# Prob & Stats Review

## STSCI/INFO/ILRST 3900: Causal Inference

**September 4, 2024**

# Agenda for Today

- Reminders and Announcements
- Probability and Statistics Review
- R/RStudio Intro
- Homework Check-in and Questions

# Reminders and Announcements

- HW 1 due Tuesday (September 10) by 5pm
  - Submit a PDF from RMarkdown via Canvas
- Office Hours throughout the week (see Syllabus or website)
  - Filippo: Monday 11am-12pm in Comstock 1187
  - Shira: Wednesday 3-4pm in  in Comstock 1187
  - See Ed Discussion for Zoom links/info

# Probability and Statistics Review

- Expectation
- Variance
- Conditional Expectation
- Independence
- Bernoulli Random Variables
- Law of Total Expectation
- Confidence Intervals
- Regression (OLS, logistic)

# Expectation

## (Expected Value, Population Mean, Average)

- Notation: $E(X), \mu$
- The **expected value** of a *finite* random variable

$$\mu = E(X) := \sum_{i=1}^{N} x_i \cdot P(x_i) \ \text{ where } \ P(x_i) := \text{Prob}(X = x_i)$$

# Expectation

## (Expected Value, Population Mean, Average)

- The **expected value** of a *countable* random variable, i.e. the (long run) average

$$E(X) = \sum_{i=1}^{\infty} x_i \cdot P(x_i)$$

- For $n$ independent and identically distributed (**i.i.d.**) random variables $X_1, \cdots, X_N$

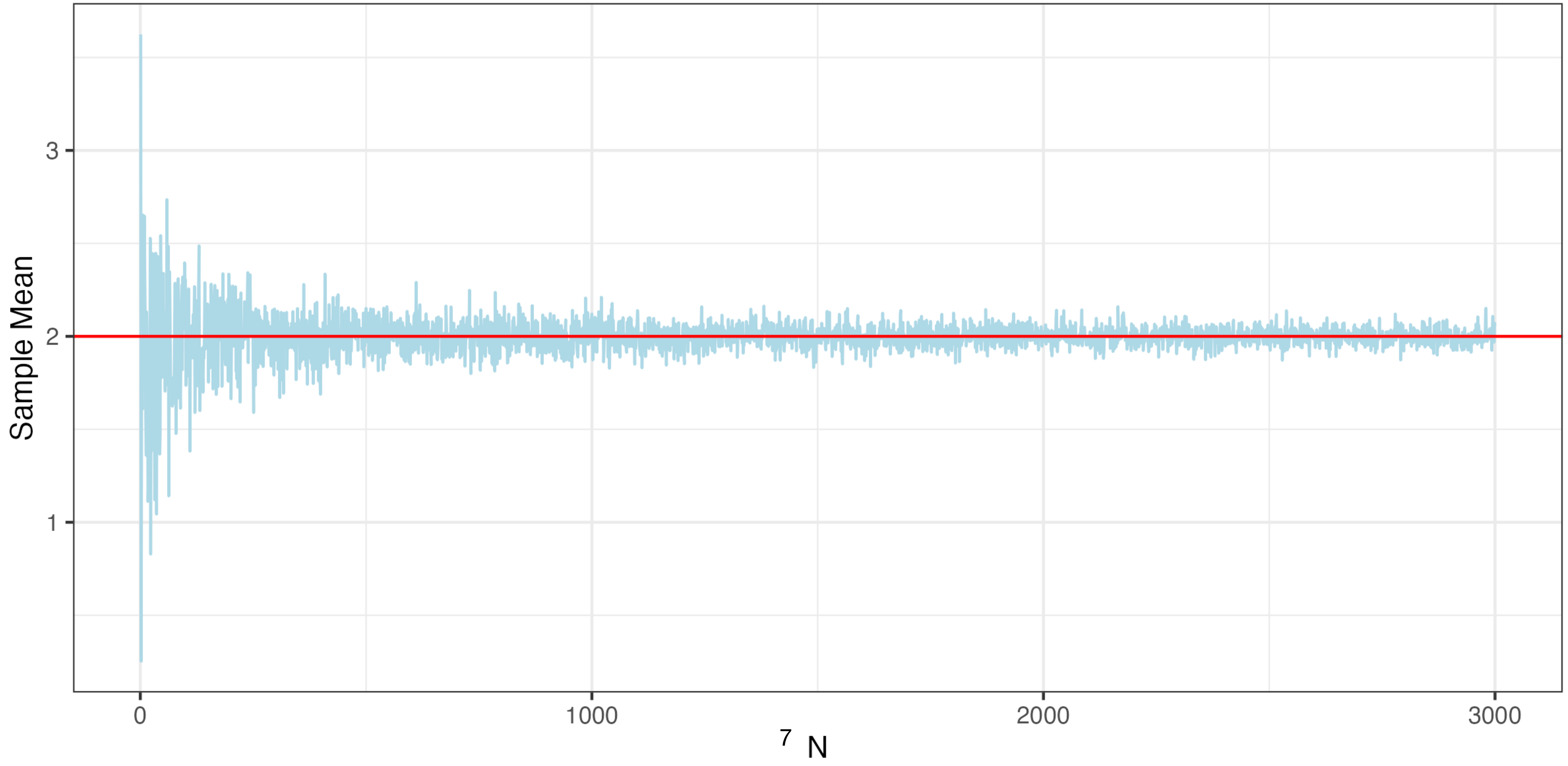
the **sample mean** is $\bar{X} = \dfrac{1}{N} \sum_{i=1}^{N} X_i$

- **Law of Large Numbers (LLN)**: the sample mean converges to the expected value (population mean) as $N \to \infty$

- Example: R (compute the sample mean for larger and larger N)

# Expectation

- $X_i$ are random draws from $\sim \mathcal{N}(2, 5)$ (a Normal r.v. with mean 2, variance 5)
- How quickly does the sample mean converge to the population mean?



# Variance

## Describes the spread of the data

- Notation: $V(X), Var(X), \sigma^2$
- Variance is the average of the squared differences from the mean
- For a random variable $X$ with expected value $\mu := E(X)$, the variance is

$$\sigma^2 = Var(X) := E\left[\left(X - \mu\right)^2\right] = E\left[X^2\right] - \mu^2$$

More explicitly, $\boxed{Var(X) = \sum_{i=1}^{n} P(x_i) \cdot (x_i - \mu)^2}$ where $P(x_i) := \text{Prob}(X = x_i)$

# Sample (Empirical) Variance

## For a finite dataset or finite sample

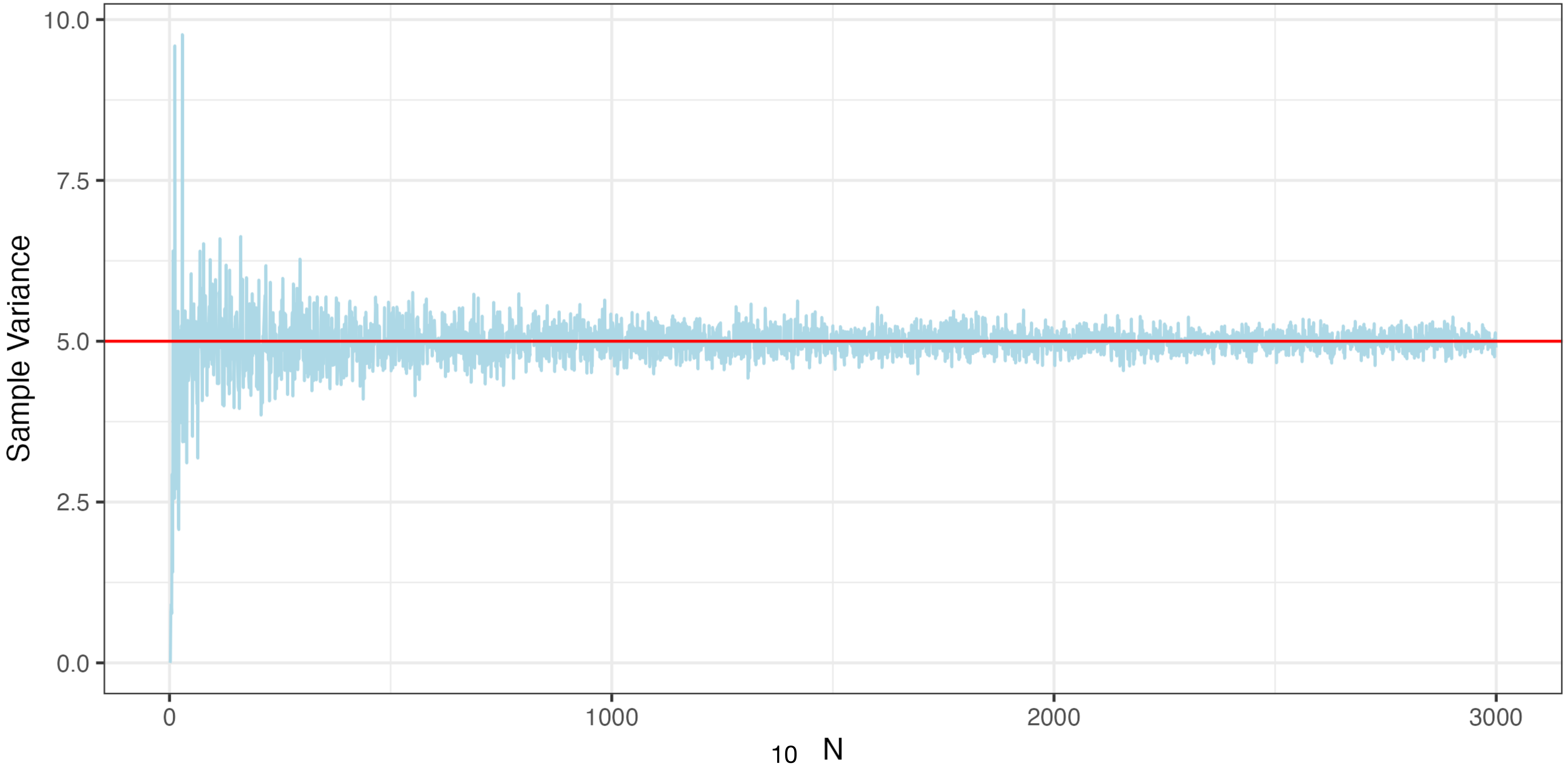
- In practice, you can compute the variance of a finite dataset as

$$\sigma^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{X}^2 \text{ where } \bar{X} := \frac{1}{N}\sum_{i=1}^{N} x_i$$

- You don't need to have the formula memorized, just be aware of it
- Likely you'll never have to explicitly compute it this way, just use an R function

# Sample Variance

- $X_i$ are random draws from $\sim \mathcal{N}(2, 5)$ (a Normal r.v. with mean 2, variance 5)
- How quickly does the sample variance converge to the population variance?



# Conditional Expectation

- Notation: $E(X \mid Y)$
- The expected value given a set of "conditions"
- Read as "the expectation of $X$ given (or conditioned on) $Y$"

$$E(X \mid Y) = \sum_{i=1}^{n} x_i \cdot P(X = x_i \mid Y)$$

$$\text{where } P(X = x_i \mid Y) = \frac{P(X = x_i \text{ and } Y)}{P(Y)}$$

# Conditional Expectation

## Example: Roll a fair dice

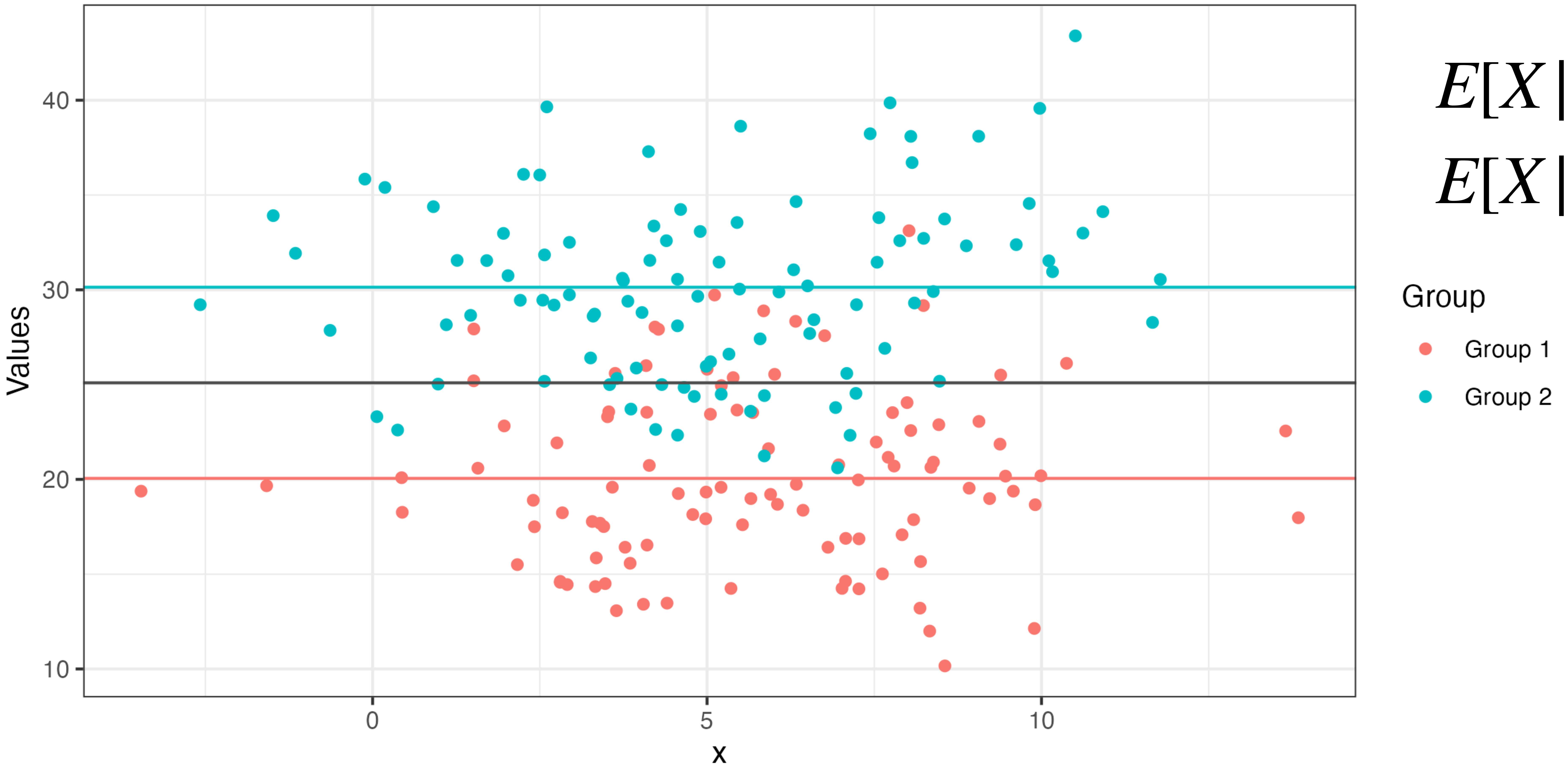
- Let $A = 1$ if you roll an even number, $0$ otherwise.
- Let $B = 1$ if you roll a prime number, $0$ otherwise. Then,

$$E[A] = \sum_{i=1}^{6} a_i \cdot P(a_i) = \frac{0+1+0+1+0+1}{6} = \frac{1}{2}$$

and the conditional expectation of $A$ given $B = 1$ (i.e. we rolled 2, 3, or 5)

$$E[A \,|\, B = 1] = \sum_{i=1}^{3} a_i \cdot P(a_i \,|\, B = 1) = \frac{1+0+0}{3} = \frac{1}{3}$$

# Conditional Expectation - Visualized




$$E[X] = 25$$

$$E[X \,|\, \text{group 1}] = 20$$

$$E[X \,|\, \text{group 2}] = 30$$

# Independence

- Notation: $\perp$ , $X \perp Y$
- Two random variables are **independent** if the outcome of one does not give any information about the outcome of the other
- Events $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$
- Recall: $P(A \cap B) = P(A \mid B)P(B)$
- If $A \perp B$, then $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$

# Independence

## Example: Dice

- Suppose you roll two fair dice. Let $A$ be the value of the first die and let $B$ be the value of the second die.

- If I say that $A = 3$, does that give you any info about what the value of B is?

- We can show that the **events** $\{A = 3\}$ and $\{B = 3\}$ are independent:

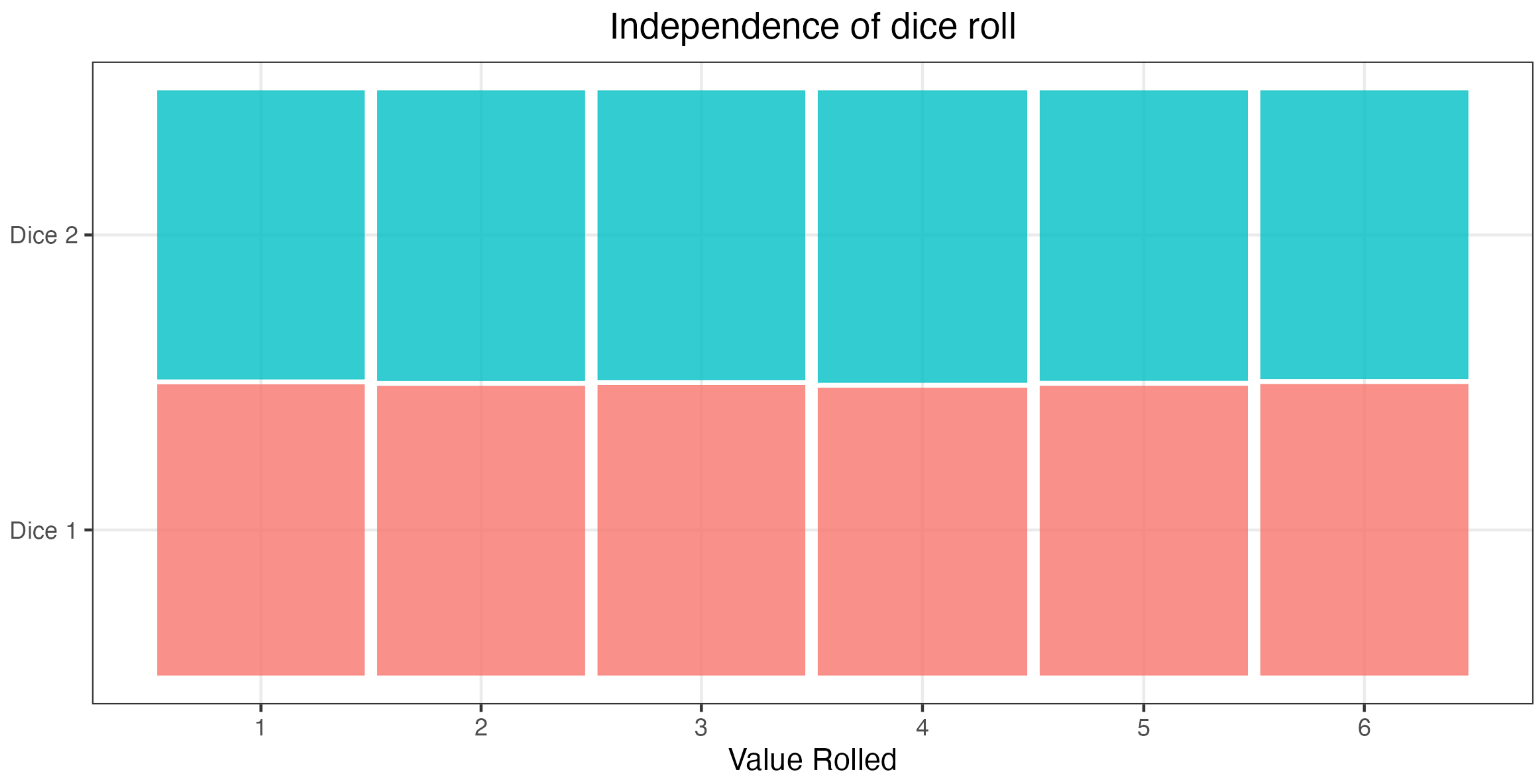
$$\begin{aligned} P(\{A = 3\} \cap \{B = 3\}) &= P(\{A = 3\} \mid \{B = 3\}) \cdot P(\{B = 3\}) \\ &= \frac{1}{6} \cdot \frac{1}{6} \\ &= P(\{A = 3\}) \cdot P(\{B = 3\}) \end{aligned}$$

- To show $A \perp B$, you would show this holds for all values of $A$ and $B$

# Independence

## Example: Dice

- If we simulate 100k dice rolls, we see that the joint probability of each combination is equal to the individual probabilities multiplied.



# Bernoulli Random Variables

## A binary/dichotomous random variable

- Notation: $B(p)$, Bernoulli$(p)$, $\mathscr{B}(p)$
- Takes the value $1$ with probability (w.p.) $p$, and the value $0$ w.p. $q := 1 - p$
- Let $X \sim B(p)$
  - “Let $X$ be a Bernoulli random variable with mean $p$”
  - $E(X) = p$ and $Var(X) = p(1-p) = pq$
- Cool fact: $E(X) = P(X = 1) = p$

# Law of Total Expectation

## (i.e. law of iterated expectations, tower rule)

- Useful property (or “trick) that will be used in class

$$E(X) = E\Big(E(X \mid Y)\Big)$$

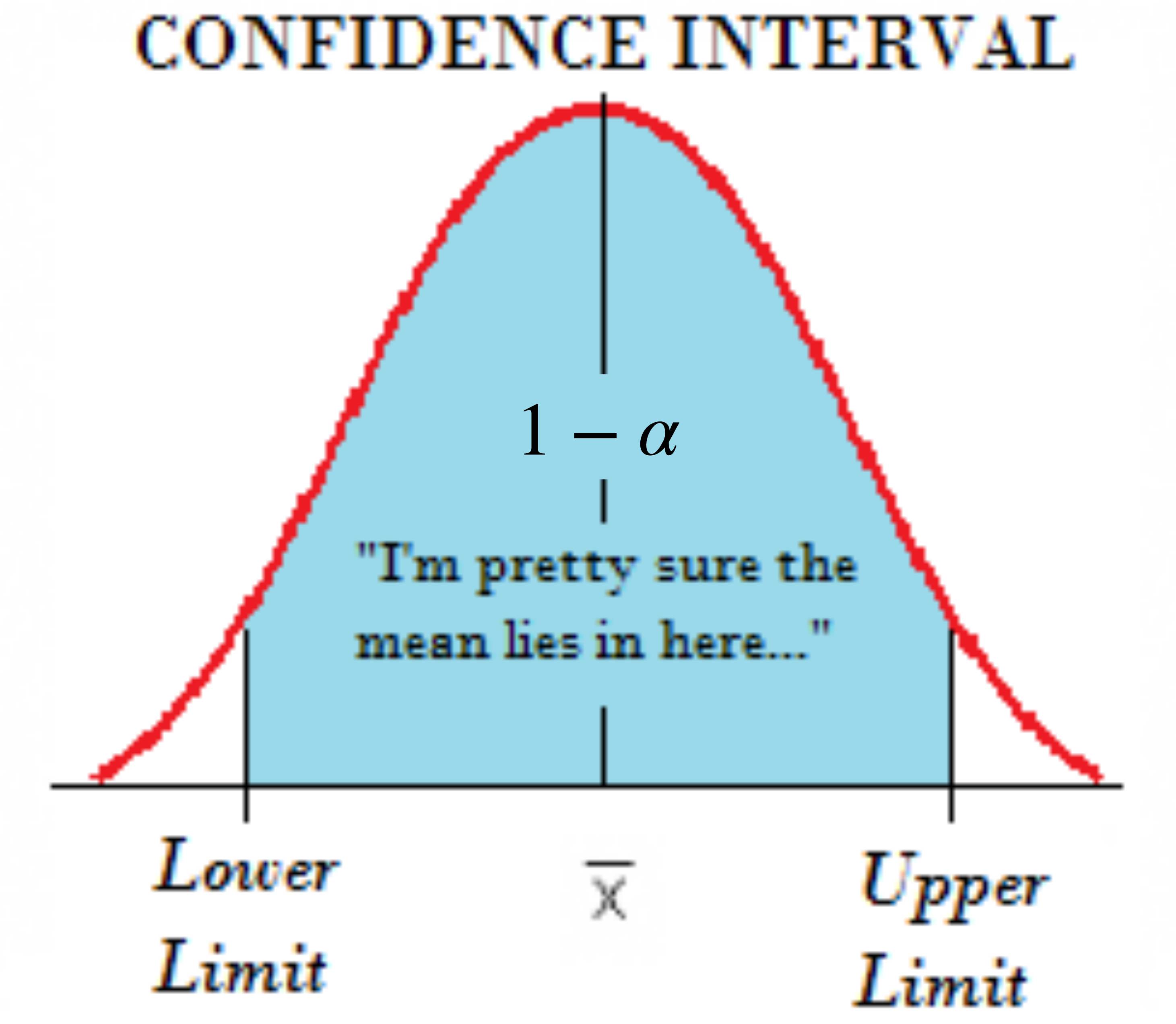
- Don’t worry too much about the technical details, just add to your toolbox :)

# Confidence Intervals

- A set of values that contains the real parameter with probability $1 - \alpha$
- Define $CI = [L, U]$ then $P(L \leq \mu \leq U) = 1 - \alpha$
- Usually $1 - \alpha$ is $95\,\%$ or $99\,\%$
- *Example*: $X_i$ are random draws from $\sim \mathcal{N}(2, 5)$
- Estimating expectation of a random variable using sample mean:

$$\hat{E}(X) = \hat{\mu} = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

# Confidence Intervals




- $\bar{X}$ is an estimate for $\mu$ with some uncertainty
- $P(\mu \leq \bar{X} - c) = P(\mu \geq \bar{X} + c) = \dfrac{\alpha}{2}$
- $P\left(\dfrac{\bar{X}-\mu}{\sigma/\sqrt{N}} \leq \dfrac{\mu - c - \mu}{\sigma/\sqrt{N}}\right) \Rightarrow -c = Z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{N}}$
- $Z_{\frac{\alpha}{2}}$ is the the critical value of the Normal distribution (For example in R: `qnorm(0.025)`)
- $CI = \bar{X} \pm Z_{\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{N}}$

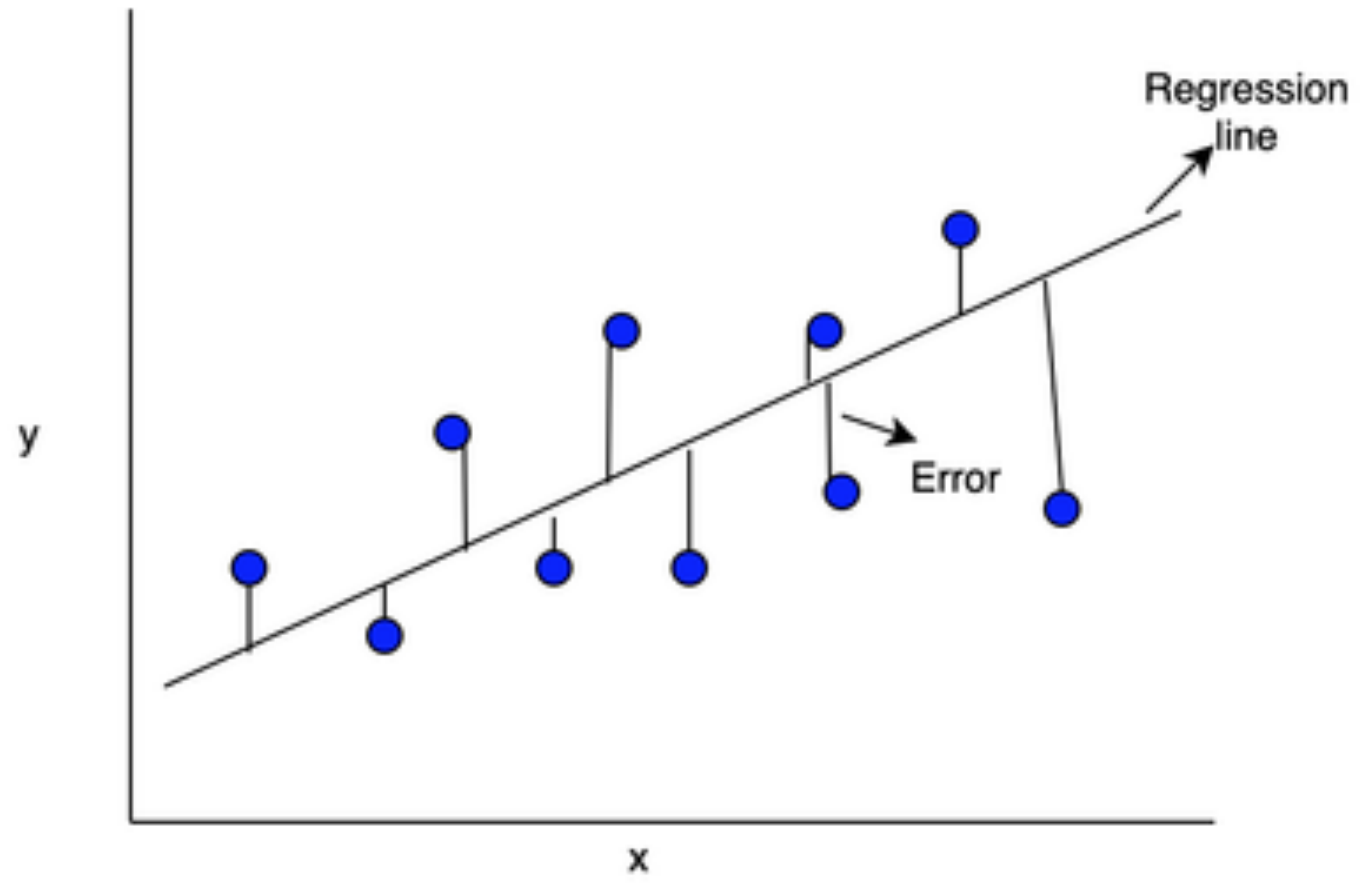
# Regression

- Estimates the relationships between $X$ and $Y$ where
- $Y$- the dependent variable, outcome/response
- $X$- independent variable, regressor/explanatory
- Main types of regression: Linear and Logistic

# Regression

## Linear Regression

- Assume data was generated: $Y_i = \alpha + \beta X_i + \varepsilon_i$ for $i = 1,\ldots,N$
- $\alpha, \beta$ are the coefficients where $\alpha$ is the intercept and $\beta$ the slope



# Regression

## Linear Regression

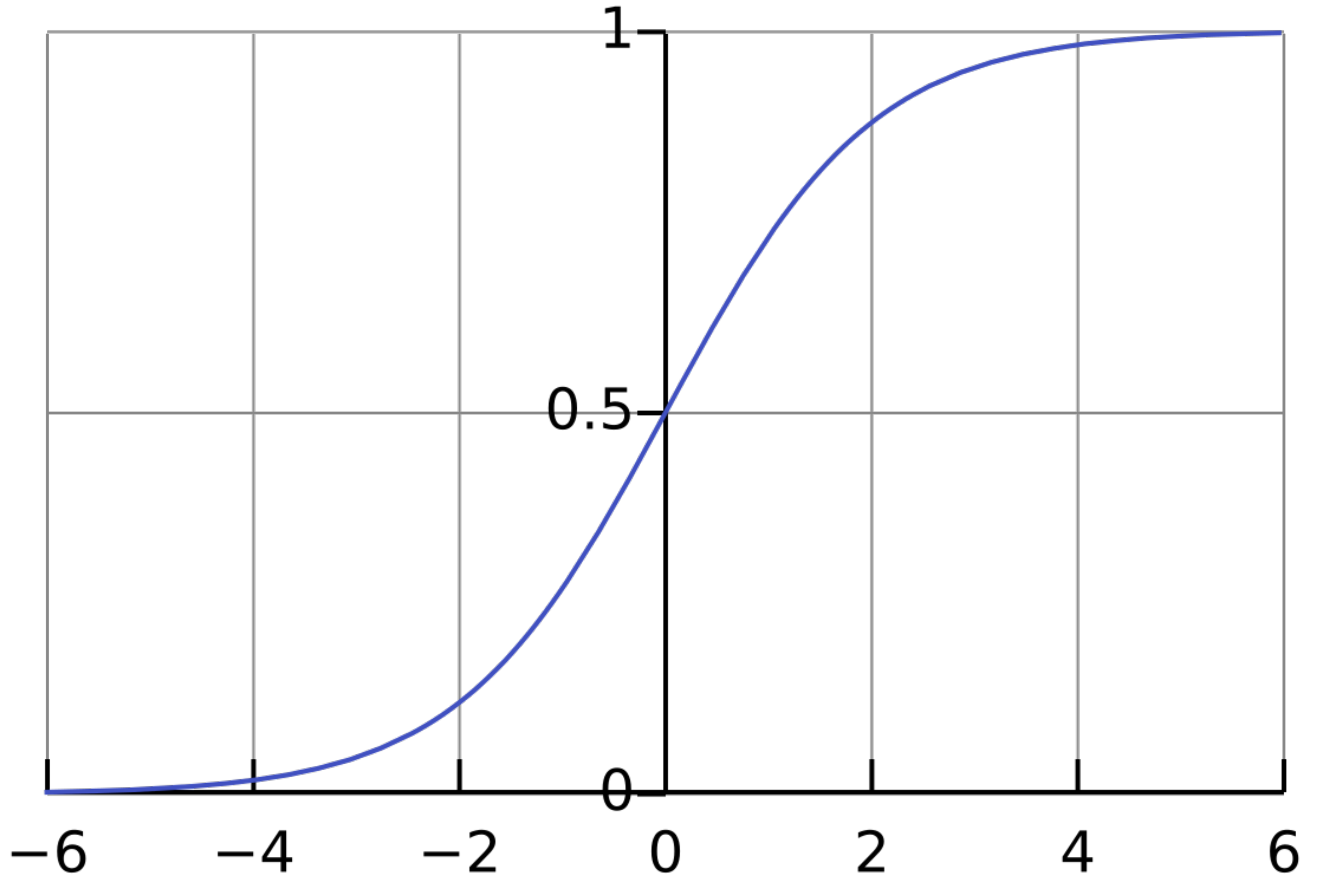
- Using ordinary least squares (OLS) to estimate $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$
- Minimizes sum of squared errors: $(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{a,b} \sum_{i=1}^{N} \left(Y_i - (a + bX_i)\right)^2$
- $\dfrac{\partial}{\partial a} SSE = \sum_{i=1}^{N} -2(Y_i - a - bX_i) \qquad \Rightarrow \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
- $\dfrac{\partial}{\partial b} SSE = \sum_{i=1}^{N} -2(Y_i - (\bar{Y} - b\bar{X}) - bX_i)X_i = \sum_{i=1}^{N} -2\left[(Y_i - \bar{Y})X_i - b(X_i - \bar{X})X_i\right]$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

# Regression

## Logistic Regression

- $Y_i$- the outcome variable is binary for $i = 1,\ldots,N$
- Use a link function to estimate $P(Y_i = 1) := p_i$ that satisfies $\mathbb{R} \to (0,1)$
  - Most common- logistic function: $\sigma(t) = \dfrac{1}{1+e^{-t}}$
- In a linear model we estimate $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$
- In logistic model we estimate $\hat{p}_i = \dfrac{1}{1+e^{-(\hat{\alpha}+\hat{\beta}X_i)}}$
- $\alpha + \beta X_i = \ln\left(\dfrac{p_i}{1-p_i}\right)$



# Regression

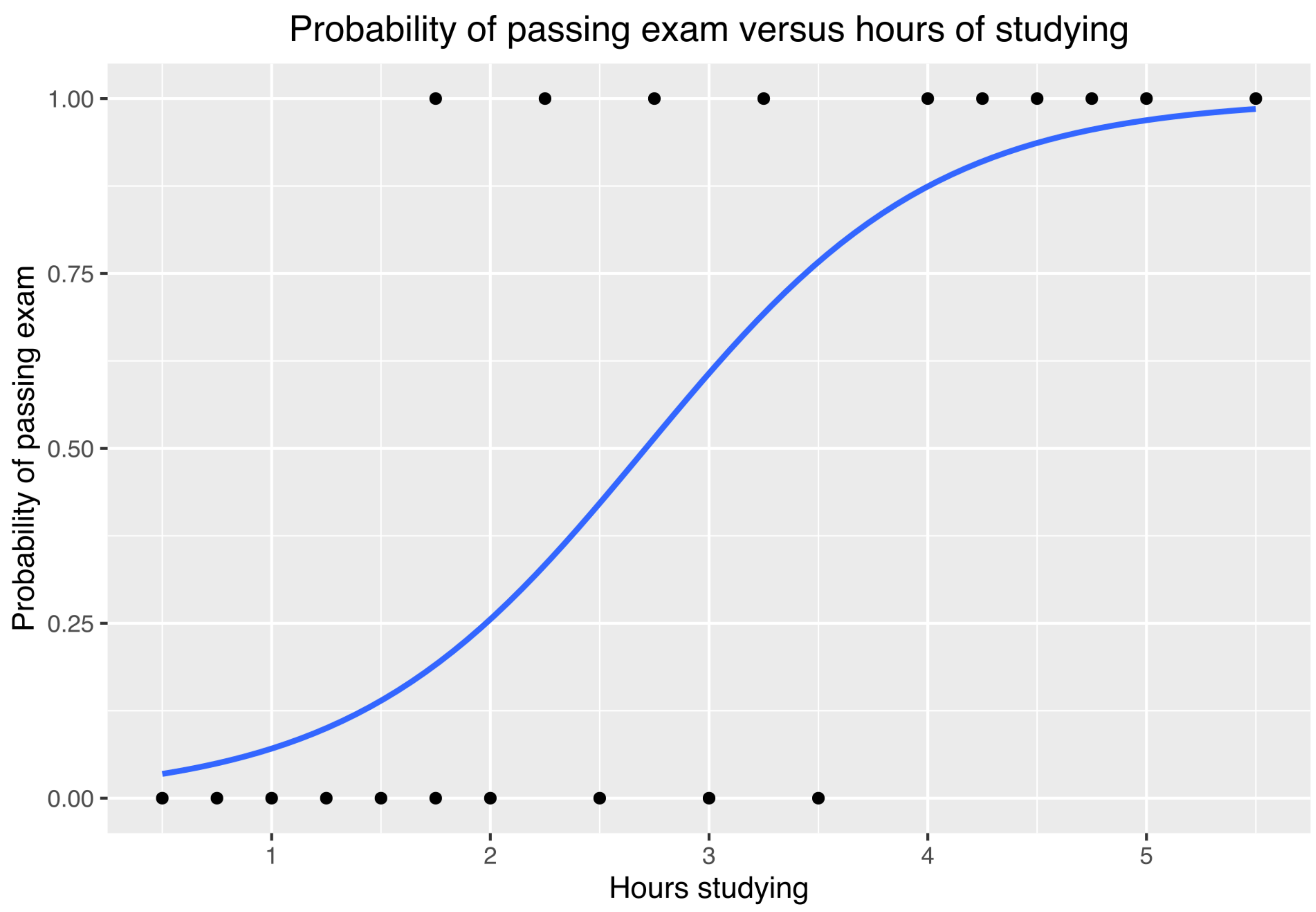
## Logistic Regression

- Odds ratio: $\dfrac{p_i}{1-p_i} = \dfrac{P(Y_i = 1)}{P(Y_1 = 0)}$
  - For example: $\dfrac{P(\text{Passing exam})}{P(\text{Not passing})} = \dfrac{3/4}{1/4}$ the odds ratio is $3:1$
- To estimate $\hat{\alpha}, \hat{\beta}$ we use maximum likelihood estimates (MLE)
- Likelihood function: $L(a, b; y) = \prod_{i=1}^{N} P(Y_i = y_i) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{(1-y_i)}$
- Log likelihood: $l(a, b; y) = \sum_{i=1}^{N} y_i \ln(p_i) + (1-y_i)\ln(1-p_i) = \sum_{i=1}^{N} \ln(1-p_i) + y_i \ln\left(\dfrac{p_i}{1-p_i}\right)$

# Regression

## Logistic Regression

- Log likelihood: $l(a,b;y) = \sum_{i=1}^{N} \ln(1-p_i) + y_i \ln\left(\frac{p_i}{1-p_i}\right) = \sum_{i=1}^{N} -\ln(1+e^{a+bX_i}) + y_i(a+bX_i)$
- To find MLE we solve $\dfrac{\partial}{\partial(a,b)} l(a,b;y) = 0$
- No close form solution
  iterative method such as:
  gradient descent or Newton–Raphson



# R

- R is an open-source programming language
- Used for statistical computing and creating plots
- Download and install R

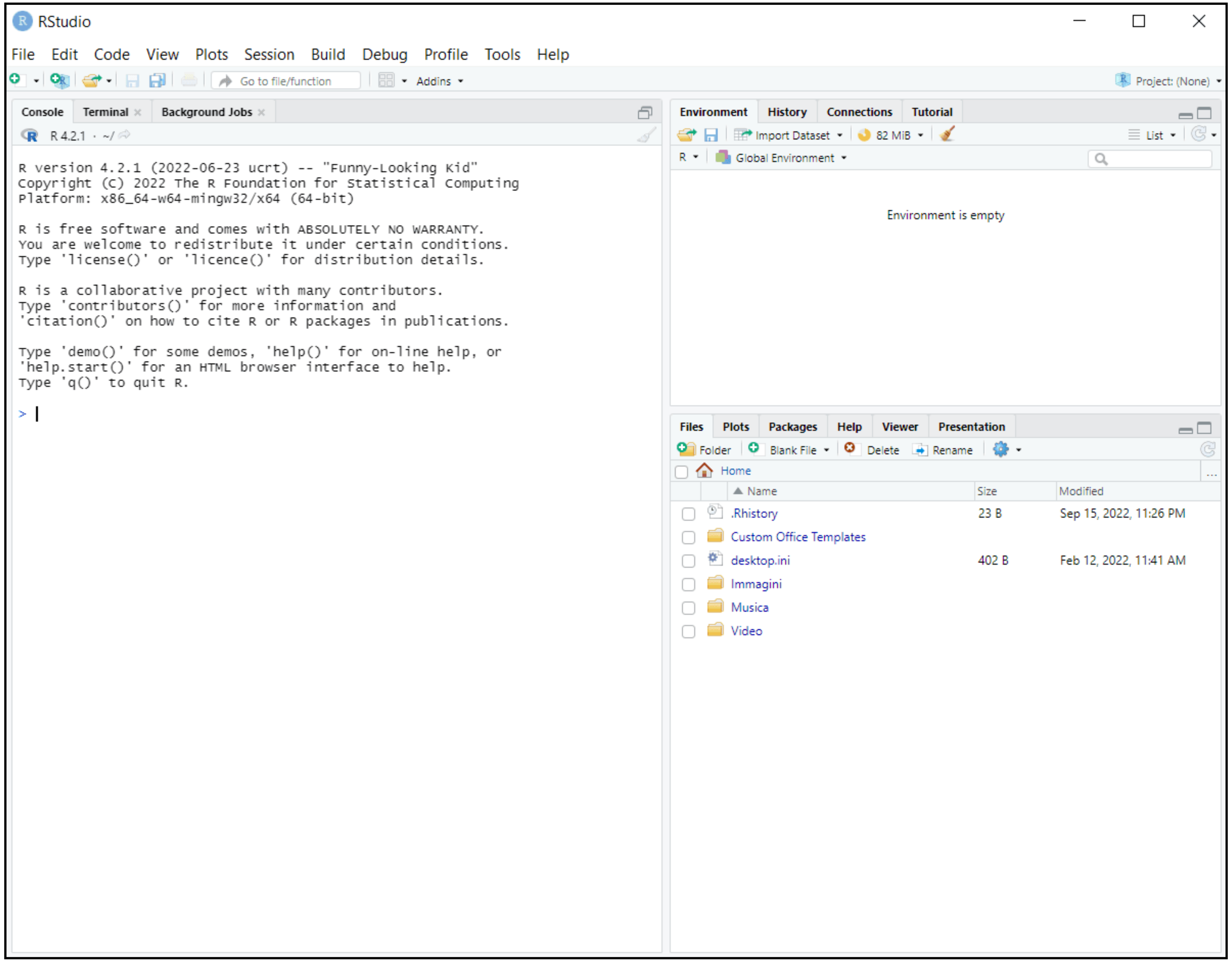
https://cran.r-project.org/

# RStudio

- RStudio is an open-source IDE (integrated development environment)
- Download and install RStudio (scroll down for earlier versions)

https://posit.co/download/rstudio-desktop/

# RStudio


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console Terminal Background Jobs
R 4.2.1 · ~/
R version 4.2.1 (2022-06-23 ucrt) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment History Connections Tutorial
Import Dataset
82 MiB
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer Presentation
Folder
Blank File
Delete
Rename
Home
Name
Size
Modified
.Rhistory
23 B
Sep 15, 2022, 11:26 PM
Custom Office Templates
desktop.ini
402 B
Feb 12, 2022, 11:41 AM
Immagini
Musica
Video

# RStudio

## Quick Demo

- Console- calculator, create variable
- Environment
- Files
- Plots
- Help
- Script

# R Markdown

- install.packages("rmarkdown")
- install.packages(“knitr")
- Download HW 1 and open in RStudio
- R Markdown tutorial

https://www.rforecology.com/post/how-to-use-rmarkdown-part-one/

- Subscripts and superscripts: to get $Y_i^a$ inline use *`$Y_{i}^{a}$`*

# Questions

- Homework Check-in
- R/RStudio