Week 2: Random Variables

Brandon Stewart[1]

Princeton

September 10–12, 2024

[1] Many of these slides are adapted from Adam Glynn and Jens Hainmueller. Some additional individual slides come from Justin Grimmer, Erin Hartman, Ian Lundberg, and Matt Salganik. 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
  - initial introduction to random variables
- This Week

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

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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 with regression

Random Variable

Intuition:
A function that maps outcomes of our random generative process to numbers.

Random Variable

Intuition:
A function that maps outcomes of our random generative process to numbers.

- a random variable is a container or placeholder for a quantity that has yet to be determined by a random generative process
- just as ordinary algebraic variables represent unknown or unspecified quantities

Random Variable

Intuition:
A function that maps outcomes of our random generative process to numbers.

- a random variable is a container or placeholder for a quantity that has yet to be determined by a random generative process
- just as ordinary algebraic variables represent unknown or unspecified quantities

Formally:

Definition (Random Variable)

A random variable is a function $X : \Omega \to \mathbb{R}$ such that,
$\forall r \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \leq r\} \in S$.

Event Space

Formally, a set S of subsets of Ω is an event space if it satisfies the following:

- Nonempty: $S \neq \emptyset$
- Closed under complements: if $A \in S$, then $A^C \in S$.
- Closed under countable unions: if $A_1, A_2, A_3, \dots \in S$, then $A_1 \cup A_2 \cup A_3 \cup \dots \in S$

Event Space

Formally, a set S of subsets of Ω is an event space if it satisfies the following:

- Nonempty: $S \neq \emptyset$
- Closed under complements: if $A \in S$, then $A^C \in S$.
- Closed under countable unions: if $A_1, A_2, A_3, \dots \in S$, then $A_1 \cup A_2 \cup A_3 \cup \dots \in S$

A technical note (beyond this course):

- The set that satisfies these properties is formally known as a σ-algebra or σ-field on Ω.
- For finite sample spaces, the event space is typically the power set (that is, the set of all subsets) of all events.
- But, this is not always true. For some sample spaces Ω, it is impossible to define the probability of every subset of Ω in a manner consistent with the axioms of probability.

Event Space

Formally, a set S of subsets of Ω is an event space if it satisfies the following:

- Nonempty: $S \neq \emptyset$
- Closed under complements: if $A \in S$, then $A^C \in S$.
- Closed under countable unions: if $A_1, A_2, A_3, \dots \in S$, then $A_1 \cup A_2 \cup A_3 \cup \cdots \in S$

A technical note (beyond this course):

- The set that satisfies these properties is formally known as a σ-algebra or σ-field on Ω.
- For finite sample spaces, the event space is typically the power set (that is, the set of all subsets) of all events.
- But, this is not always true. For some sample spaces Ω, it is impossible to define the probability of every subset of Ω in a manner consistent with the axioms of probability.

Simplification for This Course

In more formal settings you will often see probability spaces defined in terms of a probability triple:

- The sample space Ω
- the σ-algebra
- the probability measure $\mathbf{P}$

Simplification for This Course

In more formal settings you will often see probability spaces defined in terms of a probability triple:

- The sample space Ω
- the σ-algebra
- the probability measure $\mathbf{P}$

In the following definition:

Definition (Random Variable)

A random variable is a function $X : \Omega \to \mathbb{R}$ such that, $\forall r \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \leq r\} \in S$.

Everything after "such that" is the necessary and sufficient regularity condition to ensure that X is a measurable function.

Simplification for This Course

In more formal settings you will often see probability spaces defined in terms of a probability triple:

- The sample space Ω
- the σ-algebra
- the probability measure $\mathbf{P}$

In the following definition:

Definition (Random Variable)

A random variable is a function $X : \Omega \to \mathbb{R}$ such that, $\forall r \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \leq r\} \in S$.

Everything after "such that" is the necessary and sufficient regularity condition to ensure that X is a measurable function.

We will (1) omit the discussion of σ-algebra and measurable spaces and (2) suppress the notation $X(\omega)$ working instead just with X.

Our First Example: NH Ballot Order

Candidates:

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

$$X = \left\{ \phantom{\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array}} \right.$$

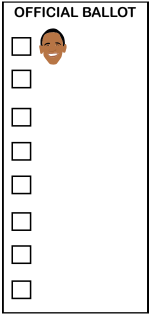
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \end{cases}$$



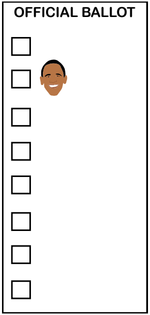
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \end{cases}$$



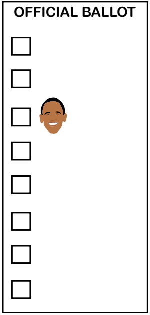
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \end{cases}$$



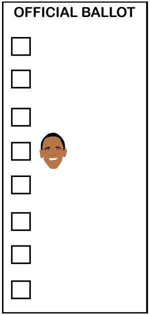
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \\ 4 & \text{if E} \end{cases}$$



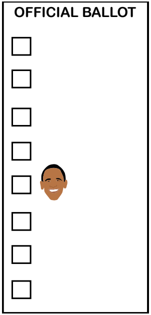
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \\ 4 & \text{if E} \\ 5 & \text{if D} \end{cases}$$



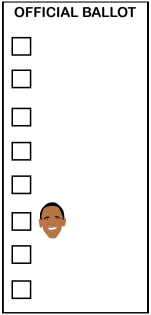
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \\ 4 & \text{if E} \\ 5 & \text{if D} \\ 6 & \text{if C} \end{cases}$$



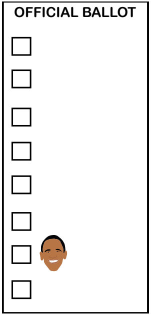
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \\ 4 & \text{if E} \\ 5 & \text{if D} \\ 6 & \text{if C} \\ 7 & \text{if S,T,U,V,W,X,Y,Z,A,B} \end{cases}$$



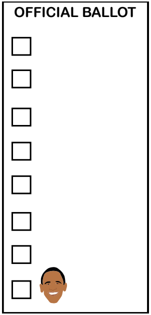
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

Our First Example: NH Ballot Order

Candidates:

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

$$X = \begin{cases} 1 & \text{if L,M,N,O} \\ 2 & \text{if H,I,J,K} \\ 3 & \text{if F,G} \\ 4 & \text{if E} \\ 5 & \text{if D} \\ 6 & \text{if C} \\ 7 & \text{if S,T,U,V,W,X,Y,Z,A,B} \\ 8 & \text{if P,Q,R} \end{cases}$$



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

Characterizing Distributions

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

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.
- For example, we will (shortly) summarize the expectation ($E[X]$) and variance ($V[X]$) of a random variable

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.
- For example, we will (shortly) summarize the expectation ($E[X]$) and variance ($V[X]$) of a random variable

A function of a random variable uses the value of X as an input into another function. We will denote it with parentheses.

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.
- For example, we will (shortly) summarize the expectation ($E[X]$) and variance ($V[X]$) of a random variable

A function of a random variable uses the value of X as an input into another function. We will denote it with parentheses.

- Let $g : U \to \mathbb{R}$ be a function, where $X(\Omega) \subseteq U \subseteq \mathbb{R}$. Then if $g(X) : \Omega \to \mathbb{R}$ is a random variable, we say that g is a function of X.

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.
- For example, we will (shortly) summarize the expectation ($E[X]$) and variance ($V[X]$) of a random variable

A function of a random variable uses the value of X as an input into another function. We will denote it with parentheses.

- Let $g : U \to \mathbb{R}$ be a function, where $X(\Omega) \subseteq U \subseteq \mathbb{R}$. Then if $g(X) : \Omega \to \mathbb{R}$ is a random variable, we say that g is a function of X.
- Because $g(X)$ is a function of a random variable, it is, itself, a random variable too!

Operators and Functions on Random Variables

An operator A on a random variable maps the function X to a real number, denoted by $A[X]$.

- An operator on a random variable returns a value characterizing some feature of X. Because it is a descriptive characteristic, we will denote it with square brackets.
- For example, we will (shortly) summarize the expectation ($E[X]$) and variance ($V[X]$) of a random variable

A function of a random variable uses the value of X as an input into another function. We will denote it with parentheses.

- Let $g : U \to \mathbb{R}$ be a function, where $X(\Omega) \subseteq U \subseteq \mathbb{R}$. Then if $g(X) : \Omega \to \mathbb{R}$ is a random variable, we say that g is a function of X.
- Because $g(X)$ is a function of a random variable, it is, itself, a random variable too!

The notation can look confusing, so ask questions if it becomes unclear! But we will return to the key subtleties of this point by the end of lecture!

Our First Operator: Expectation

Our First Operator: Expectation

The expected value of a random variable X is denoted by $E[X]$ and is a measure of central tendency of X. The expectation is (roughly) the average of all of the values weighted by probability of occurrence.

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

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

Our First Operator: Expectation

The expected value of a random variable X is denoted by $E[X]$ and is a measure of central tendency of X. The expectation is (roughly) the 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

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

Note that $E[\cdot]$ is an operator. It takes a random variable as an input and returns a number that describes the distribution (not a random variable!).

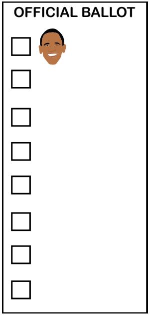
What did we expect for Obama's NH position?

Candidates:

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

$4/26 \quad \times 1$

$+$



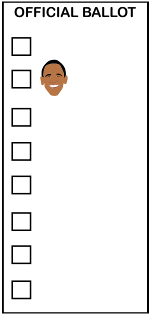
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



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 did we expect for Obama's NH position?

Candidates:

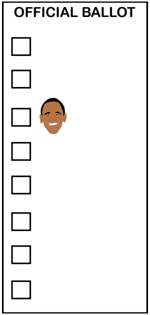
- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

$4/26 \quad \times 1$

$4/26 \quad \times 2$

$2/26 \quad \times 3$

$+$



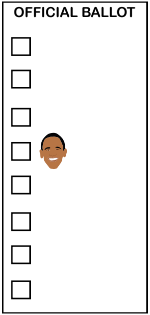
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



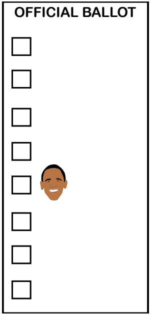
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



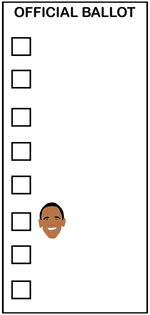
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



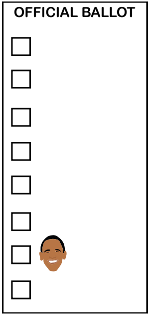
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



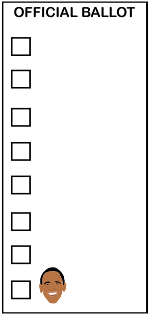
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

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

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



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:

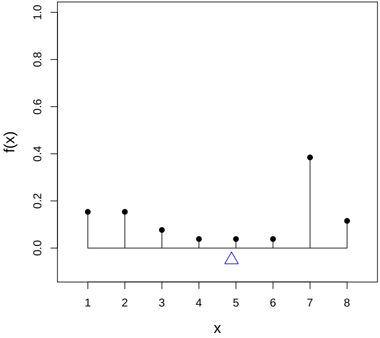
- Joe Biden
- Hillary Clinton
- Chris Dodd
- John Edwards
- Mike Gravel
- Dennis Kucinich
- Barack Obama
- Bill Richardson

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

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

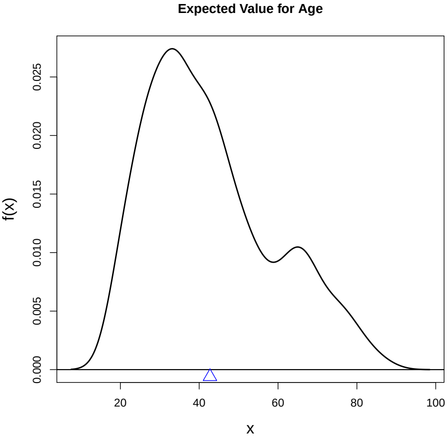
Interpreting Discrete Expected Value

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 $x_1, \dots, x_N$ 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 $x_1, \dots, x_N$ 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 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).

Let X and Y be random variables with finite expectations. Then, $\forall a, b, c \in \mathbb{R}$,

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

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 (analogous for continuous).

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 (analogous for continuous).

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 (analogous for continuous).

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)

LOTUS: Artifical Example, evil dictator

Imagine that you are the advisor to a evil dictator.

LOTUS: Artifical Example, evil dictator

Imagine that you are the advisor to a evil dictator. This dictator wants people to have more kids so that she can have more power.

LOTUS: Artifical Example, evil dictator

Imagine that you are the advisor to a evil dictator. This dictator wants people to have more kids so that she can have more power. As a thank you to her people she decides to give every household money based on the number of kids they have according to this formula:

$$\text{Payout} = (\#\ \text{kids})^2$$

LOTUS: Artifical Example, evil dictator

Imagine that you are the advisor to a evil dictator. This dictator wants people to have more kids so that she can have more power. As a thank you to her people she decides to give every household money based on the number of kids they have according to this formula:

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

- What is $E[X]$?
- What is $g(E[X])$?
- What is $E[g(X)]$?
- What does each mean?

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$
$$E[X] = 0 * 0.3 + 1 * 0.2 + 2 * 0.2 + 3 * 0.3$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$
$$E[X] = 0 * 0.3 + 1 * 0.2 + 2 * 0.2 + 3 * 0.3$$
$$E[X] = 0 + 0.2 + 0.4 + 0.9$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$
$$E[X] = 0 * 0.3 + 1 * 0.2 + 2 * 0.2 + 3 * 0.3$$
$$E[X] = 0 + 0.2 + 0.4 + 0.9$$
$$E[X] = 1.5$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$
$$E[X] = 0 * 0.3 + 1 * 0.2 + 2 * 0.2 + 3 * 0.3$$
$$E[X] = 0 + 0.2 + 0.4 + 0.9$$
$$E[X] = 1.5$$

$$g(E[X]) = 1.5^2 = 2.25$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

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

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

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

$$E[g(X)] = 0^2 P(X = 0) + 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3)$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

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

$$E[g(X)] = 0^2 P(X = 0) + 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3)$$
$$E[g(X)] = 0 * 0.3 + 1 * 0.2 + 4 * 0.2 + 9 * 0.3$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\text{ kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

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

$$E[g(X)] = 0^2 P(X = 0) + 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3)$$
$$E[g(X)] = 0 * 0.3 + 1 * 0.2 + 4 * 0.2 + 9 * 0.3$$
$$E[g(X)] = 0 + 0.2 + 0.8 + 2.7$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X = x)$	0.3	0.2	0.2	0.3

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

$$E[g(X)] = 0^2 P(X = 0) + 1^2 P(X = 1) + 2^2 P(X = 2) + 3^2 P(X = 3)$$
$$E[g(X)] = 0 * 0.3 + 1 * 0.2 + 4 * 0.2 + 9 * 0.3$$
$$E[g(X)] = 0 + 0.2 + 0.8 + 2.7$$
$$E[g(X)] = 3.7$$

LOTUS: Example, evil dictator

$$\text{Payout} = (\#\ \text{kids})^2$$

Distribution of kids per household

x	0	1	2	3
$P(X=x)$	0.3	0.2	0.2	0.3

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

$$E[g(X)] = 0^2 P(X=0) + 1^2 P(X=1) + 2^2 P(X=2) + 3^2 P(X=3)$$
$$E[g(X)] = 0 * 0.3 + 1 * 0.2 + 4 * 0.2 + 9 * 0.3$$
$$E[g(X)] = 0 + 0.2 + 0.8 + 2.7$$
$$E[g(X)] = 3.7$$

Notice that $E[g(x)]$ is like a weighted average of $g(X)$ where the density is the weight.

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

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

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
$p_X(x)$	?	?	?	?
$\widehat{p}_X(x)$	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
$p_X(x)$	?	?	?	?
$\widehat{p}_X(x)$	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
$p_Y(y)$	?	?	?	?	?
$\widehat{p}_Y(y)$	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
$\widehat{p}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{p}_X(x)$	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
$\widehat{p}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{p}_X(x)$	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
$\widehat{p}_Y(y)$	0.03	0.20	0.40	0.28	0.10	1.00
$y\widehat{p}_Y(y)$	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
$\widehat{p}_X(x)$	0.02	0.27	0.43	0.28	1.00
$x\widehat{p}_X(x)$	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
$\widehat{p}_Y(y)$	0.03	0.20	0.40	0.28	0.10	1.00
$y\widehat{p}_Y(y)$	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.

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

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

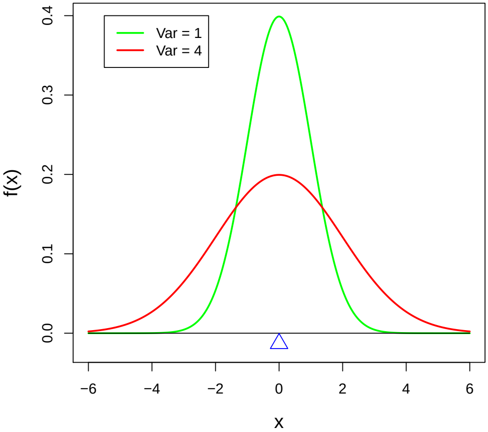
For a discrete random variable X

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

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—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—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—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: 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

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

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

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

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

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

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

$$
\begin{array}{rrl}
 & 4/26 & \times (1 - 4.88)^2 \\
 & 4/26 & \times (2 - 4.88)^2 \\
 & 2/26 & \times (3 - 4.88)^2 \\
 & 1/26 & \times (4 - 4.88)^2 \\
 & 1/26 & \times (5 - 4.88)^2 \\
 & 1/26 & \times (6 - 4.88)^2 \\
 & 10/26 & \times (7 - 4.88)^2 \\
+ & 3/26 & \times (8 - 4.88)^2 \\
\hline
 & & \color{red}{6.79}
\end{array}
$$

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

$$
\begin{array}{rrl}
 & 4/26 & \times (1-4.88)^2 \\
 & 4/26 & \times (2-4.88)^2 \\
 & 2/26 & \times (3-4.88)^2 \\
 & 1/26 & \times (4-4.88)^2 \\
 & 1/26 & \times (5-4.88)^2 \\
 & 1/26 & \times (6-4.88)^2 \\
 & 10/26 & \times (7-4.88)^2 \\
+ & 3/26 & \times (8-4.88)^2 \\
\hline
 & & 6.79
\end{array}
$$

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?

Moments

We can generalize the idea of an expectation to one of moments which are expectations of functions of a random variable X.

Moments

We can generalize the idea of an expectation to one of moments which are expectations of functions of a random variable X.

- Raw moments: The expectation of a higher polynomial order of X. The j^{th} raw moment is $E[X^j]$.
  - The expectation is the first raw moment.
  - Raw moments provide summary information about a distribution, describing elements of its shape and location.

Moments

We can generalize the idea of an expectation to one of moments which are expectations of functions of a random variable X.

- Raw moments: The expectation of a higher polynomial order of X. The j^{th} raw moment is $E[X^j]$.
  - The expectation is the first raw moment.
  - Raw moments provide summary information about a distribution, describing elements of its shape and location.
- Central moments: The expectation of a higher order polynomial of X that has been "centered" on $E[X]$. The j^{th} central moment of X is $E[(X - E[X])^j]$. Sometimes $X - E[X]$ is referred to as a "deviation" from $E[X]$.
  - Generally provides more useful information about the spread and shape of a distribution than the raw moments.

Moments

We can generalize the idea of an expectation to one of moments which are expectations of functions of a random variable X.

- Raw moments: The expectation of a higher polynomial order of X. The j^{th} raw moment is $E[X^j]$.
  - The expectation is the first raw moment.
  - Raw moments provide summary information about a distribution, describing elements of its shape and location.
- Central moments: The expectation of a higher order polynomial of X that has been "centered" on $E[X]$. The j^{th} central moment of X is $E[(X - E[X])^j]$. Sometimes $X - E[X]$ is referred to as a "deviation" from $E[X]$.
  - Generally provides more useful information about the spread and shape of a distribution than the raw moments.
  - Second central moment: variance
  - Third central moment (standardized): skew (asymmetry around the expectation)
  - Fourth central moment (standardized): kurtosis or "tailedness"

Another way to characterize distributions is with their moment-generating function.

Expected Value as Mean Squared Error Minimizer

But why is expectation a good description?

Expected Value as Mean Squared Error Minimizer

But why is expectation a good description?

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

But why is expectation a good description?

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

But why is expectation a good description?

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

But why is expectation a good description?

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

But why is expectation a good description?

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:

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

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:

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

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:

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

Indicator Random Variables

Indicator Random Variables

An indicator random variable I_A or $I(A)$ for an event A is defined to be 1 if A occurs and 0 otherwise.

Indicator Random Variables

An indicator random variable I_A or $I(A)$ for an event A is defined to be 1 if A occurs and 0 otherwise.

Definition (Indicator Random Variables)

Let A and B be events. Then the following properties hold:

1. $(I_A)^k = I_A$ for any positive integer k.
2. $I_{A^c} = 1 - I_A$
3. $I_{A \cap B} = I_A I_B$
4. $I_{A \cup B} = I_A + I_B - I_A I_B$

The Fundamental Bridge

The Fundamental Bridge

Theorem (Fundamental bridge between probability and expectation)

There is a one-to-one correspondence between events and indicator random variables, and the probability of an event A is the expected value of its indicator random variable I_A:

$$P(A) = E(I_A)$$

(Blitzstein and Hwang Theorem 4.4.2)

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).

Going Deeper:

Blitzstein, Joseph K., and Hwang, Jessica. (2019). *Introduction to Probability*. CRC Press. `http://stat110.net/`

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).

Going Deeper:

Blitzstein, Joseph K., and Hwang, Jessica. (2019). *Introduction to Probability*. CRC Press. `http://stat110.net/`

Next:

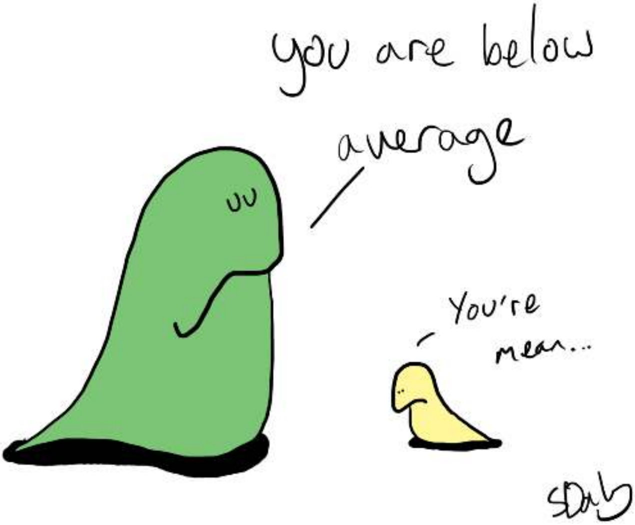
- properties of joint and conditional distributions
- famous distributions

Fun with

Fun with Averages

Fun with Averages

Central Tendency



The Story of Averages

Measurements

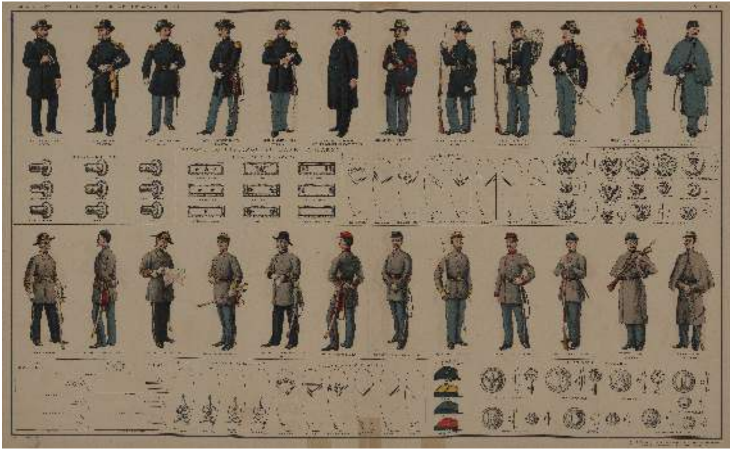
MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE D'OBSERVATIONS calculé.
Pouces.							
33	3	5	0,5000			0,5000	7
34	18	31	0,4995	52	50	0,4993	29
35	81	141	0,4964	42,5	42,5	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

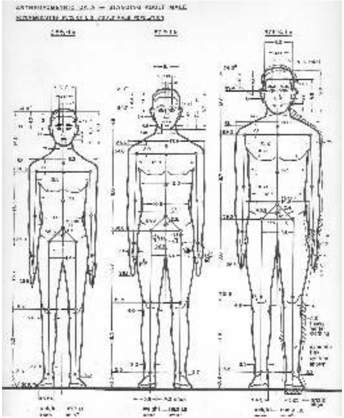
The determination of the average man is not merely a matter of speculative curiosity; it may be of the most important service to the science of man and the social system. It ought necessarily to precede every other inquiry into social physics, since it is, as it were, the basis. The average man, indeed, 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, A treatise on man and the development of his faculties, 1842.

The Military Takes to the Idea



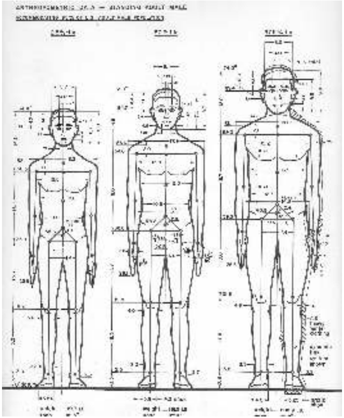
The Problem with Averages

The Average Man



Gilbert Daniels, 1950: On 10 dimensions that matter most for design of planes, how many are average on all 10 dimensions?

The Average Man



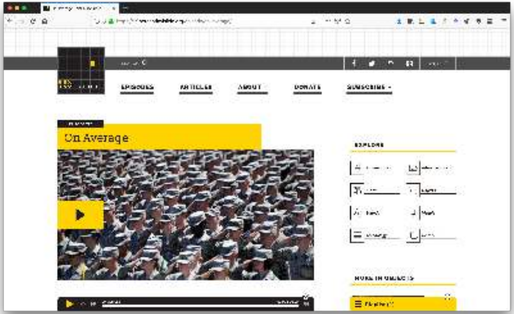
Gilbert Daniels, 1950: On 10 dimensions that matter most for design of planes, how many are average on all 10 dimensions? 0 of 4,063
One step deeper: high-dimensional statistics breaks lots of intuitions.
`https://arbital.com/p/5n1/`

Here's a pilot who is not average

Kim Campbell won the Distinguished Flying Cross for her heroics in Bagdad. She is not average.

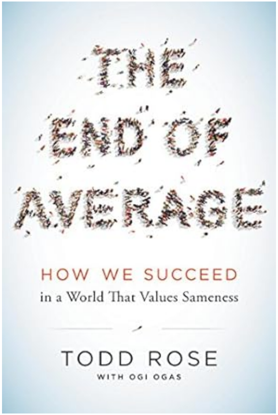
https://commons.wikimedia.org/wiki/File:Kim_campbell_a10.jpg

On averages

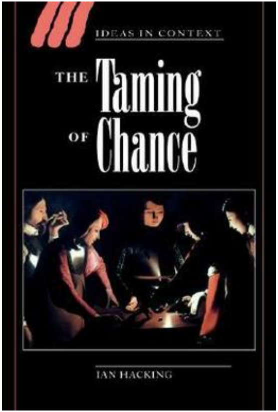


https://99percentinvisible.org/episode/on-average/

On averages



Fit creates opportunity



Intellectual history of averages

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
  - initial introduction to random variables
- This Week

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

- Last Week
  - welcome and outline of course
  - described uncertain outcomes with probability
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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
  - initial introduction to random variables
- This Week
  - 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 with regression

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.

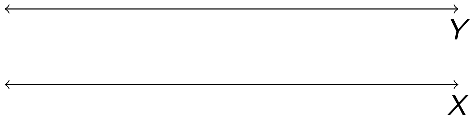
Understanding Joint Distributions Mathematically

Understanding Joint Distributions Mathematically

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

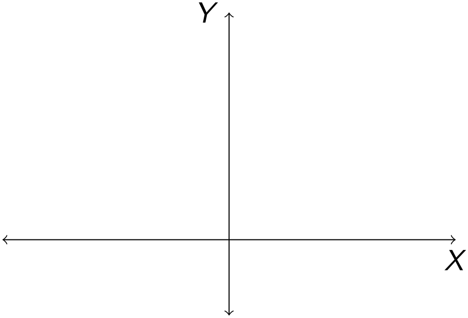
Understanding Joint Distributions Mathematically

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



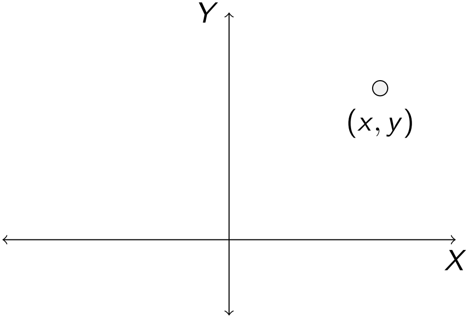
Understanding Joint Distributions Mathematically

- 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

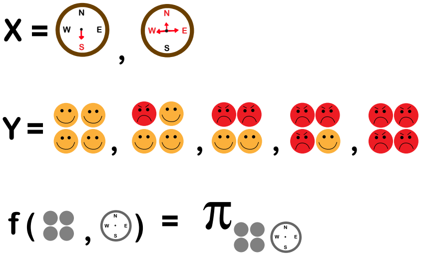
- Recall the list experiment about racial prejudice. Suppose we define $X = 0$ (Non-southern), 1 (Southern) and $Y =$ "number of angering items" for a randomly selected respondent receiving the treatment list.

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.

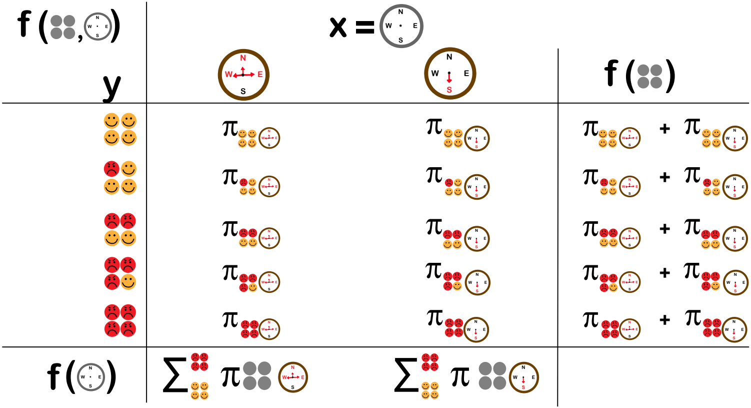
Example Joint Distribution: Binary X, Discrete Y

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}	$\pi_{00} + \pi_{01}$
$f_X(x)$	$\sum_{y=0}^{4} \pi_{y0}$	$\sum_{y=0}^{4} \pi_{y1}$	

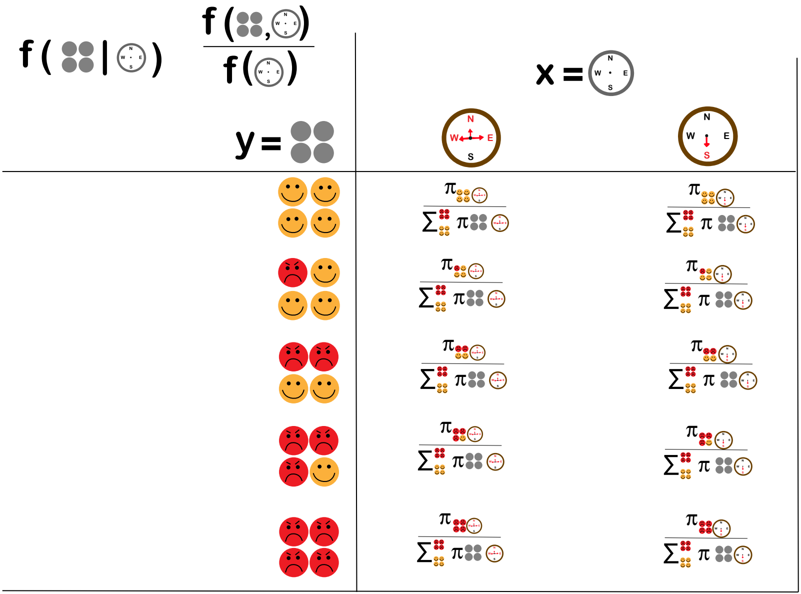
Example Joint Distribution: Binary X, Discrete Y

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

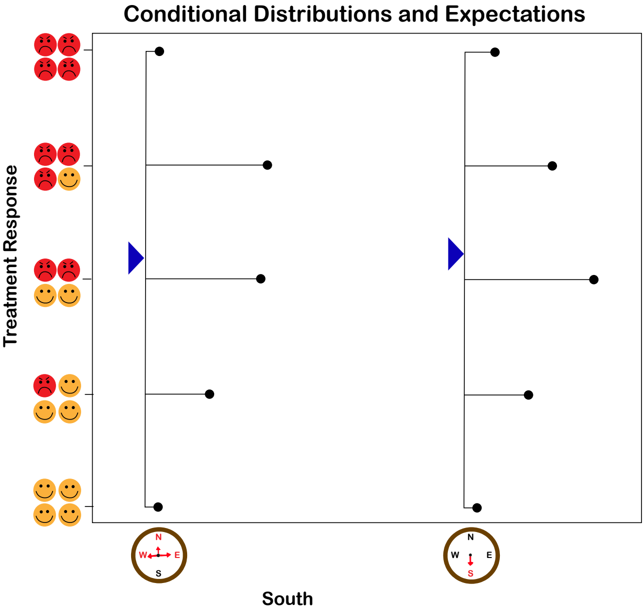


Discrete Conditional Distribution

Discrete Conditional Distribution

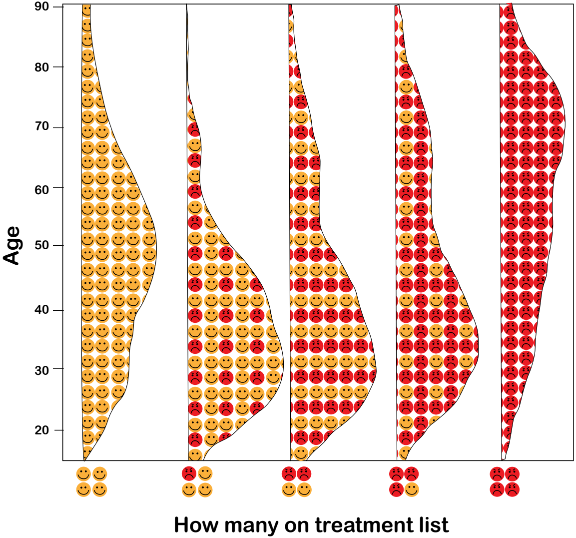


Discrete Conditional Distribution

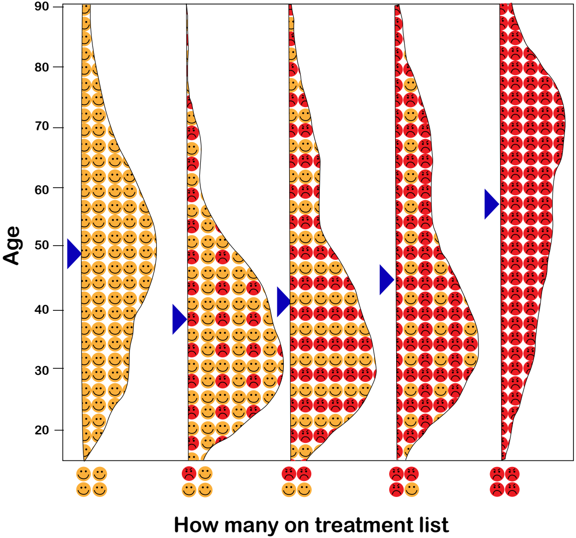


Example: Continuous Conditional Distribution

Example: Continuous Conditional Distribution



Conditional Expectation Function (coming soon!)



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	0.03	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:

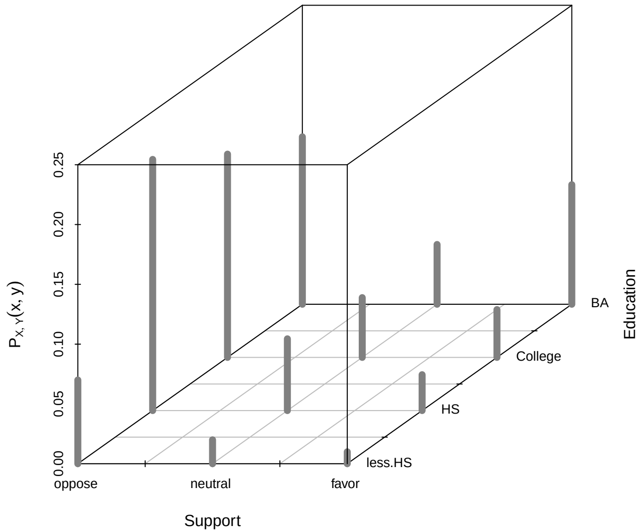
$$p_{X,Y}(x,y) \;=\; P(X = x \text{ and } Y = y)$$

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	0.03	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

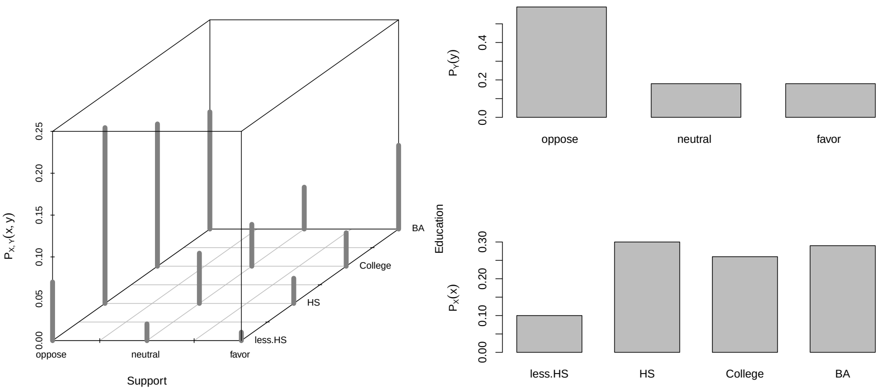
Given the joint PMF $p_{X,Y}(x,y)$ can we recover the marginal PMF $p_Y(y)$ (distribution over a single variable)?

		X: Education				
		less HS	HS	College	BA	$p_Y(y)$
Y: Support	oppose	0.07	0.21	0.17	0.14	0.62
	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	0.03	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	0.03	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

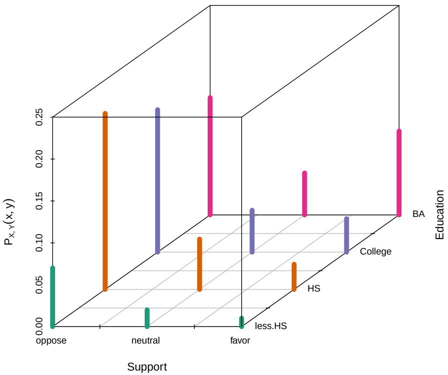


Figure: Joint

Joint and Conditional Probability Mass Functions

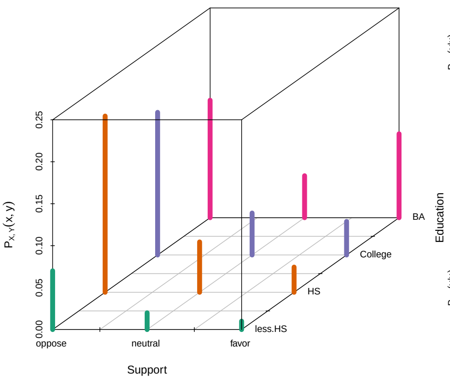


Figure: Joint

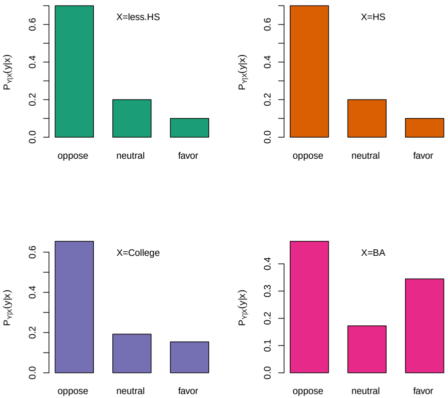


Figure: Conditional

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:

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

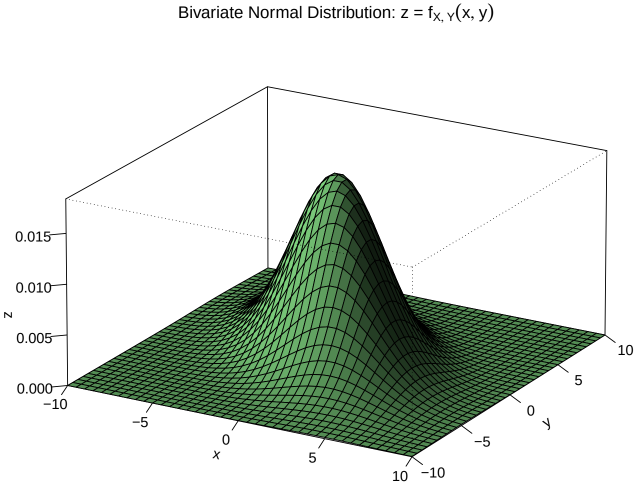
where

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

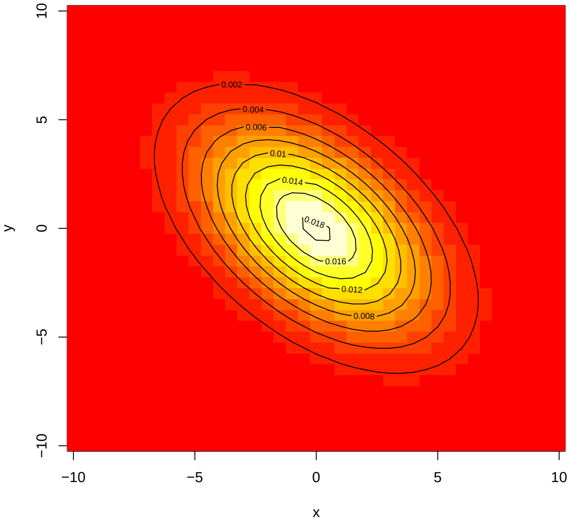
Restrictions:

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

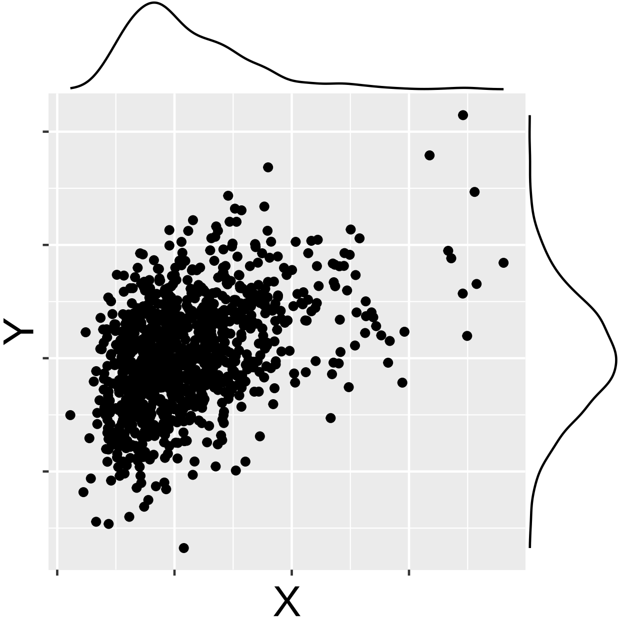
3D Plot of a Joint Probability Density Function



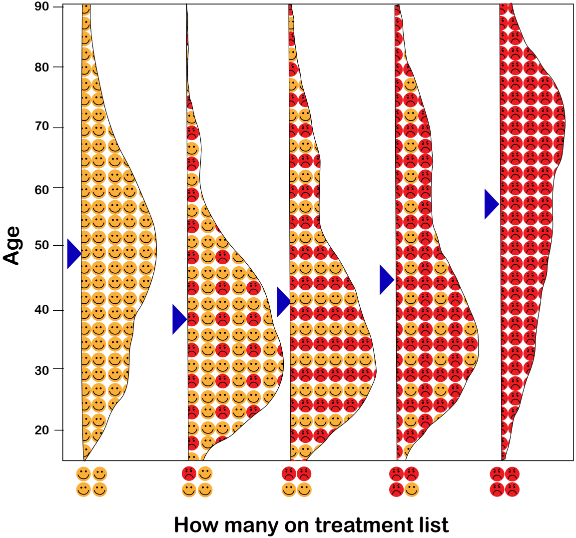
Contour Plot of a Joint Probability Density Function



From Joint to Marginal PDF



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:

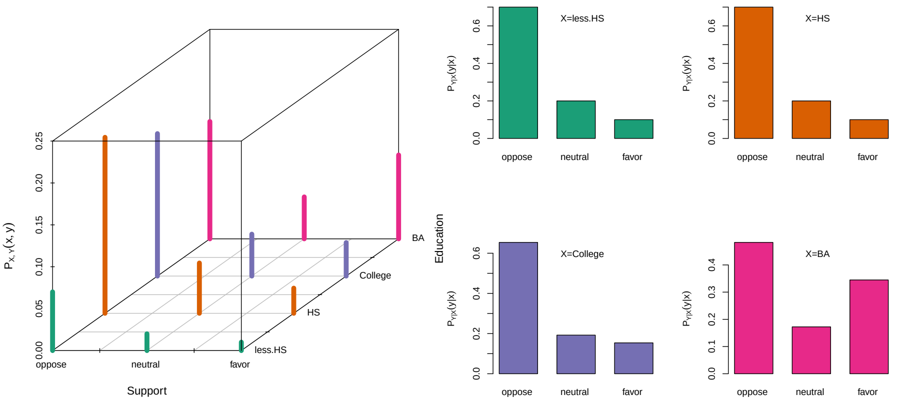
$$E[Y|X = x] = \sum_y y\, P(Y = y|X = x) = \sum_y y\, p_{Y|X}(y|x)$$

Definition (Conditional Expectation (Continuous))

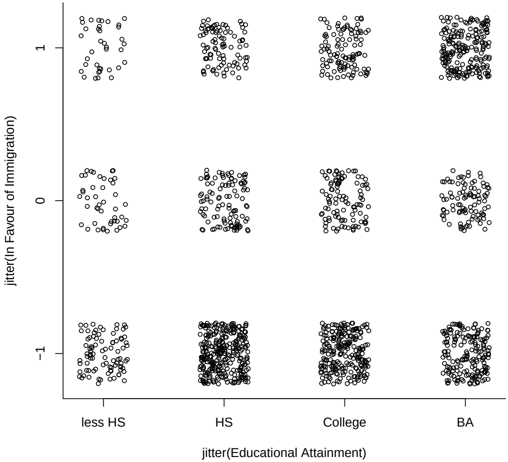
Let Y and X be continuous random variables. The conditional expectation of Y given $X = x$ is given by:

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

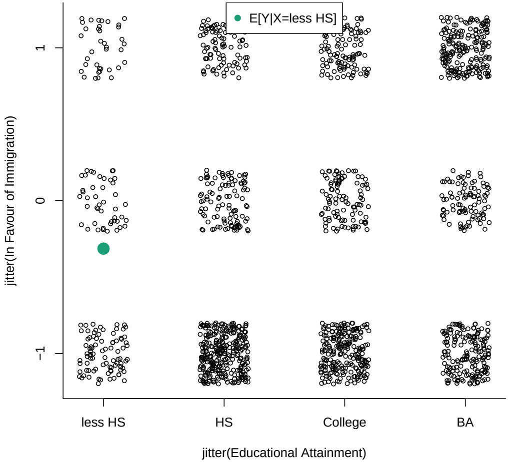
Joint and Conditional Probability Mass Functions



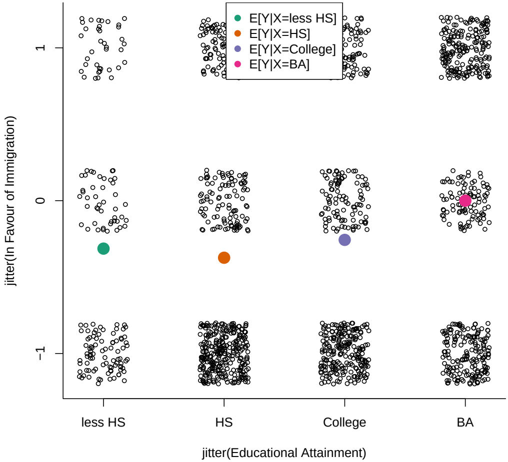
Conditional PMF $P_{Y|X}(y|x)$



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,

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

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)

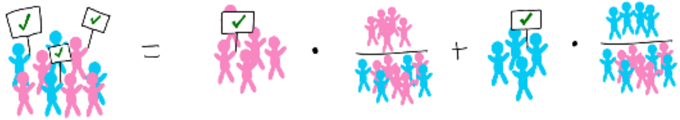
For two random variables X and Y,

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

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 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):

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

Conditional Variance

Definition

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

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

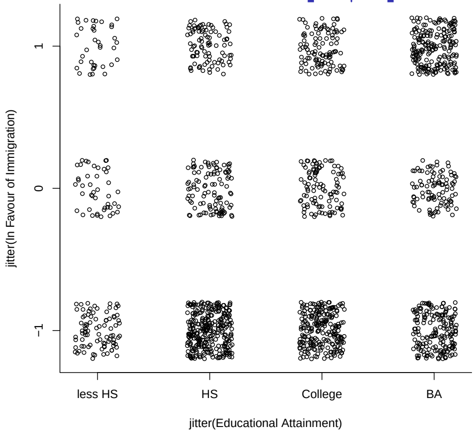
A useful related result is the law of total variance (Eve's Law):

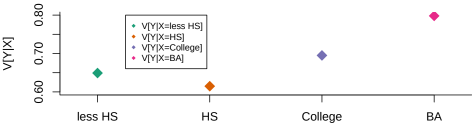
$$\underbrace{V[Y]}_{\text{Total variance}} \;=\; \underbrace{E[V[Y|X]]}_{\text{Average of Group Variances}} \;+\; \underbrace{V[E[Y|X]]}_{\text{Variance in Group Averages}}$$

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 (perhaps frustratingly we still call it an operator and use $E[\cdot|\cdot]$).

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 (perhaps frustratingly we still call it an operator and use $E[\cdot|\cdot]$).
- $E[X|Y]$ 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 (perhaps frustratingly we still call it an operator and use $E[\cdot|\cdot]$).
- $E[X|Y]$ 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)

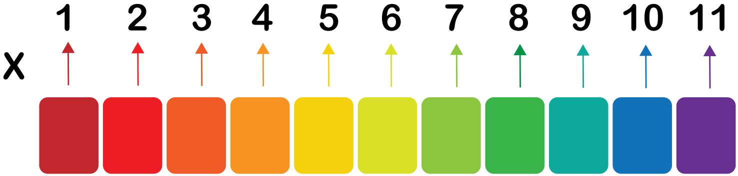
Important Subtleties in Pictures

Sample space

Important Subtleties in Pictures

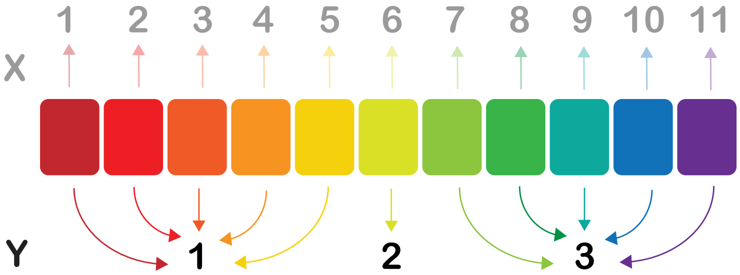
Sample space

Important Subtleties in Pictures



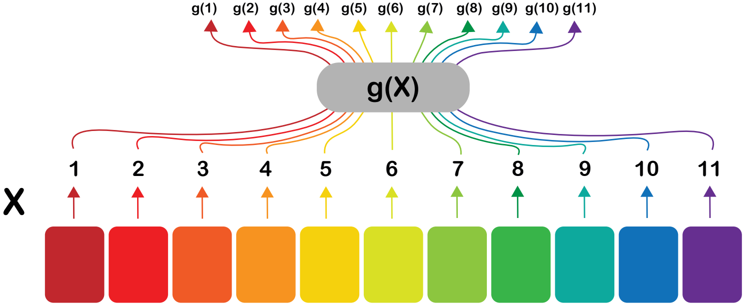
Random variable

Important Subtleties in Pictures



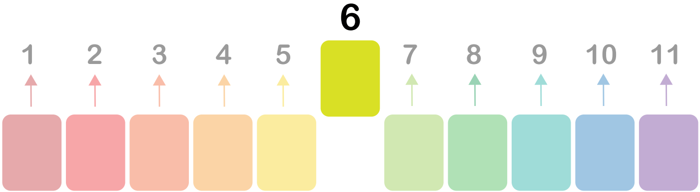
Random variable

Important Subtleties in Pictures



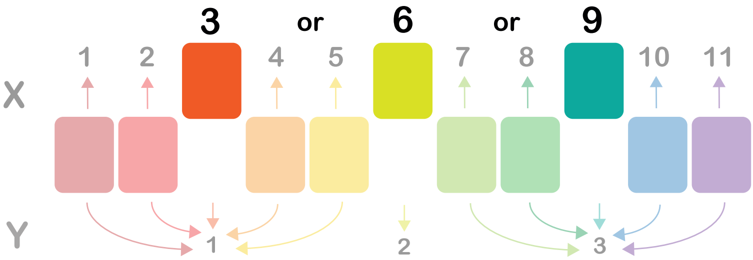
Function of a random variable is a random variable

Important Subtleties in Pictures



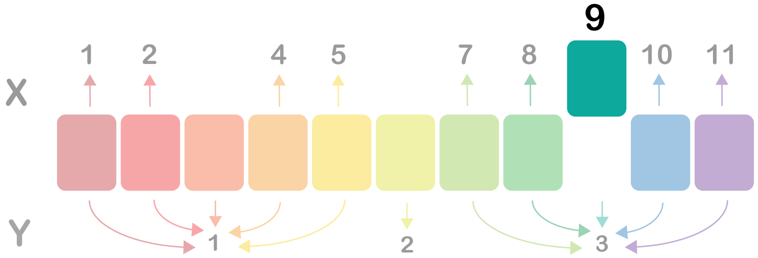
$E[X]$

Important Subtleties in Pictures



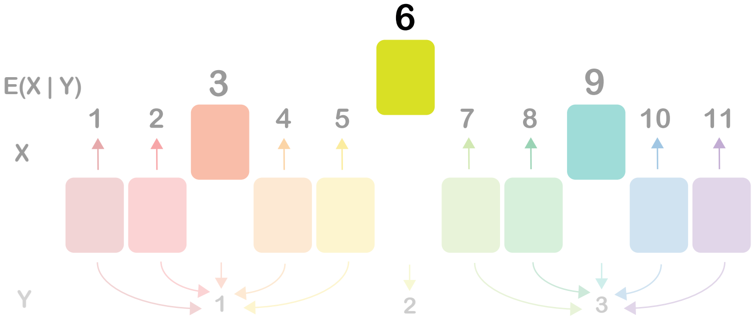
$$E[X|Y]$$

Important Subtleties in Pictures



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

Important Subtleties in Pictures



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

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 $Y \perp\!\!\!\perp X$.

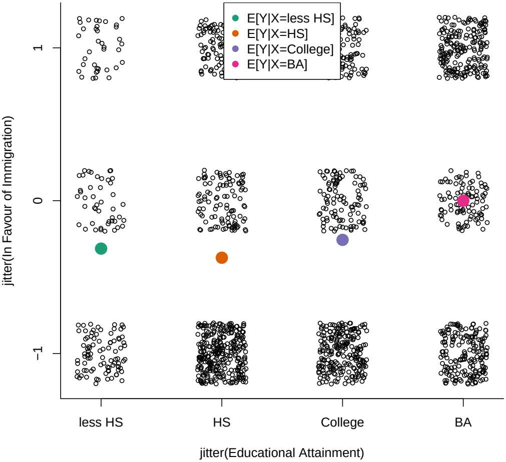
Independence implies

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

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

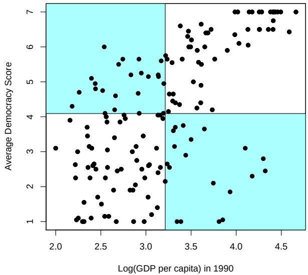
Definition

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)

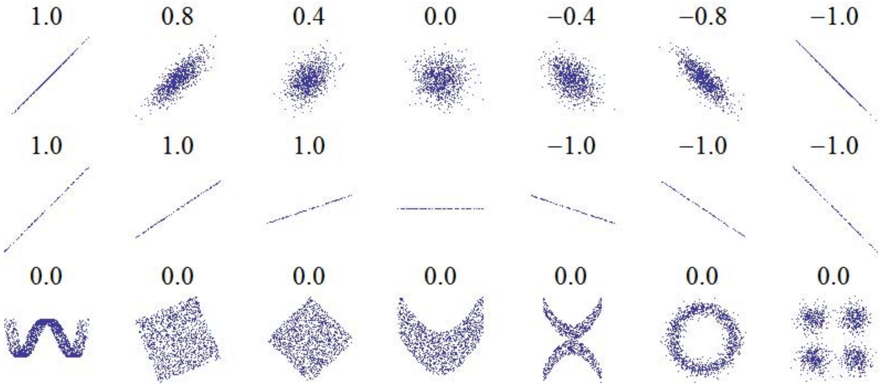
The correlation between two random variables X and Y is defined as

$$Cor[X, Y] = \frac{Cov[X, Y]}{\sqrt{V[X]V[Y]}} = \frac{Cov[X, Y]}{SD[X]SD[Y]}.$$

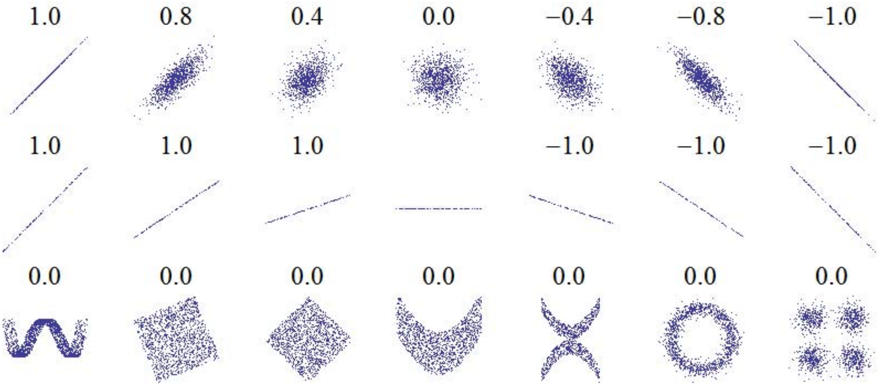
- $Cor[X, Y]$ is a standardized measure of linear association between X and Y.
- Always satisfies: $-1 \leq Cor[X, Y] \leq 1$.

Correlation is *Linear*

Correlation is *Linear*

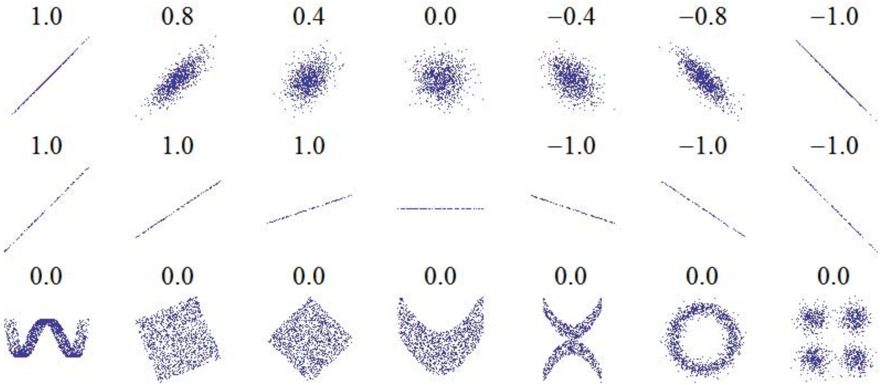


Correlation is *Linear*



- $Cor[X, Y] = \pm 1$ 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

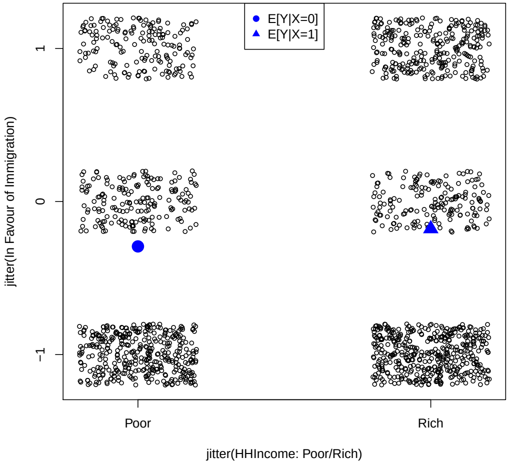
$$f_{Y|X,Z}(y \mid x, z) \; = \; f_{Y|Z}(y \mid z)$$

- 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

$$E[Y|X = x, Z = z] = E[Y|Z = z].$$

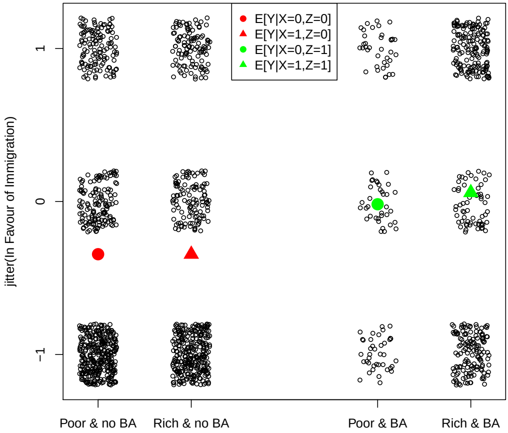
Is $Y \perp\!\!\!\perp X$?

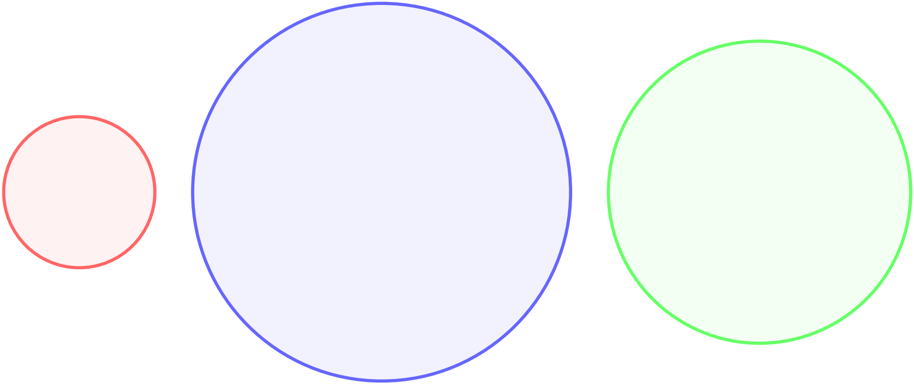
Example: X = wealth, Y = support for immigration, Z = education.



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

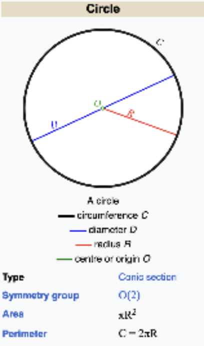
Example: X = wealth, Y = support for immigration, Z = education.

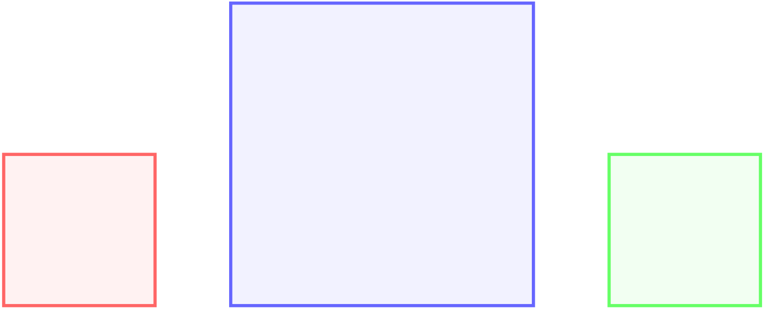




These circles are all different, but they are from the same “family”

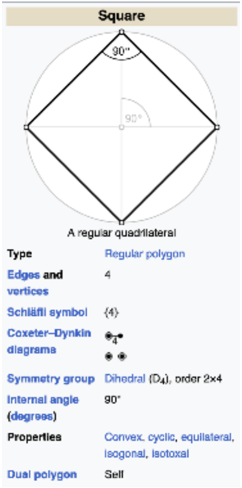
- The family of circles is defined by one parameter: usually the radius
- If something is a circle with radius r, then you know other things about it

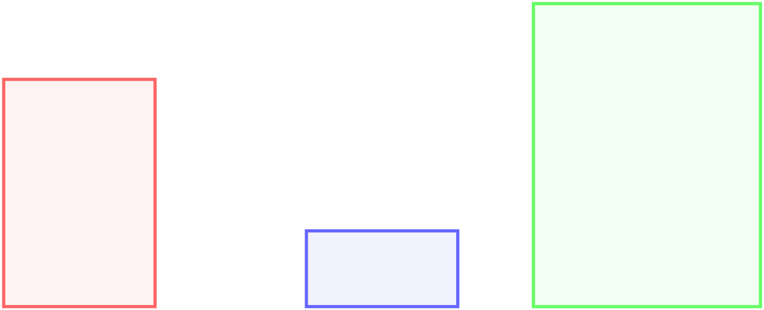


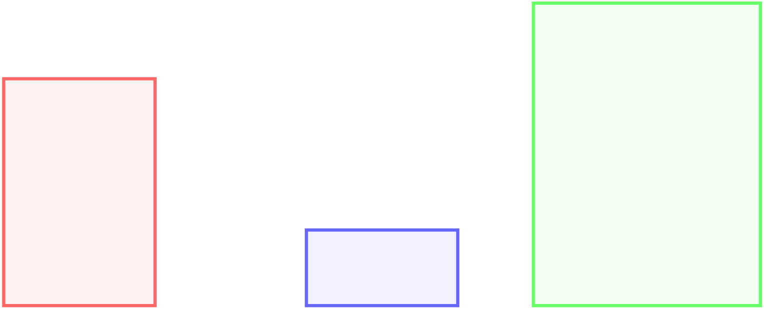


These squares are all different, but they are from the same “family”.

- The family of squares is defined by one parameter: usually the length
- If something is a square with length l, then you know other things about it

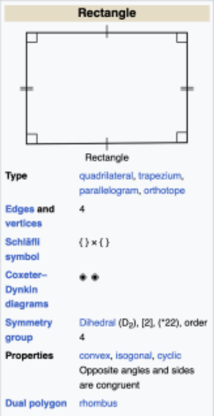






These rectangles are all different, but they are from the same “family”.

- The family of rectangle is defined by two parameters: length and width
- If something is a rectangle with length l and width w then you know other things about it



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,

$$P(X = x) \;=\; \pi^x (1 - \pi)^{1-x}$$

for $x \in \{0, 1\}$ and $P(X = x) = 0$ otherwise.
We will (equivalently) say that

$$X \;\sim\; \text{Bernoulli}(\pi)$$

$\sim$ 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

$$f_X(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Then X is a normally distributed random variable with parameters μ and σ^2.
Equivalently, we'll write

$$X \quad \sim \quad \text{Normal}(\mu, \sigma^2)$$

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

$$Z \sim \text{Normal}(0, 1)$$

We'll call the cumulative distribution function of Z,

$$F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2/2)dz$$

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

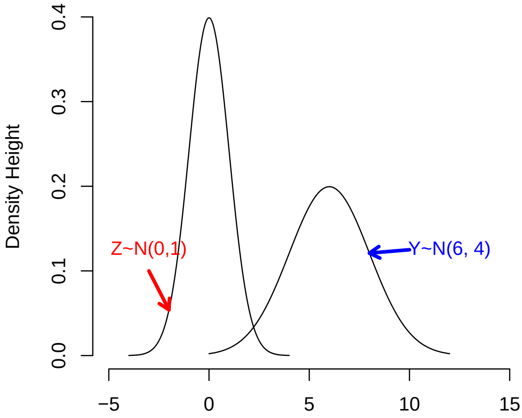
Suppose $Z \sim \text{Normal}(0, 1)$.

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

$$f_{X_1,X_2}(\boldsymbol{x}) \quad = \quad (2\pi)^{-N/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

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

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-2/2} 1^{-1/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{x} - \mathbf{0})' \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (\boldsymbol{x} - \mathbf{0})\right)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 + x_2^2)\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_1^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2}\right) \end{aligned}$$

$\rightsquigarrow$ 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,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \ldots & \text{cov}(X_1, X_N) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \ldots & \text{cov}(X_2, X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_N, X_1) & \text{cov}(X_N, X_2) & \ldots & \text{var}(X_N) \end{pmatrix}$$

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,

$$
\boldsymbol{\Sigma} = \begin{pmatrix}
\text{var}(X_1) & \text{cov}(X_1, X_2) & \ldots & \text{cov}(X_1, X_N) \\
\text{cov}(X_2, X_1) & \text{var}(X_2) & \ldots & \text{cov}(X_2, X_N) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(X_N, X_1) & \text{cov}(X_N, X_2) & \ldots & \text{var}(X_N)
\end{pmatrix}
$$

A nice extra property: the marginals are normal!

Universality of the Uniform

Universality of the Uniform

Theorem (Universality of the Uniform)

- *Regardless of the distribution of* X*,* $F(X) \sim \textit{Uniform}(0, 1)$
- *For a random variable* X *with CDF* F *and a Uniform random variable* U*,* $F^{-1}(U) \sim X$

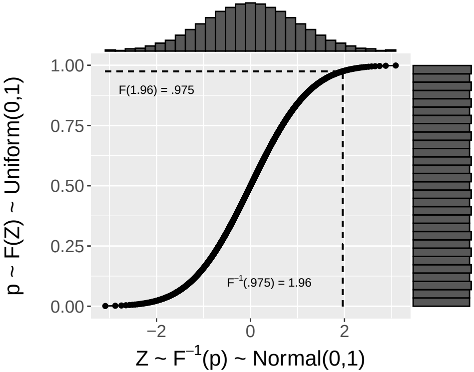
Blitzstein and Hwang Theorem 5.3.1

Universality of the Uniform

Theorem (Universality of the Uniform)

- *Regardless of the distribution of X, $F(X) \sim Uniform(0,1)$*
- *For a random variable X with CDF F and a Uniform random variable U, $F^{-1}(U) \sim X$*

Blitzstein and Hwang Theorem 5.3.1

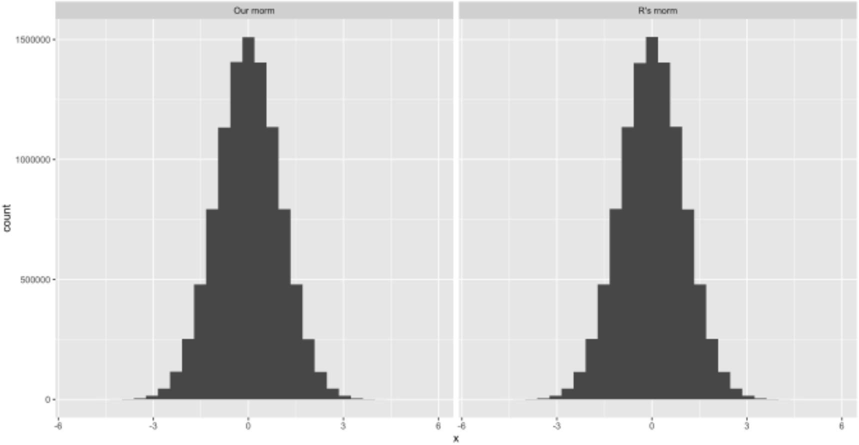


Self-written `rnorm`

Self-written `rnorm`

```
draw.rnorm <- function(n, mean = 0, sd = 1) {
  u <- runif(n)
  z <- qnorm(u, mean=mean, sd=sd)
  return(z)
}
```

Did we succeed?



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!

Fun with

Fun with Connections

Fun with Connections

Exponential: $-\log$ Uniform

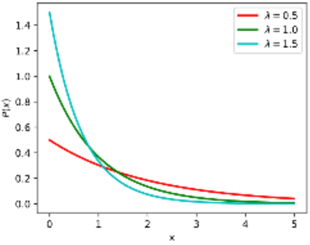
Figure credit: Wikipedia

The Uniform distribution is defined on the interval $(0, 1)$. Suppose we wanted a distribution defined on all positive numbers.

Definition

X follows an **exponential** distribution with rate parameter λ if

$$X \sim -\frac{1}{\lambda} log(U)$$



Exponential: $-\log$ Uniform

The exponential is often used for wait times. For instance, if you're waiting for shooting stars, the time until a star comes might be exponentially distributed.
Key properties:

- Memorylessness: Expected remaining wait time does not depend on the time that has passed
- $E(X) = \frac{1}{\lambda}$
- $V(X) = \frac{1}{\lambda^2}$

Exponential-uniform connection

Suppose $X_1, X_2 \overset{iid}{\sim} Expo(\lambda)$. What is the distribution of $\frac{X_1}{X_1+X_2}$?

Exponential-uniform connection

Suppose $X_1, X_2 \overset{iid}{\sim} Expo(\lambda)$. What is the distribution of $\frac{X_1}{X_1+X_2}$? The proportion of the wait time that is represented by X_1 is uniformly distributed over the interval, so

$$\frac{X_1}{X_1 + X_2} \sim Uniform(0, 1)$$

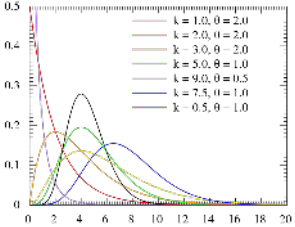
Gamma: Sum of independent Exponentials

Figure credit: Wikipedia

Definition

Suppose we are waiting for a shooting stars, with the time between stars $X_1, \ldots, X_a \overset{iid}{\sim} Expo(\lambda)$. The distribution of time until the ath shooting star is

$$G \sim \sum_{i=1}^{a} X_i \sim Gamma(a, \lambda)$$



Gamma: Properties

Properties of the *Gamma*(a, λ) distribution include:

- $E(G) = \frac{a}{\lambda}$
- $V(G) = \frac{a}{\lambda^2}$

Beta: Uniform order statistics

Suppose we draw $U_1, \ldots, U_k \sim \mathit{Uniform}(0,1)$, and we want to know the distribution of the jth order statistic, $U_{(j)}$. Using the Uniform-Exponential connection, we could also think of these $U_{(j)}$ as being the location of the jth Exponential in a series of $k+1$ Exponentials. Thus,

$$U_{(j)} \sim \frac{\sum_{i=1}^{j} X_i}{\sum_{i=1}^{j} X_i + \sum_{i=j+1}^{k+1} X_i} \sim \mathit{Beta}(j, k-j+1)$$

This defines the **Beta distribution**.
Can we name the distribution at the top of the fraction?

Beta: Uniform order statistics

Suppose we draw $U_1, \dots, U_k \sim Uniform(0,1)$, and we want to know the distribution of the jth order statistic, $U_{(j)}$. Using the Uniform-Exponential connection, we could also think of these $U_{(j)}$ as being the location of the jth Exponential in a series of $k+1$ Exponentials. Thus,

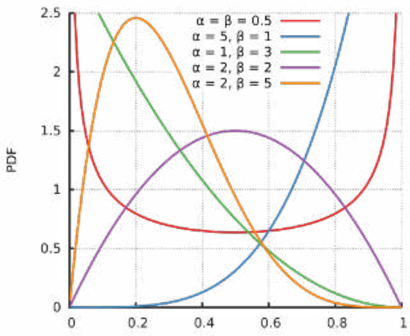
$$U_{(j)} \sim \frac{\sum_{i=1}^{j} X_i}{\sum_{i=1}^{j} X_i + \sum_{i=j+1}^{k+1} X_i} \sim Beta(j, k-j+1)$$

This defines the **Beta distribution**.
Can we name the distribution at the top of the fraction?

$$\sim \frac{G_j}{G_j + G_{k-j+1}}$$

What do Betas look like?

Figure credit: Wikipedia



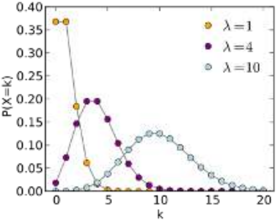
Poisson: Number of Exponential events in a time interval

Figure credit: Wikipedia

Definition

Suppose the time between shooting stars is distributed $X \sim Expo(\lambda)$. Then, the number of shooting stars in an interval of time t is distributed

$$Y_t \sim Poisson(\lambda t)$$



Poisson: Number of Exponential events in a time interval

Properties of the Poisson:

- If $Y \sim Pois(\lambda t)$, then $V(Y) = E(Y) = \lambda t$
- Number of events in disjoint intervals are independent

χ_n^2: A particular Gamma

Definition

We define the **chi-squared distribution** with n degrees of freedom as

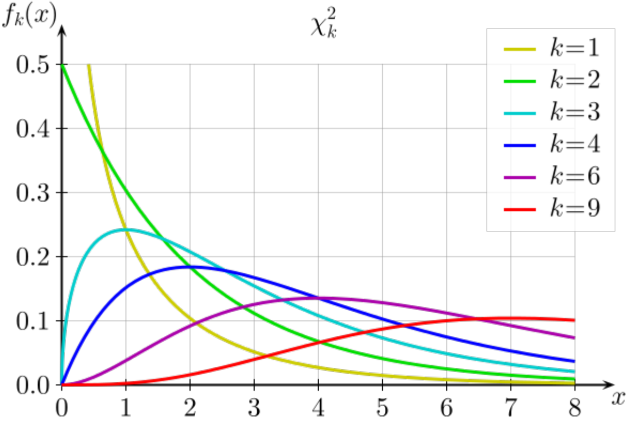
$$\chi_n^2 \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

More commonly, we think of it as the sum of a series of independent squared Normals, $Z_1, \ldots, Z_n \overset{iid}{\sim} Normal(0, 1)$:

$$\chi_n^2 \sim \sum_{i=1}^{n} Z_i^2$$

χ^2_n: A particular Gamma

Figure credit: Wikipedia

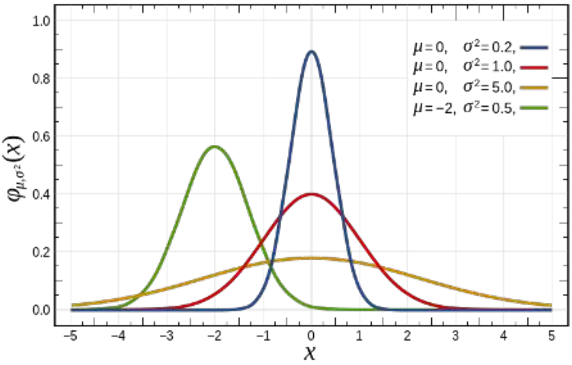


Normal: Square root of χ^2_1 with a random sign

Figure credit: Wikipedia

Definition

Z follows a **Normal** distribution if $Z \sim S\sqrt{\chi^2_1}$, where S is a random sign with equal probability of being 1 or -1.



Normal: An alternate construction

Note: This is far above and beyond what you need to understand for the course!

Box-Muller Representation of the Normal

Let $U_1, U_2 \overset{iid}{\sim} Uniform$. Then

$$Z_1 \equiv \sqrt{-2\log U_2}\cos(2\pi U_1)$$

$$Z_2 \equiv \sqrt{-2\log U_2}\sin(2\pi U_1)$$

so that $Z_1, Z_2 \overset{iid}{\sim} N(0,1)$

What is this?

- **Inside the square root:** $-2\log U_2 \sim Expo\left(\frac{1}{2}\right) \sim Gamma\left(\frac{2}{2}, \frac{1}{2}\right) \sim \chi^2_2$
- **Inside the cosine and sine:** $2\pi U_1$ is a uniformly distributed angle in polar coordinates, and the cos and sin convert this to the cartesian x and y components, respectively.
- **Altogether in polar coordinates:** The square root part is like a radius distributed $\chi \sim |Z|$, and the second part is an angle.

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):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{y\log\mu - \exp(\log\mu) - \log y!\right\}$$

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):

$$P(Y_i = y \mid \mu) \;=\; \exp\left\{y \log\mu - \exp(\log\mu) - \log y!\right\}$$

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):

$$P(Y_i = y \mid \mu) \;=\; \exp\{y\log\mu - \exp(\log\mu) - \log y!\}$$

$\implies \theta = \log\mu$, $\phi = 1$, $a(\phi) = \phi, b(\theta) = \exp(\theta)$, 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 $\mathbb{V}(Y_i) = \frac{d^2 b(\theta_i)}{d\theta_i^2} = \exp(\theta_i) = \mu_i$