A/B testing and design

A/B TESTING IN R



Lauryn Burleigh
Data Scientist

datacamp

A/B testing

- Statistical technique
- Data-driven decisions
- Test differences and relationships
- Ideal for:
  - Optimizing marketing campaigns, testing websites and advertisements

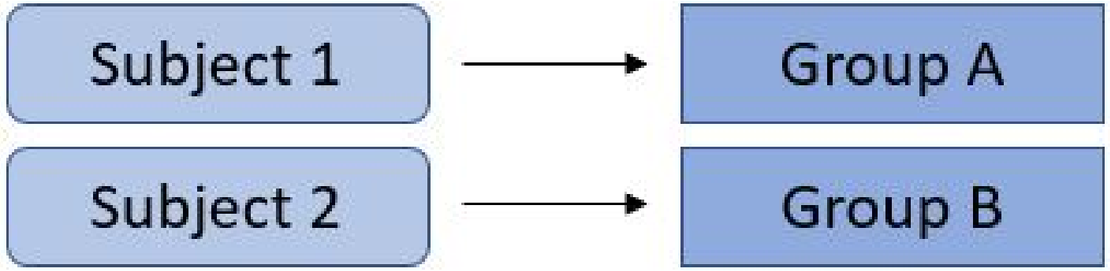
In this course:

- Analyze A/B data
- Make decisions and predictions



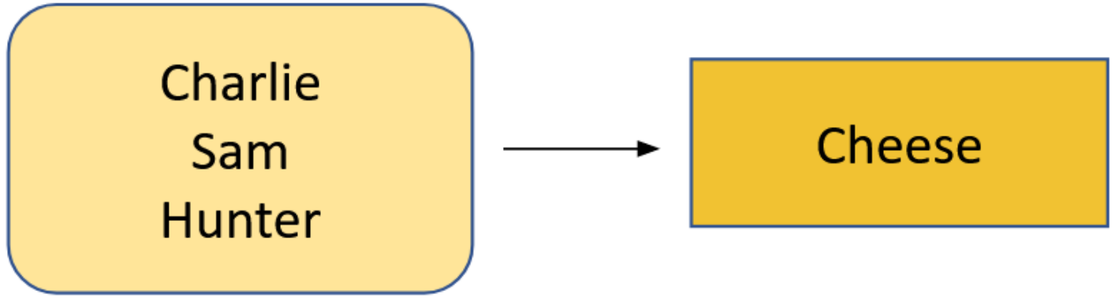
A/B design

- Impact of a condition:
  - Time to eat Cheese or Pepperoni pizza
- Group A: Cheese pizza
- Group B: Pepperoni pizza
- Between-Subjects design



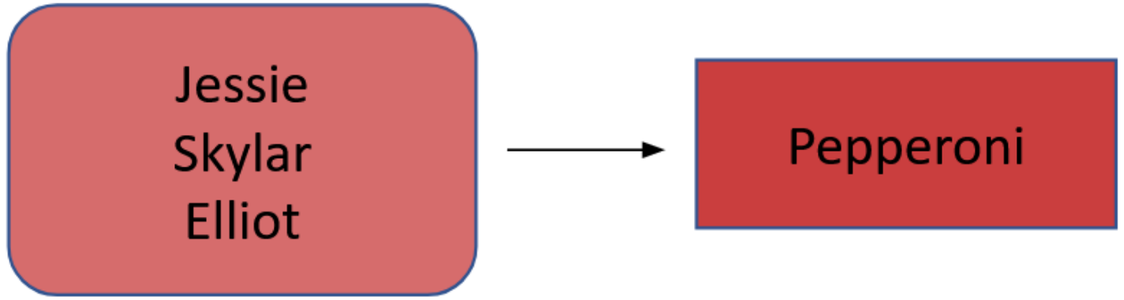
Control group

Group A



- Control group

Group B



- Experimental group

Data frame organization

Wide Format

- Groups in different columns
- NAs because of between-subjects design

ID	Cheese	Pepperoni
01	5.21	NA
02	3.75	NA
03	6.32	NA
04	NA	6.53
05	NA	7.01
06	NA	6.98

Long Format

- Groups denoted in one column
- **Preferred format**

ID	Topping	Time
01	Cheese	5.21
02	Cheese	3.75
03	Cheese	6.32
04	Pepperoni	7.64
05	Pepperoni	7.98
06	Pepperoni	5.62

Wide to long format

ID	Cheese	Pepperoni
01	5.21	NA
02	3.75	NA
03	NA	6.53
04	NA	7.01

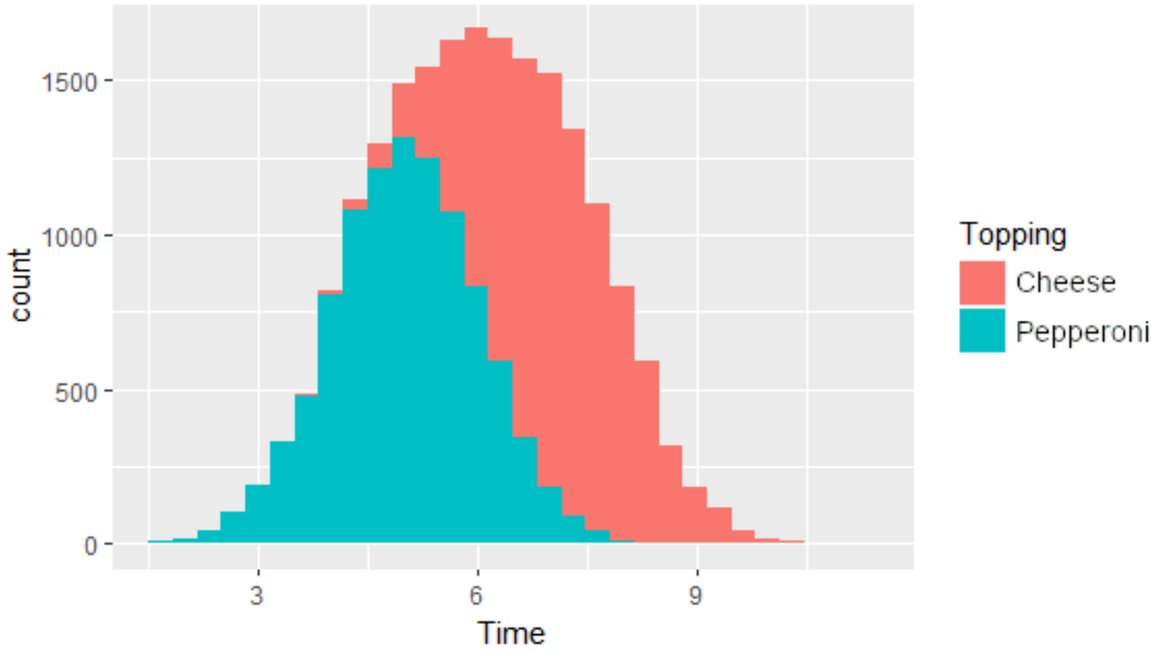
ID	Topping	Time
01	Cheese	5.21
02	Cheese	3.75
03	Pepperoni	7.64
04	Pepperoni	7.98

```
library(tidyr)
longdf <- Pizza %>% pivot_longer(cols = c("Cheese", "Pepperoni"),
           names_to = "Topping", values_to = "Time") %>%
           na.omit()
```

Visualize groupings

```
library(ggplot2)

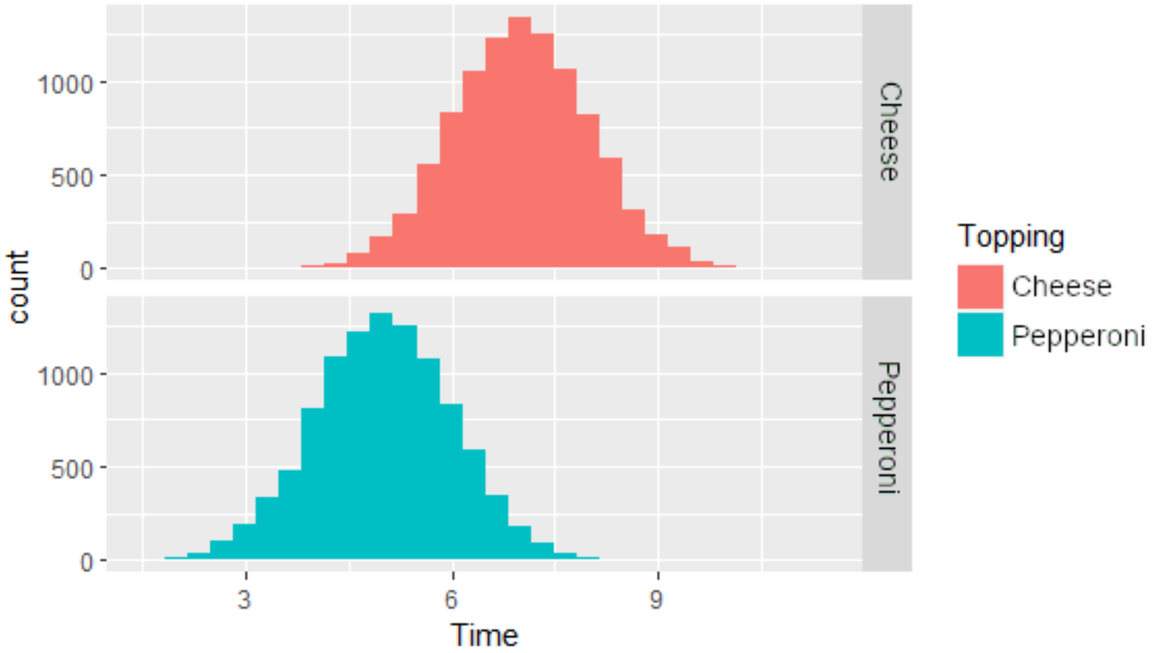
ggplot(Pizza, aes(x = Time,
                  fill = Topping)) +
geom_histogram()
```



Separate groupings

```
library(ggplot2)

ggplot(Pizza, aes(x = Time,
                  fill = Topping)) +
geom_histogram() + facet_grid(Topping~.)
```



A/B test hypotheses

Compare measure between groups

- Cheese pizza is enjoyed a different amount than Pepperoni pizza
- Cheese pizza is eaten at a different speed than Pepperoni pizza

Difference between conditions

Relationship of measure in groups

- Enjoyment is associated with speed of eating
- Both Cheese and Pepperoni groups: enjoyment of pizza is associated with speed of eating pizza
- One group (Cheese): enjoyment of Cheese pizza is associated with speed of eating Cheese pizza

Trend between measures

Let's practice!

A/B TESTING IN R

Considerations in A/B testing

A/B TESTING IN R



Lauryn Burleigh
Data Scientist

datacamp

A/B testing considerations

Only use A/B if...

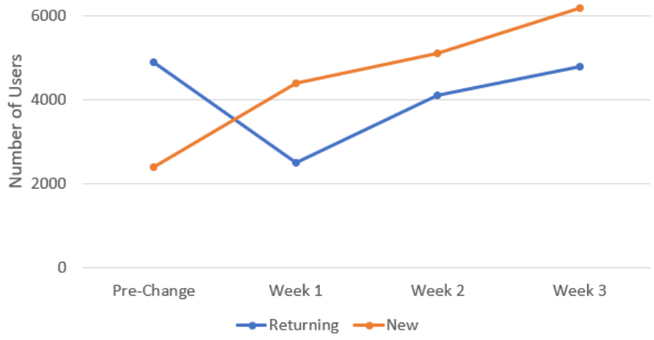
- Subjects/traffic are meaningful
- Time available for design and tests
- Clear hypothesis

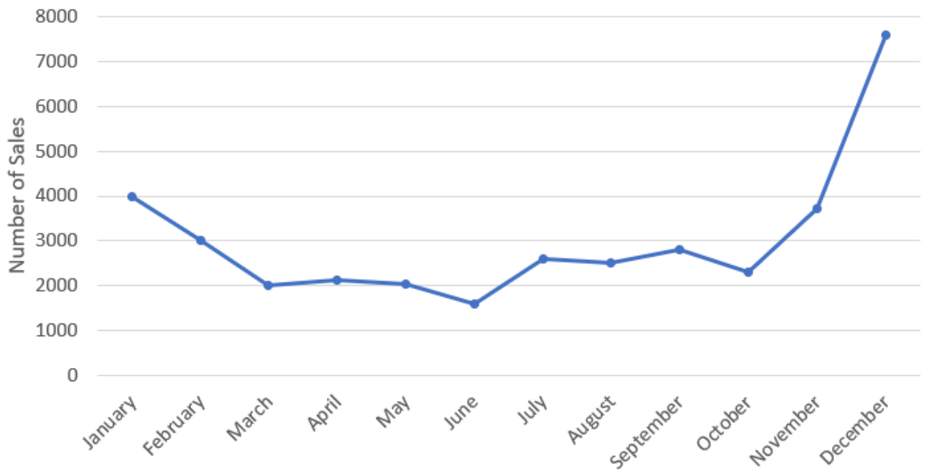
A/B test considerations

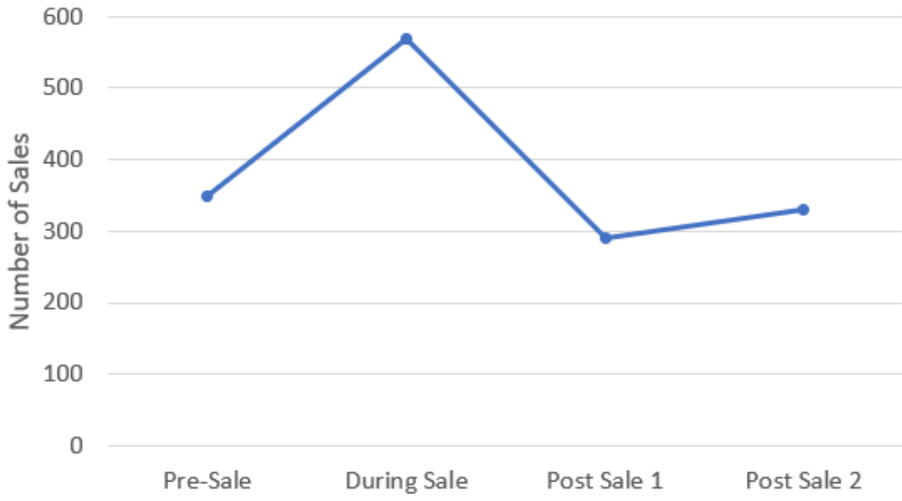
- Data fluctuations
- Number of variables
- Regression to the mean

Fluctuations in data

- For accuracy, need a representation of whole population
- Fluctuations impact results
  - Change in subjects
  - Day of week
  - Holidays
  - Public regard

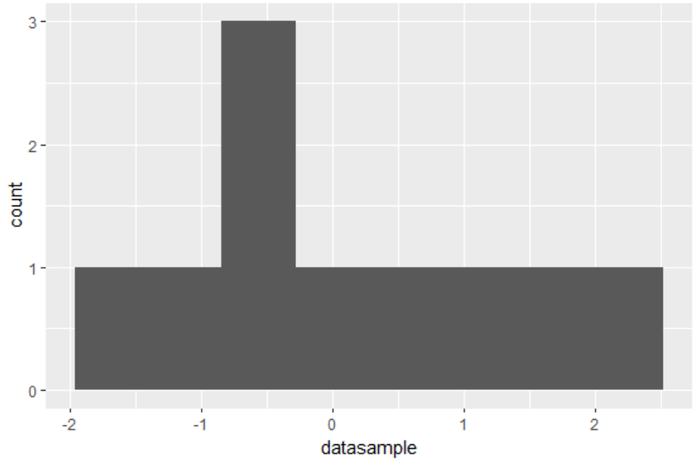


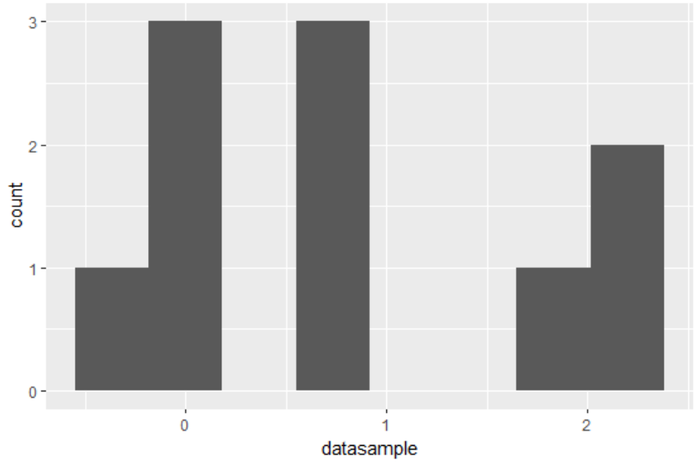


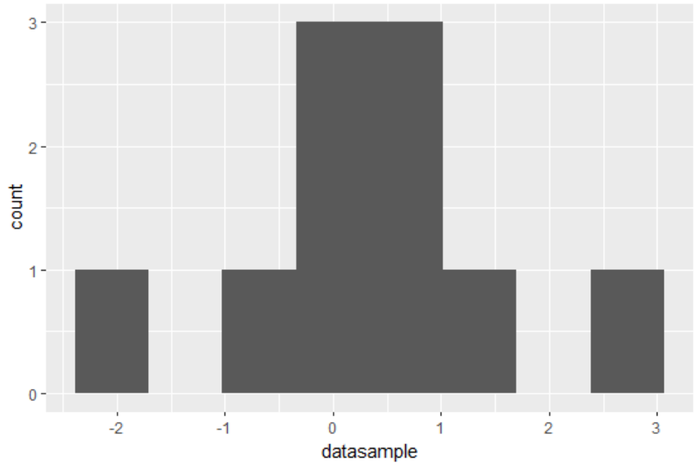


Example of fluctuations

```
normaldist <- rnorm(10000)
datasample <- sample(normaldist, 10)
ggplot() +
    aes(datasample) +
    geom_histogram(bins = 8)
```

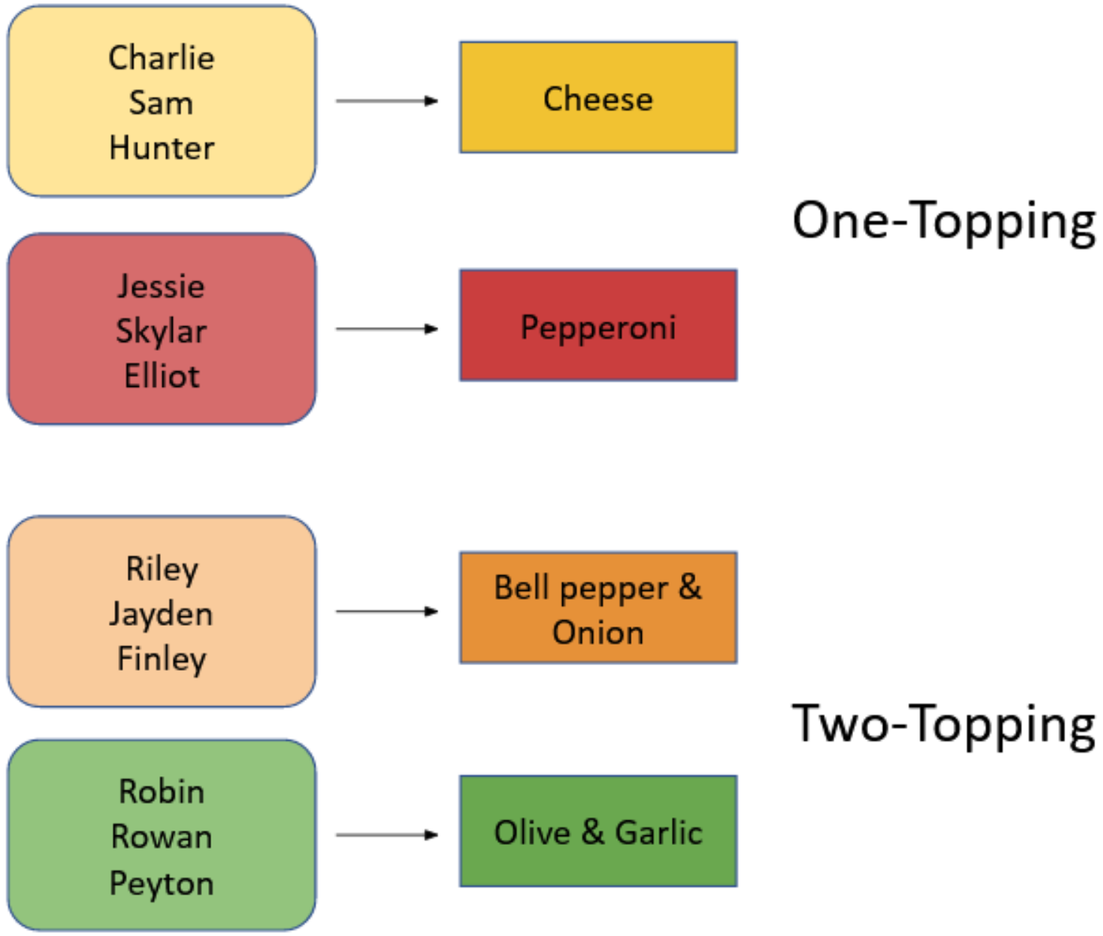






Number of variables

- One variable - ideal
- One topping/one variable
  - Assessing individual topping/condition
  - Cheese: control variable
  - Pepperoni: topping variable
- Two topping/multiple variables
  - Assessing combinations
  - Pairings: bell pepper & onions, olives & garlic
  - No control



Variables and type I error

- More variables -> more analyses
  - **Greater Type I error rate**
- Common significance: 5%
  - alpha: 0.05
  - 5% chance of a Type I error
- Confidence level: 100 - significance
  - Used to calculate family-wise error rate
    - 1 - Probability of no false positives

Calculate family-wise error rate

- Significance: 5%
- Confidence level: 95%
- Tests to run: 10

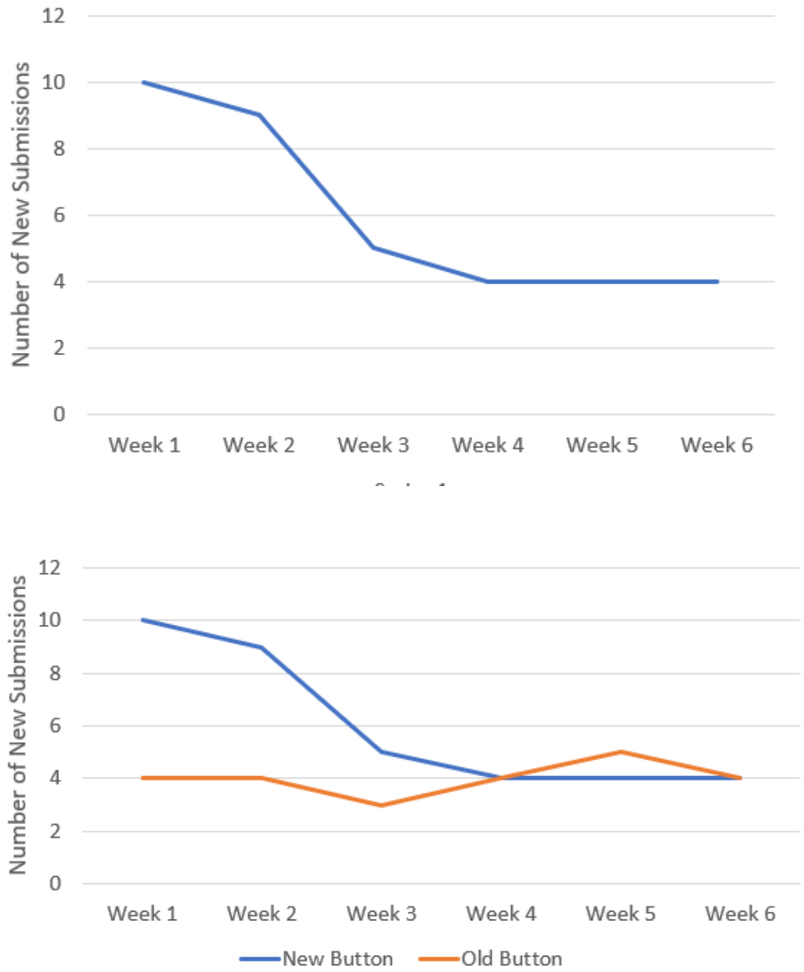
```
1 - (1-0.05)^10
1 - (0.95)^10
```

```
0.40126306076
```

40%

Regression to the mean

- Extreme values averaging out with additional data
- Type I error risk
- Compare to a control group



Regression to the mean

- Small sample sizes are inaccurate
- Type I error risk
- More data -> true mean

ID	Enjoyment [1-10]
01	1
02	10
03	9

```
mean(c(1, 10))
```

```
5.5
```

```
mean(c(1, 10, 9))
```

```
6.6666
```

Let's practice!

A/B TESTING IN R

Power and sample size

A/B TESTING IN R

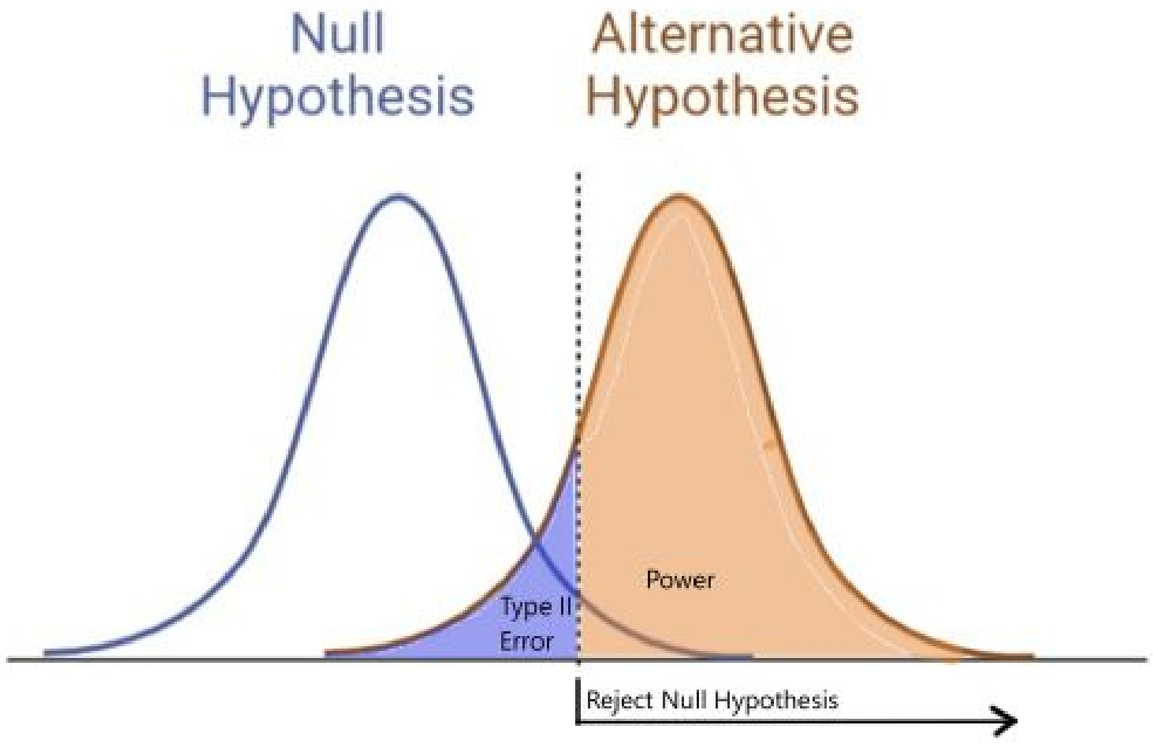


Lauryn Burleigh
Data Scientist

datacamp

Power defined

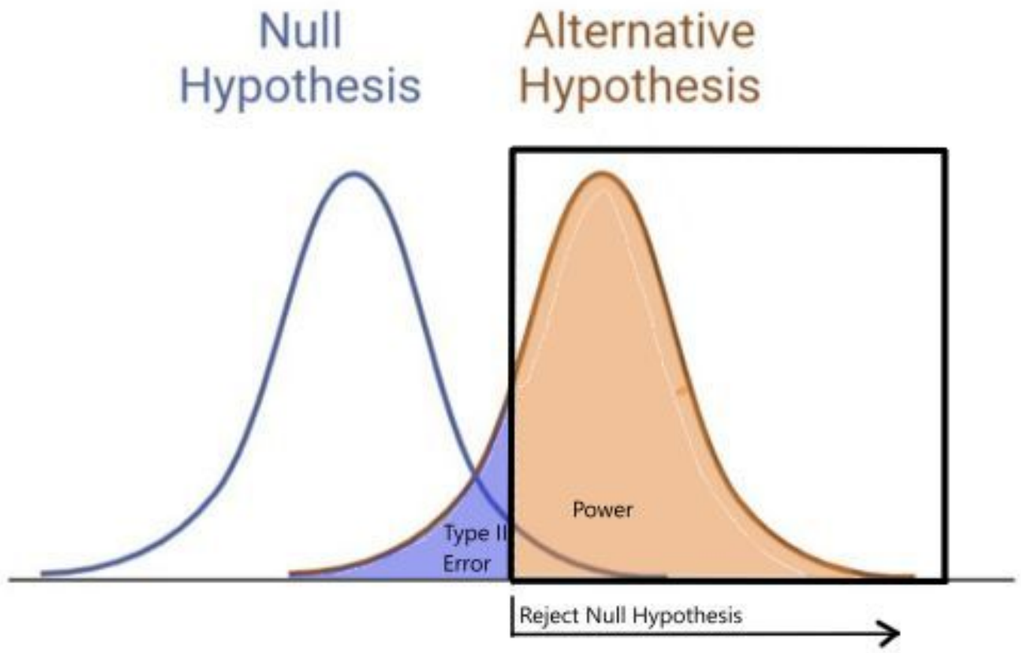
- Probability of rejecting the null hypothesis when **false**
- Not making a Type II error of not rejecting the null
- Ideal: small Type II error-rate, large power



Power benefits

Test usefulness

- Rejecting null hypothesis when it should be

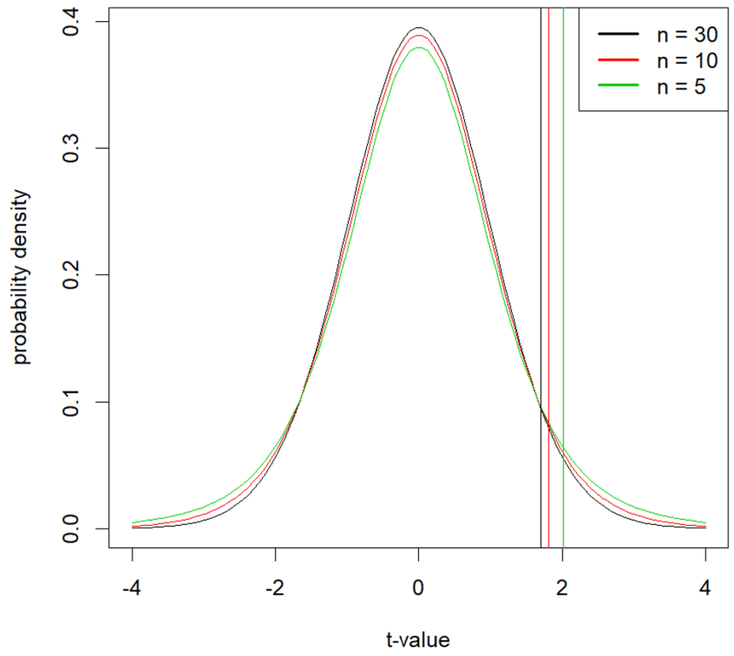


Power

- Determine sample size needed
- Check if results can be trusted

Sample size

- Estimated effect size
- Estimated power (commonly 0.8)
- Alpha (commonly 0.05)

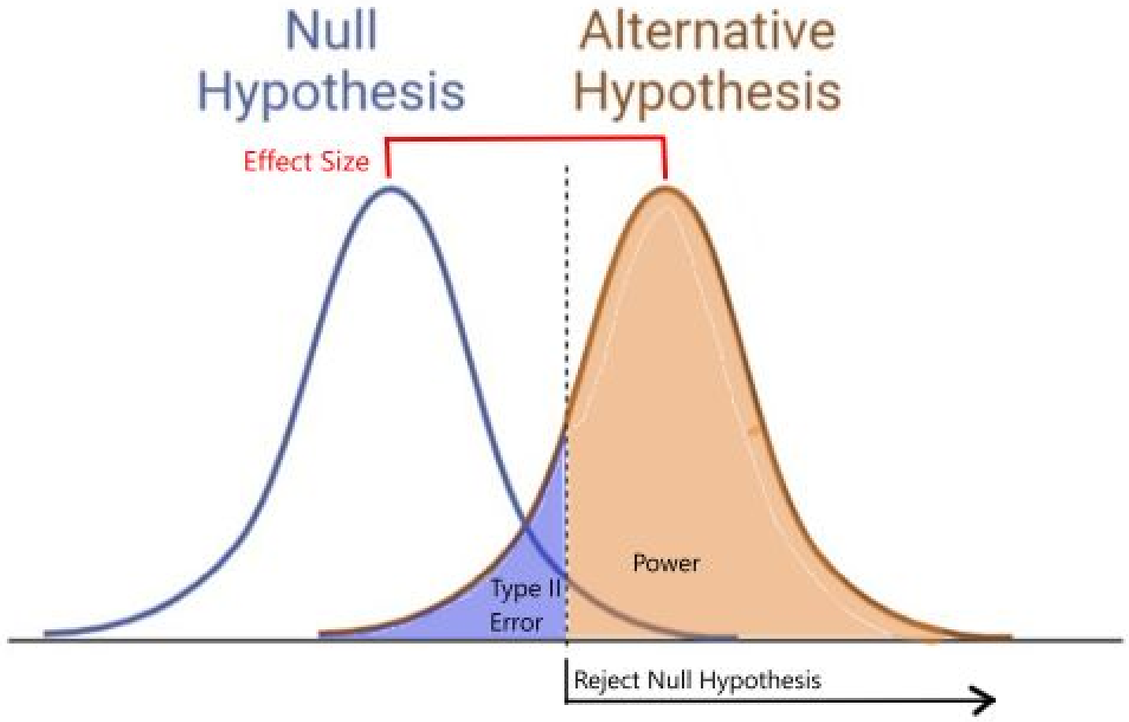



```
library(pwr)
pwr.t.test(d = .8, power = 0.8,
           sig.level = 0.05,
           type = "one.sample",
           alternative = "two.sided")
```

```
     One-sample t test power calculation
              n = 14.30276
              d = 0.8
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Effect size

- Expected size of effect
- Mean of control group - mean of experimental group



Prior to analysis

Find effect size with:

- Background information
- Preliminary data

After analysis

Find effect size with:

- Full data set

Power analysis of test

Higher power = higher probability to correctly reject null hypothesis

Three aspects needed:

- Sample size
- Effect size
- Alpha

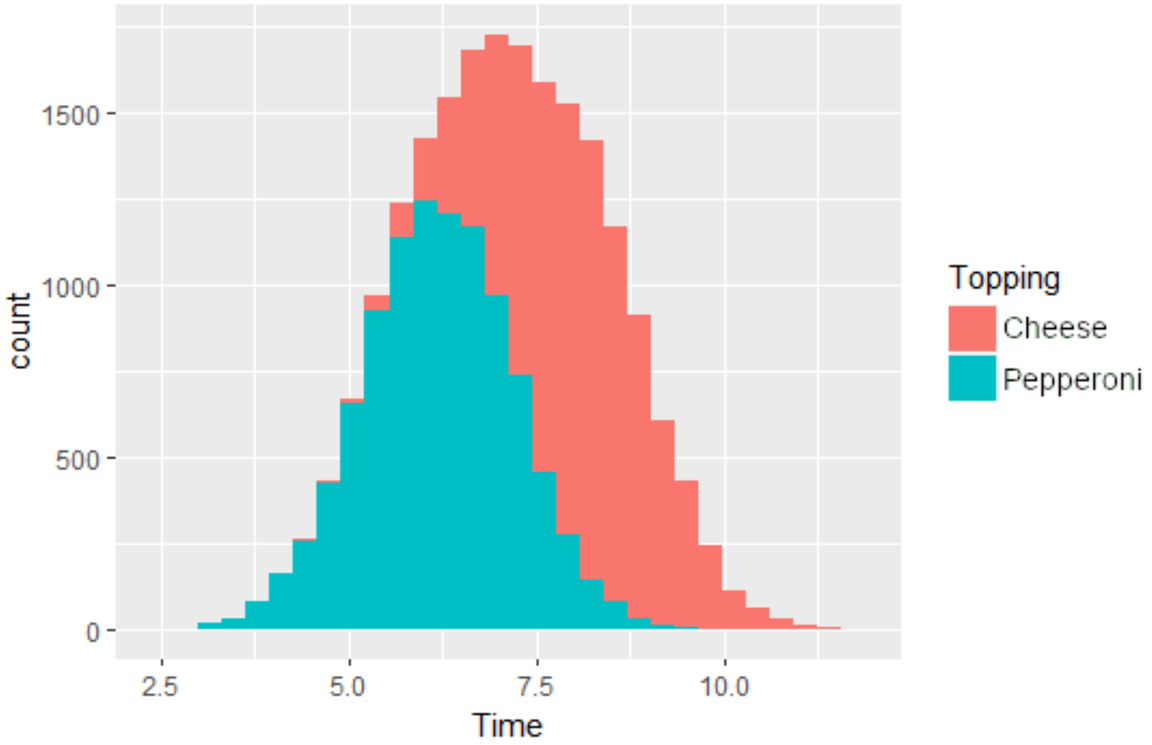
```
library(pwr)
pwr.t.test(n = 20, sig.level = 0.045,
           d = .81, type = "one.sample")
```

```
     One-sample t test power calculation
              n = 20
              d = 0.81
      sig.level = 0.045
          power = 0.9223189
    alternative = two.sided
```

Pizza distributions

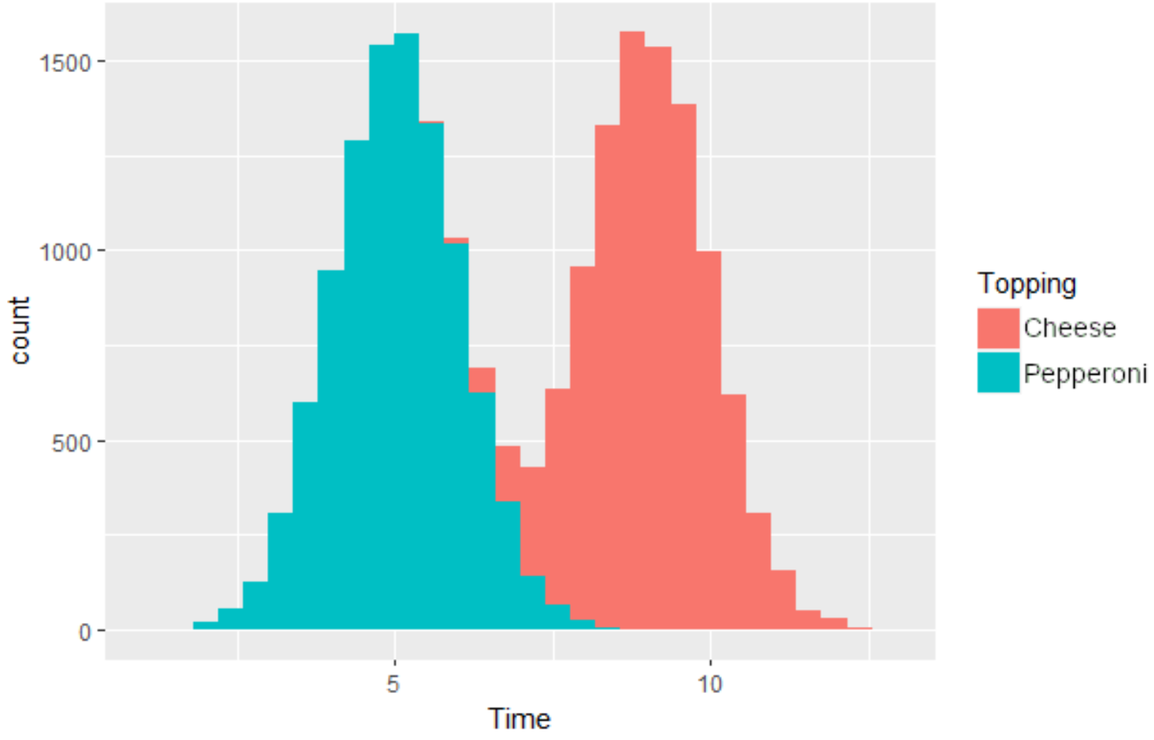
Similar distributions

No significant difference

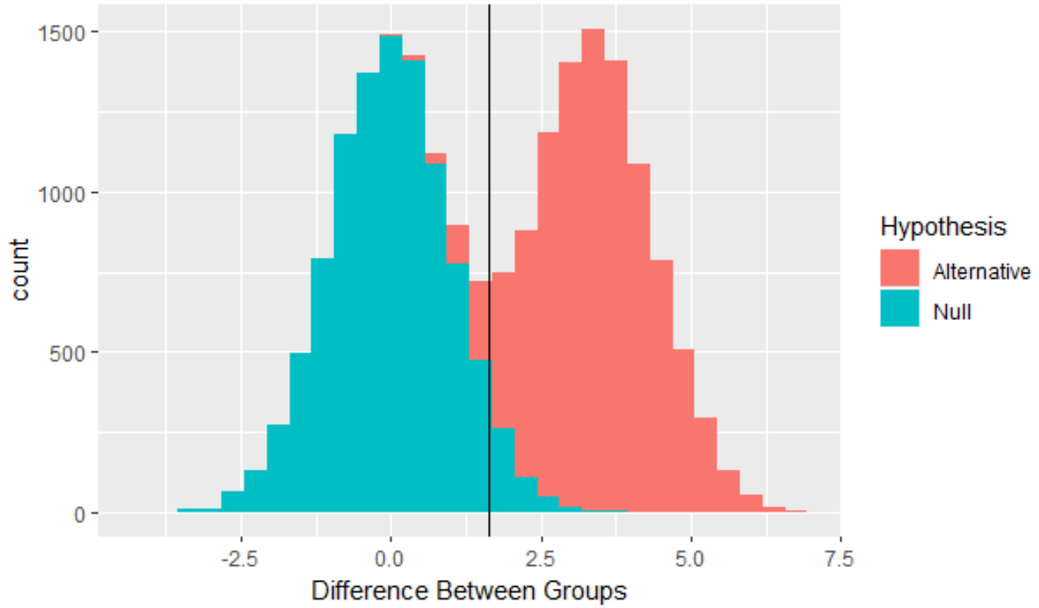


Different distributions

Likely significant difference



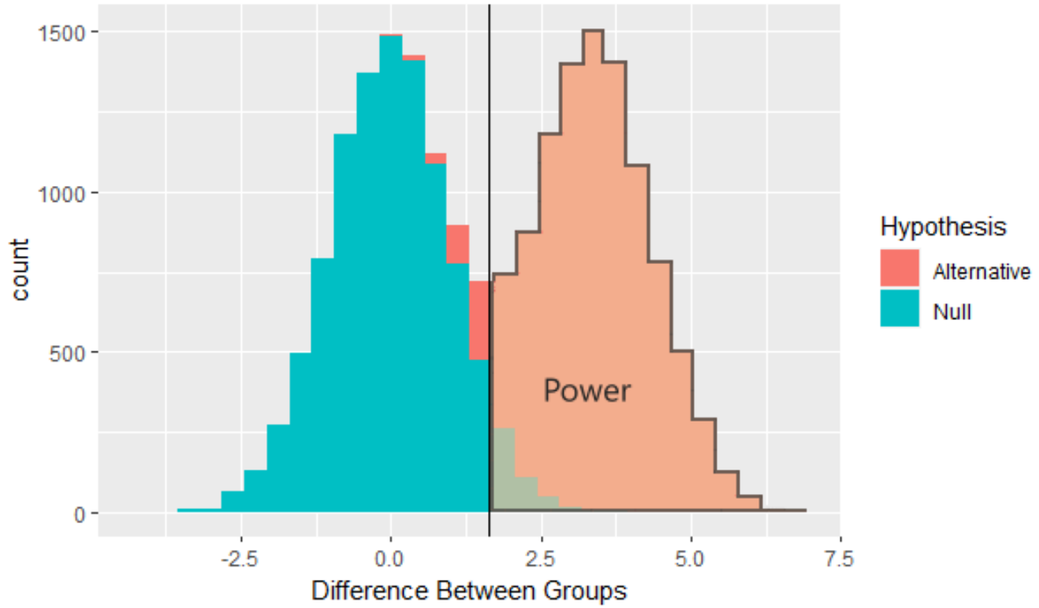
Pizza hypotheses




```
library(ggplot2)
ggplot(HypDists,
       aes(x = Time, fill = Hypothesis)) +
  geom_histogram() +
  xlab("Difference Between Groups") +
  geom_vline(xintercept = 1.64)
```

- Similar: left of rejection value
- Different: right of rejection value

Pizza power




```
library(ggplot2)
ggplot(HypDists,
       aes(x = Time, fill = Hypothesis))
  geom_histogram() +
  xlab("Difference Between Groups") +
  geom_vline(xintercept = 1.64)
```

- Similar: left of rejection value
- Different: right of rejection value
- Power: probability of not incorrectly finding same topping distributions (Type II error)

Let's practice!

A/B TESTING IN R