Week 2: Random Variables

Brandon Stewart[1]

Princeton

September 7–11, 2020

[1]These slides are heavily influenced by Adam Glynn, Justin Grimmer, Jens Hainmueller and Ian Lundberg. Many illustrations by Shay O'Brien.

Where We've Been and Where We're Going...

- Last Week

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Example: Ballot Order

Example: Ballot Order

Evidence suggests that candidates gain a small advantage from ballot order.

Example: Ballot Order

Evidence suggests that candidates gain a small advantage from ballot order.

As a response, in 2008 New Hampshire chose a letter from the alphabet and then listed the candidates in alphabetical order starting with that letter.

Example: Ballot Order

Evidence suggests that candidates gain a small advantage from ballot order.

As a response, in 2008 New Hampshire chose a letter from the alphabet and then listed the candidates in alphabetical order starting with that letter.

We can use probability to assess the "fairness" of this process.

Example: Ballot Order

Evidence suggests that candidates gain a small advantage from ballot order.

As a response, in 2008 New Hampshire chose a letter from the alphabet and then listed the candidates in alphabetical order starting with that letter.

We can use probability to assess the "fairness" of this process.

We will do this by introducing a random variable X to be Barack Obama's position on the 2008 New Hampshire primary ballot.

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

Imagine two coin flips

$$\Omega = \Big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\Big\}$$

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

Imagine two coin flips

$$\Omega = \big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\big\}$$

we could define a random variable $X(\omega)$ to be the function that returns the number of heads for each element (ω) of the sample space (Ω).

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

Imagine two coin flips

$$\Omega = \Big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\Big\}$$

we could define a random variable $X(\omega)$ to be the function that returns the number of heads for each element (ω) of the sample space (Ω).

- $X(\{heads, heads\}) = 2$
- $X(\{heads, tails\}) = 1$
- $X(\{tails, heads\}) = 1$
- $X(\{tails, tails\}) = 0$

What is a Random Variable?

Intuition: functions that map outcomes to numbers.

Formal: X is a function that maps the sample space to the real numbers.

Imagine two coin flips

$$\Omega = \Big\{\{heads, heads\}, \{heads, tails\}, \{tails, heads\}, \{tails, tails\}\Big\}$$

we could define a random variable $X(\omega)$ to be the function that returns the number of heads for each element (ω) of the sample space (Ω).

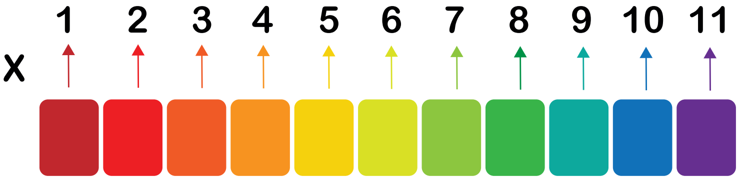
- $X(\{heads, heads\}) = 2$
- $X(\{heads, tails\}) = 1$
- $X(\{tails, heads\}) = 1$
- $X(\{tails, tails\}) = 0$

We will generally suppress the function notation and just refer to X.

A Visual Example

A Visual Example

A Visual Example



A Brief Note on Notation

A Brief Note on Notation

- We almost always use capital roman letters for the “name” of the random variable such as X.

A Brief Note on Notation

- We almost always use capital roman letters for the "name" of the random variable such as X.
- We refer to a fixed value with a lower case letter x.

A Brief Note on Notation

- We almost always use capital roman letters for the "name" of the random variable such as X.
- We refer to a fixed value with a lower case letter x.
- So we might write $P(X = x)$ to be the probability that the number of heads we observe is equal to some fixed value x.

A Brief Note on Notation

- We almost always use capital roman letters for the "name" of the random variable such as X.
- We refer to a fixed value with a lower case letter x.
- So we might write $P(X = x)$ to be the probability that the number of heads we observe is equal to some fixed value x.
- We will sometimes write out the mapping from the sample space to the random variable. For example,

$$X = \begin{cases} 1 \text{ if heads} \\ 0 \text{ if tails} \end{cases}$$

A Brief Note on Notation

- We almost always use capital roman letters for the “name” of the random variable such as X.
- We refer to a fixed value with a lower case letter x.
- So we might write $P(X = x)$ to be the probability that the number of heads we observe is equal to some fixed value x.
- We will sometimes write out the mapping from the sample space to the random variable. For example,

$$X = \begin{cases} 1 \text{ if heads} \\ 0 \text{ if tails} \end{cases}$$

- Other times the sample space is already numeric so its more obvious (e.g. how many minutes until the train arrives).

Quick FAQ

Quick FAQ

- Why have random variables at all?

Quick FAQ

- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.

Quick FAQ

- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.
- Why are they <u>random</u> variables?

Quick FAQ

- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.
- Why are they <u>random</u> variables?
  they are realizations of a stochastic process (i.e. randomness in the outcome, not the mapping).

Quick FAQ

- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.
- Why are they <u>random</u> variables?
  they are realizations of a stochastic process (i.e. randomness in the outcome, not the mapping).
- Is it really easier this way? It seems hard.

Quick FAQ

- Why have random variables at all?
  it makes the math easier, even across very different sample spaces.
- Why are they <u>random</u> variables?
  they are realizations of a stochastic process (i.e. randomness in the outcome, not the mapping).
- Is it really easier this way? It seems hard.
  random variables are about bridging the abstract math and the concrete world. that can be hard, but it is super important and better than the alternative!

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \left\{ \right.$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

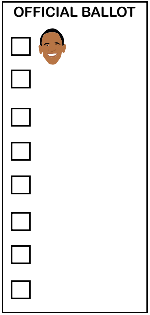
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

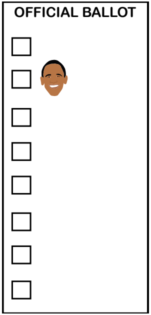
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

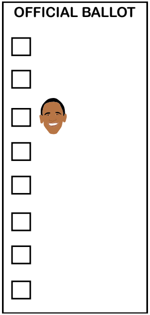
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ \\ \\ \\ \\ \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

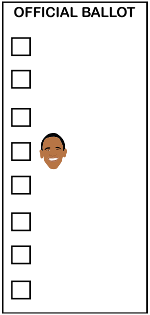
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ \\ \\ \\ \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

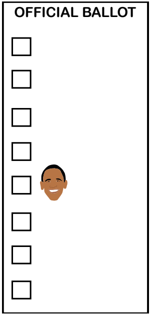
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

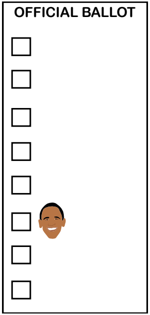
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

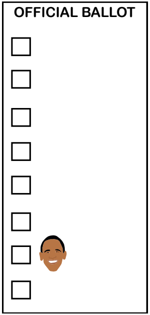
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

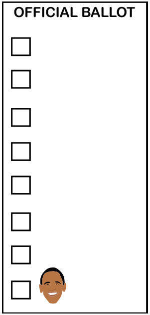
X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

NH Ballot Order Example

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$X = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

X is a random variable indicating Obama's position on the ballot. Highlighted letters are those leading to a given ballot position. Highlighted individual is first.

Discrete Distributions

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.
- For discrete distributions, the random variable X takes on a finite, or a countably infinite number of values.

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.
- For discrete distributions, the random variable X takes on a finite, or a countably infinite number of values.
- A common shorthand is to think of discrete random variables taking on distinct values.

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.
- For discrete distributions, the random variable X takes on a finite, or a countably infinite number of values.
- A common shorthand is to think of discrete random variables taking on distinct values.
- A probability mass function (PMF) and a cumulative distribution function (CDF) are two common ways to define the probability distribution for a discrete random variable.

Discrete Distributions

- The distribution of a random variable specifies the probability of all events associated with that random variable.
- For discrete distributions, the random variable X takes on a finite, or a countably infinite number of values.
- A common shorthand is to think of discrete random variables taking on distinct values.
- A probability mass function (PMF) and a cumulative distribution function (CDF) are two common ways to define the probability distribution for a discrete random variable.
- Probability mass functions provide a compact way to represent information about how likely various outcomes are.

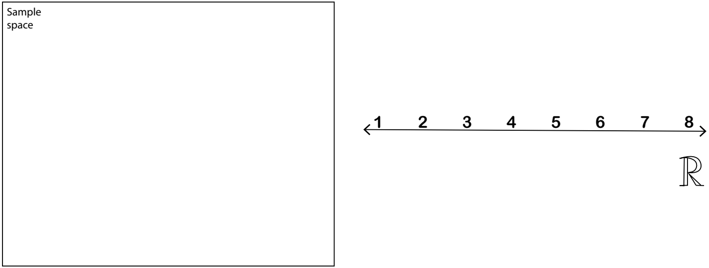
Where do Distributions Come From?

Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.

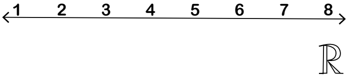
Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.



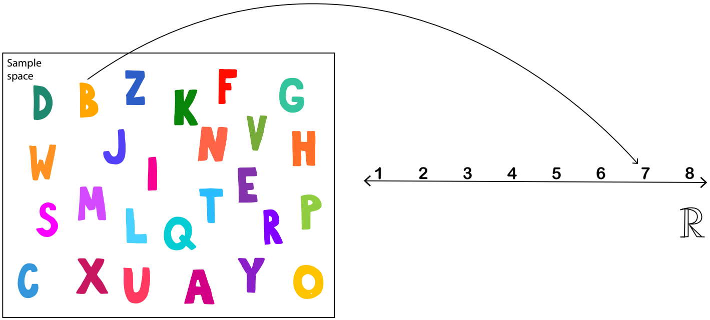
Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.



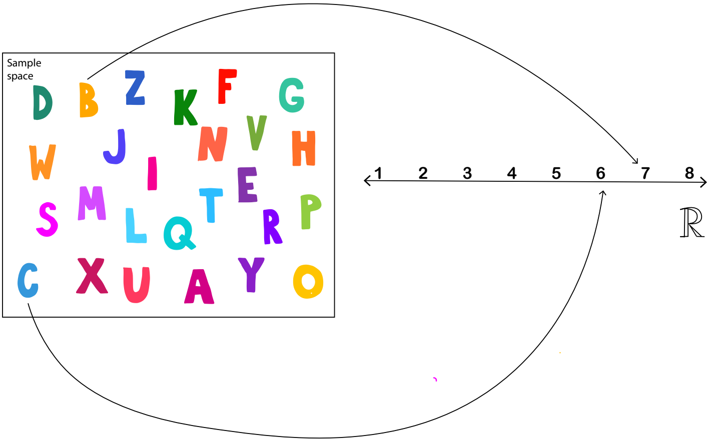
Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.



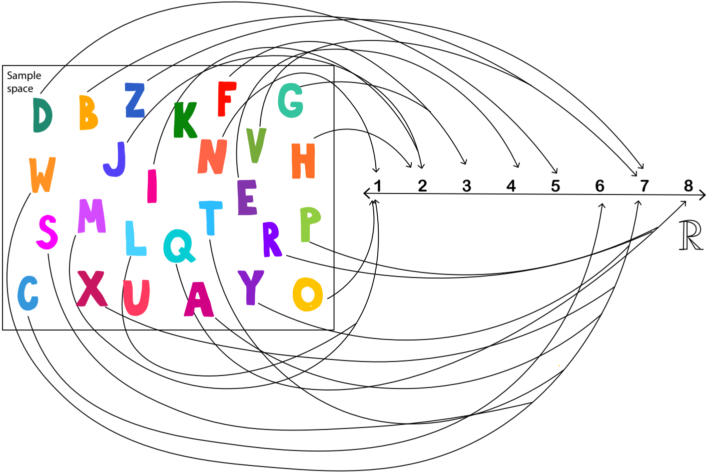
Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.



Where do Distributions Come From?

The probabilities associated with each realization of the random variables come from the underlying stochastic realization of the sample space.

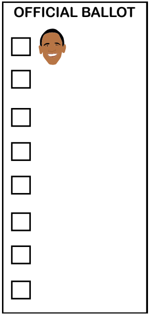


Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

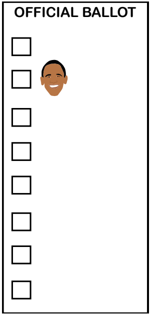
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

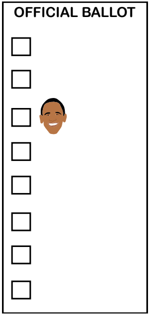
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

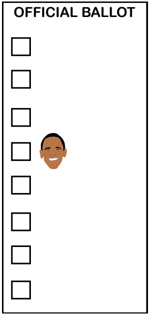
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

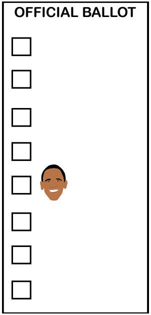
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \\ 1/26 & x = 5 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

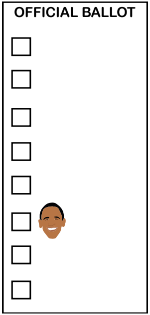
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \\ 1/26 & x = 5 \\ 1/26 & x = 6 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

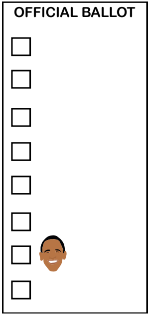
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \\ 1/26 & x = 5 \\ 1/26 & x = 6 \\ 10/26 & x = 7 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

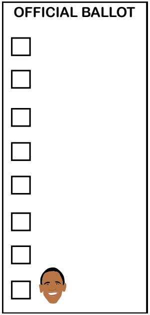
Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Example: New Hampshire

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$p_X(x) = \begin{cases} 4/26 & x = 1 \\ 4/26 & x = 2 \\ 2/26 & x = 3 \\ 1/26 & x = 4 \\ 1/26 & x = 5 \\ 1/26 & x = 6 \\ 10/26 & x = 7 \\ 3/26 & x = 8 \end{cases}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Probability of the random variable equaling a number is just the probability of the underlying event (subset of the sample space).

Discrete Probability Mass Functions

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.
- Later we will drop the subscript when it is clear from context.

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.
- Later we will drop the subscript when it is clear from context.

Three key properties:

- this will always be non-negative.

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.

- Later we will drop the subscript when it is clear from context.

Three key properties:

- this will always be non-negative.
- the support of X is the set of values where the PMF is non-zero.

Discrete Probability Mass Functions

Definition (Probability Mass Function)

The probability mass function (PMF) of a discrete random variable X is the function p_X given by,

$$p_X(x) = P(X = x)$$

Understanding the Notation:

- $X = x$ is defining an event.

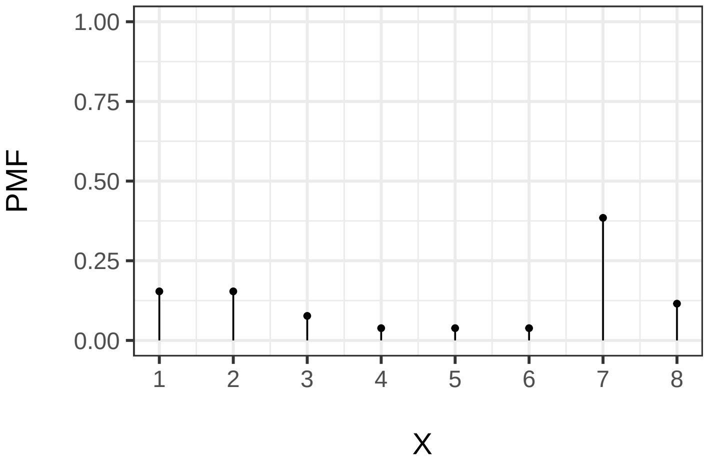
  More formally we might say, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ which can be read as the set of realizations ω in the sample space Ω such that the function $X(\omega)$ returns the fixed value x.

- Later we will drop the subscript when it is clear from context.

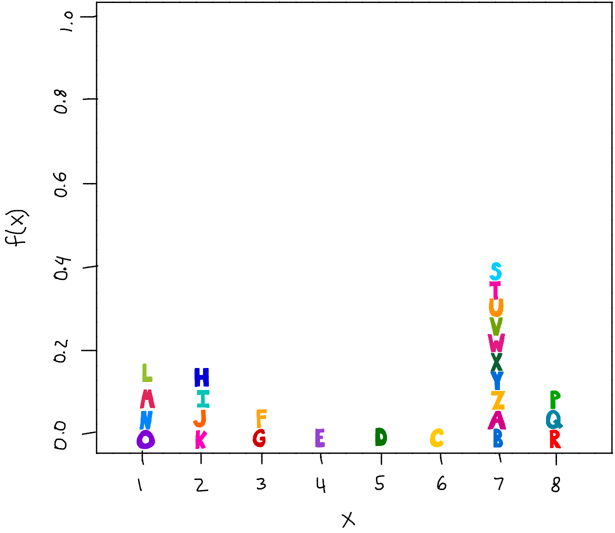
Three key properties:

- this will always be non-negative.
- the support of X is the set of values where the PMF is non-zero.
- $\sum_x p_X(x) = 1$.

NH Obama Ballot Position PMF Plot



NH Obama Ballot Position PMF Plot



Cumulative Distribution Function

Cumulative Distribution Function

Definition (Cumulative Distribution Function)

The cumulative distribution function (CDF) of a random variable X is the function F_X given by,

$$F_X(x) = P(X \leq x)$$

Cumulative Distribution Function

Definition (Cumulative Distribution Function)

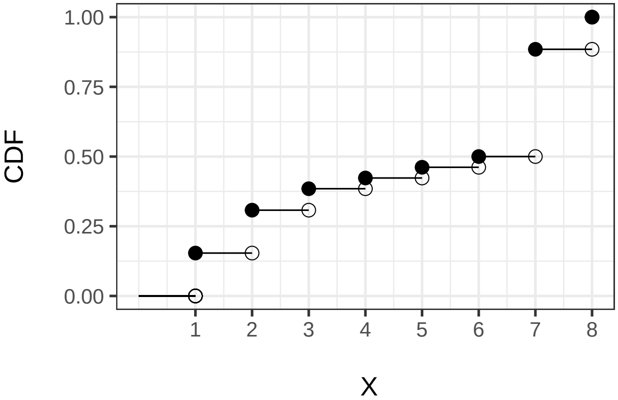
The cumulative distribution function (CDF) of a random variable X is the function F_X given by,

$$F_X(x) = P(X \leq x)$$

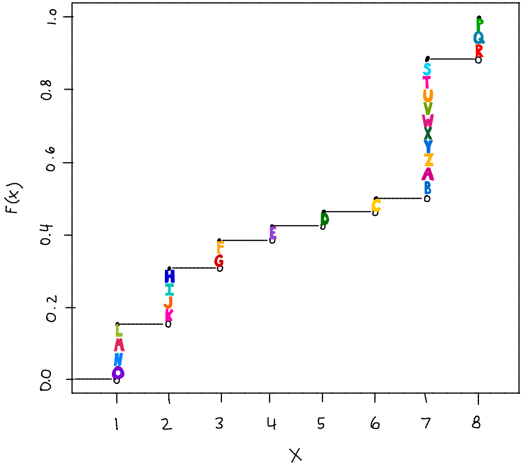
Key properties:

- non-decreasing
- right-continuous
- converges to 0 and 1 in the limits

NH Obama Ballot Position CDF Plot



NH Obama Ballot Position CDF Plot



Example Discrete Distributions

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

  - Discrete Uniform distribution: Let X be a random variable that puts equal probability on each value that X can take:

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

  - Discrete Uniform distribution: Let X be a random variable that puts equal probability on each value that X can take:

$$p_X(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

  - Discrete Uniform distribution: Let X be a random variable that puts equal probability on each value that X can take:

$$p_X(x) = \begin{cases} 1/k & \text{for } x = 1, \dots, k \\ 0 & \text{otherwise} \end{cases}$$

- We can summarize these distributions with one number

Example Discrete Distributions

- A major advantage of random variables is that they often have a distribution with a known form (that comes with known results!)
  - Bernoulli distribution: Let X be a binary variable with $P(X = 1) = \pi$ and, thus, $P(X = 0) = 1 - \pi$, where $\pi \in [0, 1]$. It has PMF:

$$p_X(x) = \pi^x (1 - \pi)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

  - Discrete Uniform distribution: Let X be a random variable that puts equal probability on each value that X can take:

$$p_X(x) = \begin{cases} 1/k & \text{for } x = 1, \dots, k \\ 0 & \text{otherwise} \end{cases}$$

- We can summarize these distributions with one number
- We will return to this in the last video of the week.

We Covered. . .

We Covered. . .

- The definition of a random variable.

We Covered...

- The definition of a random variable.
- Probability mass functions and cumulative distribution functions.

We Covered...

- The definition of a random variable.
- Probability mass functions and cumulative distribution functions.

Next time continuous random variables.

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Definition of Random Variables

2. Continuous Distribution
   - Defining a Continuous Random Variable
   - Probability Density Functions and Cumulative Distribution Functions
   - Subtleties of the Continuous Setting

3. Expectation as a Measure of Central Tendency

4. Variance as a Measure of Dispersion

5. Joint and Conditional Distributions

6. Characterizing Conditional Distributions

7. Independence and Covariance

Continuous Distributions

Continuous Distributions

- Continuous random variables take on an uncountably infinite number of values.

Continuous Distributions

- Continuous random variables take on an uncountably infinite number of values.
- This is often a useful approximation when variables take many values.

Continuous Distributions

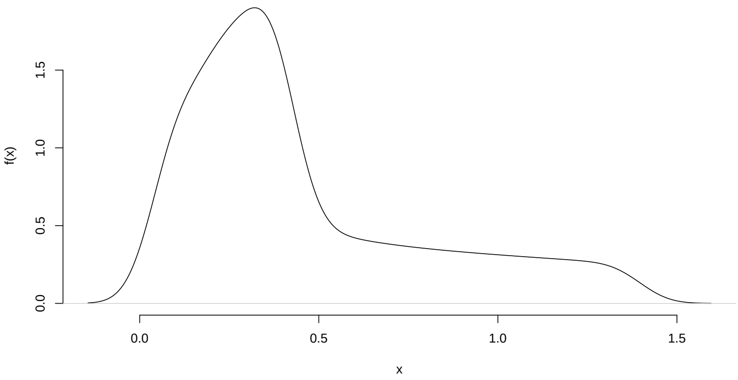
- Continuous random variables take on an uncountably infinite number of values.
- This is often a useful approximation when variables take many values.
- A probability density function (PDF) and a cumulative distribution function (CDF) are two common ways to define the distribution for a continuous random variable.

Continuous Distributions

- Continuous random variables take on an uncountably infinite number of values.
- This is often a useful approximation when variables take many values.
- A probability density function (PDF) and a cumulative distribution function (CDF) are two common ways to define the distribution for a continuous random variable.
- They are similar to the discrete case with a few subtle differences.

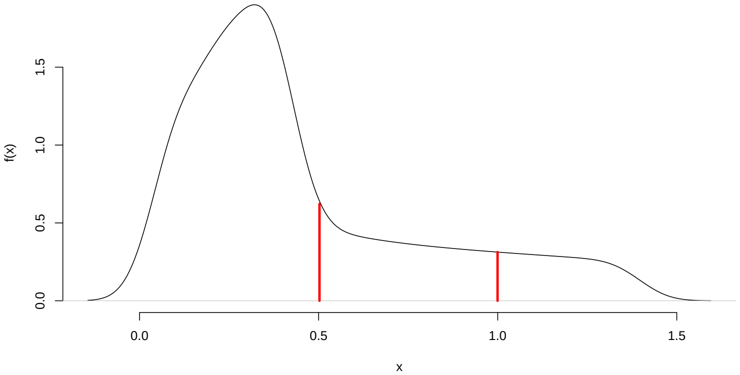
Calculus Review: Integration

Suppose we have some function $f(x)$



Calculus Review: Integration

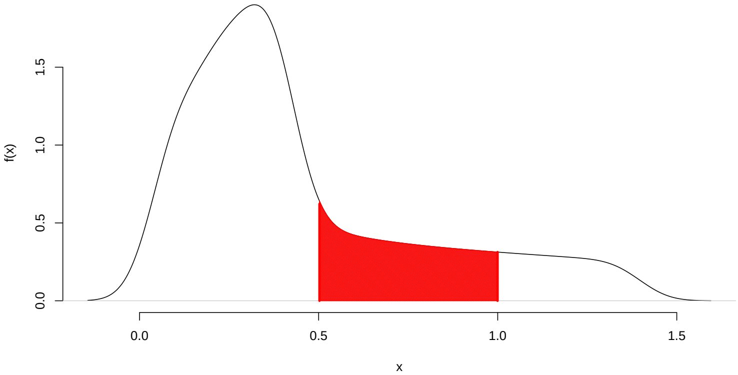
Suppose we have some function $f(x)$



What is the area under $f(x)$ between $\frac{1}{2}$ and 1?

Calculus Review: Integration

Suppose we have some function $f(x)$



What is the area under $f(x)$ between $\frac{1}{2}$ and 1?

Area under curve $= \int_{1/2}^{1} f(x)dx = F(1) - F(1/2)$

Continuous Random Variable

Continuous Random Variable

A continuous random variable has a continuous cumulative distribution function (CDF) which, as in the discrete case, defines the probability that $P(X \leq x)$.

Continuous Random Variable

A continuous random variable has a continuous cumulative distribution function (CDF) which, as in the discrete case, defines the probability that $P(X \leq x)$.

Definition (Continuous Distribution)

A random variable has a continuous distribution if its CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. (Blizstein and Hwang Definition 5.1.1)

Probability Density Function (PDF)

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Definition (Probability density function)

For a continuous random variable X with CDF F_X, the probability density function of X is the derivative f of the CDF, given by $f_X(x) = \frac{d}{dx}F_X(x)$

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Definition (Probability density function)

For a continuous random variable X with CDF F_X, the probability density function of X is the derivative f of the CDF, given by $f_X(x) = \frac{d}{dx}F_X(x)$

Key Properties:

- non-negative

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Definition (Probability density function)

For a continuous random variable X with CDF F_X, the probability density function of X is the derivative f of the CDF, given by $f_X(x) = \frac{d}{dx}F_X(x)$

Key Properties:

- non-negative
- integrates to 1. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Probability Density Function (PDF)

The probability density function is the analog of the probability mass function for discrete random variables.

Definition (Probability density function)

For a continuous random variable X with CDF F_X, the probability density function of X is the derivative f of the CDF, given by $f_X(x) = \frac{d}{dx} F_X(x)$

Key Properties:

- non-negative
- integrates to 1. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- for any measurable set of real numbers B,

$$P(X \in B) = \int_B f_X(x)dx$$

Defining the CDF in terms of the PDF

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

pdf

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

pdf $\xrightarrow{\text{Integrate}}$

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

$$\text{pdf} \xrightarrow{\text{Integrate}} \text{cdf}$$

Defining the CDF in terms of the PDF

Definition (CDF of a Continuous Random Variable)

For a continuous random variable X define its cumulative distribution function $F_X(x)$ as,

$$F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(x)dx$$

$$\begin{tikzcd} \text{pdf} \arrow[r, bend left, "\text{Integrate}"] & \text{cdf} \arrow[l, bend left, "\text{Differentiate}"] \end{tikzcd}$$

A Visual Example

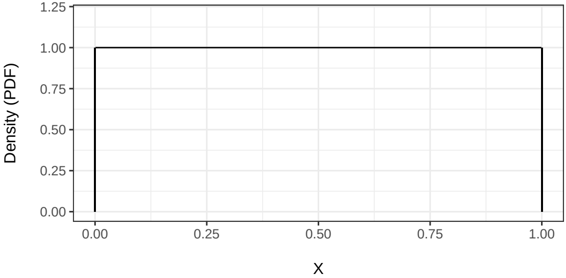
Imagine you choose a number completely at random between 0 and 1 with all equally sized sets of values being equally likely.

A Visual Example

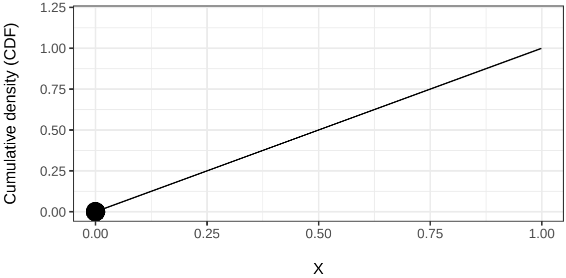
Imagine you choose a number completely at random between 0 and 1 with all equally sized sets of values being equally likely. This is a standard uniform distribution which has the CDF,

$$F_X(x) = x$$

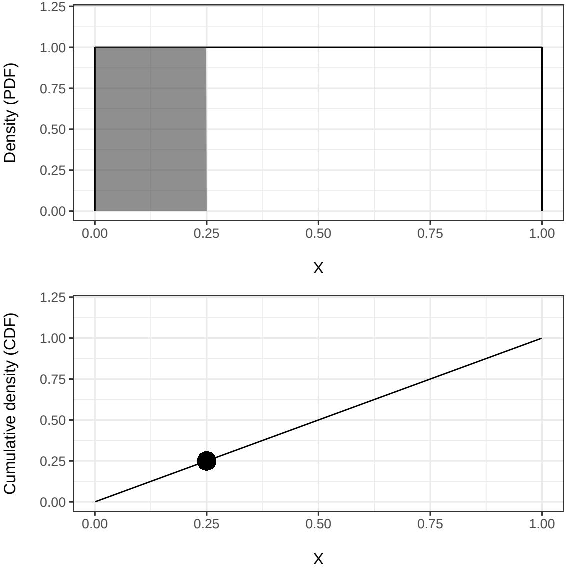
with support over $[0, 1]$.

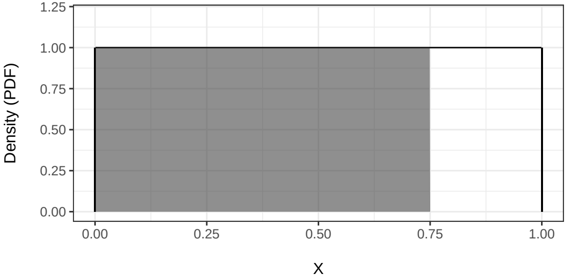


Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X

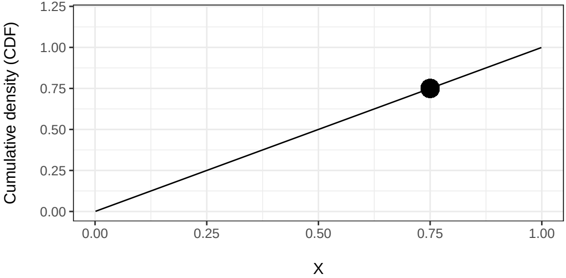


Cumulative density (CDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X

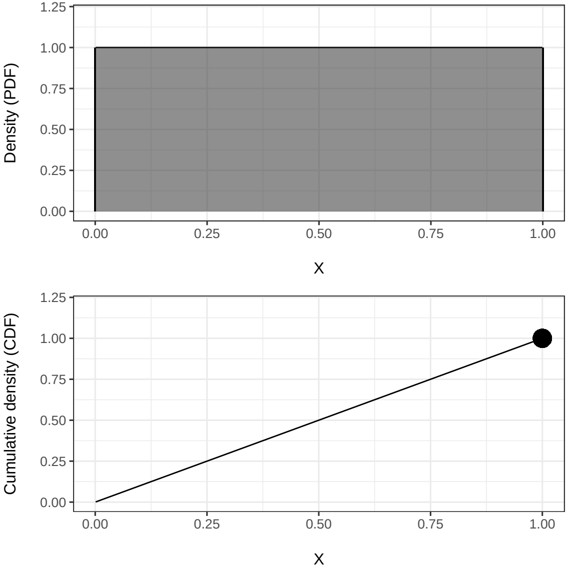

1.25
1.00
0.75
0.50
0.25
0.00
Density (PDF)
0.00
0.25
0.50
0.75
1.00
X
1.25
1.00
0.75
0.50
0.25
0.00
Cumulative density (CDF)
0.00
0.25
0.50
0.75
1.00
X



Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X

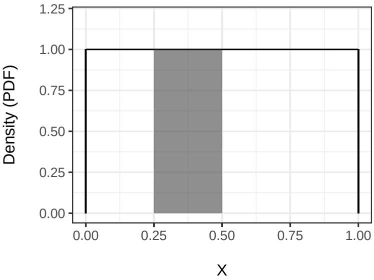


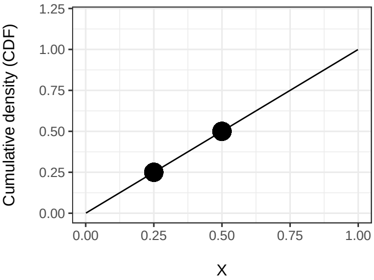
Cumulative density (CDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X


Density (PDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X
Cumulative density (CDF)
1.25
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
X

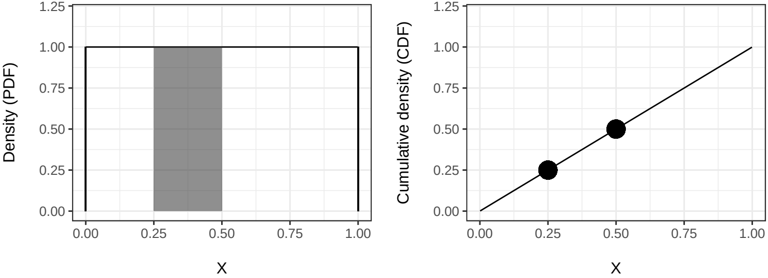
What is the probability that the number is between 0.25 and 0.5?

What is the probability that the number is between 0.25 and 0.5?





What is the probability that the number is between 0.25 and 0.5?



$$F_X(.5) - F_X(.25) = .25$$

The Core PMF/PDF Difference

The Core PMF/PDF Difference

The probability mass function provides the probability of a set of outcomes by summing over the probability mass function evaluated at each of those outcomes.

The Core PMF/PDF Difference

The probability mass function provides the probability of a set of outcomes by summing over the probability mass function evaluated at each of those outcomes.

The probability density function for continuous variables provides the probability of a set of outcomes by integrating over a range of values.

The Core PMF/PDF Difference

The probability mass function provides the probability of a set of outcomes by summing over the probability mass function evaluated at each of those outcomes.

The probability density function for continuous variables provides the probability of a set of outcomes by integrating over a range of values.

This means (perhaps counterintuitively) that a probability density function:

- can return a value greater than 1
- assigns the probability of any exact value is zero.

The Core PMF/PDF Difference

The probability mass function provides the probability of a set of outcomes by summing over the probability mass function evaluated at each of those outcomes.

The probability density function for continuous variables provides the probability of a set of outcomes by integrating over a range of values.

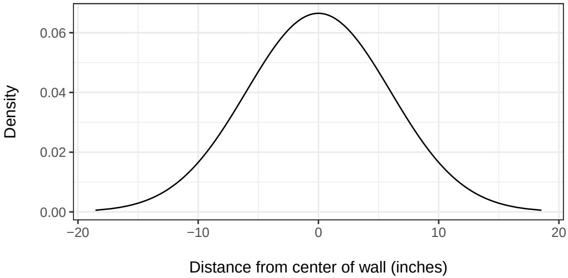
This means (perhaps counterintuitively) that a probability density function:

- can return a value greater than 1
- assigns the probability of any exact value is zero.

Let's explain!

A Numerical Example

Let's suppose we have someone throwing darts and we measure how far they are from the center of the wall in inches. In this case, perhaps the darts will be distributed with the following PDF.

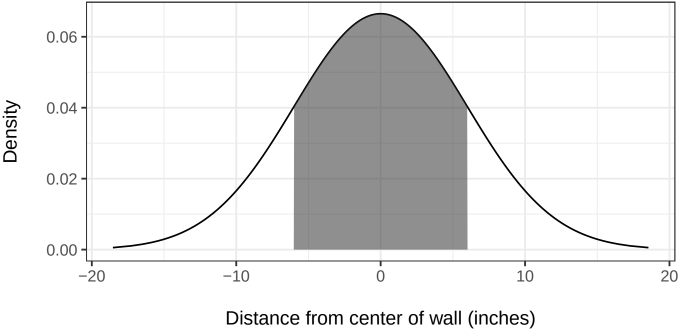


A Numerical Example

How would we calculate the probability that a dart lands within 6 inches of the center of the wall?

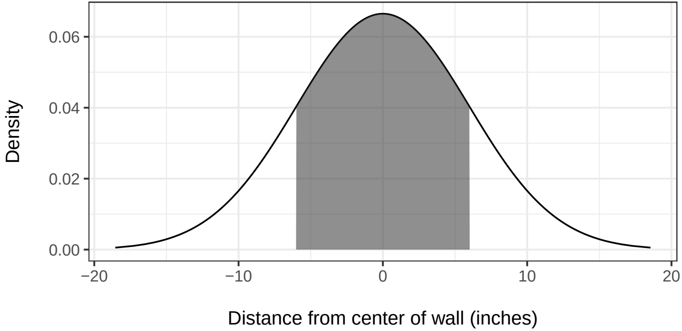
A Numerical Example

How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



A Numerical Example

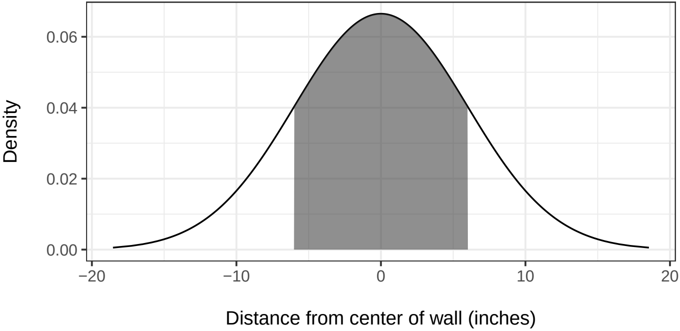
How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



$$P(X \in (-6, 6)) \quad =$$

A Numerical Example

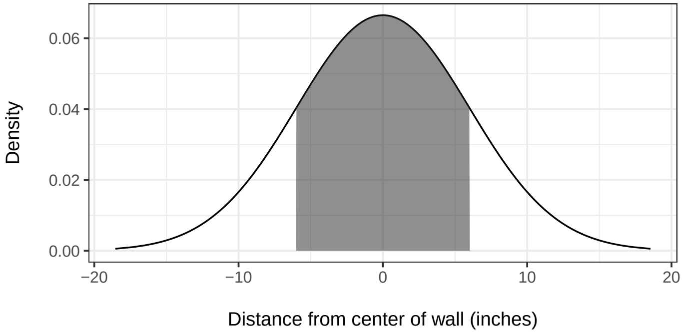
How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



$$P(X \in (-6,6)) \quad = \quad \int_{-6}^{6} f_X(x)dx$$

A Numerical Example

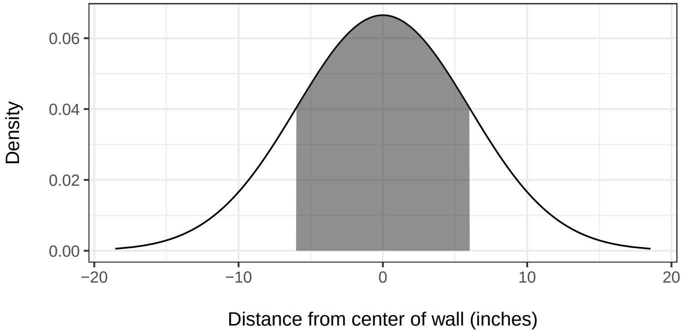
How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



$$\begin{aligned} P(X \in (-6,6)) &= \int_{-6}^{6} f_X(x)dx \\ &= F_X(6) - F_X(-6) \end{aligned}$$

A Numerical Example

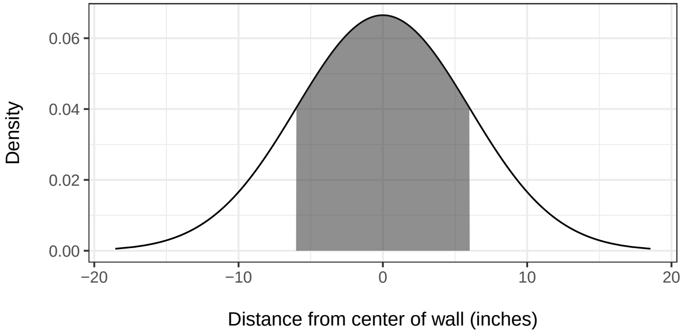
How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



$$
\begin{aligned}
P(X \in (-6, 6)) &= \int_{-6}^{6} f_X(x)dx \\
&= F_X(6) - F_X(-6) \\
&= P(X < 6) - P(X < -6)
\end{aligned}
$$

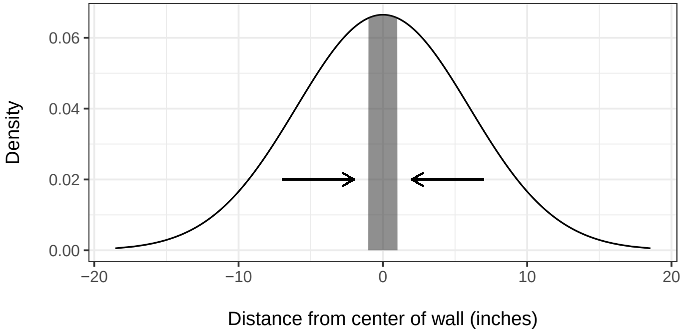
A Numerical Example

How would we calculate the probability that a dart lands within 6 inches of the center of the wall?



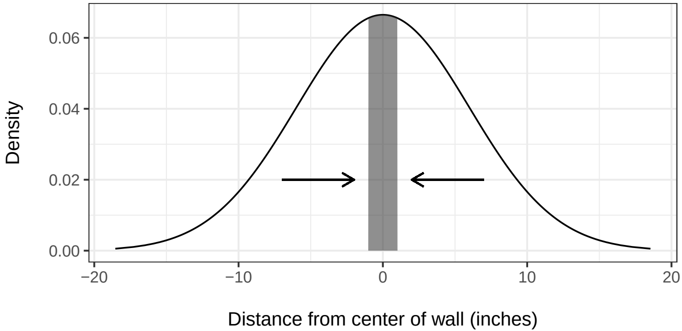
$$
\begin{aligned}
P(X \in (-6, 6)) &= \int_{-6}^{6} f_X(x)dx \\
&= F_X(6) - F_X(-6) \\
&= P(X < 6) - P(X < -6) \\
&= 0.683
\end{aligned}
$$

One inch?



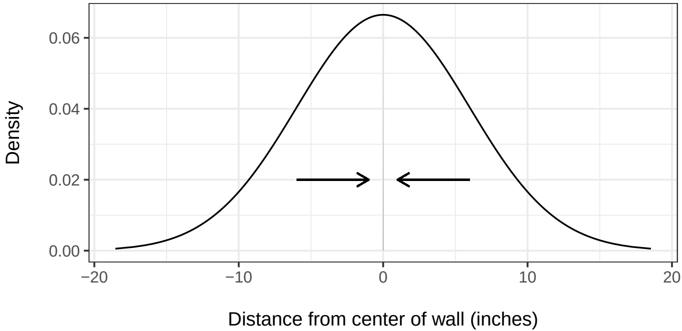
Density
0.06
0.04
0.02
0.00
−20
−10
0
10
20
Distance from center of wall (inches)

One inch?

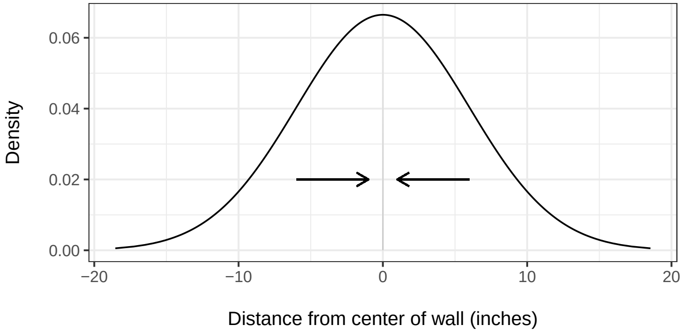


$$P(X \in (-1, 1)) = 0.0664135$$

1/100th of an inch?

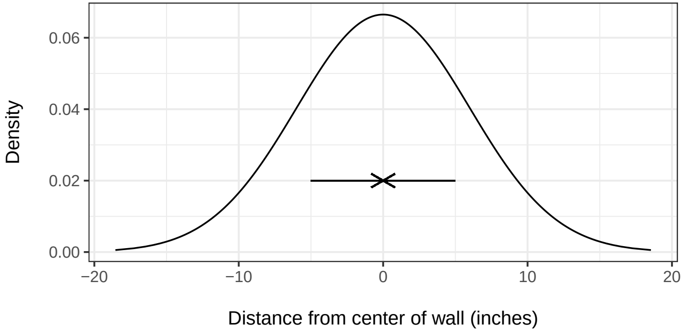


1/100th of an inch?

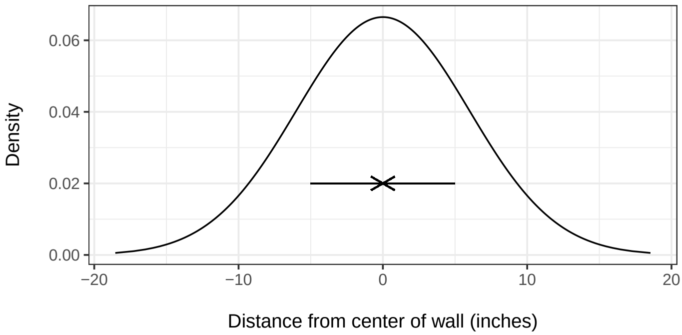


$$P(X \in (-.01, .01)) = 0.0006649037$$

A perfect bullseye?



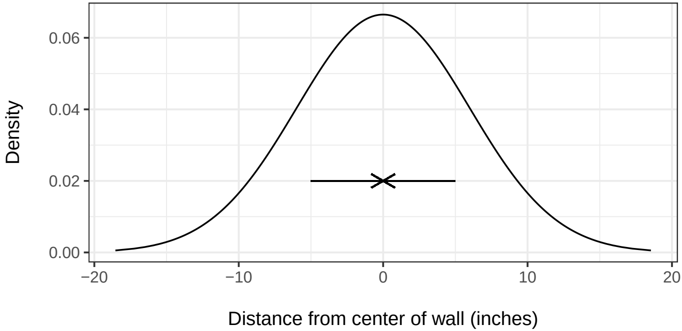
A perfect bullseye?



$$P(X = 0) = 0$$

The probability that a continuous variable takes on a discrete value is 0!

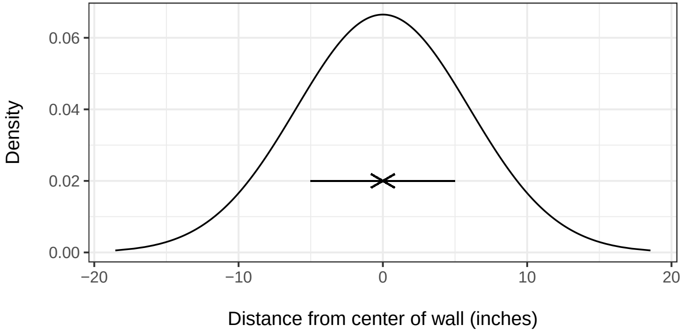
A perfect bullseye?



$$P(X = 0) = 0$$

The probability that a continuous variable takes on a discrete value is 0!
Why?

A perfect bullseye?



$$P(X = 0) = 0$$

The probability that a continuous variable takes on a discrete value is 0!
Why?
Because the width of the range we are calculating is zero, the area is zero.

The Practical Upshot

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.
- The probability mass function (PMF) is for discrete variables and returns a probability.

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.
- The probability mass function (PMF) is for discrete variables and returns a probability.
- The probability density function (PDF) is for continuous variables and provides a probability when integrated over a subset of the support.

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.
- The probability mass function (PMF) is for discrete variables and returns a probability.
- The probability density function (PDF) is for continuous variables and provides a probability when integrated over a subset of the support.
- Reconciling the continuous/discrete divide is the purview of measure theory which is a layer deeper than we are going to go in this class.

The Practical Upshot

- The cumulative distribution function (CDF) has the same interpretation between discrete and continuous.
- The probability mass function (PMF) is for discrete variables and returns a probability.
- The probability density function (PDF) is for continuous variables and provides a probability when integrated over a subset of the support.
- Reconciling the continuous/discrete divide is the purview of measure theory which is a layer deeper than we are going to go in this class.
- As with discrete random variables there are common families of distributions (last video of the week).

We Covered. . .

We Covered. . .

- the definition of a continuous random variable

We Covered. . .

- the definition of a continuous random variable
- probability density functions and their interpretation

We Covered. . .

- the definition of a continuous random variable
- probability density functions and their interpretation
- cumulative distribution functions

We Covered...

- the definition of a continuous random variable
- probability density functions and their interpretation
- cumulative distribution functions

Next time we will describe how to characterize a distribution.

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Characterizing Distributions

Distributions have all kinds of wonky shapes. How do we characterize what they look like?

Expectation

Expectation

The expected value of a random variable X is denoted by $E[X]$ and is a measure of **central tendency** of X. Roughly speaking, an expected value is like a weighted average of all of the values weighted by probability of occurrence.

The expected value of a *discrete* random variable X is defined as

$$E[X] = \sum_{\text{all } x} x \cdot p_X(x).$$

Expectation

The expected value of a random variable X is denoted by $E[X]$ and is a measure of **central tendency** of X. Roughly speaking, an expected value is like a weighted average of all of the values weighted by probability of occurrence.

The expected value of a *discrete* random variable X is defined as

$$E[X] = \sum_{\text{all } x} x \cdot p_X(x).$$

The expected value of a *continuous* random variable X is defined as

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx.$$

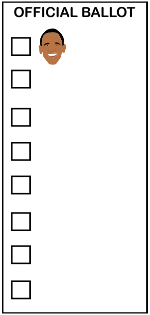
What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$4/26 \quad \times 1$

$+$



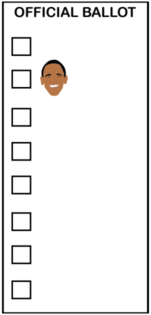
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$4/26 \quad \times 1$$
$$4/26 \quad \times 2$$
$$+$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

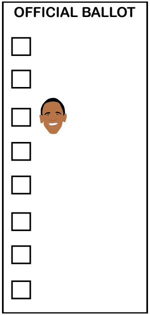
What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$4/26 \quad \times 1$$
$$4/26 \quad \times 2$$
$$2/26 \quad \times 3$$

$$+$$



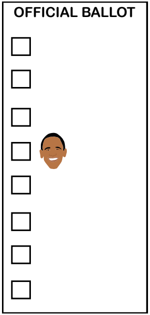
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times 1 \\
4/26 & \times 2 \\
2/26 & \times 3 \\
1/26 & \times 4 \\
 & \\
+ & \\
\hline
\end{array}
$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

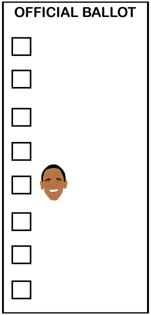
What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$4/26 \quad \times 1$
$4/26 \quad \times 2$
$2/26 \quad \times 3$
$1/26 \quad \times 4$
$1/26 \quad \times 5$

$+$ ____________



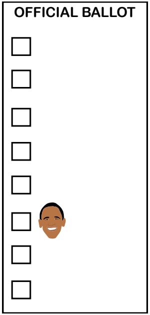
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times 1 \\
4/26 & \times 2 \\
2/26 & \times 3 \\
1/26 & \times 4 \\
1/26 & \times 5 \\
1/26 & \times 6 \\
+ & \\
\hline
\end{array}
$$



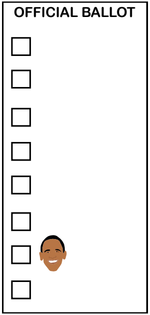
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{rl}
4/26 & \times 1 \\
4/26 & \times 2 \\
2/26 & \times 3 \\
1/26 & \times 4 \\
1/26 & \times 5 \\
1/26 & \times 6 \\
10/26 & \times 7 \\
+ & \\
\hline
\end{array}
$$



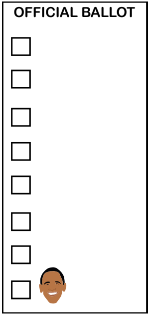
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$\begin{array}{rrr} & 4/26 & \times 1 \\ & 4/26 & \times 2 \\ & 2/26 & \times 3 \\ & 1/26 & \times 4 \\ & 1/26 & \times 5 \\ & 1/26 & \times 6 \\ & 10/26 & \times 7 \\ + & 3/26 & \times 8 \\ \hline \end{array}$$



A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What did we expect for Obama's NH position?

Candidates:

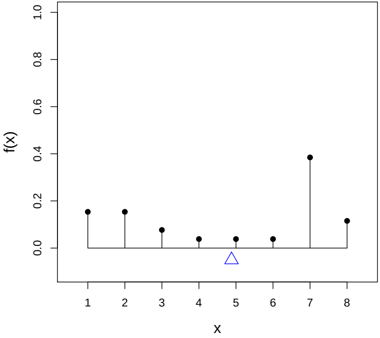
- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$\begin{array}{rrr} & 4/26 & \times 1 \\ & 4/26 & \times 2 \\ & 2/26 & \times 3 \\ & 1/26 & \times 4 \\ & 1/26 & \times 5 \\ & 1/26 & \times 6 \\ & 10/26 & \times 7 \\ + & 3/26 & \times 8 \\ \hline & & 4.88 \end{array}$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

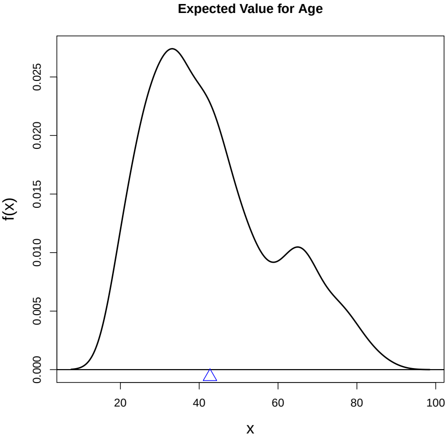
Interpreting Discrete Expected Value

The expected value for a discrete random variable is the balance point of the mass function.



Interpreting Continuous Expected Value

The expected value for a continuous random variable is the balance point of the density function.



Why the Expected Value (Balance Point)?

Why the Expected Value (Balance Point)?

- It is the probabilistic equivalent of the sample average (mean).

Why the Expected Value (Balance Point)?

- It is the probabilistic equivalent of the sample average (mean).
- It is a reasonable measure for the "center" of the data.

Why the Expected Value (Balance Point)?

- It is the probabilistic equivalent of the sample average (mean).
- It is a reasonable measure for the "center" of the data.
- We have some intuition about balance points.

Why the Expected Value (Balance Point)?

- It is the probabilistic equivalent of the sample average (mean).
- It is a reasonable measure for the "center" of the data.
- We have some intuition about balance points.
- It has some useful and convenient properties.

Population Mean as an Expected Value

Population Mean as an Expected Value

Let $x_1, \ldots, x_N$ be our population. Then the population mean is the following

Population Mean as an Expected Value

Let $x_1, \ldots, x_N$ be our population. Then the population mean is the following

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Population Mean as an Expected Value

Let $x_1, \ldots, x_N$ be our population. Then the population mean is the following

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

This can be re-written in the following form:

$$\bar{x} = \sum_{i=1}^{N} \left\{ \frac{1}{N} x_i \right\}$$

Population Mean as an Expected Value

Let $x_1, \ldots, x_N$ be our population. Then the population mean is the following

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

This can be re-written in the following form:

$$\bar{x} = \sum_{i=1}^{N}\left\{\frac{1}{N}x_i\right\}$$

Note how this resembles the definition of discrete expected value. If all values distinct (i.e. $x_i \neq x_j$ for all $i \neq j$).

Population Mean as an Expected Value

Let $x_1, \ldots, x_N$ be our population. Then the population mean is the following

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

This can be re-written in the following form:

$$\bar{x} = \sum_{i=1}^{N} \left\{ \frac{1}{N} x_i \right\}$$

Note how this resembles the definition of discrete expected value. If all values distinct (i.e. $x_i \neq x_j$ for all $i \neq j$).

$$\bar{x} = \sum_{\text{all } x_i} x_i f_X(x_i), \text{ where } f_X(x_i) = \frac{1}{N}$$

Three properties of expectation:

- Additivity
- Homogenity
- LOTUS

Property 1 of Expected Value: Additivity

Property 1 of Expected Value: Additivity

Expectations of sums are sums of expectations.

Property 1 of Expected Value: Additivity

Expectations of sums are sums of expectations.

Suppose we have k random variables $X_1, \ldots, X_k$. If $E[X_i]$ exists for all $i = 1, \ldots, k$, then

Property 1 of Expected Value: Additivity

Expectations of sums are sums of expectations.

Suppose we have k random variables $X_1, \ldots, X_k$. If $E[X_i]$ exists for all $i = 1, \ldots, k$, then

$$E\left[\sum_{i=1}^{k} X_i\right] = E[X_1] + \cdots + E[X_k]$$

Property 1 of Expected Value: Additivity

Expectations of sums are sums of expectations.

Suppose we have k random variables $X_1, \ldots, X_k$. If $E[X_i]$ exists for all $i = 1, \ldots, k$, then

$$E\left[\sum_{i=1}^{k} X_i\right] = E[X_1] + \cdots + E[X_k]$$

Property 2 of Expected Value: Homogeneity

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.
- The expectation of a constant times a random variable is the constant times the expectation of the random variable.

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.
- The expectation of a constant times a random variable is the constant times the expectation of the random variable.

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.
- The expectation of a constant times a random variable is the constant times the expectation of the random variable.

Suppose a and b are constants and X is a random variable. Then

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.
- The expectation of a constant times a random variable is the constant times the expectation of the random variable.

Suppose a and b are constants and X is a random variable. Then

$$E[b] = b$$
$$E[aX] = aE[X]$$
$$E[aX + b] = aE[X] + b$$

Property 2 of Expected Value: Homogeneity

- The expected value of a constant is the constant.
- The expectation of a constant times a random variable is the constant times the expectation of the random variable.

Suppose a and b are constants and X is a random variable. Then

$$E[b] = b$$
$$E[aX] = aE[X]$$
$$E[aX + b] = aE[X] + b$$

Together properties 1 and 2 are linearity (and this is sometimes presented as Linearity of Expectations).

Property 3 of Expected Value: LOTUS

Property 3 of Expected Value: LOTUS

Law of the Unconscious Statistician: If $g(X)$ is a function of a discrete random variable, then

$$E[g(X)] = \sum_x g(x) f_X(x),$$

Property 3 of Expected Value: LOTUS

Law of the Unconscious Statistician: If $g(X)$ is a function of a discrete random variable, then

$$E[g(X)] = \sum_x g(x) f_X(x),$$

essentially the expected value of the transformation of the random variable is just the weighted average of the transformed outcomes.

Property 3 of Expected Value: LOTUS

Law of the Unconscious Statistician: If $g(X)$ is a function of a discrete random variable, then

$$E[g(X)] = \sum_x g(x) f_X(x),$$

essentially the expected value of the transformation of the random variable is just the weighted average of the transformed outcomes.

This means we can calculate the expected value of $g(X)$ without explicitly knowing the distribution of $g(X)$.

Property 3 of Expected Value: LOTUS

Law of the Unconscious Statistician: If $g(X)$ is a function of a discrete random variable, then

$$E[g(X)] = \sum_x g(x) f_X(x),$$

essentially the expected value of the transformation of the random variable is just the weighted average of the transformed outcomes.

This means we can calculate the expected value of $g(X)$ without explicitly knowing the distribution of $g(X)$.
Why the name LOTUS? "because this can be done very easily and mechanically, perhaps in a state of unconsciousness." (Blitzstein and Hwang, Sec 4.5)

Summary of Expected Value Properties

Summary of Expected Value Properties

The three properties:

1) Additivity: expectation of sums are sums of expectations

$$E[X+Y] = E[X] + E[Y]$$

Summary of Expected Value Properties

The three properties:

1) Additivity: expectation of sums are sums of expectations

$$E[X + Y] = E[X] + E[Y]$$

2) Homogeneity: expected value of a constant is the constant

$$E[aX + b] = aE[X] + b$$

Summary of Expected Value Properties

The three properties:

1) Additivity: expectation of sums are sums of expectations

$$E[X + Y] = E[X] + E[Y]$$

2) Homogeneity: expected value of a constant is the constant

$$E[aX + b] = aE[X] + b$$

3) LOTUS: Law of the Unconscious Statistician

$$E[g(X)] = \sum_x g(x) f_X(x)$$

Two common misunderstandings about expected value

Two common misunderstandings about expected value

- $E[g(X)] = g(E[X])$ <u>only if</u> $g(\cdot)$ is a linear function

Two common misunderstandings about expected value

- $E[g(X)] = g(E[X])$ only if $g(\cdot)$ is a linear function
- $E[XY] = E[X]E[Y]$ only if X and Y are independent

Example: Assessing Racial Prejudice

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:
    1. the federal government increasing the tax on gasoline.
    2. professional athletes getting million-dollar salaries.
    3. large corporations polluting the environment.

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:
    1. the federal government increasing the tax on gasoline.
    2. professional athletes getting million-dollar salaries.
    3. large corporations polluting the environment.
  - second half, add a fourth item

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:
    1. the federal government increasing the tax on gasoline.
    2. professional athletes getting million-dollar salaries.
    3. large corporations polluting the environment.
  - second half, add a fourth item
    4. a black family moving in next door

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:
    1. the federal government increasing the tax on gasoline.
    2. professional athletes getting million-dollar salaries.
    3. large corporations polluting the environment.
  - second half, add a fourth item
    4. a black family moving in next door
  - use the answers to infer the proportion upset by the fourth item.

Example: Assessing Racial Prejudice

- We often want to ask sensitive questions which a survey respondent is unlikely to honestly answer
- A list experiment asks respondents how many items on a list they agree with
  - for example, what proportion of people would be upset by a black family moving in next door to them (Kuklinski et al 1997).
  - randomly split survey into two halves
  - first half ask how many of the following items upset you:
    1. the federal government increasing the tax on gasoline.
    2. professional athletes getting million-dollar salaries.
    3. large corporations polluting the environment.
  - second half, add a fourth item
    4. a black family moving in next door
  - use the answers to infer the proportion upset by the fourth item.
- We can use random variables!

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.
- $A = 1$ if angered by a black family moving in next door.
- $A = 0$ if not angered by a black family moving in next door.

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.
- $A = 1$ if angered by a black family moving in next door.
- $A = 0$ if not angered by a black family moving in next door.

$$E[Y] \quad = \quad E[X + A]$$

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.
- $A = 1$ if angered by a black family moving in next door.
- $A = 0$ if not angered by a black family moving in next door.

$$\begin{aligned} E[Y] &= E[X + A] \\ &= E[X] + E[A] \end{aligned}$$

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.
- $A = 1$ if angered by a black family moving in next door.
- $A = 0$ if not angered by a black family moving in next door.

$$\begin{aligned} E[Y] &= E[X+A] \\ &= E[X]+E[A] \\ E[Y]-E[X] &= E[A] \end{aligned}$$

Identifying the Percent Angry

Let's make this mathematical with random variables. This is the first step in defining an estimator and assessing its performance (more next week!).

Assume that $Y = X + A$, where for a randomly sampled respondent,

- $Y =$ the number of angering items on full list.
- $X =$ the number of angering items on baseline list.
- $A = 1$ if angered by a black family moving in next door.
- $A = 0$ if not angered by a black family moving in next door.

$$\begin{aligned} E[Y] &= E[X + A] \\ &= E[X] + E[A] \\ E[Y] - E[X] &= E[A] \end{aligned}$$

So if we know $E[Y]$ and $E[X]$ we can get the expected proportion angered by our item without knowing the individual status of anyone!

Racial Prejudice Example (Kuklinski et al, 1997)

Racial Prejudice Example (Kuklinski et al, 1997)

$X = \#$ of angering items on the baseline list for Southerners:

x	0	1	2	3
$f_X(x)$	?	?	?	?
$\widehat{f}_X(x)$	0.02	0.27	0.43	0.28
$\widehat{F}_X(x)$	0.02	0.29	0.72	1.00

Racial Prejudice Example (Kuklinski et al, 1997)

$X = \#$ of angering items on the baseline list for Southerners:

x	0	1	2	3
$f_X(x)$	?	?	?	?
$\widehat{f}_X(x)$	0.02	0.27	0.43	0.28
$\widehat{F}_X(x)$	0.02	0.29	0.72	1.00

$Y = \#$ of angering items on the treatment list for Southerners:

y	0	1	2	3	4
$f_Y(y)$	?	?	?	?	?
$\widehat{f}_Y(y)$	0.02	0.20	0.40	0.28	0.10
$\widehat{F}_Y(y)$	0.02	0.22	0.62	0.90	1.00

Racial Prejudice Example

$X = \#$ of angering items on the baseline list for Southerners:

x	0	1	2	3	Sum
$\widehat{f}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{f}_X(x)$	0.00	0.27	0.86	0.84	1.97

Racial Prejudice Example

$X = \#$ of angering items on the baseline list for Southerners:

x	0	1	2	3	Sum
$\widehat{f}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{f}_X(x)$	0.00	0.27	0.86	0.84	1.97

$Y = \#$ of angering items on the treatment list for Southerners:

y	0	1	2	3	4	Sum
$\widehat{f}_Y(y)$	0.03	0.20	0.40	0.28	0.10	1.00
$y\widehat{f}_Y(y)$	0.00	0.20	0.80	0.84	0.40	2.24

Racial Prejudice Example

$X = \#$ of angering items on the baseline list for Southerners:

x	0	1	2	3	Sum
$\widehat{f}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{f}_X(x)$	0.00	0.27	0.86	0.84	1.97

$Y = \#$ of angering items on the treatment list for Southerners:

y	0	1	2	3	4	Sum
$\widehat{f}_Y(y)$	0.03	0.20	0.40	0.28	0.10	1.00
$y\widehat{f}_Y(y)$	0.00	0.20	0.80	0.84	0.40	2.24

$$\widehat{E[A]} = 2.24 - 1.97 = 0.27$$

On List Experiments in Research

On List Experiments in Research

When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments

GRAEME BLAIR *University of California, Los Angeles*
ALEXANDER COPPOCK *Yale University*
MARGARET MOOR *Yale University*

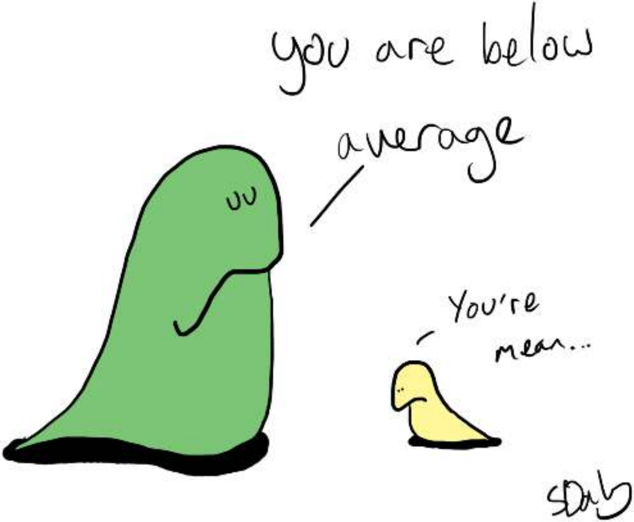
Eliciting honest answers to sensitive questions is frustrated if subjects withhold the truth for fear that others will judge or punish them. The resulting bias is commonly referred to as social desirability bias, a subset of what we label sensitivity bias. We make three contributions. First, we propose a social reference theory of sensitivity bias to structure expectations about survey responses on sensitive topics. Second, we explore the bias-variance trade-off inherent in the choice between direct and indirect measurement technologies. Third, to estimate the extent of sensitivity bias, we meta-analyze the set of published and unpublished list experiments (a.k.a., the item count technique) conducted to date and compare the results with direct questions. We find that sensitivity biases are typically smaller than 10 percentage points and in some domains are approximately zero.

Fun with

Fun with Averages

Fun with Averages

Central Tendency



The Story of Averages

Measurements

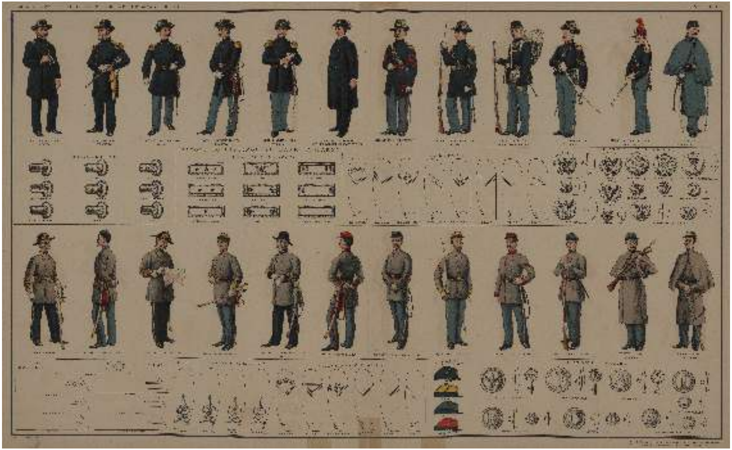
MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE D'OBSERVATIONS calculé.
Pouces.							
33	3	5	0,5000			0,5000	7
34	18	31	0,4995	52	50	0,4993	29
35	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4823	33,5	34,5	0,4854	323
37	420	732	0,4501	26,0	26,5	0,4531	732
38	749	1305	0,3769	18,0	18,5	0,3799	1333
39	1073	1867	0,2464	10,5	10,5	0,2466	1838
			0,0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	934	1628	0,2913	13	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000

Social Physics

Social Physics

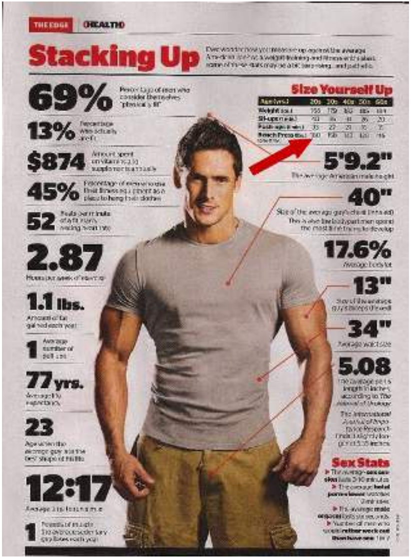
The determination of the average man is not merely a matter of speculative curiosity; it may be of the most important service to the science of man and the social system. It ought necessarily to precede every other inquiry into social physics, since it is, as it were, the basis. The average man, indeed, is in a nation what the centre of gravity is in a body; it is by having that central point in view that we arrive at the apprehension of all the phenomena of equilibrium and motion
- Quetelet

The Military Takes to the Idea



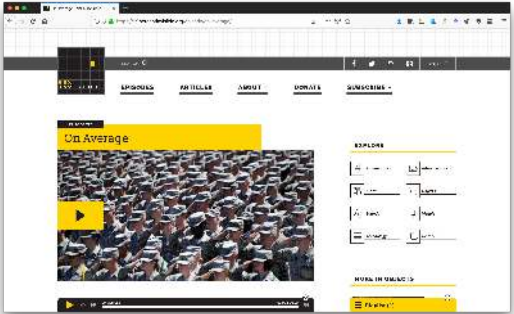
The Problem with Averages

The Average Man



The Face of the Average Man

On averages



https://99percentinvisible.org/episode/on-average/

We Covered. . .

We Covered...

- Expectations (definitions, properties etc.)

We Covered...

- Expectations (definitions, properties etc.)
- A short history of the average

We Covered...

- Expectations (definitions, properties etc.)
- A short history of the average

Next time: variance as a measure of a distribution's dispersion!

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Variance: A Measure of Dispersion

Variance: A Measure of Dispersion

Expectation told us about the central tendency of a random variable, but what about dispersion?

Variance: A Measure of Dispersion

Expectation told us about the central tendency of a random variable, but what about dispersion?

The expected value of a function $g()$ of the random variable X is denoted by $E[g(X)]$ and measures the central tendency of $g(X)$.

Variance: A Measure of Dispersion

Expectation told us about the central tendency of a random variable, but what about dispersion?
The expected value of a function $g()$ of the random variable X is denoted by $E[g(X)]$ and measures the central tendency of $g(X)$.

The variance is a special case of this, and the variance of a random variable X (a measure of its dispersion) is given by

$$\begin{aligned} V[X] &= E[(X - E[X])^2] \\ &= E[X^2] - E[X]^2 \end{aligned}$$

It is the expectation of the squared distances from the mean.

For a discrete random variable X

$$V[X] = \sum_{\text{all } x} (x - E[X])^2 p_X(x)$$

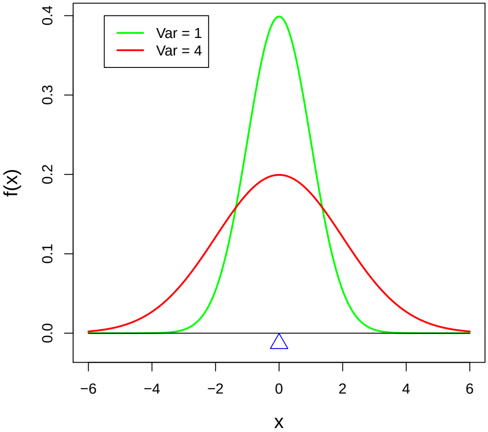
For a discrete random variable X

$$V[X] = \sum_{\text{all } x} (x - E[X])^2 p_X(x)$$

For a continuous random variable X

$$V[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x) dx$$

Variance Measures the Spread of a Distribution



Why the Variance?

Why the Variance?

- It is a reasonable measure for the "spread" of a distribution.

Why the Variance?

- It is a reasonable measure for the "spread" of a distribution.
- The Normal distribution—more later this week—is completely determined by its expected value (location) and variance (spread).

Why the Variance?

- It is a reasonable measure for the “spread” of a distribution.
- The Normal distribution—more later this week—is completely determined by its expected value (location) and variance (spread).
- The square root of the variance is the standard deviation.

Why the Variance?

- It is a reasonable measure for the “spread” of a distribution.
- The Normal distribution—more later this week—is completely determined by its expected value (location) and variance (spread).
- The square root of the variance is the standard deviation.
- The variance and standard deviation have some useful properties.

Property 1 of Variance: Behavior with Constants

Suppose a and b are constants and X is a random variable. Then

Property 1 of Variance: Behavior with Constants

Suppose a and b are constants and X is a random variable. Then

- The variance of a constant is zero.

Property 1 of Variance: Behavior with Constants

Suppose a and b are constants and X is a random variable. Then

- The variance of a constant is zero.
- The variance of a constant times a random variable is the constant squared times the variance of the random variable.

Property 1 of Variance: Behavior with Constants

Suppose a and b are constants and X is a random variable. Then

- The variance of a constant is zero.
- The variance of a constant times a random variable is the constant squared times the variance of the random variable.

$$V[b] = 0$$
$$V[aX] = a^2 V[X]$$
$$V[aX + b] = a^2 V[X] + 0$$

Property 2 of Variance: Additivity for Independent Random Variables

Property 2 of Variance: Additivity for Independent Random Variables

Variances of sums of independent RVs are sums of variances.

Property 2 of Variance: Additivity for Independent Random Variables

Variances of sums of independent RVs are sums of variances.

Suppose we have k independent random variables $X_1, \ldots, X_k$. If $V[X_i]$ exists for all $i = 1, \ldots, k$, then

$$V\left[\sum_{i=1}^{k} X_i\right] = V[X_1] + \cdots + V[X_k]$$

NB: Technically independence is sufficient but not necessary.

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$4/26 \quad \times (1 - 4.88)^2$

$+$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times (1 - 4.88)^2 \\
4/26 & \times (2 - 4.88)^2 \\
+ & \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times (1 - 4.88)^2 \\
4/26 & \times (2 - 4.88)^2 \\
2/26 & \times (3 - 4.88)^2 \\
 & \\
+ & \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times (1 - 4.88)^2 \\
4/26 & \times (2 - 4.88)^2 \\
2/26 & \times (3 - 4.88)^2 \\
1/26 & \times (4 - 4.88)^2 \\
 & \\
+ & \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{ll}
4/26 & \times (1-4.88)^2 \\
4/26 & \times (2-4.88)^2 \\
2/26 & \times (3-4.88)^2 \\
1/26 & \times (4-4.88)^2 \\
1/26 & \times (5-4.88)^2 \\
 & \\
+ & \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$\begin{array}{ll} 4/26 & \times (1 - 4.88)^2 \\ 4/26 & \times (2 - 4.88)^2 \\ 2/26 & \times (3 - 4.88)^2 \\ 1/26 & \times (4 - 4.88)^2 \\ 1/26 & \times (5 - 4.88)^2 \\ 1/26 & \times (6 - 4.88)^2 \\ & \\ + & \\ \hline \end{array}$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{rl}
4/26 & \times (1-4.88)^2 \\
4/26 & \times (2-4.88)^2 \\
2/26 & \times (3-4.88)^2 \\
1/26 & \times (4-4.88)^2 \\
1/26 & \times (5-4.88)^2 \\
1/26 & \times (6-4.88)^2 \\
10/26 & \times (7-4.88)^2 \\
+ & \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{rrl}
 & 4/26 & \times (1-4.88)^2 \\
 & 4/26 & \times (2-4.88)^2 \\
 & 2/26 & \times (3-4.88)^2 \\
 & 1/26 & \times (4-4.88)^2 \\
 & 1/26 & \times (5-4.88)^2 \\
 & 1/26 & \times (6-4.88)^2 \\
 & 10/26 & \times (7-4.88)^2 \\
+ & 3/26 & \times (8-4.88)^2 \\
\hline
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$
\begin{array}{rrl}
 & 4/26 & \times (1-4.88)^2 \\
 & 4/26 & \times (2-4.88)^2 \\
 & 2/26 & \times (3-4.88)^2 \\
 & 1/26 & \times (4-4.88)^2 \\
 & 1/26 & \times (5-4.88)^2 \\
 & 1/26 & \times (6-4.88)^2 \\
 & 10/26 & \times (7-4.88)^2 \\
+ & 3/26 & \times (8-4.88)^2 \\
\hline
 & & 2.93
\end{array}
$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

What was the variance of Obama's NH position?

Candidates:

- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$$\begin{array}{rcl} & 4/26 & \times (1 - 4.88)^2 \\ & 4/26 & \times (2 - 4.88)^2 \\ & 2/26 & \times (3 - 4.88)^2 \\ & 1/26 & \times (4 - 4.88)^2 \\ & 1/26 & \times (5 - 4.88)^2 \\ & 1/26 & \times (6 - 4.88)^2 \\ & 10/26 & \times (7 - 4.88)^2 \\ + & 3/26 & \times (8 - 4.88)^2 \\ \hline & & 2.93 \end{array}$$

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Does variance matter for fairness?

One Step Deeper: Moments

One Step Deeper: Moments

Definition

Suppose X is a random variable with pdf f_X. Define,

$$E[X^n] \;=\; \int_{-\infty}^{\infty} x^n f_X(x)dx$$

We will call X^n the n^{th} moment of X

One Step Deeper: Moments

Definition

Suppose X is a random variable with pdf f_X. Define,

$$E[X^n] \;=\; \int_{-\infty}^{\infty} x^n f_X(x)dx$$

We will call X^n the n^{th} moment of X

- $V(X) = \text{Second Moment} - \text{First Moment}^2$

One Step Deeper: Moments

Definition

Suppose X is a random variable with pdf f_X. Define,

$$E[X^n] \ = \ \int_{-\infty}^{\infty} x^n f_X(x) dx$$

We will call X^n the n^{th} moment of X

- $V(X) = \text{Second Moment} - \text{First Moment}^2$
- We are assuming that the integral converges.

One Step Deeper: Moments

Definition

Suppose X is a random variable with pdf f_X. Define,

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

We will call X^n the n^{th} moment of X

- $V(X) =$ Second Moment $-$ First Moment^2
- We are assuming that the integral converges.
- Another way to characterize distributions is with their moment-generating function.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.
Suppose we want to pick a single number (c) that summarizes a random variable X. What we mean by summarizes determines the best choice of c.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.
Suppose we want to pick a single number (c) that summarizes a random variable X. What we mean by summarizes determines the best choice of c.

Generally speaking we want a summary that is in the "center" of the data, i.e. that is as close as possible to all possible datapoints. Again though, the choice turns on what we mean by close.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.

Suppose we want to pick a single number (c) that summarizes a random variable X. What we mean by summarizes determines the best choice of c.

Generally speaking we want a summary that is in the "center" of the data, i.e. that is as close as possible to all possible datapoints. Again though, the choice turns on what we mean by close.

Two notions of closeness:

- Mean Squared Error: $E\left[(X-c)^2\right]$
  This leads to choosing the mean of X.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.
Suppose we want to pick a single number (c) that summarizes a random variable X. What we mean by summarizes determines the best choice of c.

Generally speaking we want a summary that is in the "center" of the data, i.e. that is as close as possible to all possible datapoints. Again though, the choice turns on what we mean by close.
Two notions of closeness:

- Mean Squared Error: $E\left[(X-c)^2\right]$
  This leads to choosing the mean of X.
- Mean Absolute Error: $E[|X-c|]$
  This leads to choosing the median of X.

Expected Value as Mean Squared Error Minimizer

Now we can return to the question of why expectation? and offer one technical answer.
Suppose we want to pick a single number (c) that summarizes a random variable X. What we mean by summarizes determines the best choice of c.

Generally speaking we want a summary that is in the "center" of the data, i.e. that is as close as possible to all possible datapoints. Again though, the choice turns on what we mean by close.
Two notions of closeness:

- Mean Squared Error: $E\left[(X-c)^2\right]$
  This leads to choosing the mean of X.
- Mean Absolute Error: $E[|X-c|]$
  This leads to choosing the median of X.

Let's prove the first result (see Blitzstein and Hwang 2014 Theorem 6.1.4 on pg 245 for this proof and the proof on mean absolute error).

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X - c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \quad (1)$$

We choose c to minimize this term.

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$.

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$.

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Now to prove the identity:

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Now to prove the identity:

$$V[X] = V[X - c] \qquad \text{(Prop 1 of Variance)}$$

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Now to prove the identity:

$$\begin{aligned} V[X] &= V[X-c] && \text{(Prop 1 of Variance)} \\ &= E\left[(X-c)^2\right] - (E[X-c])^2 && \text{(Defn of Variance)} \end{aligned}$$

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Now to prove the identity:

$$\begin{aligned} V[X] &= V[X-c] && \text{(Prop 1 of Variance)} \\ &= E\left[(X-c)^2\right] - (E[X-c])^2 && \text{(Defn of Variance)} \\ &= E\left[(X-c)^2\right] - (\mu - c)^2 && \text{(Linearity of Exp)} \end{aligned}$$

Proof of Mean as Mean Squared Error Minimizer

Let X be a random variable and $E[X] = \mu$. We want to show that the value of c that minimizes the mean squared error $E\left[(X-c)^2\right]$ is the mean, μ (Blitzstein and Hwang Theorem 6.1.4).

We will prove the following identity below:

$$E\left[(X-c)^2\right] = V[X] + (\mu - c)^2 \tag{1}$$

We choose c to minimize this term. The choice cannot affect $V[X]$. Setting $c = \mu$ sets $(\mu - c)^2 = 0$ and any other choice makes $(\mu - c)^2 > 0$. Therefore (assuming the identity holds), $c = \mu$ minimizes Eq 1.

Now to prove the identity:

$$\begin{aligned} V[X] &= V[X-c] && \text{(Prop 1 of Variance)} \\ &= E\left[(X-c)^2\right] - (E[X-c])^2 && \text{(Defn of Variance)} \\ &= E\left[(X-c)^2\right] - (\mu - c)^2 && \text{(Linearity of Exp)} \\ V[X] + (\mu - c)^2 &= E\left[(X-c)^2\right] \end{aligned}$$

We Covered. . .

We Covered…

- Variance (definitions, properties etc.)

We Covered...

- Variance (definitions, properties etc.)
- A tiny preview of moments

We Covered. . .

- Variance (definitions, properties etc.)
- A tiny preview of moments
- Motivation of the expectation as a minimizer of mean squared error.

We Covered...

- Variance (definitions, properties etc.)
- A tiny preview of moments
- Motivation of the expectation as a minimizer of mean squared error.

Next time: Joint Distributions!

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Joint Distributions

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$
- We also talked about the conditional probability of A given that B occurred.

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$
- We also talked about the conditional probability of A given that B occurred.
- We also need to think about more than one random variable at the same time.

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$
- We also talked about the conditional probability of A given that B occurred.
- We also need to think about more than one random variable at the same time.
- The joint distribution of two (or more) variables describes the pairs of observations that we are more or less likely to see.

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$
- We also talked about the conditional probability of A given that B occurred.
- We also need to think about more than one random variable at the same time.
- The joint distribution of two (or more) variables describes the pairs of observations that we are more or less likely to see.
- The conditional distribution describes one random variable given knowledge of another.

Joint Distributions

- We've talked about joint probabilities of events—what was the probability of A and B occurring: $P(A \cap B)$
- We also talked about the conditional probability of A given that B occurred.
- We also need to think about more than one random variable at the same time.
- The joint distribution of two (or more) variables describes the pairs of observations that we are more or less likely to see.
- The conditional distribution describes one random variable given knowledge of another.
- We will start with a visual preview, then step back to go through the math more concretely.

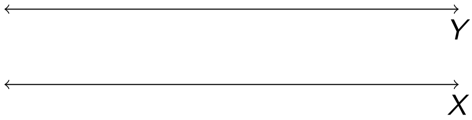
Understanding Joint Distributions Mathematically

Understanding Joint Distributions Mathematically

- Consider two random variables now, X and Y, each on the real line, $\mathbb{R}$.

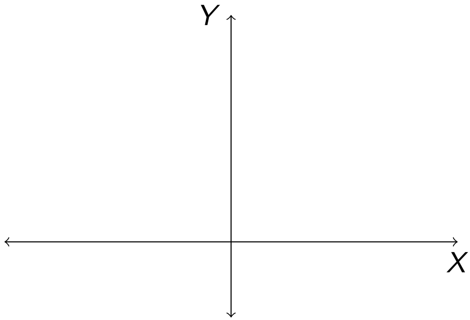
Understanding Joint Distributions Mathematically

- Consider two random variables now, X and Y, each on the real line, $\mathbb{R}$.



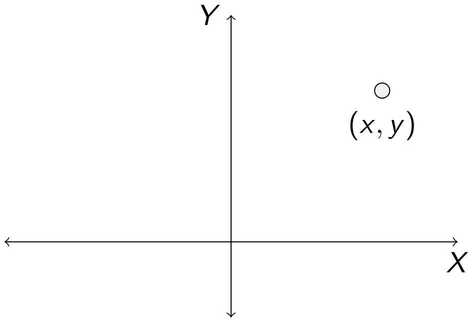
Understanding Joint Distributions Mathematically

- Consider two random variables now, X and Y, each on the real line, $\mathbb{R}$.
- The pair form a two-dimensional space, or $\mathbb{R} \times \mathbb{R}$



Understanding Joint Distributions Mathematically

- Consider two random variables now, X and Y, each on the real line, $\mathbb{R}$.
- The pair form a two-dimensional space, or $\mathbb{R} \times \mathbb{R}$
- One realization of the random variable is a point in that space



Example: Racial Prejudice

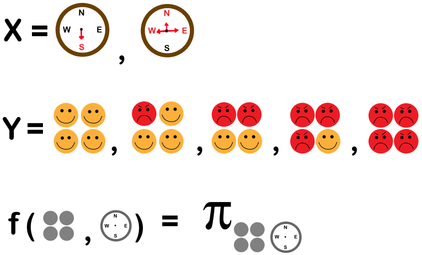
- Recall the list experiment about racial prejudice. Suppose we define $X = 0$ (Non-southern), 1 (Southern) and $Y =$ "number of angering items" for a randomly selected respondent receiving the treatment list.

Example: Racial Prejudice

- Recall the list experiment about racial prejudice. Suppose we define $X = 0$ (Non-southern), 1 (Southern) and $Y =$ "number of angering items" for a randomly selected respondent receiving the treatment list.
- Furthermore, we define the probability that this respondent will have the values $X = x$ and $Y = y$ to be $p_{Y,X}(y,x) = \pi_{yx}$

Example: Racial Prejudice

- Recall the list experiment about racial prejudice. Suppose we define $X = 0$ (Non-southern), 1 (Southern) and $Y =$ "number of angering items" for a randomly selected respondent receiving the treatment list.
- Furthermore, we define the probability that this respondent will have the values $X = x$ and $Y = y$ to be $p_{Y,X}(y, x) = \pi_{yx}$



Example Joint Distribution: Binary X, Discrete Y

Example Joint Distribution: Binary X, Discrete Y

Although we cannot observe the responses for the entire population, we can imagine what they might look like as a joint distribution.

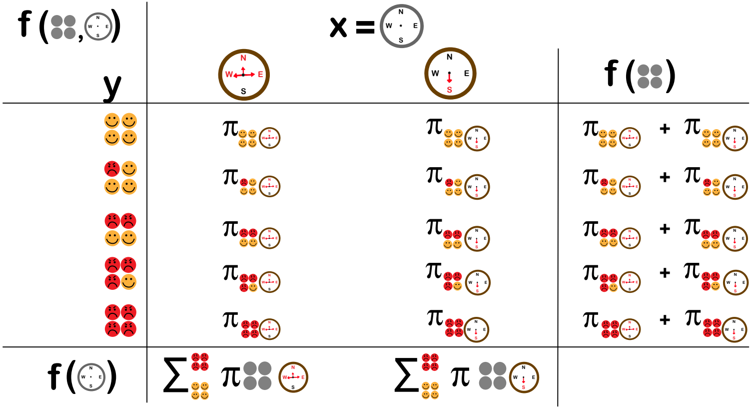
Example Joint Distribution: Binary X, Discrete Y

Although we cannot observe the responses for the entire population, we can imagine what they might look like as a joint distribution.

$f_{X,Y}(x,y)$	x		
y	0	1	$f_Y(y)$
0	π_{00}	π_{01}	$\pi_{00}+\pi_{01}$
1	π_{10}	π_{11}	$\pi_{00}+\pi_{01}$
2	π_{20}	π_{21}	$\pi_{00}+\pi_{01}$
3	π_{30}	π_{31}	$\pi_{00}+\pi_{01}$
4	π_{40}	π_{41}	$\pi_{00}+\pi_{01}$
$f_X(x)$	$\sum_{y=0}^{4}\pi_{y0}$	$\sum_{y=0}^{4}\pi_{y1}$	

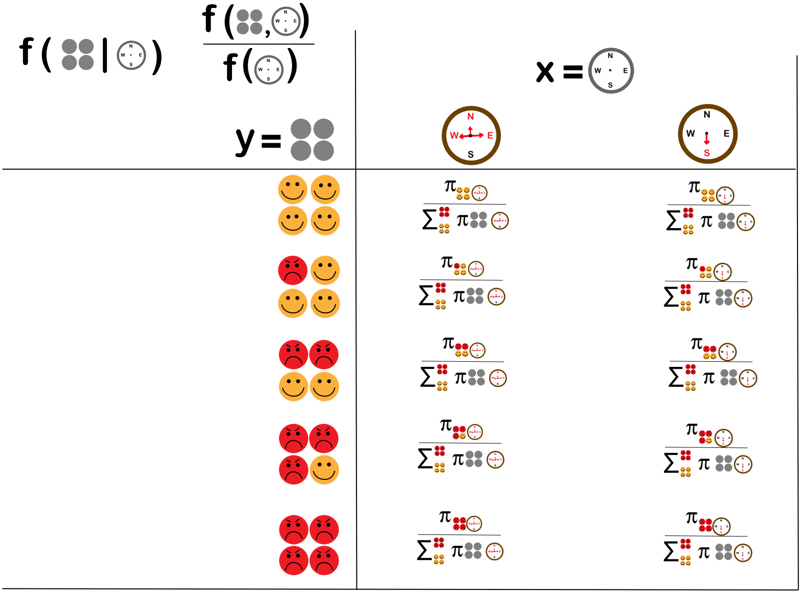
Example Joint Distribution: Binary X, Discrete Y

Although we cannot observe the responses for the entire population, we can imagine what they might look like as a joint distribution.

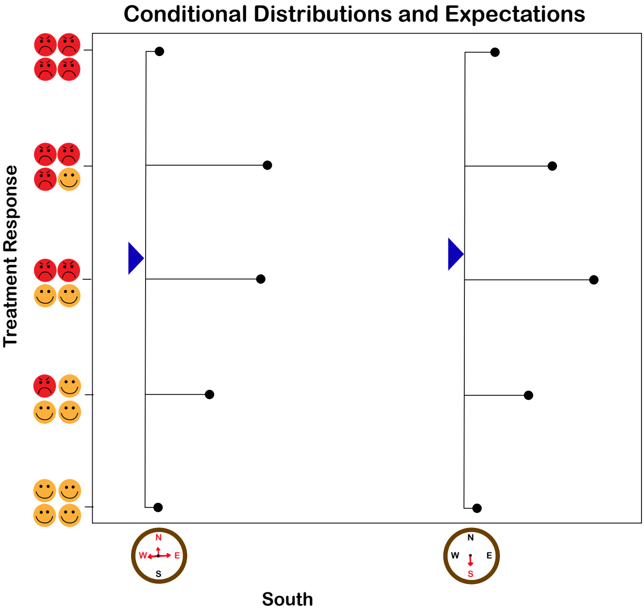


Discrete Conditional Distribution

Discrete Conditional Distribution

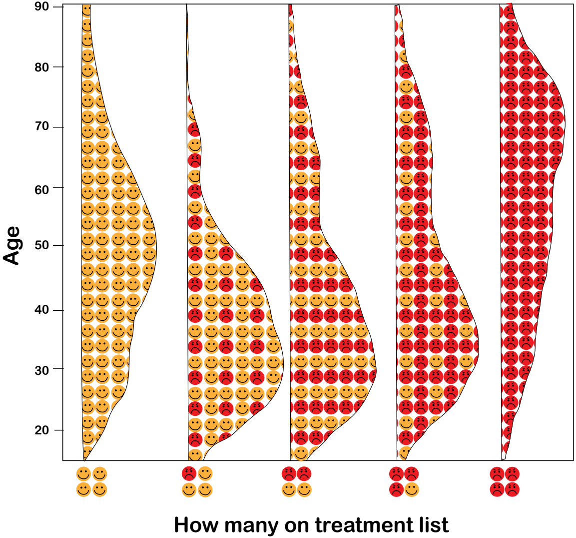


Discrete Conditional Distribution

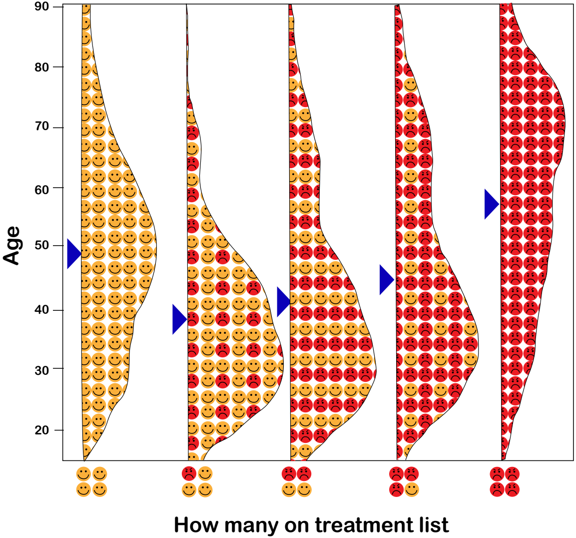


Example: Continuous Conditional Distribution

Example: Continuous Conditional Distribution



Conditional Expectation Function—next time!



Joint Probability Mass Function

Definition

For two discrete random variables X and Y the joint Probability Mass Function (PMF) $P_{X,Y}(x, y)$ gives the probability that $X = x$ and $Y = y$ for all x and y:

$$p_{X,Y}(x, y) \;=\; P(X = x \text{ and } Y = y)$$

Restrictions:

Joint Probability Mass Function

Definition

For two discrete random variables X and Y the joint Probability Mass Function (PMF) $P_{X,Y}(x, y)$ gives the probability that $X = x$ and $Y = y$ for all x and y:

$$p_{X,Y}(x, y) = P(X = x \text{ and } Y = y)$$

Restrictions:

- $p_{X,Y}(x, y) \geq 0$ and $\sum_x \sum_y p_{X,Y}(x, y) = 1$.

Joint Probability Mass Function

Definition

For two discrete random variables X and Y the joint Probability Mass Function (PMF) $p_{X,Y}(x,y)$ gives the probability that $X = x$ and $Y = y$ for all x and y:

$$p_{X,Y}(x,y) \; = \; P(X = x \text{ and } Y = y)$$

Should the U.S. allow more immigrants to come and live here?

		X: Education			
		less HS	HS	College	BA
Y: Support	oppose	0.07	0.22	0.18	0.15
	neutral	0.02	0.06	0.05	0.05
	favor	0.01	0.03	0.04	0.11

Joint Probability Mass Function

Definition

For two discrete random variables X and Y the joint Probability Mass Function (PMF) $p_{X,Y}(x,y)$ gives the probability that $X = x$ and $Y = y$ for all x and y:

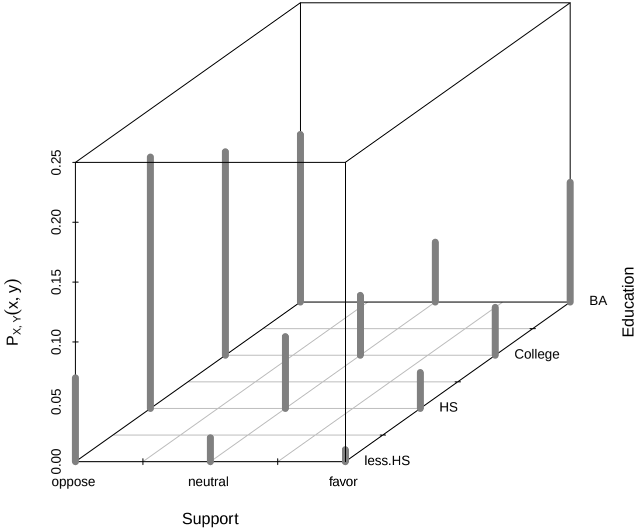
$$p_{X,Y}(x,y) \; = \; P(X = x \text{ and } Y = y)$$

Should the U.S. allow more immigrants to come and live here?

		X: Education			
		less HS	HS	College	BA
Y: Support	oppose	0.07	0.22	0.18	0.15
	neutral	0.02	0.06	0.05	0.05
	favor	0.01	0.03	0.04	0.11

With discrete random variables this is very similar to thinking about a cross-tab, with frequencies/ probabilities in the cells instead of raw numbers.

Joint Probability Mass Function



From Joint to Marginal PMF

Given the joint PMF $p_{X,Y}(x, y)$ can we recover the marginal PMF $p_Y(y)$ (distribution over a single variable)?

		X: Education				
		less HS	HS	College	BA	
Y: Support	oppose	0.07	0.21	0.17	0.14	
	neutral	0.02	0.06	0.05	0.05	
	favor	0.01	0.03	0.04	0.10	

From Joint to Marginal PMF

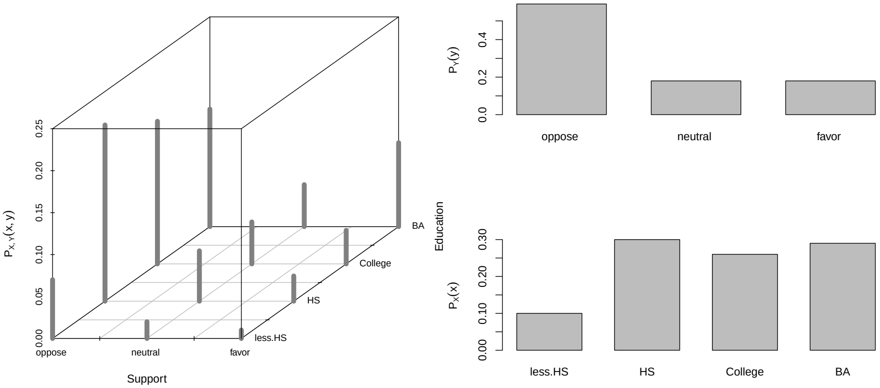
Given the joint PMF $p_{X,Y}(x,y)$ can we recover the marginal PMF $p_Y(y)$ (distribution over a single variable)?

		X: Education				
		less HS	HS	College	BA	$p_Y(y)$
	oppose	0.07	0.21	0.17	0.14	0.62
Y: Support	neutral	0.02	0.06	0.05	0.05	0.19
	favor	0.01	0.03	0.04	0.10	0.19

To obtain $p_Y(y)$ we marginalize the joint probability function $p_{X,Y}(x,y)$ over X:

$$p_Y(y) = \sum_x p_{X,Y}(x,y) = \sum_x P(X=x, Y=y)$$

Joint and Marginal Probability Mass Functions



Why Does Marginalization Work?

Why Does Marginalization Work?

Begin with discrete case.

Why Does Marginalization Work?

Begin with discrete case. Consider jointly distributed discrete random variables, X and Y. We'll suppose they have joint pmf,

$$P(X = x, Y = y) = p_{X,Y}(x, y)$$

Why Does Marginalization Work?

Begin with discrete case. Consider jointly distributed discrete random variables, X and Y. We'll suppose they have joint pmf,

$$P(X = x, Y = y) = p_{X,Y}(x, y)$$

Suppose that the distribution allocates its mass at $x_1, x_2, \ldots, x_M$ and $y_1, y_2, \ldots, y_N$.

Why Does Marginalization Work?

Begin with discrete case. Consider jointly distributed discrete random variables, X and Y. We'll suppose they have joint pmf,

$$P(X = x, Y = y) = p_{X,Y}(x, y)$$

Suppose that the distribution allocates its mass at $x_1, x_2, \ldots, x_M$ and $y_1, y_2, \ldots, y_N$.

Define the conditional mass function $P(X = x|Y = y)$ as,

$$\begin{aligned} P(X = x|Y = y) &\equiv p_{X|Y}(x|y) \\ &= p_{X,Y}(x, y)/p_Y(y) \end{aligned}$$

Why Does Marginalization Work?

Begin with discrete case. Consider jointly distributed discrete random variables, X and Y. We'll suppose they have joint pmf,

$$P(X = x, Y = y) = p_{X,Y}(x, y)$$

Suppose that the distribution allocates its mass at $x_1, x_2, \ldots, x_M$ and $y_1, y_2, \ldots, y_N$.

Define the conditional mass function $P(X = x|Y = y)$ as,

$$\begin{aligned} P(X = x|Y = y) &\equiv p_{X|Y}(x|y) \\ &= p_{X,Y}(x, y)/p_Y(y) \end{aligned}$$

Then it follows that:

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$$

Why Does Marginalization Work?

Begin with discrete case. Consider jointly distributed discrete random variables, X and Y. We'll suppose they have joint pmf,

$$P(X = x, Y = y) = p_{X,Y}(x, y)$$

Suppose that the distribution allocates its mass at $x_1, x_2, \ldots, x_M$ and $y_1, y_2, \ldots, y_N$.

Define the conditional mass function $P(X = x|Y = y)$ as,

$$\begin{aligned} P(X = x|Y = y) &\equiv p_{X|Y}(x|y) \\ &= p_{X,Y}(x, y)/p_Y(y) \end{aligned}$$

Then it follows that:

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$$

Marginalizing over y to get $p_X(x)$ is then,

$$p_X(x_j) = \sum_{i=1}^{N} p_{X|Y}(x_j|y_i)p_Y(y_i)$$

A Table

	$Y = 0$	$Y = 1$	
$X = 0$	$p(0,0)$	$p(0, 1)$	$p_X(0)$
$X = 1$	$p(1,0)$	$p(1,1)$	$p_X(1)$
	$p_Y\ (0)$	$p_Y\ (1)$	

A Table

	Y = 0	Y = 1	
X = 0	0.01	0.05	?
X = 1	0.25	0.69	?
	0.26	0.74	

$$
\begin{aligned}
p_X(0) &= P(X=0|Y=0)P(Y=0) + P(X=0|Y=1)P(Y=1) \\
&= \frac{0.01}{0.26} \times 0.26 + \frac{0.05}{0.74} \times 0.74 \\
&= 0.06
\end{aligned}
$$

A Table

	Y = 0	Y = 1	
X = 0	0.01	0.05	?
X = 1	0.25	0.69	?
	0.26	0.74	

$$\begin{aligned} p_X(0) &= P(X=0|Y=0)P(Y=0) + P(X=0|Y=1)P(Y=1) \\ &= \frac{0.01}{0.26} \times 0.26 + \frac{0.05}{0.74} \times 0.74 \\ &= 0.06 \end{aligned}$$

$$\begin{aligned} p_X(1) &= P(X=1|Y=0)P(Y=0) + P(X=1|Y=1)P(Y=1) \\ &= \frac{0.25}{0.26} \times 0.26 + \frac{0.69}{0.74} \times 0.74 \\ &= 0.94 \end{aligned}$$

Conditional PMF

Conditional PMF

Definition

The conditional PMF of Y given X, $p_{Y|X}(y|x)$, is the PMF of Y when X is known to be at a particular value $X = x$:

$$p_{Y|X}(y|x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

Conditional PMF

Definition

The conditional PMF of Y given X, $p_{Y|X}(y|x)$, is the PMF of Y when X is known to be at a particular value $X = x$:

$$p_{Y|X}(y|x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

Key relationships:

- $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$ (multiplicative rule)
- $p_{Y|X}(y|x) = p_{X|Y}(x|y)p_Y(y)/p_X(x)$ (Bayes' rule)

Conditional PMF

Definition

The conditional PMF of Y given X, $p_{Y|X}(y|x)$, is the PMF of Y when X is known to be at a particular value $X = x$:

$$p_{Y|X}(y|x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

Key relationships:

- $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$ (multiplicative rule)
- $p_{Y|X}(y|x) = p_{X|Y}(x|y)p_Y(y)/p_X(x)$ (Bayes' rule)

Conditional PMFs are just like ordinary PMFs, but refer to a universe where the "conditioning event" $(X = x)$ is known to have occurred.

Conditional PMF

Definition

The conditional PMF of Y given X, $p_{Y|X}(y|x)$, is the PMF of Y when X is known to be at a particular value $X = x$:

$$p_{Y|X}(y|x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Key relationships:

- $p_{X,Y}(x, y) = p_{Y|X}(y|x)p_X(x)$ (multiplicative rule)
- $p_{Y|X}(y|x) = p_{X|Y}(x|y)p_Y(y)/p_X(x)$ (Bayes' rule)

Conditional PMFs are just like ordinary PMFs, but refer to a universe where the "conditioning event" $(X = x)$ is known to have occurred.

Conditional distributions are key in statistical modeling because they inform us how the distribution of Y varies across different levels of X.

From Joint to Conditional: $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$

Table: Joint PMF $p_{X,Y}(x,y)$ and Marginal PMFs $p_X(x), p_Y(y)$

		Education				
	$p_{X,Y}(x,y)$	less HS	HS	College	BA	$p_Y(y)$
Support	oppose	0.07	0.22	0.18	0.15	0.62
	neutral	0.02	0.06	0.05	0.05	0.19
	favor	0.01	0.03	0.04	0.11	0.19
	$p_X(x)$	0.11	0.32	0.27	0.31	1.00

From Joint to Conditional: $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$

Table: Joint PMF $p_{X,Y}(x,y)$ and Marginal PMFs $p_X(x), p_Y(y)$

		Education				
	$p_{X,Y}(x,y)$	less HS	HS	College	BA	$p_Y(y)$
Support	oppose	0.07	0.22	0.18	0.15	0.62
	neutral	0.02	0.06	0.05	0.05	0.19
	favor	0.01	0.03	0.04	0.11	0.19
	$p_X(x)$	0.11	0.32	0.27	0.31	1.00

Table: Conditional PMF $p_{Y|X}(y|x)$

		Education				
	$p_{Y|X}(y|x)$	less HS	HS	College	BA	
Support	oppose	0.70	0.70	0.65	0.48	0.62
	neutral	0.20	0.20	0.19	0.17	0.19
	favor	0.10	0.10	0.15	0.34	0.19
		1.00	1.00	1.00	1.00	1.00

Joint and Conditional Probability Mass Functions

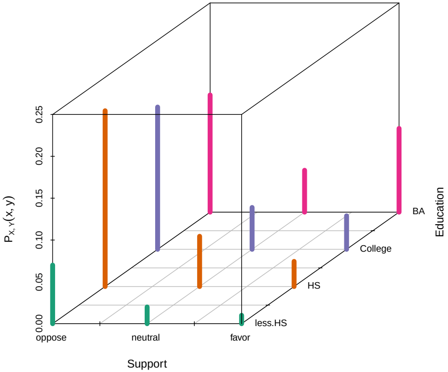


Figure: Joint

Joint and Conditional Probability Mass Functions

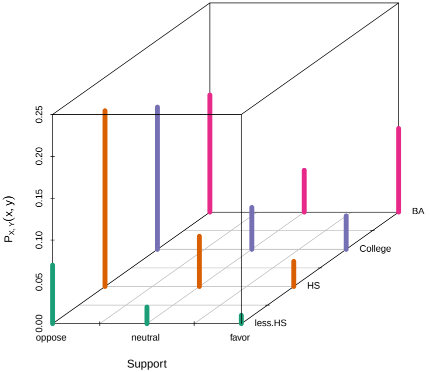


Figure: Joint

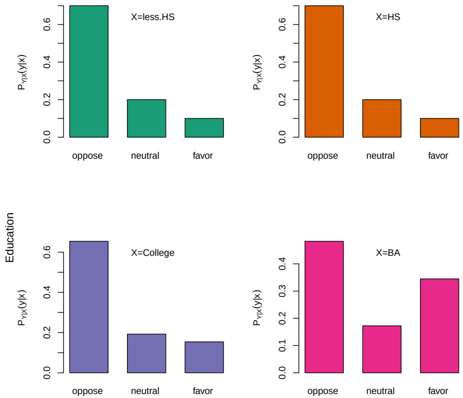


Figure: Conditional

Joint Probability Density Function

Definition

For two continuous random variables X and Y the joint PDF $f_{X,Y}(x, y)$ gives the density height where $X = x$ and $Y = y$ for all x and y.

Joint Probability Density Function

Definition

For two continuous random variables X and Y the joint PDF $f_{X,Y}(x, y)$ gives the density height where $X = x$ and $Y = y$ for all x and y.

The multiplicative rule:

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x)$$

where

- $f_{Y|X}(y|x)$: Conditional PDF of Y given $X = x$
- $f_X(x)$: Marginal PDF of X

Restrictions:

Joint Probability Density Function

Definition

For two continuous random variables X and Y the joint PDF $f_{X,Y}(x,y)$ gives the density height where $X = x$ and $Y = y$ for all x and y.

The multiplicative rule:

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x)$$

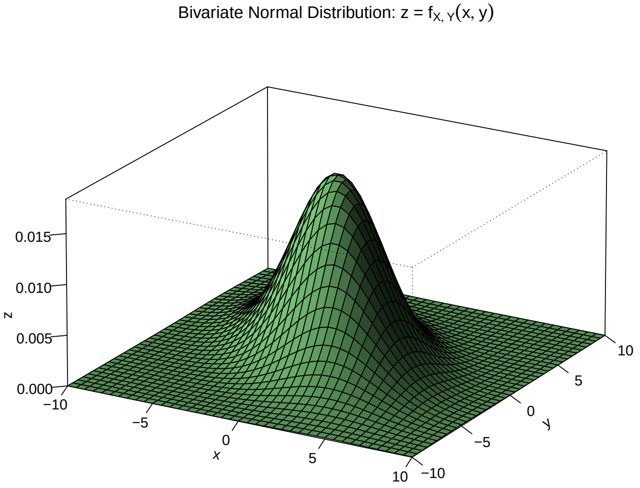
where

- $f_{Y|X}(y|x)$: Conditional PDF of Y given $X = x$
- $f_X(x)$: Marginal PDF of X

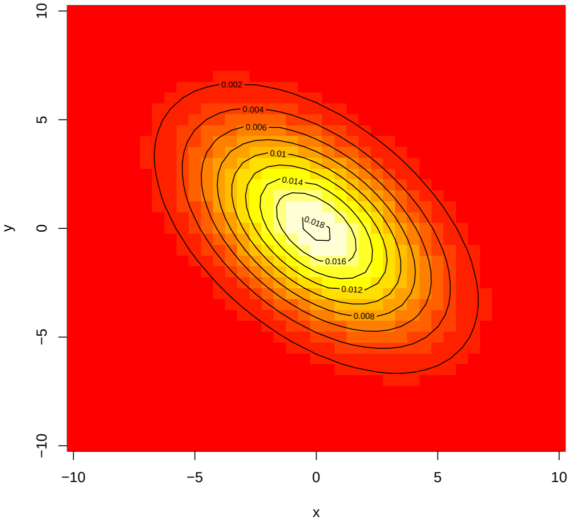
Restrictions:

- $\int_x \int_y f_{X,Y}(x,y)\, dy\, dx = 1$

3D Plot of a Joint Probability Density Function



Contour Plot of a Joint Probability Density Function



From Joint to Marginal PDF

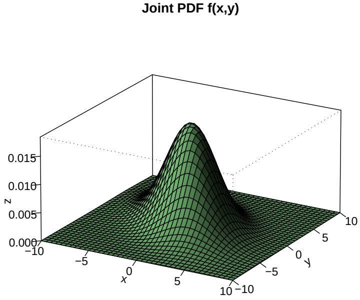
How can we obtain $f_Y(y)$ from $f_{X,Y}(x,y)$?

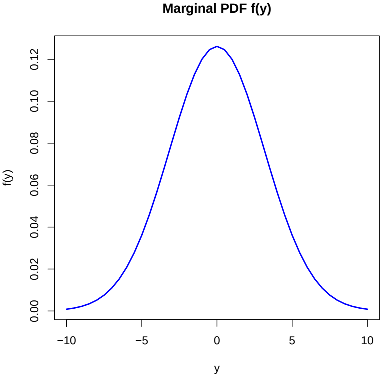
From Joint to Marginal PDF

How can we obtain $f_Y(y)$ from $f_{X,Y}(x,y)$?

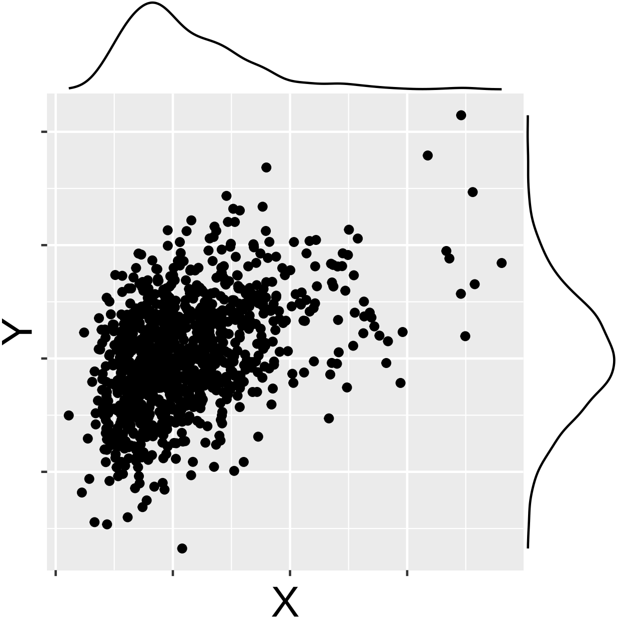
We marginalize the joint probability function $f_{X,Y}(x,y)$ over X:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$





From Joint to Marginal PDF



We Covered. . .

We Covered. . .

- Joint distributions for discrete and continuous random variables.

We Covered...

- Joint distributions for discrete and continuous random variables.
- Conditional distributions.

We Covered...

- Joint distributions for discrete and continuous random variables.
- Conditional distributions.
- Marginalization

We Covered. . .

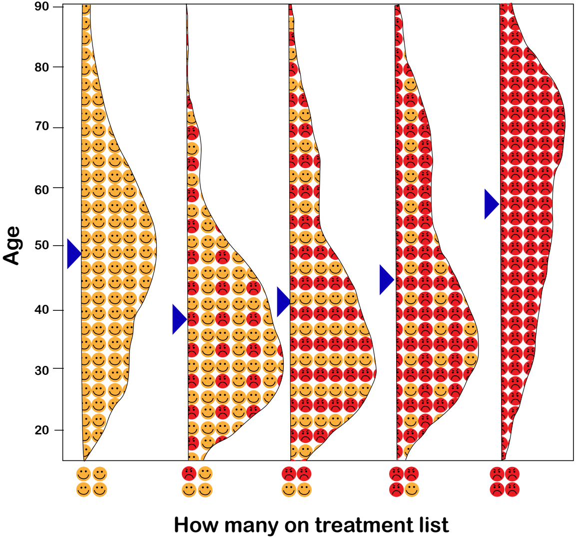
- Joint distributions for discrete and continuous random variables.
- Conditional distributions.
- Marginalization

Next time: Characterizing Conditional Distributions!

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Remember this?



Conditioning on X

Conditioning on X

- A common goal in statistical modeling is to characterize the conditional distribution of the outcome variable $f_{Y|X}(y|x)$ across different levels of $X = x$.

Conditioning on X

- A common goal in statistical modeling is to characterize the conditional distribution of the outcome variable $f_{Y|X}(y|x)$ across different levels of $X = x$.

- Typically, we summarize the conditional distributions with a few parameters such as the conditional mean of $E[Y|X = x]$ and the conditional variance $V[Y|X = x]$

Conditioning on X

- A common goal in statistical modeling is to characterize the conditional distribution of the outcome variable $f_{Y|X}(y|x)$ across different levels of $X = x$.

- Typically, we summarize the conditional distributions with a few parameters such as the conditional mean of $E[Y|X = x]$ and the conditional variance $V[Y|X = x]$

- Moreover, we are often interested in estimating $E[Y|X]$, i.e. the conditional expectation function that describes how the conditional mean of Y varies across all possible values of X.

Conditional Expectation

Definition (Conditional Expectation (Discrete))

Let Y and X be discrete random variables. The conditional expectation of Y given $X = x$ is defined as:

$$E[Y|X = x] = \sum_y y\, P(Y = y|X = x) = \sum_y y\, p_{Y|X}(y|x)$$

Conditional Expectation

Definition (Conditional Expectation (Discrete))

Let Y and X be discrete random variables. The conditional expectation of Y given $X = x$ is defined as:

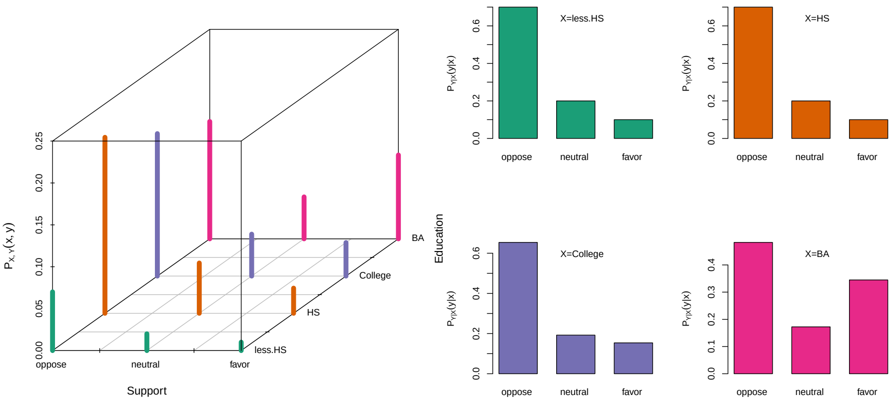
$$E[Y|X = x] = \sum_y y\, P(Y = y|X = x) = \sum_y y\, p_{Y|X}(y|x)$$

Definition (Conditional Expectation (Continuous))

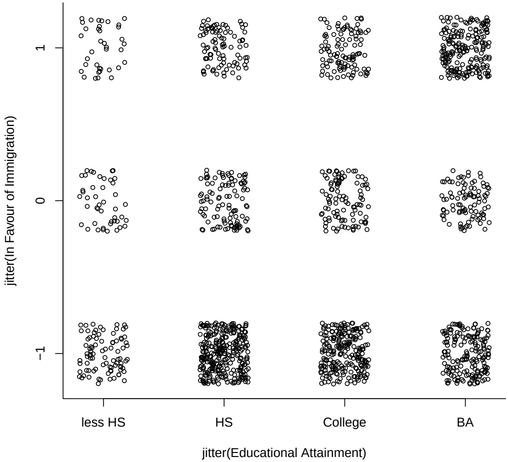
Let Y and X be continuous random variables. The conditional expectation of Y given $X = x$ is given by:

$$E[Y|X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)dy$$

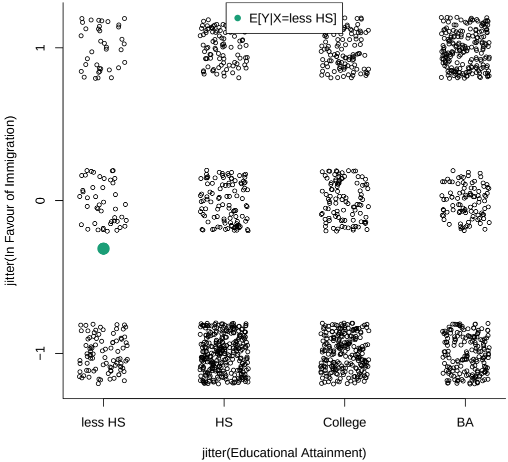
Joint and Conditional Probability Mass Functions



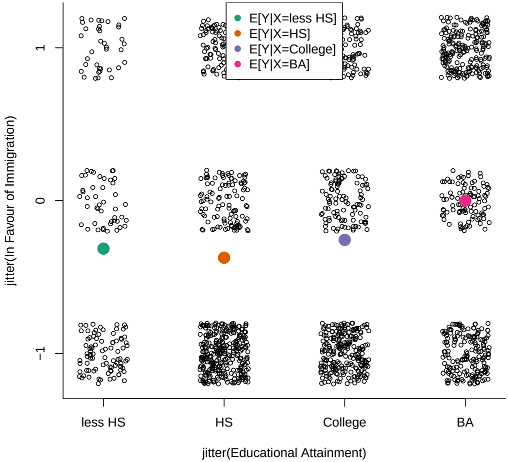
Conditional PMF $P_{Y|X}(y|x)$



Conditional Expectation $E[Y|X = 1]$



Conditional Expectation Function $E[Y|X]$



Law of Iterated Expectations

Law of Iterated Expectations

Theorem (Law of Iterated Expectations/Adam's Law)

For two random variables X and Y,

$$E[Y] = E[E[Y|X]] = \begin{cases} \sum_{\text{all } x} E[Y|X=x] \cdot p_X(x) & (\textit{discrete } X) \\ \int_{-\infty}^{\infty} E[Y|X=x] \cdot f_X(x) dx & (\textit{continuous } X) \end{cases}$$

Note that the outer expectation is taken with respect to the distribution of X.

Law of Iterated Expectations

Theorem (Law of Iterated Expectations/Adam's Law)

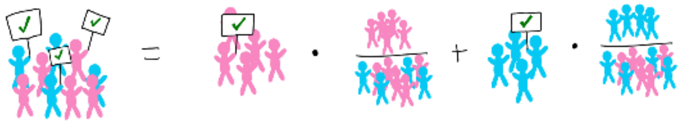
For two random variables X and Y,

$$E[Y] = E[E[Y|X]] = \begin{cases} \sum_{\text{all } x} E[Y|X=x] \cdot p_X(x) & (\textit{discrete } X) \\ \int_{-\infty}^{\infty} E[Y|X=x] \cdot f_X(x)dx & (\textit{continuous } X) \end{cases}$$

Note that the outer expectation is taken with respect to the distribution of X.

Example: Y (support) and $X \in \{1, 0\}$ (AfAm). Then, the LIE tells us:

$$\begin{aligned} \underbrace{E[Y]}_{\text{Average Support}} &= E[E[Y|X]] \\ &= \underbrace{E[Y|X=1]}_{\text{Average Support}|\text{AfAm}^c} \cdot \underbrace{p_X(1)}_{P(\text{AfAm}^c)} + \underbrace{E[Y|X=0]}_{\text{Average Support}|\text{AfAm}} \cdot \underbrace{p_X(0)}_{P(\text{AfAm})} \end{aligned}$$



Properties of Conditional Expectation

Properties of Conditional Expectation

Conditional expectations have some convenient properties

Properties of Conditional Expectation

Conditional expectations have some convenient properties

1. $E[c(X)|X] = c(X)$ for any function $c(X)$.

Properties of Conditional Expectation

Conditional expectations have some convenient properties

1. $E[c(X)|X] = c(X)$ for any function $c(X)$.
    - Basically, any function of X is a constant with regard to the conditional expectation. If we know X, then we also know X^2, for instance.

Properties of Conditional Expectation

Conditional expectations have some convenient properties

1. $E[c(X)|X] = c(X)$ for any function $c(X)$.
   - Basically, any function of X is a constant with regard to the conditional expectation. If we know X, then we also know X^2, for instance.
2. $E[(Y - E[Y|X])^2] \leq E[(Y - g(X))^2]$
   (given $E[Y^2] < \infty$ and $E[g(X)^2] < \infty$ for some function g)

Properties of Conditional Expectation

Conditional expectations have some convenient properties

1. $E[c(X)|X] = c(X)$ for any function $c(X)$.
   - Basically, any function of X is a constant with regard to the conditional expectation. If we know X, then we also know X^2, for instance.
2. $E[(Y - E[Y|X])^2] \leq E[(Y - g(X))^2]$
   (given $E[Y^2] < \infty$ and $E[g(X)^2] < \infty$ for some function g)
   - This says that the conditional expectation is the function of X that minimizes the squared prediction error for Y across any possible function of X.

Properties of Conditional Expectation

Conditional expectations have some convenient properties

1. $E[c(X)|X] = c(X)$ for any function $c(X)$.
   - Basically, any function of X is a constant with regard to the conditional expectation. If we know X, then we also know X^2, for instance.
2. $E[(Y - E[Y|X])^2] \leq E[(Y - g(X))^2]$
   (given $E[Y^2] < \infty$ and $E[g(X)^2] < \infty$ for some function g)
   - This says that the conditional expectation is the function of X that minimizes the squared prediction error for Y across any possible function of X.
   - This is analogous to the result we saw a few videos ago about the mean.

Conditional Variance

Conditional Variance

Conditional expectation gives us information about the central tendency of a random variable given another random variable.

Conditional Variance

Conditional expectation gives us information about the central tendency of a random variable given another random variable.

We also want to know the conditional variance to understand our uncertainty about the conditional distribution.

Conditional Variance

Conditional expectation gives us information about the central tendency of a random variable given another random variable.

We also want to know the conditional variance to understand our uncertainty about the conditional distribution.

Remember, the conditional distribution of $Y|X$ is basically like any other probability distribution, so we are going to want to summarize the center *and* spread.

Conditional Variance

Definition

The conditional variance of Y given $X = x$ is defined as:

$$V[Y|X = x] = \begin{cases} \displaystyle\sum_{\text{all } y} (y - E[Y|X = x])^2 P_{Y|X}(y|x) & \text{(discrete } Y\text{)} \\ \displaystyle\int_{-\infty}^{\infty} (y - E[Y|X = x])^2 f_{Y|X}(y|x) dy & \text{(continuous } Y\text{)} \end{cases}$$

Conditional Variance

Definition

The conditional variance of Y given $X = x$ is defined as:

$$V[Y|X = x] \;=\; \begin{cases} \displaystyle\sum_{\text{all } y} (y - E[Y|X = x])^2 P_{Y|X}(y|x) & \text{(discrete } Y\text{)} \\ \displaystyle\int_{-\infty}^{\infty} (y - E[Y|X = x])^2 f_{Y|X}(y|x) dy & \text{(continuous } Y\text{)} \end{cases}$$

A useful related result is the law of total variance (Eve's Law):

$$\underbrace{V[Y]}_{\text{Total variance}} \quad = \quad \underbrace{E[V[Y|X]]}_{\text{Average of Group Variances}} \quad + \quad \underbrace{V[E[Y|X]]}_{\text{Variance in Group Averages}}$$

Conditional Variance

Definition

The conditional variance of Y given $X = x$ is defined as:

$$V[Y|X=x] \;=\; \begin{cases} \displaystyle\sum_{\text{all } y} (y - E[Y|X=x])^2 P_{Y|X}(y|x) & \text{(discrete } Y\text{)} \\ \displaystyle\int_{-\infty}^{\infty} (y - E[Y|X=x])^2 f_{Y|X}(y|x) dy & \text{(continuous } Y\text{)} \end{cases}$$

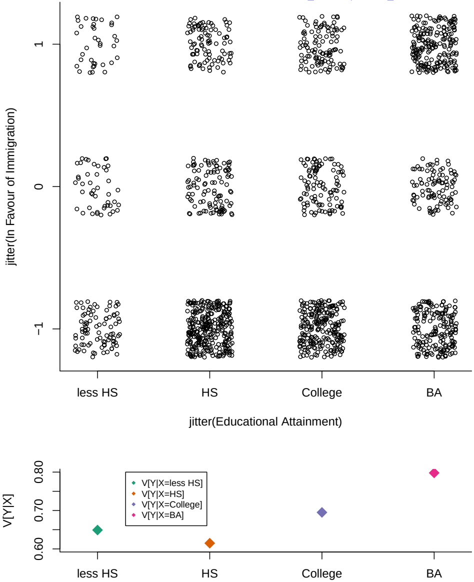
A useful related result is the law of total variance (Eve's Law):

$$\underbrace{V[Y]}_{\text{Total variance}} \;=\; \underbrace{E[V[Y|X]]}_{\text{Average of Group Variances}} \;+\; \underbrace{V[E[Y|X]]}_{\text{Variance in Group Averages}}$$

Example: Y (support) and $X \in \{1, 0\}$ (group). The LTV says that the total variance in support can be decomposed into two parts:

1. On average, how much support varies within groups (within variance)
2. How much average support varies between groups (between variance)

Conditional Variance Function $V[Y|X]$



Subtleties

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.
- $E[X]$ is a constant though (we sometimes refer to $E[\cdot]$ as an operator to make clear it doesn't behave the same as $g(\cdot)$).
- Why?

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.
- $E[X]$ is a constant though (we sometimes refer to $E[\cdot]$ as an operator to make clear it doesn't behave the same as $g(\cdot)$).
- Why? There is no longer anything stochastic in $E[X]$. Take the discrete case: $E[X] = \sum_x x p_X(x)$. Note that this is entirely in terms of realized values.

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.
- $E[X]$ is a constant though (we sometimes refer to $E[\cdot]$ as an operator to make clear it doesn't behave the same as $g(\cdot)$).
- Why? There is no longer anything stochastic in $E[X]$. Take the discrete case: $E[X] = \sum_x x p_X(x)$. Note that this is entirely in terms of realized values.
- By contrast, $E[X|Y]$ is random.

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.
- $E[X]$ is a constant though (we sometimes refer to $E[\cdot]$ as an operator to make clear it doesn't behave the same as $g(\cdot)$).
- Why? There is no longer anything stochastic in $E[X]$. Take the discrete case: $E[X] = \sum_x x p_X(x)$. Note that this is entirely in terms of realized values.
- By contrast, $E[X|Y]$ is random.
- $E[X|Y]$ is a function into which one can plug a value of $Y = y$ and get the expectation of X conditional on that value. Thus the randomness 'comes from' Y.

Subtleties

- It is important to distinguish between what is random/stochastic and what is constant. However, this can be tricky at first.
- If X is a random variable, generally a function of X ($g(X)$) is also a random variable.
- $E[X]$ is a constant though (we sometimes refer to $E[\cdot]$ as an operator to make clear it doesn't behave the same as $g(\cdot)$).
- Why? There is no longer anything stochastic in $E[X]$. Take the discrete case: $E[X] = \sum_x x p_X(x)$. Note that this is entirely in terms of realized values.
- By contrast, $E[X|Y]$ is random.
- $E[X|Y]$ is a function into which one can plug a value of $Y = y$ and get the expectation of X conditional on that value. Thus the randomness 'comes from' Y.

Let's look at this in pictures.
(If you want to know more: Blitzstein and Hwang pg 392-393)

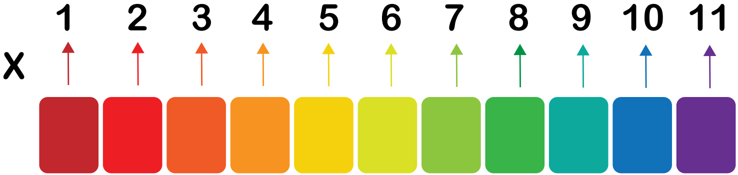
Important Subtleties in Pictures

Sample space

Important Subtleties in Pictures

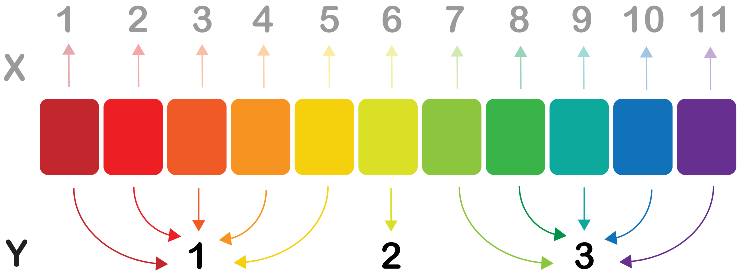
Sample space

Important Subtleties in Pictures



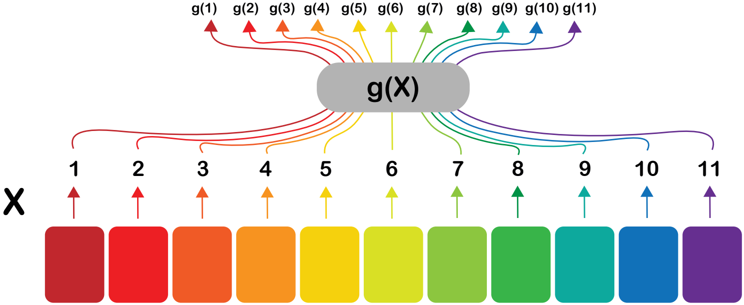
Random variable

Important Subtleties in Pictures



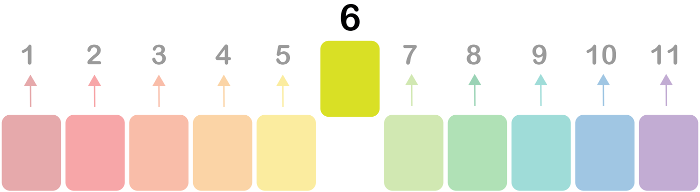
Random variable

Important Subtleties in Pictures



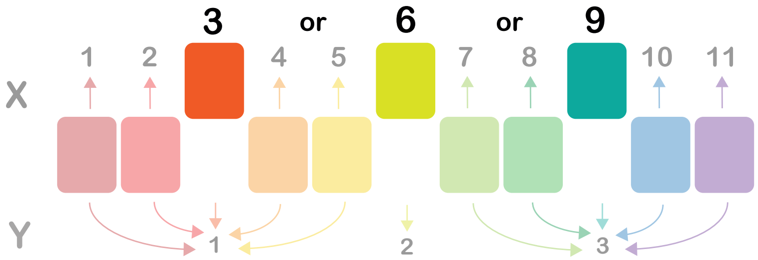
Function of a random variable is a random variable

Important Subtleties in Pictures



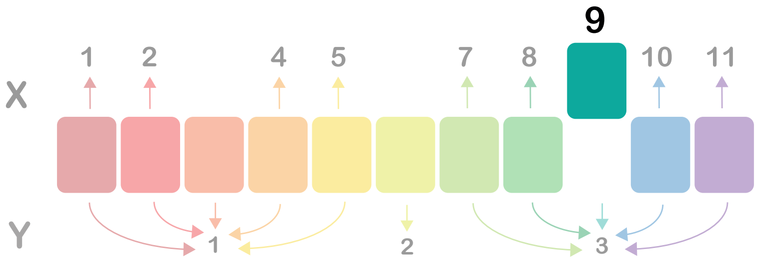
$E[X]$

Important Subtleties in Pictures



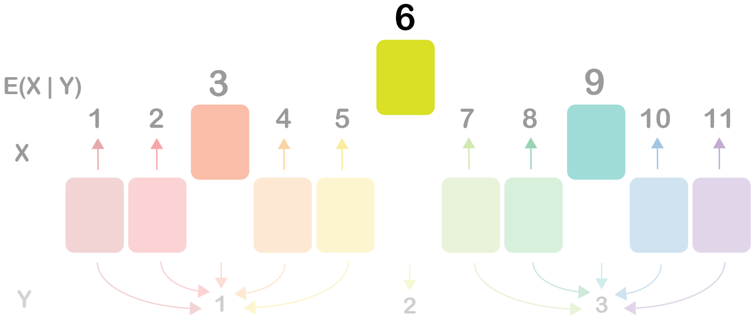
$$E[X|Y]$$

Important Subtleties in Pictures



$$E[X|Y=3]$$

Important Subtleties in Pictures



$$E[E[X|Y]] = E[X]$$

We Covered...

We Covered. . .

- Conditional Expectations

We Covered...

- Conditional Expectations
- Conditional Variance

We Covered. . .

- Conditional Expectations
- Conditional Variance
- Law of Iterated Expectation

We Covered. . .

- Conditional Expectations
- Conditional Variance
- Law of Iterated Expectation

Next time: Independence and Covariance!

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Independence

Definition (Independence of Random Variables)

Two random variables Y and X are independent if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for all x and y. We write this as $Y \perp\!\!\!\perp X$.

Independence

Definition (Independence of Random Variables)

Two random variables Y and X are independent if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all x and y. We write this as $Y \perp\!\!\!\perp X$.

Independence implies

$$f_{Y|X}(y|x) = f_Y(y)$$

and thus

$$E[Y|X = x] =$$

Independence

Definition (Independence of Random Variables)

Two random variables Y and X are independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all x and y. We write this as $Y \perp\!\!\!\perp X$.

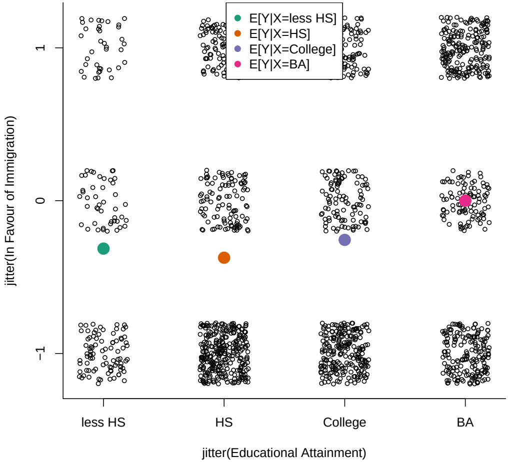
Independence implies

$$f_{Y|X}(y|x) = f_Y(y)$$

and thus

$$E[Y|X = x] = E[Y]$$

Is $Y \perp\!\!\!\perp X$?



Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$E[XY] \;=$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$E[XY] = \sum_{\text{all } x} \sum_{\text{all } y} x\, y\, p_{X,Y}(x, y)$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$\begin{aligned} E[XY] &= \sum_{\text{all } x}\sum_{\text{all } y} x\,y\,p_{X,Y}(x,y) \\ &= \sum_{\text{all } x}\sum_{\text{all } y} x\,y\,p_X(x)p_Y(y) \end{aligned}$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$\begin{aligned}
E[XY] &= \sum_{\text{all } x} \sum_{\text{all } y} x\, y\, p_{X,Y}(x, y) \\
&= \sum_{\text{all } x} \sum_{\text{all } y} x\, y\, p_X(x) p_Y(y) \\
&= \sum_{\text{all } x} x\, p_X(x) \sum_{\text{all } y} y\, p_Y(y)
\end{aligned}$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$\begin{aligned}
E[XY] &= \sum_{\text{all } x}\sum_{\text{all } y} x\, y\, p_{X,Y}(x, y) \\
&= \sum_{\text{all } x}\sum_{\text{all } y} x\, y\, p_X(x) p_Y(y) \\
&= \sum_{\text{all } x} x\, p_X(x) \sum_{\text{all } y} y\, p_Y(y) \\
&= E[X]E[Y]
\end{aligned}$$

Expected Values with Independent Random Variables

If random variables X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

Proof: For discrete X and Y,

$$\begin{aligned}
E[XY] &= \sum_{\text{all } x}\sum_{\text{all } y} x\, y\, p_{X,Y}(x, y) \\
&= \sum_{\text{all } x}\sum_{\text{all } y} x\, y\, p_X(x) p_Y(y) \\
&= \sum_{\text{all } x} x\, p_X(x) \sum_{\text{all } y} y\, p_Y(y) \\
&= E[X]E[Y]
\end{aligned}$$

We can prove the continuous case by following the same steps, with $\sum$ replaced by $\int$.

Covariance

Covariance

Definition

The covariance of X and Y is defined as:

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Covariance

Definition

The covariance of X and Y is defined as:

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Covariance measures the linear association between two random variables

Covariance

Definition

The covariance of X and Y is defined as:

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Covariance measures the linear association between two random variables .
- If $Cov[X, Y] > 0$, observing an X value greater than $E[X]$ makes it more likely to also observe a Y value greater than $E[Y]$, and vice versa.

Covariance

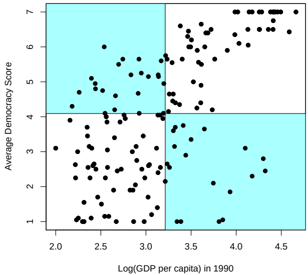
Definition

The covariance of X and Y is defined as:

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Covariance measures the linear association between two random variables .
- If $Cov[X, Y] > 0$, observing an X value greater than $E[X]$ makes it more likely to also observe a Y value greater than $E[Y]$, and vice versa.

  - Points in upper right and lower left quadrants (relative to the means) add to the covariance.
  - Points in the upper left and lower right quadrants subtract from the covariance.



Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$?

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= \end{aligned}$$

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$
\begin{aligned}
Cov[X, Y] &= E[XY] - E[X]E[Y] \\
&= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\
&= 0.
\end{aligned}
$$

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$?

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X,Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X,Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X,Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$?

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

What is $Cov[X, Y]$?

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

What is $Cov[X, Y]$?

$$\begin{aligned} Cov[X, Y] &= E[XX^2] - E[X]E[X^2] = E[X^3] - E[X]E[X^2] \\ &= \end{aligned}$$

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

What is $Cov[X, Y]$?

$$\begin{aligned} Cov[X, Y] &= E[XX^2] - E[X]E[X^2] = E[X^3] - E[X]E[X^2] \\ &= E[X] - E[X]E[X^2] = \end{aligned}$$

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

What is $Cov[X, Y]$?

$$\begin{aligned} Cov[X, Y] &= E[XX^2] - E[X]E[X^2] \; = \; E[X^3] - E[X]E[X^2] \\ &= E[X] - E[X]E[X^2] \; = \; 0 - 0 \cdot E[X^2] \; = \; 0. \end{aligned}$$

Covariance and Independence

Does $X \perp\!\!\!\perp Y$ imply $Cov[X, Y] = 0$? Yes!

Proof:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \quad \text{(independence)} \\ &= 0. \end{aligned}$$

Does $Cov[X, Y] = 0$ imply $X \perp\!\!\!\perp Y$? No!

Counterexample: Suppose $X \in \{-1, 0, 1\}$ with $p_X(x) = 1/3$ and $Y = X^2$.

Is $X \perp\!\!\!\perp Y$? No, because $p_{Y|X}(y \mid x) \neq p_Y(y)$
(Learning about X gives meaningful information about Y.)

What is $Cov[X, Y]$?

$$\begin{aligned} Cov[X, Y] &= E[XX^2] - E[X]E[X^2] \;=\; E[X^3] - E[X]E[X^2] \\ &= E[X] - E[X]E[X^2] \;=\; 0 - 0 \cdot E[X^2] \;=\; 0. \end{aligned}$$

Therefore, $X \perp\!\!\!\perp Y \implies Cov[X, Y] = 0$, but not vice versa.

Important Identities for Variances and Covariances

Important Identities for Variances and Covariances

1. For random variables X and Y and constants a, b and c,

$$V[aX + bY + c] = a^2 V[X] + b^2 V[Y] + 2ab\, Cov[X, Y]$$

Important Identities for Variances and Covariances

1. For random variables X and Y and constants a, b and c,

$$V[aX + bY + c] = a^2 V[X] + b^2 V[Y] + 2ab\, Cov[X, Y]$$

2. Important special cases:

$$V[X + Y] = V[X] + V[Y] + 2Cov[X, Y]$$
$$V[X - Y] = V[X] + V[Y] - 2Cov[X, Y]$$

Important Identities for Variances and Covariances

1. For random variables X and Y and constants a, b and c,

$$V[aX + bY + c] = a^2 V[X] + b^2 V[Y] + 2ab\, Cov[X, Y]$$

2. Important special cases:

$$V[X + Y] = V[X] + V[Y] + 2Cov[X, Y]$$
$$V[X - Y] = V[X] + V[Y] - 2Cov[X, Y]$$

3. Furthermore, if X and Y are independent,

$$V[X \pm Y] = V[X] + V[Y]$$

Important Identities for Variances and Covariances

1. For random variables X and Y and constants a, b and c,

$$V[aX + bY + c] = a^2 V[X] + b^2 V[Y] + 2ab\, Cov[X, Y]$$

2. Important special cases:

$$V[X + Y] = V[X] + V[Y] + 2Cov[X, Y]$$

$$V[X - Y] = V[X] + V[Y] - 2Cov[X, Y]$$

3. Furthermore, if X and Y are independent,

$$V[X \pm Y] = V[X] + V[Y]$$

Proof: Plug in to the definition of variance and expand (try it yourself!)

Correlation

Correlation

- $Cov[X, Y]$ depends not only on the strength of (linear) association between X and Y, but also the scale of X and Y.

Correlation

- $Cov[X, Y]$ depends not only on the strength of (linear) association between X and Y, but also the scale of X and Y.
- Can we have a pure measure of association that is scale-independent?

Correlation

- $Cov[X, Y]$ depends not only on the strength of (linear) association between X and Y, but also the scale of X and Y.
- Can we have a pure measure of association that is scale-independent?

Definition (Correlation)

The correlation between two random variables X and Y is defined as

$$Cor[X, Y] = \frac{Cov[X, Y]}{\sqrt{V[X]V[Y]}} = \frac{Cov[X, Y]}{SD[X]SD[Y]}.$$

Correlation

- $Cov[X, Y]$ depends not only on the strength of (linear) association between X and Y, but also the scale of X and Y.
- Can we have a pure measure of association that is scale-independent?

Definition (Correlation)

The correlation between two random variables X and Y is defined as

$$Cor[X, Y] = \frac{Cov[X, Y]}{\sqrt{V[X]V[Y]}} = \frac{Cov[X, Y]}{SD[X]SD[Y]}.$$

- $Cor[X, Y]$ is a standardized measure of linear association between X and Y.

Correlation

- $Cov[X, Y]$ depends not only on the strength of (linear) association between X and Y, but also the scale of X and Y.
- Can we have a pure measure of association that is scale-independent?

Definition (Correlation)

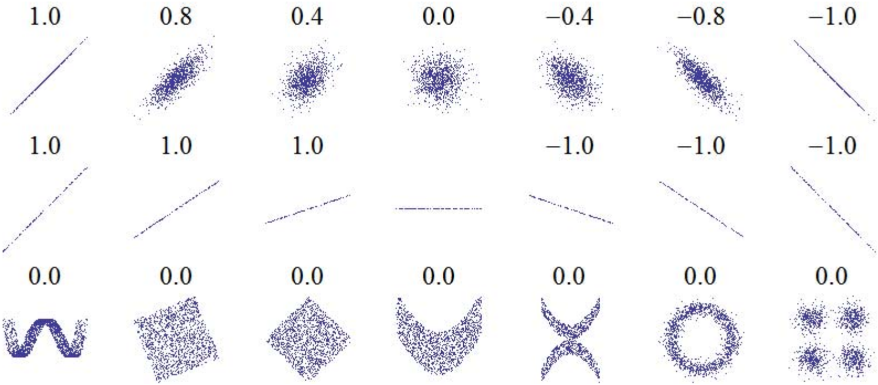
The correlation between two random variables X and Y is defined as

$$Cor[X, Y] = \frac{Cov[X, Y]}{\sqrt{V[X]V[Y]}} = \frac{Cov[X, Y]}{SD[X]SD[Y]}.$$

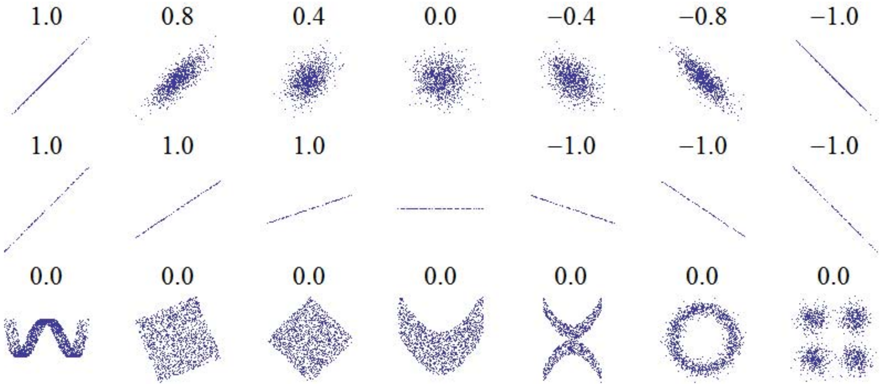
- $Cor[X, Y]$ is a standardized measure of linear association between X and Y.
- Always satisfies: $-1 \leq Cor[X, Y] \leq 1$.

Correlation is *Linear*

Correlation is *Linear*

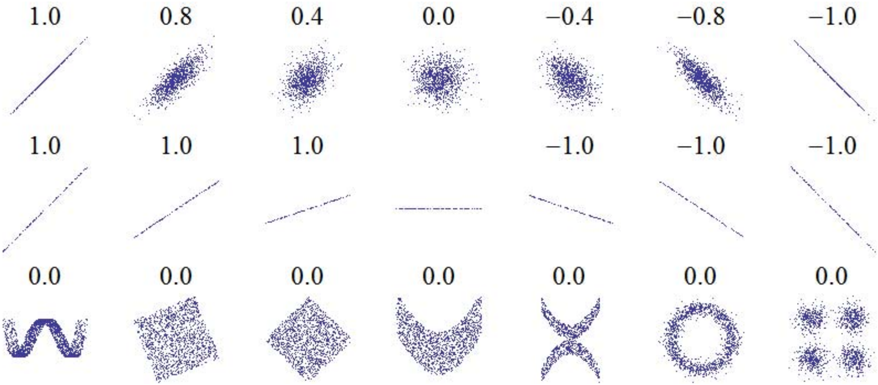


Correlation is *Linear*



- $Cor[X, Y] = \pm 1$ iff $Y = aX + b$ where $a \neq 0$.

Correlation is *Linear*



- $Cor[X, Y] = \pm 1$ iff $Y = aX + b$ where $a \neq 0$.
- Like covariance, correlation measures the linear association between X and Y.

Conditional Independence

Conditional Independence

Definition (Conditional Independence of Random Variables)

Random variables Y and X are conditionally independent given Z iff

$$f_{X,Y|Z}(x,y|z) \; = \; f_{Y|Z}(y|z) \cdot f_{X|Z}(x|z)$$

for all x, y, and z. This is often written as $Y \perp\!\!\!\perp X \mid Z$.

Conditional Independence

Definition (Conditional Independence of Random Variables)

Random variables Y and X are conditionally independent given Z iff

$$f_{X,Y|Z}(x, y|z) \ = \ f_{Y|Z}(y|z) \cdot f_{X|Z}(x|z)$$

for all x, y, and z. This is often written as $Y \perp\!\!\!\perp X \mid Z$.

- Can also be written as

$$f_{Y|X,Z}(y \mid x, z) \ = \ f_{Y|Z}(y \mid z)$$

Conditional Independence

Definition (Conditional Independence of Random Variables)

Random variables Y and X are conditionally independent given Z iff

$$f_{X,Y|Z}(x, y|z) \;=\; f_{Y|Z}(y|z) \cdot f_{X|Z}(x|z)$$

for all x, y, and z. This is often written as $Y \perp\!\!\!\perp X \mid Z$.

- Can also be written as

$$f_{Y|X,Z}(y \mid x, z) \;=\; f_{Y|Z}(y \mid z)$$

- Interpretation: Once we know Z, X contains no meaningful information about likely values of Y.
  (Z has all the information about Y contained in X, if any.)

Conditional Independence

Definition (Conditional Independence of Random Variables)

Random variables Y and X are conditionally independent given Z iff

$$f_{X,Y|Z}(x, y|z) \;=\; f_{Y|Z}(y|z) \cdot f_{X|Z}(x|z)$$

for all x, y, and z. This is often written as $Y \perp\!\!\!\perp X \mid Z$.

- Can also be written as

$$f_{Y|X,Z}(y \mid x, z) \;=\; f_{Y|Z}(y \mid z)$$

- Interpretation: Once we know Z, X contains no meaningful information about likely values of Y.
  (Z has all the information about Y contained in X, if any.)
- $Y \perp\!\!\!\perp X \mid Z$ implies

$$E[Y|X = x, Z = z] =$$

Conditional Independence

Definition (Conditional Independence of Random Variables)

Random variables Y and X are conditionally independent given Z iff

$$f_{X,Y|Z}(x, y|z) \;=\; f_{Y|Z}(y|z) \cdot f_{X|Z}(x|z)$$

for all x, y, and z. This is often written as $Y \perp\!\!\!\perp X \mid Z$.

- Can also be written as

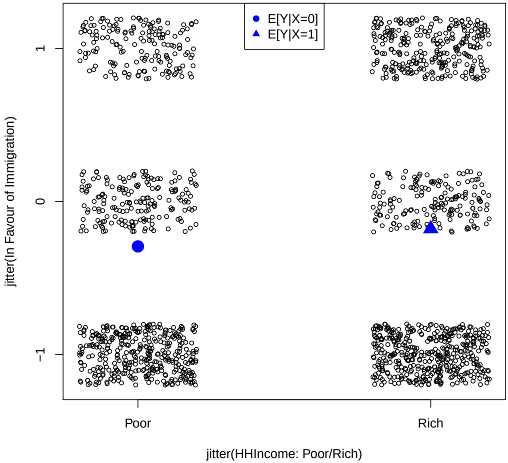
$$f_{Y|X,Z}(y \mid x, z) \;=\; f_{Y|Z}(y \mid z)$$

- Interpretation: Once we know Z, X contains no meaningful information about likely values of Y.
  (Z has all the information about Y contained in X, if any.)
- $Y \perp\!\!\!\perp X \mid Z$ implies

$$E[Y|X = x, Z = z] = E[Y|Z = z].$$

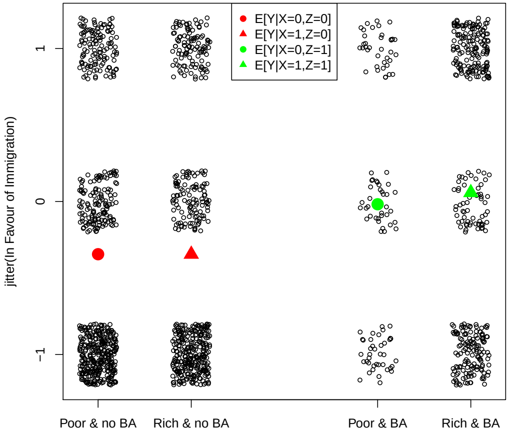
Is $Y \perp\!\!\!\perp X$?

Example: X = wealth, Y = support for immigration, Z = education.



Is $Y \perp\!\!\!\perp X | Z$?

Example: X = wealth, Y = support for immigration, Z = education.



We Covered...

We Covered...

- Independence

We Covered. . .

- Independence
- Covariance and Correlation

We Covered. . .

- Independence
- Covariance and Correlation
- Conditional Independence

We Covered...

- Independence
- Covariance and Correlation
- Conditional Independence

Next time: Famous Distributions!

Where We've Been and Where We're Going...

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability.
- This Week
  - define random variables
  - summarize random variables using expectation and variance
  - properties of joint and conditional distributions
  - famous distributions
- Next Week
  - estimating these features from data
  - estimating uncertainty
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

Distributions

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.
- We can work with arbitrary pmf/pdfs but we will often work with particular families of distributions.

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.
- We can work with arbitrary pmf/pdfs but we will often work with particular families of distributions.
  - members of the same family have similar forms determined by parameters

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.
- We can work with arbitrary pmf/pdfs but we will often work with particular families of distributions.
  - members of the same family have similar forms determined by parameters
  - the parameters determine the shape of the distribution

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.
- We can work with arbitrary pmf/pdfs but we will often work with particular families of distributions.
  - members of the same family have similar forms determined by parameters
  - the parameters determine the shape of the distribution
- When we can work with an existing set of distributions, it makes calculations simpler

Distributions

- We like random variables because they take complex real world phenomena and represent them with a common mathematical infrastructure.
- We can work with arbitrary pmf/pdfs but we will often work with particular families of distributions.
  - members of the same family have similar forms determined by parameters
  - the parameters determine the shape of the distribution
- When we can work with an existing set of distributions, it makes calculations simpler
- Examples: Bernoulli, Binomial, Gamma, Normal, Poisson, t-distribution

Bernoulli Random Variable

Definition

Suppose X is a random variable, with $X \in \{0, 1\}$ and $P(X = 1) = \pi$.
Then we will say that X is Bernoulli random variable,

$$P(X = x) \quad = \quad \pi^x (1 - \pi)^{1-x}$$

for $x \in \{0, 1\}$ and $P(X = x) = 0$ otherwise.
We will (equivalently) say that

$$X \quad \sim \quad \text{Bernoulli}(\pi)$$

$\sim$ means equality in distribution (not values!). Often $X \sim \text{Bernoulli}(\pi)$ would be read 'X is distributed Bernoulli with parameter π'

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$\begin{aligned} E[X] &= 1 \times P(X = 1) + 0 \times P(X = 0) \\ &= \pi + 0(1 - \pi) = \pi \end{aligned}$$

$E[X] = \pi$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$
\begin{aligned}
E[X] &= 1 \times P(X = 1) + 0 \times P(X = 0) \\
&= \pi + 0(1 - \pi) = \pi \\
\text{var}(X) &= E[X^2] - E[X]^2
\end{aligned}
$$

$E[X] = \pi$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$\begin{aligned} E[X] &= 1 \times P(X=1) + 0 \times P(X=0) \\ &= \pi + 0(1-\pi) = \pi \\ \text{var}(X) &= E[X^2] - E[X]^2 \\ E[X^2] &= 1^2 P(X=1) + 0^2 P(X=0) \end{aligned}$$

$E[X] = \pi$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$\begin{aligned}
E[X] &= 1 \times P(X=1) + 0 \times P(X=0) \\
&= \pi + 0(1-\pi) = \pi \\
\text{var}(X) &= E[X^2] - E[X]^2 \\
E[X^2] &= 1^2 P(X=1) + 0^2 P(X=0) \\
&= \pi
\end{aligned}$$

$E[X] = \pi$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$
\begin{aligned}
E[X] &= 1 \times P(X=1) + 0 \times P(X=0) \\
&= \pi + 0(1-\pi) = \pi \\
\text{var}(X) &= E[X^2] - E[X]^2 \\
E[X^2] &= 1^2 P(X=1) + 0^2 P(X=0) \\
&= \pi \\
\text{var}(X) &= \pi - \pi^2
\end{aligned}
$$

$E[X] = \pi$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$\begin{aligned}
E[X] &= 1 \times P(X = 1) + 0 \times P(X = 0) \\
&= \pi + 0(1 - \pi) = \pi \\
\text{var}(X) &= E[X^2] - E[X]^2 \\
E[X^2] &= 1^2 P(X = 1) + 0^2 P(X = 0) \\
&= \pi \\
\text{var}(X) &= \pi - \pi^2 \\
&= \pi(1 - \pi)
\end{aligned}$$

$E[X] = \pi$
$\text{var}(X) = \pi(1 - \pi)$

Bernoulli Random Variable Mean and Variance

Suppose $X \sim \text{Bernoulli}(\pi)$

$$\begin{aligned} E[X] &= 1 \times P(X = 1) + 0 \times P(X = 0) \\ &= \pi + 0(1 - \pi) = \pi \\ \text{var}(X) &= E[X^2] - E[X]^2 \\ E[X^2] &= 1^2 P(X = 1) + 0^2 P(X = 0) \\ &= \pi \\ \text{var}(X) &= \pi - \pi^2 \\ &= \pi(1 - \pi) \end{aligned}$$

$E[X] = \pi$
$\text{var}(X) = \pi(1 - \pi)$

Importantly, we can also just look this up!

Normal/Gaussian Random Variables

Definition

Suppose X is a random variable with $X \in \mathbb{R}$ and density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Then X is a normally distributed random variable with parameters μ and σ^2.
Equivalently, we'll write

$$X \sim \text{Normal}(\mu, \sigma^2)$$

Expected Value/Variance of Normal Distribution

Z is a standard normal distribution if

Expected Value/Variance of Normal Distribution

Z is a standard normal distribution if

$$Z \sim \text{Normal}(0, 1)$$

Expected Value/Variance of Normal Distribution

Z is a standard normal distribution if

$$Z \sim \text{Normal}(0, 1)$$

We'll call the cumulative distribution function of Z,

Expected Value/Variance of Normal Distribution

Z is a standard normal distribution if

$$Z \quad \sim \quad \text{Normal}(0, 1)$$

We'll call the cumulative distribution function of Z,

$$F_Z(x) \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2/2) dz$$

Expected Value/Variance of Normal Distribution

Z is a standard normal distribution if

$$Z \sim \text{Normal}(0, 1)$$

We'll call the cumulative distribution function of Z,

$$F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2/2) dz$$

Proposition

Scale/Location. If $Z \sim N(0, 1)$, then $X = aZ + b$ is,

$$X \sim Normal(b, a^2)$$

Intuition

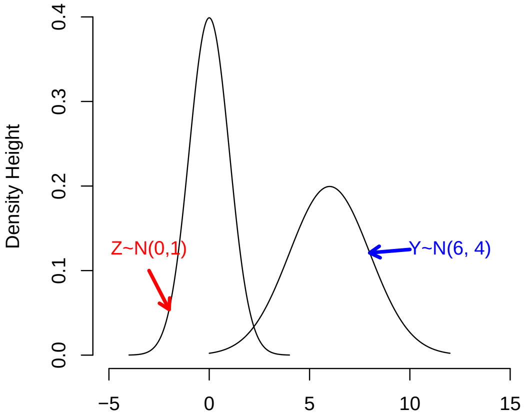
Suppose $Z \sim \text{Normal}(0, 1)$.

Intuition

Suppose $Z \sim \text{Normal}(0, 1)$.
$Y = 2Z + 6$

Intuition

Suppose $Z \sim \text{Normal}(0, 1)$.
$Y = 2Z + 6$
$Y \sim \text{Normal}(6, 4)$



Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

Expectation and Variance

Assume we know:

$$
\begin{aligned}
E[Z] &= 0 \\
Var(Z) &= 1
\end{aligned}
$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$
E[Y] = E[\sigma Z + \mu]
$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \\ &= \mu \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \\ &= \mu \\ Var(Y) &= Var(\sigma Z + \mu) \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \\ &= \mu \\ Var(Y) &= Var(\sigma Z + \mu) \\ &= \sigma^2 Var(Z) + Var(\mu) \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \\ &= \mu \\ Var(Y) &= Var(\sigma Z + \mu) \\ &= \sigma^2 Var(Z) + Var(\mu) \\ &= \sigma^2 + 0 \end{aligned}$$

Expectation and Variance

Assume we know:

$$\begin{aligned} E[Z] &= 0 \\ Var(Z) &= 1 \end{aligned}$$

This implies that, for $Y \sim \text{Normal}(\mu, \sigma^2)$

$$\begin{aligned} E[Y] &= E[\sigma Z + \mu] \\ &= \sigma E[Z] + \mu \\ &= \mu \\ Var(Y) &= Var(\sigma Z + \mu) \\ &= \sigma^2 Var(Z) + Var(\mu) \\ &= \sigma^2 + 0 \\ &= \sigma^2 \end{aligned}$$

Multivariate Normal

Definition

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ is a vector of random variables. If $\boldsymbol{X}$ has pdf

$$f_{X_1,X_2}(\boldsymbol{x}) \quad = \quad (2\pi)^{-N/2}\det\left(\boldsymbol{\Sigma}\right)^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Then we will say $\boldsymbol{X}$ has a Multivariate Normal Distribution,

$$\boldsymbol{X} \quad \sim \quad \text{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Multivariate Normal Distribution

Consider the (bivariate) special case where $\boldsymbol{\mu} = (0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Multivariate Normal Distribution

Consider the (bivariate) special case where $\boldsymbol{\mu} = (0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$f(x_1, x_2) = (2\pi)^{-2/2} 1^{-1/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \mathbf{0})' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x} - \mathbf{0})\right)\right)$$

Multivariate Normal Distribution

Consider the (bivariate) special case where $\boldsymbol{\mu} = (0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-2/2} 1^{-1/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \mathbf{0})' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x} - \mathbf{0})\right)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \end{aligned}$$

Multivariate Normal Distribution

Consider the (bivariate) special case where $\boldsymbol{\mu} = (0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-2/2} 1^{-1/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \mathbf{0})' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x} - \mathbf{0})\right)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \end{aligned}$$

Multivariate Normal Distribution

Consider the (bivariate) special case where $\boldsymbol{\mu} = (0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-2/2} 1^{-1/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \boldsymbol{0})' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x} - \boldsymbol{0})\right)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \end{aligned}$$

$\rightsquigarrow$ product of univariate standard normally distributed random variables

Properties of the Multivariate Normal Distribution

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$

$$\begin{aligned} E[\boldsymbol{X}] &= \boldsymbol{\mu} \\ \text{cov}(\boldsymbol{X}) &= \boldsymbol{\Sigma} \end{aligned}$$

So that,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \ldots & \text{cov}(X_1, X_N) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \ldots & \text{cov}(X_2, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_N, X_1) & \text{cov}(X_N, X_2) & \ldots & \text{var}(X_N) \end{pmatrix}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{y \log\mu - \exp(\log\mu) - \log y!\right\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

Example: Poisson(μ):

$$P(Y_i = y \mid \mu) \;=\; \exp\{y \log \mu - \exp(\log \mu) - \log y!\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{y\log\mu - \exp(\log\mu) - \log y!\right\}$$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y; \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

Example: Poisson(μ):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{ y \log \mu - \exp(\log \mu) - \log y! \right\}$$

$\implies \theta = \log \mu$, $\phi = 1$, $a(\phi) = \phi, b(\theta) = \exp(\theta)$, and $c = -\log y!$

One Step Deeper: Exponential Family

Nearly every distribution we will discuss is in the exponential family. An exponential family distribution has the density of the following form:

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Example: Poisson(μ):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{y\log\mu - \exp(\log\mu) - \log y!\right\}$$

$\Longrightarrow \theta = \log\mu$, $\phi = 1$, $a(\phi) = \phi, b(\theta) = \exp(\theta)$, and $c = -\log y!$

Many other examples, including: Normal, Bernoulli/binomial, Gamma, multinomial, exponential, negative binomial, beta, uniform, chi-squared, etc.

This slide and the following based on material from Teppei Yamamoto

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function
- In the Poisson model, $\theta_i = \log \mu_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$

One Step Deeper: Properties of the Exponential Family

- Mean is a function of θ and given by

$$\mathbb{E}(Y) \equiv \mu = b'(\theta)$$

- Variance is a function of θ and ϕ and given by

$$\mathbb{V}(Y) \equiv V = b''(\theta)a(\phi)$$

- Common forms of $a(\phi)$: 1 (Poisson, Bernoulli), ϕ (normal, Gamma), and ϕ/ω_i (binomial)
- $b''(\theta)$ is called the variance function
- In the Poisson model, $\theta_i = \log \mu_i$, $a(\phi) = 1$ and $b(\theta_i) = \exp(\theta_i)$
  $\Rightarrow \mathbb{E}(Y_i) = \frac{db(\theta_i)}{d\theta_i} = \exp(\theta_i) = \mu_i$ and $\mathbb{V}(Y_i) = \frac{d^2 b(\theta_i)}{d\theta_i^2} = \exp(\theta_i) = \mu_i$

Summary

Summary

- Random variables and probability distributions provide useful models of the world

Summary

- Random variables and probability distributions provide useful models of the world
- We can characterize distributions in terms of their expectation (location) and variance (spread).

Summary

- Random variables and probability distributions provide useful models of the world
- We can characterize distributions in terms of their expectation (location) and variance (spread).
- Joint and conditional distributions capture the relationship between random variables.

Summary

- Random variables and probability distributions provide useful models of the world
- We can characterize distributions in terms of their expectation (location) and variance (spread).
- Joint and conditional distributions capture the relationship between random variables.
- There is a common set of famous distributions such as the Normal distribution.

This Week in Review

- Random Variables!
- Expectation and Variance!
- Distributions!

This Week in Review

- Random Variables!
- Expectation and Variance!
- Distributions!

Going Deeper:

Blitzstein, Joseph K., and Hwang, Jessica. (2019). *Introduction to Probability*. CRC Press. http://stat110.net/

This Week in Review

- Random Variables!
- Expectation and Variance!
- Distributions!

Going Deeper:

Blitzstein, Joseph K., and Hwang, Jessica. (2019). *Introduction to Probability*. CRC Press. `http://stat110.net/`

Next week: inference!