Statistics Review

EC320, Set 02

Andrew Dickinson

Spring 2024

Admin

Admin

1. R and RStudio Install
2. **Lab**

- Lab module on Canvas homepage
- First recording available now

1. **Koans**

- First Koans due next Friday
- Get started now

Stopping point after lecture 02

1. **PS01**

- First problem set will be assigned next week
- Due next Tuesday (04/16)

1. **Textbook**

Motivation

The focus of our course is **regression analysis**–part of the fundamental toolkit for learning from data.

The **underlying theory** is critical to grasp the mechanics and pitfalls

- Make us better practitioners and savvier consumers of science.

Today: Review the essential concepts from Math 243

Warning.

The following review is a lot packed in very briefly though you *should* have learned much of it before. But that being said, it will be overwhelming for most.

Notation

Notation

Data on a variable X are a sequence of n observations, indexed by i:

$$\{x_i : 1, \ldots, n\}.$$

Ex. $n = 5$

i	x_i
1	8
2	9
3	4
4	7
5	2

- i indicates the row number.
- n is the number of rows.
- x_i is the value of X for row i.

Summation

The **summation operator** adds a sequence of numbers over an index:

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n.$$

> The sum of x_i from 1 to n.

i	x_i
1	7
2	4
3	10
4	3

$$\sum_{i=1}^{4} x_i = 7 + 4 + 10 + 3 = 23$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \frac{1}{4}\sum_{i=1}^{4} x_i = \ 6$$

Summation

The **summation operator** adds a sequence of numbers over an index:

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n.$$

> The sum of x_i from 1 to n.

i	c
1	2
2	2
3	2
4	2

$$\sum_{i=1}^{4} x_i = 7 + 4 + 10 + 3 = 23$$

$$\text{sample average}\left\{ \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \frac{1}{4}\sum_{i=1}^{4} x_i = \ 6 \right.$$

Summation: Rule 01

For any constant c,

$$\sum_{i=1}^{n} c = nc.$$

i	c
1	2
2	2
3	2
4	2

$$\sum_{i=1}^{4} 2 = 4 \times 2$$
$$= 8$$

Summation: Rule 02

For any constant c,

$$\sum_{i=1}^{n} c x_i = c \sum_{i=1}^{n} x_i .$$

i	x_i
1	8
2	9
3	4
4	7
5	2

$$\sum_{i=1}^{3} 2x_i = 2 \times 7 + 2 \times 4 + 2 \times 10$$

$$= 14 + 8 + 20 = 42$$

$$2 \sum_{i=1}^{3} x_i = 2(7 + 4 + 10) = 42$$

Summation: Rule 03

If $\{(x_i, y_i) : 1, \ldots, n\}$ is a set of n pairs, and a and b are constants, then

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

i	a	x_i	b	y_i
1	2	7	1	4
2	2	4	1	2

$$\sum_{i=1}^{2}(2x_i + y_i) = 18 + 10 = 28 \quad (1)$$

$$2\sum_{i=1}^{2} x_i + \sum_{i=1}^{2} y_i = 2 \times 11 + 6 = 28 \quad (2)$$

Summation: Caution 01

The **sum of the ratios** is not the **ratio of the sums:**

$$\sum_{i=1}^{n} x_i / y_i \neq \left(\sum_{i=1}^{n} x_i \right) \Bigg/ \left(\sum_{i=1}^{n} y_i \right)$$

Ex.

If $n = 2$, then $\frac{x_1}{y_1} + \frac{x_2}{y_2} \neq \frac{x_1 + x_2}{y_1 + y_2}$

Summation: Caution 02

The **sum of squares** is not the **square of the sums:**

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2$$

Ex.

If $n = 2$, then $x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

Cartesian coordinate system

Cartesian plane: 2-D plane defined by two perpendicular number lines:

- x-axis (*horizontal*)
- y-axis (*vertical*)

Using these axes, any point in the plane is described using an ordered pair of numbers (x, y)

Cartesian coordinate system

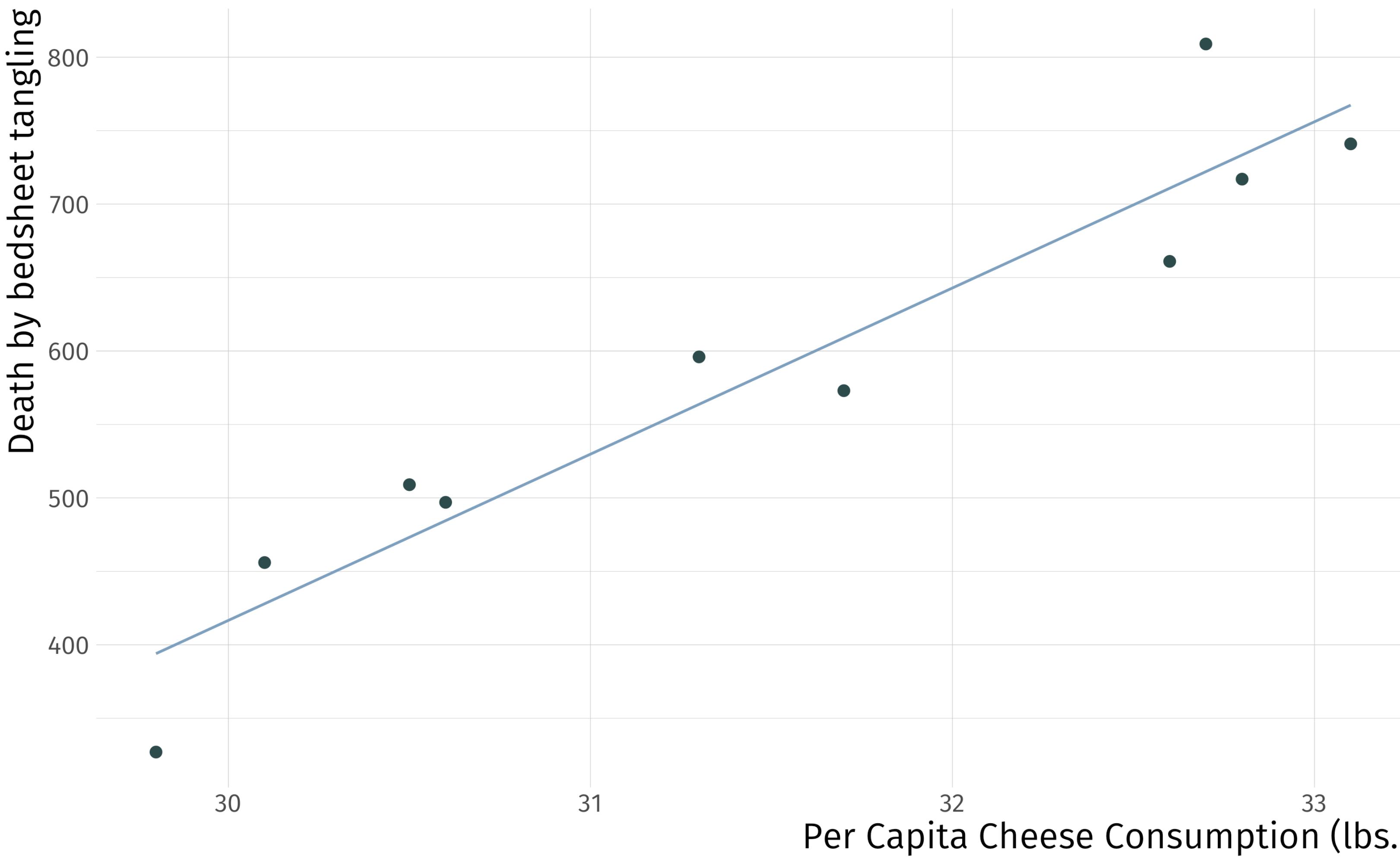
A particular line on this plane takes the form

$$y = a + bx$$

where a is known as the intercept and b is the slope.

Any incremental unit increase in x results in y increasing by b.

Ex.



Death by bedsheet tangling
800
700
600
500
400
30
31
32
33
Per Capita Cheese Consumption (lbs.)

Basic probability

Essential definitions

Experiment:

> Any procedure that is *infinitely repeatable* and has a *well-defined set of outcomes*.

Ex. Flip a coin 10 times and record the number of heads.

Random Variable:

> A variable with *numerical values determined by an experiment or a random phenomenon*.

- Describes the sample space of an experiment.

Essential definitions

Sample Space:

> The set of potential outcomes an experiment could generate

Ex. The sum of two dice is an integer from 2 to 12.

Event:

> A subset of the sample space or a combination of outcomes.

Ex. Rolling a two or a four.

Random variables

Notation: Capital letters for random variables (*e.g.*, X, Y, or Z) and lowercase letters for particular outcomes (*e.g.*, x, y, or z).

Experiment

Flipping a coin.

Events:

Heads or tails.

Random Variable: (X)

Receive \$1 if heads, $x_i = 1$, pay \$1 if tails, $x_i = -1$

Sample Space:

$\{-1, 1\}$

Discrete random variables

> A random variable that takes a countable set of values.

Bernoulli (*binary*) random variable

> Random variable that takes values of either 1 or 0.

- Characterized by $P(X = 1)$, "the probability of success."
- Probabilities sum to 1: $P(X = 1) + P(X = 0) = 1$

More generally, if

$$P(X = 1) = \theta$$

then

$$P(X = 0) = 1 - \theta$$

for some $\theta \in [0, 1]$

Discrete Random Variables: Probabilities

We describe a discrete random variable by listing its possible values with associated probabilities.

If X takes on k possible values $\{x_1, \ldots, x_k\}$, then the probabilities $p_1, p_2, \ldots, p_k$ are defined by

$$p_j = P(X = x_j), \quad j = 1, 2, \ldots, k,$$

where

$$p_j \in [0, 1]$$

and

$$p_1 + p_2 + \cdots + p_k = 1.$$

Discrete Random Variables

Probability density function (pdf)

> The *pdf* of X summarizes possible outcomes and associated probabilities:

$$f(x_j) = p_j, \quad j = 1, 2, \ldots, k.$$

Ex. 2020 Presidential election: 538 electoral votes at stake.

- $\{X : 0, 1, \ldots, 538\}$ is the number of votes won.
- Unlikely that one will win 0 or 538 votes: $f(0) \approx 0$ and $f(538) \approx 0$.
- Nonzero probability of winning an exact majority: $f(270) > 0$.

Discrete random variables *Ex.*

Basketball player goes to the foul line to shoot two free throws.

- X is the number of shots made (either 0, 1, or 2).
- The pdf of X is $f(0) = 0.3$, $f(1) = 0.4$, $f(2) = 0.3$.[1]

Use the pdf to calculate the probability of the **event** that the player makes *at least one shot*, *i.e.*, $P(X \geq 1)$.

$$P(X \geq 1) = P(X = 1) + P(X = 2) = 0.4 + 0.3 = 0.7$$

1. Note: the probabilities sum to 1

Continuous random variables

> A random variable that takes any real value with *zero* probability.

Wait, what?! The variable takes so many values that we can't count all possibilities, so the probability of any one particular value is zero.

Measurement is discrete (*e.g.*, dollars and cents), but variables with many possible values are best treated as continuous.

- *e.g.*, electoral votes, height, wages, temperature, *etc.*

Continuous random variables

Probability density functions also describe continuous random variables.

Difference between continuous and discrete **PDFs**

- Interested in the probability of events within a *range* of values.
- *e.g.* What is the probability of more than 1 inch of rain tomorrow?

Distributions

Distributions

> Function that represents all outcomes of a random variable and the corresponding probabilities.

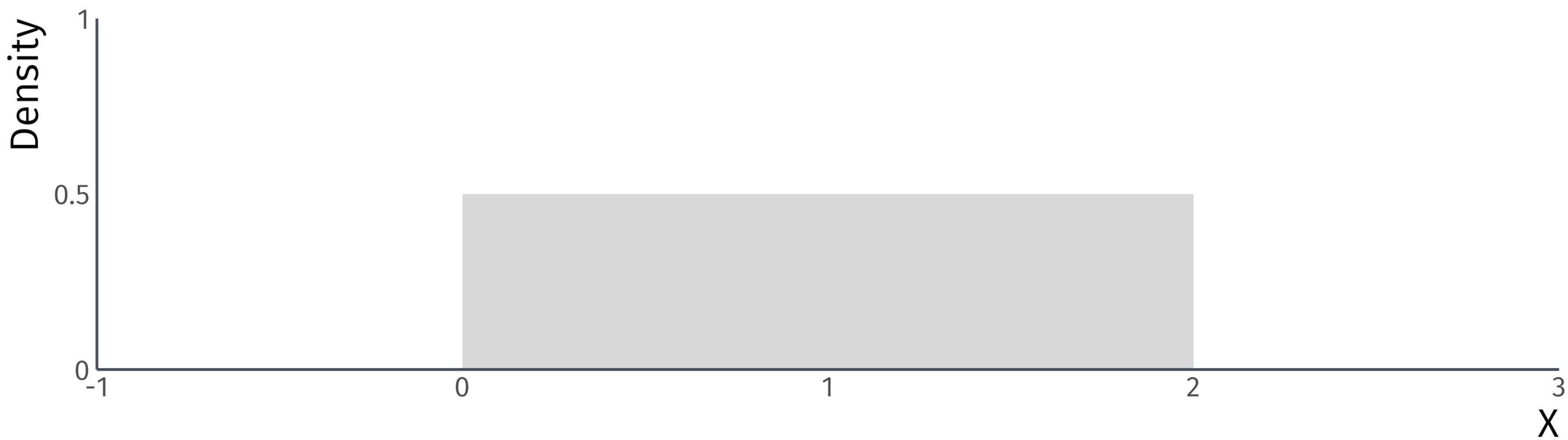
- Summary that describes the spread of data points in a set
- Essential for making inferences and assumptions from data

Key Takeaway: The shape of a distribution provides valuable information

Uniform distribution

The probability density function of a variable uniformly distributed between 0 and 2 is

$$f(x) = \begin{cases} \frac{1}{2} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

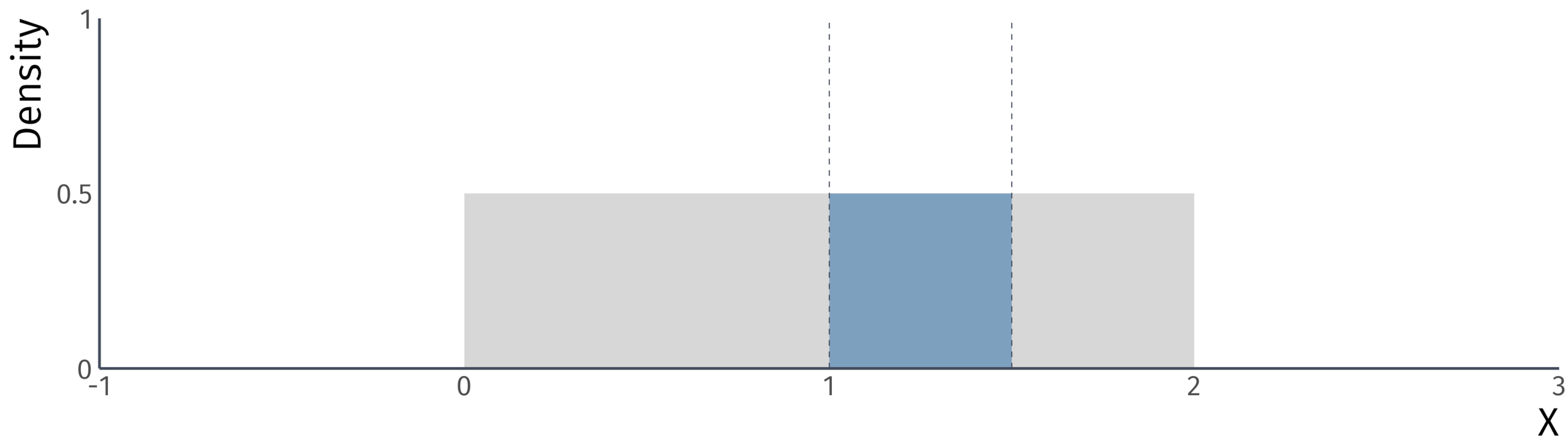


Uniform distribution

By definition, the area under $f(x)$ is equal to 1.

The **shaded area** illustrates the probability of the event $1 \leq X \leq 1.5$.

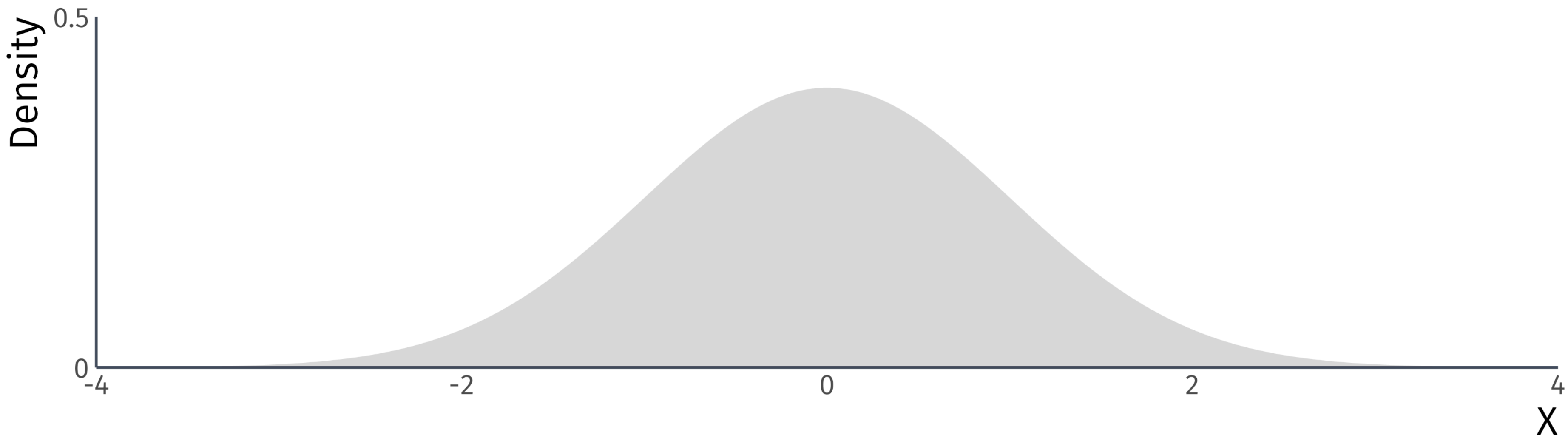
$$P(1 \leq X \leq 1.5) = (1.5 - 1) \times 0.5 = 0.25$$



Normal Distribution

The **"bell curve"**

- Symmetric: mean and median occur at the same point (*i.e.*, no skew).
- Low-probability events in tails; high-probability events near center.

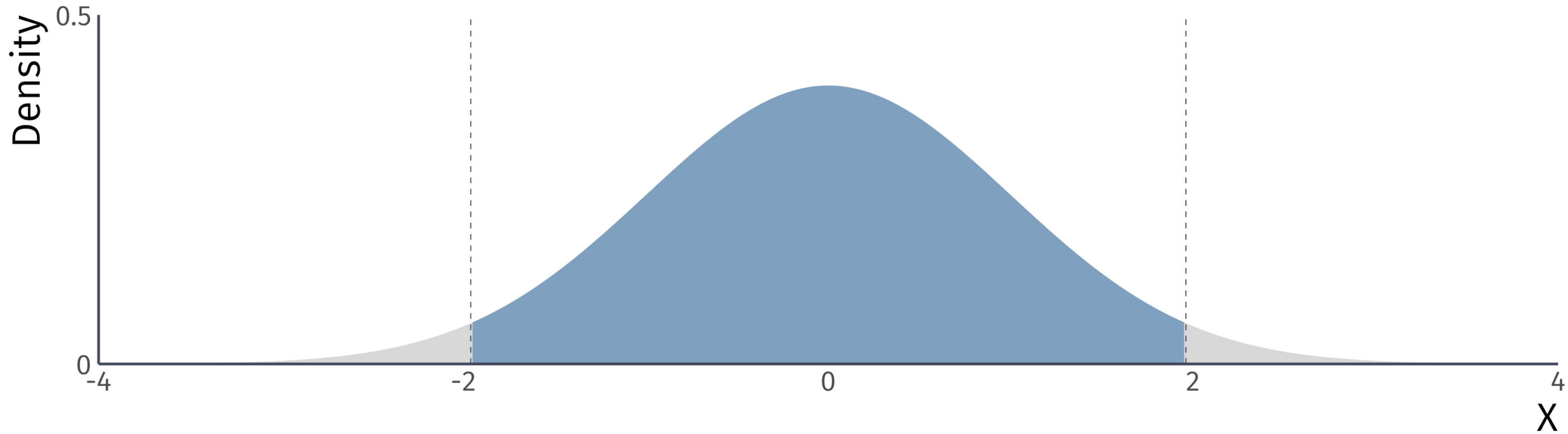


Normal Distribution

The **shaded area** illustrates the probability of the event $-2 \leq X \leq 2$.

- "Find area under curve" = use integral calculus (or, in practice, R).

$$P(-2 \leq X \leq 2) \approx 0.95$$



Normal Distribution

Continuous distribution where x_i takes the value of any real number ($\mathbb{R}$)

- Domain spans the entire real line
- Centered on the distribution mean μ

Rule 1: The probability that the random variable takes a value x_i is 0 for any $x_i \in \mathbb{R}$

Rule 2: The probability that the random variable falls between $[x_i, x_j]$ range, where $x_i \neq x_j$, is the area under $p(x)$ between those two values

The area above represents $p(x) = 0.95$. The values $\{-1.96, 1.96\}$ represent the 95% confidence interval for μ.

Moments

Moments

> Quantitative measures used to describe the shape and characteristics of a probability distribution[1]

Summarize and understand the important features of a distribution

First moment: **Mean**

Second moment: **Variance**

Third moment: Skewness

Fourth moment: Kurtosis

$\vdots$

1. See this video for a more in depth description of moments (click here)

Expected Value

> Describes the *central tendency* of distribution in a single number.[1]

Density functions describe the entire distribution, but sometimes we just want a summary.

Other summary statistics we may be interested in include

- Median
- Standard deviation
- 25th percentile
- 75th percentile

1. *Central tendency* = typical value to expect upon drawing from the distribution

Expected Value (discrete)

> The expected value of a discrete random variable X is the weighted average of its k values $\{x_1, \dots, x_k\}$ and their associated probabilities:

$$
\begin{aligned}
E(X) &= x_1 P(x_1) + x_2 P(x_2) + \cdots + x_k P(x_k) \\
&= \sum_{j=1}^{k} x_j P(x_j).
\end{aligned}
$$

AKA: **Population mean**

Expected Value ***Ex.***

Rolling a six-sided die once can take values $\{1, 2, 3, 4, 5, 6\}$, each with equal probability. *What is the expected value of a roll?*

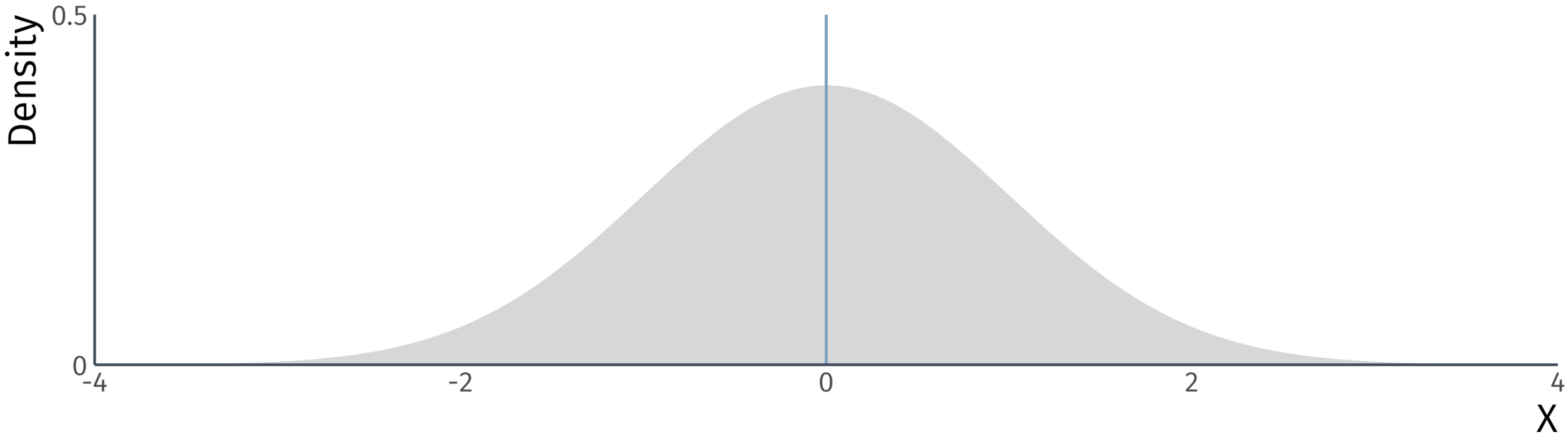
$$
\begin{aligned}
E(\text{Roll}) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} \\
+ 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5
\end{aligned}
$$

Note: The **EV** can be a number that isn't a possible outcome of X.

Expected value (continuous)

If X is a continuous random variable and $f(x)$ is its probability density function, then the expected value of X[1] is

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$



1. Note: x represents the particular values of X.

Expected value: Rule 01

For any constant c, $E(c) = c$. ***Ex.***

- $E(5) = 5$.
- $E(1) = 1$.
- $E(4700) = 4700$.

Expected value: Rule 02

For any constants a and b, $E(aX + b) = aE(X) + b$.

Ex. Suppose X is the high temperature in degrees Celsius in Eugene during August. The long-run average is $E(X) = 28$. If Y is the temperature in degrees Fahrenheit, then $Y = 32 + \frac{9}{5}X$. What is $E(Y)$?

$$E(Y) = 32 + \frac{9}{5}E(X) = 32 + \frac{9}{5} \times 28 = 82.4$$

Expected value: Rule 03

If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then

$$E(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) \;=\; a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n)$$

In English, **the expected value of the sum = the sum of expected values**.

Expected value: Rule 03

The expected value of the sum = the sum of expected values.

Ex. Suppose that a coffee shop sells X_1 small, X_2 medium, and X_3 large caffeinated beverages in a day. The quantities sold are random with expected values $E(X_1) = 43$, $E(X_2) = 56$, and $E(X_3) = 21$. The prices of small, medium, and large beverages are 1.75, 2.50, and 3.25 dollars. What is expected revenue?

$$
\begin{aligned}
E(1.75X_1 + 2.50X_2 + 3.35X_3) &= 1.75E(X_1) + 2.50E(X_2) + 3.25E(X_3) \\
&= 1.75(43) + 2.50(56) + 3.25(21) \\
&= 283.5
\end{aligned}
$$

Expected value: Caution

Previously, we found that the expected value of rolling a six-sided die is $E\left(\text{Roll}\right) = 3.5$.

- If we square this number, we get $\left[E(\text{Roll})\right]^2 = 12.25$.

Is $\left[E\left(\text{Roll}\right)\right]^2$ the same as $E\left(\text{Roll}^2\right)$?

$$
\begin{aligned}
E\left(\text{Roll}^2\right) &= 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} \\
&\quad + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} \\
&\approx 15.167 \neq 12.25.
\end{aligned}
$$

No!

Expected value: Caution

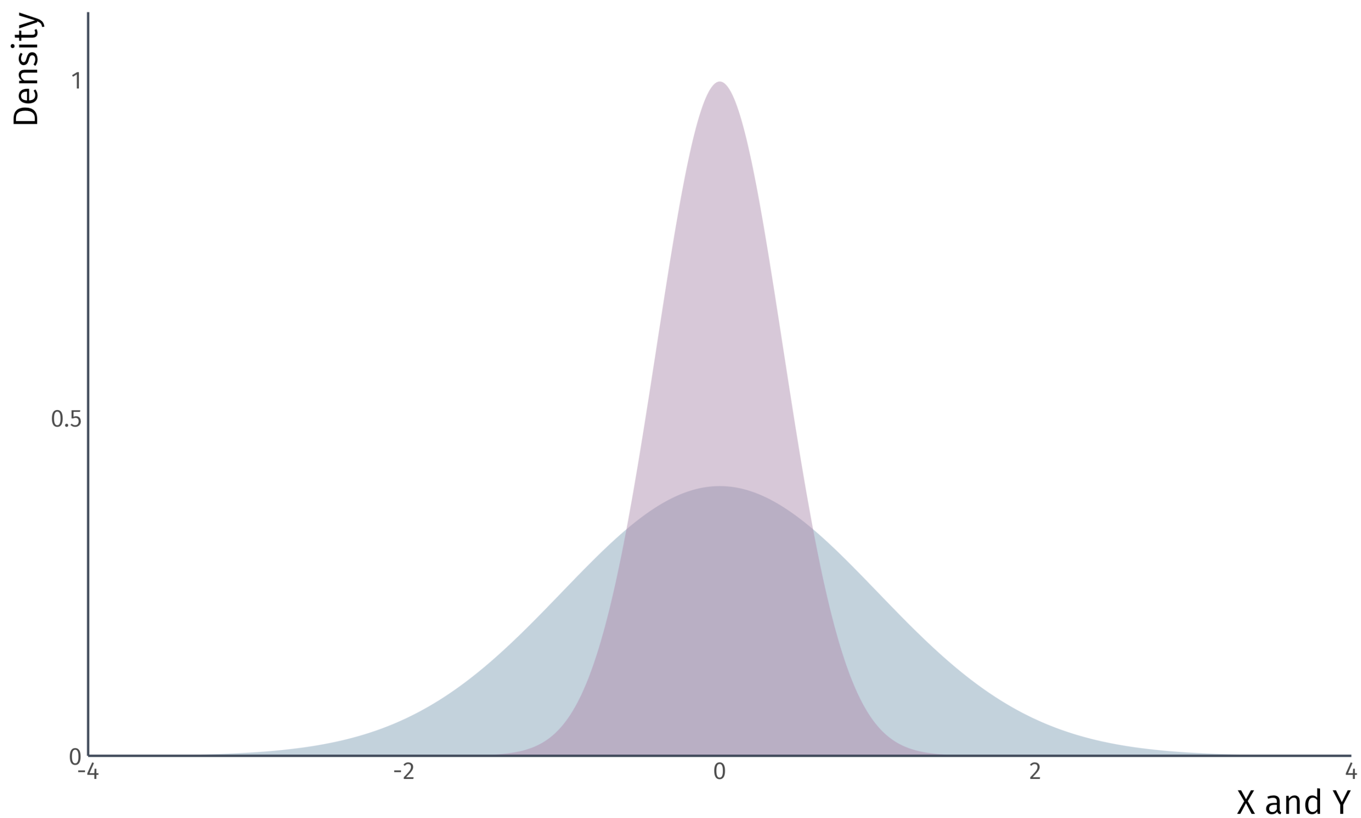
Except in special cases, **the transformation of an expected value** is note **the expected value of a transformed random variable**.

For some function $g(\cdot)$, it is typically the case that

$$g\left(E(X)\right) \neq E\left(g(X)\right).$$

Variance

Random variables X and Y share the same population mean, but are distributed differently.



Variance (σ^2)

> Tells us how far X deviates from μ, *on average*:

$$\text{Var}(X) \equiv E\left[(X-\mu)^2\right] = \sigma_X^2$$

Where: $\mu = E(X)$.

How tightly is a random variable distributed about its mean?

Describe the distance of X from its population mean μ as the squared difference: $(X-\mu)^2$.

- Distributing the terms above yields $\sigma^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$.

Variance: Rule 01

$\mathbf{Var}(X) = \mathbf{0} \iff X$ is a constant.

- A random variable that never deviates from its mean has zero variance.

Wait what? How can a random variable be a constant?? Because a constant fits the technical definition of a random variable[1]. It's just not-so-random

1. See this link for a more technical explanation

Variance: Rule 02

For any constants a and b, $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

Ex. Suppose X is the high temperature in degrees Celsius in Eugene during August. If Y is the temperature in degrees Fahrenheit, then $Y = 32 + \frac{9}{5}X$. What is $\text{Var}(Y)$?

$$\text{Var}(Y) = (\frac{9}{5})^2 \text{Var}(X) = \frac{81}{25} \text{Var}(X)$$

Standard Deviation (σ)

> The positive square root of the variance:

$$\text{sd}(X) = +\sqrt{\text{Var}(X)} = \sigma$$

Rule 01: For any constant c, $\text{sd}(c) = 0$.

Rule 02: For any constants a and b, $\text{sd}(aX + b) = |a| \, \text{sd}(X)$.

Note: The same as variance, almost

Standardizing a random variable

When we're working with a random variable X with an unfamiliar scale, it is useful to **standardize** it by defining a new variable Z:

$$Z \equiv \frac{X - \mu}{\sigma}$$

Z has mean 0 and standard deviation 1. How?

- First, some simple trickery: $Z = aX + b$, where $a \equiv \frac{1}{\sigma}$ and $b \equiv -\frac{\mu}{\sigma}$.
- $E(Z) = aE(X) + b = \mu\frac{1}{\sigma} - \frac{\mu}{\sigma} = 0$.
- $\mathrm{Var}(Z) = a^2\mathrm{Var}(X) = \frac{1}{\sigma^2}\sigma^2 = 1$.

Covariance

> For two random variables X and Y, the covariance is defined as the expected value (or mean) of the product of their deviations from their individual expected values:

$$\text{Cov}(X, Y) \equiv E\left[(X - \mu_X)(Y - \mu_Y)\right] = \sigma_{XY}$$

Idea: Characterize the relationship between random variables X and Y.

- **Positive correlation:** When $\sigma_{XY} > 0$, then X is above its mean when Y is above its mean, *on average*.
- **Negative correlation:** When $\sigma_{XY} < 0$, then X is below its mean when Y is above its mean, *on average*.

Covariance: Rule 01

Statistical independence:

> If X and Y are independent, then $E(XY) = E(X)E(Y)$.

- If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Caution:

- $\text{Cov}(X, Y) = 0$ **does not imply** that X and Y are independent.
- $\text{Cov}(X, Y) = 0$ means that X and Y are *uncorrelated*.

Covariance: Rule 02

For any constants a, b, c, and d,

$$\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$$

Correlation Coefficient

A problem with covariance is that it is sensitive to units of measurement.

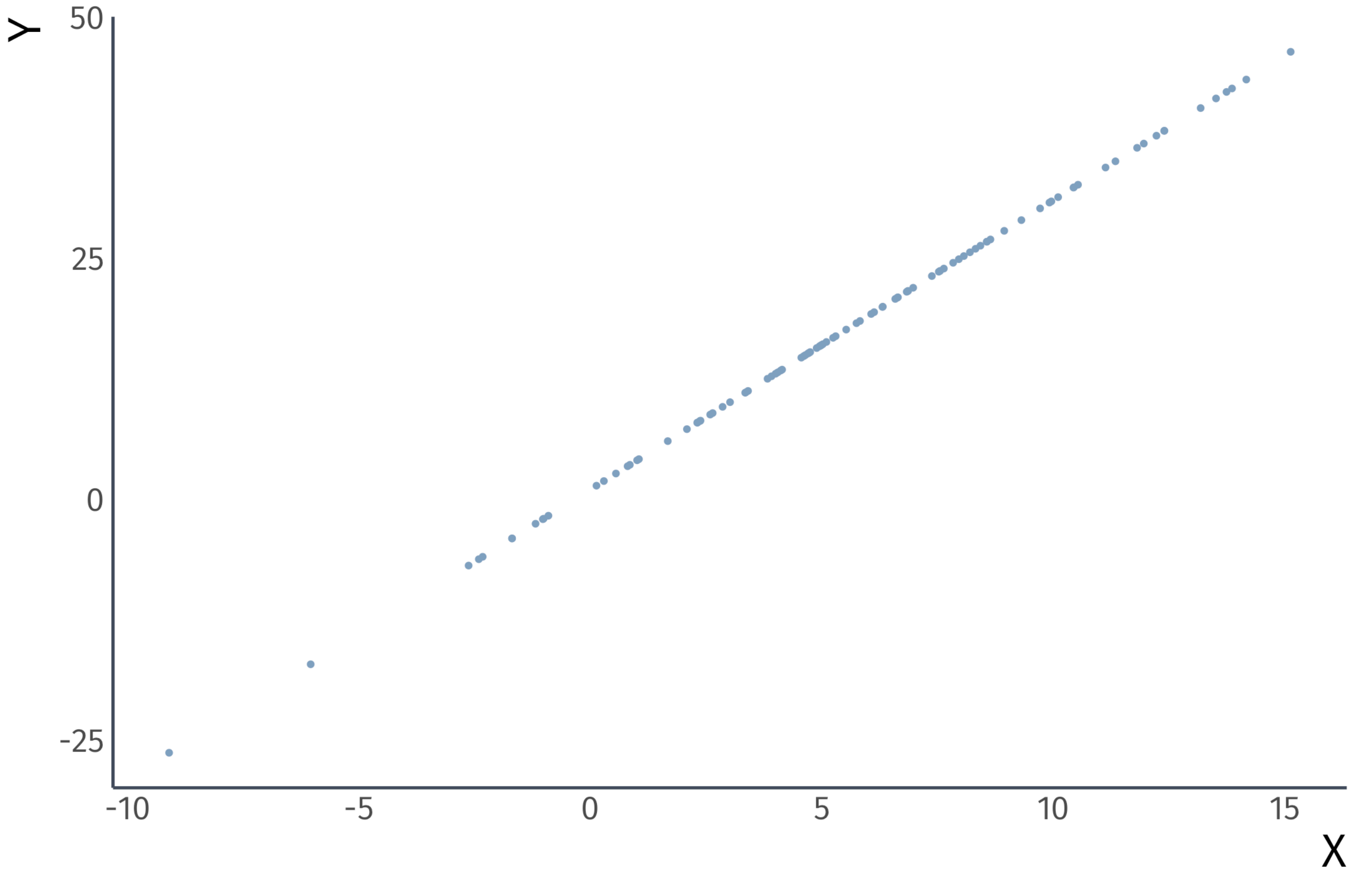
The **correlation coefficient** solves this problem by rescaling the covariance:

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{sd}(X) \times \text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

- Also denoted as ρ_{XY}.
- $-1 \leq \text{Corr}(X, Y) \leq 1$
- Invariant to scale: if I double Y, $\text{Corr}(X, Y)$ will not change.

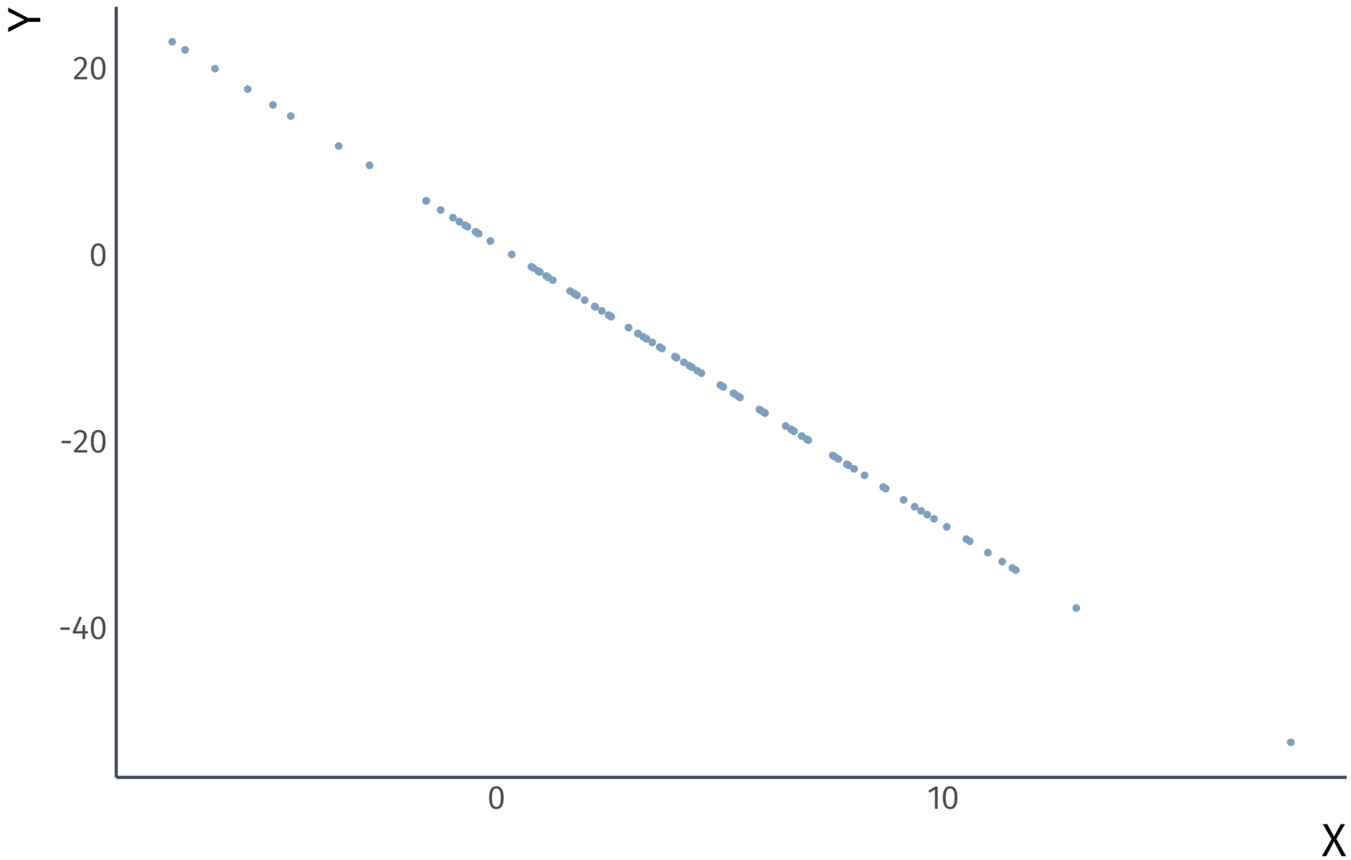
Correlation Coefficient

Perfect positive correlation: $\text{Corr}(X, Y) = 1$.



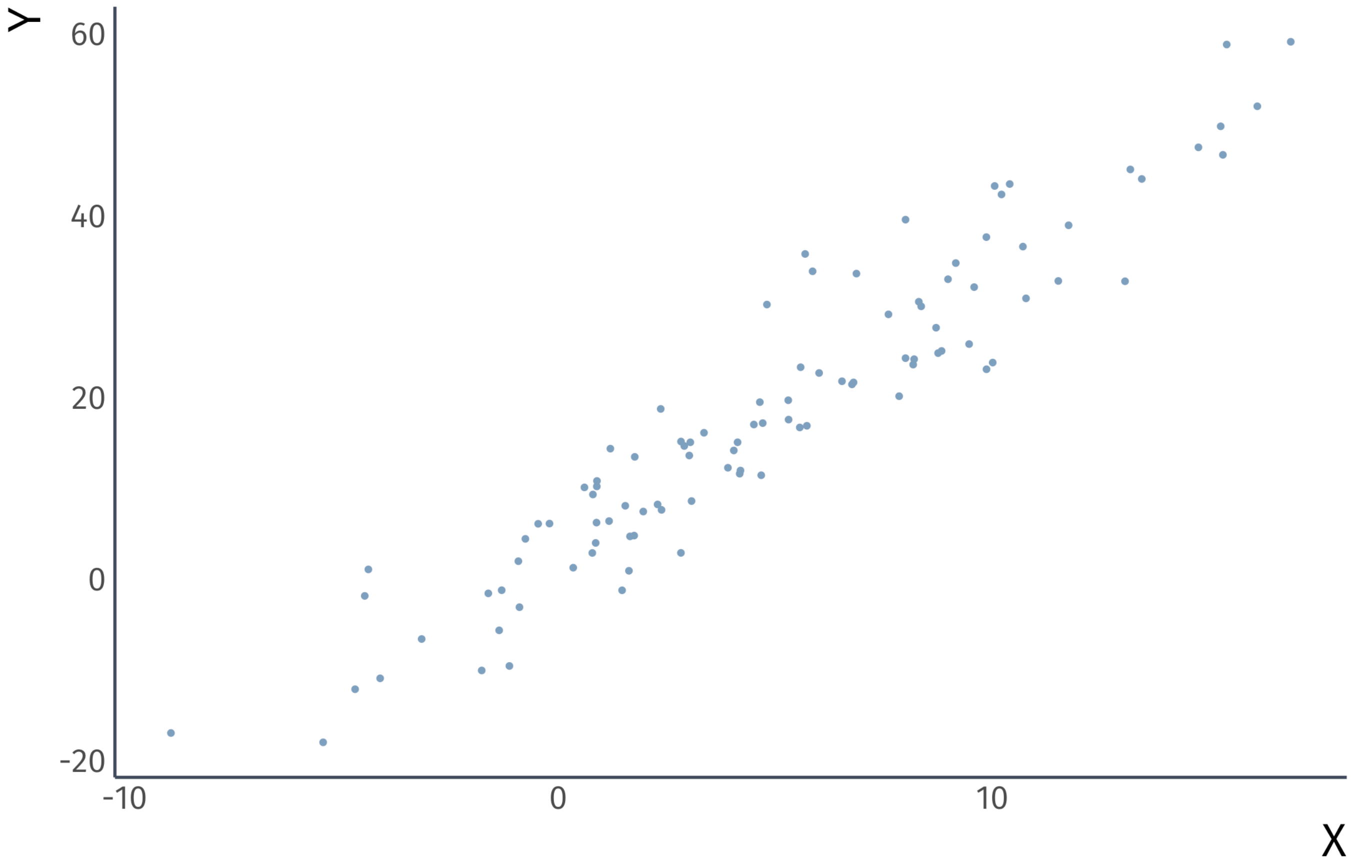
Correlation Coefficient

Perfect negative correlation: $\mathbf{Corr}(X, Y) = -1$.



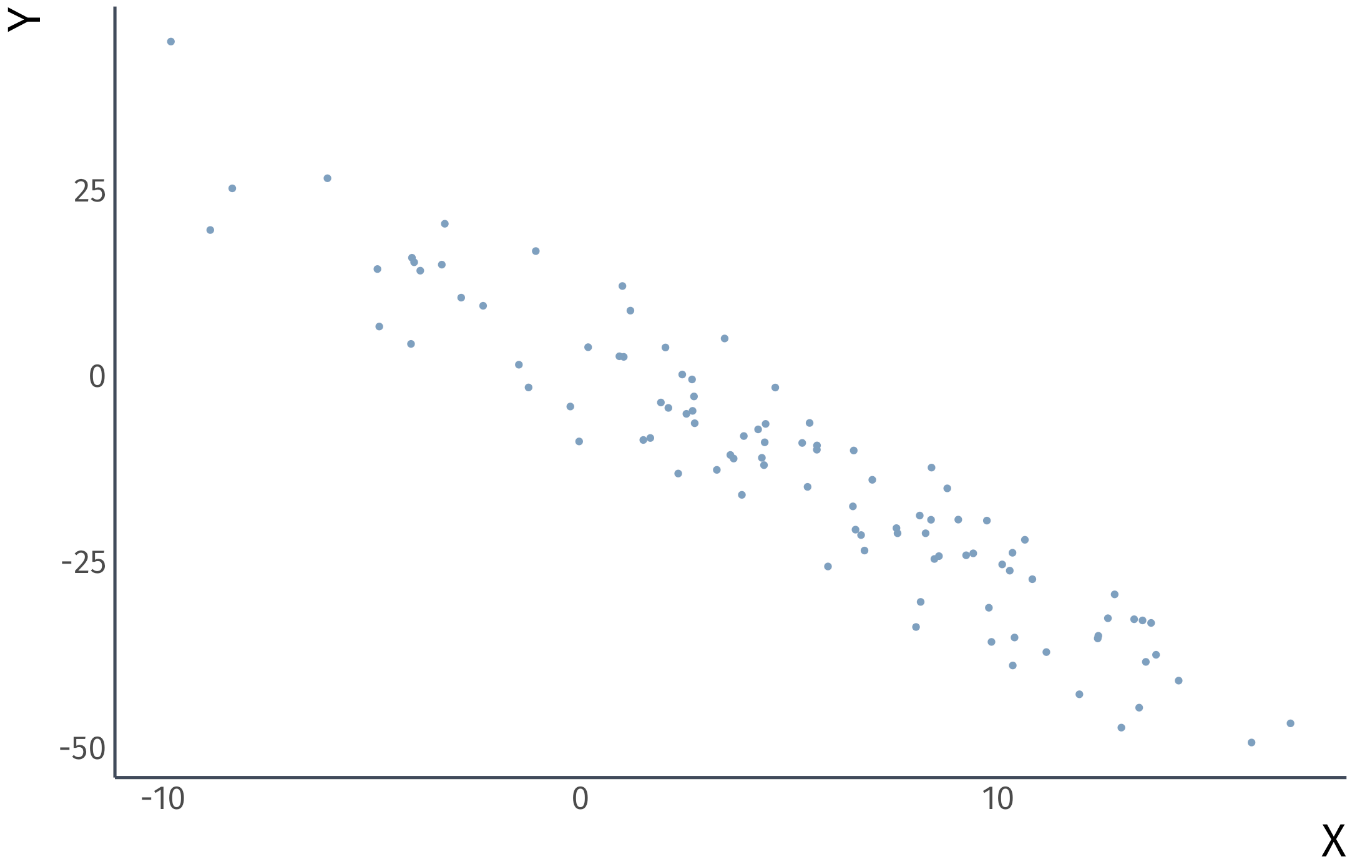
Correlation Coefficient

Positive correlation: $\mathbf{Corr}(X, Y) > \mathbf{0}$.



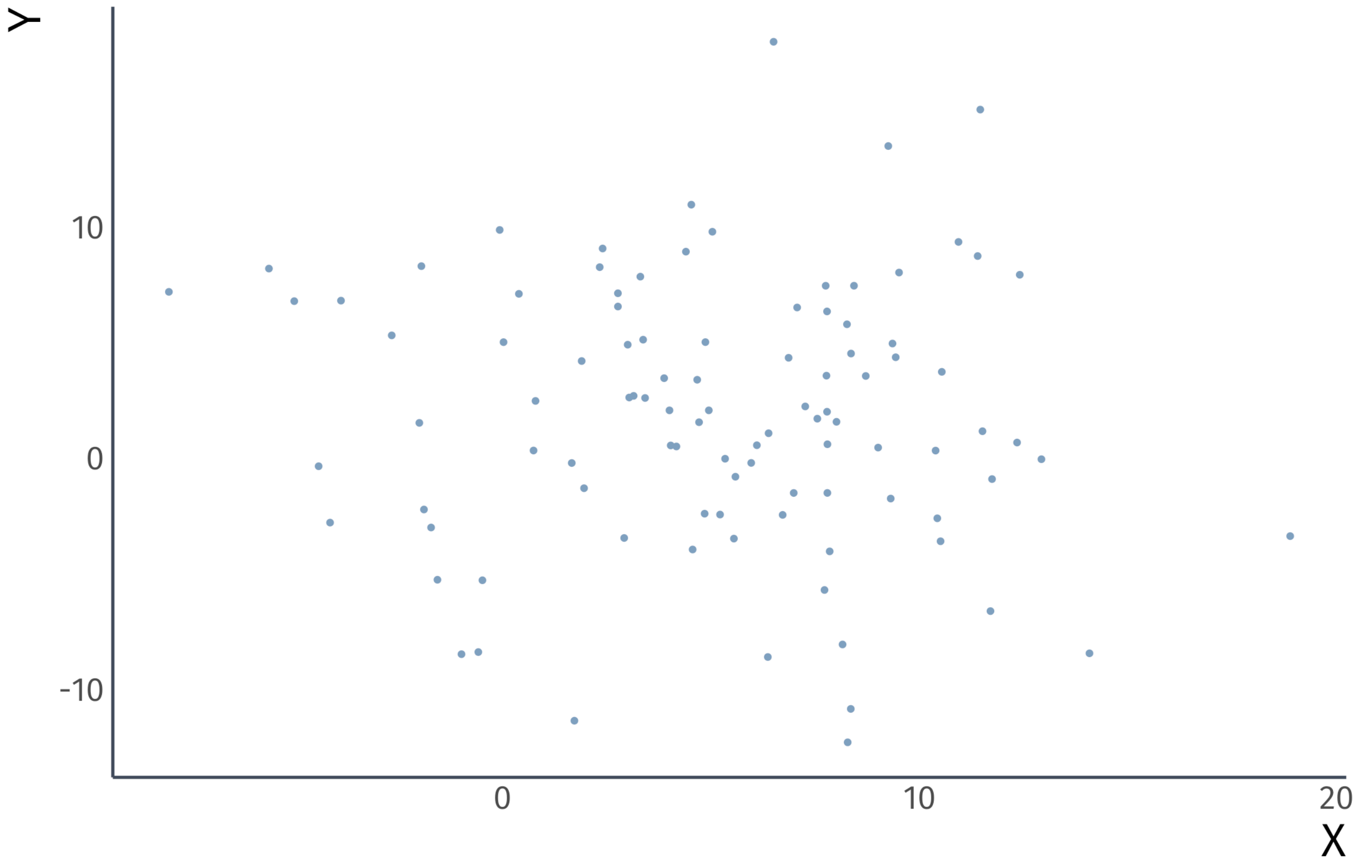
Correlation Coefficient

Negative correlation: $\text{Corr}(X, Y) < 0$.



Correlation Coefficient

No correlation: $\text{Corr}(X, Y) = 0$.



Variance: Rule 03

For constants a and b,

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \, \text{Cov}(X, Y).$$

- If X and Y are uncorrelated, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

- If X and Y are uncorrelated, then

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

Expanded proof

Estimators

Estimators

Why do we estimate things? *Because we can't measure everything*

Suppose we want to know the average height of the population in the US

- We have a sample 1 million Americans

How can we use these data to estimate the height of the population?

Estimators

Estimand:

> Quantity that is to be estimated in a statistical analysis

Estimator:

> A rule (or formula) for estimating an unknown population parameter given a sample of data.

Estimate:

> A specific numerical value that we obtain from the sample data by applying the estimator.

Estimators ***Ex.***

Suppose we want to know the average height of the population in the US

- We have a sample 1 million Americans

Estimand: The population mean (μ)

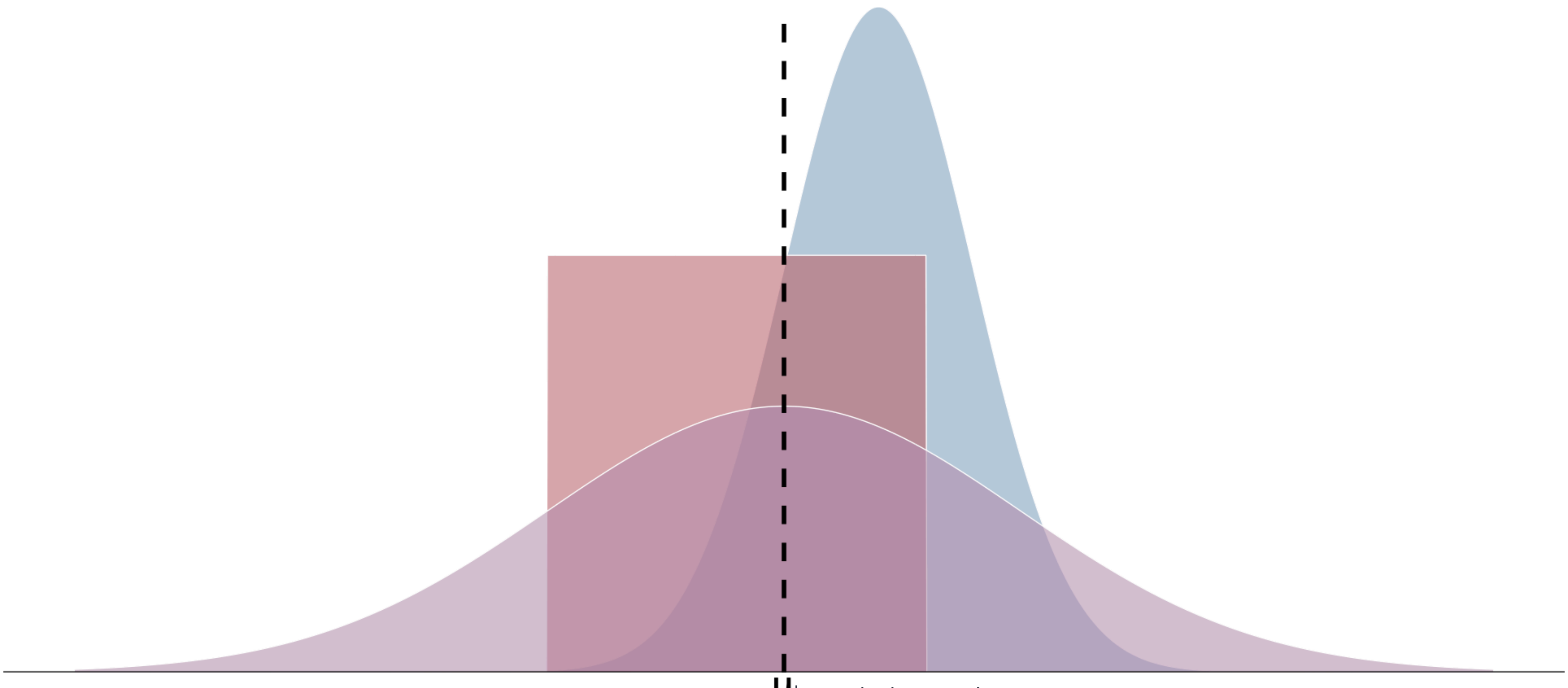
Estimator: The sample mean ($\bar{X}$)

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

Estimate: The sample mean ($\hat{\mu} = 5'6''$)

Properties of estimators

Imagine that we want to estimate an unknown parameter μ, and we know the distributions of three competing estimators. *Which one should we use?*



Properties of estimators

Question: *What properties make an estimator reliable?*

Answer (1): **Unbiasedness**

On average, does the estimator tend toward the correct value?

More formally: Does the mean of estimator's distribution equal the parameter it estimates?

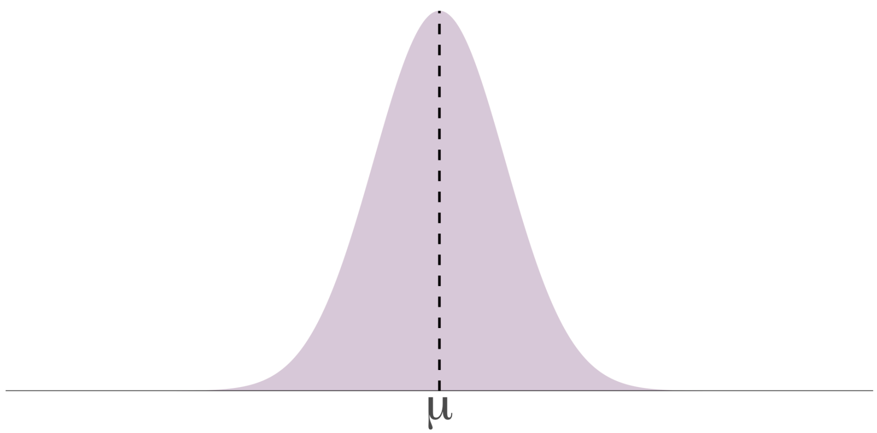
$$\text{Bias}_\mu\left(\hat{\mu}\right) = E\left[\hat{\mu}\right] - \mu$$

Properties of estimators

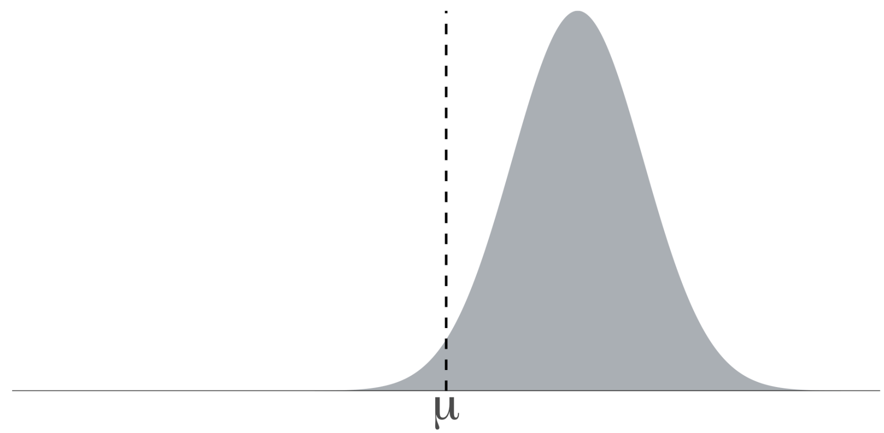
Question What properties make an estimator reliable?

AO1: **Unbiasedness**

Unbiased estimator: $E\left[\hat{\mu}\right] = \mu$



Biased estimator $E\left[\hat{\mu}\right] \neq \mu$



Unbiasedness example

Is the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$ an unbiased estimator of the population mean $E(x_i) = \mu$?

$$\begin{aligned} E\left[\hat{\mu}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[x_i\right] \qquad \} \quad \textbf{rule 3} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \qquad \} \quad \textbf{by definition} \\ &= \mu \end{aligned}$$

Properties of estimators

Question What properties make an estimator reliable?

A02: **Efficiency** *(low variance)*

The central tendencies (means) of competing distributions are not the only things that matter. We also care about the variance of an estimator.

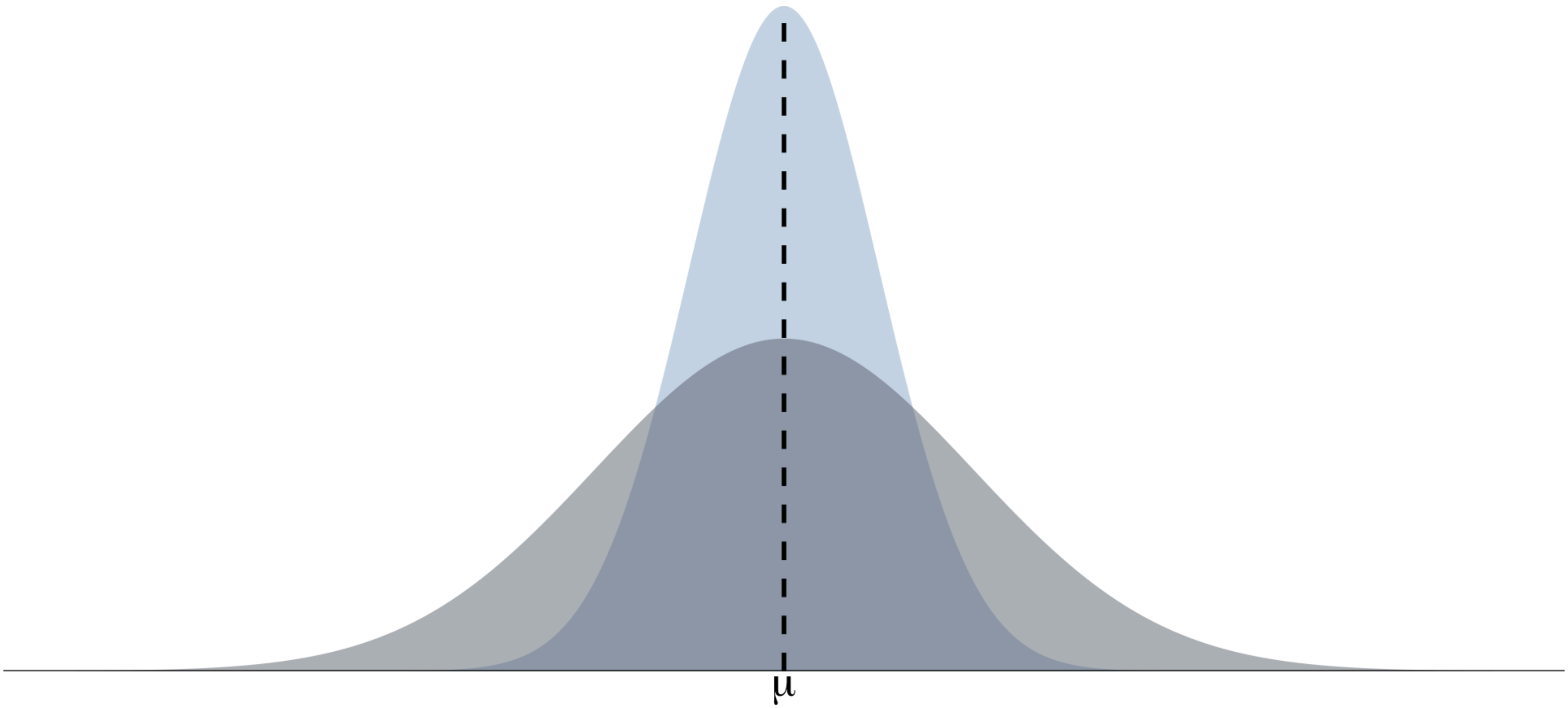
$$\text{Var}\left(\hat{\mu}\right) = E\left[\left(\hat{\mu} - E\left[\hat{\mu}\right]\right)^2\right]$$

Lower variance estimators estimate closer to the mean in each sample

Properties of estimators

Question: *What properties make an estimator reliable?*

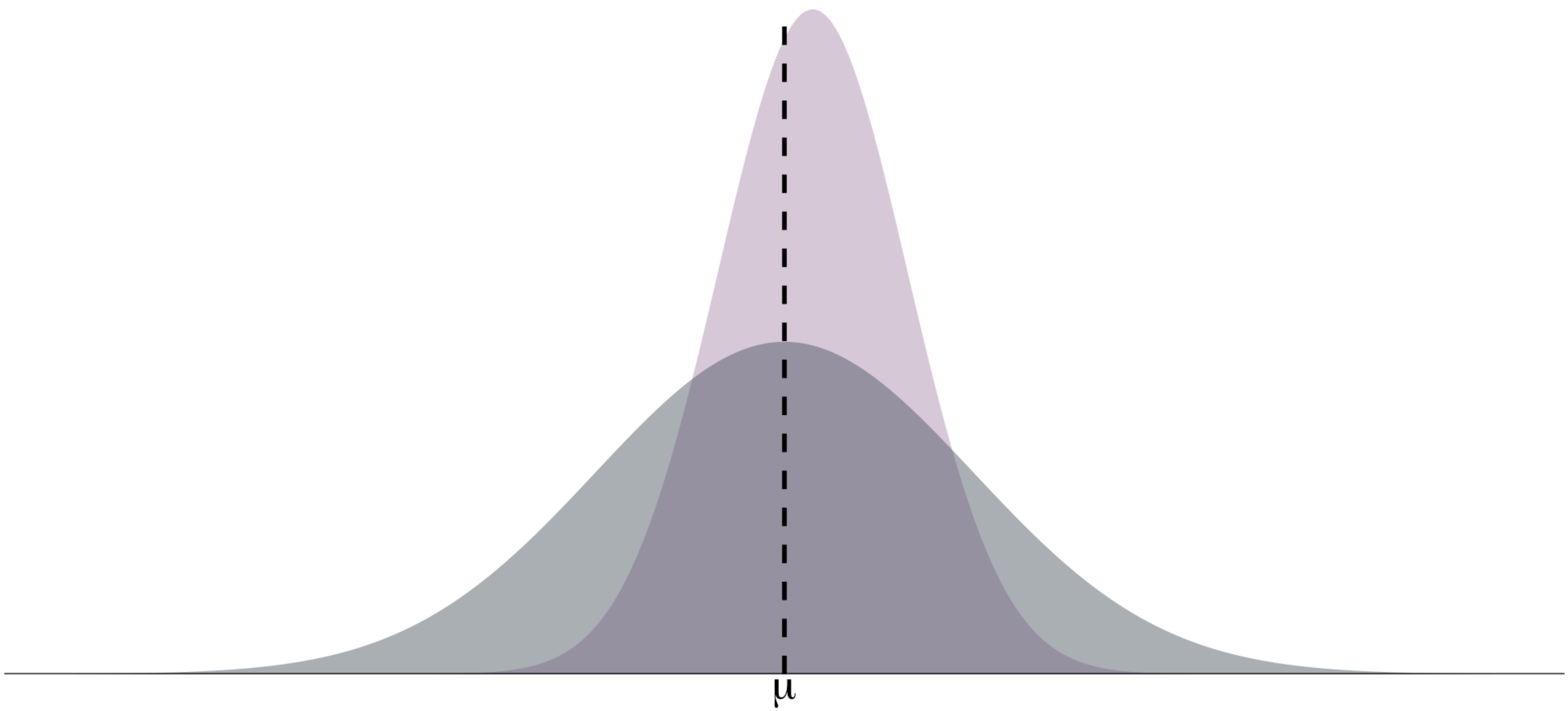
A02: **Efficiency** *(low variance)*



The bias-variance tradeoff

Should we be willing to take a bit of bias to reduce the variance

In economics/causal inference we emphasize unbiasedness



Unbiased estimators

In addition to the sample mean, there are several other unbiased estimators we will use often.

- **Sample variance** estimates variance σ^2.
- **Sample covariance** estimates covariance σ_{XY}.
- **Sample correlation** estimates the pop. correlation coefficient ρ_{XY}.

Unbiased estimators

Sample variance, S_X^2, is an unbiased estimator of the pop. variance σ^2

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Sample covariance, S_{XY}, is an unbiased estimator of the pop. covariance, σ_{XY}

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Unbiased estimators

Sample correlation r_{XY} is an unbiased estimator of the pop. correlation coefficient ρ_{XY}

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2}\sqrt{S_Y^2}}.$$

Sampling

Sampling

Population:

> A group of items or events we would like to know about.

Ex. Americans, games of chess, cats in Eugene, etc.

Parameter[1]

> a value that describes that population

Ex. Mean height of American, average length of a chess game, median weight of the kitties

1. :Parameter of interest is the parameter that the researcher seeks to learn about

Sampling

Sample:

> A survey of a subset of the population.

Ex. Respondents to a survey, random sample of econ students at the UO

Often we aim to draw observations randomly from the population

- Advantageous as it becomes a **representative sample** of the population...

Sampling distributions

Focus: Populations vs Samples

- How can we make inferences about a **population** based on a small **sample** of the population?
- How do we learn about an unknown population parameter of interest?

Challenge: Usually missing data of the entire population.

Solution: Sample from the population and estimate the parameter.

- Draw n observations from the population, then use an estimator.

Sampling distributions

There are myriad ways to produce a sample,[1] but we will restrict our attention **to simple random sampling**, where

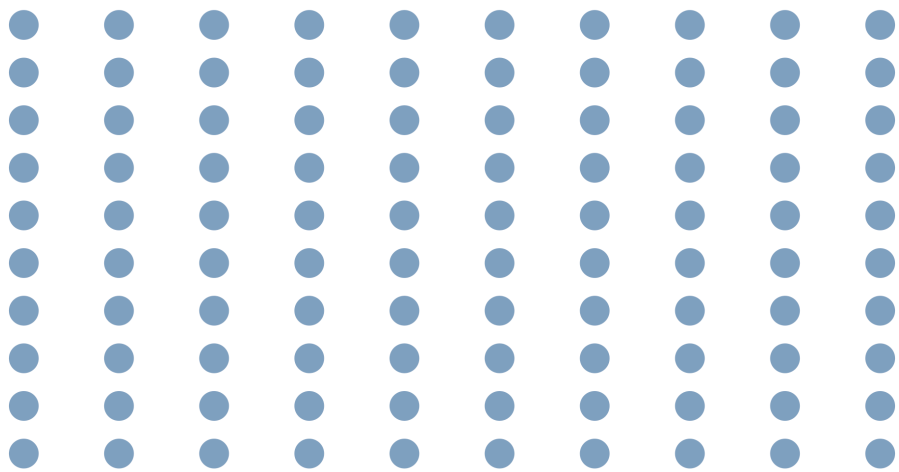
1. Each observation is a random variable.
2. The n random variables are independent.

Life becomes much simpler for the econometrician.

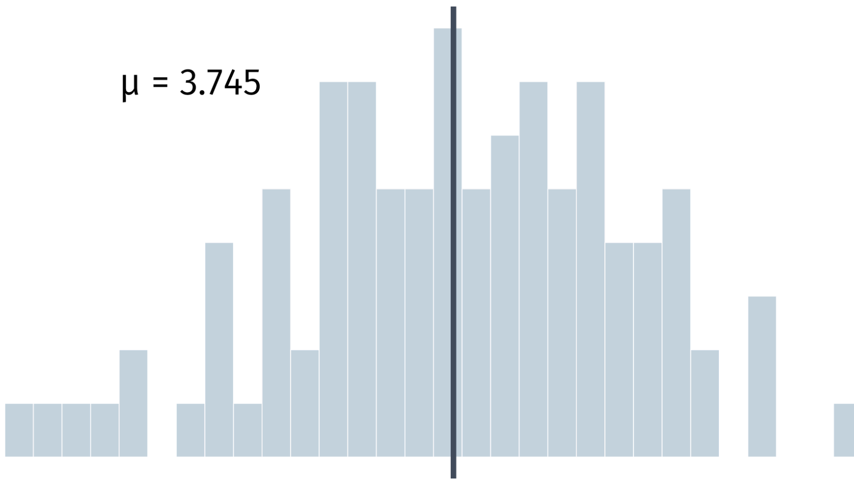
1. Only a subset of these can help produce reliable statistics.

Population *vs.* sample

Question: *Why do we care about population vs. sample?*



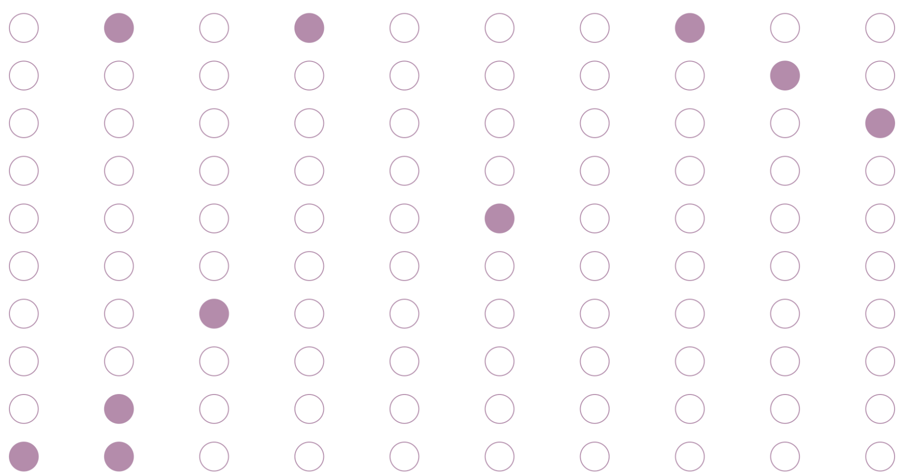
Population



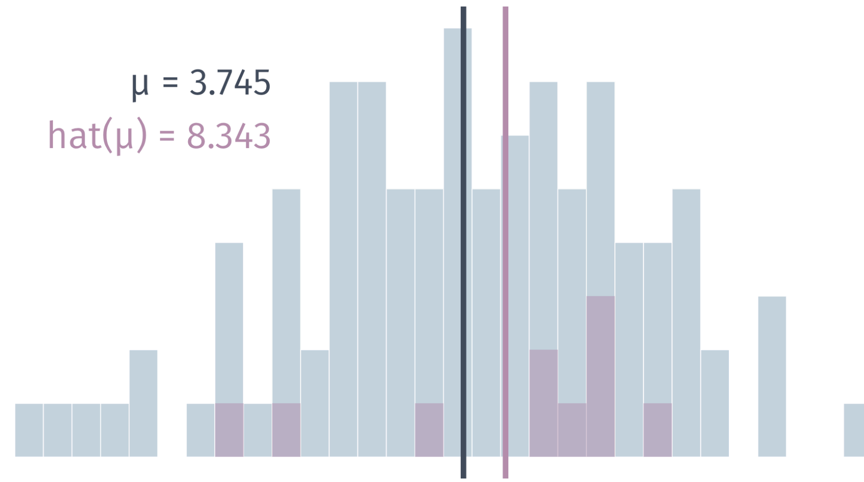
Population relationship

Population vs sample

Question: *Why do we care about population vs. sample?*



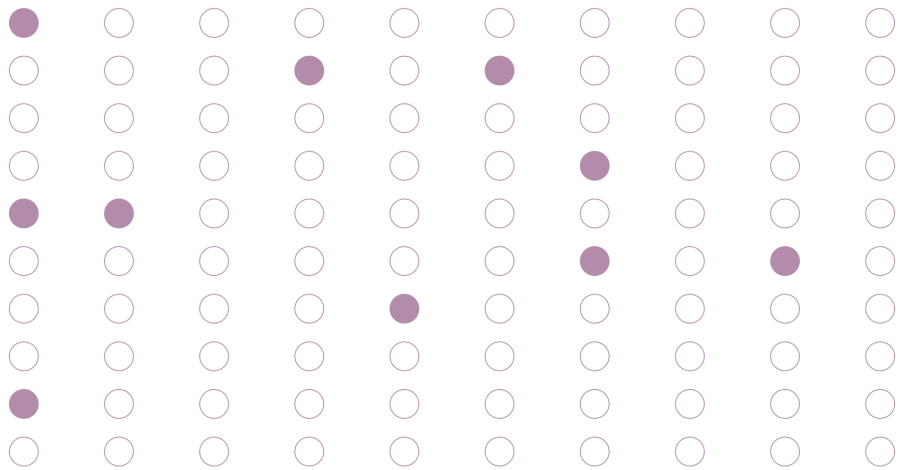
10 random individuals



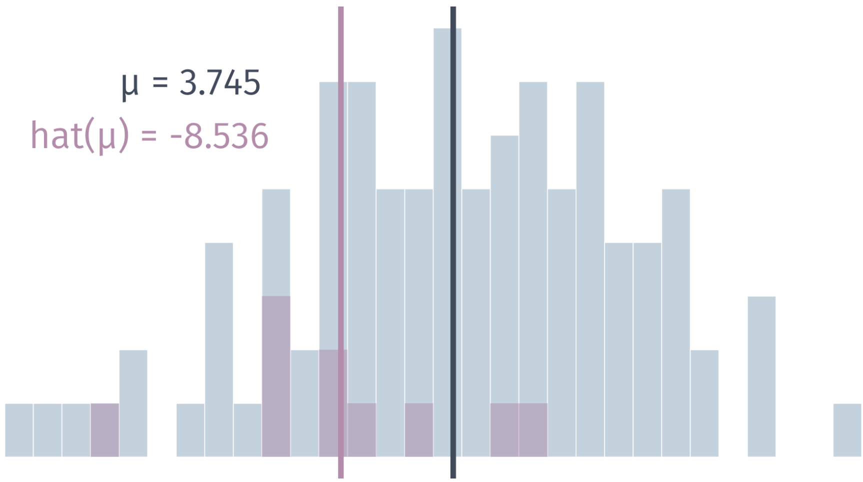
Population relationship

Population *vs* sample

Question: *Why do we care about population vs. sample?*



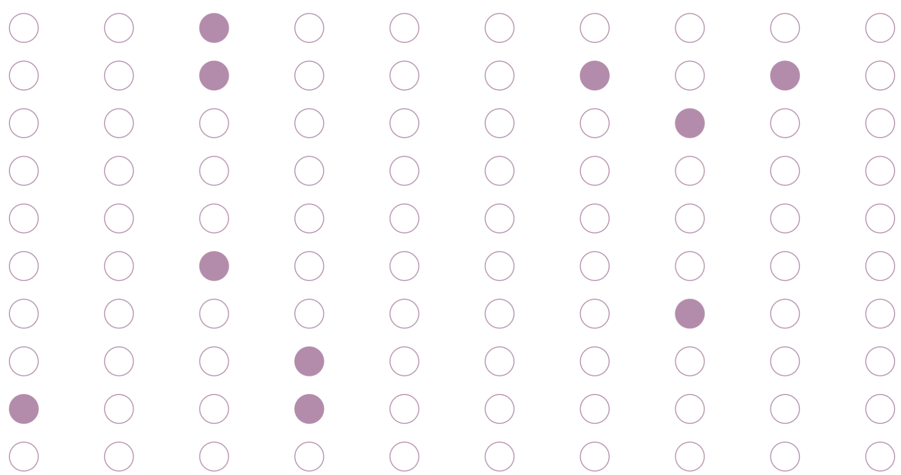
10 random individuals



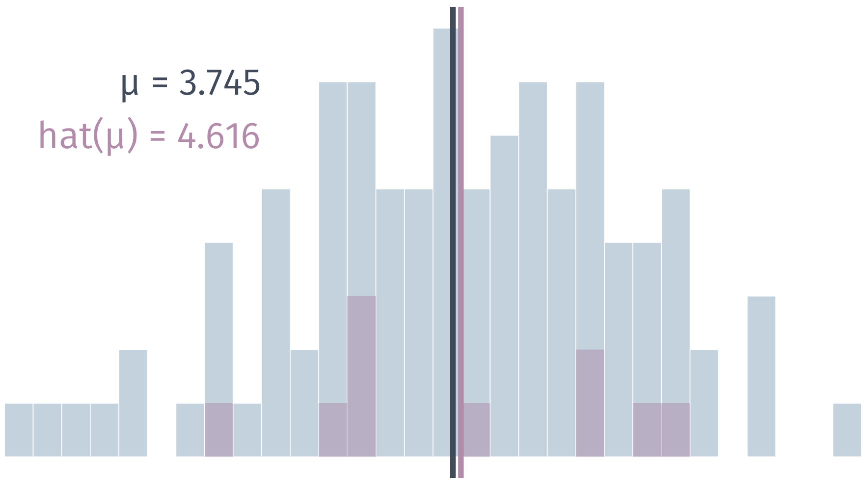
Population relationship

Population *vs* sample

Question: *Why do we care about population vs. sample?*



10 random individuals



Population relationship

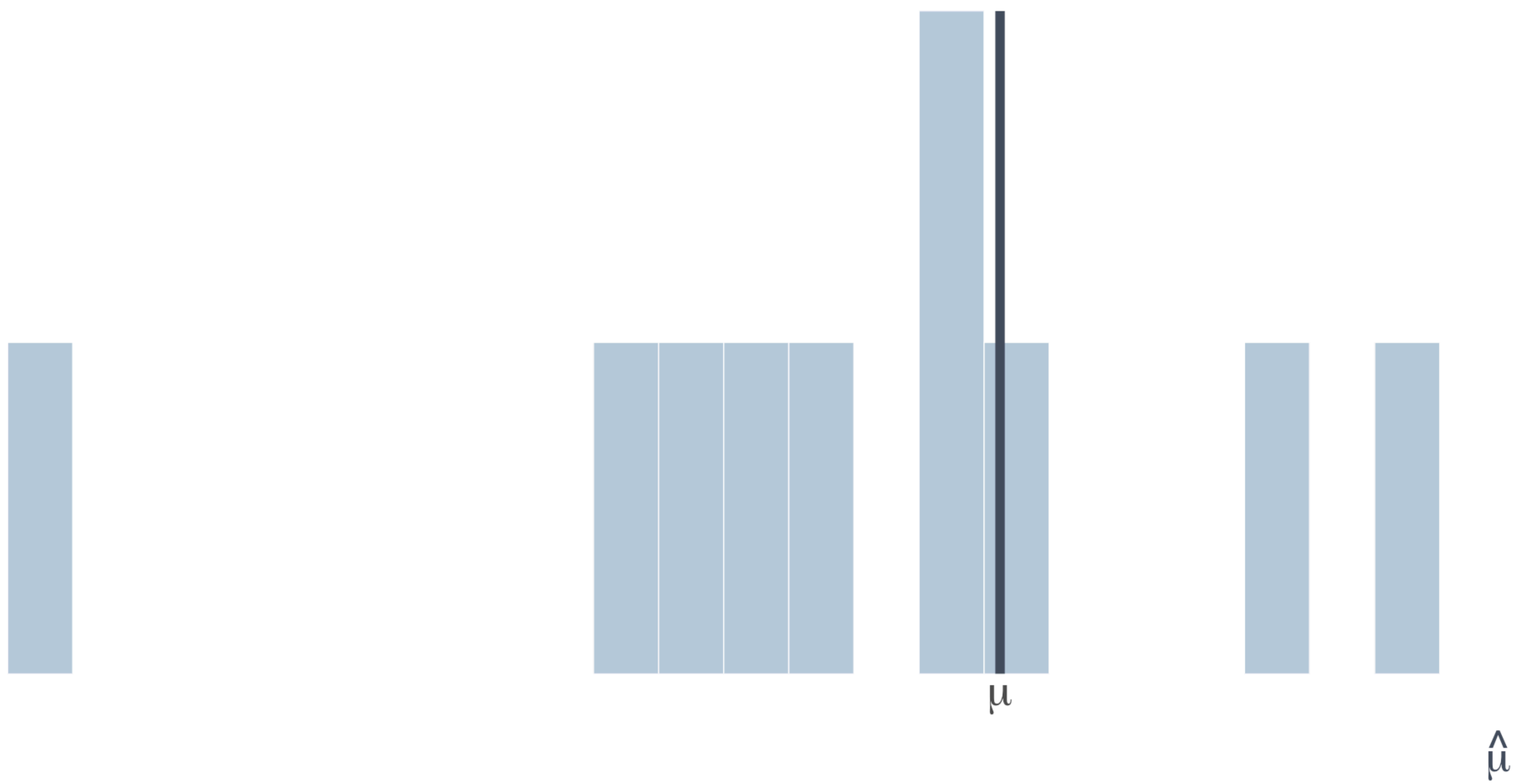
Let's repeat this **10,000 times** and then plot the estimates.

(This exercise is called a Monte Carlo simulation.)

How in the world do I do that

- Show the code

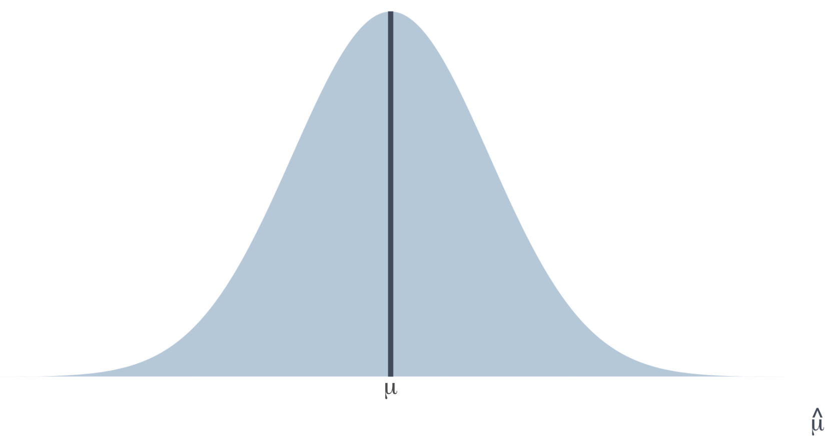
10 100 1,000 10,000 100,000



Regular resampling means of 10 obs at a time

Population *vs.* sample

Question: *Why do we care about population vs. sample?*



As the number of samples approach infinity

On average, the mean of the samples are close to the population mean

- Some individual samples can miss the mark.
- The difference between individual samples and the population creates **uncertainty**

Population *vs.* sample

Question: *Why do we care about population vs. sample?*

Answer: Uncertainty matters.

- $\hat{\mu}$ is a random variable that depends on the sample.
- We don't know if our sample is representative of the population.
- Individual sample means can be biased
- We have to keep track of this uncertainty.

Population distributions

Consider the following argument (this slide scrolls down)

Suppose we have some estimator $\hat{\theta}$ for a parameter θ:

- θ is unobserved, but assume $\hat{\theta}$ follows a probability distribution $p(\hat{\theta})$
- We hypothesize some value, say $\theta = 2.5$
- We use our estimator $\hat{\theta}$ to calculate an estimate. $\hat{\theta} = 45$
- If we make an ***assumption*** of the distribution of $\hat{\theta}$, we can calculate the probability of getting $\hat{\theta} = 45$ when $\theta = 2.5$ is true.
- For sake of argument, let's say that the probability that $\theta = 2.5$ if we observe $\theta = 45$ is less than 0.001

We can say

> *if θ really was 2.5, then the probability of getting $\hat{\theta} = 45$ is super super low. Thus the probability that θ is actually 2.5 is super super low".*

- We can make statements about the true value of θ just by knowing the distribution of our preferred estimator $\hat{\theta}$

But what distribution should we be assuming?

The Central Limit Theorem

Theorem

> *Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with mean $E[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2 < \infty$, let $\bar{X}$ be the sample mean. Then, as $n \to \infty$, the function $\frac{\sqrt{n}(\bar{X}-\mu)}{S_x}$ converges to a* Normal Distribution *with mean 0 and variance 1.*

- CLT states that when $n \to \infty$, the sample mean will be normally distributed.

- The Law of Large Number (LLN) states that as $n \to \infty$, the sample converges on the population mean.

The Central Limit Theorem

Some interesting YouTube links:

- A more in depth explanation + visualization
- What is so special about the normal distribution?

Data types

Data

There are **two** broad types of data

1. **Experimental data**

> Data generated in controlled, laboratory settings[1]

Ideal for **causal identification**, but difficult to obtain

- Logistically intractable
- Expensive
- Morally repugnant

1. Note: Experiments can often occur outside the lab (eg randomized control trials and A/B testing)

Data

There are **two** broad types of data

1. **Experimental data**
2. **Observational data**

> Data generated in non-experimental settings

Types of observational data:

- Surveys
- Census
- Administrative data
- Environmental data
- Transaction data
- Text and image data

Commonly used though poses challenges to **causal identification**

Data types: Cross sectional

> Sample of individuals from a population at a point in time

Ideally collected using **random sampling**

- **random sampling** $+$ **sufficient sample size** $=$ **representative sample**
- Non-random sampling is more common and difficult to work with

Note: Used extensively in applied microeconomics[1] and is the main focus of this course

1. Applied microeconomics = Labor, health, education, public finance, development, industrial organization, and urban economics

Data types: Time series

> Observations of variables over time

Ex.

- Quarterly GDP
- Daily stock prices
- Annual infant mortality rates

Complication: Observations are not independent draws

- eg GDP this quarter is highly correlated to GDP last quarter

More advanced methods needed[1]

1. See EC 421 and EC 422

Data types: Pooled cross sectional

> Cross sections from different points in time

Useful for studying relationship that change over time.

Again, requires more advanced methods[1]

1. See EC 421 and many of the 400-level applied classes

Data types: Panel data

> Time series for each cross sectional unit

Ex. Daily attendance across my class

Can control for `unobserved` characteristics

Again, requires more advanced methods[1]

1. See EC 421 and many of the 400-level applied classes

Data types: Messy data

Analysis ready dataset are rare. Most data are *messy*

Data wrangling is a non-trivial part of an economist or data scientist/analyst's job

R has a suite of packages that facilitate data wrangling:

- The `tidyverse`: `readr`, `tidyr`, `dplyr`, `ggplot2` + others

Table of Contents

Admin

Review

Appendix

Variance: Rule 03 Expanded

Back to Variance Rule 03

The variance of a random variable X is defined as:

$$\text{Var}(X) = E[(X - \mu_X)^2]$$

$\text{Cov}(X, Y)$ is defined as:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

For two random variables X and Y, the variance of their sum $X + Y$ is:

$$\text{Var}(X + Y) = E[((X + Y) - (\mu_X + \mu_Y))^2]$$

Expanding the squared term, we get:

$$\begin{aligned}\text{Var}(X+Y) &= E[(X-\mu_X+Y-\mu_Y)^2] \\ &= E[(X-\mu_X)^2+2(X-\mu_X)(Y-\mu_Y)+(Y-\mu_Y)^2] \\ &= E[(X-\mu_X)^2]+E[2(X-\mu_X)(Y-\mu_Y)]+E[(Y-\mu_Y)^2] \\ &= \text{Var}(X)+2\text{Cov}(X,Y)+\text{Var}(Y)\end{aligned}$$

If X and Y are uncorrelated, then $\text{Cov}(X,Y)=0$, and the above simplifies to:

$$\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$$

Similarly, the variance of the difference $X-Y$ is:

$$\text{Var}(X-Y)=E[((X-Y)-(\mu_X-\mu_Y))^2]$$

Expanding the squared term, just like before:

$$\begin{aligned}\text{Var}(X-Y) &= E[(X-\mu_X-(Y-\mu_Y))^2] \\ &= E[(X-\mu_X)^2-2(X-\mu_X)(Y-\mu_Y)+(Y-\mu_Y)^2] \\ &= \text{Var}(X)-2\text{Cov}(X,Y)+\text{Var}(Y)\end{aligned}$$

Again, if X and Y are uncorrelated, $\text{Cov}(X,Y)=0$, and we have:

$$\text{Var}(X-Y)=\text{Var}(X)+\text{Var}(Y)$$