Hypothesis testing

(Module 6)

Statistics (MAST20005) &
Elements of Statistics
(MAST90058)

School of Mathematics and Statistics
University of Melbourne

Semester 2, 2022

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

Outline

Preface
A cautionary word
A motivating example
Classical hypothesis testing (Neyman-Pearson)
Hypotheses
Tests & statistics
Errors (Type I, Type II)
Significance level & power
Alternative formulations
Significance testing (Fisher)
Modern hypothesis testing
Common scenarios
Single proportion
Two proportions
Single mean
Single variance

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

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**

Outline

Preface
A cautionary word
A motivating example
Classical hypothesis testing (Neyman-Pearson)
Hypotheses
Tests & statistics
Errors (Type I, Type II)
Significance level & power
Alternative formulations
Significance testing (Fisher)
Modern hypothesis testing
Common scenarios
Single proportion
Two proportions
Single mean
Single variance

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

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: $Y \sim \mathrm{Bi}(200, p)$, 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

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 $\theta \neq \theta_0$.

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:

reject H_0 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 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)

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!

Significance level

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

- 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:

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

- 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!

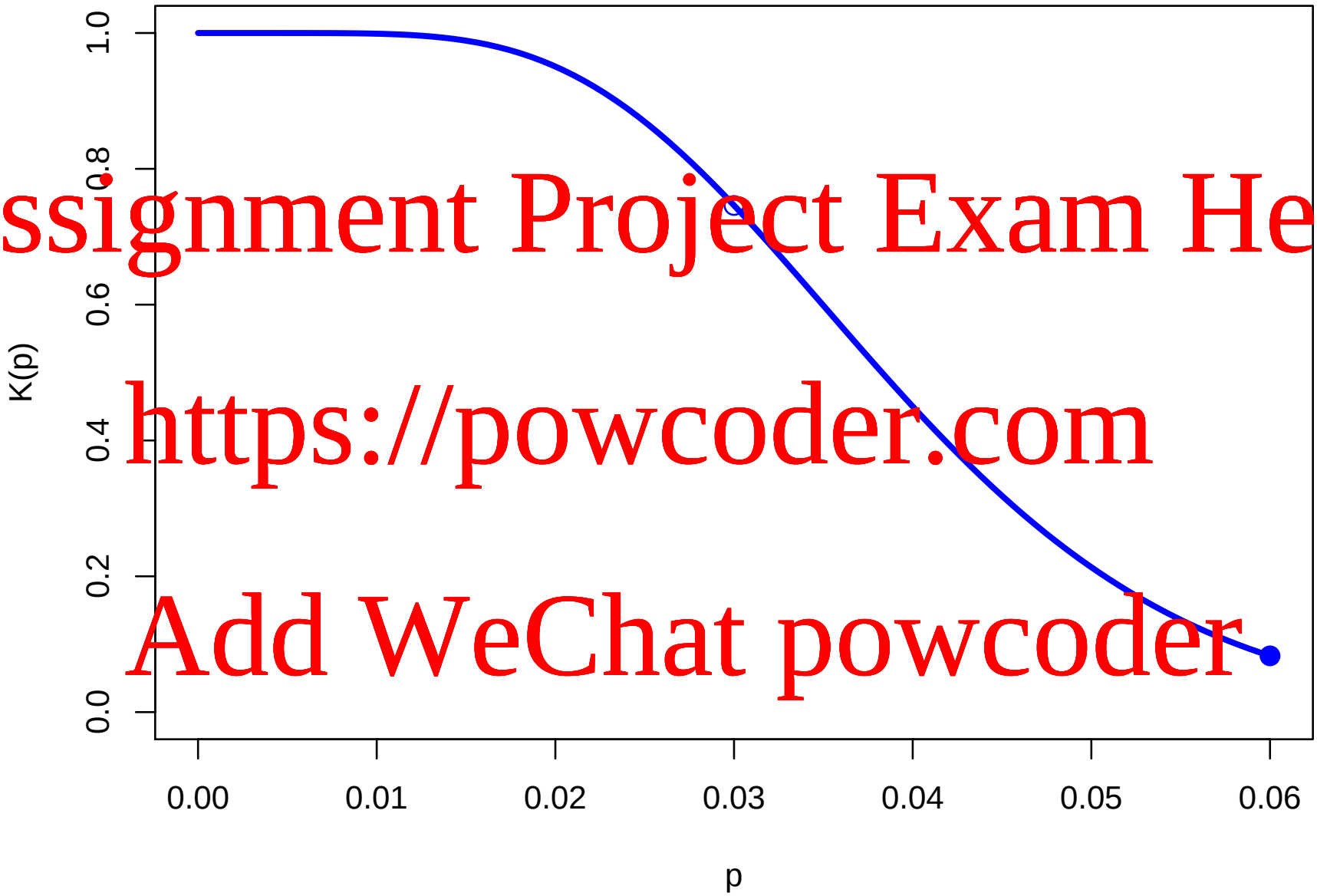
Power

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

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

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

For our 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 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

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.

Outline

Preface
A cautionary word
Something example
Classical hypothesis testing (Neyman-Pearson)
Hypotheses
Tests & statistics
Errors (Type I, Type II)
Significance level & power
Alternative formulations
Significance testing (Fisher)
Modern hypothesis testing
Common scenarios
Single proportion
Two proportions
Single mean
Single variance

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)

Outline

Preface
A cautionary word
A motivating example
Classical hypothesis testing (Neyman-Pearson)
Hypotheses
Test statistics
Errors (Type I, Type II)
Significance level & power
Alternative formulations
Significance testing (Fisher)
Modern hypothesis testing
Common scenarios
Single proportion
Two proportions
Single mean
Single variance

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...

Outline

Preface
- A cautionary word
- Sweetening example

Classical hypothesis testing (Neyman-Pearson)
- Hypotheses
- Tests & statistics
- Errors (Type I, Type II)
- Significance level & power
- Alternative formulations

Significance testing (Fisher)

Modern hypothesis testing

Common scenarios
- Single proportion
- Two proportions
- Single mean
- Single variance

Common scenarios: overview

Proportions:

- Single proportion
- Two proportions

Normal distribution:

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

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

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

- 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

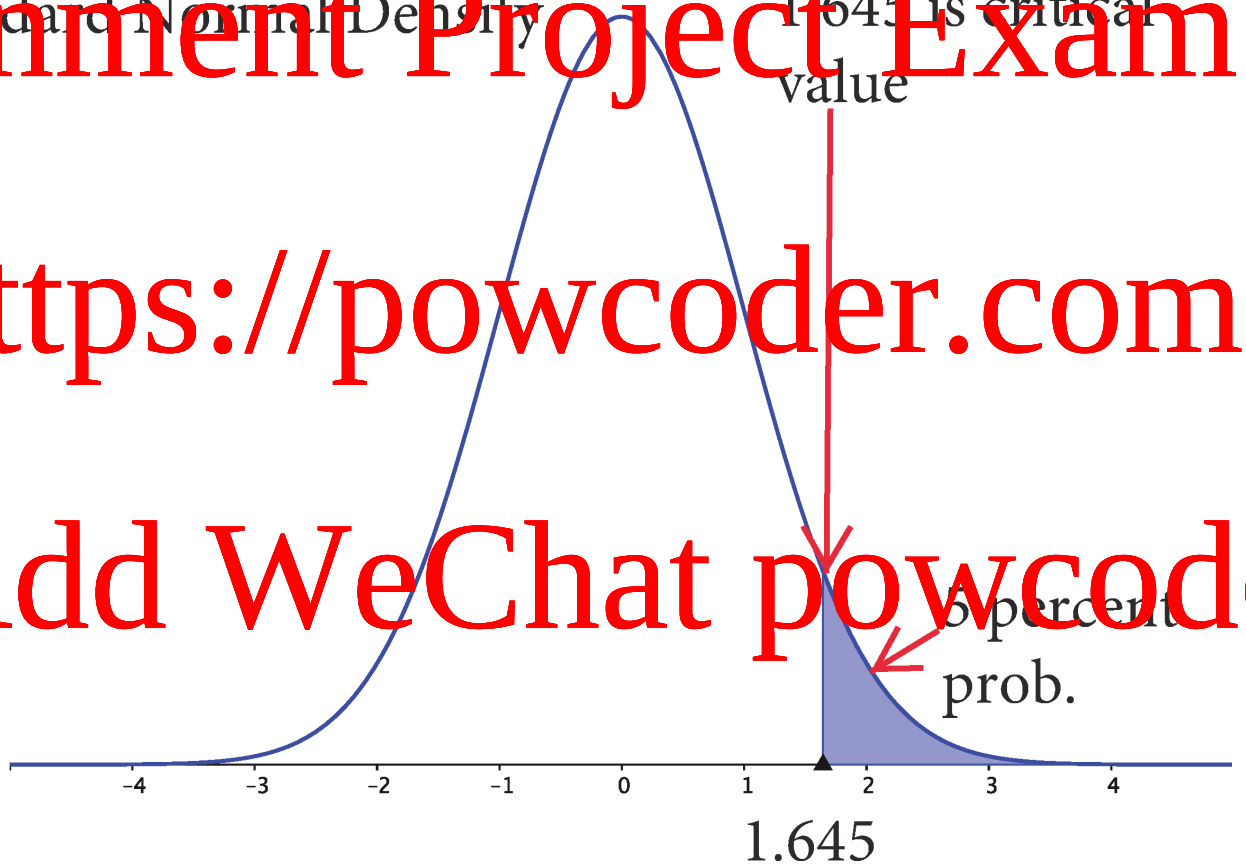
- It is more common to use standardised test statistics
- Here, report Z instead of Y and compare to $\Phi^{-1}(1-\alpha)$ instead of c
- Express Z as the standardised proportion of 6's,

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

- 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:

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

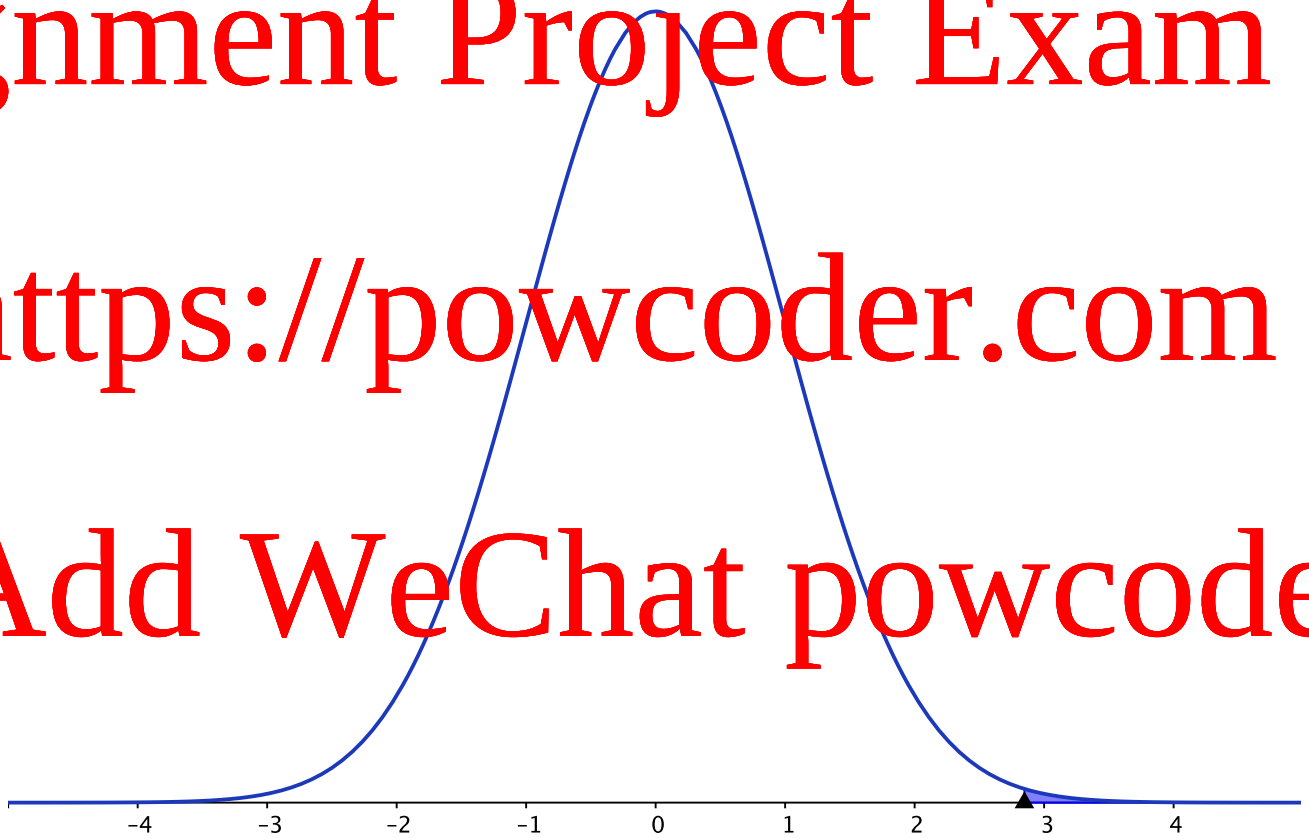
- Since $0.00226 < 0.05$, we reject H_0.

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
```

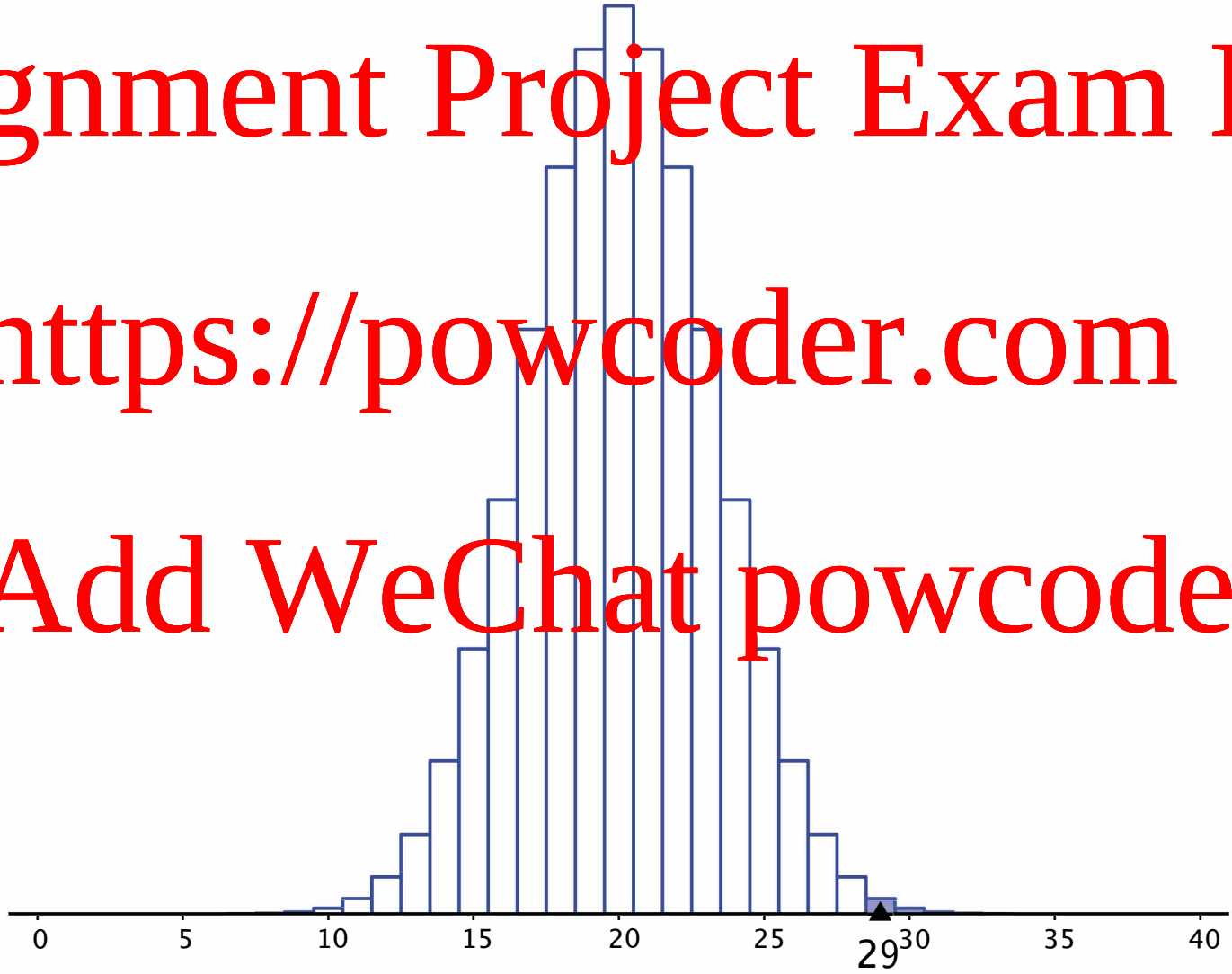
Z distributed Standard Normal



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
```

Y distributed Binomial n=40, p=0.5 pmf



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

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

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

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

- 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

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} > \Phi^{-1}(1-\alpha)$$

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
$H_0: p_1 = p_2$	$H_1: p_1 > p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} > \Phi^{-1}(1-\alpha)$
$H_0: p_1 = p_2$	$H_1: p_1 < p_2$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} < \Phi^{-1}(\alpha)$
$H_0: p_1 = p_2$	$H_1: p_1 \neq p_2$	$\lvert z \rvert = \frac{\lvert \hat{p}_1 - \hat{p}_2 \rvert}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} > \Phi^{-1}(1-\alpha/2)$

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

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 $N(\mu, \sigma^2)$ 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.

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

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).

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.

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 $n = 11$, $m = 13$, $\bar{x} = 1.03$, $s_X^2 = 0.24$, $\bar{y} = 1.66$, $s_Y^2 = 0.35$,

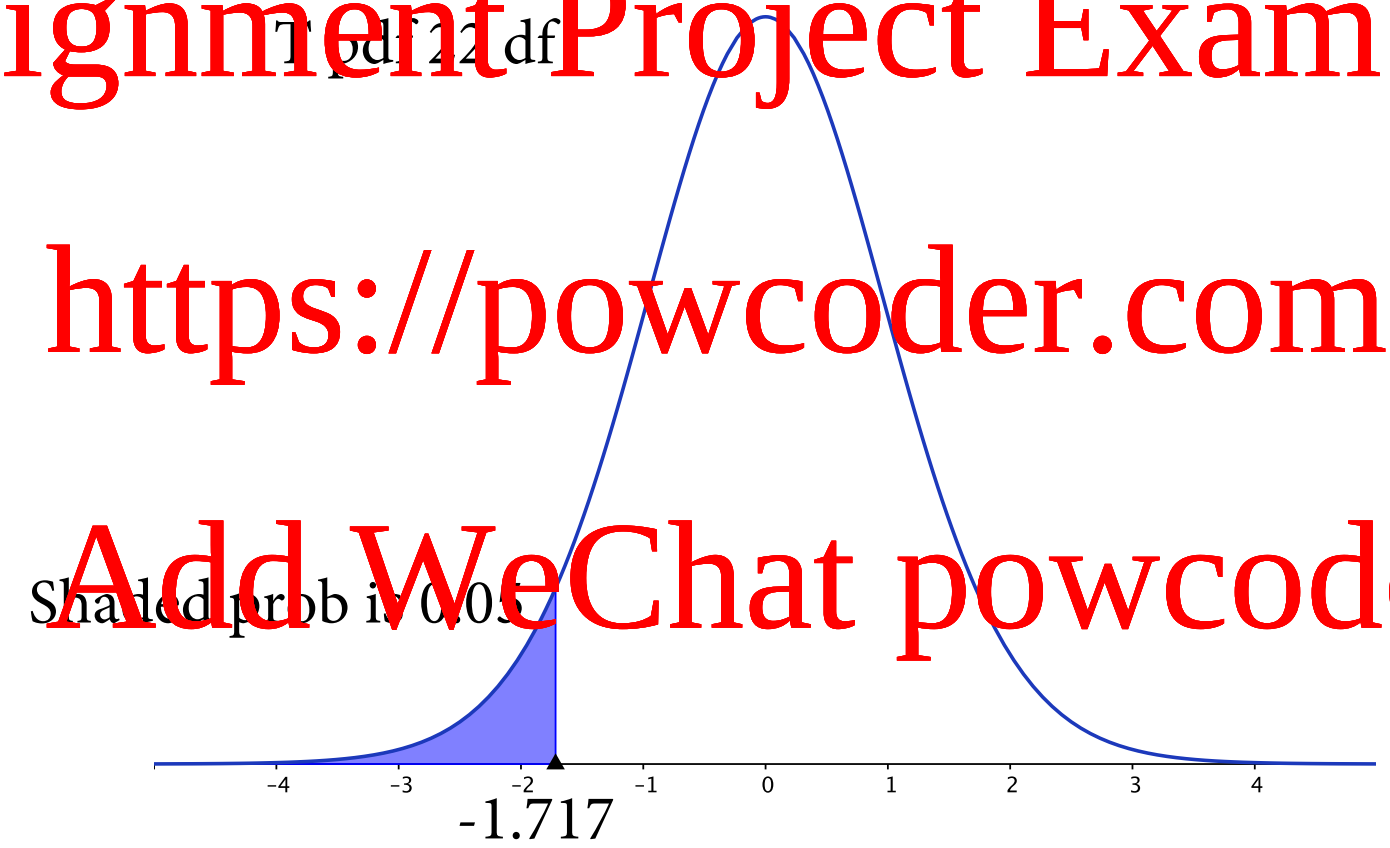
$$s_p^2 = \frac{10 \times 0.24 + 12 \times 0.35}{11 + 13 - 2} = 0.3 = 0.548^2$$

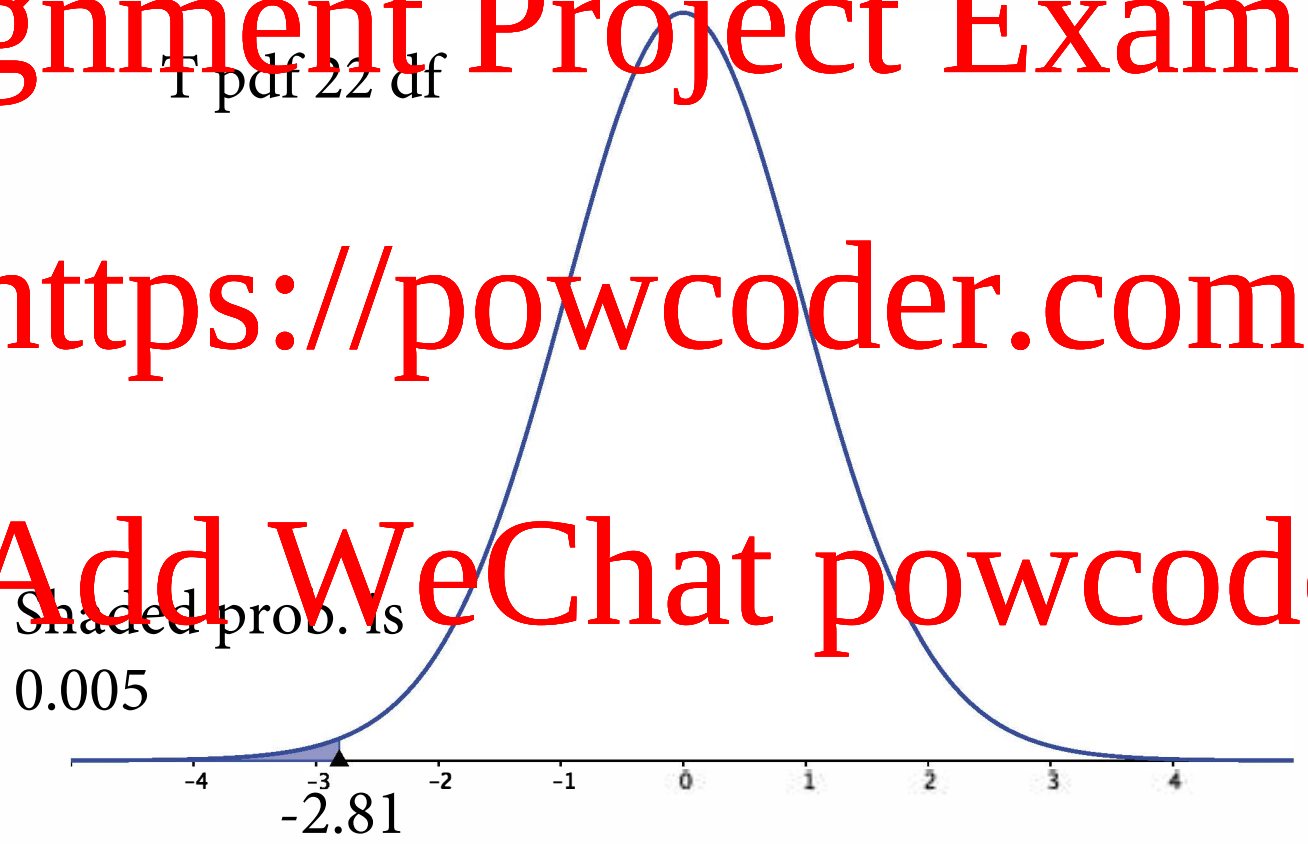
and thus

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

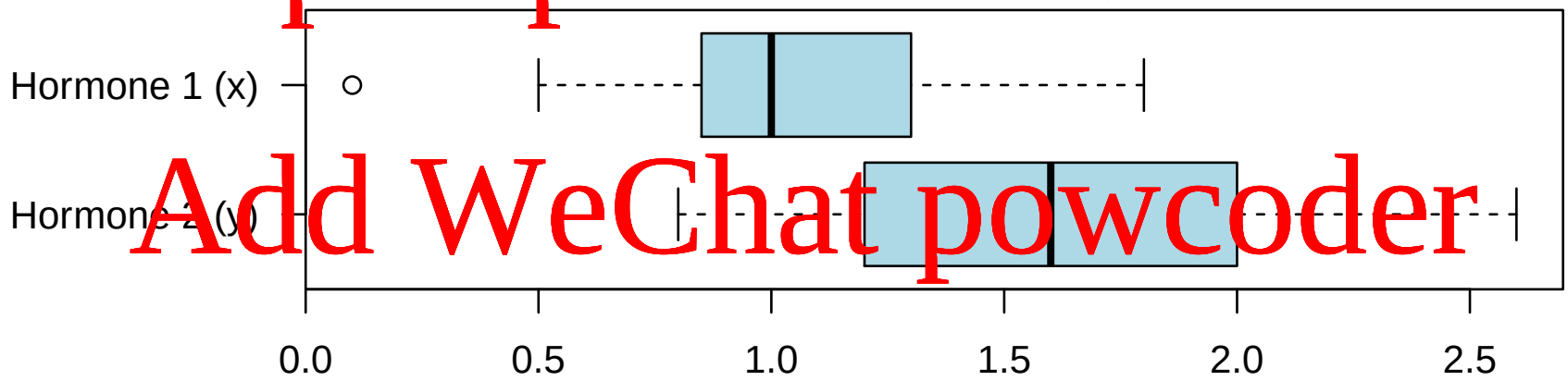
- The critical value is $c = -1.717$ (corresponding to $\alpha = 0.05$ and $n + m - 2 = 22$) so we reject H_0 and conclude that there is statistically significant evidence of less growth with hormone 1
- The p-value is

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





```
> 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)

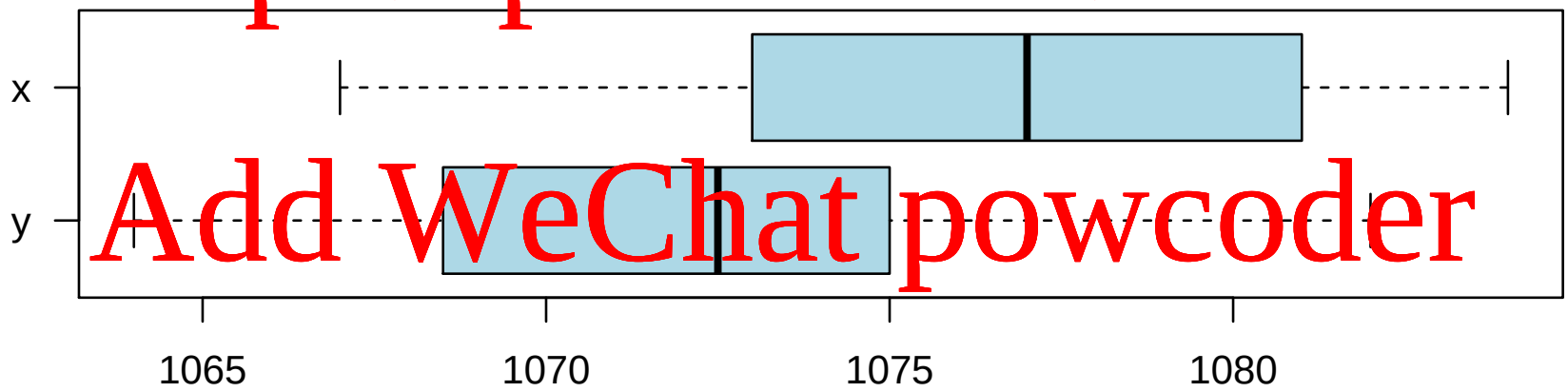
The weights (in grams) of packages filled by two methods are $X \sim \mathrm{N}(\mu_X, \sigma^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma^2)$.

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 = 2.1053, df = 22, p-value = 0.05215
alternative hypothesis:
  true difference in means is not equal to 0
95 percent confidence interval:
 -0.04438773  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.

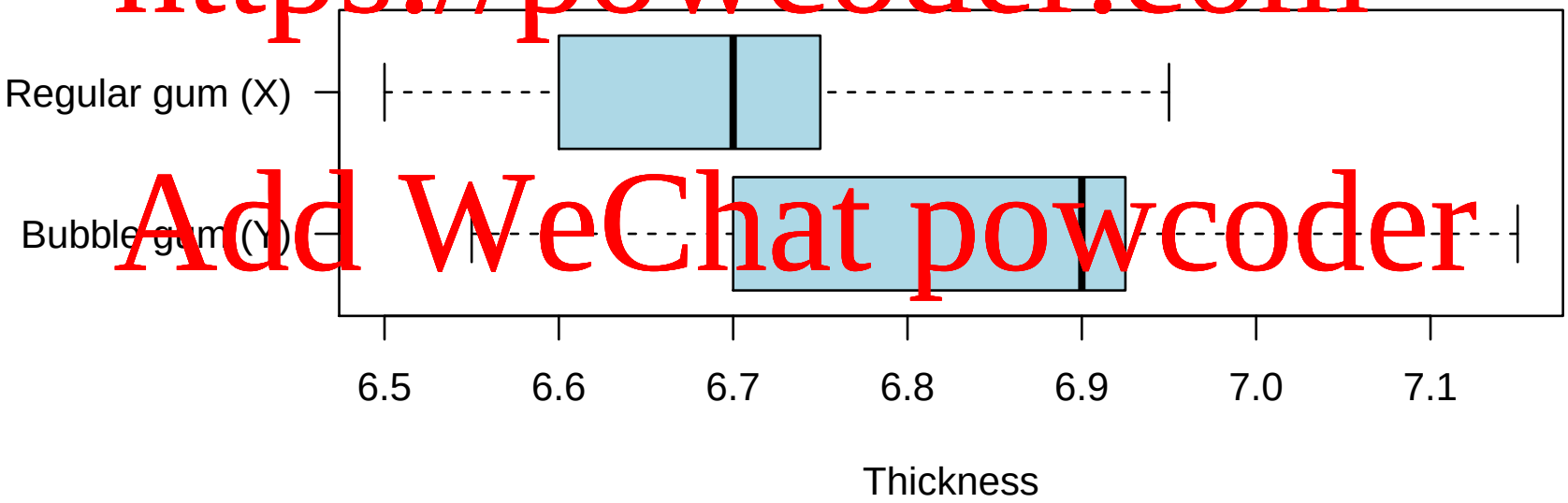
$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

We use the Welch approximation (see Module 3),

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

```
> 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:
 -0.1973427 -0.0826573
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
```

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,

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

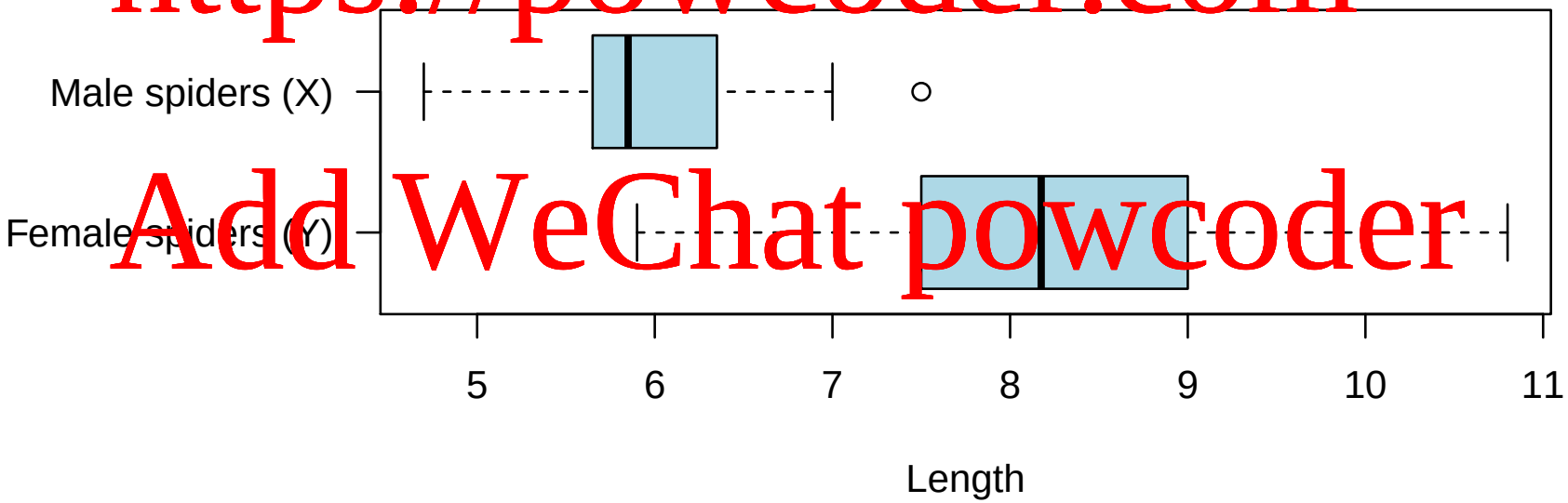
Example (normal, two variances)

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

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

```
> 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: $s_X^2 = 0.4399$, $s_Y^2 = 1.41$,

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

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
```

Outline

Preface
A cautionary word
A motivating example
Classical hypothesis testing (Neyman-Pearson)
Hypotheses
Test statistics
Errors (Type I, Type II)
Significance level & power
Alternative formulations
Significance testing (Fisher)
Modern hypothesis testing
Common scenarios
Single proportion
Two proportions
Single mean
Single variance

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'.

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...

‘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?
- Do we 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'