Type I and type II errors

2022-03-3

Type I and Type II errors

p-value the probability of observed difference or something more extreme given the null hypothesis is true

	H0 is:	
	True	False
Fail to reject	Correct	False Negative
Reject	False positive	Correct

	H0 is:	
	True	False
Fail to reject	-	Type II error
Reject	Type I error	-

Multiple testing for RNA-seq

In the analysis of large scale datasets we apply hypothesis testing of 100s, 1000s, 10000s of hypotheses.

we test 10s of genes by qPCR

In RNA-seq for each gene we test - we apply a distinct hypothesis test to each one of the genes assayed

Multiple testing for RNA-seq

In the analysis of large scale datasets we apply hypothesis testing of 100s, 1000s, 10000s of hypotheses.

we test 10s of genes by qPCR

In RNA-seq for each gene we test - we apply a distinct hypothesis test to each one of the genes assayed

we have a massive problem with multiple testing

- $\mathbf{H}_0$ the means of two populations are equal.
- $\mathbf{H}_1$ the means of two populations are not equal

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_*} \qquad (1)$$

```
?t.test
```

P-values are random variables

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Perform a t.test. (**?t.test**)

P-values are random variables

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Perform a t.test. (**?t.test**)

now do this 20000 times .... save the resulting p-value from each iteration and plot as a histogram

P-values are random variables

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Draw 100 random numbers from a normal distribution with a mean of 0 and a standard deviation of 1 (**?rnorm**)

Perform a t.test. (**?t.test**)

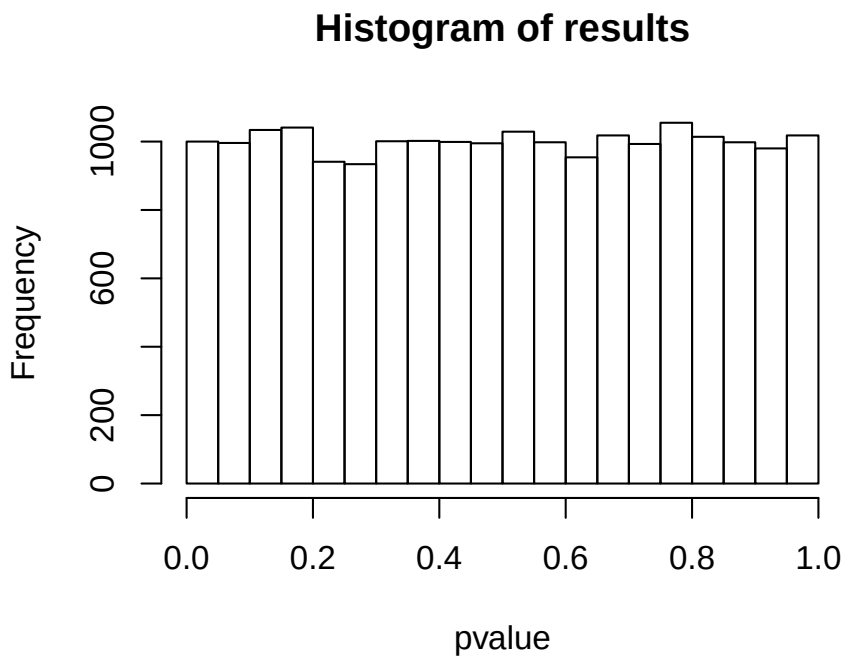
now do this 20000 times .... save the resulting p-value from each iteration and plot as a histogram

what do you expect?

Have a go ...

Distribution of p-values under the null

```
results = c()
for(i in 1:20000){
  x = rnorm(100, 0, 1)
  y = rnorm(100, 0, 1)
  results = c(results, t.test(x,y)$p.value)
}
hist(results, xlab="pvalue")
sum(results < 0.05)
```



Type I and Type II errors

p-value the probability of observed difference or something more extreme given the null hypothesis is true

	H0 is:	
	True	False
Fail to reject	Correct	False Negative
Reject	False positive	Correct

	H0 is:	
	True	False
Fail to reject	-	Type II error
Reject	Type I error (α)	-

Type I and Type II errors

p-value the probability of observed difference or something more extreme given the null hypothesis is true

expected number of false positives $= \alpha N$

Do jelly beans cause acne?

H_0 : There is no evidence of an association between eating jelly beans and having acne.

H_1 : There is evidence of an association between eating jelly beans having acne ($p < \alpha$)

https://xkcd.com/882/

Do jelly beans cause acne?

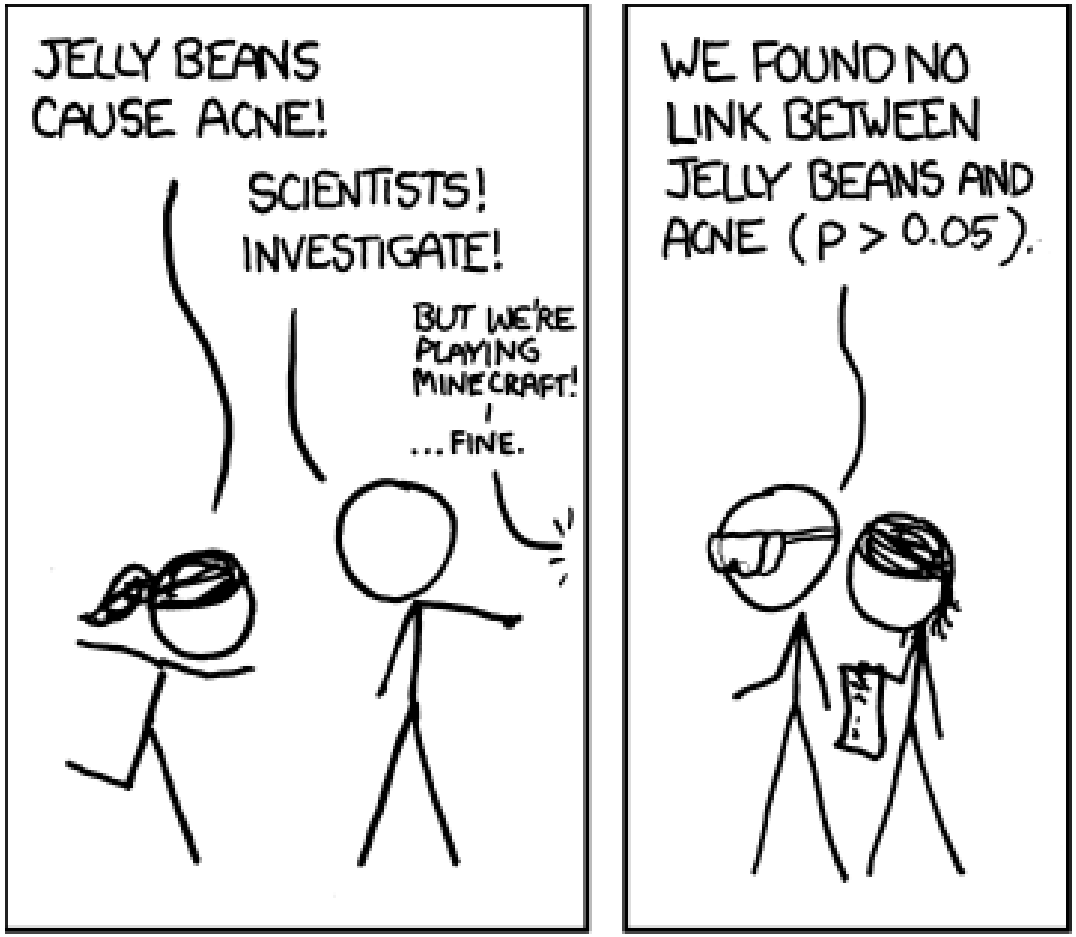
		Jelly beans +	Jelly beans -
Acne	+	a	b
	-	c	d

H_0 : There is no evidence of an association between eating jelly beans and having acne.

H_1 : There is evidence of an association between eating jelly beans having acne ($p < \alpha$)

https://xkcd.com/882/

Do jelly beans cause acne?

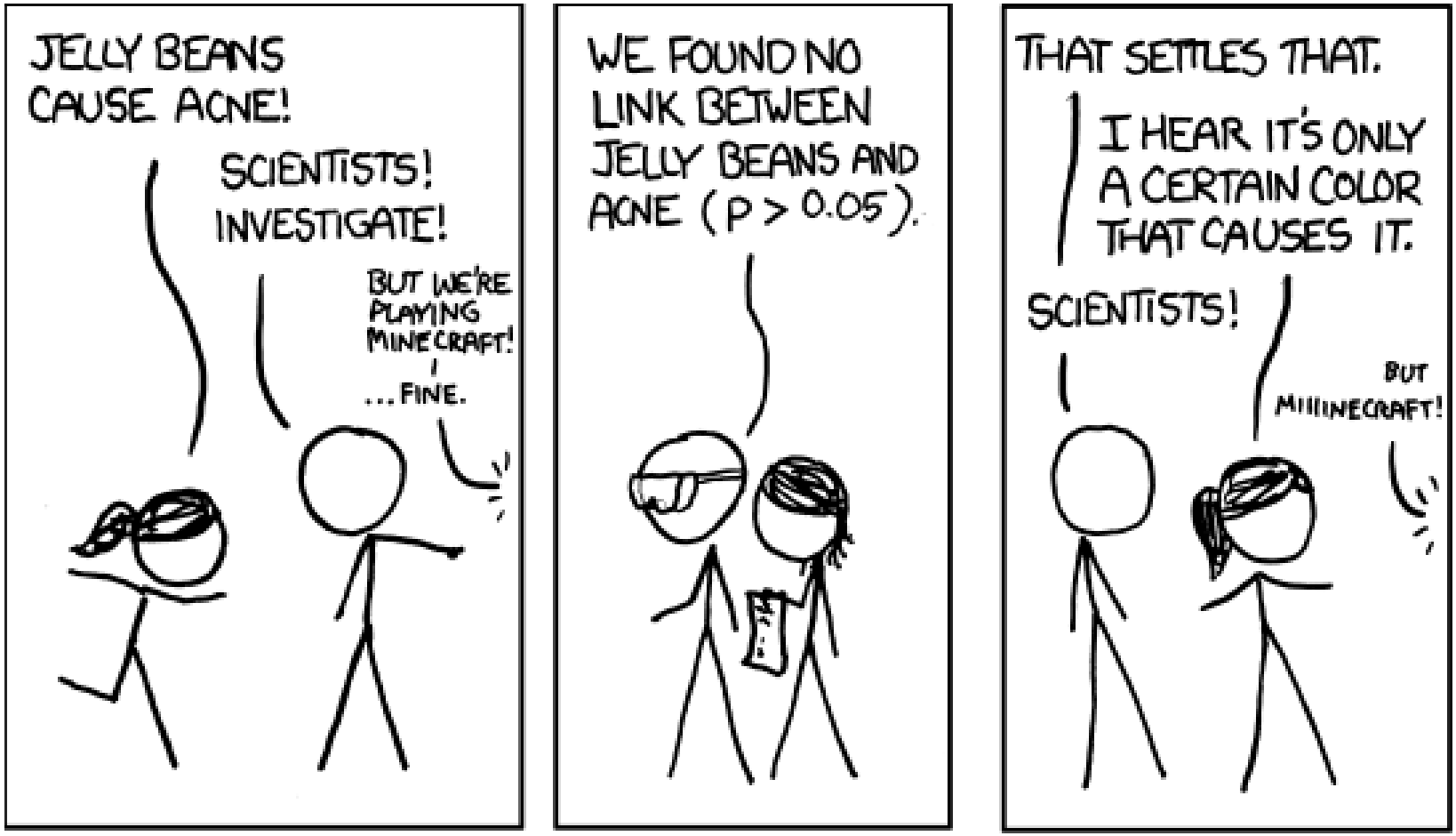


H_0 : **There is no evidence of an association between eating jelly beans and having acne.**

H_1 : There is evidence of an association between eating jelly beans having acne ($p < \alpha$)

https://xkcd.com/882/

Do jelly beans cause acne?

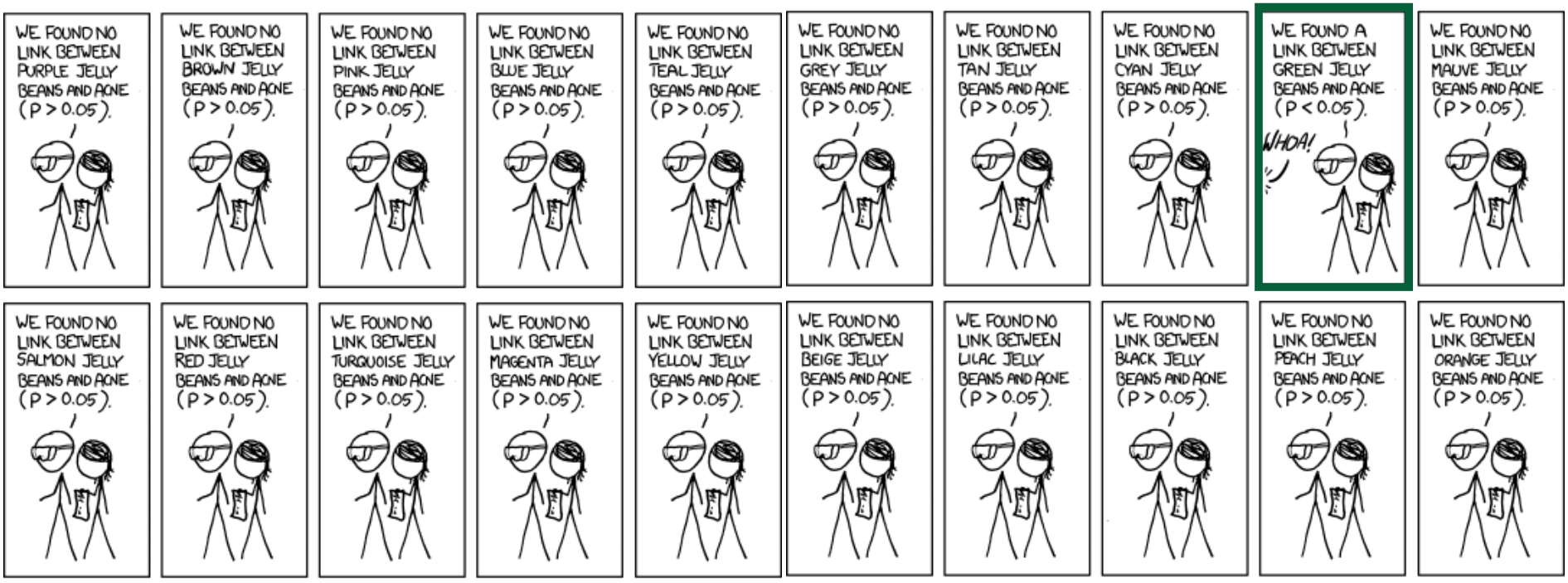


H_0 : **There is no evidence of an association between eating jelly beans and having acne.**

H_1 : There is evidence of an association between eating jelly beans having acne ($p < \alpha$)

https://xkcd.com/882/

Do jelly beans cause acne?



H_0 : There is no evidence of an association between eating {X} coloured jelly beans and having acne.

H_1 : There is evidence of an association between eating {X} coloured jelly beans having acne ($p < \alpha$)

https://xkcd.com/882/

How many crimes against science and statistics have been committed?

How many crimes against science and statistics have been committed?

- testing multiple hypotheses without proper correction
- torturing the data until some test gave us a significant result - hypothesis switching - **p-hacking**
- overstating results - ridiculous press release
- is this actually a useful/relevant question to expend time/resources on?

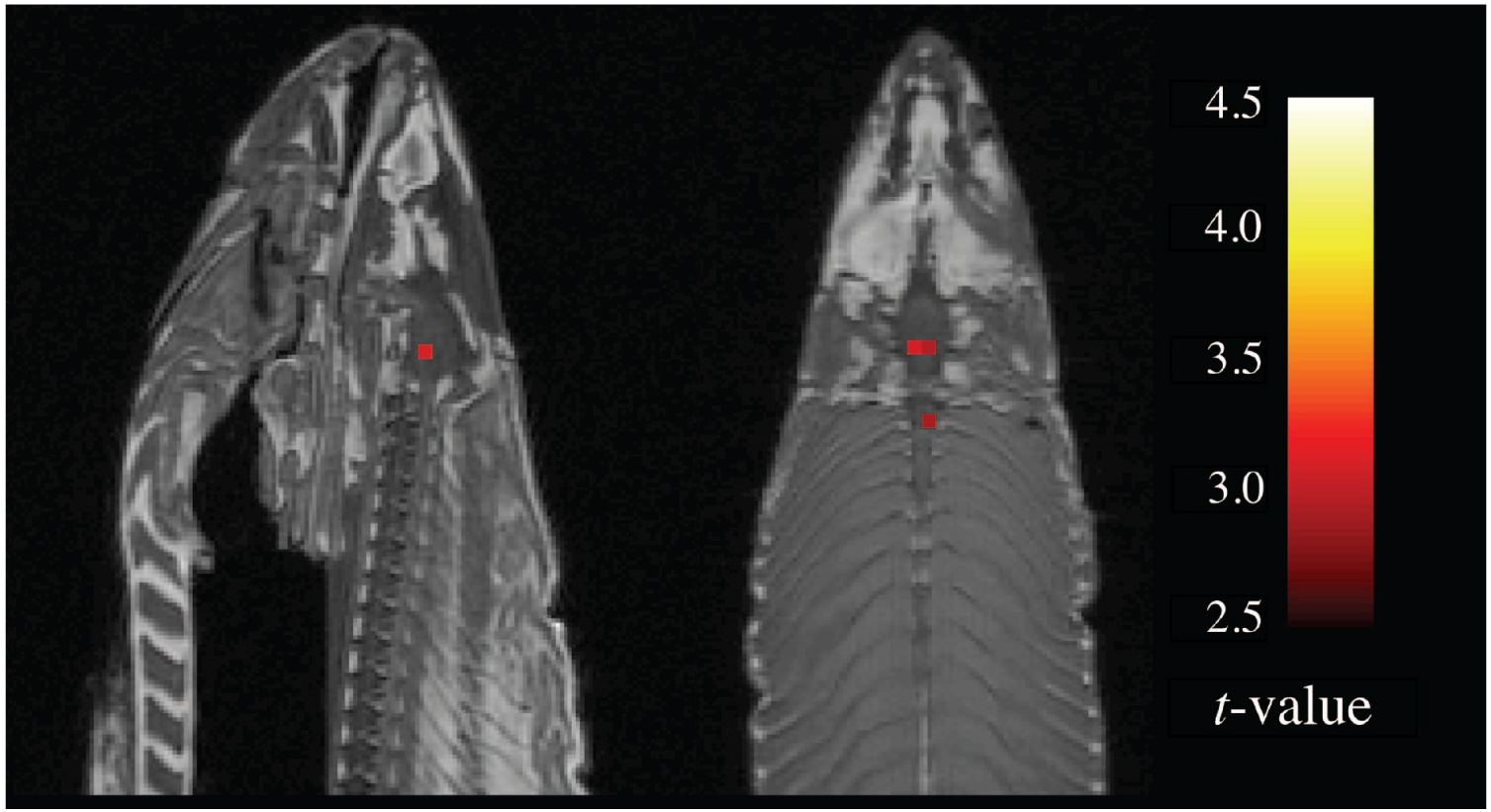
THE DEAD SALMON STUDY

The salmon measured approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.

The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing.

Bennett *et al.* et al. 2010.

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction



https://blogs.scientificamerican.com/scicurious-brain/ignobel-prize-in-neuroscience-the-dead-salmon-study/

Bennett *et al.* et al. 2010.

The family wise error rate

V = number of false positives (Type I errors)
m_0 = number of null hypotheses

$$P(V \geq 1) \leq 0.05 \tag{2}$$

$$P(V \geq 1) = 1 - P(\text{not rejecting any of the } m_0 \text{ nulls}) \tag{3}$$

$$P(V \geq 1) = 1 - (1 - \alpha)^{m_0} \tag{4}$$

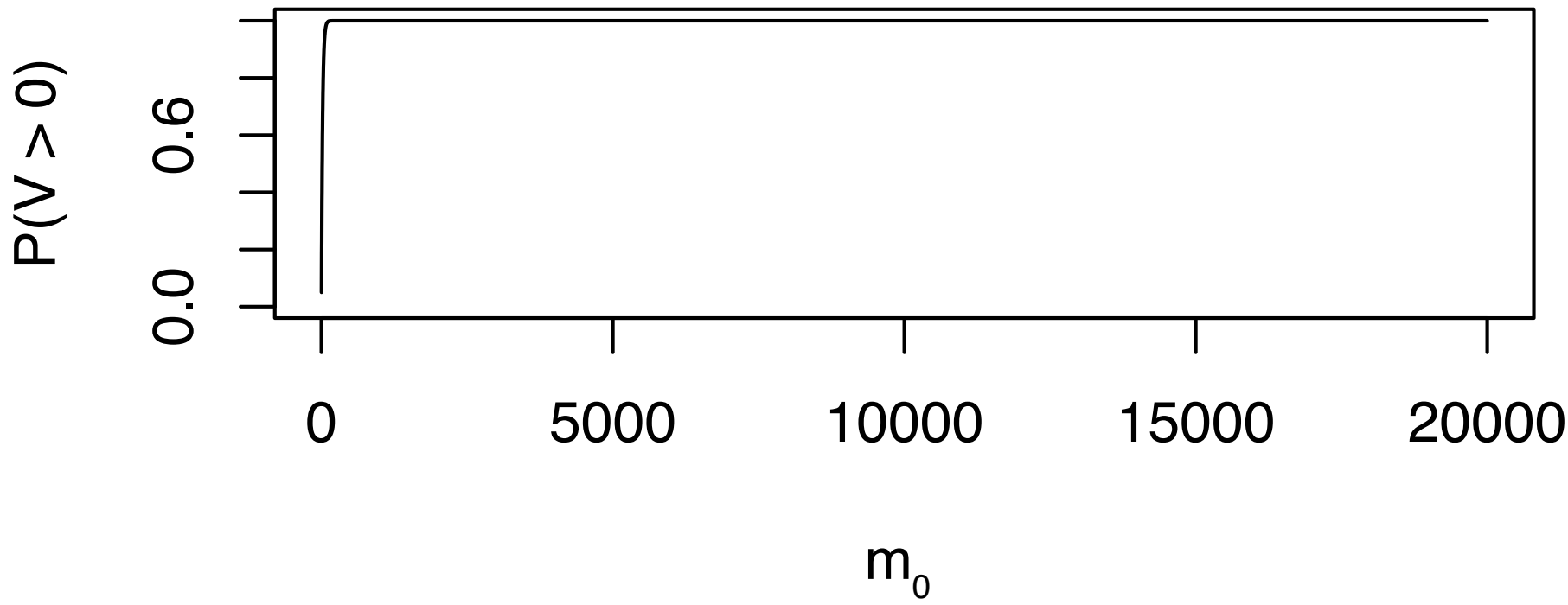
Exercise

plot P(V $>$ 0) for different values of m_0 (1-20000)

The family wise error rate

```
plot(1:20000, 1 - (1 - alpha)^seq(1:20000), type="l",
xlab="m0", ylab="P(V > 0)", ylim=c(0,1))
```

$$P(V \geq 1) = 1 - (1 - \alpha)^{m_0} \tag{5}$$



Bonferroni correction

V = number of false positives (Type I errors)

m_0 = number of null hypotheses

Bonferroni correction

V = number of false positives (Type I errors)

m_0 = number of null hypotheses

but we don't know m_0

Bonferroni correction

V = number of false positives (Type I errors)

m_0 = number of null hypotheses

but we don't know m_0

we only know the total number of tests m (i.e. $m_0 + m_1$) and that $m_0 \leq m$

Bonferroni correction

V = number of false positives (Type I errors)
m_0 = number of null hypotheses
but we don't know m_0
we only know the total number of tests m (i.e. $m_0 + m_1$) and that $m_0 \leq m$

$$P(V \geq 1) = 1 - (1 - \alpha)^m \tag{6}$$

Bonferroni correction

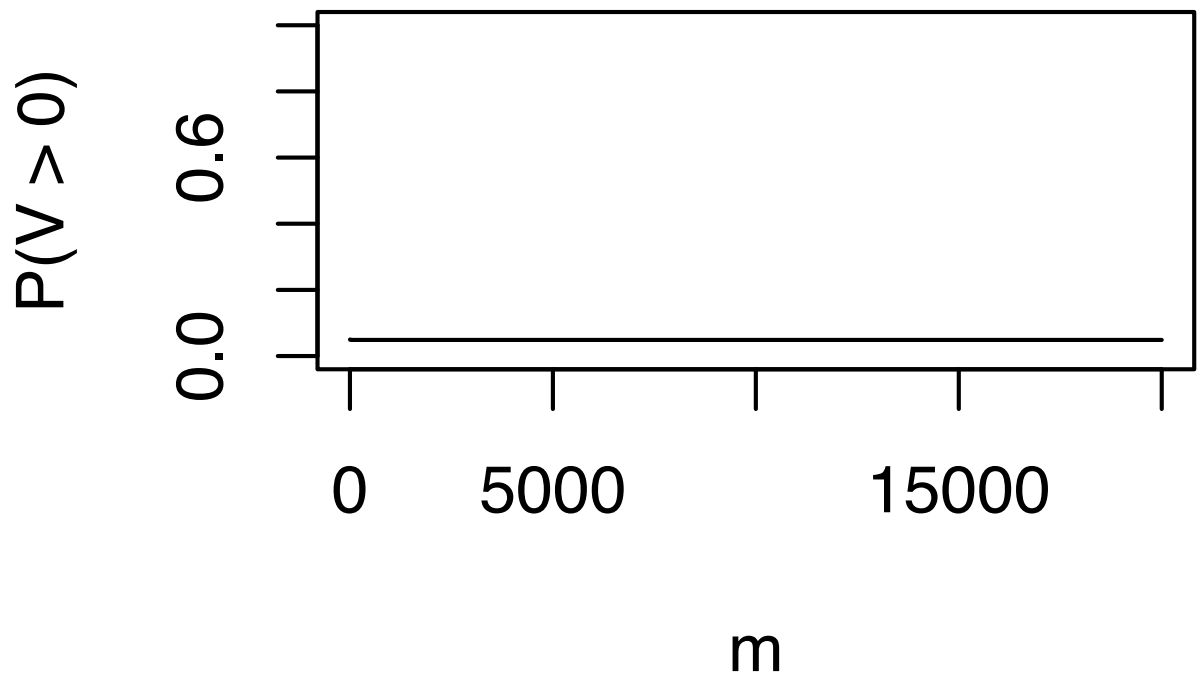
V = number of false positives (Type I errors)
m_0 = number of null hypotheses
but we don't know m_0
we only know the total number of tests m (i.e. $m_0 + m_1$) and that $m_0 \leq m$

$$P(V \geq 1) = 1 - (1 - \alpha)^m \tag{6}$$

$$P(V \geq 1) = 1 - (1 - \frac{\alpha}{m})^m \tag{7}$$

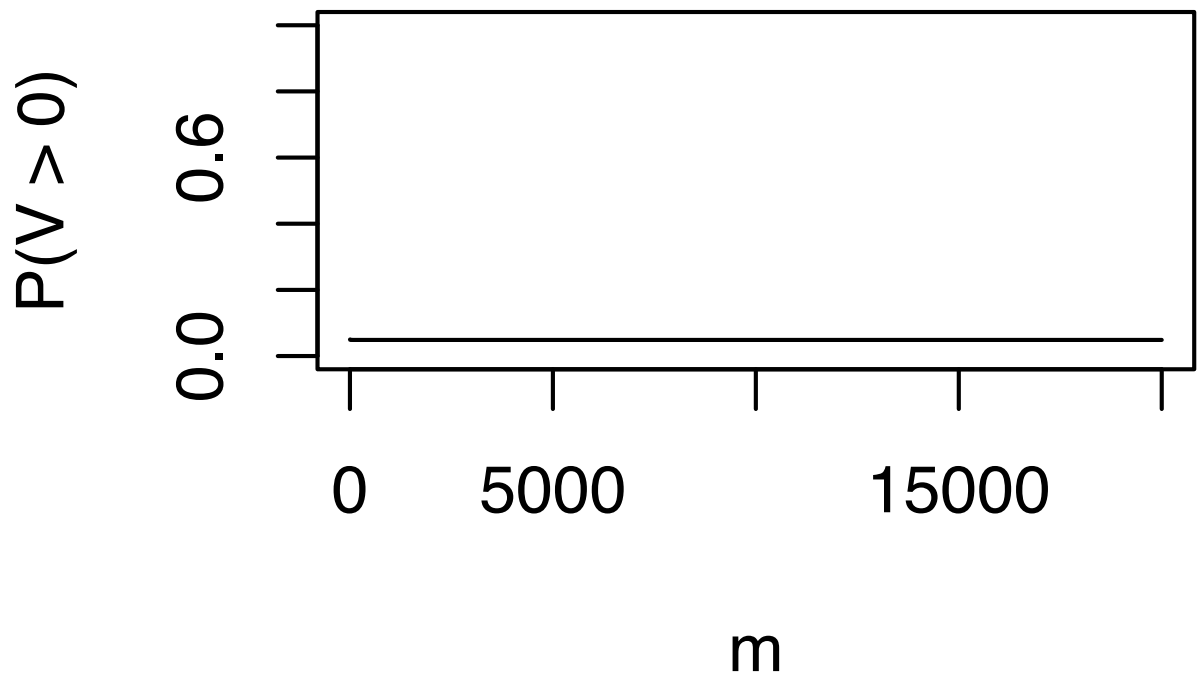
Bonferroni correction

```
plot(1:20000,
       1 - (1 - alpha/seq(1:20000))^seq(1:20000),
       type="l", xlab="m", ylab="P(V > 0)", ylim=c(0,1))
```



Bonferroni correction

```
plot(1:20000,
       1 - (1 - alpha/seq(1:20000))^seq(1:20000),
       type="l", xlab="m", ylab="P(V > 0)", ylim=c(0,1))
```



if m is large, the rejection threshold ($\frac{\alpha}{m}$) is very small.

this is pretty useless for situations where we test lots of hypotheses.... let m = 20000

α^* = 0.05 / 20000 = 2.5e-07

need to control for Type I errors (multiple testing problem) in a different way ... controlling the **False Discovery Rate**

?p.adjust

False discovery rate - FDR

What is padj?

Is it really so bad to make a False Discovery?
Relax our requirement that we never make a false discovery, as long as compared to the number of real discoveries, the false discoveries are minimal.

FDR = proportion of rejected hypotheses are wrong (number of type I errors / number of tests declared as significant)

$$FDR = E\left(\frac{V}{R}\right)$$

keep the population false positive rate below a defined threshold

V = Number of Type I errors
R = total number of hypotheses declared as significant

False discovery rate - FDR
Benjamini-Hochberg

adjusts p-values - makes them larger

- order the raw p-values $p_i \leq .... \leq p_m$ (smallest to largest)
- rank the pvalues
- start at m:
  - Calculate each individual p-value's Benjamini-Hochberg critical value, $q_i = \frac{p_i N}{i}$
  - ith FDR is the minimum previous FDR or this value (keeps it monotonic)

i = the individual p-value's rank

N = total number of tests

p_i = i th smallest P-value

FDR

$$q_i = \frac{p_i N}{i}$$

The numerator is the expected number of Type I errors if you accept all results that have P-values $\leq p_i$ or smaller.

The denominator (i) is the number of results you actually accept at the ith P-value threshold.

	p-value	rank	$q_i = \frac{p_i N}{i}$
A	0.001	1	0.00800000
B	0.008	2	0.03200000
C	0.039	3	0.06720000
D	0.041	4	0.06720000
E	0.042	5	0.06720000
F	0.060	6	0.08000000
G	0.074	7	0.08457143
H	0.205	8	0.20500000

Adjusting p-values in R in general

```
?p.adjust
```

What do you expect if data come from different distributions?

```
for( i in 1:10000 ) {
    x = rnorm(10, 0, 1)
    y = rnorm(10, 1.5, 1)
    result = c(result, t.test(x,y)$p.value)
}
hist(result)
```

p-values > 0.05 are false negatives

Statistical power

if the true effect is of a specified size and the experiment is repeated many times, what fraction of the results will be significant?

the ability of a test to correctly reject a false null hypothesis

what things does the size of a p-value depend on ?

statistical power

what things does the size of a p-value depend on ?

sample size

variability

effect size

Type I and Type II errors

p-value the probability of observed difference or something more extreme given the null hypothesis is true

	H0 is:	
	True	False
Fail to reject	Correct	False Negative
Reject	False positive	Correct

	H0 is:	
	True	False
Fail to reject	1 - α	Type II error (β)
Reject	Type I error (α)	1 - β

Power

power is equal to 1 - β

so we want to understand how many replicates/samples to have to achieve a specified **power**

typically set to 80%

given .... **variance**, **sample size**, **significance level** what is the power of this experiment to detect an **effect of a specific size**

```
for( i in 1:10000 ) {
    x = rnorm(10, 0, 1)
    y = rnorm(10, 1.5, 1)
    result = c(result, t.test(x,y)$p.value)
}
hist(result)
```

how many are false negatives?

what is the power of this test?

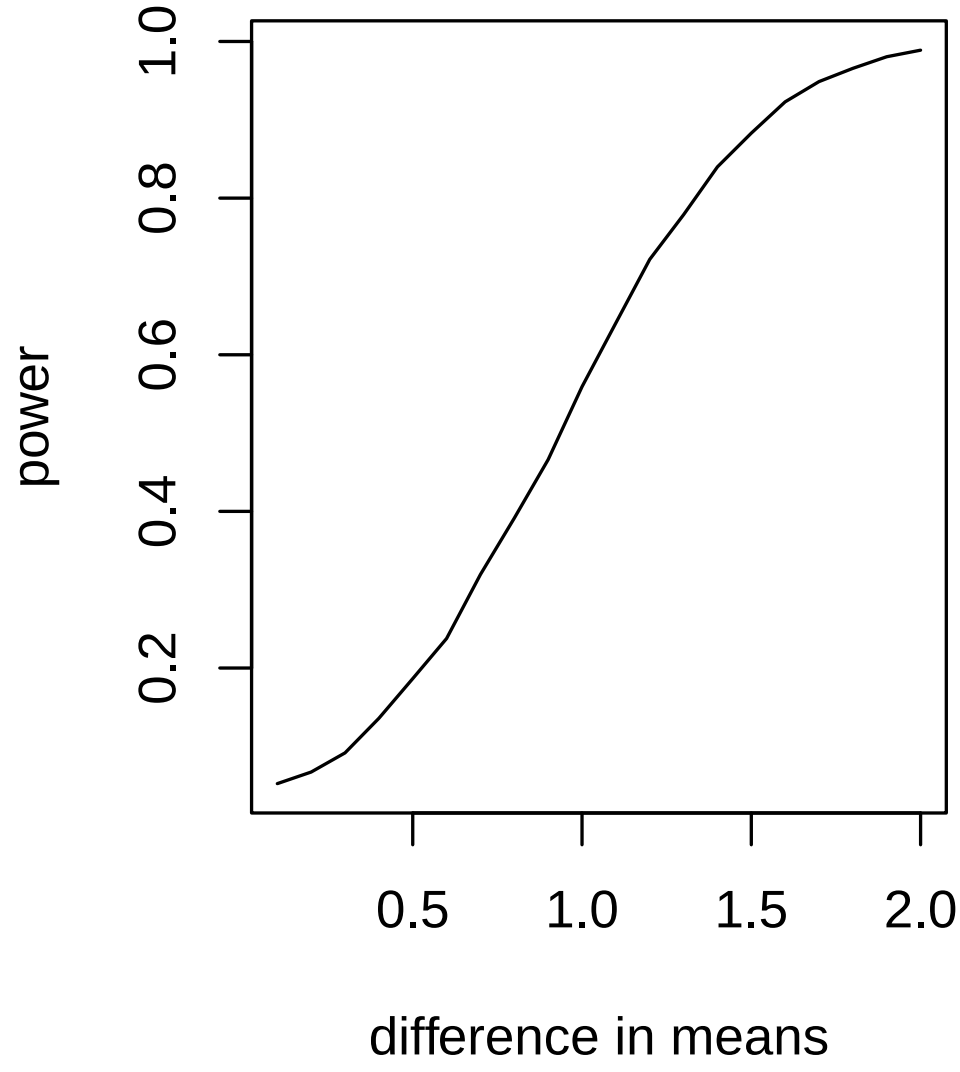
what happens if we make the difference between the means smaller or larger?

lets try plotting power as a function of the difference between two means

lets keep, everything the same but now lets see what happens if you change effect size from 0.1 to 2 in steps of 0.1

for(effectsize in seq(0.1, 2, by=0.1))

lets try plotting power as a function of the difference between two means



what happens if I change the sd

```
for( i in 1:10000 ) {
    x = rnorm(10, 0, 2)
    y = rnorm(10, 1.5, 2)
    result = c(result, t.test(x,y)$p.value)
}
hist(result)
1 - (sum(result > 0.05)/length(result))
```

how many are false negatives?

what is the power of this test?

what happens if we make the sd smaller or larger?

what happens if I change the sample size

```
for( i in 1:10000 ) {
    x = rnorm(10, 0, 1)
    y = rnorm(10, 1.5, 1)
    result = c(result, t.test(x,y)$p.value)
}
hist(result)
1 - (sum(result > 0.05)/length(result))
```

how many are false negatives?

what is the power of this test?

what happens if we make the sample size smaller or larger?

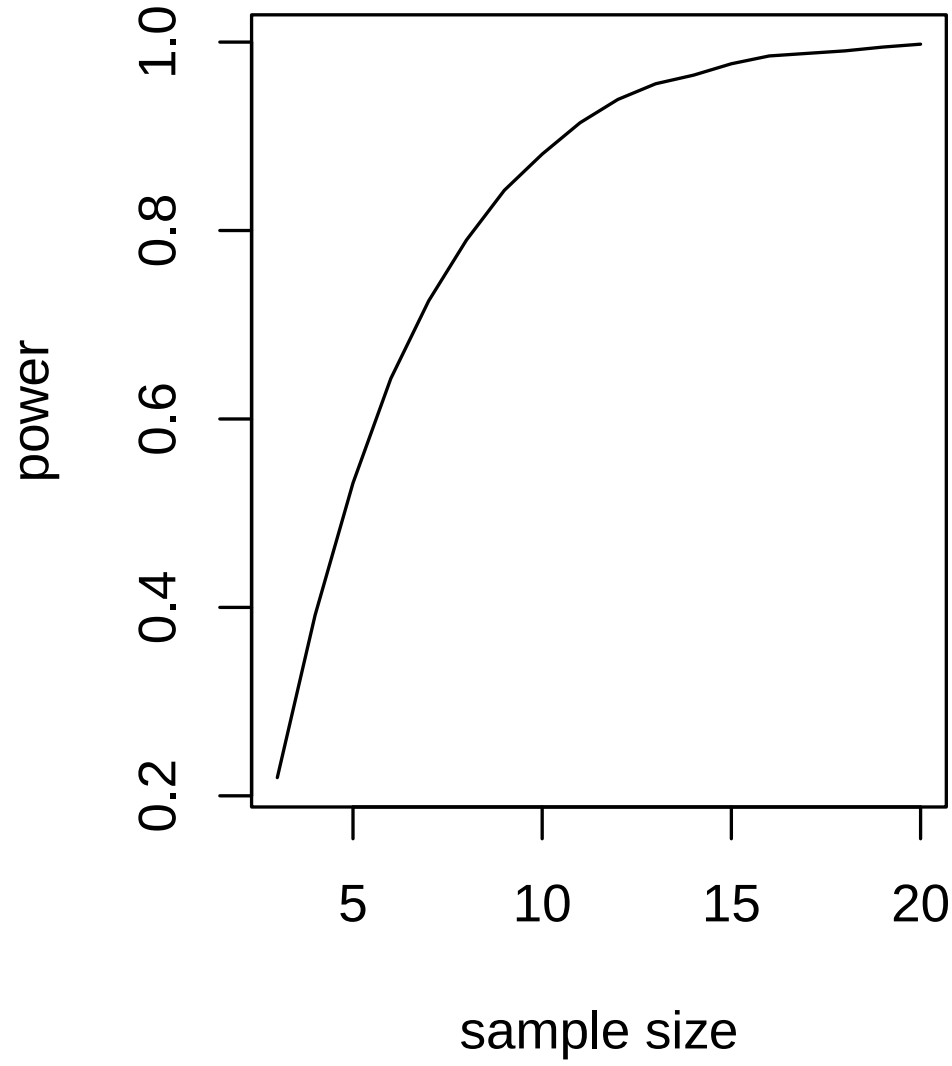
lets try plotting power as a function of sample size

effectsize = 1.5

sd = 1

samplesize in seq(3, 20, by=1)

lets try plotting power as a function of sample size



type I and II errors

the boy who cried wolf

first everyone believed that there was a wolf, when there wasn't. Next they believe there was no wolf, when there was.

replace wolf with effect

- next week: Problemset 1 answers, linear models, bootstrapping
- reading: chapters 32, 33, 34, 35
- problemset 2 available tonight or over the weekend