

Econometrics I

Lecture 2: Math Review

Paul T. Scott
NYU Stern

Fall 2021

Today's Roadmap

1 Probability Review

- ▶ Random variables and realizations
- ▶ Conditional probability and Bayes' Rule
- ▶ Means, variances and other moments
- ▶ Conditional moments

2 Statistics Review

- ▶ The Law of Large Numbers
- ▶ The Central Limit Theorem
- ▶ Hypothesis Testing
- ▶ Example: Testing the means of two RVs

3 Linear Algebra Review (very basic)

- ▶ Matrix and vector notation
- ▶ Transposes and inverses
- ▶ Matrix multiplication
- ▶ Matrix “calculus”

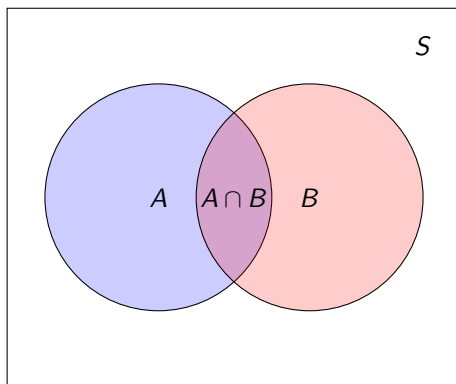
Probability Review

Outcomes, Events and Sets

To discuss probability we first need a few basic definitions:

- An **outcome** is something we *can* observe but may not know in advance
Example: For a coin flip, H (heads) is an outcome
Example: The wage of a randomly sampled worker
- A **sample space** is a set of all possible outcomes
Example: For a coin flip, $\{H, T\}$ is the sample space
Example: For two coin flips, $\{HH, HT, TH, TT\}$ is the sample space
- An **event** is any subset of the sample space
Example: For two coin flips, $\{HH, TT\}$ is the event “getting the same side both times”
- A **probability** is a function from S to $[0, 1]$ such that
 - 1 $P(E) \in [0, 1]$ for any event, E
 - 2 $P(S) = 1$
 - 3 $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$

Outcomes and Events as a Venn Diagram



- A and B are events in the sample space, S
- The intersection, $A \cap B$, is the purple part
- The union, $A \cup B$, would be everything that isn't white

Conditional Probability

- The **conditional probability** of an event, A , given, B , is the probability that A occurs if B is known to have occurred.

Example: If two dice are rolled and sum to 8, what is the probability at least one dice was a 4?

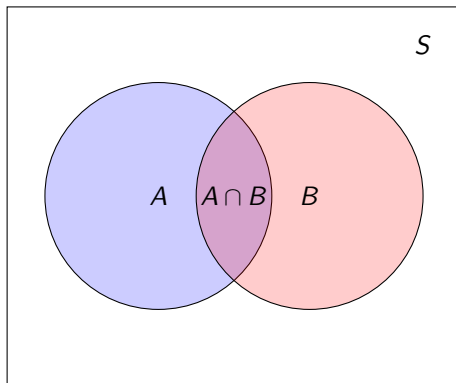
- How to calculate conditional probability?
 - ▶ The new sample space is just B and the event of both A and B happening is AB so a natural definition is:

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

- ▶ Notice that we can do the symmetric thing for $P(B|A)$ and rearrange to get Bayes' Rule:

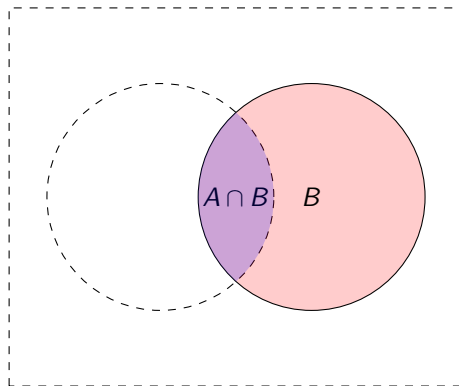
$$P(A|B)P(B) = P(B|A)P(A)$$

Conditional Probability in a Venn Diagram



- The probability of A is the relative area of A
- But what if B definitely occurred?

Conditional Probability in a Venn Diagram



- The probability of A is STILL the relative area of A
- But we only take into account the part of A “inside” B

Example: election odds

Candidate	$Pr(\text{wins primary})$	$Pr(\text{wins general})$	$Pr(\text{wins general} \text{wins primary})$
Biden	.286	.167	.584
Buttigieg	.048	.029	.604
Harris	.111	.059	.531
Sanders	.154	.083	.538
Warren	.333	.182	.547

Independence

- An important definition: A and B are **independent** if $P(A|B) = P(A)$ and vice versa
- Intuitively, independence means that B contains no information about A
- Example: Whether a coin lands heads or tails does not depend on the outcome of any prior flip.
- Is this concept obvious? *Gambler's Ruin*: people observed a roulette wheel turn up black several times and began to bet against black even though each new spin did not depend on the outcome of the previous spin!

Random Variables

- A **random variable** is any function from the sample space S to the real numbers, $\mathbb{R}$

Example: If rolling two dice, the sum is a random variable

Example: ... so is the number that comes up on one die

- We can easily extend the definition of probability to random variables:
 - ▶ The probability a random variable X is equal to x is the probability of all events so that $X(E) = x$. Formally:

$$P(X = x) = P\left(\bigcup_{E: X(E)=x} E\right)$$

Example: If rolling two dice, probability of the sum being 12 is given by:

$$P(D_1 + D_2 = 12) = P(D_1 = 6 \cap D_2 = 6) = 1/36$$

Example: If you measure 100 randomly selected men's heights, the average height is a random variable.

Discrete versus Continuous Random Variables

Intuition for probability is usually discrete (dice rolls) but a lot of randomness is best modeled as continuous (height or wages)...

- For a continuous random variable X the probability that $X = x$ is “0” since any one outcome happens with vanishingly small probability
- Instead we think about *sets* like $P(a < X < b)$
- Define the **Cumulative Distribution Function** of X to be:

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

- The continuous analog of probability for single events is the **Probability Density Function**:

$$f(x) = \frac{d}{dx}F(x)$$

- ▶ This is *not* the probability of observing x
- ▶ It can be bigger than 1!
- ▶ However, it acts like a probability in that it is a “weight”

The Mean of a Random Variable

Often we do not care about RVs *per se* but about certain properties:

- The **Mean** or **Expectation** of a random variable is the probability-weighted average outcome (denoted by $E(X)$ or μ_X)

- ▶ For discrete RVs:

$$E(X) = \sum_x P(X = x) * x$$

- ▶ For continuous RVs:

$$E(X) = \int xf(x)dx$$

- We can easily take the mean of *functions* of random variables:

$$E(g(X)) = \sum_x P(X = x)g(x)$$

Properties of Expectations

- Mean of a constant is a constant:

$$E(a) = a$$

- Linearity:

$$E(aX + bY) = aE(X) + bE(Y)$$

- NOT A Property: Swapping expectations and functions!

$$E(g(X)) \neq g(E(X))$$

- Knowledge check: is it true that $E(XY) = E(X)E(Y)$?

Variance and Covariance

- The **Variance** of a random variable is defined as follows:

$$\text{Var}(X) = E \left((X - E(X))^2 \right)$$

This is a measure of the dispersion of X

Also denoted by σ_X^2

- The **Covariance** of two random variables is defined as follows:

$$\text{Cov}(X, Y) = E \left((X - E(X))(Y - E(Y)) \right)$$

Measure of the tendency of X and Y to move in the same direction

- ▶ If X and Y tend to be far from the mean at the same time then Covariance has large magnitude
- ▶ If X and Y tend to be large at the same time then Cov will be positive
- ▶ If X tends to be large when Y is small then Cov will be negative

More on Variance and Covariance

- **NB:** The covariance and variance of random variables are two of the most important and commonly seen concepts in econometrics! Learn them!
- Useful properties:

$$\text{Cov}(aX + bY, Z) = a \times \text{Cov}(X, Z) + b \times \text{Cov}(Y, Z)$$

$$\text{Cov}(X, X) = \text{Var}(X)$$

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

- Exercise you should have done before: Prove the above!

Conditional Mean (or Conditional Expectation)

- Conditional mean: expected value of RV, X , if event, A , is *known*
Example: Expected value of a dice roll, D , if we *know* that $D \geq 4$
- The **Conditional Mean** of a discrete random variable is given by:

$$E(X|A) = \sum_x P(X = x|A) * x$$

and analogously for continuous random variables.

Example: (From above):

$$E(D|D \geq 4) = 4 \times \frac{1}{3} + 5 \times \frac{1}{3} + 6 \times \frac{1}{3} = 5$$

- The conditional mean is also linear but *treats known entities as constant*
Example:

$$E(Y \times X|Y = y) = y \times E(X|Y = y)$$

Conditional Variance

- The **Conditional Variance** of a random variable is given by:

$$\text{Var}(X|A) = E \left((X - E(X|A))^2 | A \right)$$

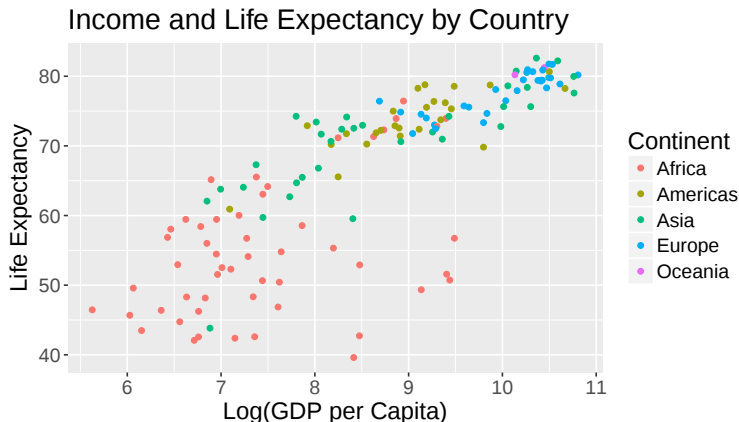
- Conditional variance has the same properties as variance but also treats constant as known:

$$\text{Var}(X \times Y | Y = y) = y^2 \times \text{Var}(X | Y = y)$$

- With two random variables, can define the **Conditional Covariance**:

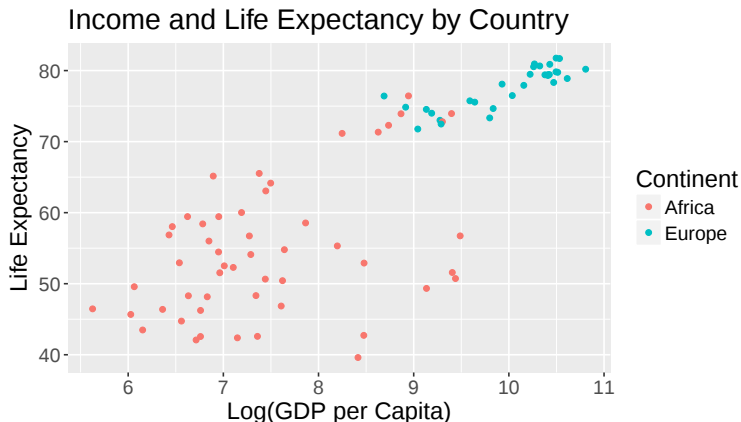
$$\text{Cov}(X, Y|A) = E \left((X - E(X|A))(Y - E(Y|A)) | A \right)$$

Conditional Variance: Real World Example



- There's clearly some upward correlation between income and health
- But how does this relationship look within continents?

Conditional Variance: Real World Example



- The variance in outcomes in Africa is huge while in Europe it's small
- If one only focused on Africa, there is barely any relationship visible

Facts and Theorems about Conditional Moments

- **Key Concept:** A mean or a variance is a *number*; the conditional mean or conditional variance is a *function*!

Example Consider a dice roll D :

- 1 $E(D) = 3.5$
 - 2 $E(D|D \geq d)$ is a function of little d !
- Important fact: If X and Y are independent:

$$E(X|Y) = E(X)$$

It should be easy to prove this yourself from the definition of the conditional mean.

Facts and Theorems about Conditional Moments, Cont'd

Key Theorem 1: The Law of Total Expectation:

$$E(E(X|A)) = E(X)$$

- In words: The weighted average of conditional means of random variable is just the unconditional mean
- *Example:* Calculating the average SAT score, S , of a college student:
 - 1 Calculate the mean across all students, $E(S)$:

$$\mu_{SAT} = \frac{1}{N} \times \sum_{EVERYONE} SAT_i$$

- 2 Calculate the mean at each university, $E(S|U = u)$ and then take the population-weighted mean at each university, $E(E(S|U = u))$:

$$\mu_{SAT} = \mu_{SAT,NYU} \times P(NYU) + \mu_{SAT,Columbia} \times P(Columbia) + \dots$$

Facts and Theorems about Conditional Moments, Cont'd

- **Key Theorem 2:** The Law of Total Variance.

$$\text{Var}(X) = E(\text{Var}(X|A)) + \text{Var}(E(X|A))$$

- ▶ In words: The variance of X is the average variance of X at different outcomes of A and the variance of the mean of X at different values of A
- ▶ *Example:* The variance in income across countries, X , is the average variance of income *within* countries, $\text{Var}(X|A)$ plus the variance of average income *between* countries, $E(X|A)$
- ▶ **Alert:** Conditional moments are very important in econometrics

Sequences of Random Variables and Limits

- In addition to random variables we can define a **sequence of random variables**, X_n as a sequence of functions from an underlying probability space to $\mathbb{R}$

Example: The sum of n dice rolls

- Sequences in calculus have a sense of convergence
- Random variables are more complicated because they are random. There are three types of convergence:

- 1 **Almost Sure Convergence** (Not going to use this, just here for completeness)
- 2 **Convergence in Probability**
- 3 **Convergence in Distribution**

Almost Sure Convergence

- X_n is said to converge almost surely to X if for any $\varepsilon > 0$

$$P\left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon\right) = 1$$

- If X is a constant (i.e., just a number), μ , then the probability of drawing a sequence X_n so that $\lim_{n \rightarrow \infty} X_n \neq \mu$ goes to 0

Similar to standard definition: for *any* given sequence, eventually that specific sequence will settle down.

Example: Let X_n be maximum value of dice roll for n throws of dice. X_n converges almost surely to 6 since the probability of *never* rolling 6 goes to zero

Convergence in Probability

- X_n is said to converge in probability to X (or we say X_n is **consistent** for X) if for any $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$$

- If X is a constant (i.e., just a number), μ , then as n gets large the probability that $X_n \neq \mu$ becomes 0
- The difference between convergence almost surely and in probability is subtle (and honestly not important for this course)
- Key takeaway: for a random variable, even as N gets large, there can be some probability that something crazy happens. Convergence concepts are a mathematical way of saying this probability vanishes.

Convergence in Distribution

- A random variable X_n converges in distribution to X if the cdfs converge:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

- Intuition: as n gets large, X_n is still a random variable (rather than converging to a number), and it behaves like X in terms of probabilities of events
- We will visualize this below when reviewing the Central Limit Theorem

Statistics Review

Basic Definitions

- We start with a **sample** of data that consists of a series of observations
- Each ob is a realization of a random variable from an underlying distribution called the **data generating process** or the **population**
 - ▶ If all observations come are independent and come from same distribution, then data is said to be **independently and identically distributed** (iid)
 - ▶ A **Simple Random Sample** is a set of independent draws from the same distribution, and is guaranteed to yield iid data
- A **Statistic** is any function from the data to $\mathbb{R}$
- A **Parameter** is a number that characterizes the population
- Some notation:
 - ▶ Index observations with a subscript: so X_i is i^{th} observation
 - ▶ Represent data as a random variable with a capital letter, X
 - ▶ Represent a realization with a lower case letter, x

Statistics are random variables!

- A random sample itself is a random variable
 - ▶ Why: Two different random samples have different numbers, and so are different realizations from the same distribution
 - ▶ Each observation is *also* a random variable

Example: If a roll a dice 5 times, ONE sample would be $\{H, H, T, T, H\}$ but another sample could be $\{H, T, T, H, T\}$.

- This also means that statistics (functions of the sample) are random variables as well

Example: The sample mean $\bar{X}$ is a random variable *before* data is observed (but a number after)

An Example with Coin Flips

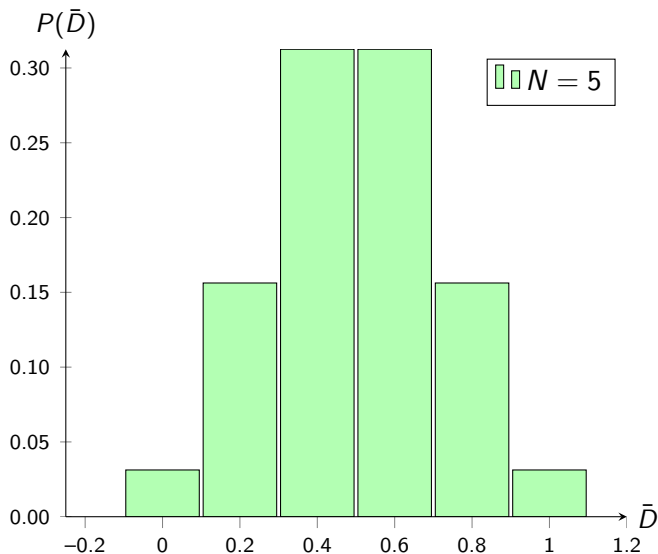
Consider a fair coin and labeling heads as 0 and tails as 1.

- Calling the dice outcome, D :

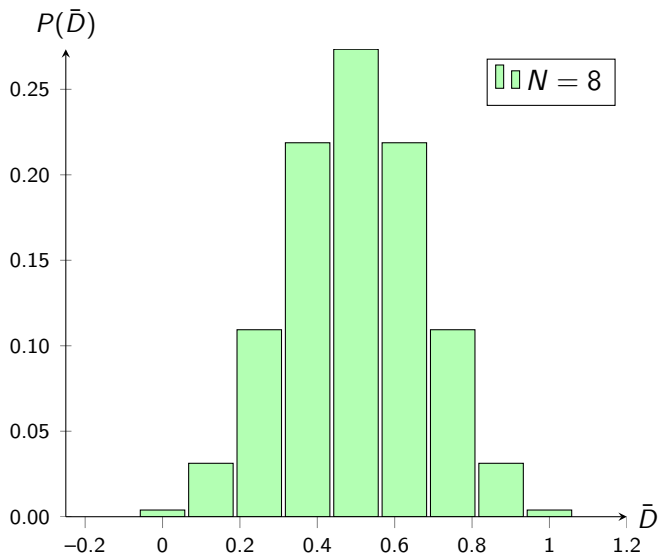
$$E(D) = .5 \times 1 + .5 \times 0 = .5$$

- Now consider an experiment of flipping the coin 5 times:
 - ▶ For each sample, $\{D_1, \dots, D_5\}$, calculate $\bar{D}$
 - ▶ Example 1: $\{H, H, T, T, H\}$ implies $\bar{D} = .4$
 - ▶ Example 2: $\{T, H, T, T, H\}$ implies $\bar{D} = .6$
- For different samples, different values of $\bar{D}$, so what is the distribution?
 - ▶ For each possible value of $\bar{D}$ need the probability of all possible flips (events) that yield that value

An Example with Coin Flips: Visualization



An Example with Coin Flips: Changing Sample Size



Sample Moments versus Population Moments

- Many statistical models contain some parameter that we wish to estimate

Examples: Mean of an RV, μ , of the correlation between X and Y , ρ_{XY}

- Statistics that estimate parameters are called estimators or estimates
 - The **sample mean** (or sample variance, etc.) is the mean *of the sample*. It is an example of a statistic
 - For a sample of data, $X_1, \dots, X_N$ the sample mean is given by,

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

- This is NOT the population mean, $E(X)$ (a parameter)
- We often denote an estimator with a “hat”: $\bar{X} = \hat{\mu}$.

How do we pick an estimator?

- Statistics and estimates do not fall from the sky
- We like or dislike different estimators based on desirable properties and whether they work with our modeling assumptions
- List of useful properties (we'll see these again):
 - ① **Unbiasedness:** $E(\hat{\mu}) = \mu$
 - ② **Consistency:** $\lim_{N \rightarrow \infty} \hat{\mu} \xrightarrow{P} \mu$
 - ③ **Efficiency:** Whether or not $Var(\hat{\mu}_X)$ is large or small.

Goal of econometrics: find estimators that have as many of these properties fulfilled as we can

The Power of Large Samples

- $\bar{X}$ matters because it is a good predictor of μ_X
- In general, we care about statistics that are informative about important parameters of the population

Sample Variance $\frac{1}{N-1} \sum_i (X_i - \bar{X})^2 \Leftrightarrow$ Population Variance

Sample Covariance $\frac{1}{N-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \Leftrightarrow$ Population Covariance

- **Theorem: The Law of Large Numbers** For a SRS, as the sample size, N , becomes large, the sample mean, $\bar{X}$ will converge in probability to μ_X

- ▶ Some technical conditions are needed for formal proof – basically need $\sigma_X^2 < \infty$
- ▶ Stronger versions of this theorem exist, but above is good enough
- ▶ Corollary: for (most) functions:

$$\frac{1}{N} \sum_{i=1}^N f(X_i) \rightarrow E(f(X))$$

- The LLN is *hugely* important because it guarantees that sample moments converge to population moments!

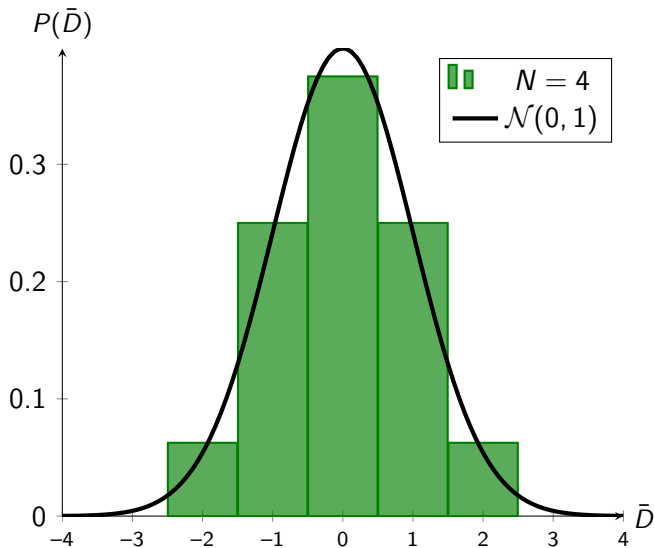
The Central Limit Theorem

- The LLN says that $\bar{X}$ will eventually get close to μ_X , but how close?
- **Theorem: The Central Limit Theorem** For a sequence of iid variables, X_i , where $E(X) = \mu$ and $Var(X) = \sigma^2$:

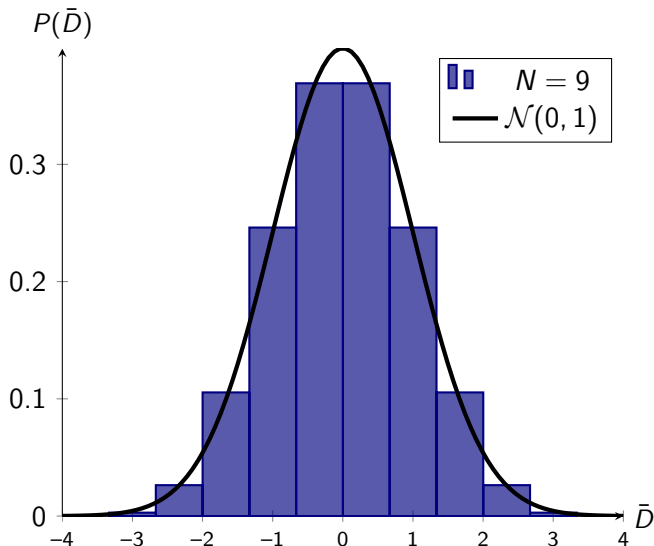
$$\lim_{n \rightarrow \infty} \sqrt{n} \times \frac{\bar{X} - \mu}{\sigma} \xrightarrow{\text{dist.}} \mathcal{N}(0, 1)$$

- Says that for n large, sample means will be approximately normally distributed NO MATTER HOW X is DISTRIBUTED
- **NB:** The exercise is treating $\bar{X}$ as a random variable, so it says that for REPEATED draws of a sample of data, the distribution of the mean across samples will look a certain way.

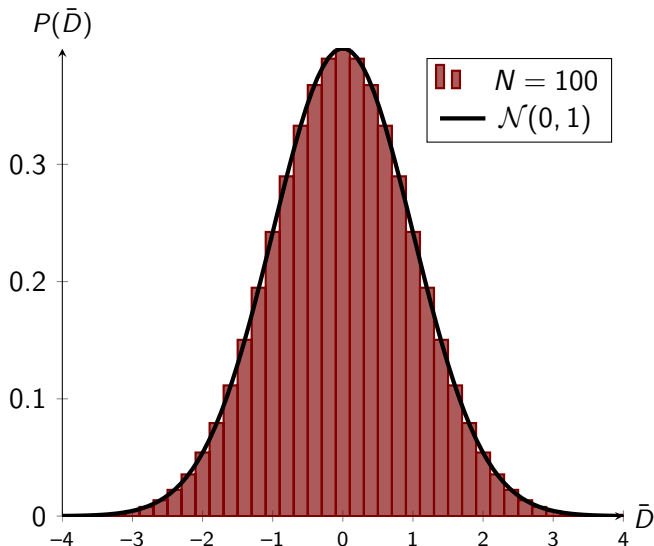
Visualizing the CLT



Visualizing the CLT



Visualizing the CLT



Interlude: The Normal Distribution...

- Normal distribution is the most important in statistics.
- Define the **Standard Normal Distribution** to be $\mathcal{N}(0, 1)$ and denote it by Z . We use Φ for the cdf of Z and ϕ for the pdf.
- Key Properties:
 - ▶ Linearity: If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are both normal with covariance σ_{XY} , then,

$$aX + bY \sim \mathcal{N}\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\right)$$

NB: Any Normal RV has cdf $\Phi((x - \mu)/\sigma)$ (called standardizing)

- ▶ Symmetry about the mean:

$$\Phi\left(\frac{x - \mu}{\sigma}\right) = 1 - \Phi\left(-\frac{x - \mu}{\sigma}\right)$$

... and its Cousins

- A squared standard normal random variable, Z^2 is called a $\chi^2(1)$ (**chi-squared** of degree 1) random variable
 - ▶ The sum of q independent $\chi(1)^2$ RVs is a $\chi^2(q)$ random variable. Called “chi-squared with q degrees of freedom”
 - ▶ Arises naturally when squaring things like sample means
- If $Z \sim \mathcal{N}(0, 1)$ is normal and V is $\chi^2(q)$ then $Z/\sqrt{V/q}$ is defined as **t-distributed** random variable with q degrees of freedom.
 - ▶ Arises naturally in stats whenever sample is drawn from a normal distribution
 - ▶ Limit as $q \rightarrow \infty$ is normal
- If U follows a t -distribution then U^2 follows an **F-distribution**

Hypothesis Testing: Introduction

- Often we are interested in making inference about a sample or samples.

Example 1: Is the mean income in two countries different?

Example 2: Is the mean of a sample greater than some number μ ?

- The central issue: data is noisy and random $\Rightarrow$ two numbers will rarely be *exactly* the same
- **Hypothesis Test:** If we *assume* a specific hypothesis is true, then the likeliness of the observed data is informative about the likeliness of the hypothesis.

Intuition: If a coin is fair, then getting 50 heads in a row is *very unlikely*
 $\Rightarrow$ the coin is probably not fair

Example: If a random variable is distributed standard normal Z then observing a number greater 3 than would only happen with .14% probability $\Rightarrow$ RV is probably *not* standard normal.

Hypothesis Testing: More Formally

- **Null Hypothesis:** Denoted H_0 . A hypothesis the researcher assumes is true (e.g., $\mu_X = 0$)
- **Alternative Hypothesis:** Denoted H_a . An alternative to the null (e.g., $\mu_X \neq 0$)
- **Test Statistic:** A function of the data that is distributed differently between the null and alternative hypotheses.

Hypothesis Testing: More Formally

- **α -Level Test:** Rejecting the null hypothesis if the test statistic occurs with less than $\alpha\%$ probability under the null.
 - ▶ α is the **Type I error:** $\alpha\%$ of the time, a null will be incorrectly rejected
 - ▶ Related concept is **Type II error:** the probability a significant result is treated as null
- *In math notation:* Given a test statistic, U , with realization u , a two-sided α -level test will reject the null if $P(U < u \cup U > u | H_0) > \alpha$
- One-sided versus Two-sided tests: If we “know” that a parameter has some restrictions (e.g., $\mu_X > 0$) then we can may only test if $u > U$ but the idea of the test is the same

An Example: Testing Means

Consider two independent samples, X_i and Y_i , of income from different countries and test if the mean income is the same.

- Step 1: Write down the hypotheses:

$$H_0 : \mu_X = \mu_Y$$

$$H_a : \mu_X \neq \mu_Y$$

- Step 2: Construct a test statistic.

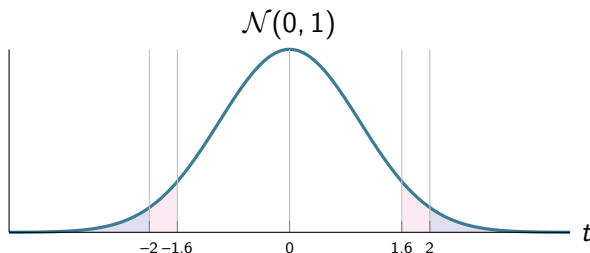
- ▶ From LLN and CLT: If n is large then $\bar{X}$ is approximately normal.
Denoted $\bar{X} \stackrel{a}{\sim} \mathcal{N}(\mu_X, \sigma^2/n)$
- ▶ Same is true for Y and since Normal RVs are linear:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{(\sigma_X^2 + \sigma_Y^2)/n}}$$

- ▶ IF H_0 true THEN $t \stackrel{a}{\sim} \mathcal{N}(0, 1)$

An Example: Testing Means (Cont'd)

- Step 3: Choose a level α and determine critical values of the test



- Step 4: Check if t lies outside of the critical region, if so reject the null hypothesis.

Intuition: t should only be in the purple region 10% of the time if the null is true. This is unlikely enough that we consider it evidence against the null.

- ▶ For 10% test, cv is 1.65, for 5% tests, cv is 1.96

On standard deviations and standard errors

- Where does the “n” come from in the CLT?
- It represents the fact that the variance of the *estimator* is shrinking as n gets big
- **Standard error**: the standard deviation of the *estimator*
 - ▶ For a sample mean: $\sigma_{\bar{X}} = \sigma_X / \sqrt{N}$
 - ▶ This relates the standard deviation and the standard error
- Often times we don't know σ_X so we have estimate it using:

$$s_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

- Technicality: Now we have an estimated parameter (s_X in the denominator, NOT σ_X)
 - ▶ Technically need to use a t-distribution with n degrees of freedom, not a normal
 - ▶ Distinction vanishes as N gets big (and most data sets have large N)
 - ▶ We will not dwell on this

A test-statistic carries a lot of information...

- An alternative to doing a test is to report a p-value
- **p-value:** *Given* the null hypothesis, what is the probability of drawing a value of $\bar{X}$ at least as far in the tails of the distribution as the *observed* value of $\bar{X}$

▶ Mathematically:

$$p = P\left(|\bar{X} - \mu_X| > |\bar{X}^{data} - \mu_X| \mid H_0\right)$$

- ▶ In principle this depends on the distribution of $\bar{X}$
- ▶ With the CLT approximation, for a two-sided p-value:

$$p \approx 2\Phi(-|t|)$$

- p-values are the smallest possible α test that would reject the null

Confidence Intervals

Another way to give us the information in a test is to construct a confidence interval:

- **Confidence Interval:** *Given a sample mean, a $\alpha\%$ CI is a set that contains the population parameter with $\alpha\%$ probability.*
- Idea:
 - 1 Pick a random null hypothesis, μ_0
 - 2 Is this reject by a $1 - \alpha$ -level test?
 - 3 If NO, put it in the confidence interval
 - 4 Do this for all possible values of μ_0
- In other words: a confidence interval is all possible hypotheses that we could not reject at $\alpha\%$ probability
- In general, also depends on the distribution of $\bar{X}$
- With CLT approximation:

$$CI_\alpha = \bar{X} \pm cv_\alpha \times \sigma_{\bar{X}}$$

Recapping

What are the main ideas to remember going forward?

- ① Statistics is about finding parameter estimates with desirable properties
 - ▶ Estimates themselves are RVs
 - ▶ Properties we like: not being wrong as often as possible
- ② The tools of the trade boil down to the CLT and the LLN
- ③ Because of randomness, to do inference we need to do hypothesis testing
 - ▶ A Hypothesis Test tells us how likely a sample is given a parameter value in the population
 - ▶ Many ways to summarize the same info: t-statistic, p-value, CI

Linear Algebra Review

Basic Definitions

- A **vector** in $\mathbb{R}^n$ is a column of numbers $(x_1, x_2, \dots, x_n)$.
- A **matrix** in $\mathbb{R}^{n \times m}$ is m columns of length n vectors (so n is the number of rows and m is the number of columns). We denote an element of a matrix by m_{ij} for row i and column j :

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$$

- For an entity on which there are many pieces of data, we store the data in a vector x_i

Example: For the USA could have $x_{USA} = (GDP_{USA}, Population_{USA}, \dots)$

- For many entities we can store all the data in a data matrix, X .

Example: For two countries:

$$X = \begin{pmatrix} GDP_{USA} & Population_{USA} \\ GDP_{Canada} & Population_{Canada} \end{pmatrix}$$

Matrix Multiplication

- For two vectors of equal length define the **dot product** as,

$$v \cdot w = \sum_{i=1}^n v_i \times w_i$$

- For two matrices, A and B of sizes $n \times m$ and $m \times k$ define the **matrix product** $C = AB$ as the $n \times k$ matrix with entries $c_{ij} = \sum_{l=1}^m a_{il} b_{lj}$
 - ▶ Easy way to remember: $(i, j)^{th}$ element of product is dot product of i^{th} row and j^{th} column of A and B respectively.
 - ▶ Not all matrices can be multiplied: left matrix must have column length equal to right matrix's row length
 - ▶ Multiplication is NOT commutative: $AB \neq BA$ even if they both exist

Transposes

- Define the **transpose** of A as the matrix A' with elements $a'_{ij} = a_{ji}$ (reverse columns and rows)
- A matrix is **symmetric** if $A' = A$
- **Important Properties:**
 - ▶ The matrix $B = A'A$ is *always* a square matrix
 - ▶ The matrix $B = A'A$ is always symmetric
 - ▶ $(A')' = A$
 - ▶ **Multiplication Rule:** $(AB)' = B'A'$
 - ▶ **Addition Rule:** $(A + B)' = A' + B'$

Inverses

- The **Identity Matrix**, I , is a matrix with 1s on the diagonal and 0s elsewhere. Clearly $AI = A$.
- Define the **left inverse** of A to be the matrix A^{-1} such that $A^{-1}A = I$
 - ▶ Can analogously define right inverse
 - ▶ Right and left inverse will NOT be the same if A is not a square matrix
 - ▶ Right and left inverse WILL be equal if A is square (then we just say inverse)
- **Important Properties:**
 - ▶ **Multiplication Rule:** $(AB)^{-1} = B^{-1}A^{-1}$
 - ▶ **Transpose Rule:** $(A')^{-1} = (A^{-1})'$
 - ▶ **Dot Product:** $v \cdot w = v'w$

Matrix Calculus

- For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ recall the definition of the derivative or Jacobian of f :

$$Df = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

- We WON'T be doing anything too complicated! But we can define two important functions given a vector x and a matrix A :
 - ▶ For Ax , $D(Ax) = A$ (as a line in 1-D calc)
 - ▶ For $x'Ax$, $D(x'Ax) = x'(A + A')$ (as a quadratic in 1-D calc)