

# SIMULATION 1

## -QUANTIFYING THE WORLD-

Lecturer: Darren Homrighausen, PhD

# SIMULATION

Simulation is a crucial part of data science

It can be used to...

- Estimate sampling distributions for intractable estimators  
(e.g. the bootstrap)
- Compute p-values when the reference distribution for a hypothesis test is unknown
- Find the power/size of a give test
- Develop intuition about or compare methods

# SIMULATION

Suppose we use R to find 20 random standard normal draws

# SIMULATION

Suppose we use R to find 20 random standard normal draws

```
x = rnorm(20)
```

How can I change **x** to have mean 20 and variance 3?

# SIMULATION

Suppose we use R to find 20 random standard normal draws

```
x = rnorm(20)
```

How can I change **x** to have mean 20 and variance 3?

```
x = x *sqrt(3) + 20
```

# SIZE, POWER, TYPE I AND II ERROR

Suppose we have a hypothesis test about a parameter  $\theta$

$$H_0 : \theta = \theta_0$$

$$H_A : \theta \neq \theta_0$$

(think of  $\theta_0$  as some number like  $\theta_0 = 0$ )

If we have a hypothesis test based around a test statistic  $T$ , we can compute a **p-value** as

$$P(|T| > t | \text{the null hypothesis } H_0 \text{ is true})$$

where  $t$  is some threshold

# SIZE, POWER, TYPE I AND II ERROR

A classic example is the T-test

$$T = \frac{\bar{Y}}{\text{SE}(\bar{Y})}$$

If  $\alpha$  is the **significance level**, then the standard threshold is

# SIZE, POWER, TYPE I AND II ERROR

A classic example is the T-test

$$T = \frac{\bar{Y}}{\text{SE}(\bar{Y})}$$

If  $\alpha$  is the **significance level**, then the standard threshold is

$$t = t_{\alpha/2, n-1}$$

(The  $\alpha/2$  quantile of the t-distribution with  $n - 1$  degrees of freedom)



# SIZE, POWER, TYPE I AND II ERROR

- **SIZE:** This is the probability we **do not** reject the null hypothesis given that it is **true**

$$\text{size} = P(\text{do not reject } H_0 | H_0 \text{ is true})$$

- **POWER:** This is the probability we **reject** the null hypothesis given that it is **false**

$$\text{power} = P(\text{reject } H_0 | H_0 \text{ is false})$$

- **TYPE I ERROR:**

$$1 - \text{size}$$

- **TYPE II ERROR:**

$$1 - \text{power}$$

# SIZE, POWER, TYPE I AND II ERROR

For any hypothesis test, we can define these (theoretical) quantities of interest

They can in some cases be exactly computed using probability theory

(Or approximately exactly computed using asymptotic theory)

However, in many cases it is impossible/difficult

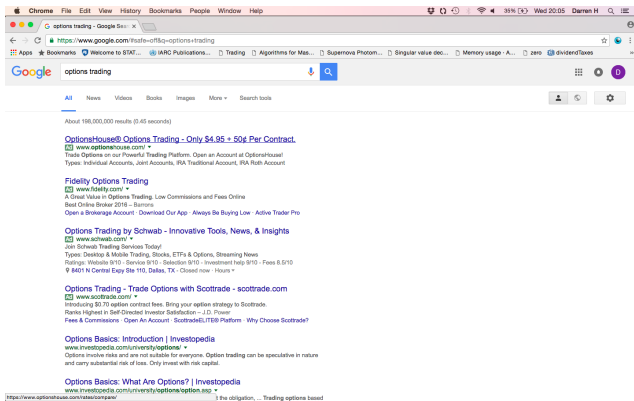
Hence, we tend to use **simulation** to compute them

# Internet advertising

# TERMINOLOGY

When discussing advertising on the internet, there are three very important concepts

- **IMPRESSIONS:** A person seeing the page with an ad
- **CLICKS:** A person clicking on that ad
- **CONVERSION:** A person doing a particular thing on the site



# ADVERTISING CONSIDERATIONS

In the particular case of Adwords companies can pay for ads in three ways

- By the click
- By the impression
- By the conversion

Companies can change how the ad is written, the keywords, the amount they are willing to pay, ...

# EXPERIMENTAL DESIGN

Companies are continually trying to tweak their ads so that more people are converted from impressions to clicks

(The rate at which this happens is called the **click-thru rate**)

This is commonly referred to as **A/B testing**, but we would know it as hypothesis testing

Let's consider a consulting problem that is really quite common

# CONSULTING PROBLEM

A new company wants to advertise on Google and has decided to pay by the **impression**

This company has come up with two different keyword sets and wants to see which one leads to the highest click-thru rate

The company wants to go with the better ad as quickly as possible

It hired our company to answer the following basic question:

How many impressions do we need to pay for before we can conclude which ad is better?

# Contingency Tables



# ANALYZING DISCRETE DATA

A **contingency table** is a cross-tabulation of discrete variables

In this particular problem, we our table would be

Ad 'A'	clicks	no clicks	impressions
Ad 'B'	clicks	no clicks	impressions

**EXAMPLE:** We pay for 1000 impressions for each ad and get the following

	clicks	no clicks	impressions
Ad 'A'	8	992	1000
Ad 'B'	5	995	1000

Can we answer which ad is better?

# CONTINGENCY TABLES

There are two main ways to answer this question

Both approaches compare the observed counts to the counts expected if there is no difference in the two ads

They differ in the approach to computing a p-value

- **CHI-SQUARED TEST:** Uses an asymptotic normal approximation
- **FISHER'S EXACT TEST:** Uses an exact hypergeometric distribution

```
contingencyTable = matrix(c(8,1000-8,5,1000-5),nrow=2,byrow=T)
> contingencyTable
  8  992
  5  995
> chisq.test(contingencyTable)$p.value
[1] 0.577861
> fisher.test(contingencyTable)$p.value
[1] 0.5797925
```