PSY 503: Foundations of Statistical Methods in Psychological Science

Statistical Models,

Probability

Suyog Chandramouli

Zoom & 311 PSH (Princeton University)

22nd September, 2025

What is a model?

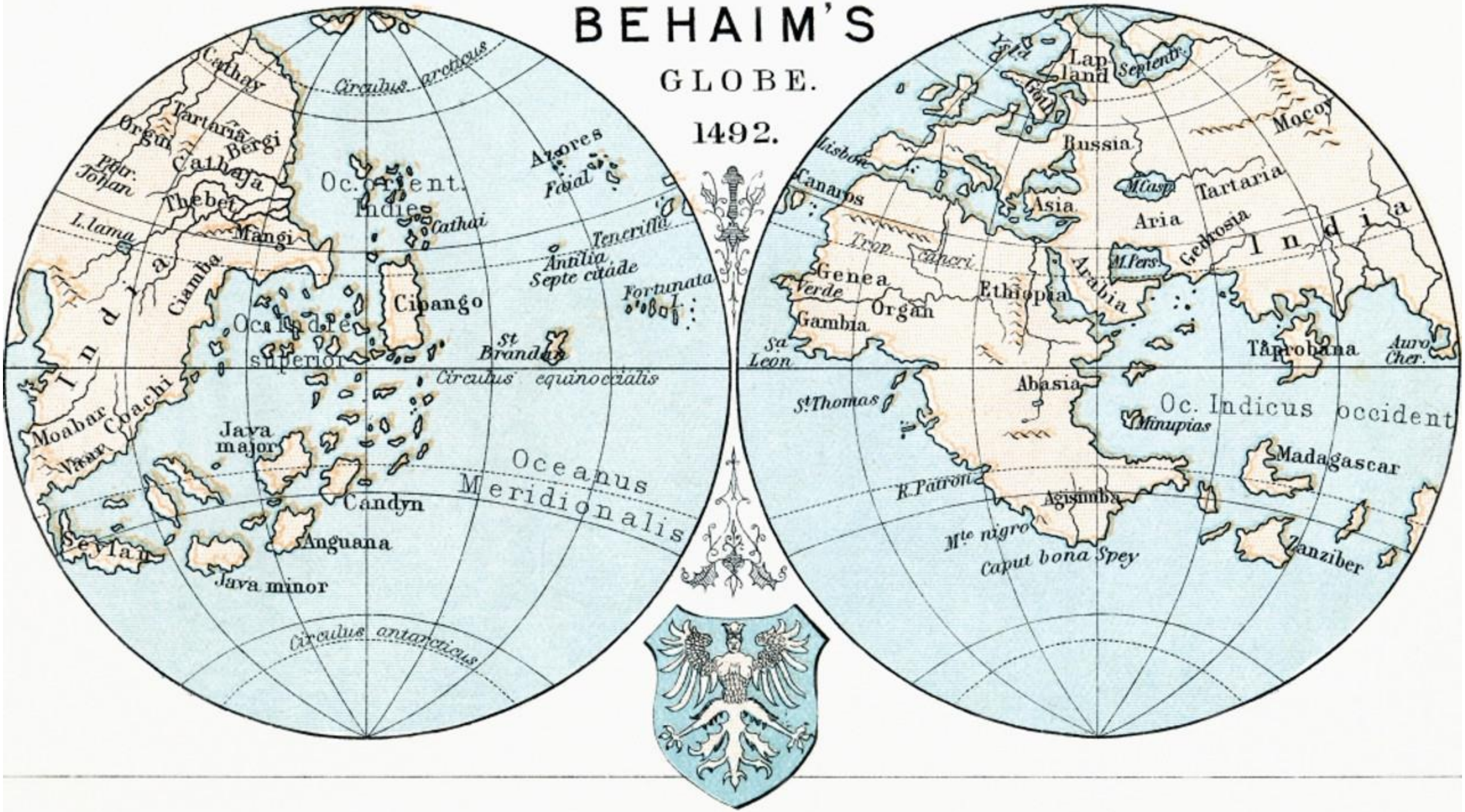
- Models are simplifications of things in the real world

What is a model?

- Models are simplifications of things in the real world

BEHAIM'S GLOBE.

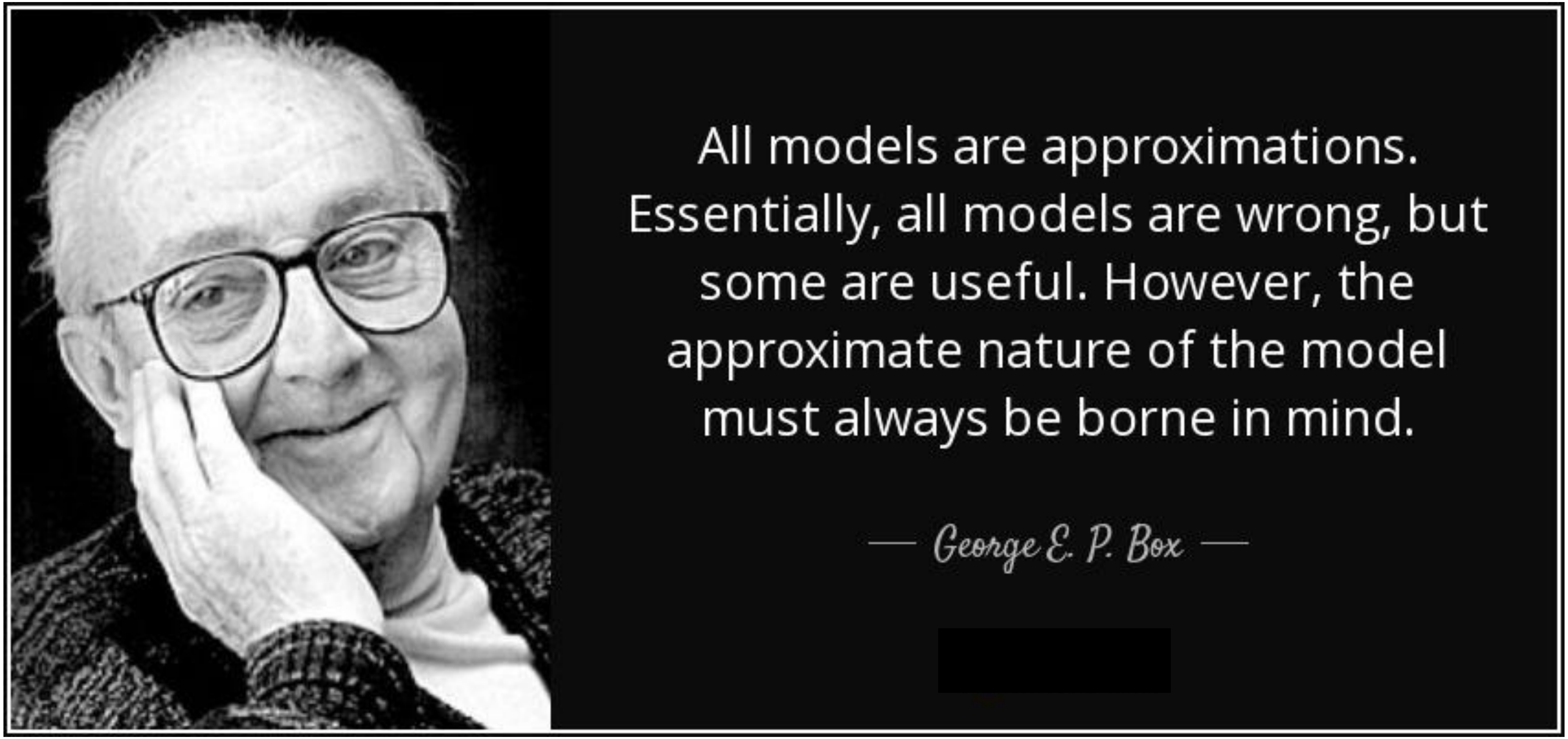
1492.




ARCTIC OCEAN
KAZAKHSTAN
MONGOLIA
CHINA
IRAN
SAUDI ARABIA
PAKISTAN
INDIA
ARABIAN SEA
JAPAN
NORTH PACIFIC
PHILIPPINE SEA
SOUTH CHINA SEA
PHILIPPINES
MALAYSIA
SRI LANKA
MALDIVES
INDONESIA
INDIAN OCEAN
AUSTRALIA

The map is not the territory.

- Our understanding is always an abstraction or simplification of the complex world around us.


All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind.
— George E. P. Box —
[redacted]

Models are not just of things, but also of processes.

- Laws of Physics
- Biology
- Climate science
- Economics

….

Psychology

Models as Golems

- Golem = animated human-like being, made from inanimate matter such as clay or mud (Clay robots)

- Powerful but mindless servants
  - Servant when used well
  - Dangerous because they follow instructions literally (no wisdom, no foresight)

- In some versions, Rabbi Judah Loew ben Bezalel built a golem to protect. But he lost control, causing innocent death

Statistical 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
  - 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!

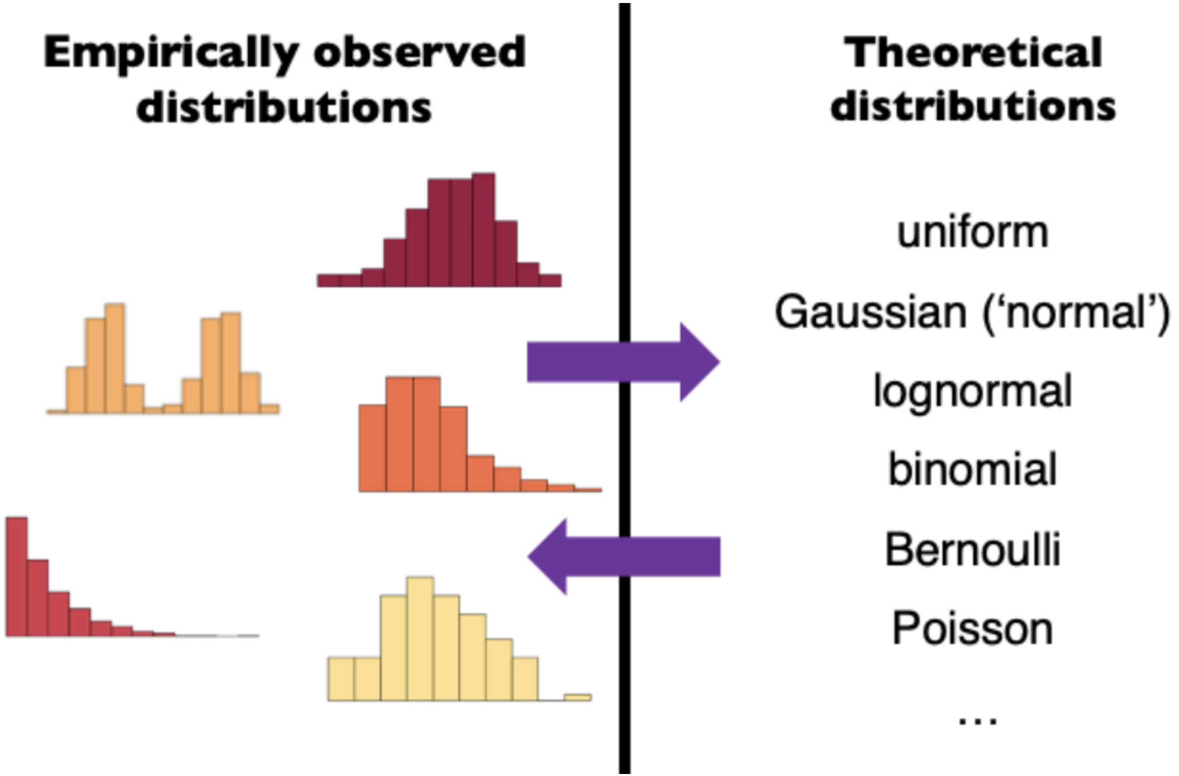
No model without assumptions

- Assumptions about data, relationships between variables, and variability
  - Violations bias results, or limit applicability
  - A lot of the statistical workflow is about checking if assumptions are met.

- True with small and large models

What is a statistical modelling?

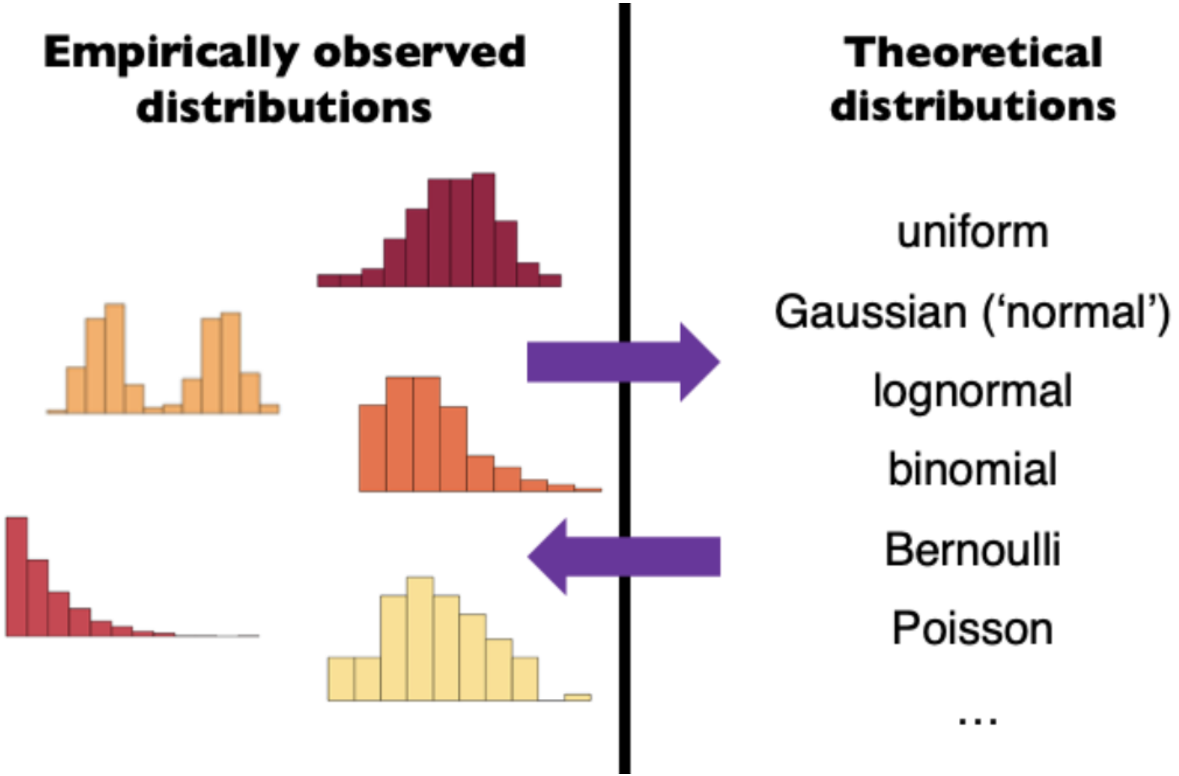
- **Statistical modeling** = “making **models** of **distributions**”



What is a statistical modelling?

- **Statistical modeling** = “making **models** of **distributions**”

  (coming up with a plausible data generating process/ DGP)



Basic Structure of a Statistical Model

$$data = model + error$$

- Data
- Model
- Use our model to ***predict*** the value of the data for any given observation:

$$\hat{data}_i = model_i$$

- Error (predicted – observed)

$$error_i = data_i - \hat{data}_i$$

Notation

- Small Roman letters
  - Individual observed data points
  - $y_1, y_2, y_3, y_4, \dots, y_n$
    - 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_1
  - The model for person 1
  - The yet-to-be-observed score of person 1

Notation

- Greek letters
  - Population parameters
  - Unobservable parameters
- μ
  - mu
    - “mew” - Used to describe means
- σ
  - Sigma
    - Used to describe a standard deviation

- Roman letters
  - Sample specific statistics
    - $\bar{X}$ - sample mean
    - s - standard deviation from the sample
  - Data estimates
    - b_0

A simple model

- Null or empty model

$$Y_i \;=\; \beta_0 \;+\; \epsilon$$

$$\mathrm{Y}_i \;=\; b_0 \;+\; e$$

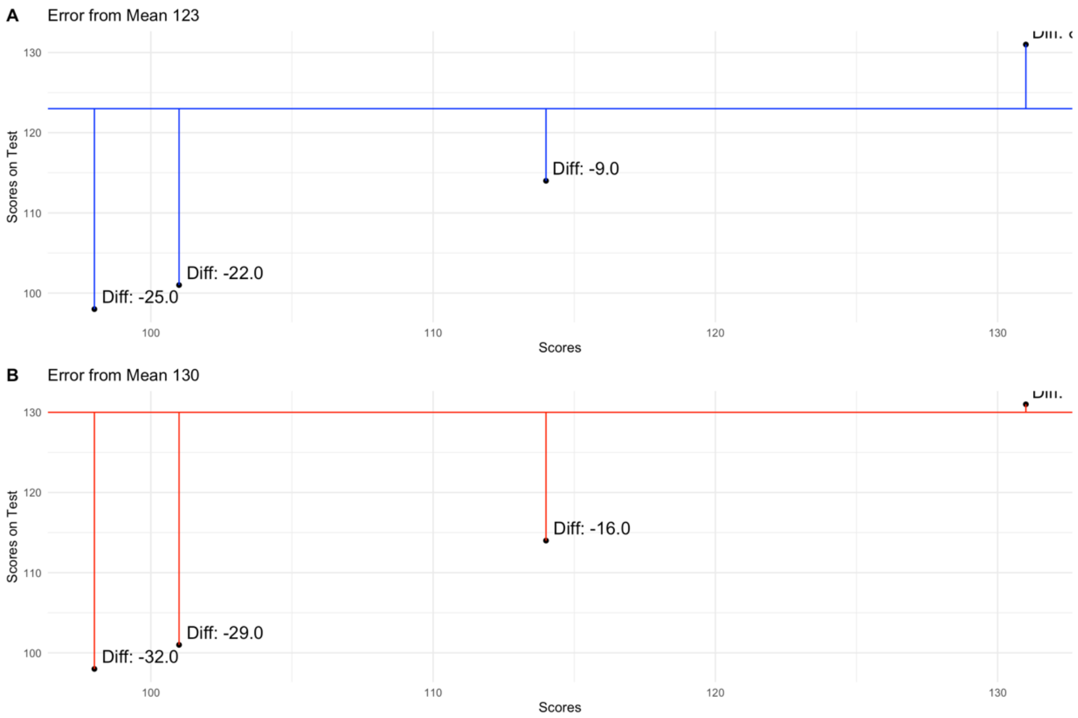
- Makes the same prediction for each observation (and we add an error sample)

A simple model: data

- Assume the following observed Scores:
  - 101
  - 114
  - 131
  - 9

Figuring out b_0

- Goal of any model is to find an “estimator” that minimizes the error
  - How we define error will determine the best estimator



Error Measures

- Sum of errors (SE)

$$SE = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

  - Ideally we'd like this to be 0

Error Measures

- Sum of absolute errors (SAE)

$$SAE = \sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

  - It gives a sense of the average magnitude of errors without considering direction

- Median is the best estimate for b_0

Error Measures

- Sum of Squared Errors (SS)
  - This measures the total squared difference between observed and predicted values
  - Most commonly used in regression analysis (what we will be using)

$$SS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The Mean

- Mean is the best estimator of b_0

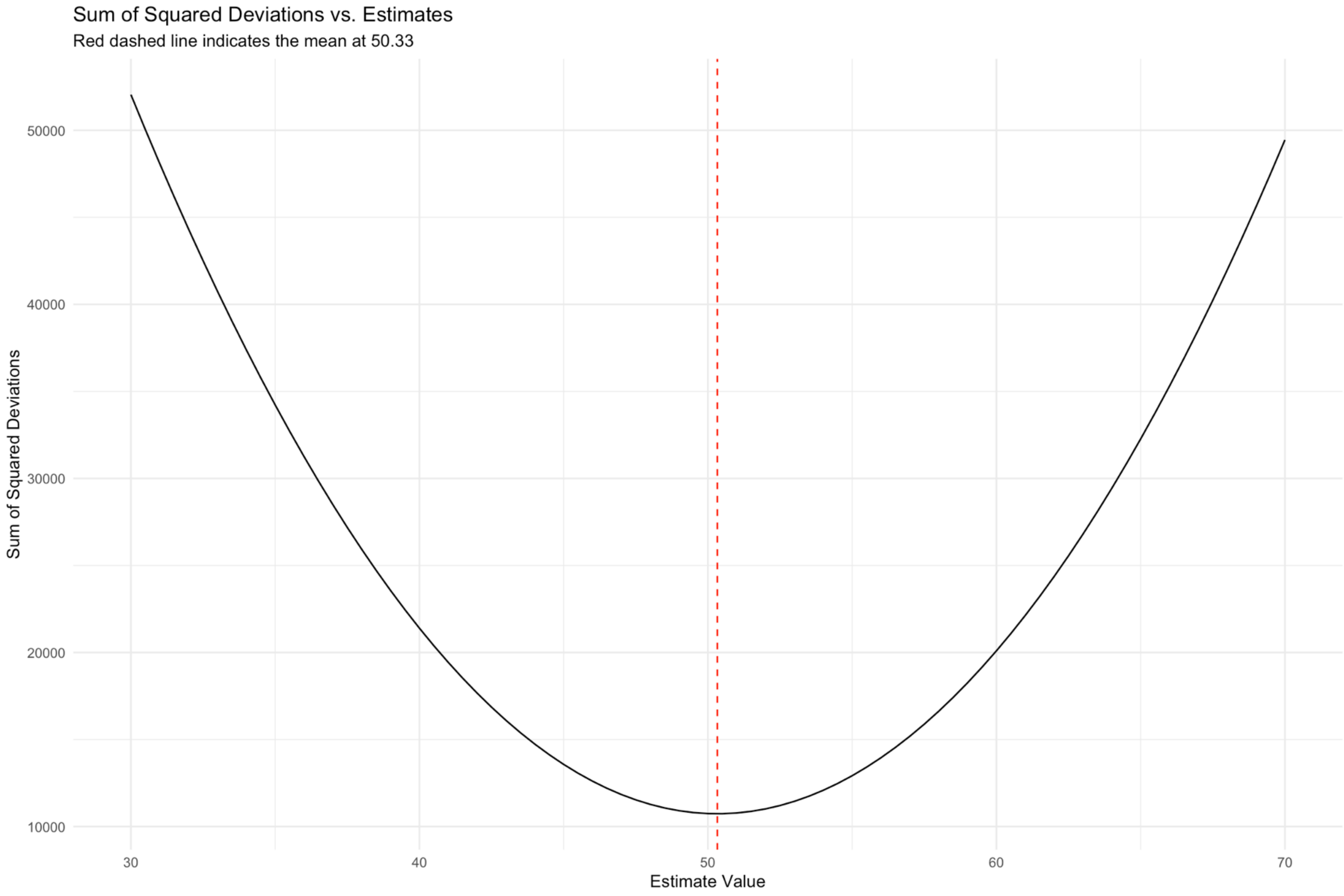
- Mean has really nice proprieties

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

- SS minimized at mean

$$SS = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

SS minimized at the mean



Why use the mean?

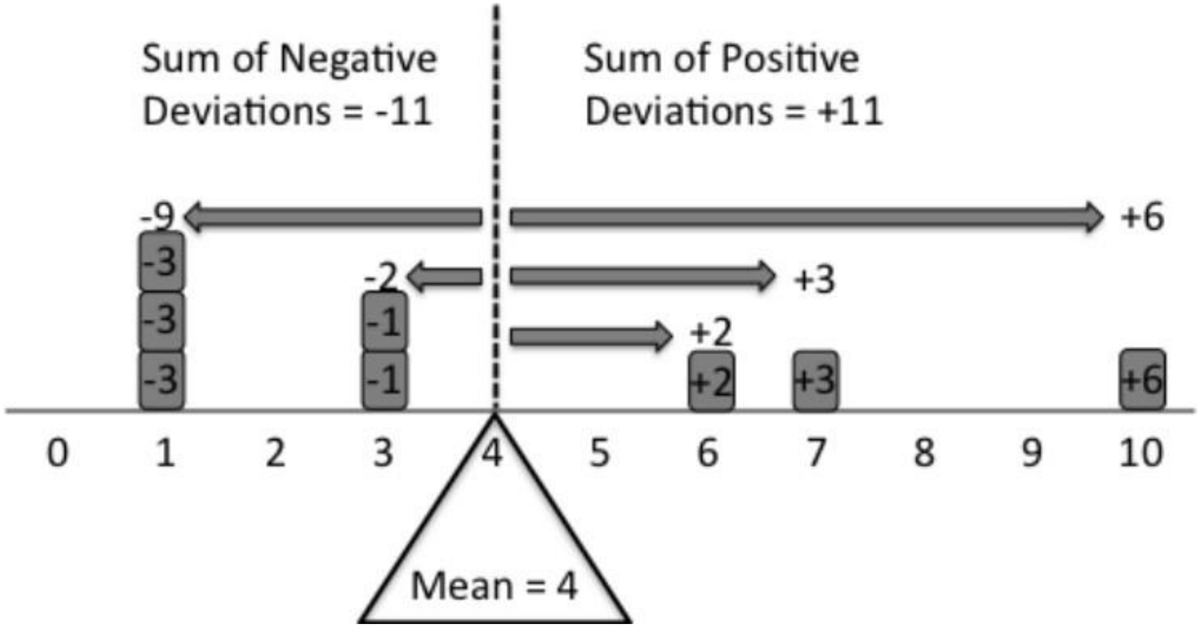
The mean is a good model for the data because it minimizes **sums of squared error.**

$$Outcome = Model + Error$$

We want a model which has minimal error.

The distance between the prediction (model) and the observed value (outcome) is the error.

The mean ensures that the positive errors and negative errors are balanced with regard to their magnitude.



Describing error

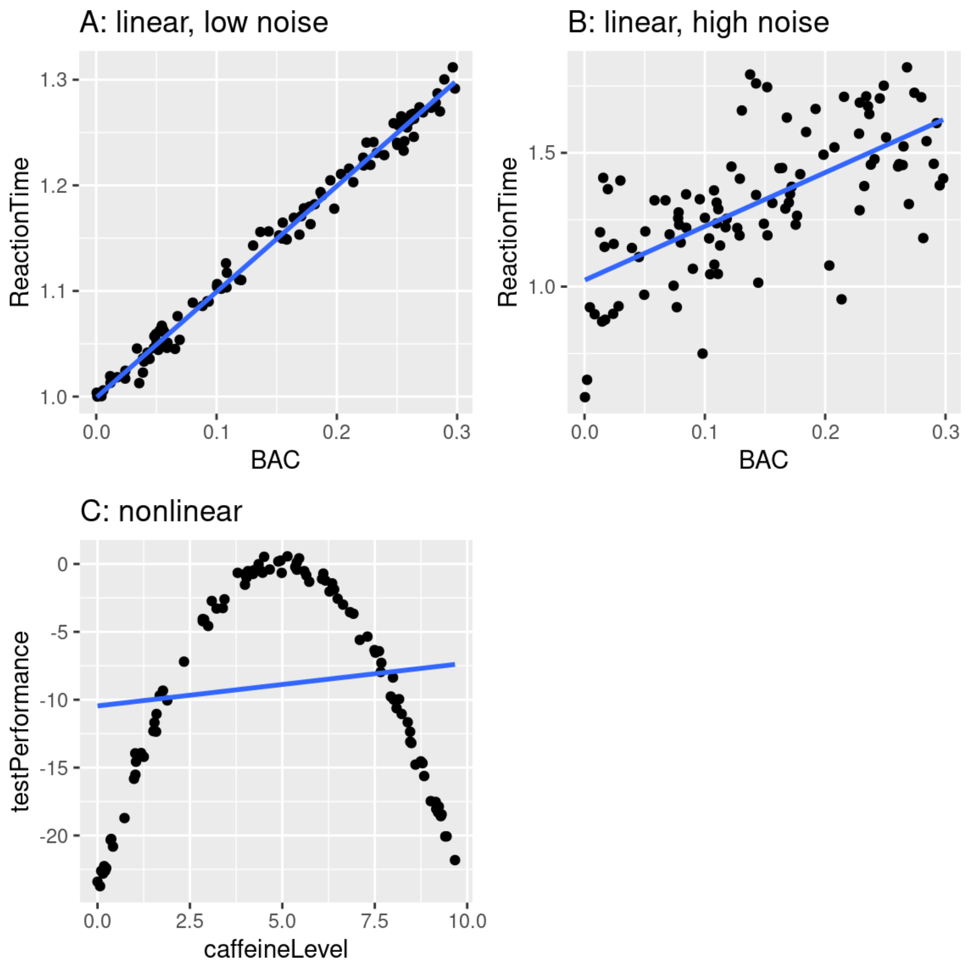
- We should have some overall description of the accuracy of model's predictions
  - SSR
    - Standard deviation

$$s^2 = \mathrm{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2$$

$$SD = \sqrt{MSE}$$

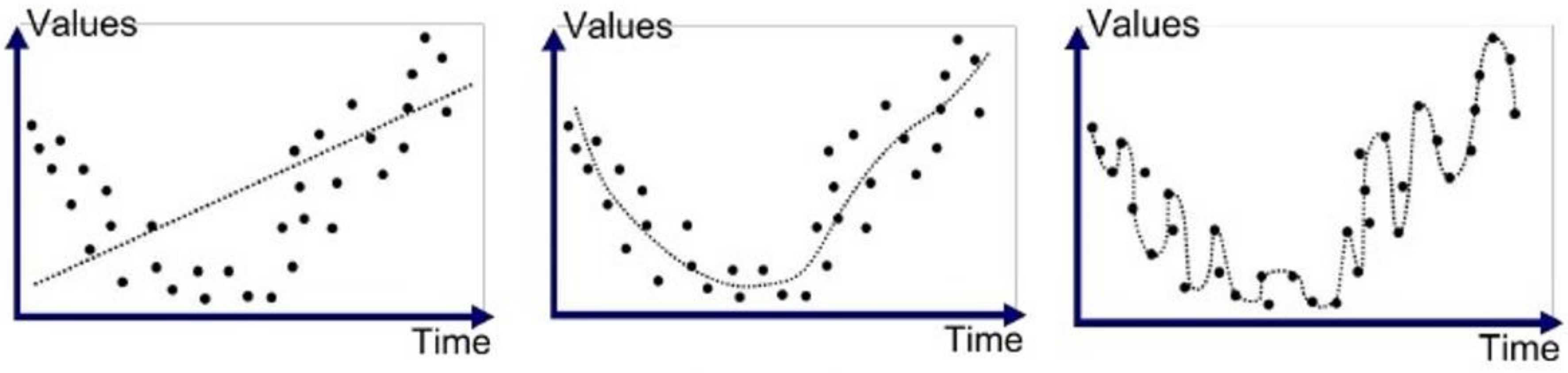
What Makes a Model "Good"

- We want it to describe our data well
- We want it to generalize to new datasets
- We want error to be as small as possible



Can a Model Be so good that it’s bad?

- Yes!
  - Overfitting
  - A model with little to no error will not generalize to new datasets

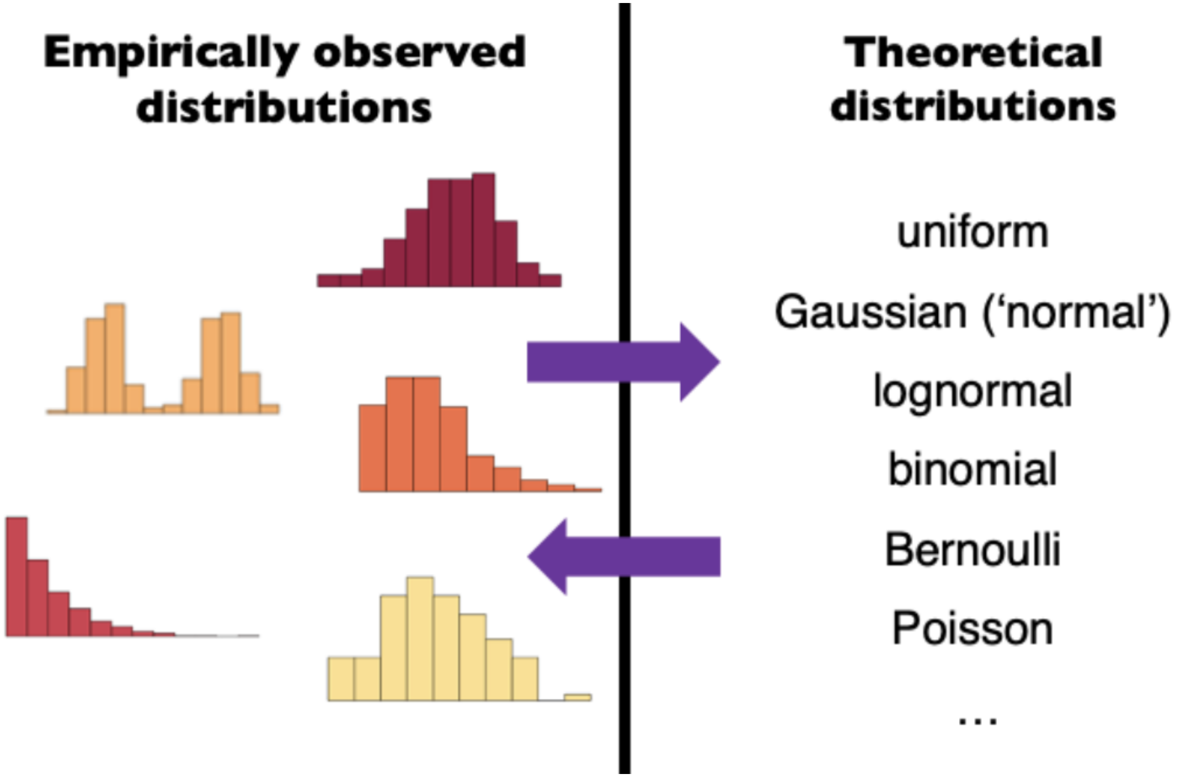


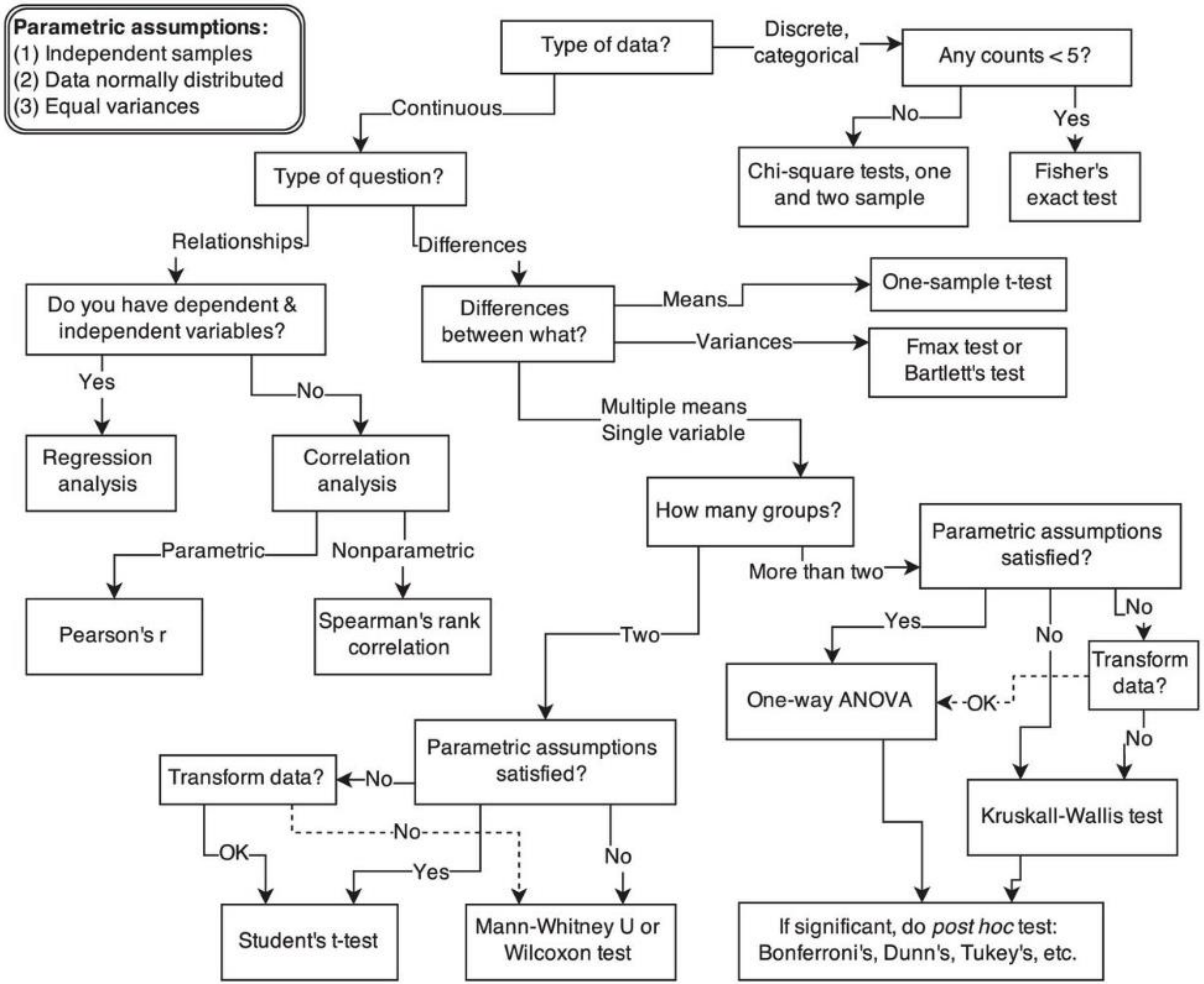
Underfitted Good Fit/Robust Overfitted

What is a statistical modelling?

- **Statistical modeling** = "making **models** of **distributions**"

  (coming up with a plausible data generating process/ DGP)





Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances
Type of data?
Discrete, categorical
Any counts < 5?
No
Yes
Chi-square tests, one and two sample
Fisher's exact test
Continuous
Type of question?
Relationships
Differences
Do you have dependent & independent variables?
Yes
No
Regression analysis
Correlation analysis
Parametric
Nonparametric
Pearson's r
Spearman's rank correlation
Differences between what?
Means
One-sample t-test
Variances
Fmax test or Bartlett's test
Multiple means Single variable
How many groups?
More than two
Parametric assumptions satisfied?
Yes
No
No
Transform data?
OK
No
One-way ANOVA
Kruskall-Wallis test
If significant, do post hoc test: Bonferroni's, Dunn's, Tukey's, etc.
Two
Parametric assumptions satisfied?
No
Transform data?
OK
No
Yes
No
Student's t-test
Mann-Whitney U or Wilcoxon test

Cookbook perspective is limiting

- We want our models to be an outcome of exploring the data, understanding relationship between our causal variables, and flexibly expressing it through statistics

- Generalized Linear Models (GLMs) provide a unified way of thinking about the several common statistical tests.

GLM

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

GLM

Model comparison approach

- Think in terms of models and not tests
- Model is determined by question, not data
- What do alternative models say about the world?

Fitting the model

We can use the `lm` function to fit the model with no predictors (Null Model / Empty model)

```
empty.model <- lm(HrsSleep2009~NULL, data =
smallNLS) empty.model
```

```
##
## Call:
## lm(formula = HrsSleep2009 ~ NULL, data =
smallNLS) ##
## Coefficients:
## (Intercept)
##        6.65
```

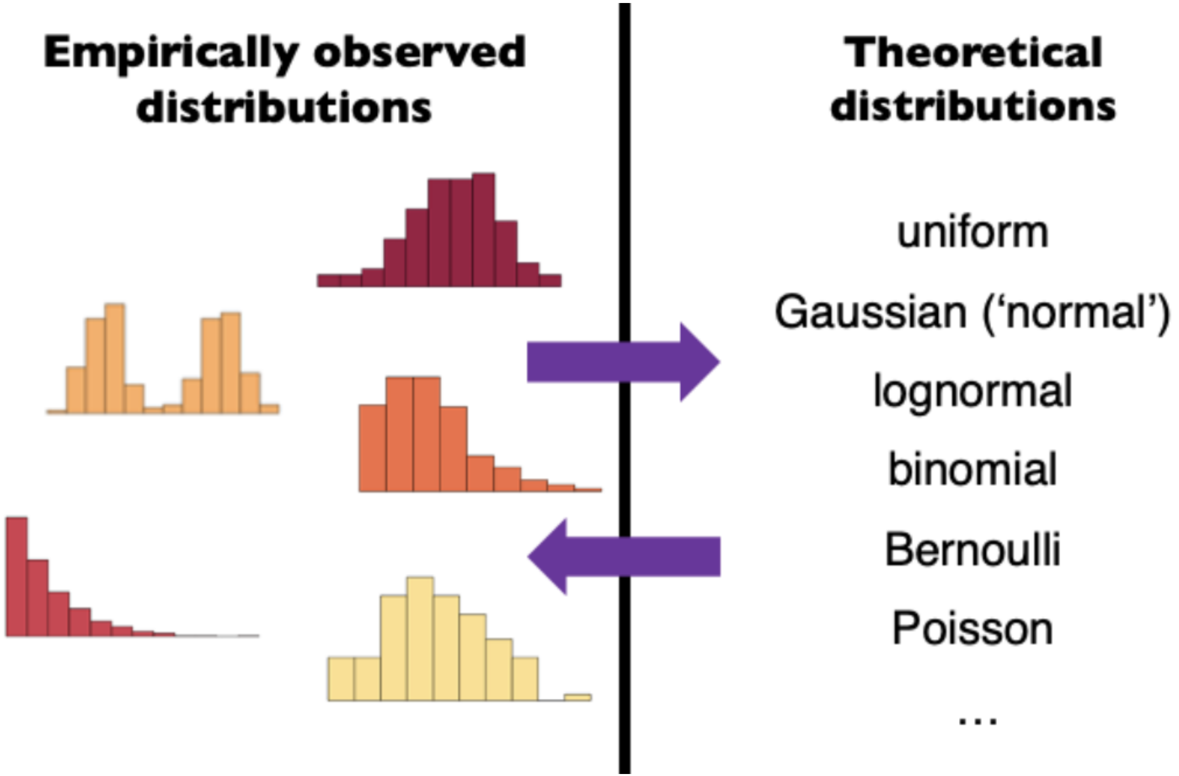
```
favstats(~HrsSleep2009, data = smallNLS)
```

```
##  min Q1 median Q3 max mean       sd  n missing
##    5  6      7  8   8 6.65 1.136708 20       0
```

What is a statistical modelling?

- **Statistical modeling** = "making **models** of **distributions**"

  (coming up with a plausible data generating process/ DGP)



Probability vs Stats

- Probability theory
  - Helps determine likelihood of different events occuring, based on knowledge of DGP.
  - Model known, Data unknown
  - Prediction
- Statistics
  - Start with observed events.
  - Data known, Model unknown
  - Determine DGP
  - Inference

Probability vs Stats

- Probability theory
  - Fair coin: P(10 heads in a row)?
  - Two dice: P(double sixes)?
  - Shuffled deck: P(5 hearts)?

  Known rules → Calculate chances

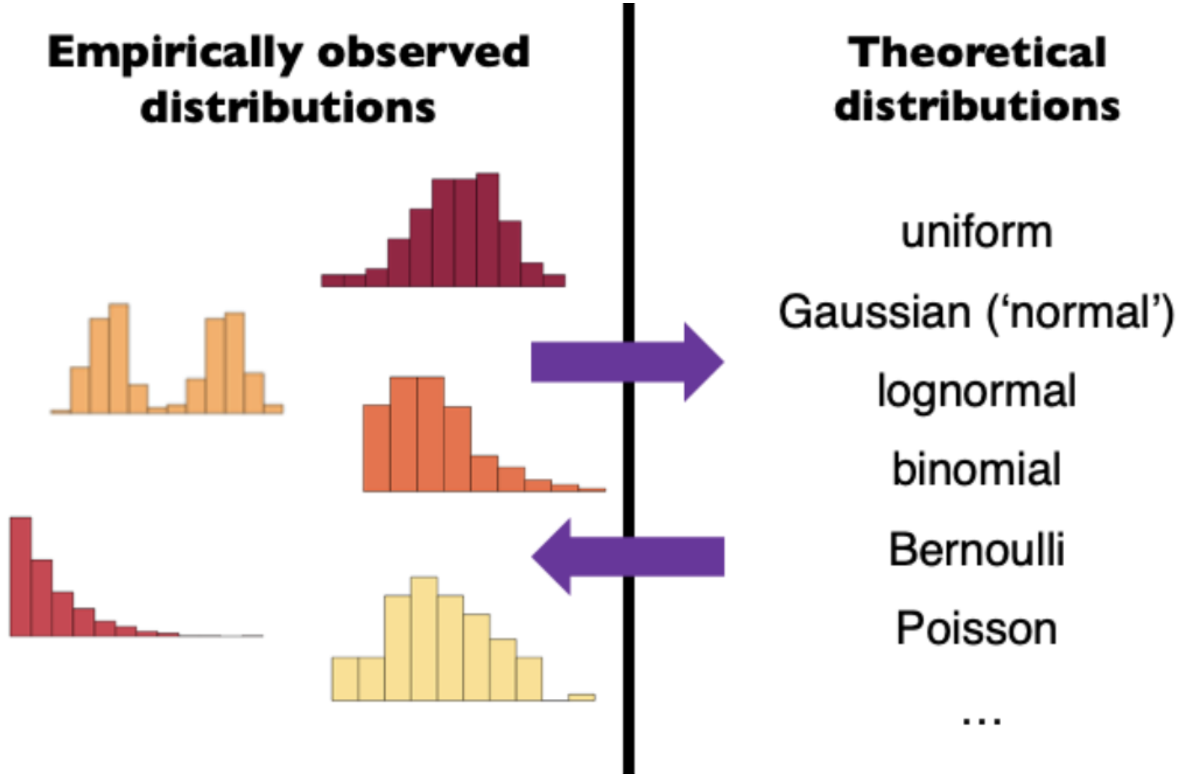
- Statistics
  - 10 heads observed: Is coin fair?
  - 5 hearts drawn: Was deck shuffled?
  - Lottery winner related to commissioner: Rigged?

Data observed → Infer truth

What is a statistical modelling?

- **Statistical modeling** = “making **models** of **distributions**”

  (coming up with a plausible data generating process/ DGP)



Distributions can be specified by numbers

- Mean

- In more detail,
  - Normal: Mean, standard deviation
  - Binomial: N, p(successes)
  - ..

Probability Basics

What is a probability?

- A number bounded between 0 and 1
- Describes the "chances" or "likelihood" of an event

Proportions and Percentages

- Percentage (%) : A ratio between event frequency, and total frequency, expressed in units of 100.

- Proportion : a decimal version (range between 0-1)

Two probability statements

- A coin has a 50% chance of landing heads

$$p(\text{heads}) = .5$$

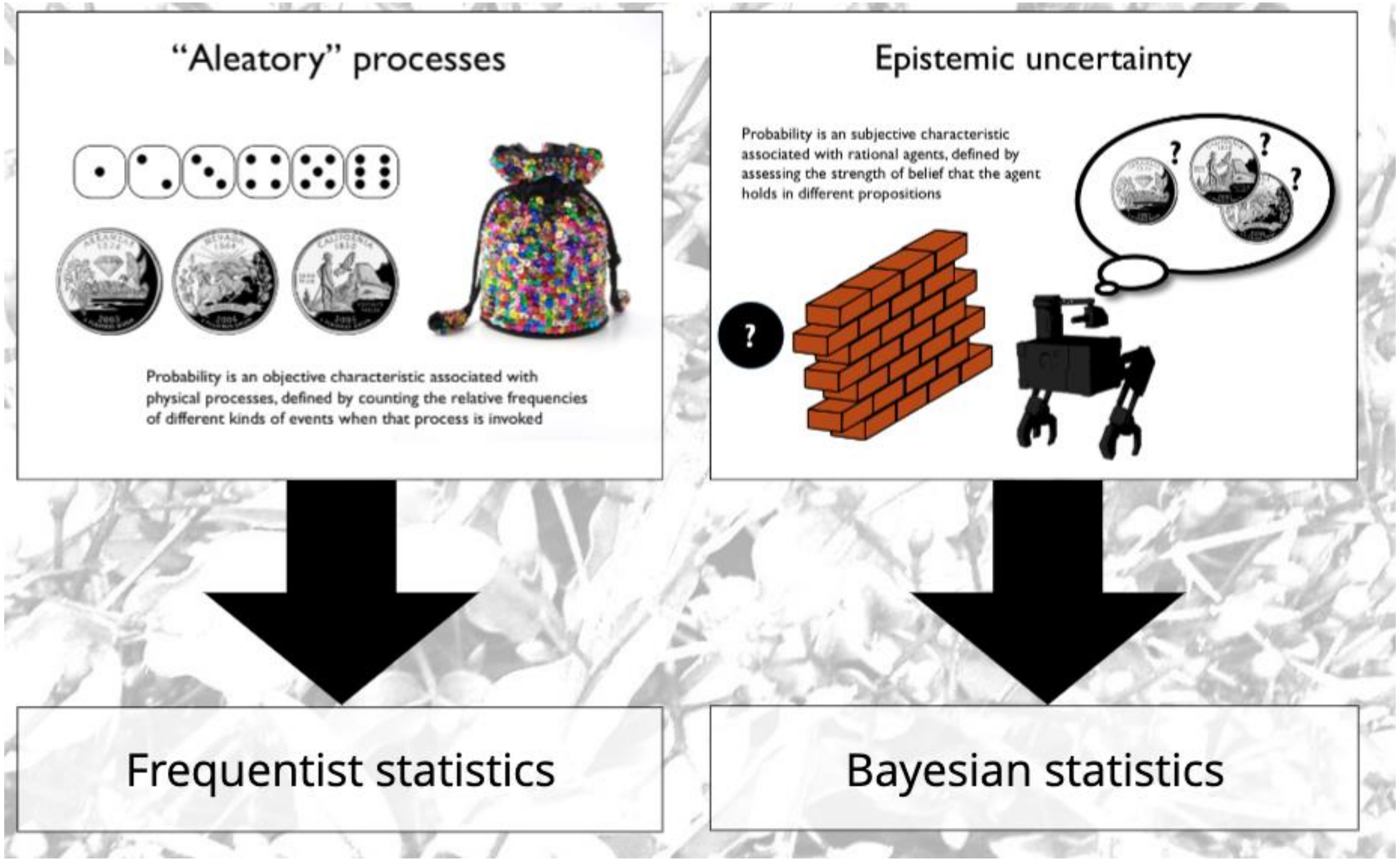
- There is a 10% chance of rain tomorrow

$$p(\text{rain tomorrow}) = .1$$

Frequentist vs. Bayesian

- Probability is defined differently depending on philosophical tradition.
- Frequentist: Long-run frequency
  - Requires repeatable events
- Bayesian: Degree of belief
- Both are valid, different tools for different purposes.

2 perspectives of probability



A fair coin

- A fair coin has a 50% chance of landing heads or tails

A fair coin

- A fair coin has a 50% chance of landing heads or tails

Discuss:

- What does this mean for a frequentist?
- What does this mean for a Bayesian?

50% chance

R used to flip a fair coin 100 times:

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,] "H"  "H"  "T"  "T"  "H"  "H"  "H"  "T"  "H"  "T"
##  [2,] "T"  "H"  "T"  "H"  "H"  "H"  "H"  "H"  "T"  "H"
##  [3,] "H"  "H"  "H"  "T"  "T"  "T"  "H"  "T"  "H"  "T"
##  [4,] "T"  "T"  "H"  "T"  "H"  "H"  "T"  "T"  "H"  "T"
##  [5,] "H"  "T"  "T"  "H"  "T"  "T"  "H"  "H"  "T"  "T"
##  [6,] "H"  "H"  "H"  "T"  "T"  "T"  "T"  "H"  "T"  "T"
##  [7,] "H"  "H"  "H"  "T"  "T"  "T"  "H"  "T"  "T"  "H"
##  [8,] "T"  "T"  "H"  "T"  "H"  "H"  "H"  "T"  "H"  "H"
##  [9,] "T"  "T"  "T"  "T"  "T"  "H"  "H"  "H"  "T"  "H"
## [10,] "T"  "H"  "H"  "T"  "T"  "H"  "T"  "T"  "H"  "T"
```

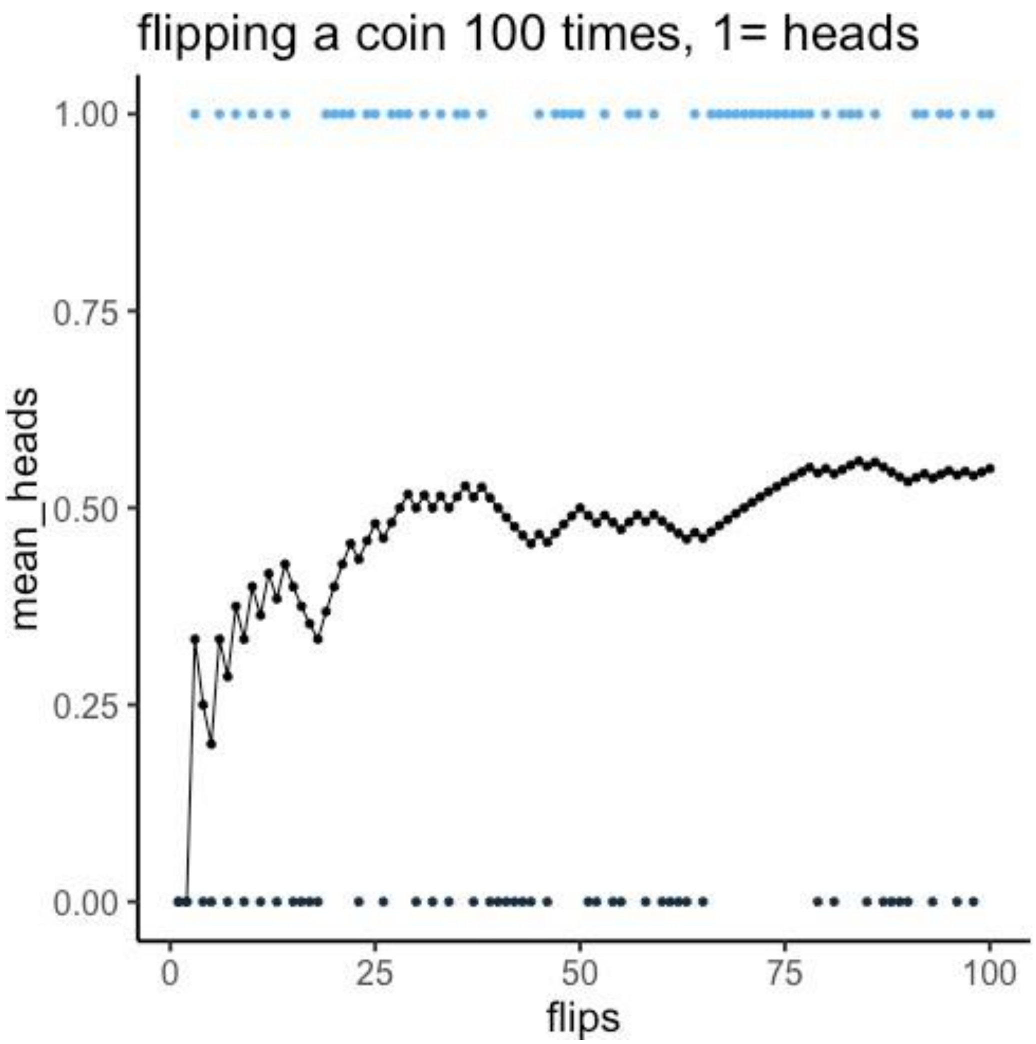
A fair coin

- A fair coin has a 50% chance of landing heads or tails

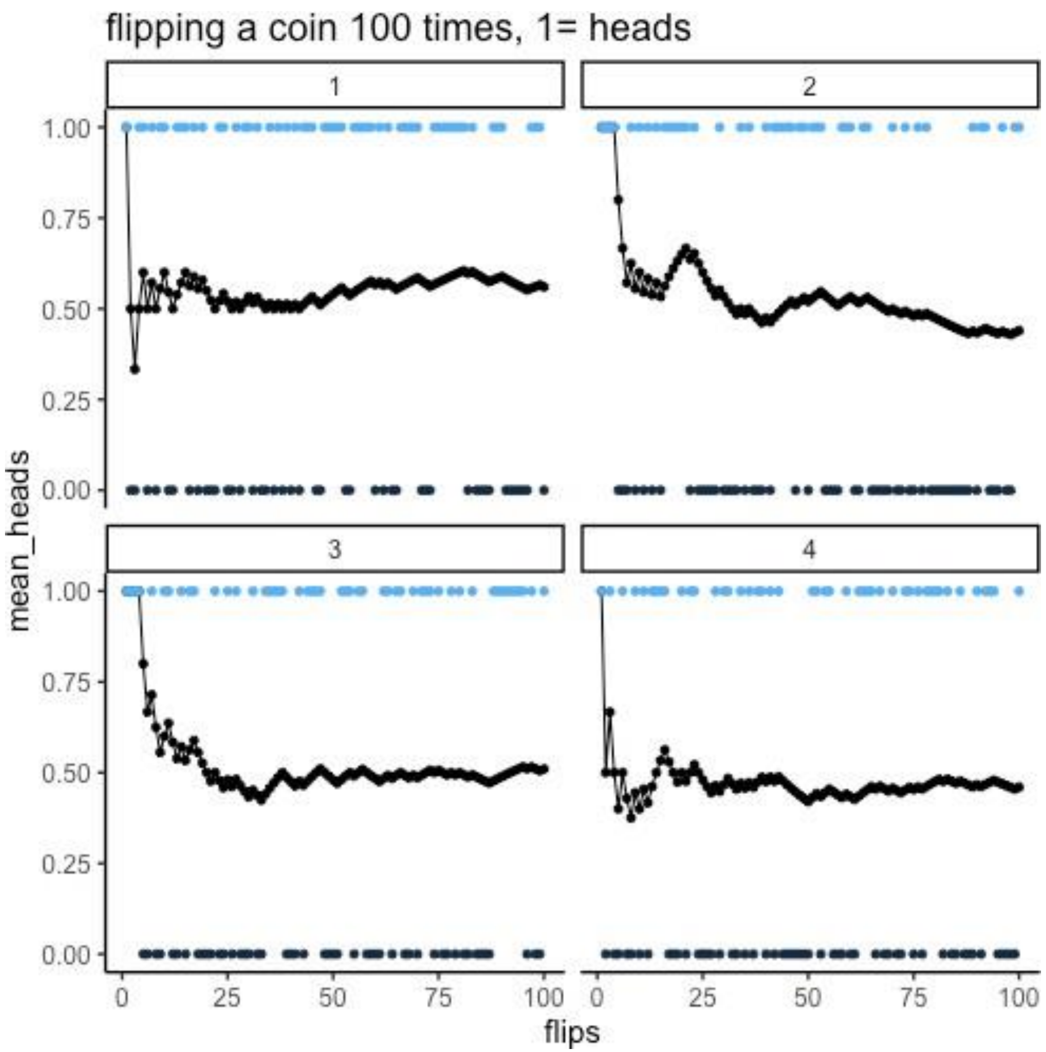
Discuss:

- What does this mean for a frequentist?
- What does this mean for a Bayesian?

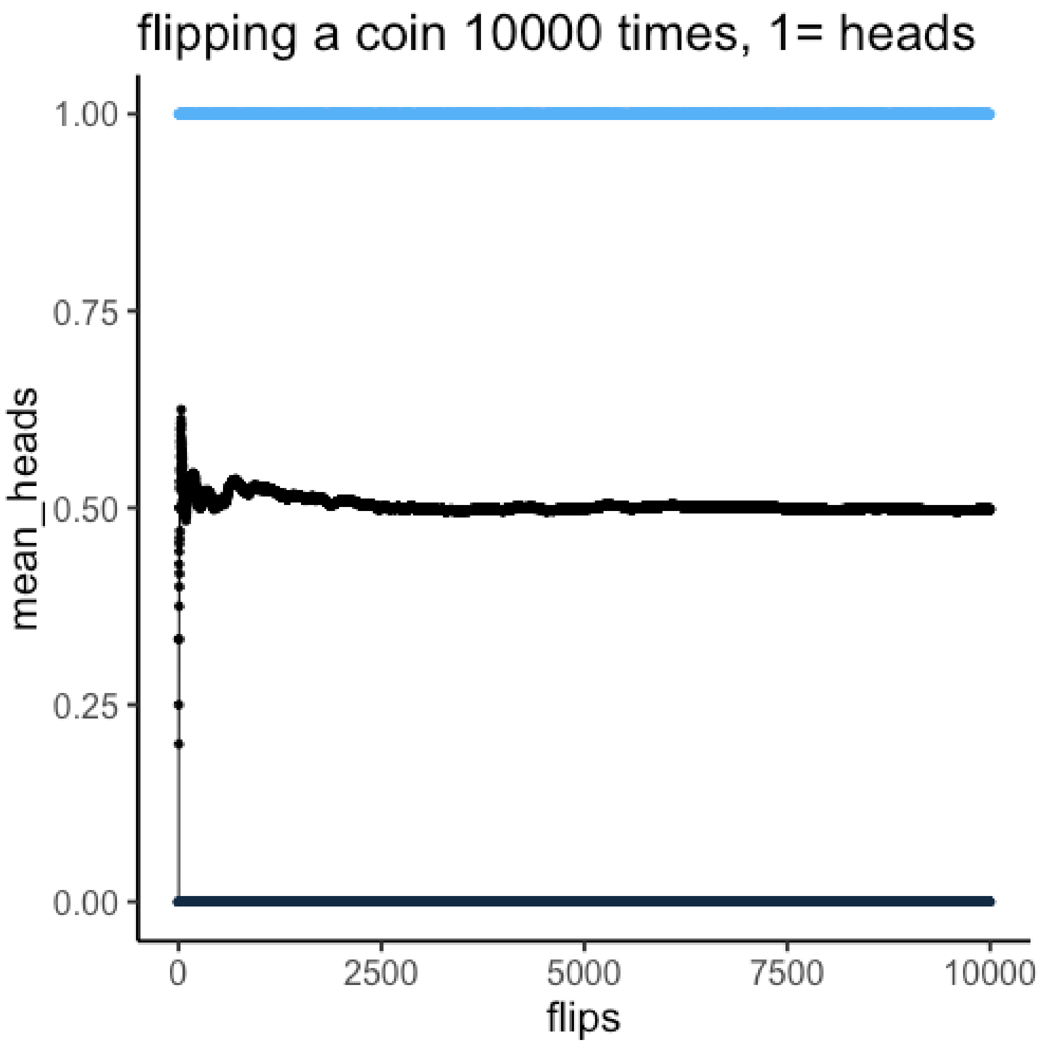
Flipping a coin 100 times



Four simulations



Flipping a coin 10000 times



coin flipping summary

1. 50% heads/tails means that **over the long run**, you should get half heads and half tails

2. When sample size (number of flips) is small, you can "randomly" get more or less than 50% heads

3. Chance is lumpy

Simulations

- Golem in action
- Generate data from your DGP
- Simulated samples converge towards population distribution with increasing sample size

“10% chance of rain tomorrow”

- Discuss:

- What is your interpretation of this statement if you’re a frequentist vs. Bayesian?

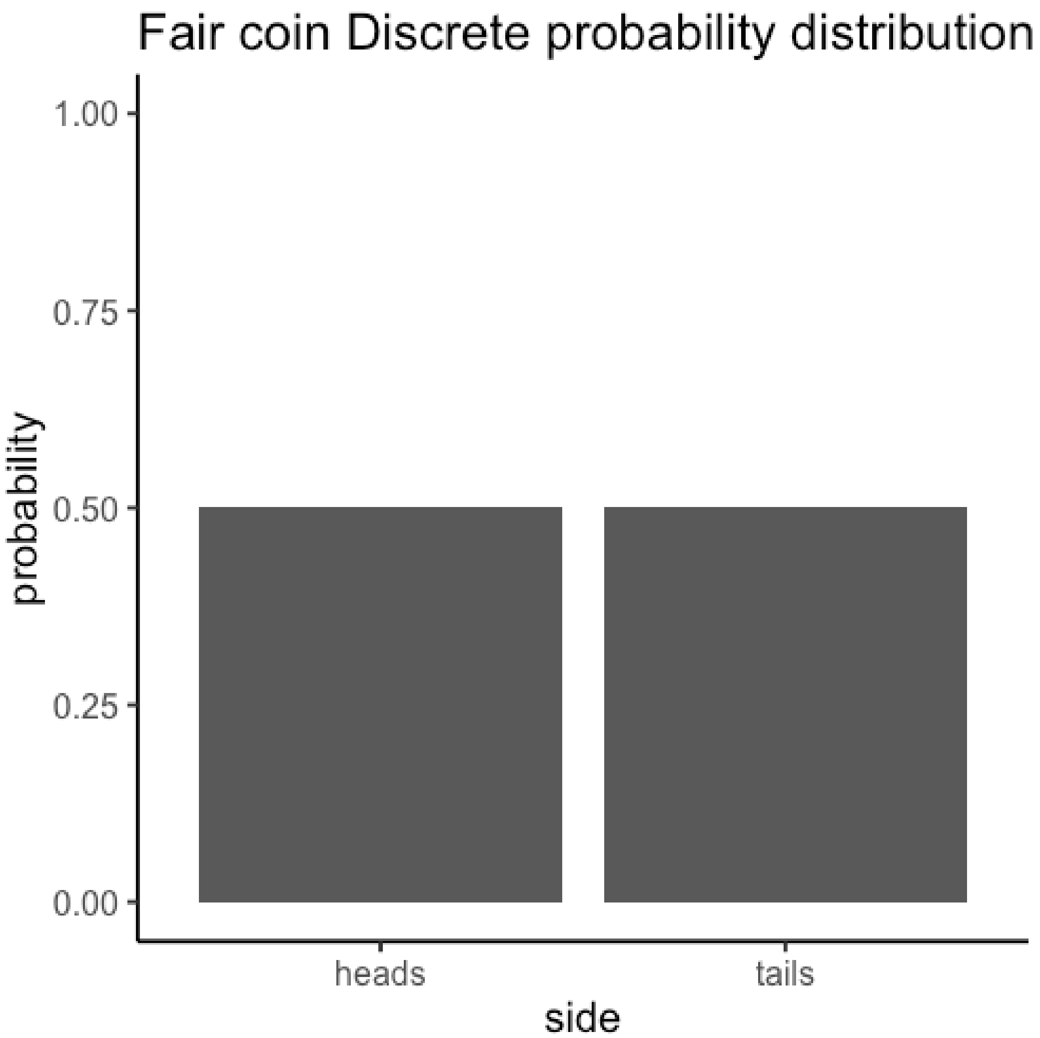
A fair coin

- **Frequentist:** If you flip this coin an infinity of times, **in the long run** half of the outcome will be heads, and half will be tails
- **Bayesian:** I am uncertain about the outcome, I can't predict what it will be.

Discrete probability distributions

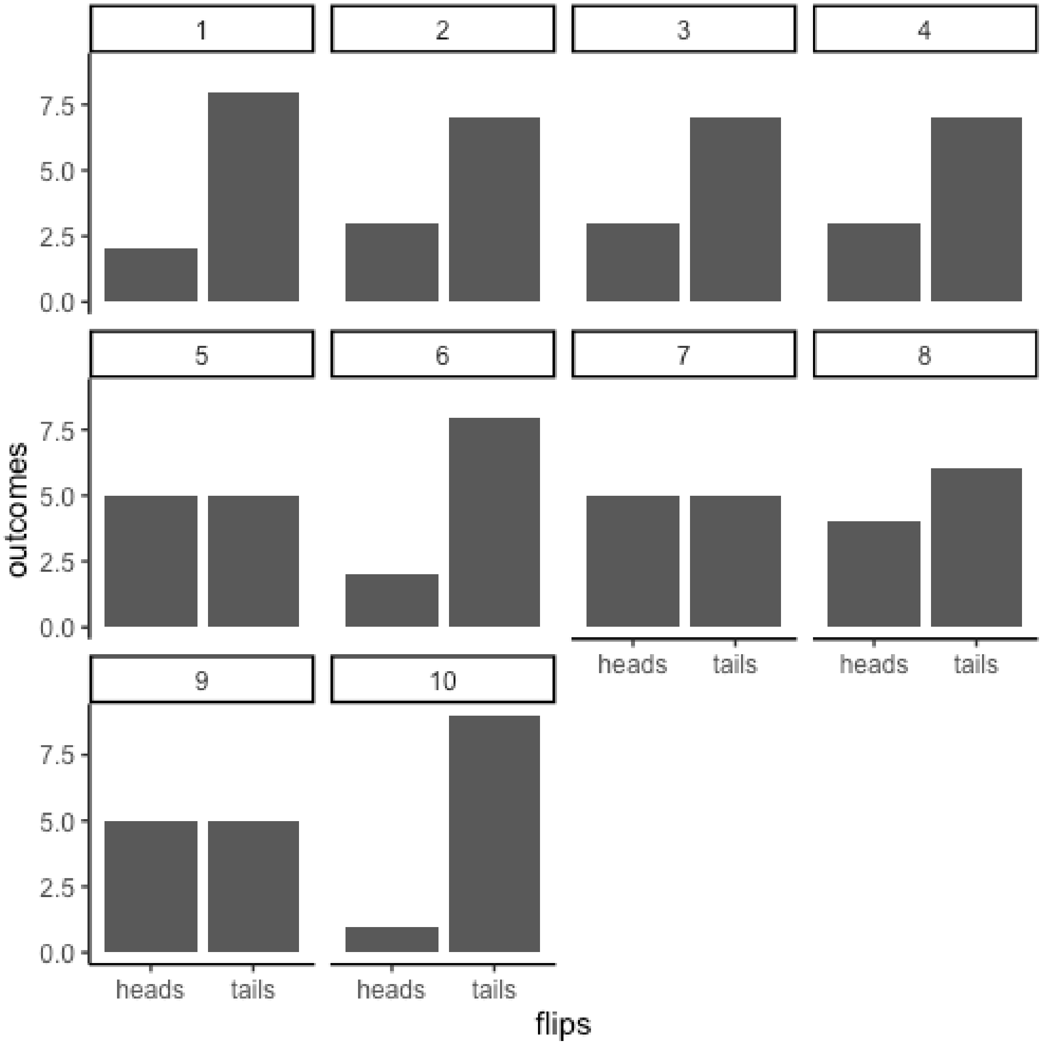
- Defines the probability of each item in a set.
- All probabilities must add up to 1

Coin flipping distribution



What can the coin flipping distribution do?

10 sets of 10 flips



Explaining Variability

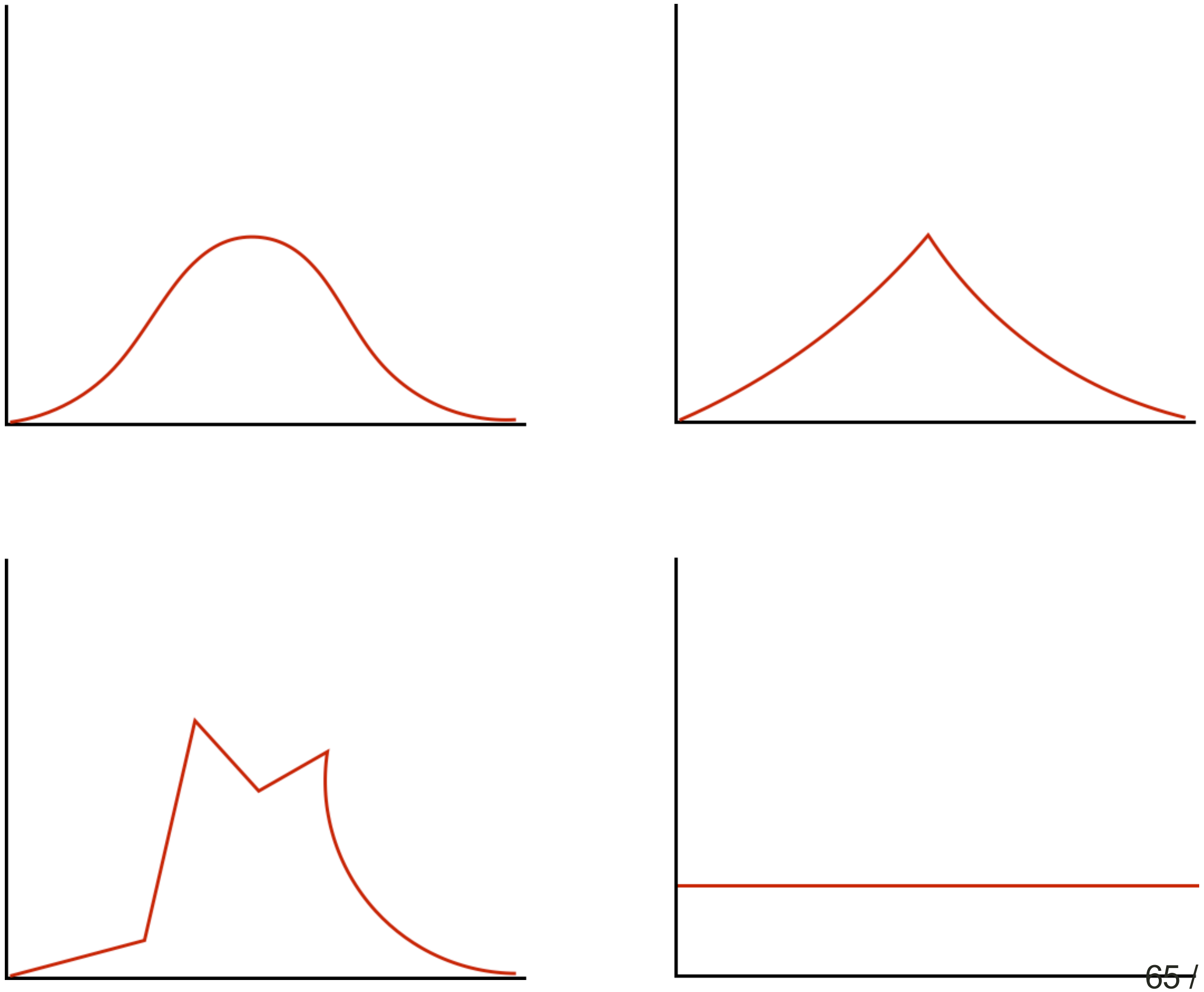
Does knowing someone's value on an explanatory variable, give us information about their value on the outcome variable?

Distributions

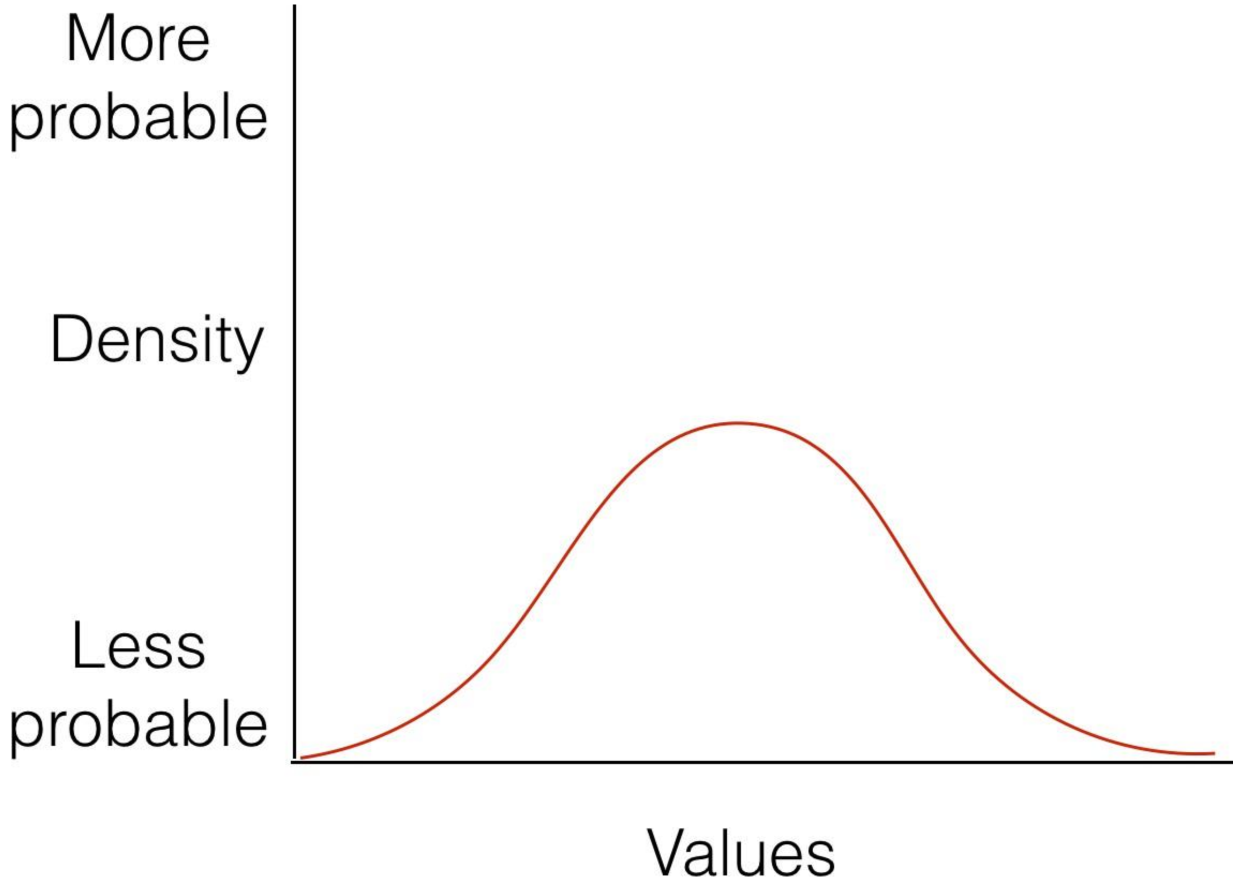
Distributions

1. A tool to define the chances of getting particular numbers
2. Distributions have shapes
3. Higher values indicate higher chance of getting a value

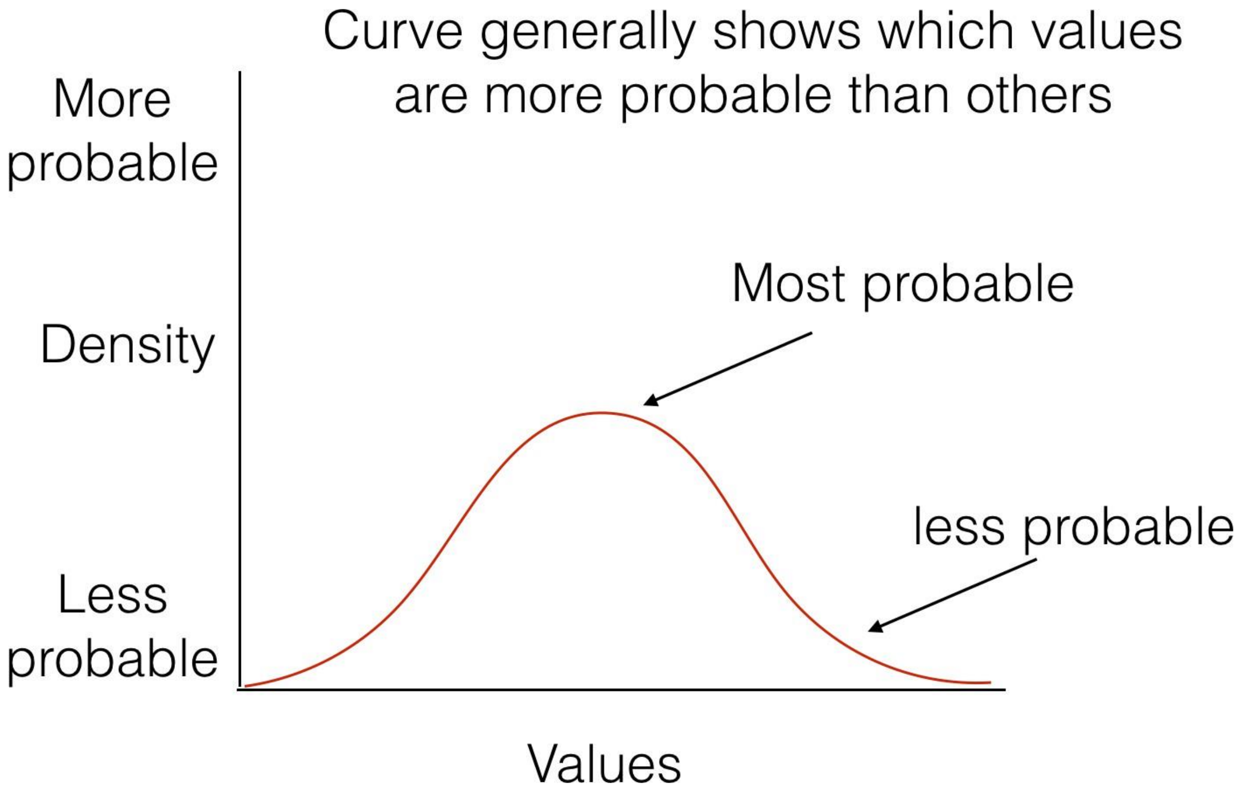
Distributions have shapes



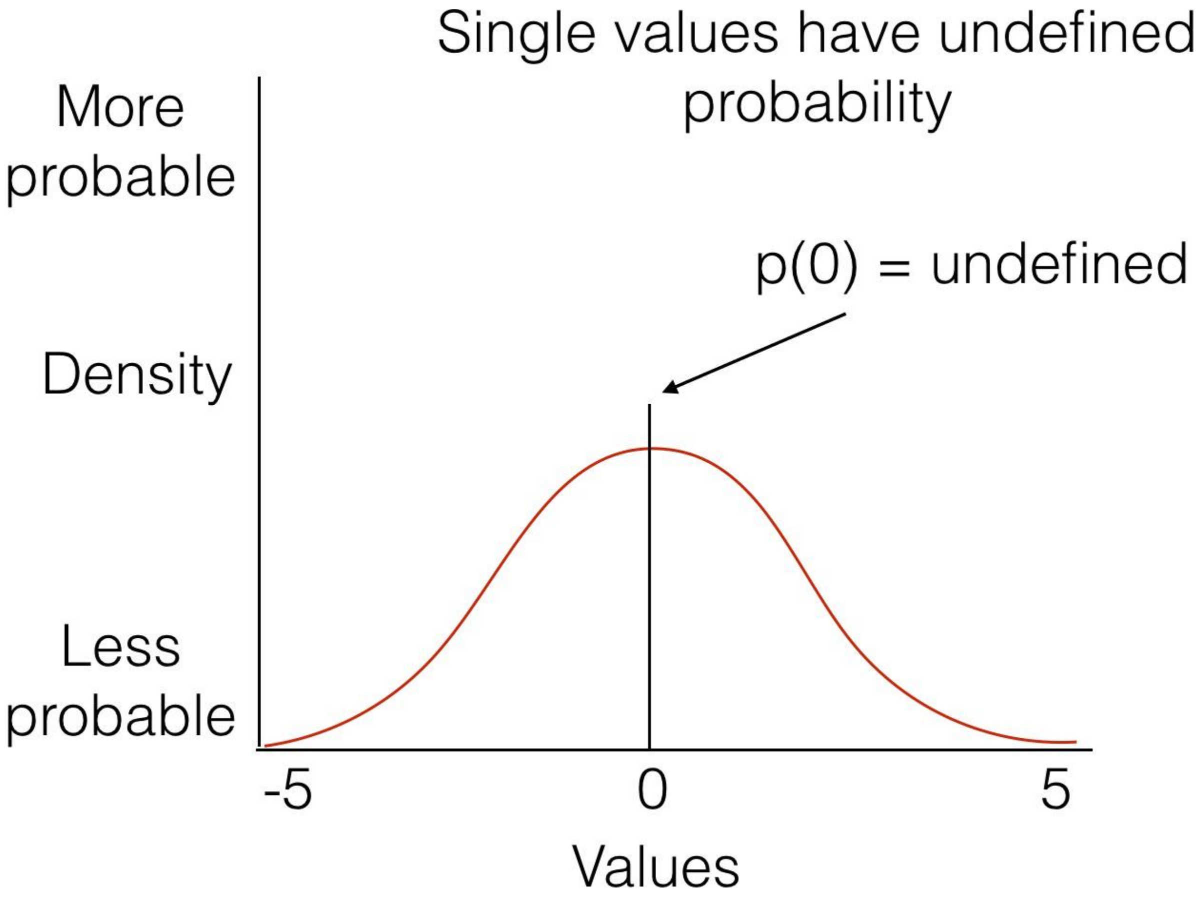
Area under the curve



Interpreting distributions

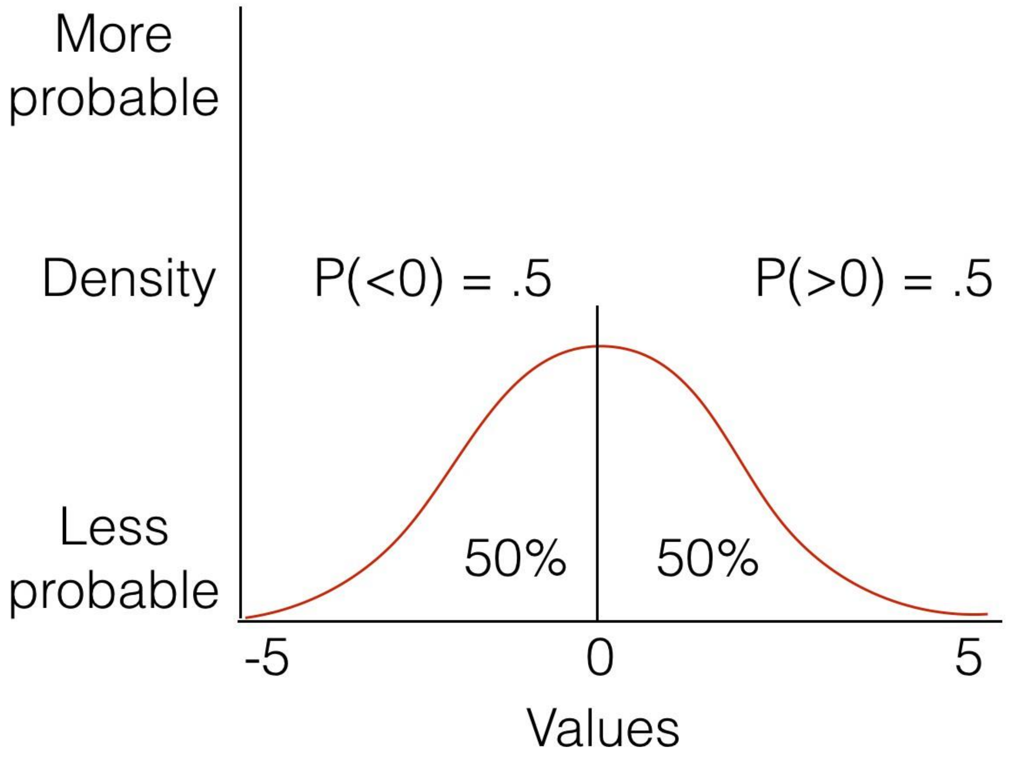


Point Estimates



Probability ranges

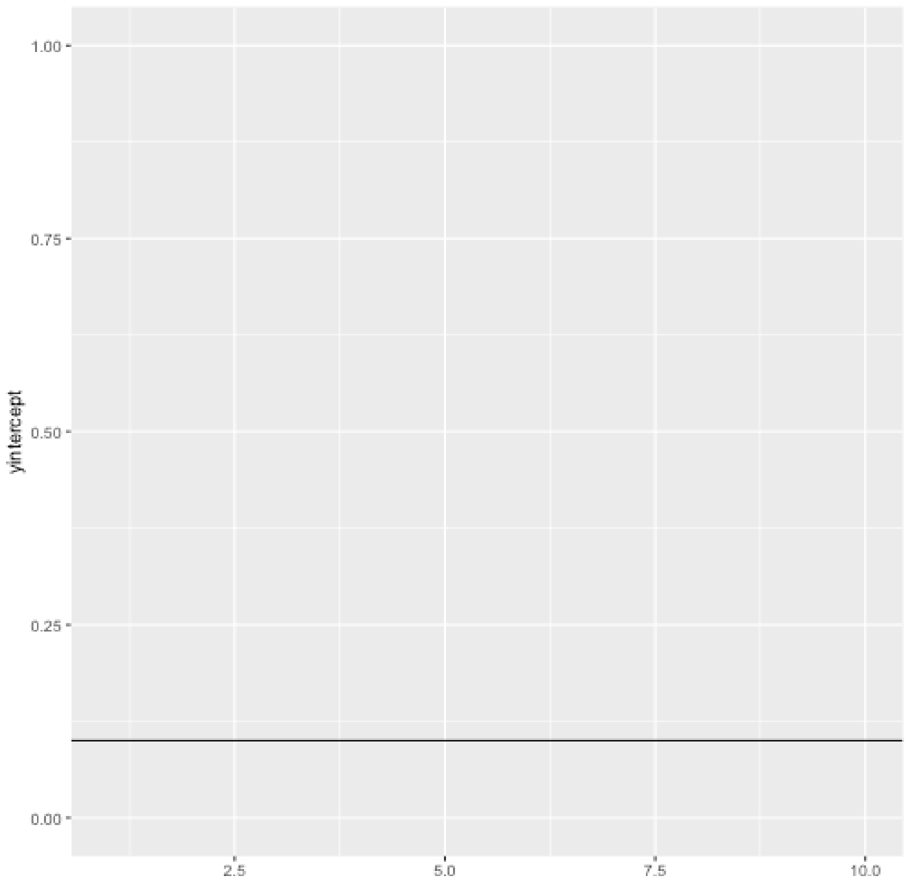
Ranges of values have defined probability



Uniform Distribution

- Definition:
  - 1. All numbers in a particular range have an equal (uniform) chance of occuring

Uniform Distribution



Sampling from a uniform

R let's you sample numbers from a uniform distribution

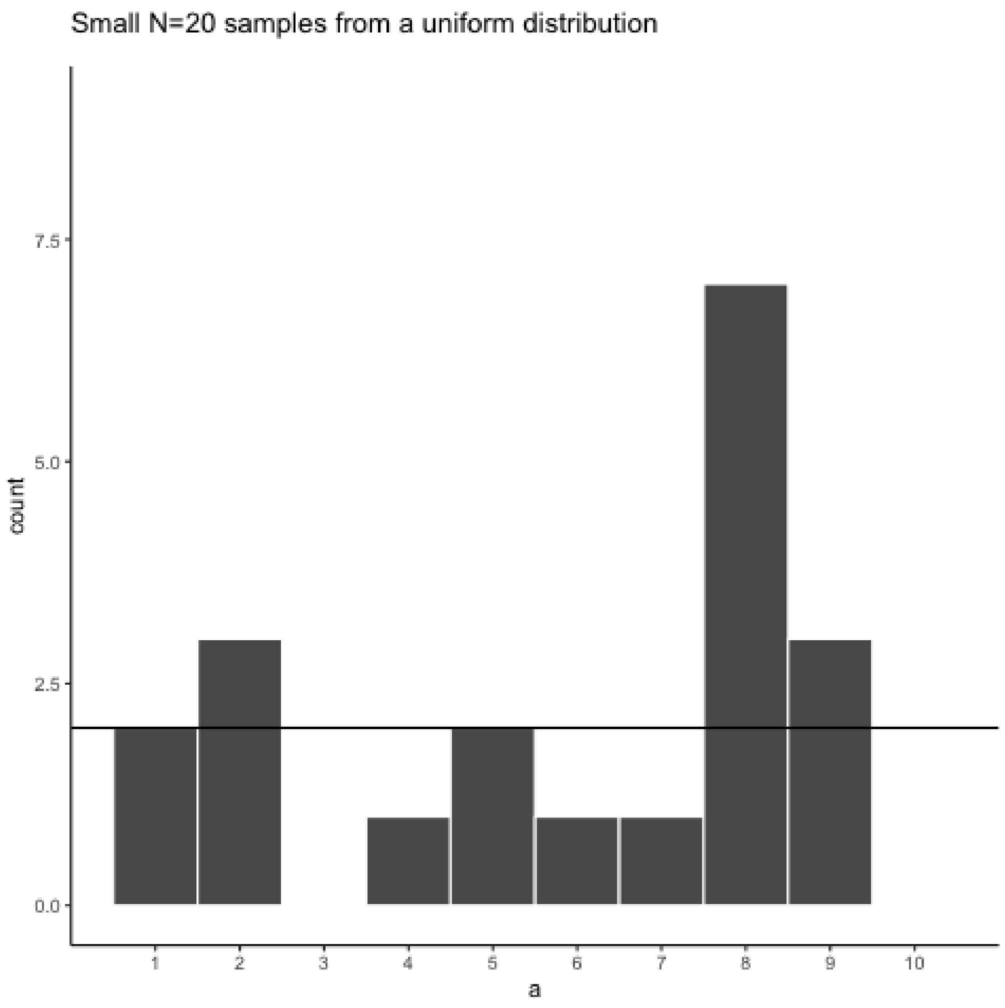
```
runif(n=3,min=0,max=10)
```

```
##[1] 0.3192023 8.1330977 1.6446916
```

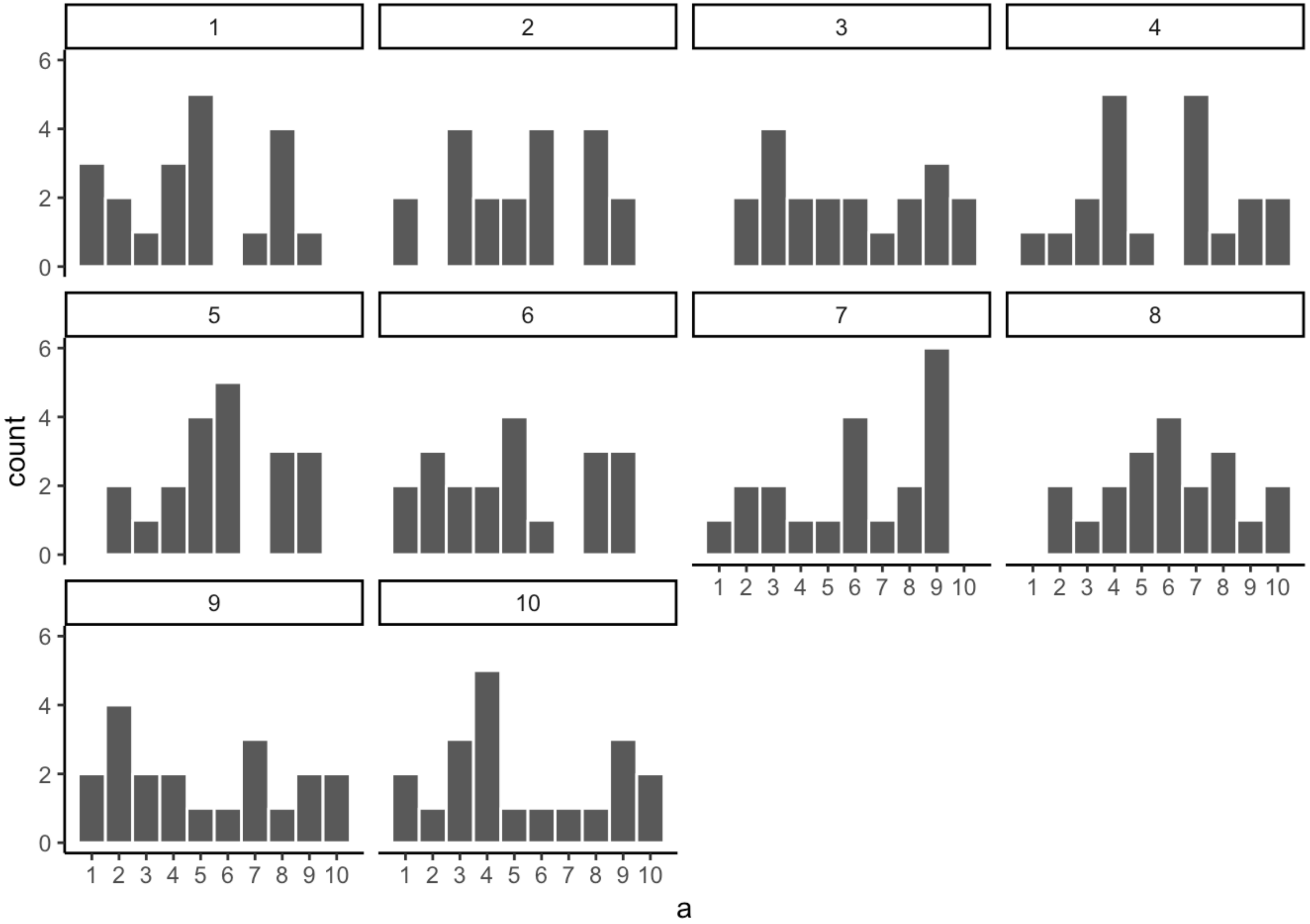
```
runif(n=3,min=0,max=10)
```

```
##[1] 7.185575 6.397575 6.017511
```

looking at samples



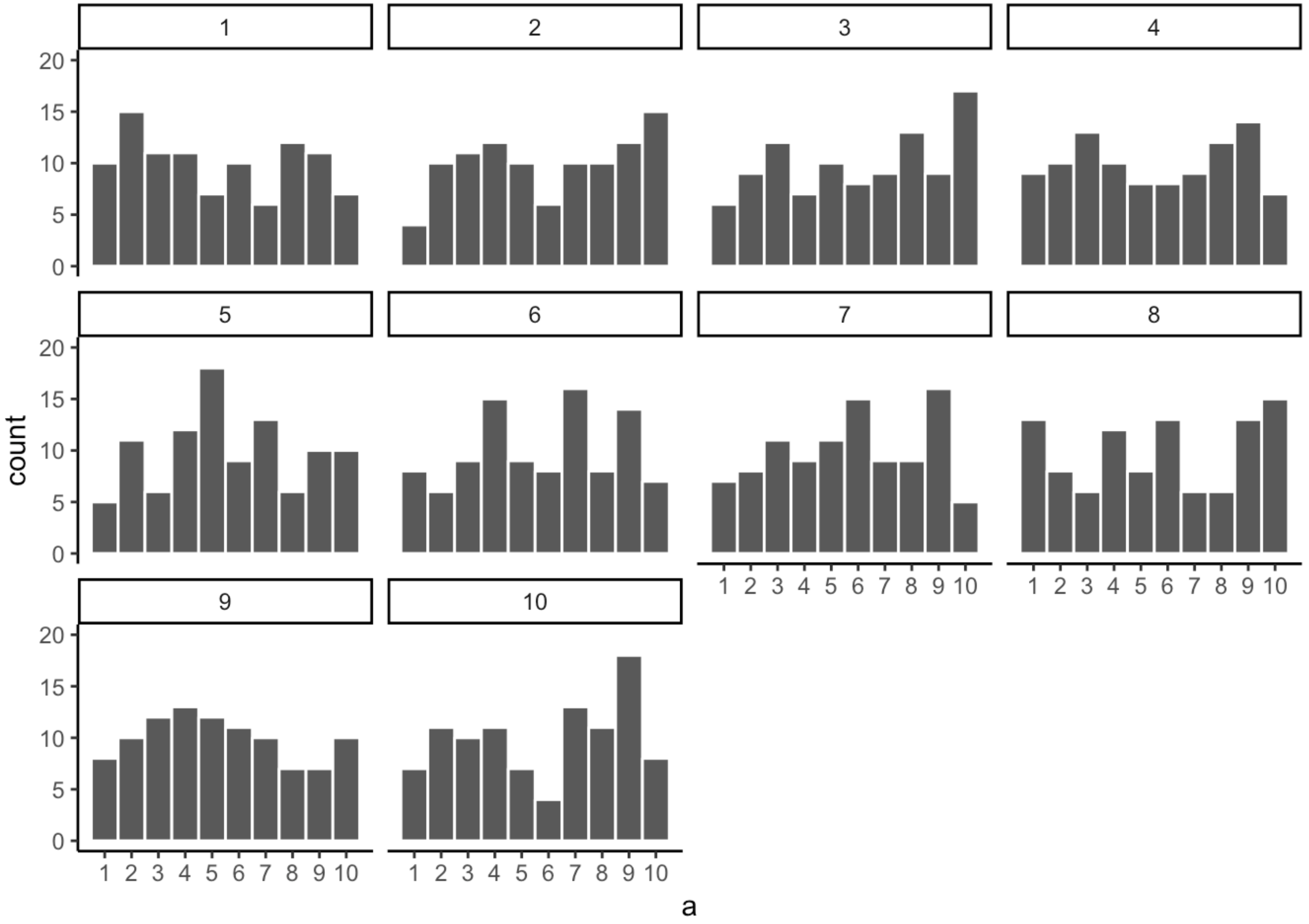
Random samples are not all the same



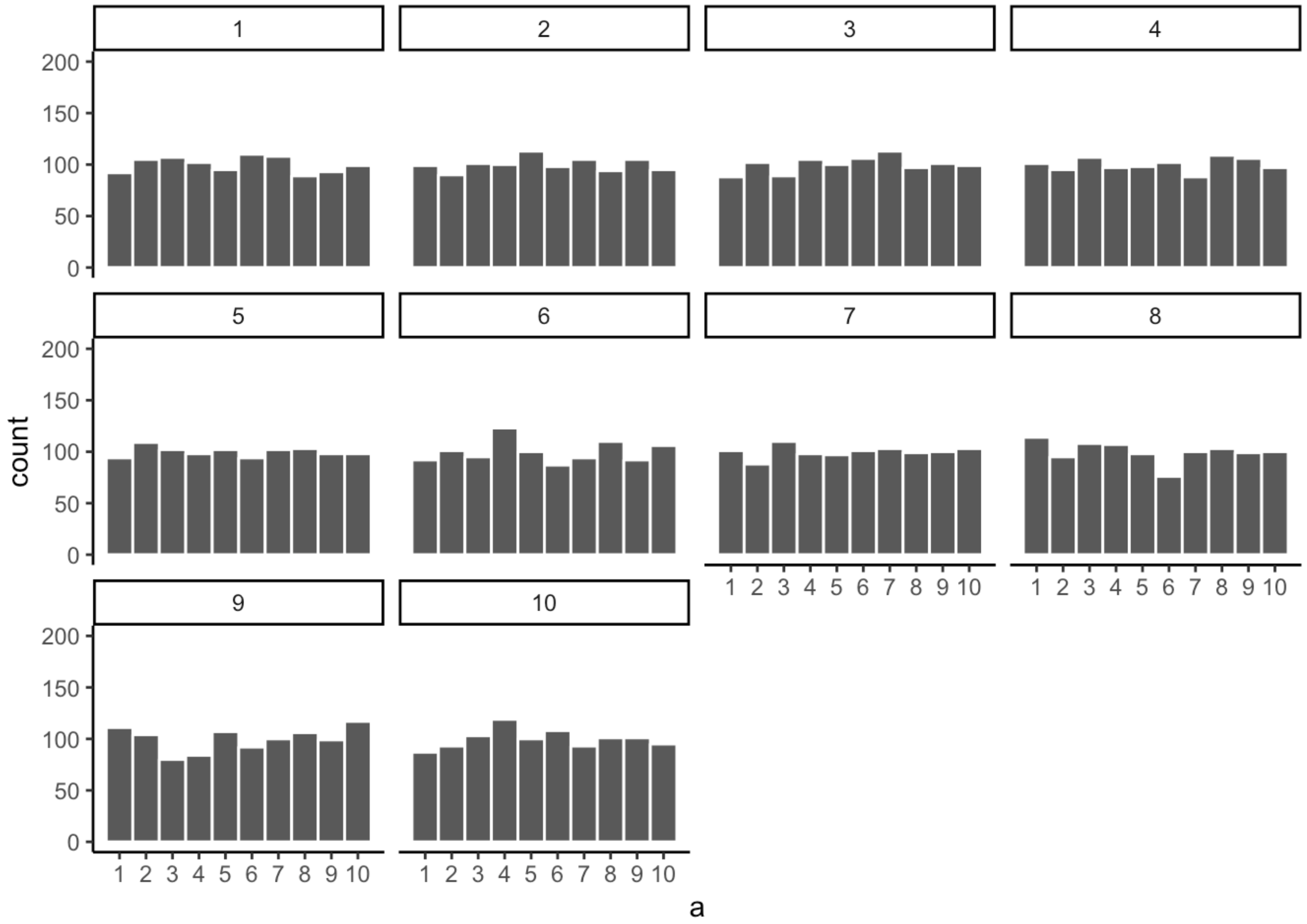
Samples estimate the distribution

1. Samples are sets of numbers taken from a distribution
2. **Samples become more like the distribution they came from, as sample size (N) increases**

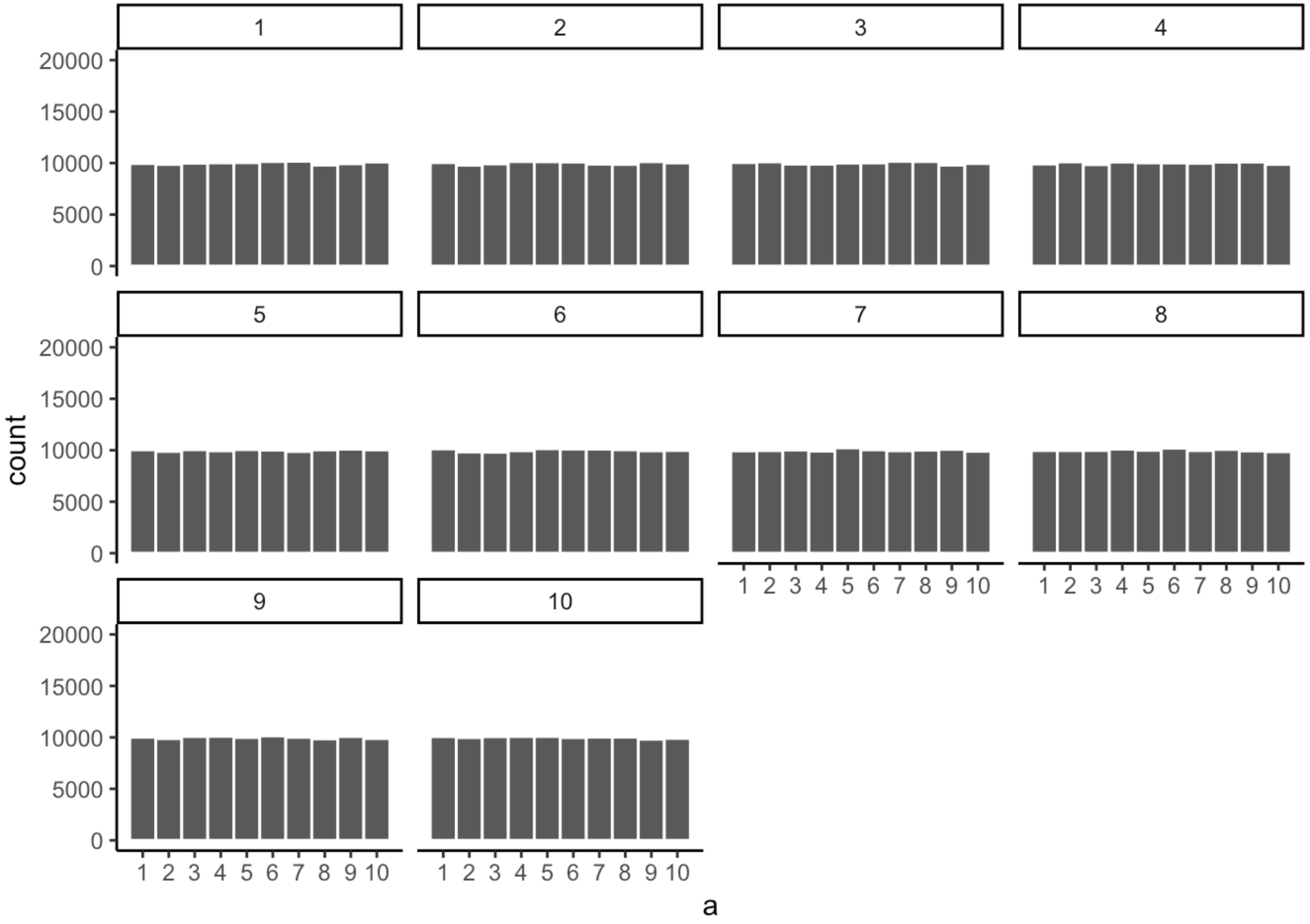
Uniform:
N=100



Uniform:
N=1,000



Uniform: N=100,000



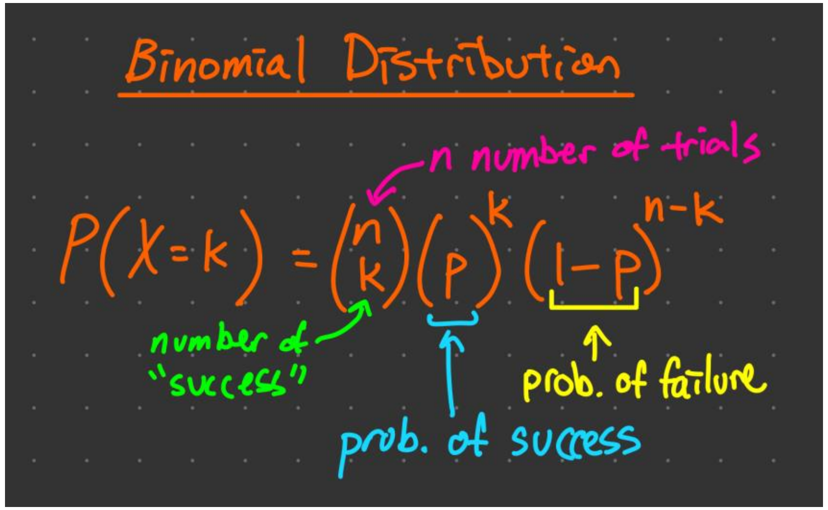
Binomial Distribution

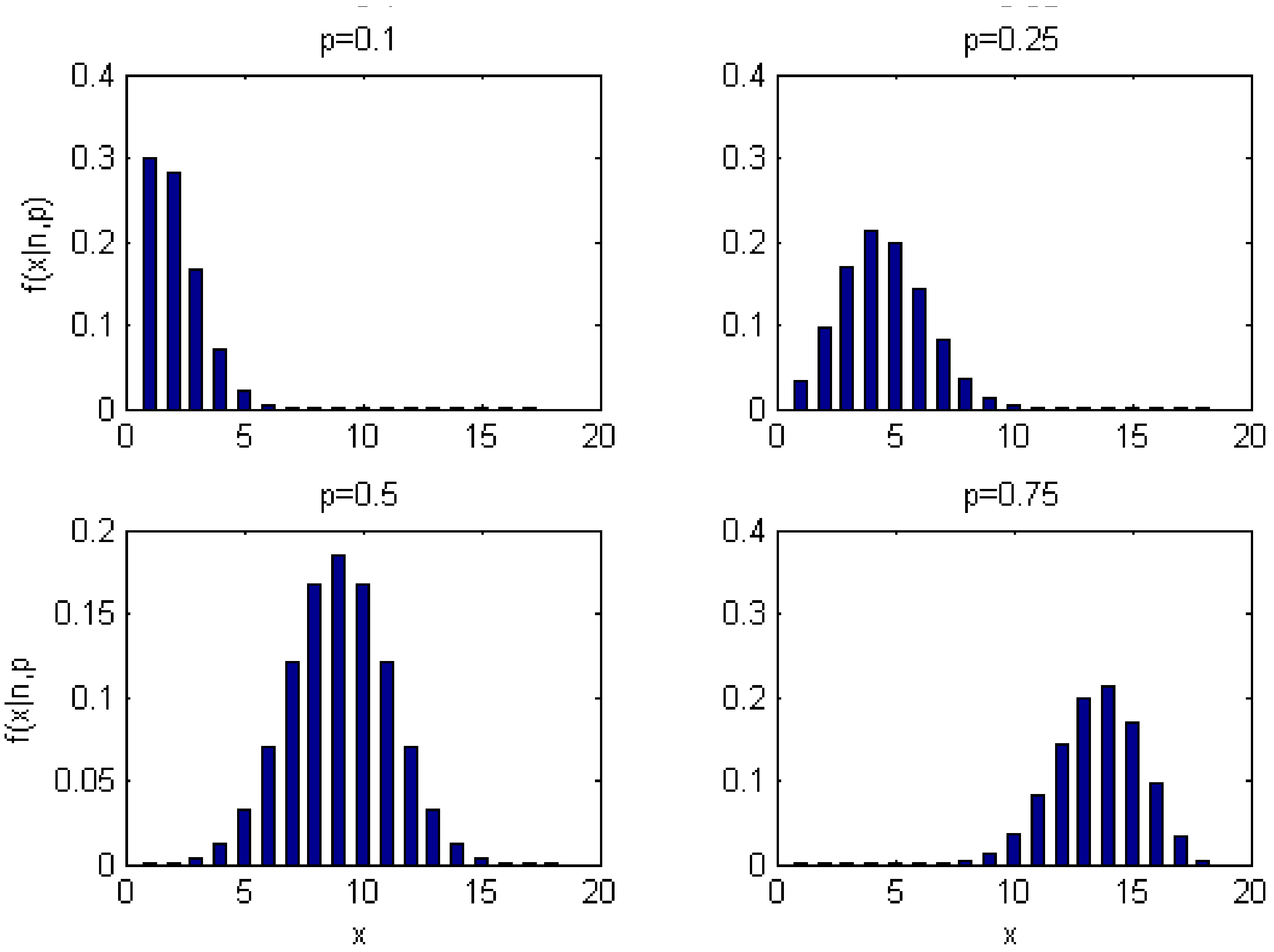
Models repeated binary trials

Example: Flipping Coins

Parameters:
n (trials),
θ or p (success probability)

X ~ Binomial(θ, N)
X = number of successes
N = number of trials
θ or p= probability of success per trial




p=0.1
p=0.25
p=0.5
p=0.75
f(x|n,p)
f(x|n,p
x
x
0.4
0.3
0.2
0.1
0
0.15
0.05
5
10
15
20

Example question:
10 coin tosses
What’s p(exactly 4 heads)?

R Functions for Binomial
dbinom(): probability of exact value
pbinom(): cumulative probability
qbinom(): quantiles
rbinom(): random generation

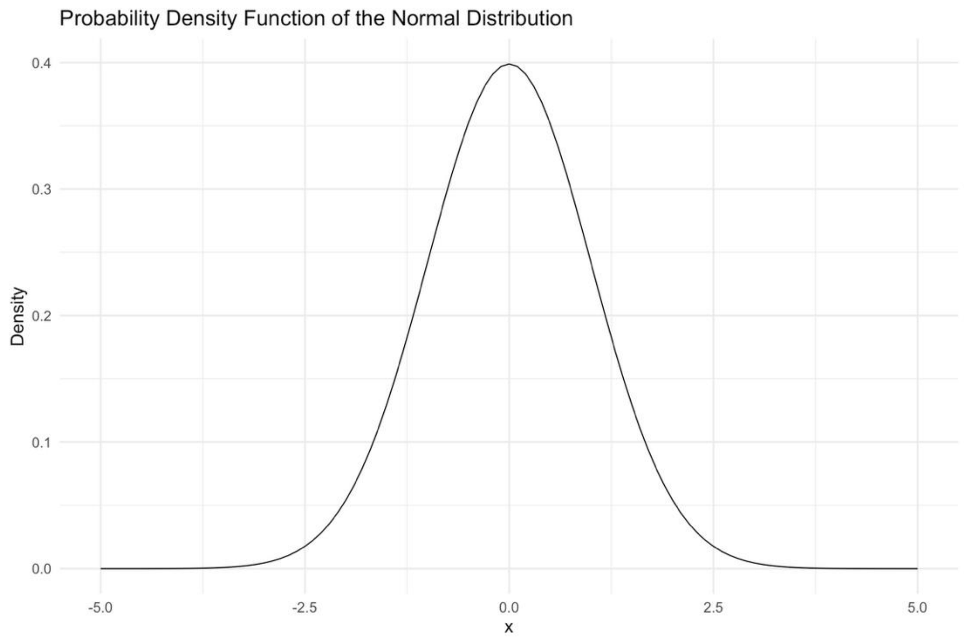
Normal Distribution

Sometimes called a Gaussian distribution

If we assume a variable is at least normally distributed can make many inferences!

Most of the statistical models assume normal distribution

Error in linear models is assumed to distributed as normal

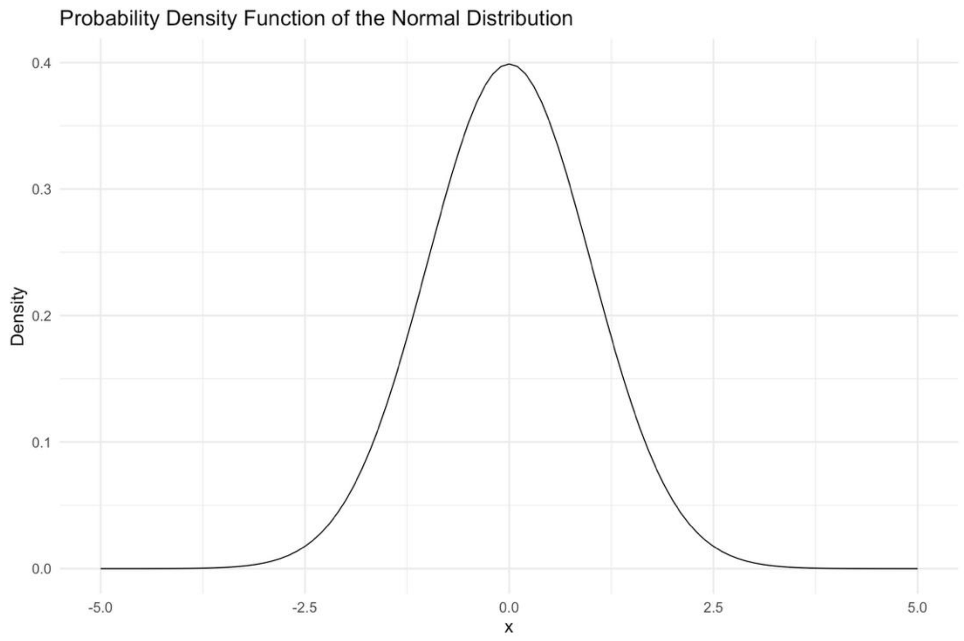


Normal Distribution

Properties

- Symmetric around mean
- Bell-shaped
- Continuous distribution
- Area under curve = 1

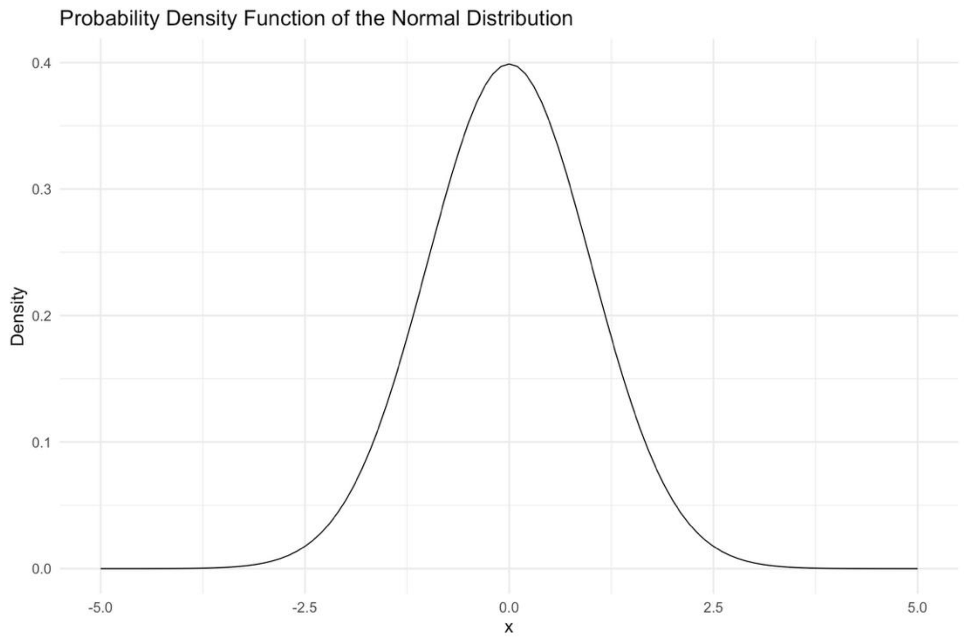
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Normal Distribution

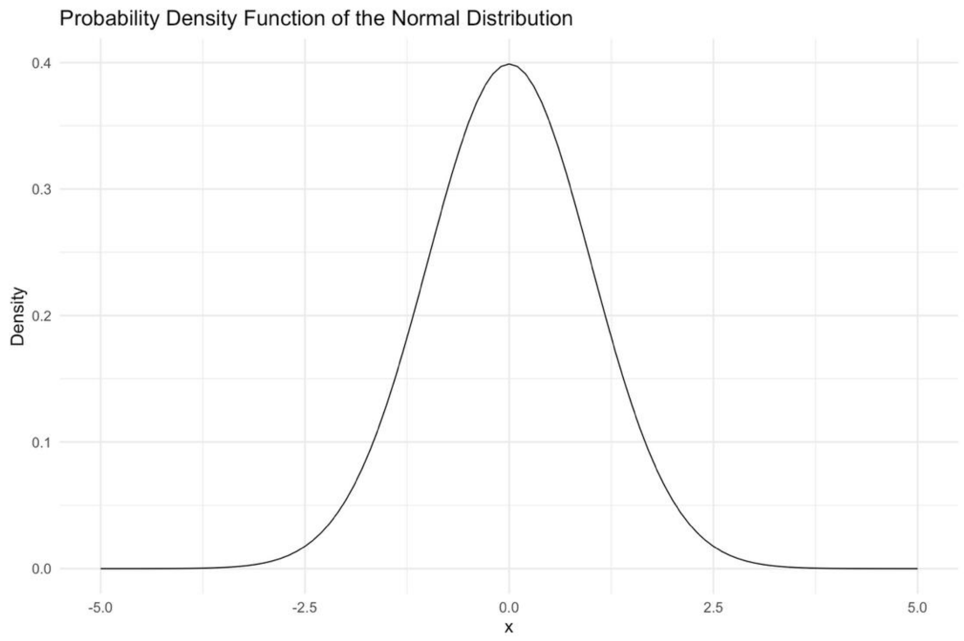
Properties

- 68.3% within 1 standard deviation
- 95.4% within 2 standard deviations
- 99.7% within 3 standard deviations



Normal Distribution

R functions
dnorm(): density
pnorm(): cumulative probability
qnorm(): quantiles
rnorm(): random generation



Normal Distribution

Source of other distributions

- Normal → (square) → Chi-square (χ^2)

  Sum of squared normal variables
  Always positive
  Skewed right
  Used in categorical data analysis

- Normal / √(Chi-square/df) → **t**

  Similar to normal but heavier tails
  Used when σ is unknown
  Degrees of freedom parameter

- $\text{Chi-square}_1/\text{Chi-square}_2$ → F

  Normal → (square) → Chi-square
  Normal / √(Chi-square/df) → t
  $\text{Chi-square}_1/\text{Chi-square}_2$ → F