Replicability

Data Science 101 Team

Multiple testing

- The classical testing strategies we have studied are okay when we are interested in testing one hypothesis
- We will see that when we consider many hypotheses, then some adjustment needs to be made
- We start by defining / reviewing the concepts of Type-I and Type-II errors

Two types of error

- **Type-I** error: false positive
- **Type-II** error: false negative
- In the standard testing framework, we bound the probability of Type-I error, the probability of rejecting the null hypothesis when this is true. This is called the **level** of the test.
- The **power** of a test is, the probability of rejecting the null hypothesis when it is false

$$\text{Power} = 1 - \mathbb{P}(\text{ type II error})$$

There is a trade-off between power and P(type I error)

If one wants to have more power (i.e. have more discoveries), then this comes at the price of more type I errors.

Why? Lowering the threshold for rejecting the null leads to more discoveries, but also to more false discoveries.

As an example, suppose X follows a normal distribution with mean μ and variance 1,

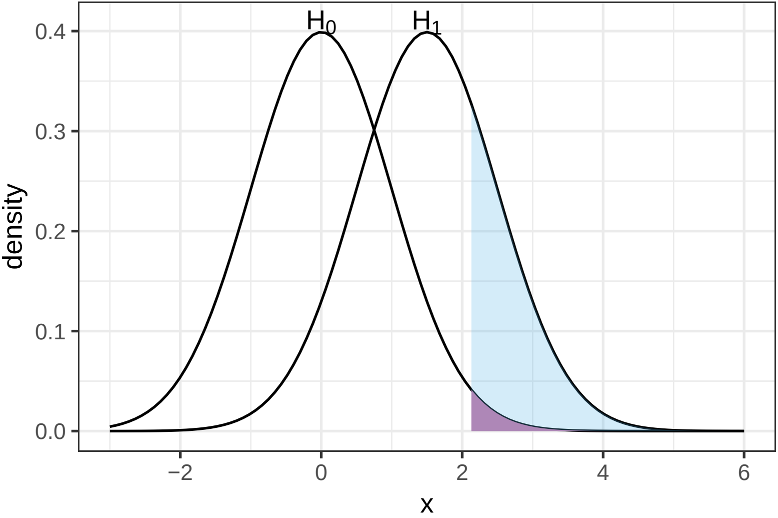
$$X \sim \mathcal{N}(\mu, 1)$$

We will visualize the power and the level (= P(type I error)) of the one-sided test

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu > 0$$

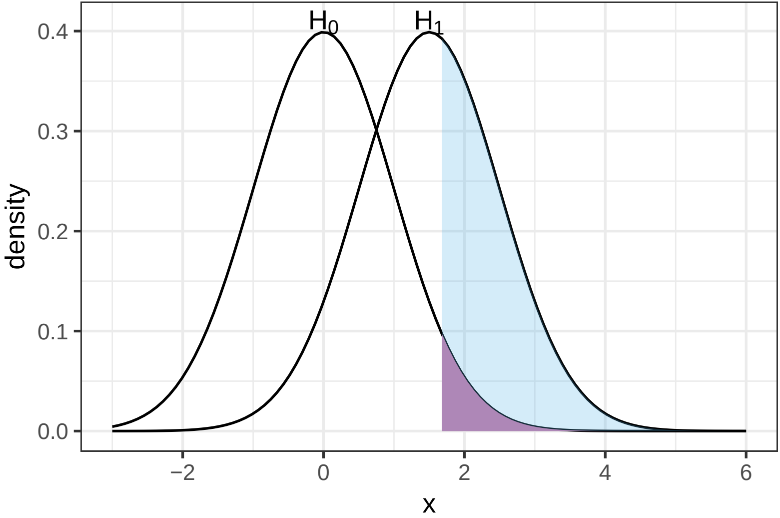
for various cut-offs.

Cut−off is 2.05



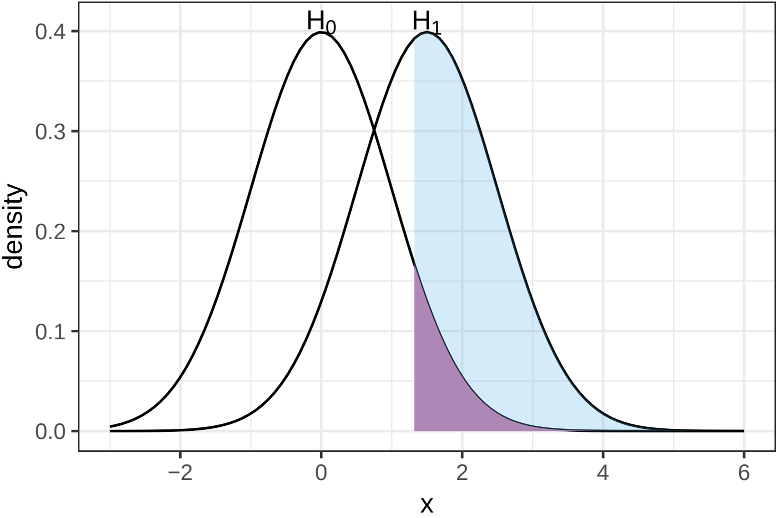
Power is 29% and $\mathbb{P}$(type I error) $= 2\%$

Cut−off is 1.65



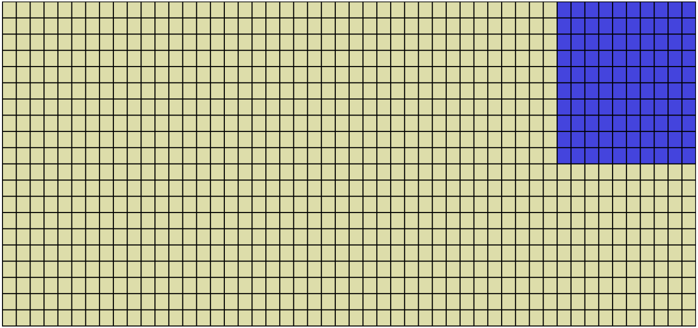
Power is 44% and $\mathbb{P}(\text{type I error}) = 5\%$

Cut−off is 1.3



Power is 59% and $\mathbb{P}$(type I error) $= 10\%$

Now consider what happens if we test many hypotheses



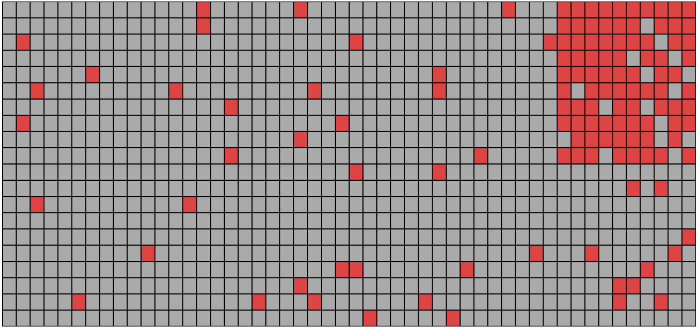
- We are interested in 1000 hypotheses
- Imagine that in truth there are 100 non null hypotheses
- Nothing going on
- Something going on

Generate some data

Suppose each test is such that $\mathbb{P}(\text{false positive}) = 0.05$ and $\mathbb{P}(\text{false negative}) = 0.2$

- How many false discoveries would you expect?
- How many false negatives?
- How many true positives?

After obtaining observations and testing each of the hypotheses

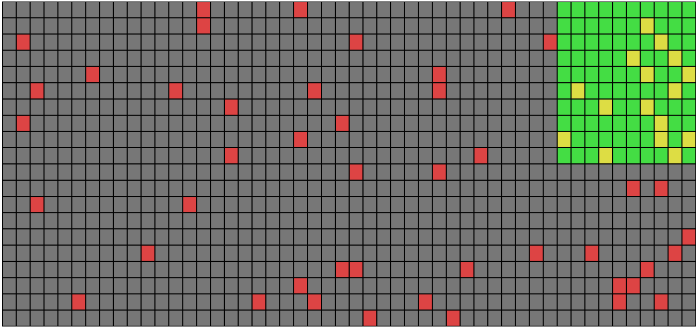


For each question, we make a decision: P(false positive)=0.05, P(false negative)=0.2.

These are the decisions we made.

- Discovery, :)
- Not a discovery, :(

Measuring errors across the entire set of hypotheses



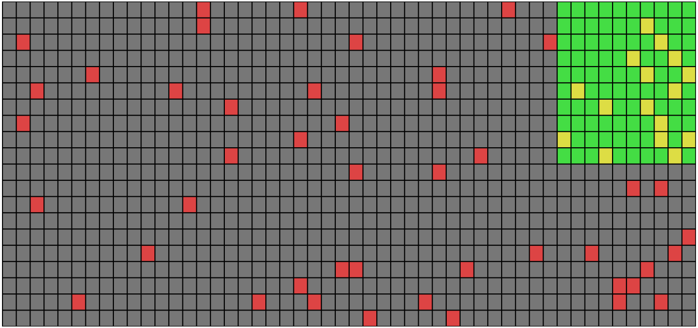
- We made 84 true discoveries
- We made 45 false discoveries
- Our *False Discovery Proportion* is 45/129=0.35.

False Discovery Proportion (FDP)

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

(When no discoveries are made, we set $\text{FDP} = 0$)

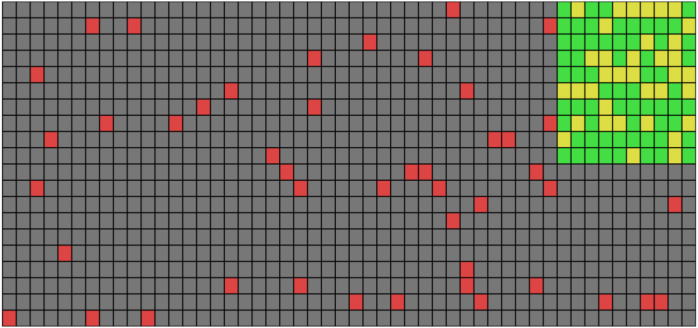
Measuring errors across the entire set of hypotheses



```
##     TD TTD TFD       FDP
## 1 129  84  45 0.3488372
```

Even with a power of 80%, a good portion of what we would report would be false.

Let's look at how the results change for a different parameter setting



For each study P(false negative)=0.4, P(false positive)=0.05.

```
##     TD TTD TFD       FDP
## 1 110  64  46 0.4181818
```

With a power of 60%, close to half of what we would report would be false!

False Discovery Rate

- We cannot observe the FDP, because this would require knowing, case by case, the true status of each hypothesis
- But there is a way to control its expected value (the FDP averaged over many experiments):

$$\mathsf{FDR} = \mathbb{E}(\mathsf{FDP})$$

Familywise error rate (FWER)

- Another possible measure of global error
- FWER: probability of making at least one wrong rejection
- It is a natural extension of the level of test (probability of Type-I error)
- It is actually the “oldest” measure of global error, but it is considered quite conservative, i.e. the price for having not a single false negative is having few true discoveries.
- With many tests, all having their individual Type-I error probability at 5%, the FWER rises quickly to 1. So wanting FWER $\leq 5\%$ requires the individual Type-I error probability to be much smaller than 5%.

Let's look at some examples:

```r
truth = rep(0, 10000) # 10,000 tests
N = 1000 # number of non-zero means
mu = 2 # signal strength
truth[1:N] = mu # Non nulls with mean 2
# Generate data
y = rnorm(10000, truth, 1)
pvalue = 2*pnorm(-abs(y)) # two-sided p values
discovery = pvalue < 0.05 # discoveries
TD = sum(discovery) # number of discoveries
TTD = sum(pvalue[1:N]<0.05) # number of true discoveries
TFD = TD - TTD # number of false discoveries
FDP = (TFD)/TD # false discovery proportion
FWER = as.numeric(TFD > 0)
data.frame(TFD, FDP, FWER)
```

```
##   TFD       FDP FWER
## 1 485 0.4894046    1
```

The same example, but controlling FWER at 5%:

Bonferroni's strategy: to control the probability of making at least one false discovery at level α, we declare a discovery when

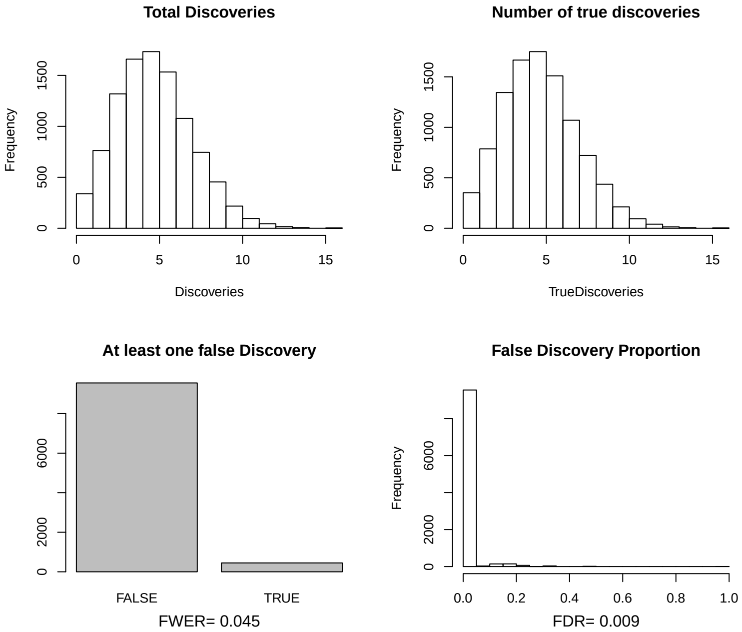
$$\text{p-value} < \frac{\alpha}{\#\text{ of tests}}$$

In our case $\alpha = 0.05$, $\#$ of tests is $10,000$.

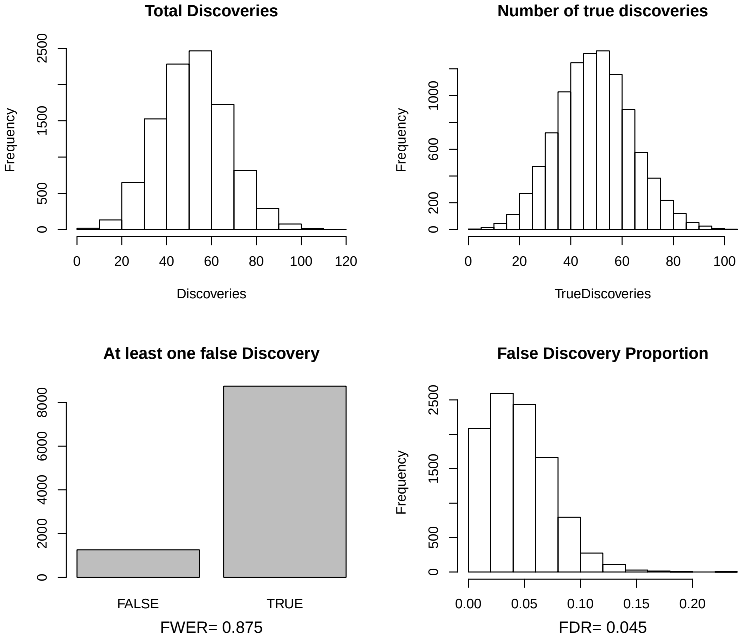
```
discovery = pvalue<0.05/10000
TD = sum(discovery)
TTD = sum(discovery[1:N])
TFD = TD - TTD
FDP = 0
if (TD>0) {FDP = (TFD)/TD}
data.frame(TFD, FDP)
```

```
##   TFD FDP
## 1   0   0
```

We can iterate this 1000 times



Following a different strategy (explained on the next slide):



FDR control

The results of the previous slide are obtained following a strategy that guarantees FDR $< q$ (with $q = 0.05$)

The strategy was introduced by Benjamini and Hochberg in 1995, together with the definition of FDR.

- It is a more liberal strategy than Bonferroni: it allows more discoveries at the price of not controlling FWER.
- It is an adaptive strategy, i.e. it depends on the data: it compares the p-values with a decreasing threshold.

Let M be the total number of hypotheses. Sort the p-values:

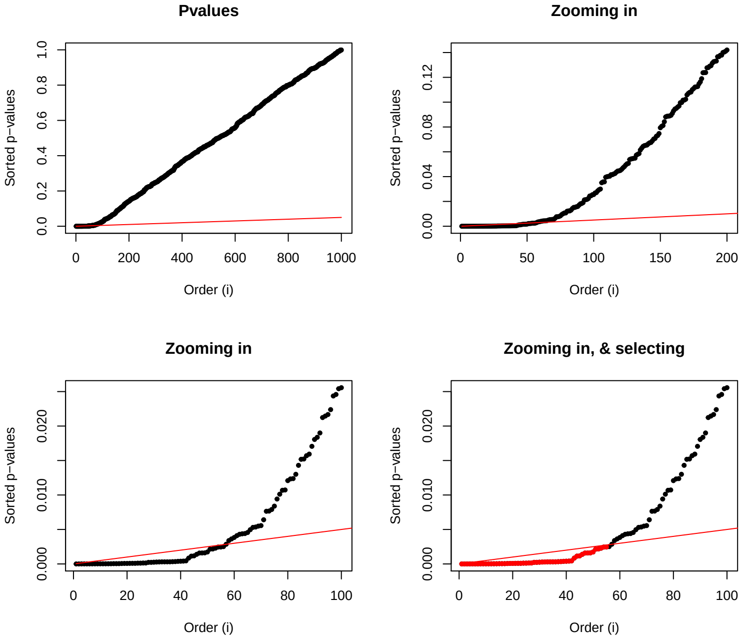
$$p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \cdots \leq p_{(M)}$$

Let j be the last value $i = 1, \ldots, M$ for which

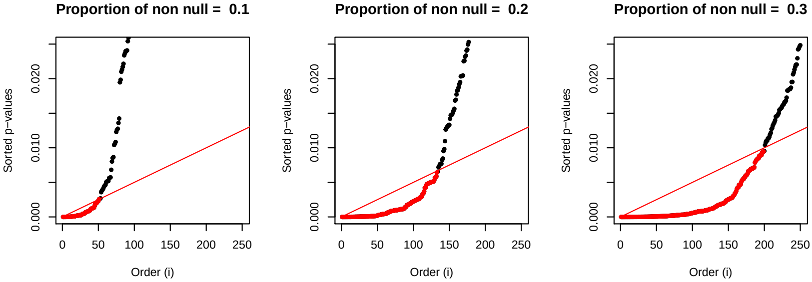
$$p_{(i)} \leq q \times \frac{i}{M}$$

- Reject all hypotheses whose p-value is $\leq p_{(j)}$

Benjamini Hochberg rule



The cut-off for rejection is ADAPTIVE: different data result in different cut-offs



Another look at BH

- We have pointed out that we cannot observe FDP because we cannot observe the numerator `TFD`.
- We *can* give a conservative estimate for TFD, the number of false discoveries: if we use a Type-I error threshold α for each individual test, then

$$\widehat{TFD}(\alpha) = \text{ Number of hypotheses} \cdot \alpha$$