



BOISE STATE UNIVERSITY

CS 533

INTRO TO DATA SCIENCE

Instructor: Jun Zhuang

PROBABILITY

Sets

A **set** $S = \{a, b, c\}$ is an unordered collection of distinct elements

No duplicates

Elements have no intrinsic order (order can be imposed from elsewhere)

Relationships:

- $a \in A$ means a is an element or member of A
- $A' \subseteq A$ means A' is a subset of A : every element of A' is also in A
- $|A|$ is the *size* (or *cardinality*) of A (can be infinite)

Set Operations

- $A \cup B$ is the **union**: all items in A , B , or both
- $A \cap B$ is the **intersection**: items in both A and B
- $A \setminus B$ is the **difference**: items in A but not B
- A^c is the **complement**: if A is a subset of some larger universe ($A \subseteq \mathcal{A}$), the set of all items *not* in A ($A^c = \mathcal{A} \setminus A$)

Events

A set E of **elementary events** – distinct individual outcomes

Coin flip: $E = \{H, T\}$

D6: $E = \{\square, \blacksquare, \blacklozenge, \blacktriangle, \blacktriangledown, \blacksquare\}$

An **event** is a set $A \subseteq E$

An event A occurs if the outcome $\xi \in A$

Elementary events are *singletons*: $A = \{H\}$ means ‘coin is heads’

E , the set of all elementary events, means ‘something happened’

$A = \{\square, \blacksquare, \blacklozenge\}$ is the event ‘rolled even number’

Logic

Set operations describe logical events:

- $A \cap B$ – both A and B happened
- $A \cup B$ – either A or B happened (or both, if compatible)
- $A \setminus B$ – A happened but not B

Can also write with logical operators:

- $A \wedge B$
- $A \vee B$
- $A \wedge \neg B$

6-sided Die

Foundational Elements

- $E = \{ \square, \square, \square, \square, \square, \square \}$
- $A_{\square} = \{ \square \}$; $A_{\square} \dots A_{\square}$ likewise
- $A_{\square}, A_{\square}, A_{\square}, A_{\square}, A_{\square}, A_{\square} \in \mathcal{F}$

Results

- All subsets of E are events
 - This happens for any discrete, finite E
 - This is the *power set* $\mathcal{P}(E)$

Probability

- $P(\{d\}) = \frac{1}{6}$ for all die values d – all values equally likely
- $P(A_{\text{even}}) = \frac{1}{2}$ - 3 of the 6 equally-likely values are even

Some Probability Facts

- $0 \leq P(A) \leq 1$ — probability is in range 0–1
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 - Why? How does this relate to additivity?
 - $P(A \cup B) \leq P(A) + P(B)$
- $P(A^c) = 1 - P(A)$
- $P(A \setminus B) = P(A) - P(A, B)$
- If $A \subseteq B$, $P(A) \leq P(B)$

Joint Probability

The **joint probability** $P(A, B)$ is the probability of both A and B occurring simultaneously.

- $P(A, B) = P(A \cap B)$
- Sometimes written $P(A; B)$

$P(D_1 = \text{4}, D_2 = \text{5})$ — die 1 is 4, die 2 is 5

Not the same as 'a 4 and a 5' – that can happen in 2 orders!

Conditional Probability

The **conditional probability** $P(B|A)$ is the probability B happened given that we know A happened.

Dice: $P(d_2 = \text{even} | d_1 = \text{even}) = \frac{1}{6}$ — first die tells us nothing

$P(d_1 = \text{even} | d_1 \text{ is even}) = \frac{1}{3}$ — there are 3 even rolls (2, 4, 6)

Conditional and Joint

We can decompose joint probabilities into conditional:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Marginal Probability

The *marginal probability* is a $P(A)$; can compute by *marginalizing* the *joint distribution* ($\mathcal{B} = \{B_1, B_2, \dots, B_n\}$ is collection of mutually exclusive events that span E):

$$P(A) = \sum_{B \in \mathcal{B}} P(A, B)$$

$D_1 \downarrow$ $D_2 \rightarrow$	□	▣	▤	▥	▦	▧	Margin
□	1/36	1/36	1/36	1/36	1/36	1/36	1/6
▣	1/36	1/36	1/36	1/36	1/36	1/36	1/6
▤	1/36	1/36	1/36	1/36	1/36	1/36	1/6
▥	1/36	1/36	1/36	1/36	1/36	1/36	1/6
▦	1/36	1/36	1/36	1/36	1/36	1/36	1/6
▧	1/36	1/36	1/36	1/36	1/36	1/36	1/6
Margin	1/6	1/6	1/6	1/6	1/6	1/6	

Independence

Two events are **independent** if knowing one tells you nothing about another:

- $P(A|B) = P(A)$ (equivalently: $P(B|A) = P(B)$)
- $P(A, B) = P(A)P(B)$

Example: $d_1 = \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$ says nothing about d_2

Exercise: prove these definitions are equivalent.

Bayes' Theorem

The Bayes' Rules can be written as:

$$P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

If we know a conditional probability, and the marginals, we can *invert the conditional*.

Wrapping Up

We can use sets to describe events or outcomes, and logical combinations.

Probability quantifies likelihood of different events.

Joint probability: multiple events.

Conditional probability describes an event conditioned on other information.



CONTINUOUS PROBABILITY AND RANDOM VARIABLES

Random Variable

Let X be the value a 6-sided die role

We call X a **random variable** – a variable w/ a random value¹

The **expected value** is

$$E[X] = \sum_{i \in 1 \dots 6} i \cdot P(i)$$

- Mean over many rolls
- Expected “return” from a roll

¹ Technically, a random variable is a function $f_x: E \rightarrow \mathbb{R}$ from elementary events to numeric values.

Expectation and Mean

If we have a sequence x_1, \dots, x_n of data points, and pick one **uniformly at random** ($P(x) = 1/n$), then

$$\begin{aligned} E[X] &= \sum_i x_i P(X = x_i) \\ &= \frac{1}{n} \sum_i x_i \\ &= \bar{x} \end{aligned}$$

Expectation = mean (weighted by probability)

Continuous Variables

We assign probabilities to **intervals**, not individual values

- $E = \mathbb{R}$ (elementary events are real numbers)
- An interval is a subset of \mathbb{R}
- \mathcal{F} is the set of intervals, their complements, and their countable unions
 - It contains infinitesimally small intervals, but not singletons

Continuous Probability

Continuous probabilities have a *distribution function*:

$$F(x) = P(X < x)$$

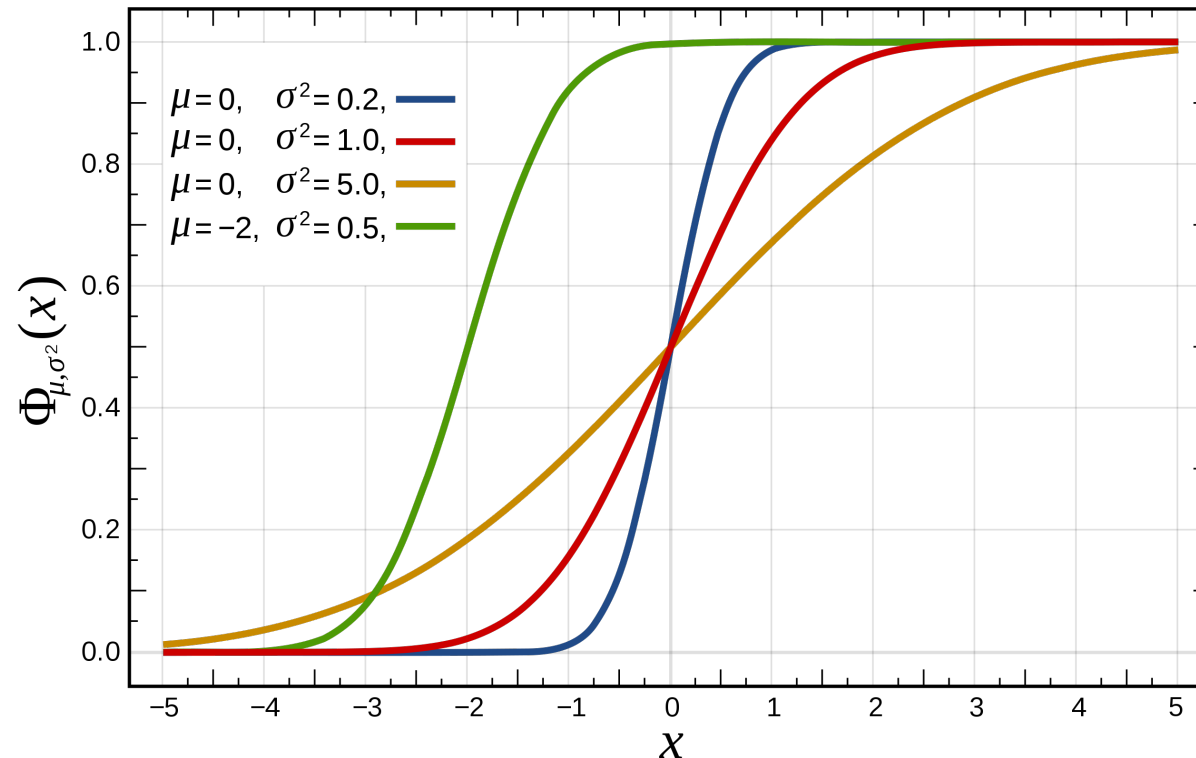


Figure from Wikipedia.

Continuous Probability

Continuous probabilities have a *distribution function*:

$$F(x) = P(X < x)$$

We can compute probabilities for any interval:

$$P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$$

The probability of an interval is the *probability mass*

Also term for probability of discrete event!

Probability Density

Distributions are often defined by a *density function* p such that

$$F(x) = \int_{-\infty}^x p(x) dx$$

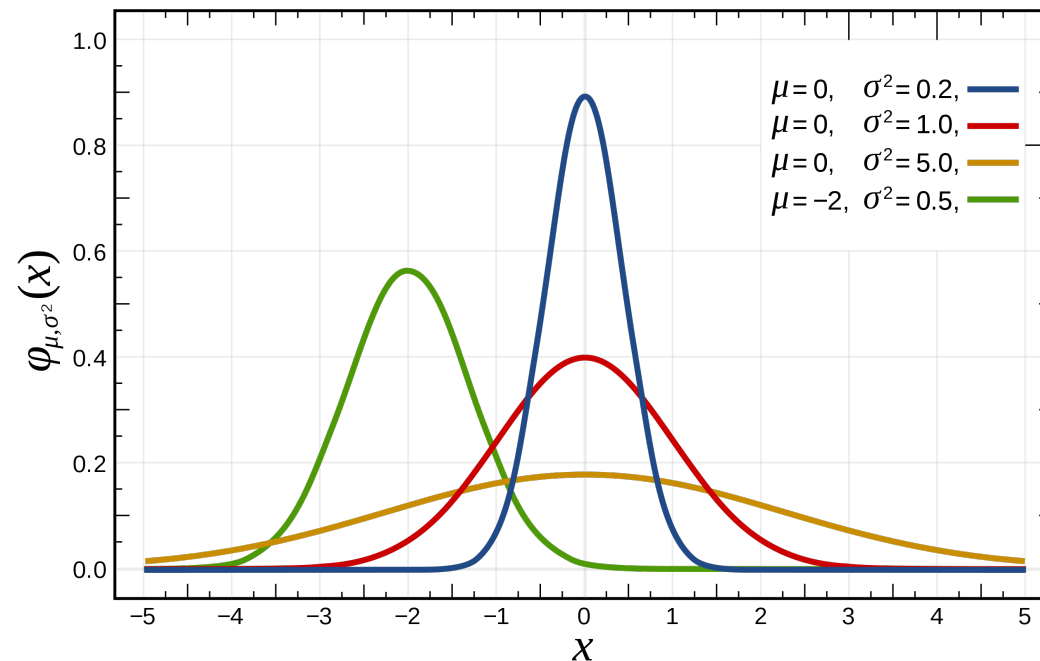


Figure from Wikipedia.

Probability Density

Distributions are often defined by a *density function* p such that

$$F(x) = \int_{-\infty}^x p(x) dx$$

- Densities can exceed 1
- Densities are **not** probabilities

- $p(x) = \lim_{\epsilon \rightarrow 0} \frac{P(x-\epsilon \leq X < x+\epsilon)}{2\epsilon}$

Continuous Expectation

Expectation of continuous random variables is mean, weighted by density:

$$E[X] = \int x p(x) dx$$



Wrapping Up

The probability of any single value of a continuous variable is effectively zero.

Instead, we use probability density, distribution functions, and assign probability to intervals.

Expectation is the *mean* of a random variable.

DISTRIBUTIONS

Numeric Distributions

We're going to focus on distributions of *numeric variables*

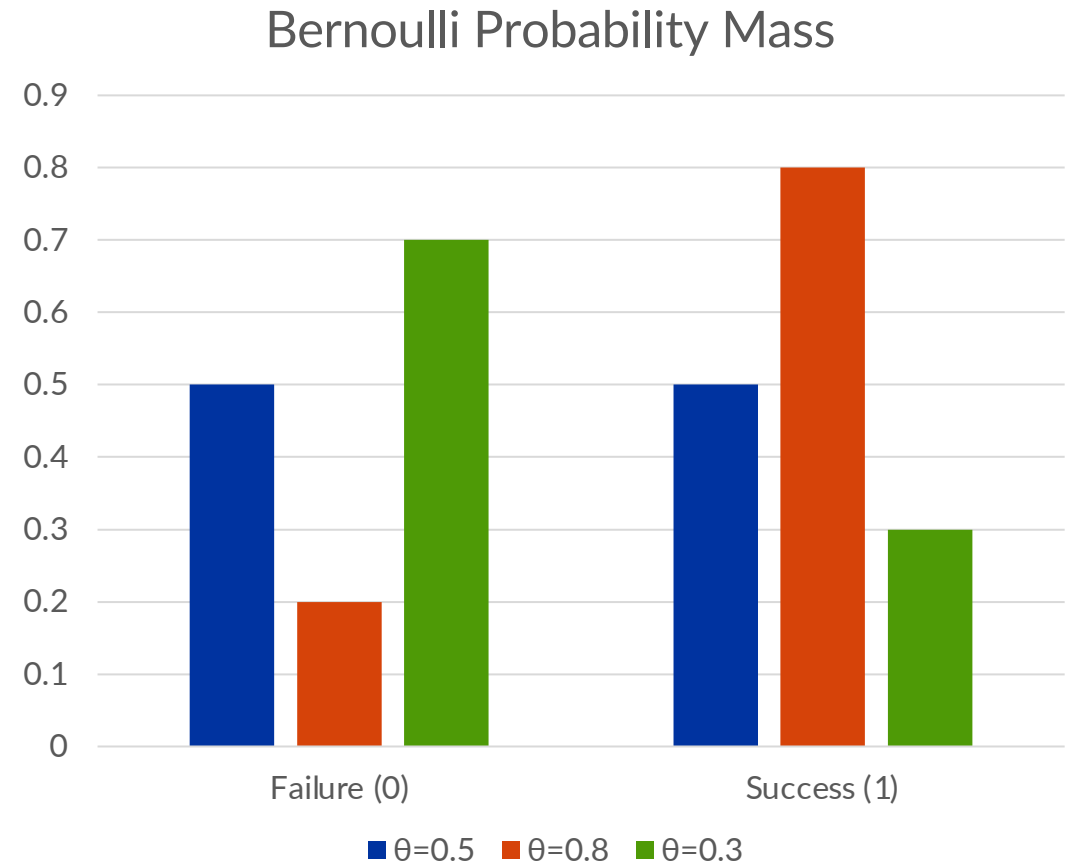
- Both discrete and continuous
- Categorical: encoded to 0, 1, etc.

Bernoulli Distribution

- Binary outcomes (0/1)
 - Success and Failure
- Parameter: θ (success prob.)

Statistics:

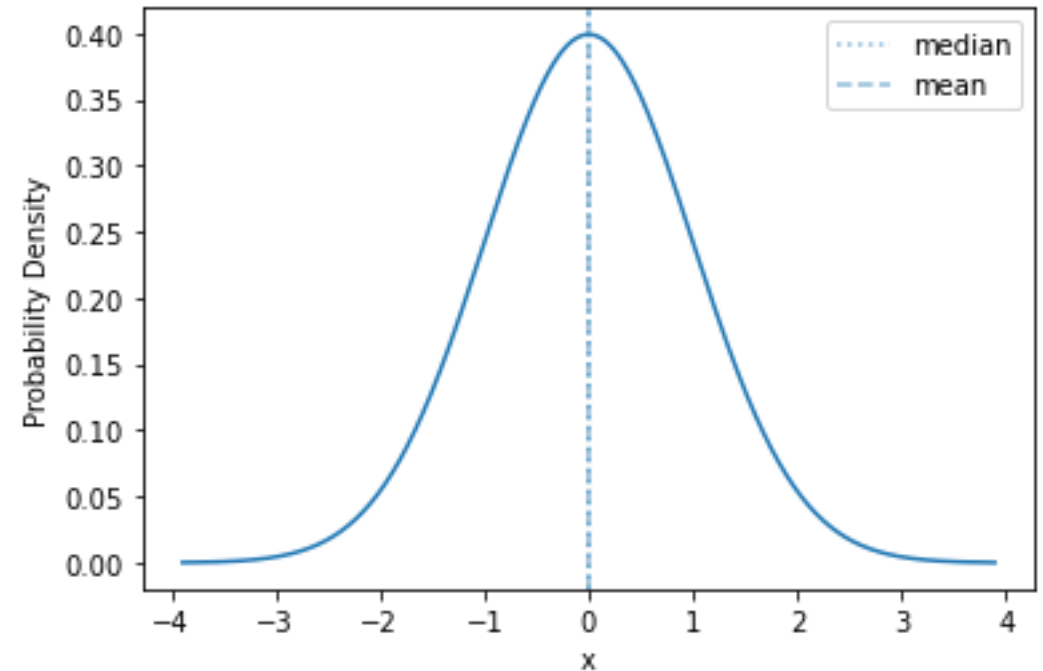
- Mean: θ
- Mode: higher-prob value



Characterizing Distributions

Parameters:

- location, scale, shape



Characterizing Distributions

Parameters:

- location, scale, shape

Density or mass function

Support (range it's defined over)

Underlying *random process*

Key Statistics

- Mean (*first moment*)
- Standard deviation or variance
 - Variance is *second central moment*
- Median (50th %ile)

Binomial

Parameters:

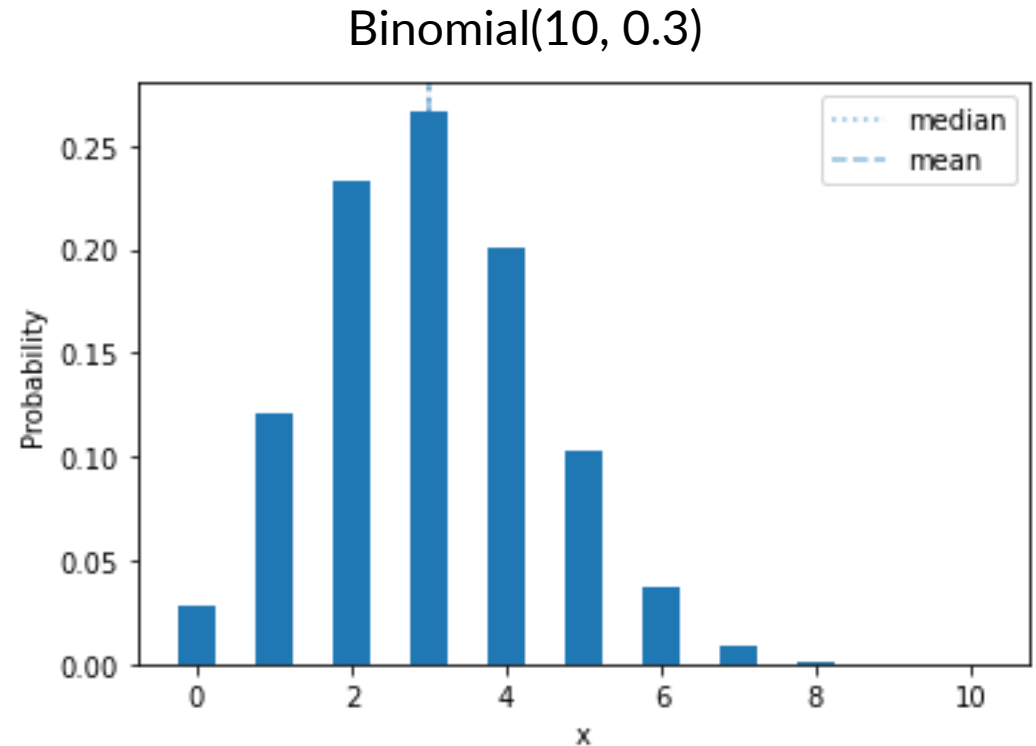
- n – number of trials
- θ – success probability

Process: n Bernoulli trials

Support: integers $[0, n]$

PMF:

$$P(y|n, \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$



Normal (Gaussian)

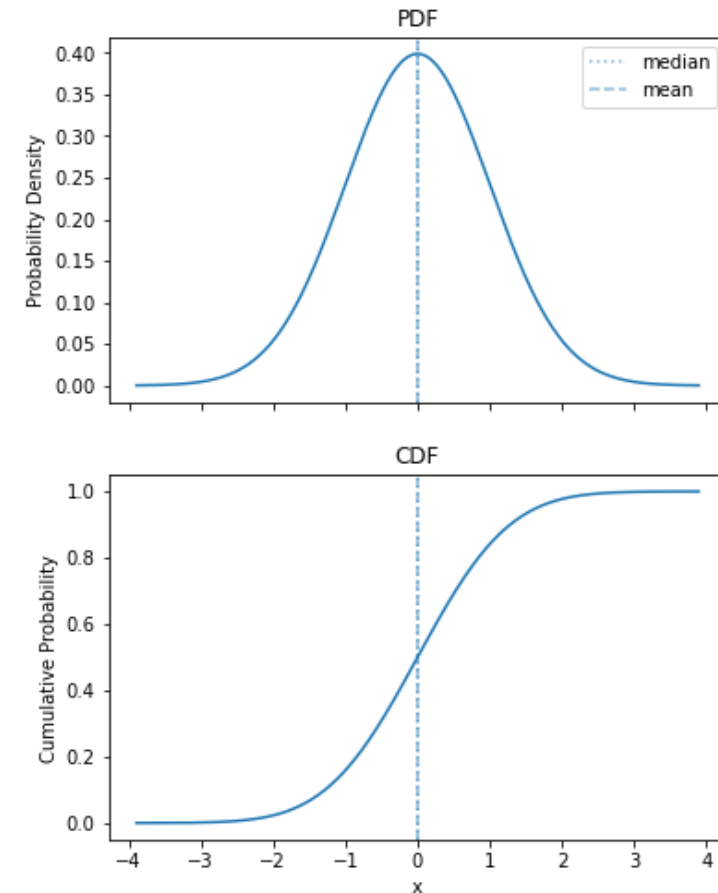
Parameters:

- μ – mean, location
- σ – standard deviation, scale

PDF:

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\mu = 0, \sigma = 1$ is **standard normal**
Normal is a **location-scale** dist.





Wrapping Up

Random variables are often described by *probability distributions*, which may have *parameters*.

There are many standard probability distributions.

See notebook + resources for more!

Photo by [JOSHUA COLEMAN](#) on [Unsplash](#)

SAMPLING AND THE DATA GENERATION PROCESS

Sampling

The inferential logic of statistics is based on *samples*

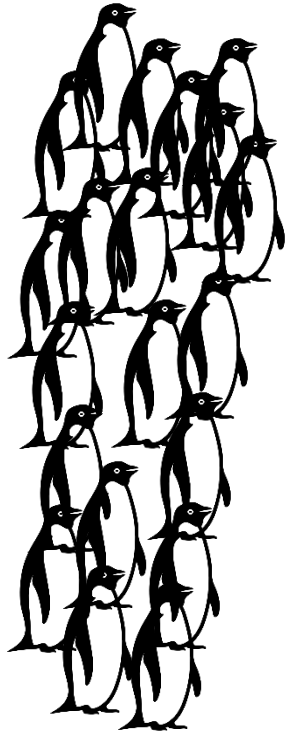
- From a distribution
 - Generate random numbers!
- From a population
 - Select them with a representative sampling strategy

Sampling the Population

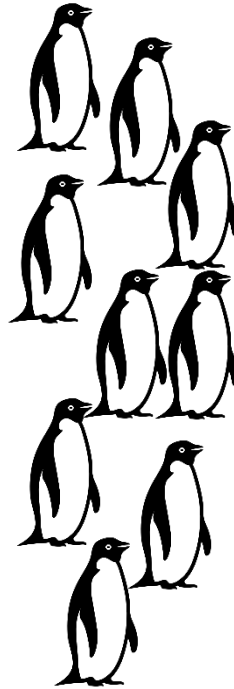
Ideal Penguin



Penguins
(Population)
Flipper length μ_{fl}



Sample of
Penguins



Statistic

$$\bar{x} = \frac{1}{n} \sum x_i$$

Is the sample **representative**?
Does it teach us about population?

Representative Samples

We need a couple of things for a sample:

- **Representative** of the population (w.r.t. parameter of interest)
 - Biases affect this (sampling, selection, response, etc.)
- **Large enough** to allow inference of parameter of interest
 - This size does not depend on population size!

Better data often better than more data!

Historically, much statistics concerned w/ efficiently using samples

Uniform Sampling

- All population members equally likely to be sampled
 - Harder than it sounds in practice
- Pros: Resulting statistical analysis relatively straightforward
- Cons: Small subgroups easy to omit!

More Strategies

Stratified Sampling

- Make sure different groups are represented, possibly equally

Oversampling

- Sample more from a minority group for in-group data
- Correct (resample or reweight) for whole-sample inference

Penguins

We have:

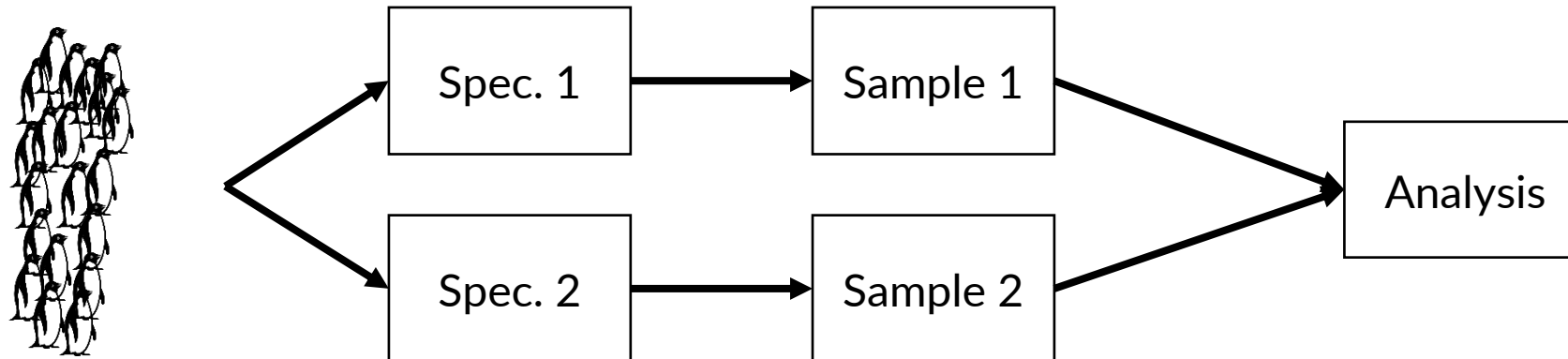
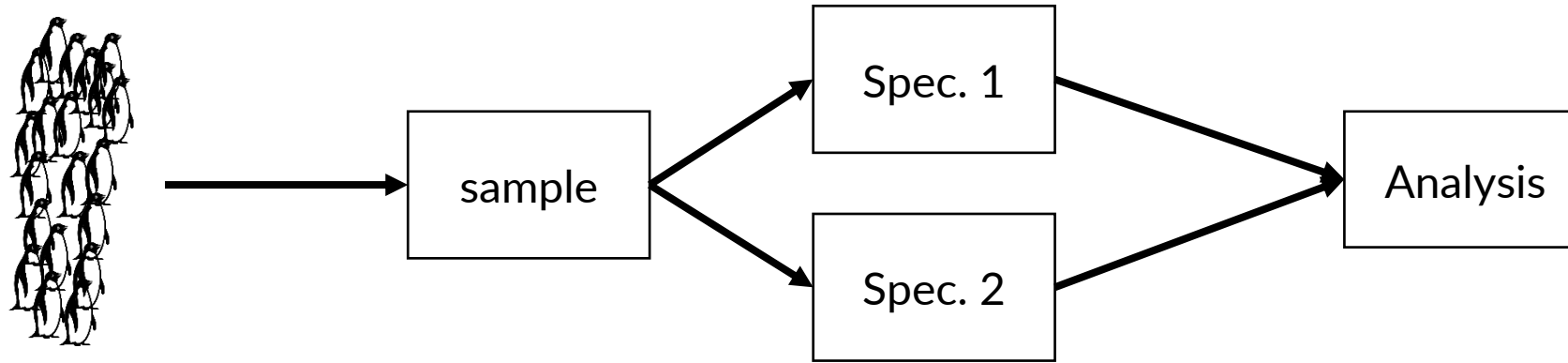
- 3 species of penguin
- Measurements for a sample within each

What is the population?

Can we answer:

- Distribution of penguin species?
- Typical measurements within a species?

Two Sampling Strategies



The Data Generating Process

How did we get our data?

1. People and movies exist
2. People find movies and watch them
3. Netflix recommends more movies
4. People maybe watch them (feed back into 2)

Reasoning about DGP helps identify sample status,

I.I.D.

Common desiderata for values, and samples:

independent and identically distributed

- Independent: one value does not affect another
- Identically distributed: all drawn from the same distribution
 - Equal mean, variance, distribution family

Uniform at random from large pop is i.i.d.

Small pop: sampling removes items! (unless with replacement)

German Tank Problem



You capture a tank.

It has serial number 2089.

How many tanks does the enemy have?

This is inference for the *max*.



Wrapping Up

Our data comes by some process.

Classically, we think about this in terms of *sampling* – how did we pick these items to analyze?

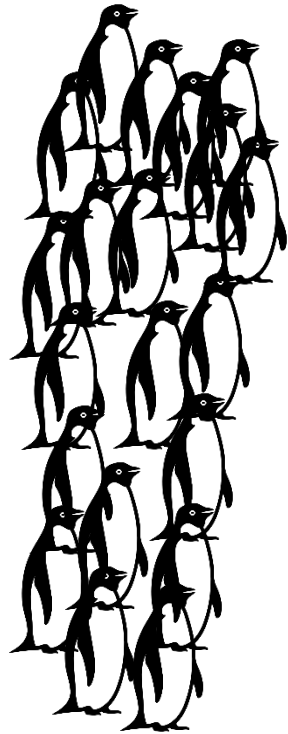
The data generation process is how our data comes into existence.

Photo by [Museums Victoria](#) on [Unsplash](#)

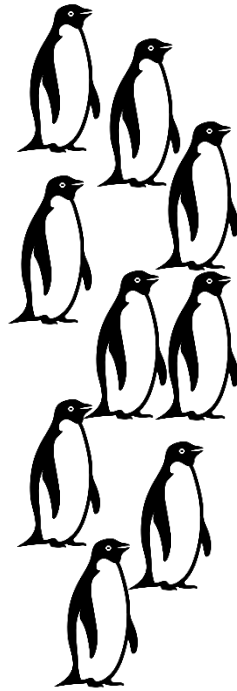
SAMPLING DISTRIBUTIONS AND CONFIDENCE INTERVALS

Sampling the Population

Penguins
(Population)
Flipper length μ_{fl}



Sample of
Penguins



Statistic

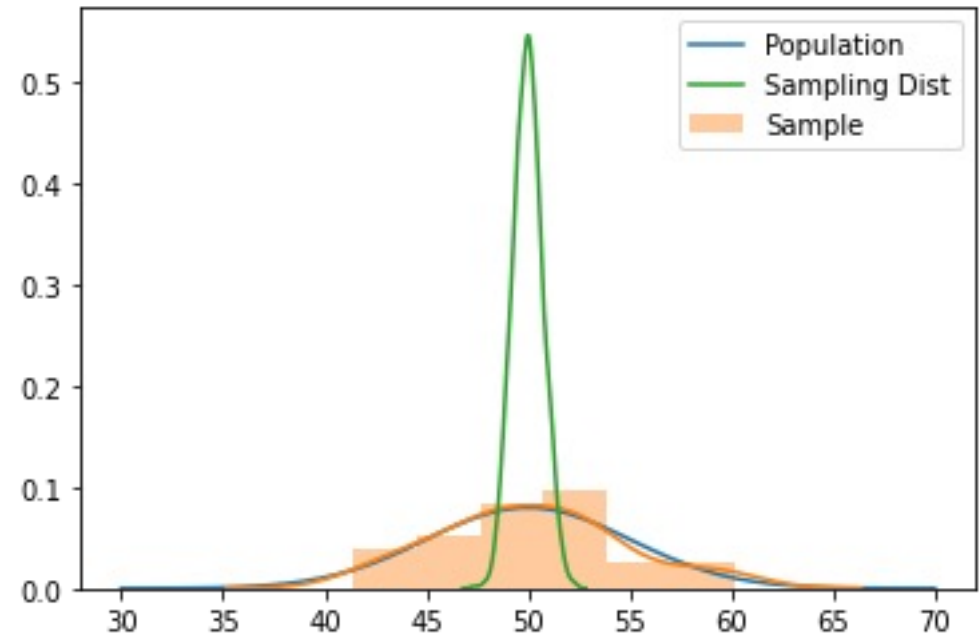
$$\bar{x} = \frac{1}{n} \sum x_i$$

Sampling Distributions

How does a statistic relate to the underlying parameter?

Classical frequentist statistics considers the outcome of **repeating the experiment**.

This gives us the **sampling distribution** of the statistic.



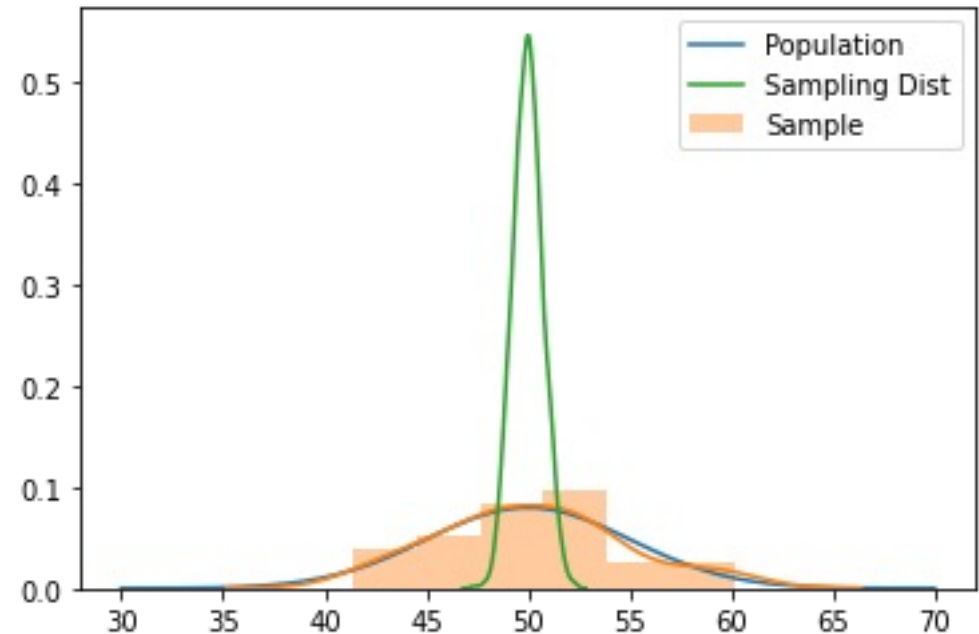
Sampling Distributions

Experiment:

1. Take sample
2. Compute statistic

On repeat:

- Do that ∞ times
- What's the distribution of the statistic?



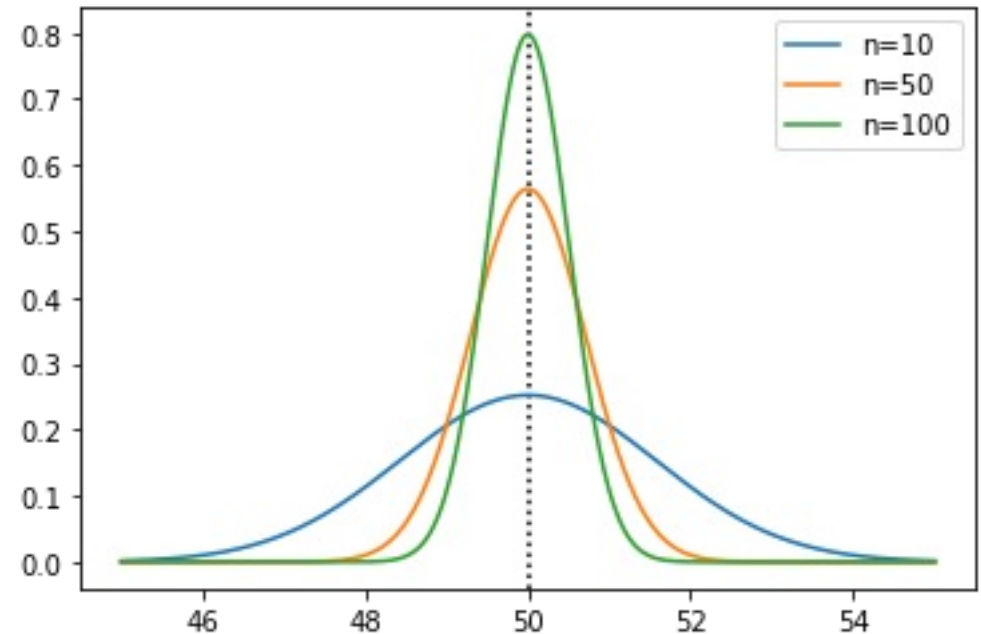
Sampling Distribution of the Sample Mean

- Take sample of size n
- Compute sample mean \bar{x}

Then:

$$\bar{x} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

X doesn't need to be normal!



Confidence Interval

95% confidence interval (CI):

$$\bar{x} \pm 1.96 \left(\frac{s}{\sqrt{n}} \right)$$

1. Take sample
2. Compute CI

This procedure returns an interval containing the true mean 95% of the time.

Chinstrap penguins:

- $\bar{x} = 195.82$
- $s = 7.13$
- $n = 68$
- $se = \frac{s}{\sqrt{n}} = 0.865$

CI: 195.82 ± 1.69

Width is estimate of precision of the estimated mean.

Comparing Confidence

Species	\bar{x}	s	95% CI
Adelie	189.95	6.54	(188.91, 190.00)
Chinstrap	195.82	7.13	(194.13, 197.52)
Gentoo	217.19	6.48	(216.04, 218.33)

The confidence intervals **do not overlap**.

This is evidence that the species have different flipper lengths.

We'll see a direct comparison later.

Interpreting Confidence

If we:

- Take a sample of size n
- Compute the statistic
- Infinitely many times

95% of the time, will be in 95% CI.

Can have other CIs (eg 99%)

CI is **not** where we're 95% sure
the parameter is!



Wrapping Up

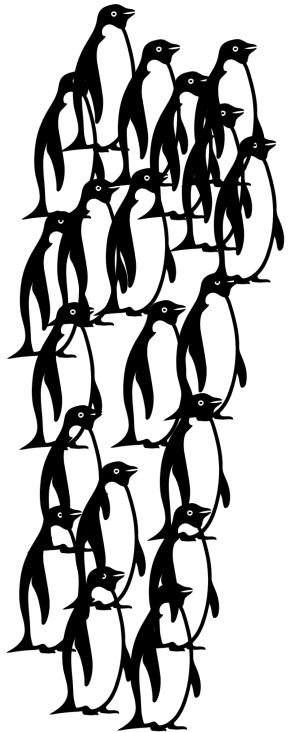
Taking a sample and computing a statistic is a random process that results in a *sampling distribution*.

We can use the sampling distribution to estimate the precision of an estimate (e.g. statistic estimating a parameter).

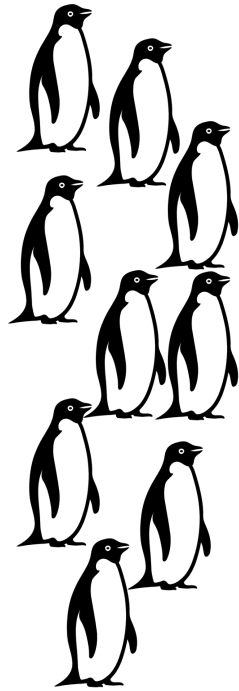
THE BOOTSTRAP

Sampling the Population

Penguins
(Population)
Flipper length μ_{fl}



Sample of
Penguins



Statistic

$$\bar{x} = \frac{1}{n} \sum x_i$$

Repeat

Sampling
Distribution


Going Beyond

Mean is well-understood: $\bar{x} \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- Estimate depends on accuracy of s (usually pretty good)
- This is a *parametric* estimate (distribution w/ params)
- Other statistics have other distributions

So, repeat?

- Take many samples, and compute sampling distribution.
- Expensive

A close-up, high-contrast image of Morpheus from the movie The Matrix. He is bald, wearing dark sunglasses, and has a serious, intense expression. The background is blurred, showing architectural details of a building.

What if I told you

you can resample a sample

The Bootstrap

We have: sample x_1, \dots, x_n

Compute new sample $\hat{x}_1, \dots, \hat{x}_n$ by resampling **with replacement**

Compute statistic from new sample

Do this B (1000-10000) times

Distribution of statistic **approximates sampling distribution**

Bootstrapping a CI

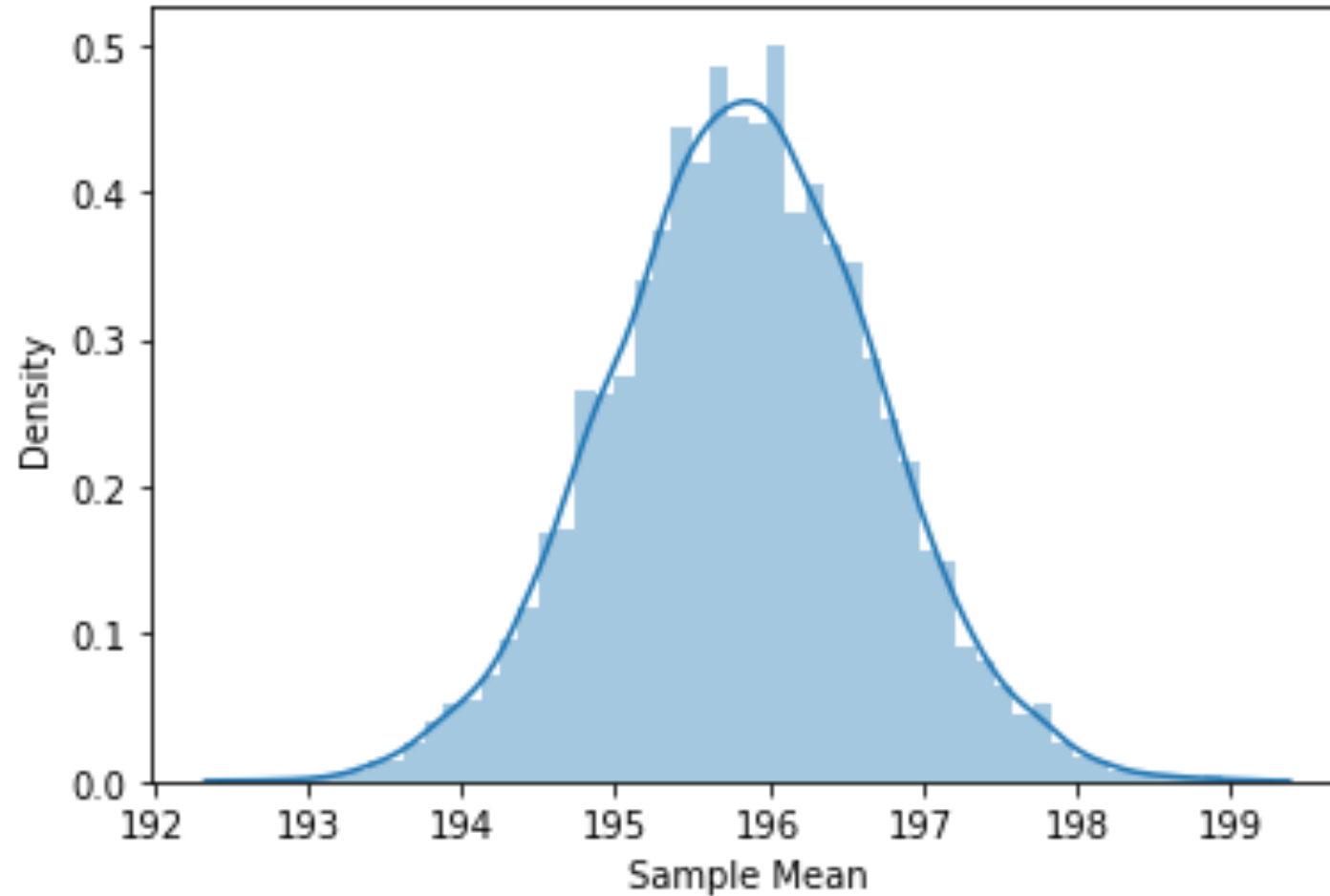
Compute bootstrap means

```
boot_means = [np.mean(rng.choice(xs, n))  
               for i in range(10000)]  
np.quantile(boot_means, [0.025, 0.975])
```

Result: `array([194.10294118, 197.52941176])`

This is what Seaborn catplot does for error bars!

Bootstrap Distribution



Fun and Games with the Bootstrap

- Estimate the sampling distribution for any statistic
- Estimate arbitrary properties of the sampling distribution
 - Mean?
 - Median?
 - Quantiles?
 - Variance?

Wrapping Up

The sampling distribution requires taking multiple samples from the population.

The **bootstrap** allows us to approximate sampling distributions by resampling a sample.



Exercise

If events A and B are independent, which of the following are true?

1. $P[A, B] = P[A]P[B]$

2. $P[B|A] = P[B]$

3. $P[A \cap B] = P[A]P[B]$

4. $P[A|B] = P[B]$

Exercise

Why don't continuous distributions have probability mass functions?

1. Because the probability of a continuous variable being any one precise value is 0.
2. Because writing down probabilities for all the continuous values would take too much space.
3. Because probability mass is only defined for discrete values.

Exercise

What is a **sampling distribution**?

1. The distribution of values resulting from repeatedly taking new samples and computing some statistic, such as a mean.
2. The distribution of values in a sample.
3. The distribution of values in the population.
4. The distribution we use to compute a sample.

Exercise

We measure a variable and compute a sample mean and 95% confidence interval of 7.2 ± 0.5 . Select the correct interpretation:

1. We are 95% sure that the true mean is in the range 6.7 to 7.7.
2. 95% of the time, the sample mean will be in the range 6.7 to 7.7.
3. We have a sampling and computational procedure that, 95% of the time, will produce an interval containing the true mean; with our sample this procedure resulted in the interval (6.7, 7.7).

Exercise

What is the key idea of the bootstrap?

1. To compute the sampling distribution by drawing new samples from the population.
2. To simulate the sampling distribution by resampling an existing sample.
3. To estimate the sampling distribution by reshuffling an existing sample.
4. To estimate the sampling distribution by randomly removing items from an existing sample.

Exercise

What does $P[A|B]$ mean?

1. The probability of either A or B happening.
2. The probability of A and B happening at the same time.
3. The probability that B happens, given that we know A happens.
4. The probability that A happens, given that we know B happens.

Exercise

What does $P[A, B]$ mean?

1. The probability of either A or B happening.
2. The probability of A and B happening at the same time.
3. The probability that B happens, given that we know A happens.
4. The probability that A happens, given that we know B happens.