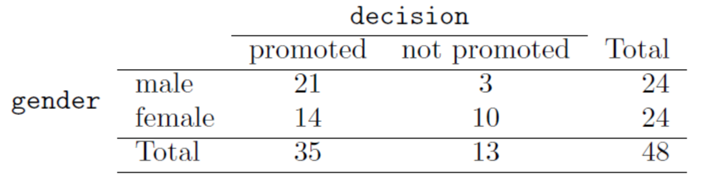
- Test the following:

$$H_0 : p_M = p_F \quad \text{versus} \quad H_1 : p_M > p_F$$

- The null is that gender has no effect on promotion decisions, with the alternative that men are more likely to be promoted.
- Run a controlled experiment: all other attributes of employees are identical.

Example (gender discrimination)

- Outcomes:



		decision		
		promoted	not promoted	Total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

- Point estimates:

$$\hat{p}_M = \frac{21}{24} \qquad \hat{p}_F = \frac{14}{24}$$

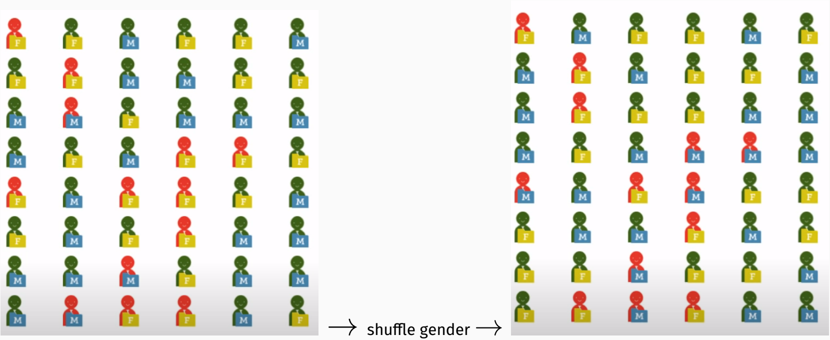
- We have $\hat{p}_M > \hat{p}_F$, but is it enough to reject H_0?
- Could use `prop.test()`, but let's see an alternative...

Permutation test

- If H_0 were true, the gender of each employee is irrelevant.
- If we ‘shuffle’ the genders around, then we shouldn’t be able to tell the difference.
- Use this idea for simulation!

Permutation test

- **Randomly permute** the gender variable
- This breaks any association (if present) between gender and promotion outcome



Permutation test p-value

- Let the test statistic be:

$$D = \hat{p}_M - \hat{p}_F$$

- From our data:

$$d_{\text{obs}} = \frac{21}{24} - \frac{14}{24} = 0.2917$$

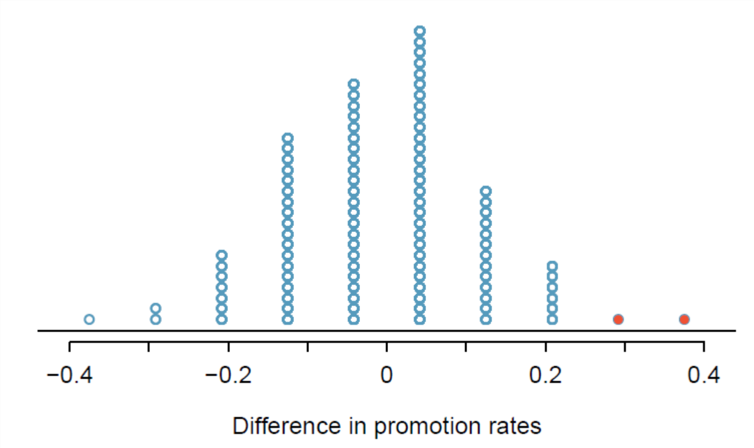
- Randomly permute gender R times (for $r = 1, 2, \ldots, R$), each time calculating:

$$d^{[r]} = \hat{p}_M^{[r]} - \hat{p}_F^{[r]}$$

- This approximates the sampling distribution of D assuming H_0.
- Use the simulated hypothetical sampling distribution to estimate the p-value:

$$\text{p-value} \approx \frac{\left(\text{number of } d^{[r]} \geqslant d_{\text{obs}}\right)}{R}$$

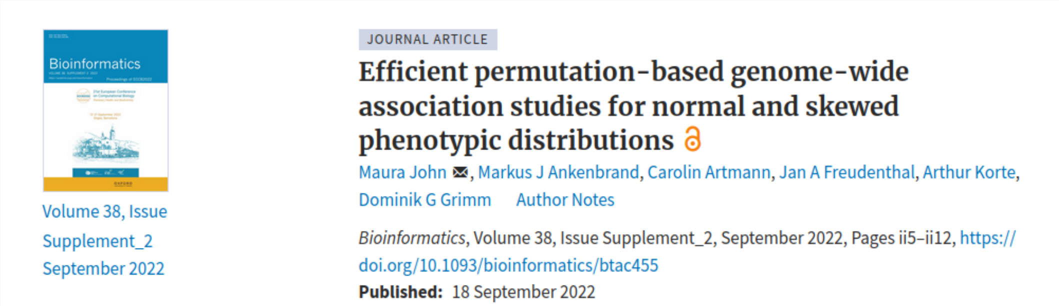
Plot of permutation test simulations



A useful general technique

- In this case, we could have just used `prop.test()`
- But the technique works in more general scenarios.
- For example...

Example (genome-wide association studies)



Volume 38, Issue Supplement_2 September 2022

JOURNAL ARTICLE

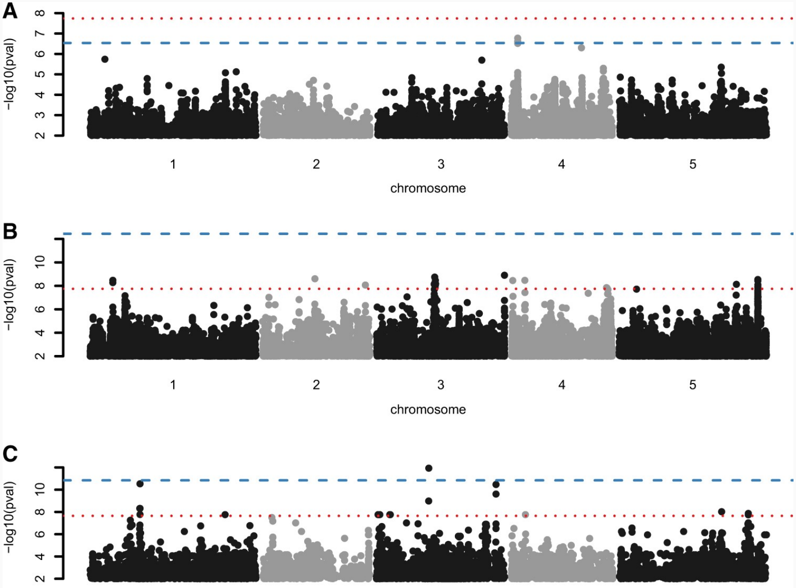
Efficient permutation-based genome-wide association studies for normal and skewed phenotypic distributions

Maura John ✉, Markus J Ankenbrand, Carolin Artmann, Jan A Freudenthal, Arthur Korte, Dominik G Grimm Author Notes

Bioinformatics, Volume 38, Issue Supplement_2, September 2022, Pages ii5–ii12, https://doi.org/10.1093/bioinformatics/btac455

Published: 18 September 2022

Example (genome-wide association studies)



Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

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.
- α = 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 study, then following this convention will usually be acceptable.

Choice of significance level

Specific fields of application can have their own conventions which are very different. For example:

- Genome-wide association studies require p-values under 10^{-8}.
- High-energy physics (particle physics) requires:
  - p-values under 0.003 (‘3 sigma’) for reporting ‘evidence of a particle’
  - p-values under 3×10^{-7} (‘5 sigma’) for reporting a ‘discovery’.

More info at: https://blogs.scientificamerican.com/observations/five-sigmawhats-that/

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 misinterpretations 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: https://dx.doi.org/10.1007/s10654-016-0149-3

‘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?
- That is, by setting up a testable hypothesis and then running an experiment to try to disprove it?
- Perhaps...but a binary decision doesn't carry much informative content
- This is a 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 unequivocal conclusions is a misconception about the nature of inference.
- But the genie is out of the bottle…

Why is hypothesis testing so popular?

- 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 stronger now.
- 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
- 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?

- 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 is helpful in designing studies, especially the concept of error probabilities (type I error, type II error, power)
- Sometimes we actually require decisions, e.g. quality control applications (such as our factory example)
- Hypothesis testing is fine 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’.

p-hacking

- An easy trap for the inexperienced: carrying out many tests, but only reporting those where you reject H_0.
- Referred to as **data dredging** or **p-hacking**.
- The reported results will typically not be ‘significant’.
- The actual type I error rate will be higher than the reported significance level, due to not correctly accounting for all of the tests that were carried out.
- Very easy to fall into this trap.
- Another example of why hypothesis testing is not considered best practice.

Outline

1 Overview

2 Hypothesis testing: key concepts

3 Common simple scenarios

4 Using simulations

5 Usage & (mis)interpretation

6 Wrap-up

Summary of approaches

Different ways we have worked with sampling distributions:

- Assumed model, mathematically derived sampling distribution
- Rely on CLT for approximate sampling distribution
- Simulate from an assumed model
- Simulate using permutations (permutation test)

How are we ‘thinking statistically’ today?

- Distinguishing between hypotheses (population) and test statistics (sample).
- Understanding the sampling distribution of test statistics.
- Exploiting mathematical theory to approximate sampling distributions (CLT).
- Exploiting the structure of the data to simulate sampling distributions (permutation test).
- Avoiding ‘black & white’ inferential conclusions when carrying out tests.