Chapter 2: Probability and Statistics

Jonathan Roth

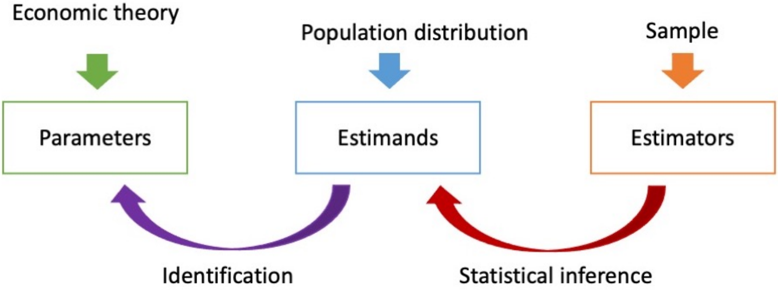
Mathematical Econometrics I
Brown University

Course Logistics

- Problem set 1 is posted. It is due on Friday September 19 at 4PM as a GradeScope submission
- TA sessions and OHs start this week. See the Canvas page for times/tentative locations
- Any logistical questions?

“Big Picture” Recap

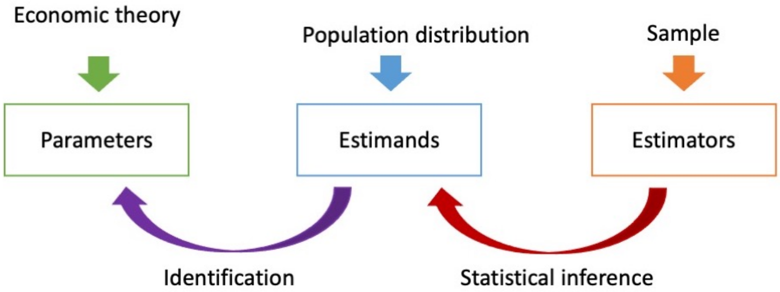
Recall our division of labor:



- **Statistics**: how does the sample data we observe relate to observable features of the population we're interested in?
- **Identification**: how do observable features of population relate to target parameters of interest?

"Big Picture" Recap

Recall our division of labor:



- **Statistics**: how does the sample data we observe relate to observable features of the population we're interested in?
- **Identification**: how do observable features of population relate to target parameters of interest?
- For both these tasks, we need a mathematical language for talking about how data is generated. Enter **probability and statistics**


I love
Statis- sticks

Outline

1. Random Variables and Probability Distributions
2. Means and Variances
3. Identification in Experiments
4. Random Sampling and Sample Means
5. Hypothesis Testing and Inference

Random Variables

- Probability theory formalizes the study of **random processes**

Random Variables

- Probability theory formalizes the study of **random processes**
- What are some examples of a random process?

Random Variables

- Probability theory formalizes the study of **random processes**
- What are some examples of a random process?
  - Flip a coin – is it heads or tails?
  - Survey a random household in US – what is their income?

Random Variables

- Probability theory formalizes the study of **random processes**
- What are some examples of a random process?
  - Flip a coin – is it heads or tails?
  - Survey a random household in US – what is their income?
- The realization of a random process is called a **random variable**.

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. “heads” and “tails” are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)
- The **sample space** is the set of all possible outcomes

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)
- The **sample space** is the set of all possible outcomes
- An **event** is a subset of the sample space; its probability is the sum of probabilities of the included outcomes

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)
- The **sample space** is the set of all possible outcomes
- An **event** is a subset of the sample space; its probability is the sum of probabilities of the included outcomes

Example: I toss two fair coins in the air

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)
- The **sample space** is the set of all possible outcomes
- An **event** is a subset of the sample space; its probability is the sum of probabilities of the included outcomes

Example: I toss two fair coins in the air

- What are the possible outcomes (sample space)?

Some Terminology

- **Outcomes** are mutually exclusive results of a random process (e.g. "heads" and "tails" are the outcomes of a single coin toss)
- The **probability** of an outcome captures its likelihood of occurring (i.e. frequency of its occurrence in repeated runs of the process)
- The **sample space** is the set of all possible outcomes
- An **event** is a subset of the sample space; its probability is the sum of probabilities of the included outcomes

Example: I toss two fair coins in the air

- What are the possible outcomes (sample space)?
- What is the probability of seeing at least one head (an event)?

Random Variables and CDFs

- A **random variable** is a numerical summary of a random process (formally, a real-valued function defined on the sample space)
  - E.g. X counts the number of heads we see

Random Variables and CDFs

- A **random variable** is a numerical summary of a random process (formally, a real-valued function defined on the sample space)
  - E.g. X counts the number of heads we see

- A real-valued random variable is characterized by its **cumulative distribution function** (CDF),

$$F(x) = Pr(X \leq x)$$

which tells us the probabilty that X is some value x or below

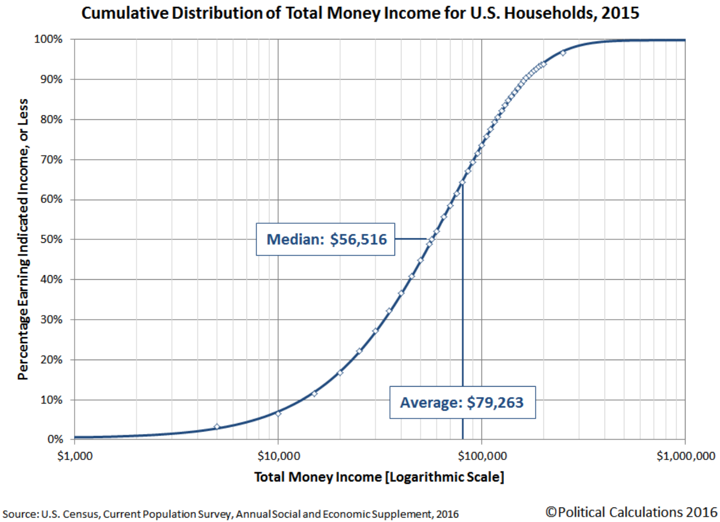
Random Variables and CDFs

- A **random variable** is a numerical summary of a random process (formally, a real-valued function defined on the sample space)
    - E.g. X counts the number of heads we see
- A real-valued random variable is characterized by its **cumulative distribution function** (CDF),

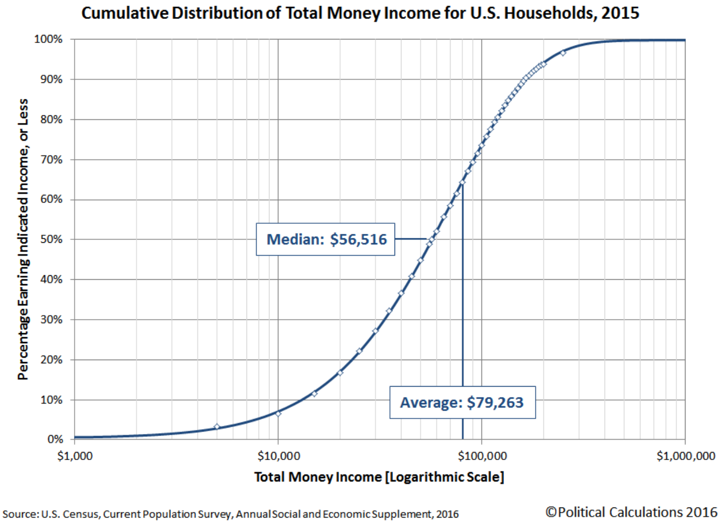
$$F(x) = Pr(X \leq x)$$

  which tells us the probabilty that X is some value x or below

    - Note: we'll typically use lower-case letters like "x" to denote realizations (i.e. non-random numbers) of random variables like "X" ...



- In 2016, around half of US households earned $56K or less



- In 2016, around half of US households earned \$56K or less
- Formally: $F(56,516) = 0.5$

We All Need Some Support...

- The **support** of a random variable X, denoted $\mathbb{X}$, is the set of values that X can take
  - If X is months in the year you were employed, $\mathbb{X} = \{0, 1, ..., 12\}$
  - If X is your income, then $\mathbb{X} = \mathbb{R}_{\geq 0}$ (approximately)

We All Need Some Support...

- The **support** of a random variable X, denoted $\mathbb{X}$, is the set of values that X can take
  - If X is months in the year you were employed, $\mathbb{X} = \{0, 1, ..., 12\}$
  - If X is your income, then $\mathbb{X} = \mathbb{R}_{\geq 0}$ (approximately)
- If the support of X is finite (e.g. $\{0, 1\}$), we say X is **discrete**
- If the support of X is a continuum (e.g. $\mathbb{R}$ or $[0, 1]$), we say X is **continuously distributed** (technically, if the CDF is differentiable)

Density and Mass Functions

- If X is discrete, we define the probability mass function (PMF) as the probability that X takes on each value in the support:

$$p(x) = Pr(X = x)$$

Density and Mass Functions

- If X is discrete, we define the probability mass function (PMF) as the probability that X takes on each value in the support:

$$p(x) = Pr(X = x)$$

  - The CDF of a discrete random variable is then

$$F(x) = \sum_{x' \leq x} p(x')$$

Density and Mass Functions

- If X is discrete, we define the probability mass function (PMF) as the probability that X takes on each value in the support:

$$p(x) = Pr(X = x)$$

  - The CDF of a discrete random variable is then

$$F(x) = \sum_{x' \leq x} p(x')$$

- For a continuous random variable, we define the probability density function (PDF) as $f(x) = \frac{d}{dx}F(x)$, implying a CDF of

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Density and Mass Functions

- If X is discrete, we define the probability mass function (PMF) as the probability that X takes on each value in the support:

$$p(x) = Pr(X = x)$$

  - The CDF of a discrete random variable is then

$$F(x) = \sum_{x' \leq x} p(x')$$

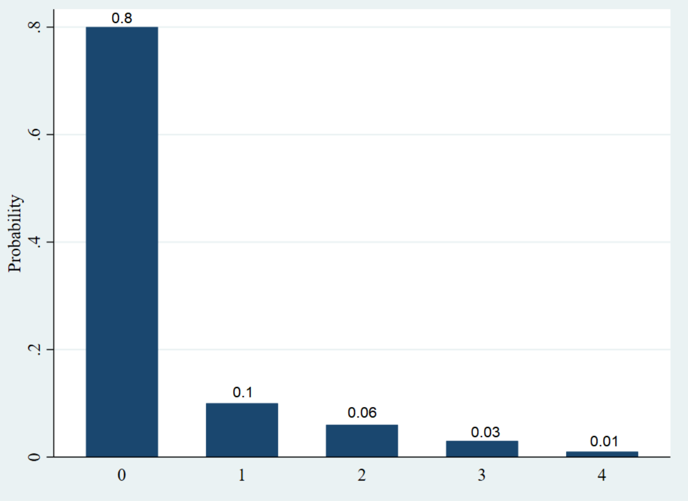
- For a continuous random variable, we define the probability density function (PDF) as $f(x) = \frac{d}{dx}F(x)$, implying a CDF of

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- Notational note: both $p(x)$ and $f(x)$ are used for PDFs/PMFs

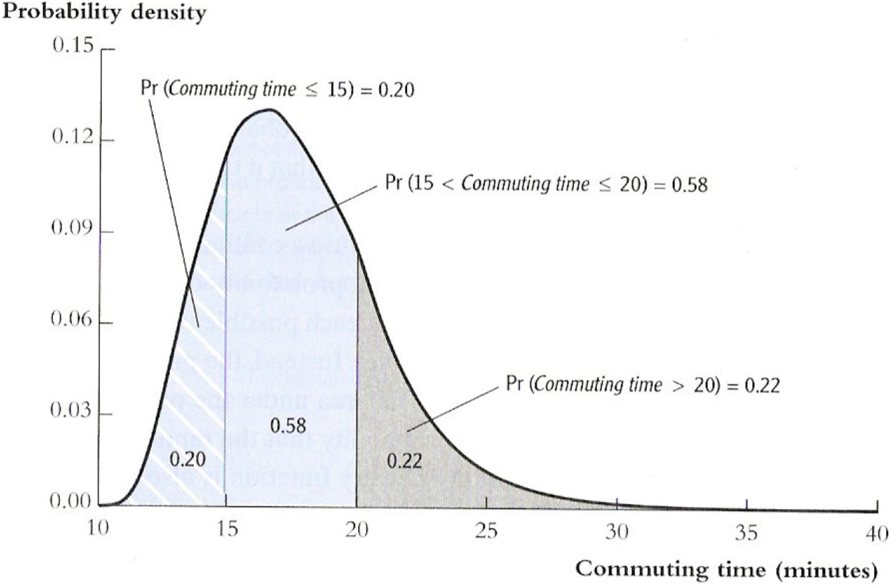
Example of a discrete random variable: number of wifi connection failures

- Here's the PMF; what is the CDF?



Example of a continuous random variable: commuting time

- Here's the PDF; what is the CDF?



Properties of PDFs/CDFs

- Key properties of CDFs $F(x) = Pr(X \leq x)$:
  - Non-decreasing: $F(x) \geq F(x')$ if $x > x'$
  - Satisfies $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$

Properties of PDFs/CDFs

- Key properties of CDFs $F(x) = Pr(X \leq x)$:
  - Non-decreasing: $F(x) \geq F(x')$ if $x > x'$
  - Satisfies $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$

- Corresponding properties of PDFs $f(x) = \frac{\partial}{\partial x} F(x)$:
  - Non-negative: $f(x) \geq 0$ for all $x \in \mathbb{X}$
  - Satisfies $\int_{x \in \mathbb{X}} f(x) dx = 1$
  - For PMFs: $\sum_{x \in \mathbb{X}} p(x) = 1$

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0, 1\}$

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0, 1\}$

- Examples: indicator for college completion, or whether a coin comes up "heads" (sometimes called a “dummy variable”)

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0,1\}$

- Examples: indicator for college completion, or whether a coin comes up "heads" (sometimes called a “dummy variable”)
- PMF:

$$p(x) = \begin{cases} 1-\pi, & x = 0 \\ \pi, & x = 1 \end{cases}$$

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

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0,1\}$

- Examples: indicator for college completion, or whether a coin comes up "heads" (sometimes called a "dummy variable")
- PMF:

  $$p(x) = \begin{cases} 1-\pi, & x = 0 \\ \pi, & x = 1 \end{cases}$$

  for some $\pi \in [0,1]$
- π is the mean (expectation) of X, which we'll define formally soon

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0,1\}$

- Examples: indicator for college completion, or whether a coin comes up "heads" (sometimes called a “dummy variable”)
- PMF:

$$p(x) = \begin{cases} 1-\pi, & x = 0 \\ \pi, & x = 1 \end{cases}$$

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

- π is the mean (expectation) of X, which we'll define formally soon
- Written $X \sim \mathit{Bernoulli}(\pi)$

Bernoulli Distributions

An important discrete distribution: Bernoulli $X \in \{0, 1\}$

- Examples: indicator for college completion, or whether a coin comes up "heads" (sometimes called a "dummy variable")
- PMF:

$$p(x) = \begin{cases} 1 - \pi, & x = 0 \\ \pi, & x = 1 \end{cases}$$

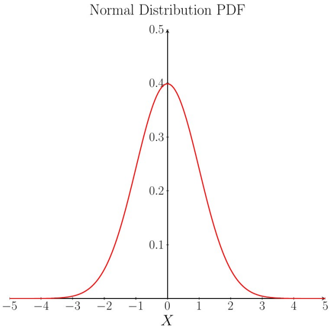
  for some $\pi \in [0, 1]$

- π is the mean (expectation) of X, which we'll define formally soon
- Written $X \sim \mathit{Bernoulli}(\pi)$
- Note this is the *only* distribution of any binary X

Normal and Uniform Distributions

An important continuous distribution: Normal $X \in \mathbb{R}$

- Example: the log of annual income (approximately)
- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$ for $x \in \mathbb{R}$ and $\sigma > 0$
- Here μ is the mean of X and σ^2 is its variance (also defined soon)
- Written $X \sim \mathrm{N}(\mu, \sigma^2)$
- Useful property: if X is normally distributed then so is $aX + b$ for any non-random a and b



Normal and Uniform Distributions

An important continuous distribution: Normal $X \in \mathbb{R}$

- Example: the log of annual income (approximately)
- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$ for $x \in \mathbb{R}$ and $\sigma > 0$
- Here μ is the mean of X and σ^2 is its variance (also defined soon)
- Written $X \sim \mathrm{N}(\mu, \sigma^2)$
- Useful property: if X is normally distributed then so is $aX + b$ for any non-random a and b

Another continuous distribution worth knowing: Uniform $X \in (a, b)$

Normal and Uniform Distributions

An important continuous distribution: Normal $X \in \mathbb{R}$

- Example: the log of annual income (approximately)
- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$ for $x \in \mathbb{R}$ and $\sigma > 0$
- Here μ is the mean of X and σ^2 is its variance (also defined soon)
- Written $X \sim \mathrm{N}(\mu, \sigma^2)$
- Useful property: if X is normally distributed then so is $aX + b$ for any non-random a and b

Another continuous distribution worth knowing: Uniform $X \in (a, b)$

- Example: random lottery numbers (approximately)

Normal and Uniform Distributions

An important continuous distribution: Normal $X \in \mathbb{R}$

- Example: the log of annual income (approximately)
- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$ for $x \in \mathbb{R}$ and $\sigma > 0$
- Here μ is the mean of X and σ^2 is its variance (also defined soon)
- Written $X \sim \mathrm{N}(\mu, \sigma^2)$
- Useful property: if X is normally distributed then so is $aX + b$ for any non-random a and b

Another continuous distribution worth knowing: Uniform $X \in (a, b)$

- Example: random lottery numbers (approximately)
- PDF: $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$

Normal and Uniform Distributions

An important continuous distribution: Normal $X \in \mathbb{R}$

- Example: the log of annual income (approximately)
- PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$ for $x \in \mathbb{R}$ and $\sigma > 0$
- Here μ is the mean of X and σ^2 is its variance (also defined soon)
- Written $X \sim \mathrm{N}(\mu, \sigma^2)$
- Useful property: if X is normally distributed then so is $aX + b$ for any non-random a and b

Another continuous distribution worth knowing: Uniform $X \in (a, b)$

- Example: random lottery numbers (approximately)
- PDF: $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$
- Written $X \sim \mathrm{U}(a, b)$; also closed under linear transformations

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?

- The CDF for the joint distribution is defined by

$$F(x,y) = Pr(X \leq x, Y \leq y),$$

the probability that $X \leq x$ and $Y \leq y$.

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?

- The CDF for the joint distribution is defined by

$$F(x, y) = Pr(X \leq x, Y \leq y),$$

  the probability that $X \leq x$ and $Y \leq y$.

- For discrete (X, Y), we define joint PMF $p(x, y) = Pr(X = x, Y = y)$

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?

- The CDF for the joint distribution is defined by

$$F(x,y) = Pr(X \leq x, Y \leq y),$$

  the probability that $X \leq x$ and $Y \leq y$.

- For discrete (X,Y), we define joint PMF $p(x,y) = Pr(X = x, Y = y)$

- For continuous (X,Y), we define joint PDF $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?
- The CDF for the joint distribution is defined by
$$F(x,y) = Pr(X \leq x, Y \leq y),$$
the probability that $X \leq x$ and $Y \leq y$.
- For discrete (X,Y), we define joint PMF $p(x,y) = Pr(X = x, Y = y)$
- For continuous (X,Y), we define joint PDF $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$

When dealing with multiple random variables, we sometimes refer to the distribution of an individual variable as its **marginal distribution**

Joint Distributions

Many interesting economic questions involve features of the **joint distribution** of two or more random variables

- E.g. How do earnings (Y) and schooling levels (X) vary together?

- The CDF for the joint distribution is defined by

$$F(x,y) = Pr(X \leq x, Y \leq y),$$

  the probability that $X \leq x$ and $Y \leq y$.

- For discrete (X,Y), we define joint PMF $p(x,y) = Pr(X = x, Y = y)$

- For continuous (X,Y), we define joint PDF $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$

When dealing with multiple random variables, we sometimes refer to the distribution of an individual variable as its **marginal distribution**

- Linked to the joint distribution by, e.g., $p(x) = \sum_{y \in \mathbb{Y}} p(x,y)$

Conditional Distributions

Combining joint and marginal distributions gives us the **conditional distribution** of one random variable given another

- Intuitively, the conditional distribution $Y|X = x$ is the distribution of Y among the sub-population with $X = x$
- Cond'l PMF $p(y \mid x) = Pr(Y = y \mid X = x) = \frac{Pr(Y=y,X=x)}{Pr(X=x)} = \frac{p(y,x)}{p(x)}$

Conditional Distributions

Combining joint and marginal distributions gives us the **conditional distribution** of one random variable given another

- Intuitively, the conditional distribution $Y|X = x$ is the distribution of Y among the sub-population with $X = x$
- Cond'l PMF $p(y \mid x) = Pr(Y = y \mid X = x) = \frac{Pr(Y=y, X=x)}{Pr(X=x)} = \frac{p(y,x)}{p(x)}$
- E.g. the distribution of earnings given college completion

Conditional Distributions

Combining joint and marginal distributions gives us the **conditional distribution** of one random variable given another

- Intuitively, the conditional distribution $Y|X = x$ is the distribution of Y among the sub-population with $X = x$
- Cond'l PMF $p(y \mid x) = Pr(Y = y \mid X = x) = \frac{Pr(Y=y, X=x)}{Pr(X=x)} = \frac{p(y,x)}{p(x)}$
- E.g. the distribution of earnings given college completion

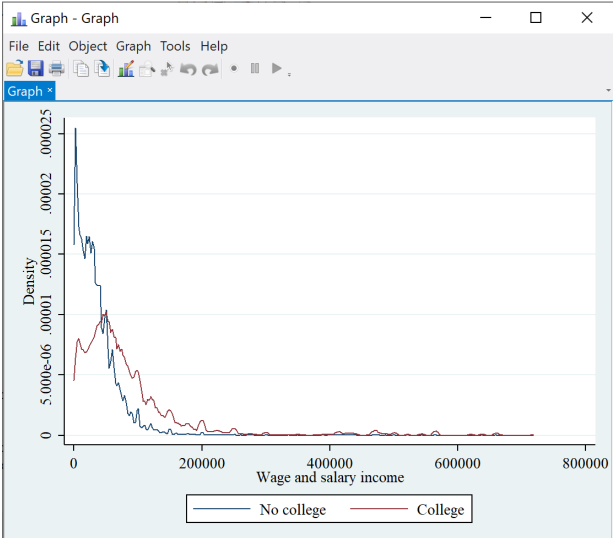
Leads immediately to *Bayes' rule*:

$$p(y \mid x) = p(x \mid y)\frac{p(y)}{p(x)}$$

Random Variables and Probability Distributions XI

Conditional PDFs of annual income given college completion:

```
twoway (kdensity incwage if educ<10) (kdensity incwage if educ>=10), ///
    xtitle("Wage and salary income") ytitle("Density") ///
    legend(label(1 "No college") label(2 "College"))
```



Independence

- An important concept in this course will be **independence**.
- Intuitively, independence says that knowing the value of X tells us nothing about the value of Y
- Formally, X, Y are independent ($X \perp\!\!\!\perp Y$) if the conditional PDF/PMF of $Y|X = x$ is the same as the unconditional one:

$$p(y \mid x) = p(y), \forall (y, x)$$

Independence

- An important concept in this course will be **independence**.
- Intuitively, independence says that knowing the value of X tells us nothing about the value of Y
- Formally, X, Y are independent ($X \perp\!\!\!\perp Y$) if the conditional PDF/PMF of $Y|X = x$ is the same as the unconditional one:

$$p(y \mid x) = p(y), \forall (y, x)$$

- Example: if D is a randomly assigned treatment, $D \perp\!\!\!\perp (Y(1), Y(0))$.

Independence

- An important concept in this course will be **independence**.

- Intuitively, independence says that knowing the value of X tells us nothing about the value of Y

- Formally, X, Y are independent ($X \perp\!\!\!\perp Y$) if the conditional PDF/PMF of $Y|X = x$ is the same as the unconditional one:

$$p(y \mid x) = p(y), \forall (y, x)$$

- Example: if D is a randomly assigned treatment, $D \perp\!\!\!\perp (Y(1), Y(0))$.
    - Understanding check: does this imply that $D \perp\!\!\!\perp Y$?

Independence

- An important concept in this course will be **independence**.
- Intuitively, independence says that knowing the value of X tells us nothing about the value of Y
- Formally, X, Y are independent ($X \perp\!\!\!\perp Y$) if the conditional PDF/PMF of $Y|X = x$ is the same as the unconditional one:

$$p(y \mid x) = p(y), \forall(y, x)$$

- Example: if D is a randomly assigned treatment, $D \perp\!\!\!\perp (Y(1), Y(0))$.
    - Understanding check: does this imply that $D \perp\!\!\!\perp Y$?
- **Conditional independence** is defined similarly. $Y \perp\!\!\!\perp X \mid W$ if

$$p(y \mid x, w) = p(y|w), \forall(y, x, w)$$

- Intuitively, X tells us nothing about Y once we know W.

Multivariate Normals

An important multivariate distribution: joint normal $\mathbf{X} \in \mathbb{R}^K$

- Parameterized by a (mean) vector $\boldsymbol{\mu} \in \mathbb{R}^K$ and a positive-definite (variance-covariance) matrix $\boldsymbol{\Sigma} \in \mathbb{R}^K \times \mathbb{R}^K$
- Note: In general I will be using **boldface** to indicate vectors/matrices

Multivariate Normals

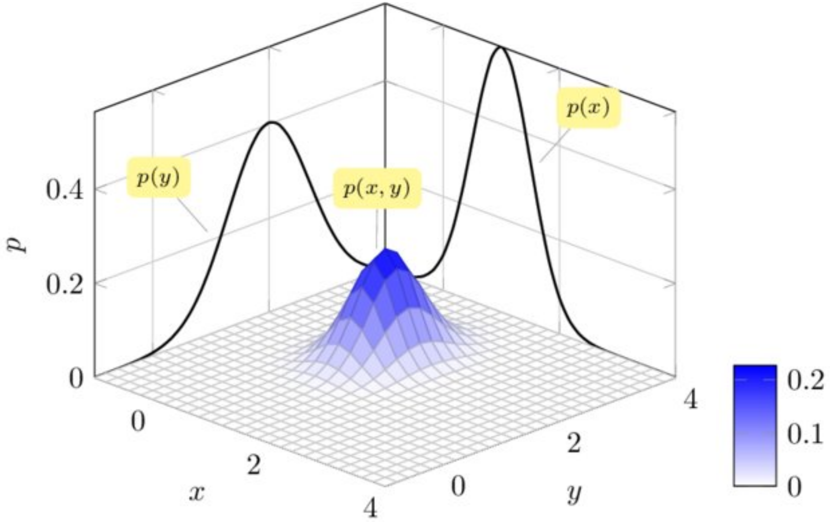
An important multivariate distribution: joint normal $\mathbf{X} \in \mathbb{R}^K$

- Parameterized by a (mean) vector $\boldsymbol{\mu} \in \mathbb{R}^K$ and a positive-definite (variance-covariance) matrix $\mathbf{\Sigma} \in \mathbb{R}^K \times \mathbb{R}^K$
- Note: In general I will be using **boldface** to indicate vectors/matrices

Many useful facts; here's a few. If $(X, Y)'$ is joint-normally distributed:

- The marginal distributions of X and Y are normal
- The conditional distributions of $X \mid Y$ and $Y \mid X$ are normal
- Any fixed linear combination $aX + bY + c$ is normally distributed

Bivariate Normal PDF



Outline

1. Random Variables and Probability Distributions✓
2. Means and Variances
3. Identification in Experiments
4. Random Sampling and Sample Means
5. Hypothesis Testing and Inference

Means

- We are often interested in the average of economic random variables (e.g. household income)

Means

- We are often interested in the average of economic random variables (e.g. household income)

- The **mean/expectation** of X is its probability-weighted typical value
  - For discrete random variables:

$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$

Means

- We are often interested in the average of economic random variables (e.g. household income)

- The **mean/expectation** of X is its probability-weighted typical value
    - For discrete random variables:

$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$

      Interpretation: long-run average of X over repeated draws

Means

- We are often interested in the average of economic random variables (e.g. household income)

- The **mean/expectation** of X is its probability-weighted typical value
  - For discrete random variables:

$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$

    Interpretation: long-run average of X over repeated draws

  - For continuous random variables, $E[X] = \int_{x \in \mathbb{X}} f(x)x dx$

Means

- We are often interested in the average of economic random variables (e.g. household income)

- The **mean/expectation** of X is its probability-weighted typical value
  - For discrete random variables:

$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$

    Interpretation: long-run average of X over repeated draws

  - For continuous random variables, $E[X] = \int_{x \in \mathbb{X}} f(x)xdx$
    Caution: may not exist if $p(x)$ puts high probability on extreme x

Means

- We are often interested in the average of economic random variables (e.g. household income)

- The **mean/expectation** of X is its probability-weighted typical value
  - For discrete random variables:
$$E[X] = \sum_{x \in \mathbb{X}} p(x)x = x_1 Pr(X = x_1) + \cdots + x_K Pr(X = x_K)$$
    Interpretation: long-run average of X over repeated draws
  - For continuous random variables, $E[X] = \int_{x \in \mathbb{X}} f(x)xdx$
    Caution: may not exist if $p(x)$ puts high probability on extreme x

- **Important fact:** The expectation operator is *linear*:
  $E[a + bX] = a + bE[X]$ for constants (a, b)
  - Easily proved from the above definitions (make sure you can!)

Calculating Means: a Simple Example

- Let X be the realization of a fair die. What is $E[X]$?

Calculating Means: a Simple Example

- Let X be the realization of a fair die. What is $E[X]$?
- By definition,

$$E[X] = Pr(X = 1) \times 1 + Pr(X = 2) \times 2 + ... + Pr(X = 6) \times 6$$

Calculating Means: a Simple Example

- Let X be the realization of a fair die. What is $E[X]$?
- By definition,

$$E[X] = Pr(X = 1) \times 1 + Pr(X = 2) \times 2 + ... + Pr(X = 6) \times 6$$

- If the die is fair, $Pr(X = 1) = ... = Pr(X = 6) = \frac{1}{6}$

Calculating Means: a Simple Example

- Let X be the realization of a fair die. What is $E[X]$?
- By definition,

$$E[X] = Pr(X = 1) \times 1 + Pr(X = 2) \times 2 + ... + Pr(X = 6) \times 6$$

- If the die is fair, $Pr(X = 1) = ... = Pr(X = 6) = \frac{1}{6}$
- Plugging this in, we have

$$E[X] = \frac{1}{6}(1 + ... + 6) = 3.5$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right]$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E\left[X\right]^2$ (why?)

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$Var(a + bX) =$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E\left[X\right]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$Var(a + bX) = E[(a + bX - E[a + bX])^2]$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$\begin{aligned} Var(a + bX) =& E[(a + bX - E[a + bX])^2] \\ =& E[(a + bX - a - bE[X])^2] \end{aligned}$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E\left[X\right]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$\begin{aligned} Var(a + bX) =& E[(a + bX - E[a + bX])^2] \\ =& E[(a + bX - a - bE[X])^2] \\ =& b^2 E[(X - E[X])^2] \end{aligned}$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$\begin{aligned} Var(a + bX) =& E[(a + bX - E[a + bX])^2] \\ =& E[(a + bX - a - bE[X])^2] \\ =& b^2 E[(X - E[X])^2] \\ =& b^2 Var(X) \end{aligned}$$

Variances

Variances measure the squared spread of a distribution:

- $Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - E[X]^2$ (why?)
- The *standard deviation* of X is $Std(X) = \sqrt{Var(X)}$; it captures the "typical" deviation of X from its mean

Variance of a linear transformation:

$$\begin{aligned} Var(a + bX) =& E[(a + bX - E[a + bX])^2] \\ =& E[(a + bX - a - bE[X])^2] \\ =& b^2 E[(X - E[X])^2] \\ =& b^2 Var(X) \end{aligned}$$

This implies that $Std(a + bX) = b \cdot Std(X)$.

- Intuitively, if I measure income in cents, the standard deviation should be 100 times if I measure it in dollars

Covariances

Covariance measures the linear association between two variables:

- $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

Covariances

Covariance measures the linear association between two variables:

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X,X) = E\left[(X - E[X])^2\right] = Var(X)$

Covariances

Covariance measures the linear association between two variables:

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X,X) = E\left[(X - E[X])^2\right] = Var(X)$
- The *correlation* between X and Y is $Corr(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$

Covariances

Covariance measures the linear association between two variables:

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X,X) = E\left[(X - E[X])^2\right] = Var(X)$
- The *correlation* between X and Y is $Corr(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$

$Cov(X,Y) > 0$ means X tends to be above its mean when Y is above its mean (and vice versa)

Covariances

Covariance measures the linear association between two variables:

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X,X) = E\left[(X - E[X])^2\right] = Var(X)$
- The *correlation* between X and Y is $Corr(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$

$Cov(X,Y) > 0$ means X tends to be above its mean when Y is above its mean (and vice versa)

- $Corr(X,Y)$ is a *unit free* (standardized) measure of linear association

Covariances

Covariance measures the linear association between two variables:

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X,X) = E\left[(X - E[X])^2\right] = Var(X)$
- The *correlation* between X and Y is $Corr(X,Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$

$Cov(X,Y) > 0$ means X tends to be above its mean when Y is above its mean (and vice versa)

- $Corr(X,Y)$ is a *unit free* (standardized) measure of linear association
- If X and Y are independent, then $Cov(X,Y) = Corr(X,Y) = 0$

Covariances

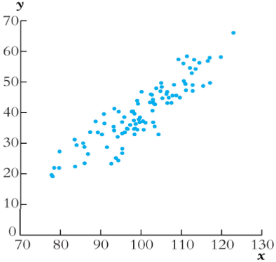
Covariance measures the linear association between two variables:

- $Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$
- Notice how $Cov(X, X) = E\left[(X - E[X])^2\right] = Var(X)$
- The *correlation* between X and Y is $Corr(X, Y) = \frac{Cov(X,Y)}{Std(X)Std(Y)}$

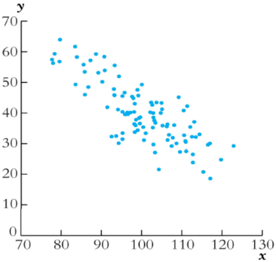
$Cov(X, Y) > 0$ means X tends to be above its mean when Y is above its mean (and vice versa)

- $Corr(X, Y)$ is a *unit free* (standardized) measure of linear association
- If X and Y are independent, then $Cov(X, Y) = Corr(X, Y) = 0$
- But <u>not</u> vice-versa! Independence is a stronger notion of association

Examples of Correlations

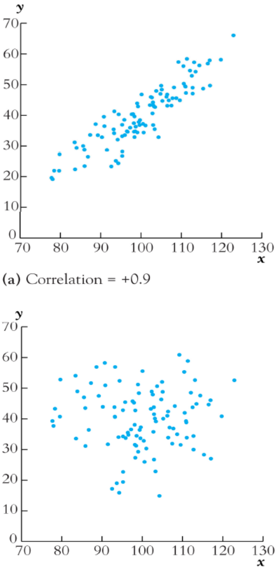


(a) Correlation = +0.9



(b) Correlation = –0.8

Examples of Correlations

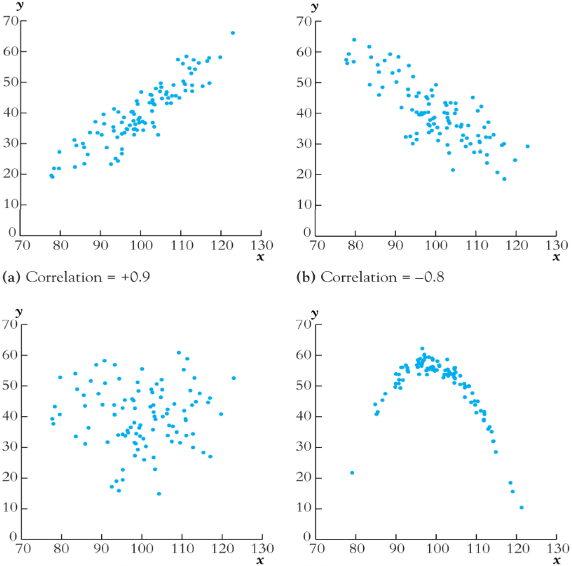


(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

Examples of Correlations



(a) Correlation = +0.9

(b) Correlation = –0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

Means/Variances of Linear Combinations

- Expectations are linear: $E[aX + bY + c] = aE[X] + bE[Y] + c$

Means/Variances of Linear Combinations

- Expectations are linear: $E[aX + bY + c] = aE[X] + bE[Y] + c$
- Variances are "quadratic":
  $Var(aX + bY + c) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$

Means/Variances of Linear Combinations

- Expectations are linear: $E[aX + bY + c] = aE[X] + bE[Y] + c$
- Variances are "quadratic":
  $Var(aX + bY + c) = a^2 Var(X) + 2abCov(X, Y) + b^2 Var(Y)$
- Covariances are linear: $Cov(aX + c, bY + d) = abCov(X, Y)$
  and $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$

Conditional Expectations

In economics we are especially interested in **conditional expectations**

- What is average of Y when $X = x$ (e.g. what are the average earnings among people who went to Brown)?

Conditional Expectations

In economics we are especially interested in **conditional expectations**

- What is average of Y when $X = x$ (e.g. what are the average earnings among people who went to Brown)?
- Conditional expectation function (CEF):
  $E[Y \mid X = x] = \sum_{y \in \mathbb{Y}} yp(y \mid x)$ if discrete
  $E[Y \mid X = x] = \int_{y \in \mathbb{Y}} yf(y \mid x)dy$ if continuous

Conditional Expectations

In economics we are especially interested in **conditional expectations**

- What is average of Y when $X = x$ (e.g. what are the average earnings among people who went to Brown)?
- Conditional expectation function (CEF):
  $E[Y \mid X = x] = \sum_{y \in \mathbb{Y}} yp(y \mid x)$ if discrete
  $E[Y \mid X = x] = \int_{y \in \mathbb{Y}} yf(y \mid x)dy$ if continuous
- We sometimes write $E[Y \mid X]$ for the random CEF evaluated at X

Conditional Expectations

In economics we are especially interested in **conditional expectations**

- What is average of Y when $X = x$ (e.g. what are the average earnings among people who went to Brown)?
- Conditional expectation function (CEF):
  $E[Y \mid X = x] = \sum_{y \in \mathbb{Y}} yp(y \mid x)$ if discrete
  $E[Y \mid X = x] = \int_{y \in \mathbb{Y}} yf(y \mid x)dy$ if continuous
- We sometimes write $E[Y \mid X]$ for the <u>random</u> CEF evaluated at X
- Say Y is *mean independent* of X when $E[Y \mid X = x] = E[Y]$ for all x

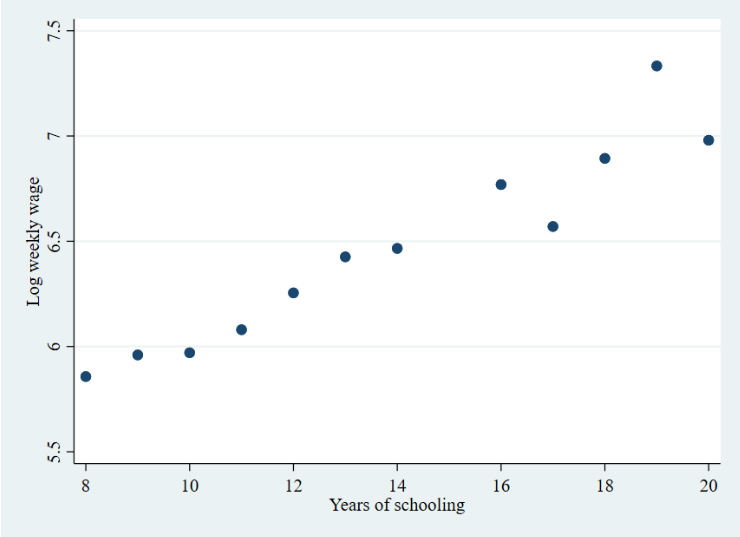
Conditional Expectations

In economics we are especially interested in **conditional expectations**

- What is average of Y when $X = x$ (e.g. what are the average earnings among people who went to Brown)?
- Conditional expectation function (CEF):
  $E[Y \mid X = x] = \sum_{y \in \mathbb{Y}} yp(y \mid x)$ if discrete
  $E[Y \mid X = x] = \int_{y \in \mathbb{Y}} yf(y \mid x)dy$ if continuous
- We sometimes write $E[Y \mid X]$ for the random CEF evaluated at X
- Say Y is *mean independent* of X when $E[Y \mid X = x] = E[Y]$ for all x
- Conditioning on X makes functions of it constant: e.g.
  $E[f(X) + g(X)Y|X = x] = f(x) + g(x)E[Y|X = x]$ for any $f(\cdot)$, $g(\cdot)$

Conditional Expectation Example

CEF of (log) annual income given years of schooling



The Big LIE

A very important result for us: the **Law of Iterated Expectations** (LIE)

Let's start with an example. Suppose I want to calculate the average height of people in the United States. The LIE says I can:

1) Compute the average height for men.

2) Compute the average height for women.

3) Average the average heights for men and women (proportional to the fraction who are women)

The Big LIE

A very important result for us: the **Law of Iterated Expectations** (LIE)

Let's start with an example. Suppose I want to calculate the average height of people in the United States. The LIE says I can:

1) Compute the average height for men.

2) Compute the average height for women.

3) Average the average heights for men and women (proportional to the fraction who are women)

Mathematically, we have

$$E[height] = P(woman)E[height|woman] + P(man)E[height|man]$$

The Big LIE

A very important result for us: the **Law of Iterated Expectations** (LIE)

Let's start with an example. Suppose I want to calculate the average height of people in the United States. The LIE says I can:

1) Compute the average height for men.

2) Compute the average height for women.

3) Average the average heights for men and women (proportional to the fraction who are women)

Mathematically, we have

$$\begin{aligned} E[height] &= P(woman)E[height|woman] + P(man)E[height|man] \\ &= E[E[height|gender]] \end{aligned}$$

The Big LIE

The formal version of the **Law of Iterated Expectations** (LIE) is:

$$E[Y] = E[E[Y \mid X]]$$

The Big LIE

The formal version of the **Law of Iterated Expectations** (LIE) is:

$$E[Y] = E[E[Y \mid X]]$$

Note that the expectation on the LHS uses $p(y)$, while the outer expectation on the RHS uses $p(x)$ and the inner expectation uses $p(y \mid x)$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$Corr(X, Y) \propto E[(X - E[X])(Y - E[Y])]$$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned} Corr(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) \mid X]] \end{aligned}$$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned} Corr(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) \mid X]] \\ &= E[(X - E[X])E[Y - E[Y] \mid X]] \end{aligned}$$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned}
Corr(X,Y) &\propto E[(X-E[X])(Y-E[Y])] \\
&= E[E[(X-E[X])(Y-E[Y]) \mid X]] \\
&= E[(X-E[X])E[Y-E[Y] \mid X]] \\
&= E[(X-E[X])(E[Y \mid X]-E[Y])]
\end{aligned}$$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned} Corr(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) \mid X]] \\ &= E[(X - E[X])E[Y - E[Y] \mid X]] \\ &= E[(X - E[X])(E[Y \mid X] - E[Y])] \\ &= 0, \text{ when } E[Y \mid X] = E[Y] \end{aligned}$$

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$
\begin{aligned}
Corr(X,Y) &\propto E[(X-E[X])(Y-E[Y])] \\
&= E[E[(X-E[X])(Y-E[Y]) \mid X]] \\
&= E[(X-E[X])E[Y-E[Y] \mid X]] \\
&= E[(X-E[X])(E[Y \mid X]-E[Y])] \\
&= 0, \text{ when } E[Y \mid X] = E[Y]
\end{aligned}
$$

Make sure you understand how we got each step!

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned} Corr(X,Y) &\propto E[(X-E[X])(Y-E[Y])] \\ &= E[E[(X-E[X])(Y-E[Y]) \mid X]] \\ &= E[(X-E[X])E[Y-E[Y] \mid X]] \\ &= E[(X-E[X])(E[Y \mid X]-E[Y])] \\ &= 0, \text{ when } E[Y \mid X]=E[Y] \end{aligned}$$

Make sure you understand how we got each step!

Converse does not hold: uncorrelated variables can be mean dependent

Mean-Independence vs. Uncorrelatedness

The LIE shows us that mean independence implies uncorrelatedness

$$\begin{aligned} Corr(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) \mid X]] \\ &= E[(X - E[X])E[Y - E[Y] \mid X]] \\ &= E[(X - E[X])(E[Y \mid X] - E[Y])] \\ &= 0, \text{ when } E[Y \mid X] = E[Y] \end{aligned}$$

Make sure you understand how we got each step!

Converse does not hold: uncorrelated variables can be mean dependent

- Also, of course, independent $\implies$ mean independent (but not $\impliedby$)

Quick Aside on Vector/Matrix Notation

Often it will be useful to work with random vectors $\mathbf{X} = [X_1, \ldots, X_N]'$

- These will always be "columns" in this course, and denoted in bold
- I will also use bold for matrices, with K columns and N rows

Quick Aside on Vector/Matrix Notation

Often it will be useful to work with random vectors $\mathbf{X} = [X_1, \ldots, X_N]'$

- These will always be "columns" in this course, and denoted in bold
- I will also use bold for matrices, with K columns and N rows

A useful reference for standard vector/matrix arithmetic/operators (e.g. transpose, inverse...) is The Matrix Cookbook

- `www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

Quick Aside on Vector/Matrix Notation

Often it will be useful to work with random vectors $\mathbf{X} = [X_1, \ldots, X_N]'$

- These will always be "columns" in this course, and denoted in bold
- I will also use bold for matrices, with K columns and N rows

A useful reference for standard vector/matrix arithmetic/operators (e.g. transpose, inverse...) is The Matrix Cookbook

- `www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

Expectations are elementwise: e.g. $E\left[\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\right] = \begin{bmatrix} E[X_{11}] & E[X_{12}] \\ E[X_{21}] & E[X_{22}] \end{bmatrix}$

- Define $Var\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{bmatrix}$ and $Cov\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}\right) = \begin{bmatrix} Cov(X_1, Y_1) & Cov(X_1, Y_2) \\ Cov(X_2, Y_1) & Cov(X_2, Y_2) \end{bmatrix}$, etc

Outline

1. Random Variables and Probability Distributions✓
2. Means and Variances ✓
3. Identification in Experiments
4. Random Sampling and Sample Means
5. Hypothesis Testing and Inference

Identification in Experiments

- The theory we've covered so far is enough to show mathematically why experiments "work", at least from an identification perspective

Identification in Experiments

- The theory we've covered so far is enough to show mathematically why experiments "work", at least from an identification perspective
- Recall the *potential outcomes* framework:
  - $Y_i(1), Y_i(0)$ are outcomes of individual i under treatment/control
  - Use these to model observed outcomes: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

Identification in Experiments

- The theory we've covered so far is enough to show mathematically why experiments "work", at least from an identification perspective
- Recall the *potential outcomes* framework:
  - $Y_i(1), Y_i(0)$ are outcomes of individual i under treatment/control
  - Use these to model observed outcomes: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$
- Suppose that we are interested in the average treatment effect:

$$ATE = E[Y_i(1) - Y_i(0)]$$

Identification in Experiments

- The theory we've covered so far is enough to show mathematically why experiments "work", at least from an identification perspective
- Recall the *potential outcomes* framework:
  - $Y_i(1), Y_i(0)$ are outcomes of individual i under treatment/control
  - Use these to model observed outcomes: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$
- Suppose that we are interested in the average treatment effect:

$$ATE = E[Y_i(1) - Y_i(0)]$$

- Suppose that for each person we assign D_i by flipping a coin. This implies that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$. Why?

Using Randomization

- By virtue of the experiment, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$.
- What is $\underbrace{E[Y_i|D_i = 1]}_{\text{Pop mean for treated units}}$?

Using Randomization

- By virtue of the experiment, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$.
- What is $\underbrace{E[Y_i|D_i = 1]}_{\text{Pop mean for treated units}}$?

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)],$$

where the first equality uses the potential outcomes model and the second equality uses (mean) independence.

Using Randomization

- By virtue of the experiment, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$.

- What is $\underbrace{E[Y_i|D_i = 1]}_{\text{Pop mean for treated units}}$?

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)],$$

  where the first equality uses the potential outcomes model and the second equality uses (mean) independence.

- Similarly, $E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0] = E[Y_i(0)]$.

Using Randomization

- By virtue of the experiment, $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$.
- What is $\underbrace{E[Y_i|D_i = 1]}_{\text{Pop mean for treated units}}$?

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)],$$

where the first equality uses the potential outcomes model and the second equality uses (mean) independence.

- Similarly, $E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0] = E[Y_i(0)]$.
- Combining these results, we can see that

$$\underbrace{E[Y_i|D_i = 1]}_{\text{Pop mean for treated}} - \underbrace{E[Y_i|D_i = 0]}_{\text{Pop mean for control}} = \underbrace{E[Y_i(1) - Y_i(0)]}_{\text{Avg treatment effect}} = \tau$$

Thus, the difference in treated/control population means in an experiment identifies the ATE!

Identification under Conditional Unconfoundedness

- Now suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristic
- Called conditional unconfoundedness or **selection on observables**

Identification under Conditional Unconfoundedness

- Now suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristic
- Called conditional unconfoundedness or **selection on observables**
- Intuitively, conditional unconfoundedness says we effectively have an experiment among people with the same value of $\mathbf{X}_i$
  - Implied by (unconditional) independence, but *weaker*: allows non-randomness through $\mathbf{X}_i$

Identification under Conditional Unconfoundedness

- Now suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristic

- Called conditional unconfoundedness or **selection on observables**

- Intuitively, conditional unconfoundedness says we effectively have an experiment among people with the same value of $\mathbf{X}_i$
  - Implied by (unconditional) independence, but *weaker*: allows non-randomness through $\mathbf{X}_i$

- Why might we believe conditional unconfoundedness?
  - Stratified experiment: we randomize among people with same value of $\mathbf{X}_i$ (e.g., we hold a lottery for each state)

Identification under Conditional Unconfoundedness

- Now suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristic

- Called conditional unconfoundedness or **selection on observables**

- Intuitively, conditional unconfoundedness says we effectively have an experiment among people with the same value of $\mathbf{X}_i$
    - Implied by (unconditional) independence, but *weaker*: allows non-randomness through $\mathbf{X}_i$

- Why might we believe conditional unconfoundedness?
    - Stratified experiment: we randomize among people with same value of $\mathbf{X}_i$ (e.g., we hold a lottery for each state)
    - "Quasi-experiment"/"natural experiment": we think D_i is (effectively) as good as random among people with same value of $\mathbf{X}_i$

Example – Hot Days and Test Scores

- Park et al (2021) study the impact of hot days (D_i) during the school year on test scores (Y_i)
  - Note Their D_i is not binary, although we could imagine a binarized treatment, e.g. $D_i = 1[Hotdays > 10]$
- Why do we think $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$?

Example – Hot Days and Test Scores

- Park et al (2021) study the impact of hot days (D_i) during the school year on test scores (Y_i)
  - Note Their D_i is not binary, although we could imagine a binarized treatment, e.g. $D_i = 1[Hotdays > 10]$
- Why do we think $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$?
  - People in places with different climates may be different on other things

Example – Hot Days and Test Scores

- Park et al (2021) study the impact of hot days (D_i) during the school year on test scores (Y_i)
    - Note Their D_i is not binary, although we could imagine a binarized treatment, e.g. $D_i = 1[Hotdays > 10]$
- Why do we think $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$?
    - People in places with different climates may be different on other things
- Park et al argue that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | X_i$, where X_i is a measure of historical weather in i's location
    - Heat in a given year is effectively random conditional on historical weather patterns.

Example – Hot Days and Test Scores

- Park et al (2021) study the impact of hot days (D_i) during the school year on test scores (Y_i)
  - Note Their D_i is not binary, although we could imagine a binarized treatment, e.g. $D_i = 1[Hotdays > 10]$
- Why do we think $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$?
  - People in places with different climates may be different on other things
- Park et al argue that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | X_i$, where X_i is a measure of historical weather in i's location
  - Heat in a given year is effectively random conditional on historical weather patterns.
- Does this seem reasonable to you?

Park et al. (2021) Abstract

Human capital generally, and cognitive skills specifically, play a crucial role in determining economic mobility and macroeconomic growth. While elevated temperatures have been shown to impair short-run cognitive performance, much less is known about whether heat exposure affects the rate of skill formation. We combine standardized achievement data for 58 countries and 12,000 US school districts with detailed weather and academic calendar information to show that the rate of learning decreases with an increase in the number of hot school days. These results provide evidence that climatic differences may contribute to differences in educational achievement both across countries and within countries by socioeconomic status and that may have important implications for the magnitude and functional form of climate damages in coupled human–natural systems.

Example – Returns to College Selectivity

- Dale and Krueger (2002) studied a Q similar to our ongoing example: What is the effect on earnings of attending a selective college?
- Clearly, $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ because students who attend selective college will tend to have different academic ability.

Example – Returns to College Selectivity

- Dale and Krueger (2002) studied a Q similar to our ongoing example: What is the effect on earnings of attending a selective college?
- Clearly, $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ because students who attend selective college will tend to have different academic ability.
- Dale and Krueger argue that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is the set of colleges to which someone applied and got admitted.
- Essentially, they argue that once we know what colleges you applied/ were admitted to, where you choose to go is effectively random

Example – Returns to College Selectivity

- Dale and Krueger (2002) studied a Q similar to our ongoing example: What is the effect on earnings of attending a selective college?
- Clearly, $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ because students who attend selective college will tend to have different academic ability.
- Dale and Krueger argue that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is the set of colleges to which someone applied and got admitted.
- Essentially, they argue that once we know what colleges you applied/ were admitted to, where you choose to go is effectively random
- Do you believe this? Why might this assumption go wrong?

Example – Returns to College Selectivity

- Dale and Krueger (2002) studied a Q similar to our ongoing example: What is the effect on earnings of attending a selective college?
- Clearly, $D_i \not\perp\!\!\!\perp (Y_i(1), Y_i(0))$ because students who attend selective college will tend to have different academic ability.
- Dale and Krueger argue that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\mathbf{X}_i$, where $\mathbf{X}_i$ is the set of colleges to which someone applied and got admitted.
- Essentially, they argue that once we know what colleges you applied/ were admitted to, where you choose to go is effectively random
- Do you believe this? Why might this assumption go wrong?
    - Students who choose to go to selective college may still differ in family background, motivation, career plans, etc.

Dale and Krueger (2002)

Abstract

Estimates of the effect of college selectivity on earnings may be biased because elite colleges admit students, in part, based on characteristics that are related to future earnings. We matched students who applied to, and were accepted by, similar colleges to try to eliminate this bias. Using the College and Beyond data set and National Longitudinal Survey of the High School Class of 1972, we find that students who attended more selective colleges earned about the same as students of seemingly comparable ability who attended less selective schools. Children from low-income families, however, earned more if they attended selective colleges.

Using Conditional Unconfoundedness

- Suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristics.

- Similar to in the experiment, we have for all $\mathbf{x}$:

$$E[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}] = E[Y_i(1)|\mathbf{X}_i = \mathbf{x}].$$

  and

$$E[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}] = E[Y_i(0)|\mathbf{X}_i = \mathbf{x}].$$

Using Conditional Unconfoundedness

- Suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristics.

- Similar to in the experiment, we have for all $\mathbf{x}$:

$$E[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}] = E[Y_i(1)|\mathbf{X}_i = \mathbf{x}].$$

  and

$$E[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}] = E[Y_i(0)|\mathbf{X}_i = \mathbf{x}].$$

- This implies that

$$\underbrace{E[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{Pop avg treated w/ } \mathbf{X}_i = \mathbf{x}} - \underbrace{E[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}]}_{\text{Pop avg control w/ } \mathbf{X}_i = \mathbf{x}} = \underbrace{E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]}_{\text{ATE with } \mathbf{X}_i = \mathbf{x}}$$

Using Conditional Unconfoundedness

- Suppose that $D_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) | \mathbf{X}_i$, where $\mathbf{X}_i$ is a vector of observable characteristics.

- Similar to in the experiment, we have for all $\mathbf{x}$:

$$E[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}] = E[Y_i(1) | \mathbf{X}_i = \mathbf{x}].$$

  and

$$E[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}] = E[Y_i(0) | \mathbf{X}_i = \mathbf{x}].$$

- This implies that

$$\underbrace{E[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{Pop avg treated w/ } \mathbf{X}_i = \mathbf{x}} - \underbrace{E[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}]}_{\text{Pop avg control w/ } \mathbf{X}_i = \mathbf{x}} = \underbrace{E[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}]}_{\text{ATE with } \mathbf{X}_i = \mathbf{x}}$$

- $E[Y_i(1) - Y_i(0) | \mathbf{X}_i = \mathbf{x}]$ is often called the *conditional average treatment effect*, written $CATE(\mathbf{x})$.

Using Conditional Unconfoundedness (cont.)

- We showed that under conditional unconfoundedness, $CATE(\mathbf{x}) = E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$ is identified.
- Is the unconditional $ATE = E[Y_i(1) - Y_i(0)]$ also identified?

Using Conditional Unconfoundedness (cont.)

- We showed that under conditional unconfoundedness, $CATE(\mathbf{x}) = E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$ is identified.

- Is the unconditional $ATE = E[Y_i(1) - Y_i(0)]$ also identified?

- Yes! Using the law of iterated expectations,

$$E[\underbrace{E[Y_i(1) - Y_i(0)|\mathbf{X}_i]}_{CATE(\mathbf{X}_i)}] = E[Y_i(1) - Y_i(0)]$$

Using Conditional Unconfoundedness (cont.)

- We showed that under conditional unconfoundedness, $CATE(\mathbf{x}) = E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$ is identified.

- Is the unconditional $ATE = E[Y_i(1) - Y_i(0)]$ also identified?

- Yes! Using the law of iterated expectations,

$$E[\underbrace{E[Y_i(1) - Y_i(0)|\mathbf{X}_i]}_{CATE(\mathbf{X}_i)}] = E[Y_i(1) - Y_i(0)]$$

- **Technical note**: Here we assume $E[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}]$ and $E[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}]$ exist for every $\mathbf{x}$

Using Conditional Unconfoundedness (cont.)

- We showed that under conditional unconfoundedness, $CATE(\mathbf{x}) = E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$ is identified.
- Is the unconditional $ATE = E[Y_i(1) - Y_i(0)]$ also identified?
- Yes! Using the law of iterated expectations,

$$E[\underbrace{E[Y_i(1) - Y_i(0)|\mathbf{X}_i]}_{CATE(\mathbf{X}_i)}] = E[Y_i(1) - Y_i(0)]$$

- **Technical note**: Here we assume $E[Y_i|D_i = 1, \mathbf{X}_i = \mathbf{x}]$ and $E[Y_i|D_i = 0, \mathbf{X}_i = \mathbf{x}]$ exist for every $\mathbf{x}$
- Requires $0 < Pr(D_i = 1|\mathbf{X}_i = \mathbf{x}]) < 1$: called an **overlap** condition
- Intuitively, we need there to be some treated and some control units for each value of X_i, in order to learn about the <u>overall</u> ATE

Learning about Population Means

- We just showed that, in an experiment, the average treatment effect is identified as the difference in population means:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1) - Y_i(0)]$$

- Similarly, under conditional unconfoundedness, the CATE is identified by a difference in conditional population means

Learning about Population Means

- We just showed that, in an experiment, the average treatment effect is identified as the difference in population means:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1) - Y_i(0)]$$

- Similarly, under conditional unconfoundedness, the CATE is identified by a difference in conditional population means

- But in practice, we don't see data for the whole population, so we don't know $E[Y_i|D_i = 1]$, $E[Y_i|D_i = 0]$, etc.

Learning about Population Means

- We just showed that, in an experiment, the average treatment effect is identified as the difference in population means:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1) - Y_i(0)]$$

- Similarly, under conditional unconfoundedness, the CATE is identified by a difference in conditional population means

- But in practice, we don't see data for the whole population, so we don't know $E[Y_i|D_i = 1]$, $E[Y_i|D_i = 0]$, etc.

- We need to learn about these **estimands** from the observed sample

- Enter **statistical inference...**

Outline

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population
- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population
- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population
- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$
  - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population
- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
    - *Independent*: Y_i is independent of Y_j for all $i \neq j$
    - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j
    - *Representative:* The distribution of Y_i is the same as the distribution from the population we care about

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population

- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$
  - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j
  - *Representative:* The distribution of Y_i is the same as the distribution from the population we care about

- *iid* and representative data is a useful baseline that's relatively easy to analyze, but it's important to realize it might not hold in practice

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population

- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$
  - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j
  - *Representative:* The distribution of Y_i is the same as the distribution from the population we care about

- *iid* and representative data is a useful baseline that's relatively easy to analyze, but it's important to realize it might not hold in practice
  - If we sample people in the same household together, not independent!

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population

- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$
  - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j
  - *Representative:* The distribution of Y_i is the same as the distribution from the population we care about

- *iid* and representative data is a useful baseline that's relatively easy to analyze, but it's important to realize it might not hold in practice
  - If we sample people in the same household together, not independent!
  - If we stratify sampling by state, not identical

Defining a Sample

- To formalize the task of statistical inference, we need to specify how our observed data is drawn from the population
- Baseline case: we observe an *independent and identically distributed* (*iid*) and *representative* sample of size N: e.g. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
  - *Independent*: Y_i is independent of Y_j for all $i \neq j$
  - *Identically distributed*: Y_i and Y_j have the same distribution for all i, j
  - *Representative:* The distribution of Y_i is the same as the distribution from the population we care about
- *iid* and representative data is a useful baseline that's relatively easy to analyze, but it's important to realize it might not hold in practice
  - If we sample people in the same household together, not independent!
  - If we stratify sampling by state, not identical
  - In the Dewey v. Truman example, not representative!

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] =$$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] =$$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) =$$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\sum_i Var(Y_i) =$$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\sum_i Var(Y_i) = \sigma^2/N,$$

where $\sigma^2 = Var(Y_i)$

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\sum_i Var(Y_i) = \sigma^2/N,$$

where $\sigma^2 = Var(Y_i)$

- Equation (1) says that $\hat{\mu}$ is *unbiased*: its average value is μ

The Mean and Variance of a Sample Average

Suppose we are interested in learning the population mean $\mu = E[Y_i]$ from an *iid* representative sample $\mathbf{Y}$ of size N

- A natural estimator is the *sample mean*: $\hat{\mu} = \frac{1}{N}\sum_i Y_i$
- $\hat{\mu}$ is a function of the random data $\mathbf{Y}$. It is thus a random variable, and it has a distribution (sometimes called a "sampling distribution")

We can use what we've learned to derive the mean and variance of $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N}\sum_i Y_i\right] = \frac{1}{N}\sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N}\sum_i Y_i\right) = \frac{1}{N^2}\sum_i Var(Y_i) = \sigma^2/N,$$

where $\sigma^2 = Var(Y_i)$

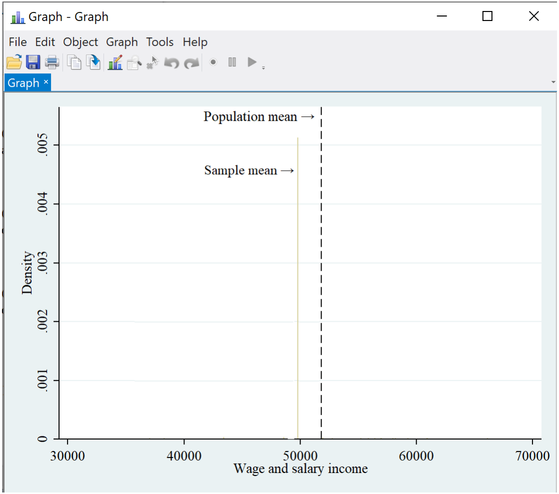
- Equation (1) says that $\hat{\mu}$ is *unbiased*: its average value is μ
- Equation (2) says that the standard deviation of $\hat{\mu}$ from its mean (i.e. μ) shrinks with the sample size N ($\approx$ *consistency*)

Random Sampling and Sample Means

Simulating unbiasedness and consistency for estimating mean income:

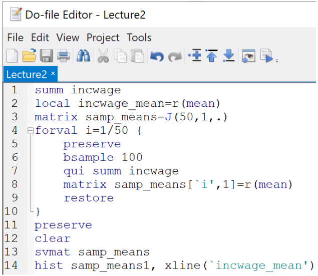
Do-file Editor - Lecture2

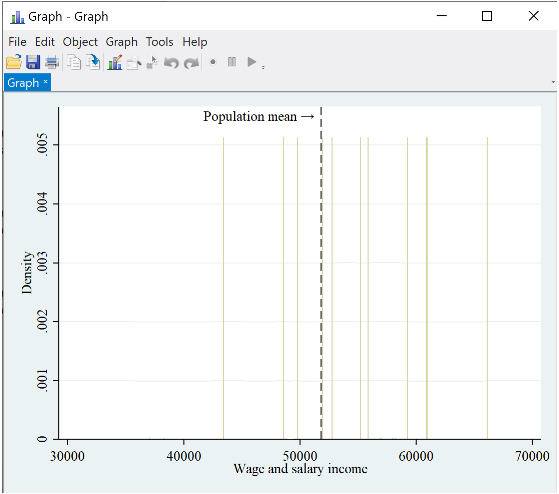
```
summ incwage
local incwage_mean=r(mean)
matrix samp_means=J(50,1,.)
forval i=1/50 {
    preserve
    bsample 100
    qui summ incwage
    matrix samp_means[`i',1]=r(mean)
    restore
}
preserve
clear
svmat samp_means
hist samp_means1, xline(`incwage_mean')
```



Simulations of Random Sampling

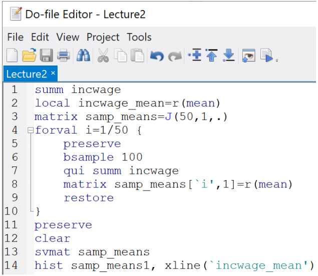
Simulating unbiasedness and consistency for estimating mean income:

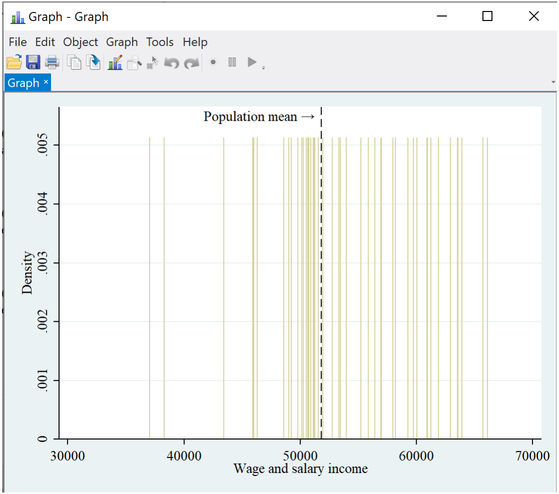




Random Sampling and Sample Means

Simulating unbiasedness and consistency for estimating mean income:



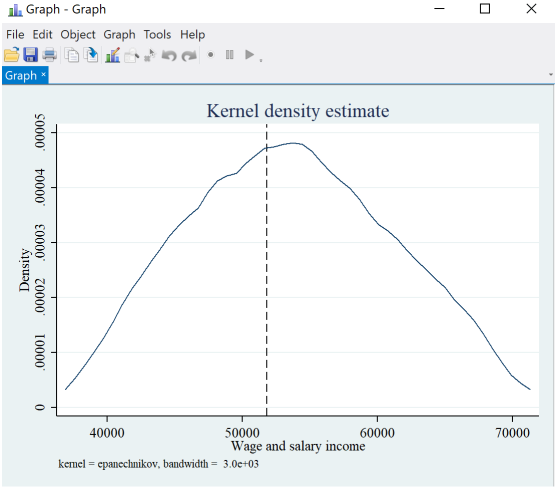


Random Sampling and Sample Means

Simulating unbiasedness and consistency for estimating mean income:

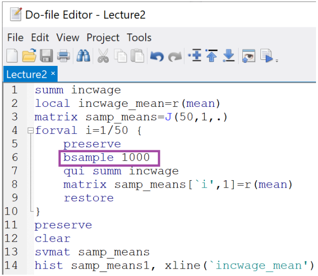
Do-file Editor - Lecture2

```
summ incwage
local incwage_mean=r(mean)
matrix samp_means=J(50,1,.)
forval i=1/50 {
    preserve
    bsample 100
    qui summ incwage
    matrix samp_means[`i',1]=r(mean)
    restore
}
preserve
clear
svmat samp_means
hist samp_means1, xline(`incwage_mean')
```

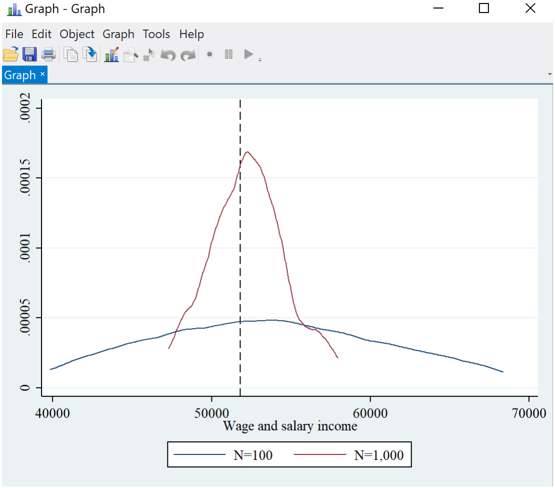


Random Sampling and Sample Means

Simulating unbiasedness and consistency for estimating mean income:




```
summ incwage
local incwage_mean=r(mean)
matrix samp_means=J(50,1,.)
forval i=1/50 {
    preserve
    bsample 1000
    qui summ incwage
    matrix samp_means[`i',1]=r(mean)
    restore
}
preserve
clear
svmat samp_means
hist samp_means1, xline(`incwage_mean')
```



Random Sampling and Sample Means

So: two reasons why $\hat{\mu} = \frac{1}{N}\sum_i Y_i$ is a good estimator of $\mu = E[Y_i]$:

- It is unbiased: $E[\hat{\mu}] = \mu$
- Its variance shrinks to zero as the sample grows: $\lim_{N\to\infty} Var(\hat{\mu}) = 0$

Random Sampling and Sample Means

So: two reasons why $\hat{\mu} = \frac{1}{N}\sum_i Y_i$ is a good estimator of $\mu = E[Y_i]$:

- It is unbiased: $E[\hat{\mu}] = \mu$
- Its variance shrinks to zero as the sample grows: $\lim_{N\to\infty} Var(\hat{\mu}) = 0$

In the next chapter we'll see another nice property of $\hat{\mu}$: when N is large, its *distribution* is approximately normal

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

- This is easiest when X_i is discrete with a small number of values x

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

- This is easiest when X_i is discrete with a small number of values x
- Natural estimator $\hat{\mu}(x) = \frac{1}{N_x} \sum_{i:X_i=x} Y_i$, where $N_x = |i : X_i = x|$ counts the number of observations with $X_i = x$

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

- This is easiest when X_i is discrete with a small number of values x
- Natural estimator $\hat{\mu}(x) = \frac{1}{N_x} \sum_{i:X_i=x} Y_i$, where $N_x = |i : X_i = x|$ counts the number of observations with $X_i = x$

Following the same derivations as before, we have

$$E[\hat{\mu}(x)] = E[Y_i \mid X_i = x] = \mu(x)$$
$$Var(\hat{\mu}(x)) = Var(Y_i \mid X_i = x)/N_x$$

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

- This is easiest when X_i is discrete with a small number of values x
- Natural estimator $\hat{\mu}(x) = \frac{1}{N_x} \sum_{i:X_i=x} Y_i$, where $N_x = |i : X_i = x|$ counts the number of observations with $X_i = x$

Following the same derivations as before, we have

$$E[\hat{\mu}(x)] = E[Y_i \mid X_i = x] = \mu(x)$$
$$Var(\hat{\mu}(x)) = Var(Y_i \mid X_i = x)/N_x$$

So this is an *unbiased* estimator which is close to the truth as $N_x \to \infty$

Random Sampling and Sample Means

Given our interest in conditional means $\mu(x) = E[Y_i \mid X_i = x]$, we might also consider conditional sample averages of Y_i given $X_i = x$

- This is easiest when X_i is discrete with a small number of values x
- Natural estimator $\hat{\mu}(x) = \frac{1}{N_x} \sum_{i:X_i=x} Y_i$, where $N_x = |i : X_i = x|$ counts the number of observations with $X_i = x$

Following the same derivations as before, we have

$$E[\hat{\mu}(x)] = E[Y_i \mid X_i = x] = \mu(x)$$
$$Var(\hat{\mu}(x)) = Var(Y_i \mid X_i = x)/N_x$$

So this is an *unbiased* estimator which is close to the truth as $N_x \to \infty$

- What if X_i is not discrete, or $N_x \not\to \infty$? Coming soon...

Outline

Hypothesis Testing – an Introduction

- We've shown that when N gets large, the sample mean $\hat{\mu}$ gets close to the population mean μ
- But what does "close" mean?
- If the sample mean of income in our data is \$50,000, is it reasonable to think the population mean could be \$55,000? What about \$70,000?

Hypothesis Testing – an Introduction

- We've shown that when N gets large, the sample mean $\hat{\mu}$ gets close to the population mean μ
- But what does "close" mean?
- If the sample mean of income in our data is \$50,000, is it reasonable to think the population mean could be \$55,000? What about \$70,000?
- **Hypothesis testing** helps us formalize the notion of "close."

Hypothesis Testing – an Introduction

- We've shown that when N gets large, the sample mean $\hat{\mu}$ gets close to the population mean μ
- But what does "close" mean?
- If the sample mean of income in our data is \$50,000, is it reasonable to think the population mean could be \$55,000? What about \$70,000?
- **Hypothesis testing** helps us formalize the notion of "close."
- It tells us whether it is likely to see a sample mean of \$50,000 if the truth is \$55,000, \$70,000, etc.

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
    - E.g. A population mean is \$55,000 would be $H_0 : \mu = 55,000$

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
    - E.g. A population mean is $55,000 would be $H_0 : \mu = 55,000$
2. Calculate how likely it would be to observe $\hat{\mu}$ at least this far from μ_0 if the null is true. This is called a **p-value**

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
    - E.g. A population mean is \$55,000 would be $H_0 : \mu = 55,000$
2. Calculate how likely it would be to observe $\hat{\mu}$ at least this far from μ_0 if the null is true. This is called a **p-value**
3. **Reject** if the *p*-value is small, i.e.: it's unlikely that we would observe a $\hat{\mu}$ so far from μ_0 if the null is true

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
   - E.g. A population mean is \$55,000 would be $H_0 : \mu = 55,000$
2. Calculate how likely it would be to observe $\hat{\mu}$ at least this far from μ_0 if the null is true. This is called a **p-value**
3. **Reject** if the *p*-value is small, i.e.: it's unlikely that we would observe a $\hat{\mu}$ so far from μ_0 if the null is true
   - A common threshold is $\alpha = 0.05$

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
    - E.g. A population mean is $55,000 would be $H_0 : \mu = 55,000$
2. Calculate how likely it would be to observe $\hat{\mu}$ at least this far from μ_0 if the null is true. This is called a **p-value**
3. **Reject** if the *p*-value is small, i.e.: it's unlikely that we would observe a $\hat{\mu}$ so far from μ_0 if the null is true
    - A common threshold is $\alpha = 0.05$
4. Form a **confidence interval** that collects all the possible values of μ_0 that we can't reject in this way

Overview of Hypothesis Testing

1. Specify a **null hypothesis** that the population mean is a particular value, $H_0 : \mu = \mu_0$.
   - E.g. A population mean is \$55,000 would be $H_0 : \mu = 55,000$
2. Calculate how likely it would be to observe $\hat{\mu}$ at least this far from μ_0 if the null is true. This is called a **p-value**
3. **Reject** if the *p*-value is small, i.e.: it's unlikely that we would observe a $\hat{\mu}$ so far from μ_0 if the null is true
   - A common threshold is $\alpha = 0.05$
4. Form a **confidence interval** that collects all the possible values of μ_0 that we can't reject in this way
   - The CI, by construction, contains the true value μ in 95% of the realizations of the data when $\alpha = 0.05$

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.
- Why consider this case?
  - $\mathrm{N}(\mu, \sigma^2/N)$ is the exact distribution of $\hat{\mu}$ if $Y_i \stackrel{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.
- Why consider this case?
  - $\mathrm{N}(\mu, \sigma^2/N)$ is the exact distribution of $\hat{\mu}$ if $Y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$
  - We'll show in the next chapter that even if Y_i is not normal, $\hat{\mu}$ is approximately normal for large N

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.
- Why consider this case?
  - $\mathrm{N}(\mu, \sigma^2/N)$ is the exact distribution of $\hat{\mu}$ if $Y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$
  - We'll show in the next chapter that even if Y_i is not normal, $\hat{\mu}$ is approximately normal for large N
  - We'll also show that with large N we can estimate σ^2 arbitrarily well

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.
- Why consider this case?
  - $\mathrm{N}(\mu, \sigma^2/N)$ is the exact distribution of $\hat{\mu}$ if $Y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$
  - We'll show in the next chapter that even if Y_i is not normal, $\hat{\mu}$ is approximately normal for large N
  - We'll also show that with large N we can estimate σ^2 arbitrarily well
- Suppose we wish to test the null hypothesis $H_0 : \mu = \mu_0$ (e.g. average income is \$55,000)
- Let $\hat{t} = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}}$, and note under the null hypothesis $\hat{t} \sim \mathrm{N}(0, 1)$

Hypothesis Testing with Normally Distributed $\hat{\mu}$

- Let's work through all the steps in the special case: $\hat{\mu} \sim \mathrm{N}(\mu, \sigma^2/N)$ with σ^2 known.
- Why consider this case?
  - $\mathrm{N}(\mu, \sigma^2/N)$ is the exact distribution of $\hat{\mu}$ if $Y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$
  - We'll show in the next chapter that even if Y_i is not normal, $\hat{\mu}$ is approximately normal for large N
  - We'll also show that with large N we can estimate σ^2 arbitrarily well
- Suppose we wish to test the null hypothesis $H_0 : \mu = \mu_0$ (e.g. average income is \$55,000)
- Let $\hat{t} = \dfrac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}}$, and note under the null hypothesis $\hat{t} \sim \mathrm{N}(0,1)$
  - The distribution of $\hat{t}$ is over repeated draws of the sample $(Y_1, ..., Y_N)$.

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0, 1)$.

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0, 1)$.

- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0,1)$.

- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

  $Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t)$

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0,1)$.
- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) =$

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0,1)$.
- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

  $$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- We define the p-value for the null $H_0 : \mu = \mu_0$ as

  $$p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0,1)$.
- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

  $$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- We define the *p*-value for the null $H_0 : \mu = \mu_0$ as

  $$p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)) = 1 - \left(\Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \right) - \Phi\left(\frac{-|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \right) \right)$$

Hypothesis Testing with Normally Distributed $\hat{\mu}$ (cont.)

- We've shown that under the null, $H_0 : \mu = \mu_0$, $\hat{t} \sim \mathrm{N}(0,1)$.
- What is $Pr(|\hat{t}| > t)$ for some $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- We define the p-value for the null $H_0 : \mu = \mu_0$ as

$$\begin{aligned} p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)) &= 1 - \left(\Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) - \Phi\left(\frac{-|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right)\right) \end{aligned}$$

- Intuitively, p is the probability we would see a $|\hat{t}|$ at least this big if the null is true.

Illustration of P-Value Construction

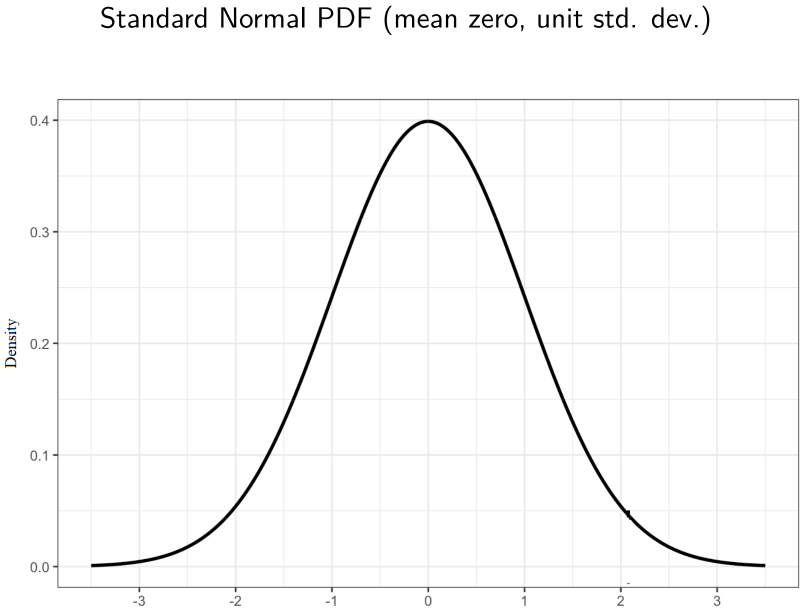


Illustration of P-Value Construction

Normalized realization of the random estimator

$$\left| \frac{\hat{\mu}(x)^{obs} - \mu_0}{\sqrt{\sigma^2(x)/N}} \right|$$

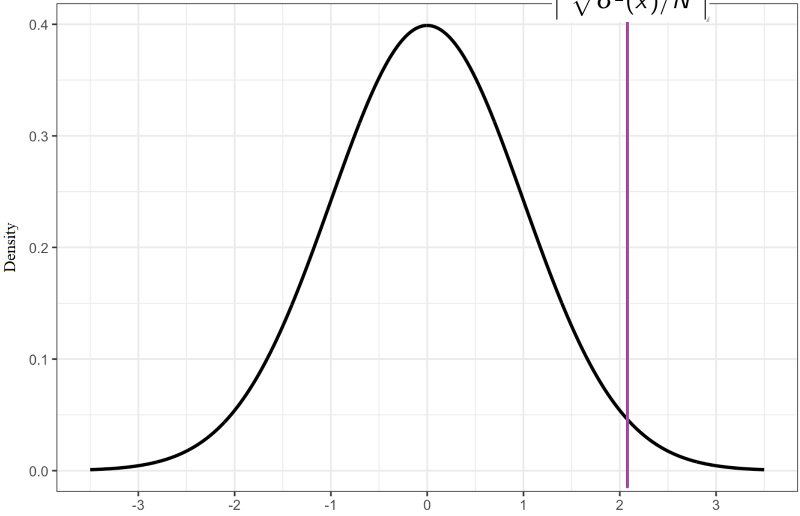
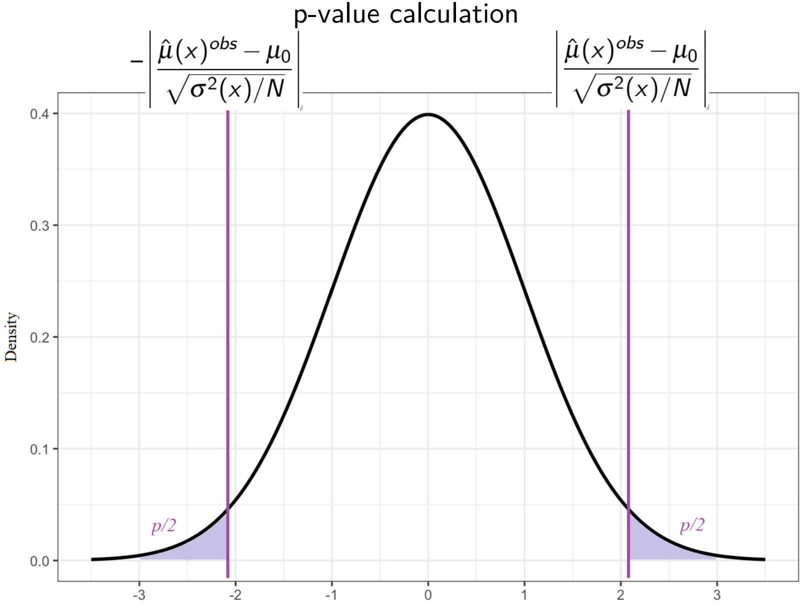


Illustration of P-Value Construction



When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.

When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.

- What does this imply about the value of μ_0 we reject/don't reject?

When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.
- What does this imply about the value of μ_0 we reject/don't reject?
- We don't reject if

$$|\hat{t}| \leq 1.96$$

When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.

- What does this imply about the value of μ_0 we reject/don't reject?

- We don't reject if

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96$$

When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.

- What does this imply about the value of μ_0 we reject/don't reject?

- We don't reject if

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96 \implies \mu_0 \in [\hat{\mu} - 1.96\sigma/\sqrt{N}, \hat{\mu} + 1.96\sigma\sqrt{N}]$$

When Do We Reject the Null?

- Recall that our p-value takes the form

$$1 - \left(\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)\right)$$

- It turns out that $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Thus, $p < 0.05$ if and only if $|\hat{t}| > 1.96$. So we reject at the 5% level if $|\hat{t}| > 1.96$.

- What does this imply about the value of μ_0 we reject/don't reject?

- We don't reject if

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96 \implies \mu_0 \in [\hat{\mu} - 1.96\sigma/\sqrt{N}, \hat{\mu} + 1.96\sigma\sqrt{N}]$$

- The interval $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ is thus the 95% confidence interval (CI)
  - It has the property that $Pr(\mu_0 \in CI) = 0.95$ when $H_0 : \mu = \mu_0$ is true

Significance and Power

- The *significance level* (or *size*) of a test is the pre-specified probability of incorrectly rejecting the null when it is true (type-I error rate)

Significance and Power

- The *significance level* (or *size*) of a test is the pre-specified probability of incorrectly rejecting the null when it is true (type-I error rate)
  - E.g. a 5% level test rejects when $p < 0.05$.

Significance and Power

- The *significance level* (or *size*) of a test is the pre-specified probability of incorrectly rejecting the null when it is true (type-I error rate)
  - E.g. a 5% level test rejects when $p < 0.05$.
- The *power* of a test is the probability of correctly rejecting the null when it is false (1 - type-II error rate)
  - The power is a function of the *alternative* hypothesis. I.e., the probability that we reject $H_0 : \mu = \mu_0$ when in fact $\mu = \mu_A$

Caution about P-Value Interpretation

- Frequentist p-values are often interpreted as "probability that H_0 is true." Is this right?

Caution about P-Value Interpretation

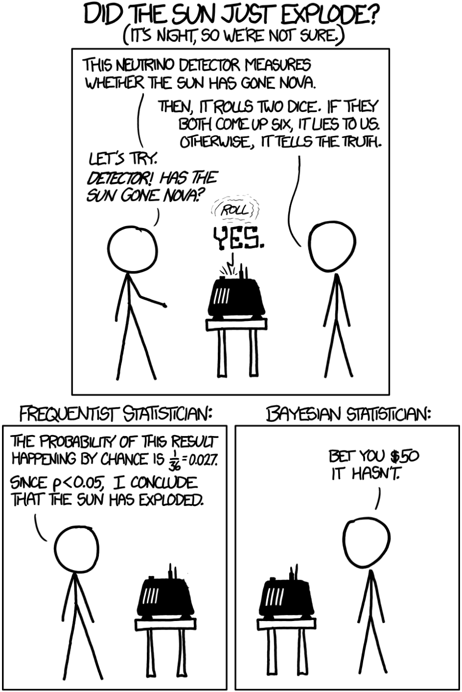
- Frequentist p-values are often interpreted as "probability that H_0 is true." Is this right? No!
  - p-value tells us the probability of getting the observed data *assuming* the null is true
  - That is, p-value tells us about $P(data|H_0)$, not $P(H_0|data)$.
  - By Bayes' rule, $P(H_0|Data) = P(Data|H_0) * P(H_0)/P(Data)$. But to formalize this, we need to take a stand on our *prior* belief that H_0 is true, $P(H_0)$.

Caution about P-Value Interpretation

- Frequentist p-values are often interpreted as "probability that H_0 is true." Is this right? No!
  - p-value tells us the probability of getting the observed data *assuming* the null is true
  - That is, p-value tells us about $P(data|H_0)$, not $P(H_0|data)$.
  - By Bayes' rule, $P(H_0|Data) = P(Data|H_0) * P(H_0)/P(Data)$. But to formalize this, we need to take a stand on our *prior* belief that H_0 is true, $P(H_0)$.
- People often interpret a $p < 0.05$ as strong evidence of an effect and $p \geq 0.05$ as evidence of no effect.
  - But $p = 0.05$ is a fairly arbitrary threshold.
  - It's better to view the p-value as a spectrum indicating how likely is the observed data given the null.

Caution about P-Value Interpretation

- Frequentist p-values are often interpreted as "probability that H_0 is true." Is this right? No!
  - p-value tells us the probability of getting the observed data *assuming* the null is true
  - That is, p-value tells us about $P(data|H_0)$, not $P(H_0|data)$.
  - By Bayes' rule, $P(H_0|Data) = P(Data|H_0) * P(H_0)/P(Data)$. But to formalize this, we need to take a stand on our *prior* belief that H_0 is true, $P(H_0)$.
- People often interpret a $p < 0.05$ as strong evidence of an effect and $p \geq 0.05$ as evidence of no effect.
  - But $p = 0.05$ is a fairly arbitrary threshold.
  - It's better to view the p-value as a spectrum indicating how likely is the observed data given the null.
  - Moreover, p-values can be large even if the null is false (low power!)


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$ = 0.027.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

nature

Published: 26 February 2015

Psychology journal bans *P* values

Chris Woolston

Nature **519**, 9 (2015) | Cite this article nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › social selection › article

Published: 26 February 2015

Psychology journal bans *P* values

Chris Woolston

Nature **519**, 9 (2015) | Cite this article

09 March 2015 This story originally asserted that "The closer to zero the P value gets, the greater the chance the null hypothesis is false." P values do not give the probability that a null hypothesis is false, they give the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It is by convention that smaller P values are interpreted as stronger evidence that the null hypothesis is false. The text has been changed to reflect this.

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?
  - Most variables are not normally distributed!

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?
  - Most variables are not normally distributed!
  - Even when they are, why would we know their variance?!

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?
  - Most variables are not normally distributed!
  - Even when they are, why would we know their variance?!
  - Have I just been wasting your time!?!

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?
  - Most variables are not normally distributed!
  - Even when they are, why would we know their variance?!
  - Have I just been wasting your time!?! No.

What About Non-Normal Data?

- So we know how to test hypotheses and make inferences on means of normal random variables when we know their variance ... so what?
  - Most variables are not normally distributed!
  - Even when they are, why would we know their variance?!
  - Have I just been wasting your time!?! No.

- We will next review powerful **asymptotic** results. These will allow us to apply similar inference tools if the sample is "large" even when Y_i is not normally distributed.