PSY 503: Foundations of Statistical Methods in Psychological Science

Models, The Standard Normal, and Z-Scores

Jason Geller, Ph.D. (he/him/his)

Princeton University

2022-09-28

Knowledge Check

Go to **www.menti.com/al1sciuuwgz7**

Define prior probability

The base rate

The probability of expected outcome before making any observations.

the study of random processes

Probability/belief that a specific outcome will happen prior to collecting new evidence/data.

Prior probability is the marginal likelihood of your unknown parameter.

P(A) (as opposed to P(A|B))-- the probability of an event not taking into account some other event that's occurred

Usually in a Bayesian sense - the belief probability of an event before being presented with data

05:00

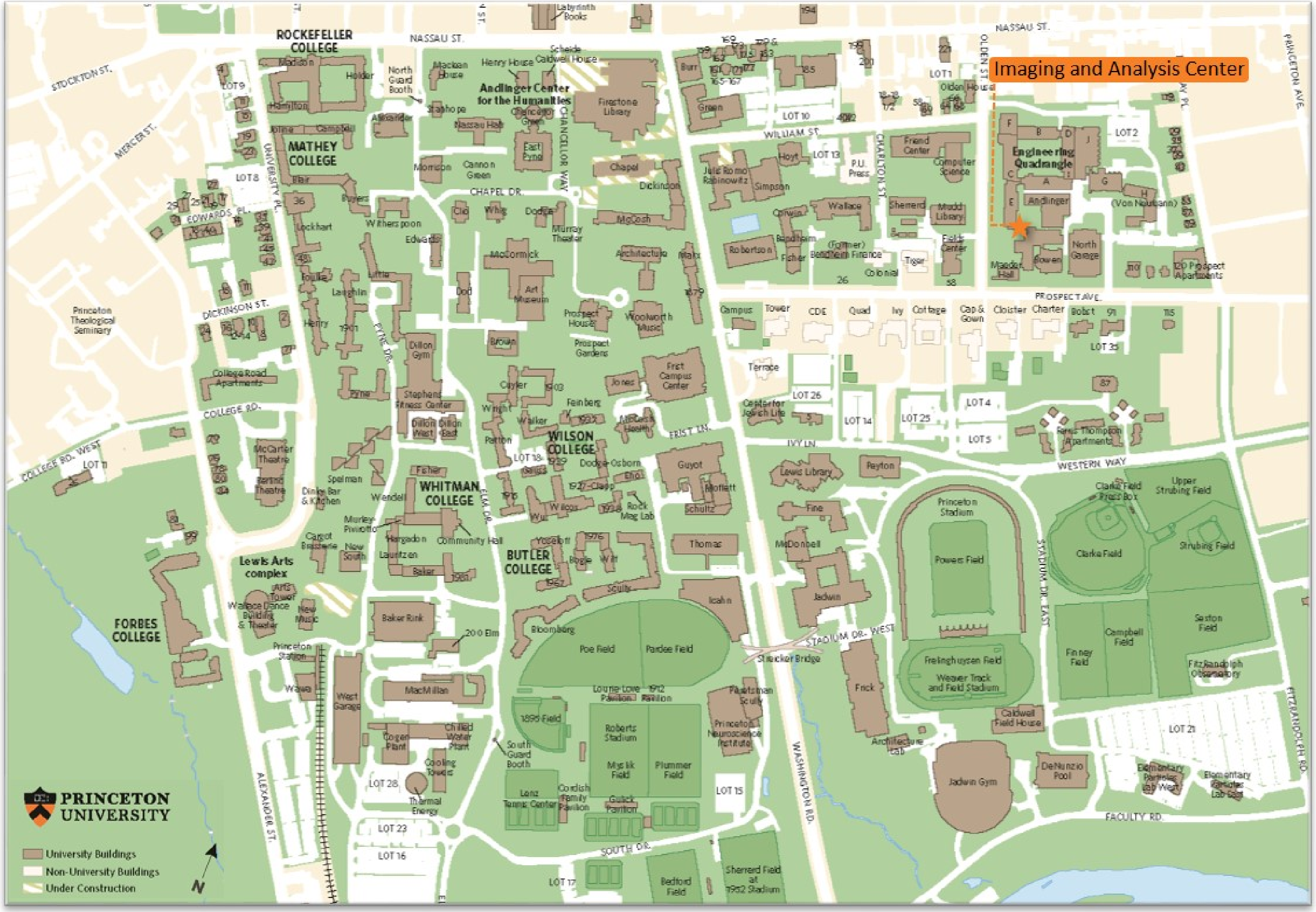
Outline

- Thinking about models
- The standard normal distribution
- Z-scores
  - How to compute Z-scores
  - Z-score practice

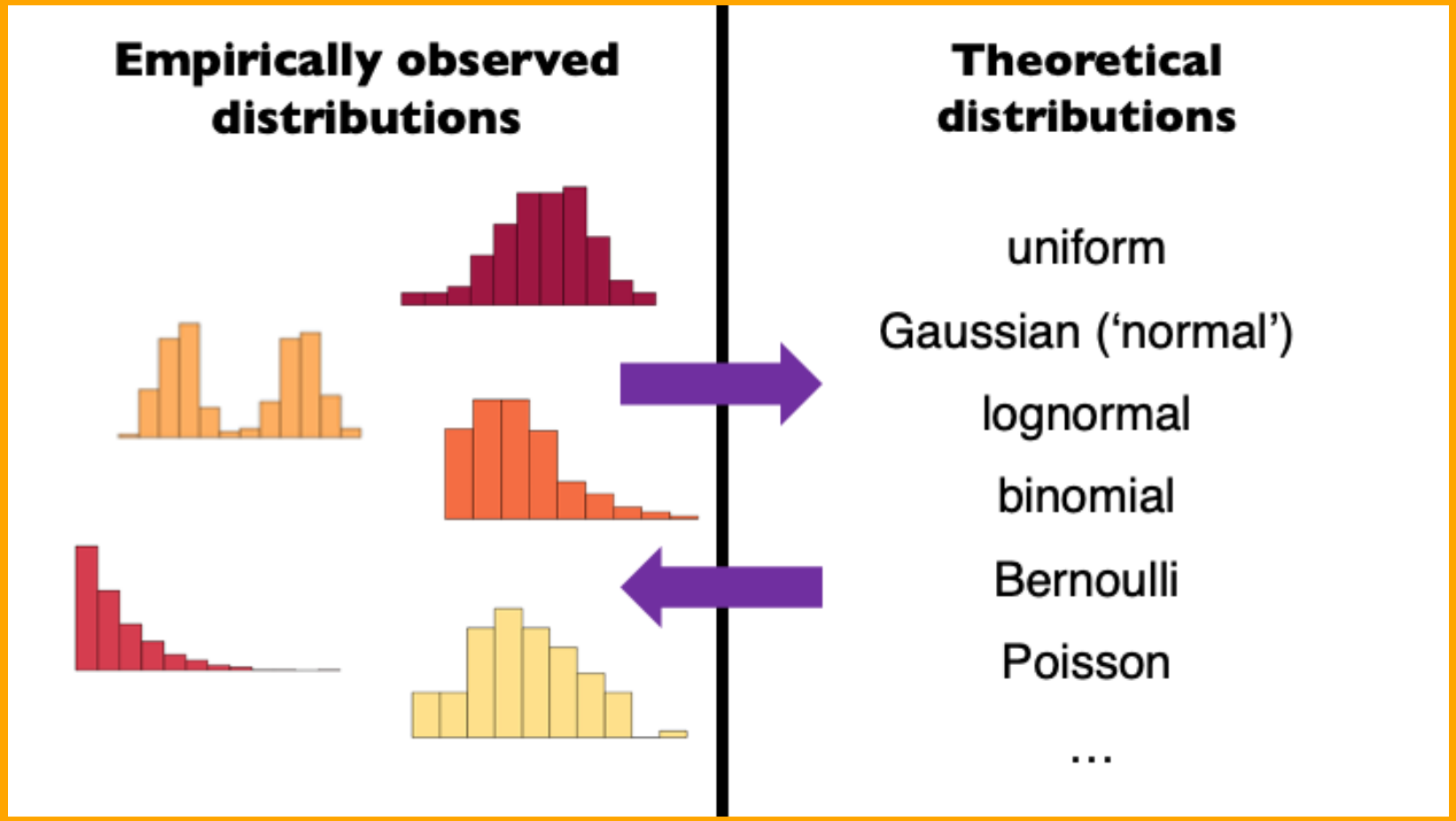
statistical modeling = "making **models** of **distributions**"

What is a model?

- Models are simplifications of things in the real world


Imaging and Analysis Center
NASSAU ST.
ROCKEFELLER COLLEGE
MATHEY COLLEGE
Andlinger Center for the Humanities
Firestone Library
Engineering Quadrangle
Andlinger
Bowen
North Garage
Friend Center
Computer Science
Mudd Library
WILLIAM ST.
PROSPECT AVE.
CHAPEL DR.
Chapel
McCosh
Art Museum
Frist Campus Center
WILSON COLLEGE
WHITMAN COLLEGE
BUTLER COLLEGE
FORBES COLLEGE
Lewis Arts complex
Dillon Gym
Princeton Stadium
Powers Field
Jadwin Gym
Princeton Neuroscience Institute
Guyot
Icahn
Frick
Lewis Library
Jadwin
Poe Field
Pardee Field
Roberts Stadium
Clarke Field
Finney Field
Campbell Field
Sexton Field
WASHINGTON RD.
FACULTY RD.
SOUTH DR.
ALEXANDER ST.
UNIVERSITY PL.
STOCKTON ST.
MERCER ST.
Princeton Theological Seminary
PRINCETON UNIVERSITY
University Buildings
Non-University Buildings
Under Construction
N

Distributions



Basic Structure of a Model

$$data = model + error$$

1. Model
2. Error (predicted - observed)
    - Use our model to predict the value of the data for any given observation:

$$\widehat{data_i} = model_i$$

$$error_i = data_i - \widehat{data_i}$$

The Golem of Prague

- The golem was a powerful clay robot
- Brought to life by writing emet (“truth”) on its forehead
- Obeyed commands literally
- Powerful, but no wisdom
- In some versions, Rabbi Judah Loew ben Bezalel built a golem to protect
- But he lost control, causing innocent deaths

Statitsical Golems

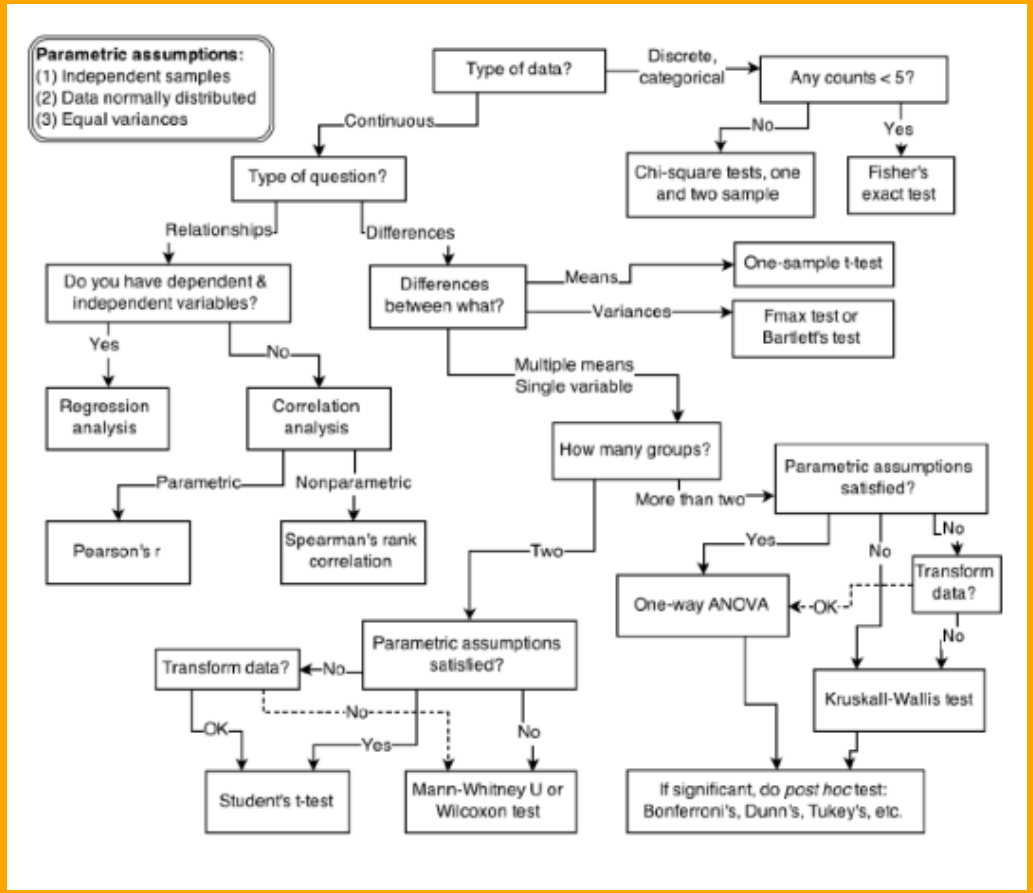
- Statistical (and scientific) models are our golems
- We build them from basic parts
- They are powerful—we can use them to understand the world and make predictions
- They are animated by “truth” (data), but they themselves are neither true nor false - The model describes the golem, not the world
- They are mindless automatons that simply run their programs
- The model doesn’t describe the world or tell us what scientific conclusion to draw—that’s on us
- We need to be careful about how we build, interpret, and apply models

Simple Example

- Experiment
  - Take 200 7-year-olds
  - Randomly assign to 2 groups
  - Control: Normal breakfast
  - Treatment: Normal breakfast + 1 packet of Smarties
  - Outcome: Age-appropriate general reasoning test
    - Norm scores: Mean 100, SD 15
- What statistical analysis do I run?

Choosing a Statitsical Model

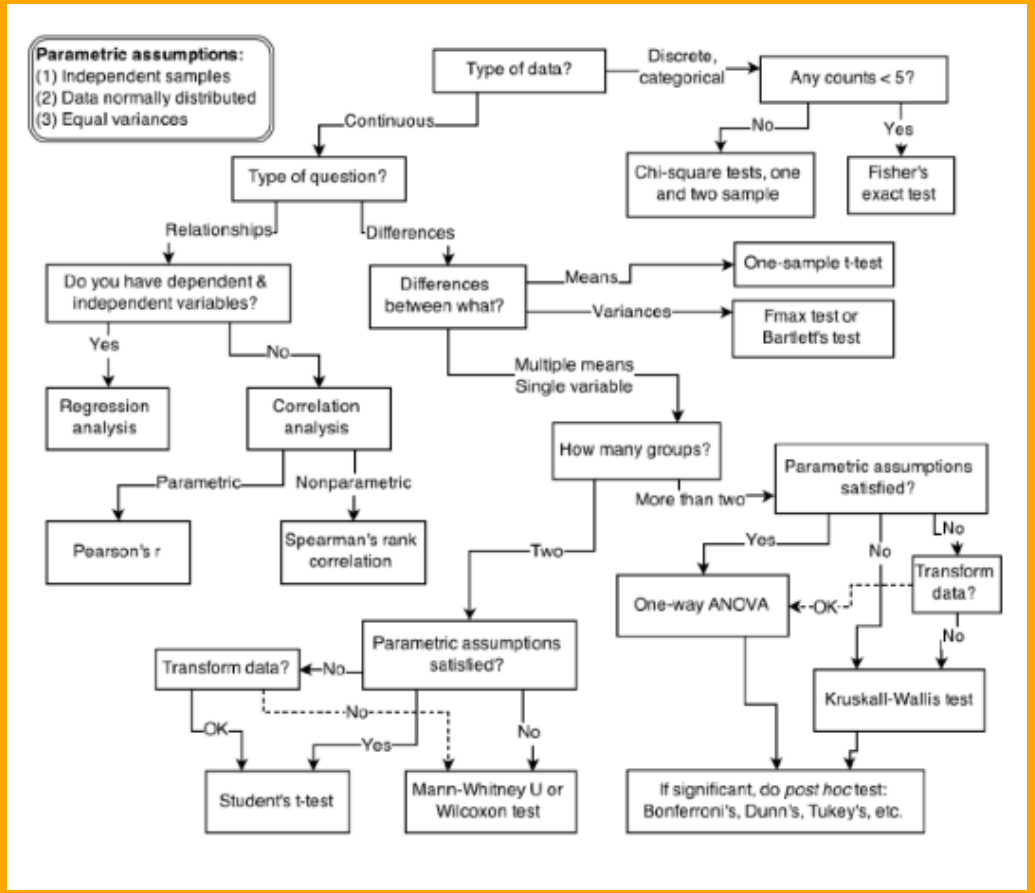
- Cookbook approach
  - My data are ordinal, what type of test do I use?



[1]Figure from Statistical Rethinking.

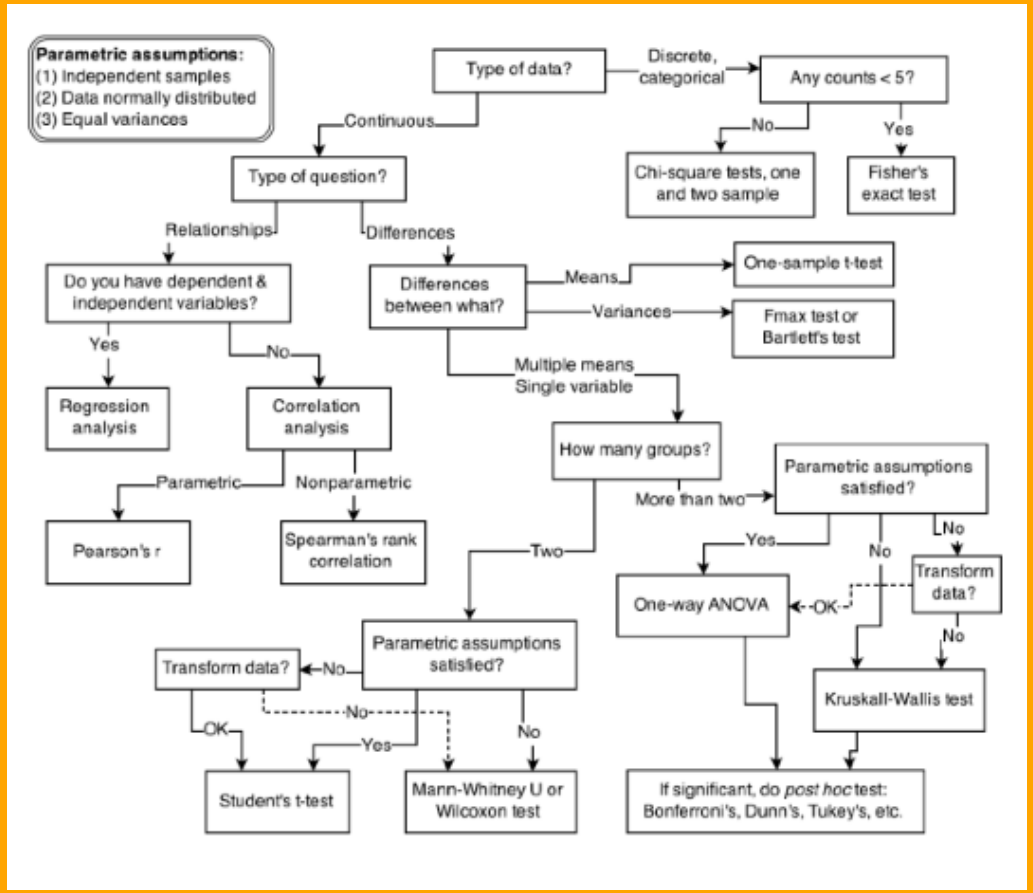
Choosing a Statistical Model

- Cookbook approach
  - My data are ordinal, what type of test do I use?
  - Every one of these tests is the same model
    - The general linear model (GLM)



Choosing a Statistical Model

- Cookbook approach
  - My data are ordinal, what type of test do I use?
  - Every one of these tests is the same model
    - The general linear model (GLM)
  - This approach makes it hard to think clearly about relationship between our question and the statistics



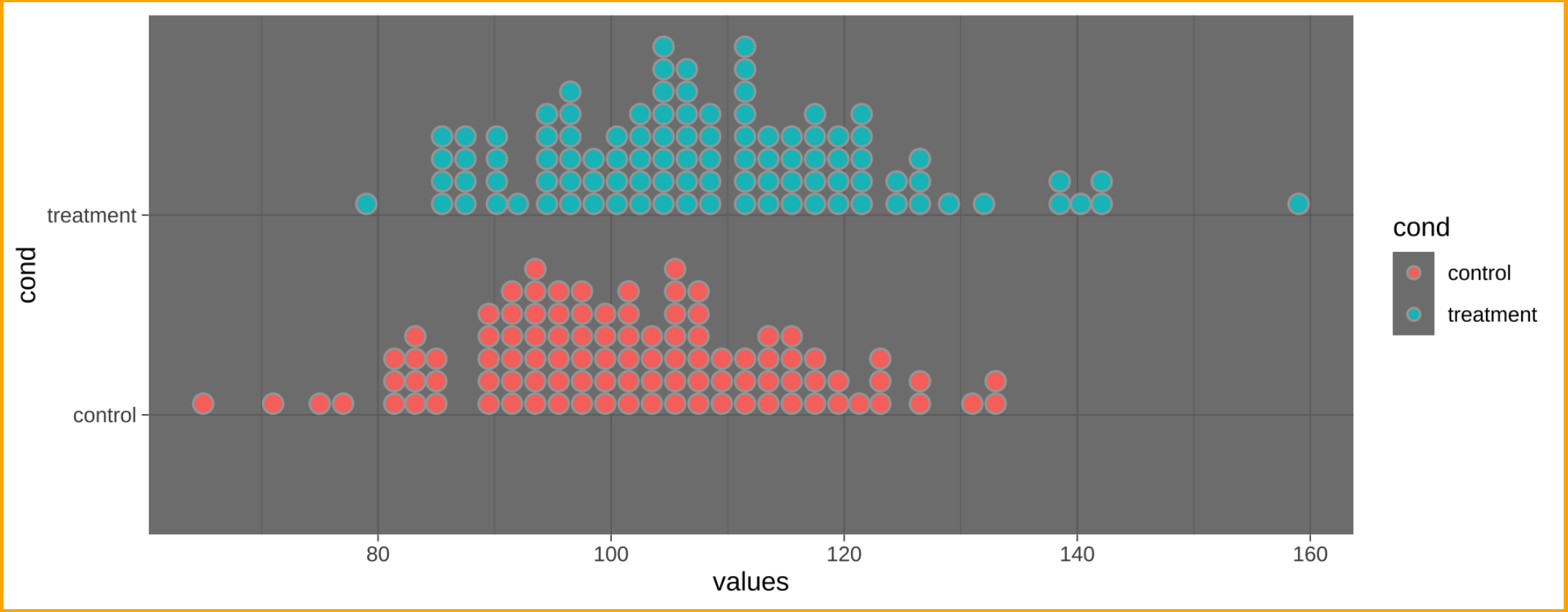
The GLM

- General mathematical framework
  - Regression all the way down
  - Highly flexible
    - Can fit qualitative (categorical) and quantitative predictors
  - Easy to interpret
  - Helps understand interrealtedness to other models
  - Easy to build to more complex models

The Data

```
library(tidyverse)
control_group= c(92, 97, 123, 101, 102, 126, 107, 81, 90, 93, 118, 105, 106, 102, 92, 127, 107, 71, 111, 93, 84


treat_group= c(99, 114, 106, 105, 96, 109, 98, 85, 104, 124, 101, 119, 86, 109, 118, 115, 112, 100, 97, 95, 112

df <- tibble(treatment=treat_group, control=control_group)

df <- df %>%
  pivot_longer(treatment:control, names_to = "cond", values_to = "values")
```

The Data


cond
treatment
control
80
100
120
140
160
values
cond
control
treatment

Bulding a Model - Notation

Small Roman Letters

- Individual observed data points
  - $y_1, y_2, y_3, y_4, ..., y_n$
    - The scores for person 1, person 2, person 3, etc.
  - y_i
    - The score for the “ith” person

Big Roman Letters

- A “random variable”
- The model for data we could observe, but haven’t yet
- Y_i
  - The model for person 1
  - The yet-to-be-observed score of person 1

Bulding a Model - Notation

Greek letters

- Unobservable parameters
    - Use to describe features of the model
- μ
    - mu
    - Pronounced “mew”
    - Used to describe means
- σ
    - Sigma
    - Pronounced “sigma”
    - Used to describe a standard deviation

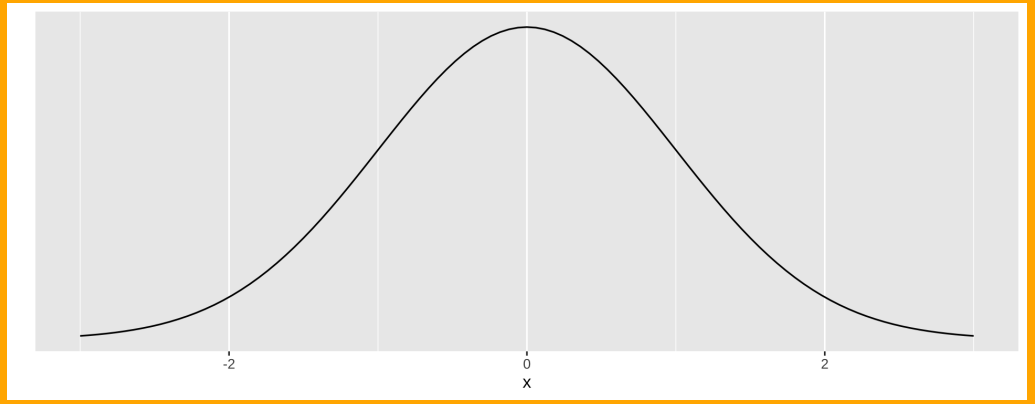
Building a Model - Normal Distribution

- Called a Gaussian model
- Many of the DVs we use are normally distributed
- If we assume a variable is at least normally distributed can make many inferences!
- Most of the statistical models assume normal distribution

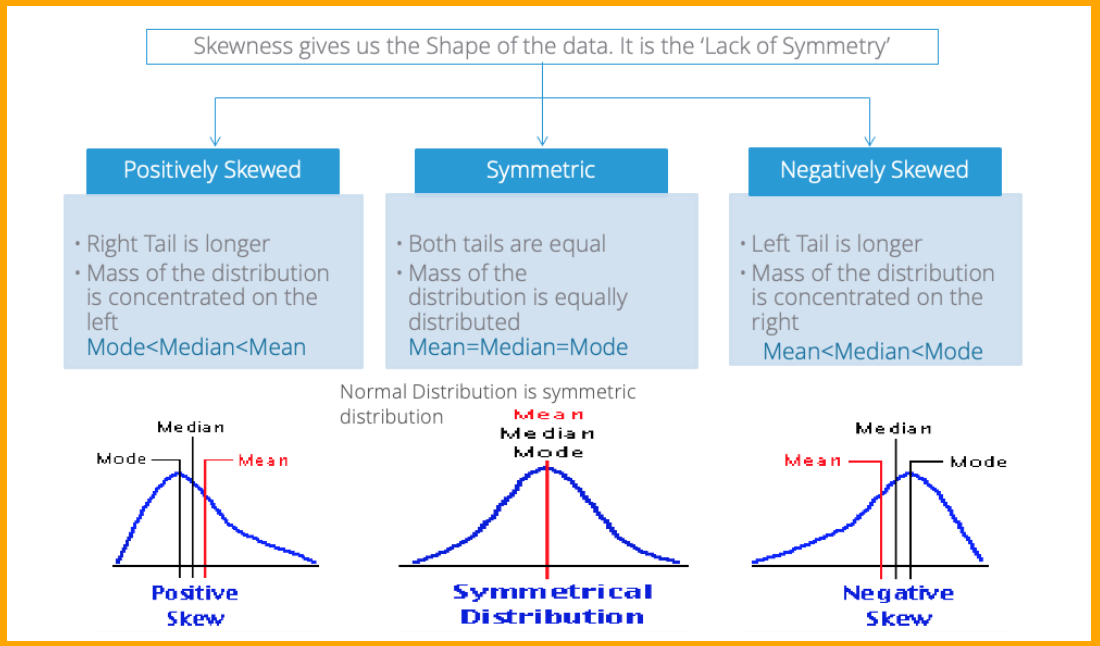
Building a Model - Normal Distribution

- Properties of a normal distribution

- Shape
  - Unimodal
  - Symmetric
  - Asymptotic



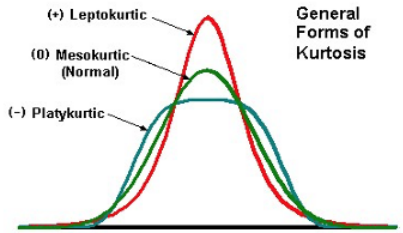
Building a Model - Normal Distribution



Building a Model - Normal Distribution

Kurtosis is defined as a measure of 'peakedness'. It is generally measured relative to Normal distribution. (Which means 'excess of kurtosis' is measured)

Mesokurtic	Leptokurtic	Platykurtic
Normal distribution is termed as mesokurtic distribution	A leptokurtic distribution has a more acute peak. (positive kurtosis)	A platykurtic distribution has a flatter peak. (negative kurtosis)



Testing for Skewness and Kurtosis

```
library(moments)

#calculate skewness
skewness(data)

#calculate kurtosis
kurtosis(data)
```

- How can we tell if bad?
  - -2 and +2 are considered acceptable in order to prove normal
  - Others suggest skewness between -2 to +2 and kurtosis is between -7 to +7

Building a Model - Normal Distribution

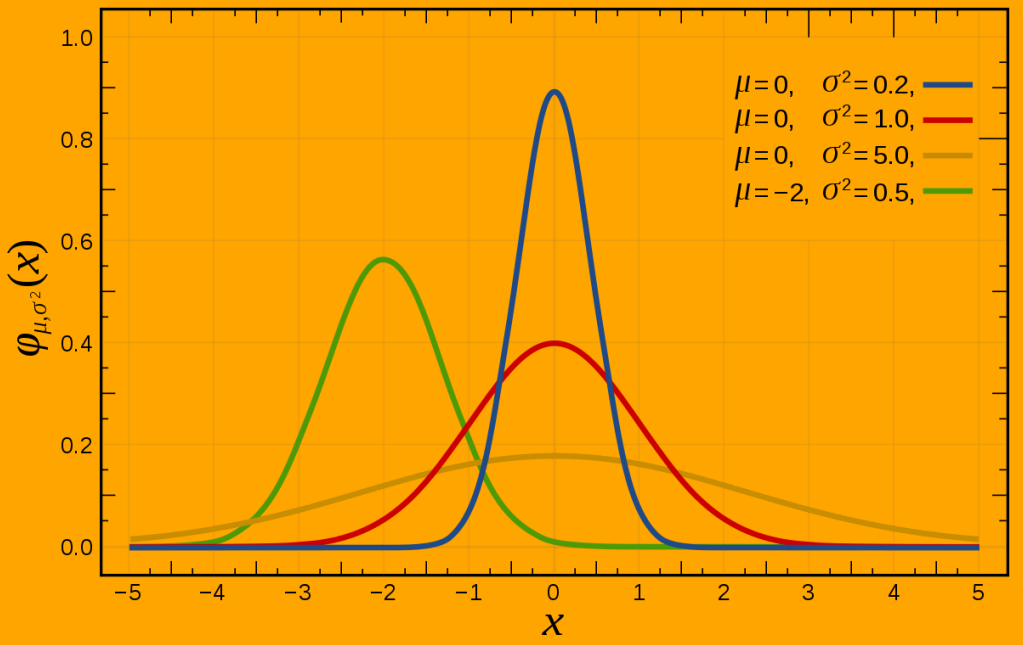
- Properties of a normal distribution
  - Empirical Rule

Building a Model - Normal Distribution

- Properties of a normal distribution
  - Empirical Rule

Building a Model - Normal Distribution

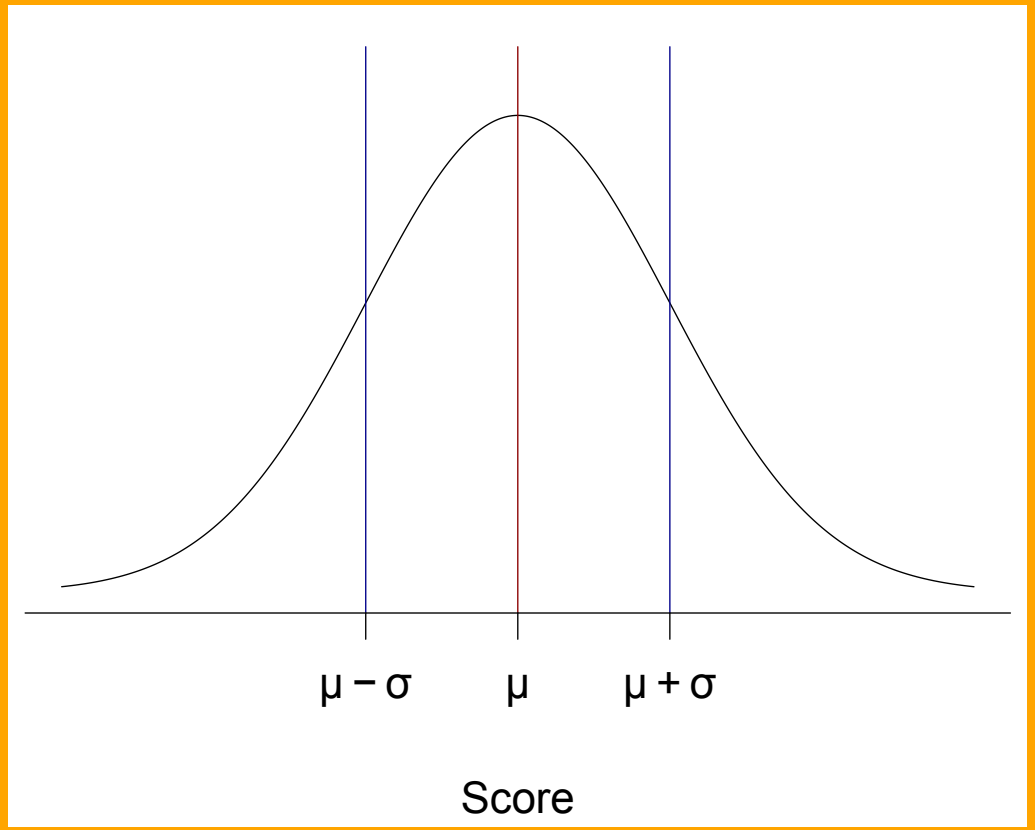
- Normal(μ, σ)
- Parameters:
  - μ Mean
  - σ Standard deviation
- Mean is the center of the distribution



Sample Mean

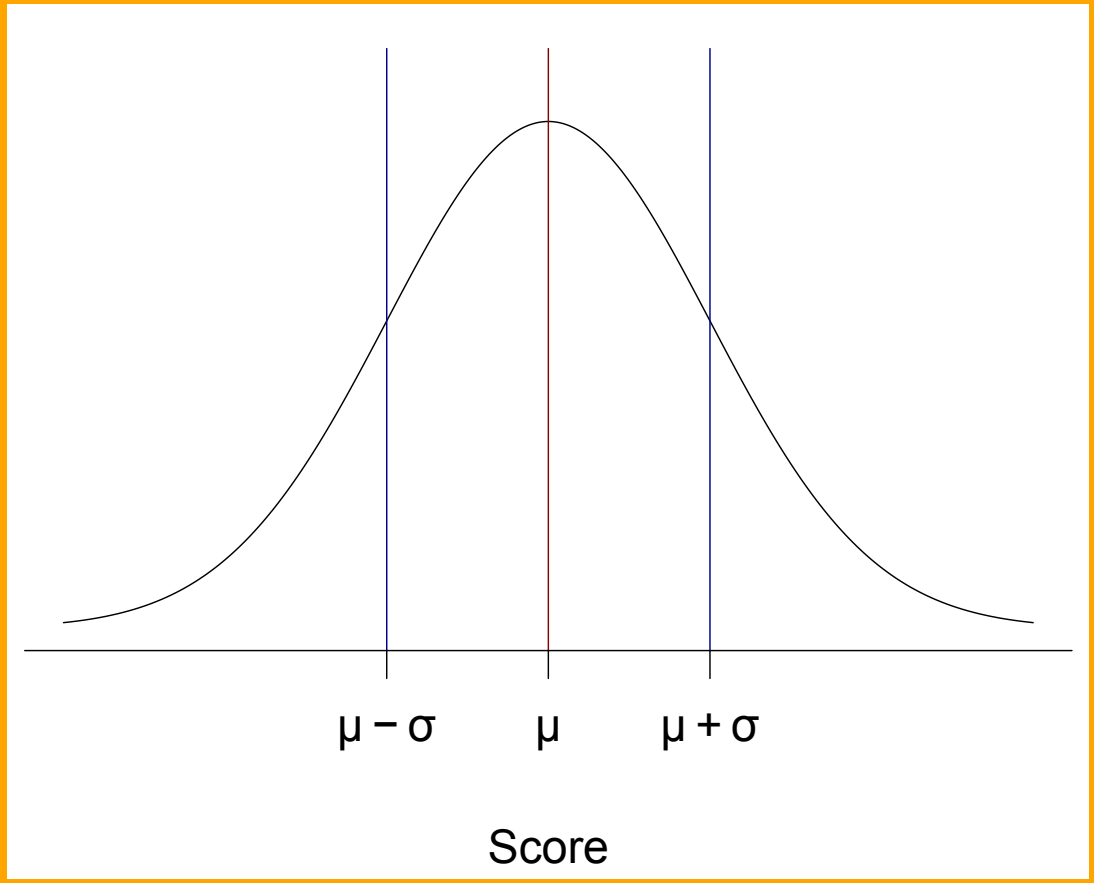
Building a Model - Normal Distribution

- Normal(μ, σ)
- Parameters:
  - μ Mean
  - σ Standard deviation
- Variance is average squared deviation from the mean Standard deviation
  - $\sigma = \sqrt{Variance}$
  - On average, how far is each point from the mean (spread)?



Building a Model - Normal Distribution

- If we say Y_1 ~ Normal(100, 15)

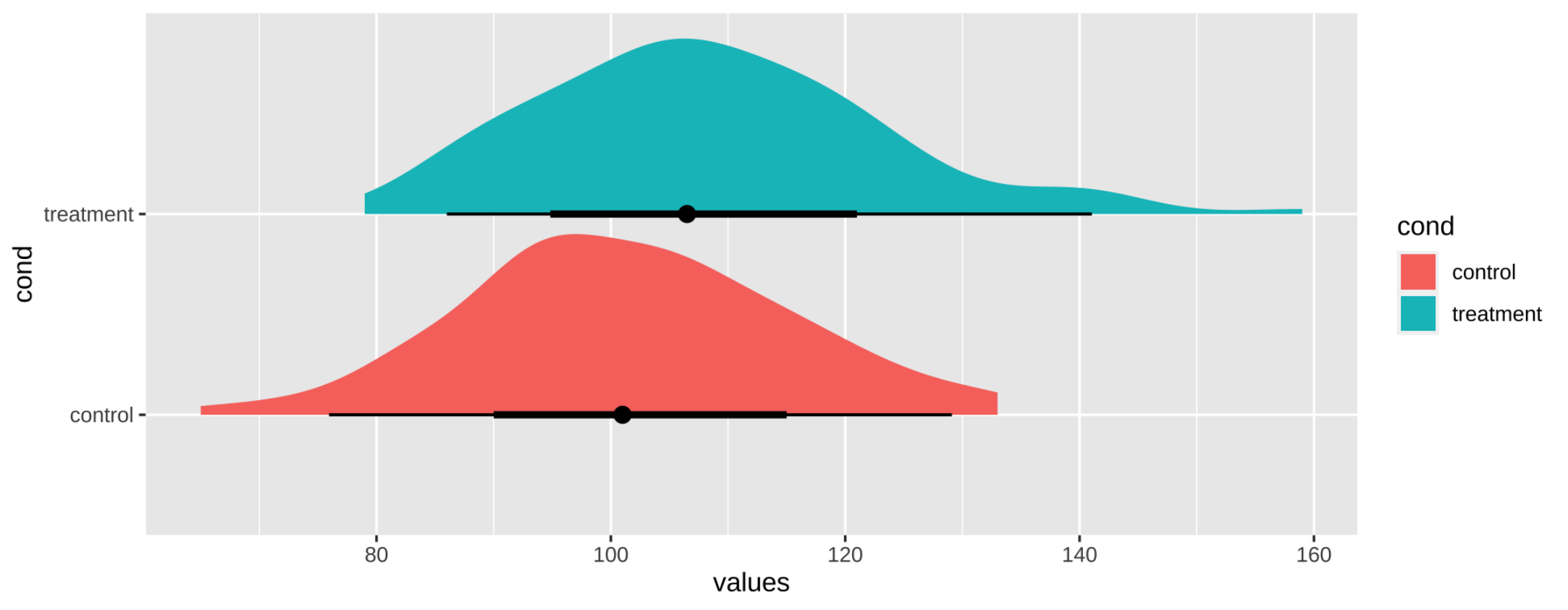


A Simple Model

- Y_1 ~ Normal(100, 15)
- Y_2 ~ Normal(100, 15)
- Y_n ~ Normal(100, 15)
- Or for all observations,
  - Y_i ~ Normal(100, 15)
- What does this model say?

1. Everyone's score comes from the same distribution
2. The average score should be around 100
3. Scores should be spread out by 15
4. Scores should follow bell-shaped curve

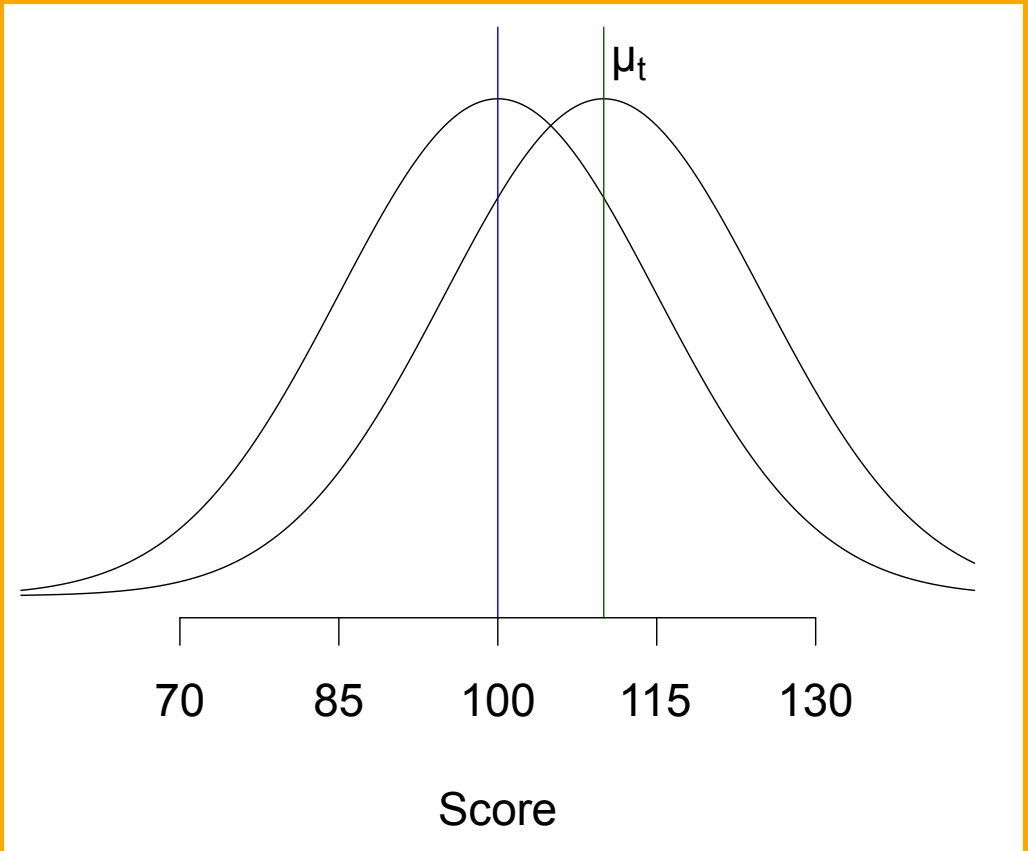
A Good Model?


treatment
control
cond
80
100
120
140
160
values
cond
control
treatment

A More Complex Model

- Allow the groups to have different means
- Add an unknown parameter
  - Something that the model will estimate
    - Y_i ,control ~ Normal(100, 15)
    - Y_i ,treatment ~ Normal($μ_t$, 15)
- What does this model say?

A More Complex Model



1. Control and treatment scores come different distributions
2. The average control group score should be around 100
3. The average treatment group score is unknown
    - Freely estimated
4. Scores should spread out by about 15 in both groups
5. Scores should follow a bell-shaped curve in both groups

Unknown Paramters

- We dont know what they are
- We need to estimate them
- Denote estimates with a hat:
  - μ_t our estimate of μt

It turns out that, for a normal distribution, the best estimate of the population mean is sample mean

μt̂ =

$$\frac{1}{n}\sum_{i=i}^{n} x_i$$

Treatment Group Sample Mean

```
mean(treat_group)
```

```
## [1] 108.43
```

Better Model?

Let's Streamline Our Notation

- Simple Model:
  - Y_i ~ Normal(100, 15)
- A More Typical Simple Model
  - Y_i ~ Normal(μ, σ)
  - μ = β_0
  - One common mean μ
  - One common SD σ

```
library(knitr)
library(broom)
#intercept only
lm(df$values~1)
```

```
##
## Call:
## lm(formula = df$values ~ 1)
##
## Coefficients:
## (Intercept)
##       104.9
```

More Complex Model

$$Y_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

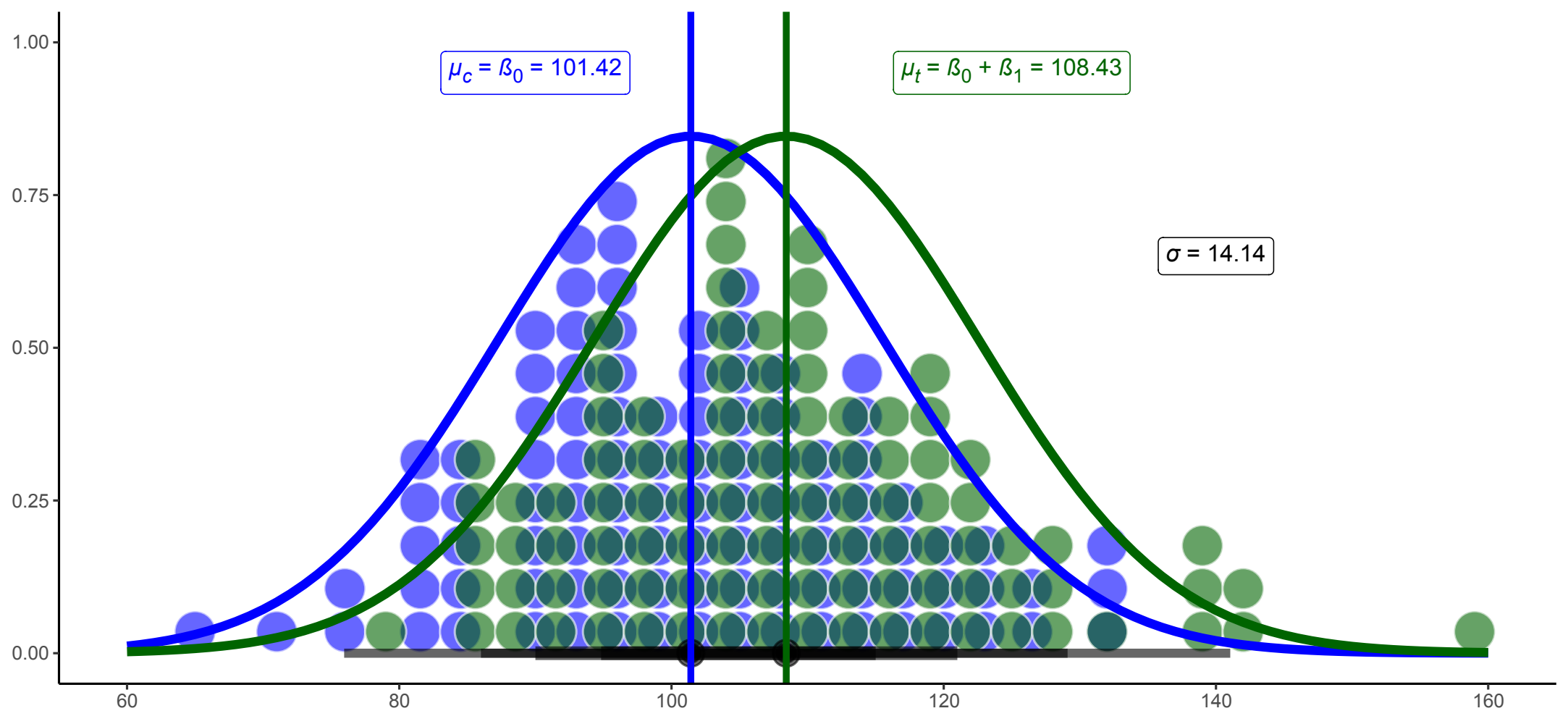
$$\mu_i = \mu control + diff * group_i$$

- Control group mean β_0
- Group mean difference β_1
- One common SD σ

```
#cond in model
lm(df$values~df$cond)
```

```
##
## Call:
## lm(formula = df$values ~ df$cond)
##
## Coefficients:
##      (Intercept)  df$condtreatment
##           101.42             7.01
```

General Linear Model



Probability and Standard Normal Distribution: Z-Scores

$$z = \frac{value - mean}{standarddeviation}$$

$$Z(x) = \frac{x - \mu}{\sigma}$$

- Z-score /standard score tells us how far away any data point is from the mean, in units of standard deviation
- Conversions
  - Solve for X

$$X = z * \sigma - \mu$$

Scaling does not change the distribution! It is just a linear transformation

Z-Scores: Example

- μ = 50 and σ = 10
- x = 43

$$Z(x) = \frac{43-50}{10}$$

Z tables

- NO MORE TABLES

Using R

- dnorm(): Z-score to density (height)
- pnorm(): Z-score to area
- qnorm(): area to Z-score

Using R

- `pnorm(): *z*-scores to area`
  - CDF
  - $P(X >= x) or P(X <= x)$

> If you cacluated a Z-score you can find the probability of a Z-score less than(lower.tail=TRUE) or greater than (lower.tail=FALSE) by using pnorm(Z).

Using R: `pnorm`

1. What is the z-score?

2. What percentage is above this z-score?

```
mu <- 70
sigma <- 10
X <- 80
```

```
z <- (X-mu)/sigma
z
```

```
## [1] 1
```

- Above

```
pnorm(1, lower.tail = FALSE)
```

```
## [1] 0.1586553
```

Using R: `pnorm`

- Percentage below IQ score of 55?

```
mu <- 100
sigma <- 15
X <- 55
z <- (X-mu)/sigma
pnorm(z, lower.tail = TRUE)
```

```
## [1] 0.001349898
```

- Percentage above IQ score of 55?

```
mu <- 100
sigma <- 15
X <- 55
z <- (X-mu)/sigma
pnorm(z, lower.tail = FALSE)
```

```
## [1] 0.9986501
```

Using R: `pnorm`

- Percentage between IQ score of 120 and 159?

```
mu <- 100
sigma <- 15
X1 <- 159
X2<-120


pnorm(X1, mu, sigma)-pnorm(X2, mu, sigma)
```

```
## [1] 0.09116933
```

Package `PnormGC`

Suppose that you have a normal random variable X μ=70 and σ=3. Probablity X will turn out to be less than 66.

```
require(tigerstats)

pnormGC(bound=66,region="below",mean=70,sd=3, graph=TRUE)
```

Package PnormGC

What about $P(X > 69)$

```
pnormGC(bound=69,region="above",
        mean=70,sd=3,graph=TRUE)
```

Package PnormGC

- The probability that X is between 68 and 72: $P(68 < X < 72)$

```
pnormGC(bound=c(68,72),region="between",
        mean=70,sd=3,graph=TRUE)
```

Using R: `qnorm`

- qnorm(): area to z-scores
- With a mean 70 and standard deviation 10, what is the score for which 5% lies above?

```
qnorm(.05, lower.tail = FALSE)
```

```
## [1] 1.644854
```

Practice `pnorm`

Suppose that BMI measures for men age 60 in a Heart Study population is normally distributed with a mean (μ) = 29 and standard deviation (σ) = 6. You are asked to compute the probability that a 60 year old man in this population will have a BMI less than 30.

- What is the z-score?
- What is probability of a Z-score less than Z? Greater than?
- What is the probability a 60 year old man in this population will have a BMI between 30 and 40.

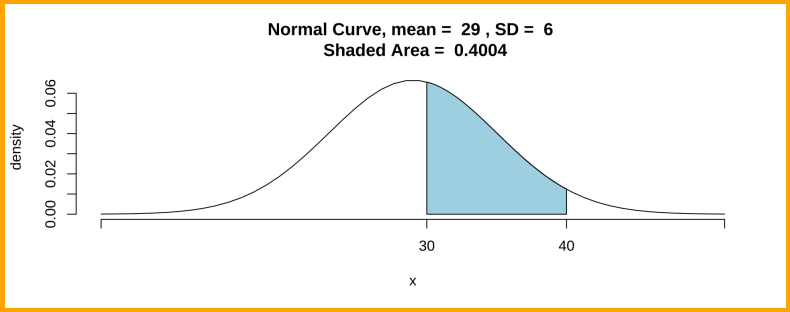
```r
mu <- 29
sigma <- 6
X <- 30
X1 <- 40
z <- (X-mu)/sigma
z2 <- (X1-mu)/sigma

pnorm(z)
```

```
## [1] 0.5661838
```

```r
pnorm(z, lower.tail = FALSE)
```

```
## [1] 0.4338162
```

```r
pnormGC(bound=c(30,40),region="between", mean=29,sd=6,graph=TRUE)
```



Practice: `qnorm`

Suppose that SAT scores are normally distributed, and that the mean SAT score is 1000, and the standard deviation of all SAT scores is 100. How high must you score so that only 10% of the population scores higher than you?

```
qnorm(.10, 1000, 100, lower.tail = FALSE)
```

```
## [1] 1128.155
```

Z-scores In Practice

- Scaling your measures so they are are comparable

```
library(datawizard)
x=c(5, 6, 7, 8, 9, 10, 15, 16)
datawizard::standardize(x)
```

```
## [1] -1.1150879 -0.8672906 -0.6194933 -0.3716960 -0.1238987  0.1238987  1.3628852
## [8]  1.6106825
## attr(,"center")
## [1] 9.5
## attr(,"scale")
## [1] 4.035556
## attr(,"robust")
## [1] FALSE
```

Measures Related to Z

- IQ
  - $\mu = 100\, \sigma = 15$
- SAT
  - $\mu = 500\, \sigma = 100$
- T-score
  - $\mu = 50\, \sigma = 10$
- New score = New SD(z) + New mean