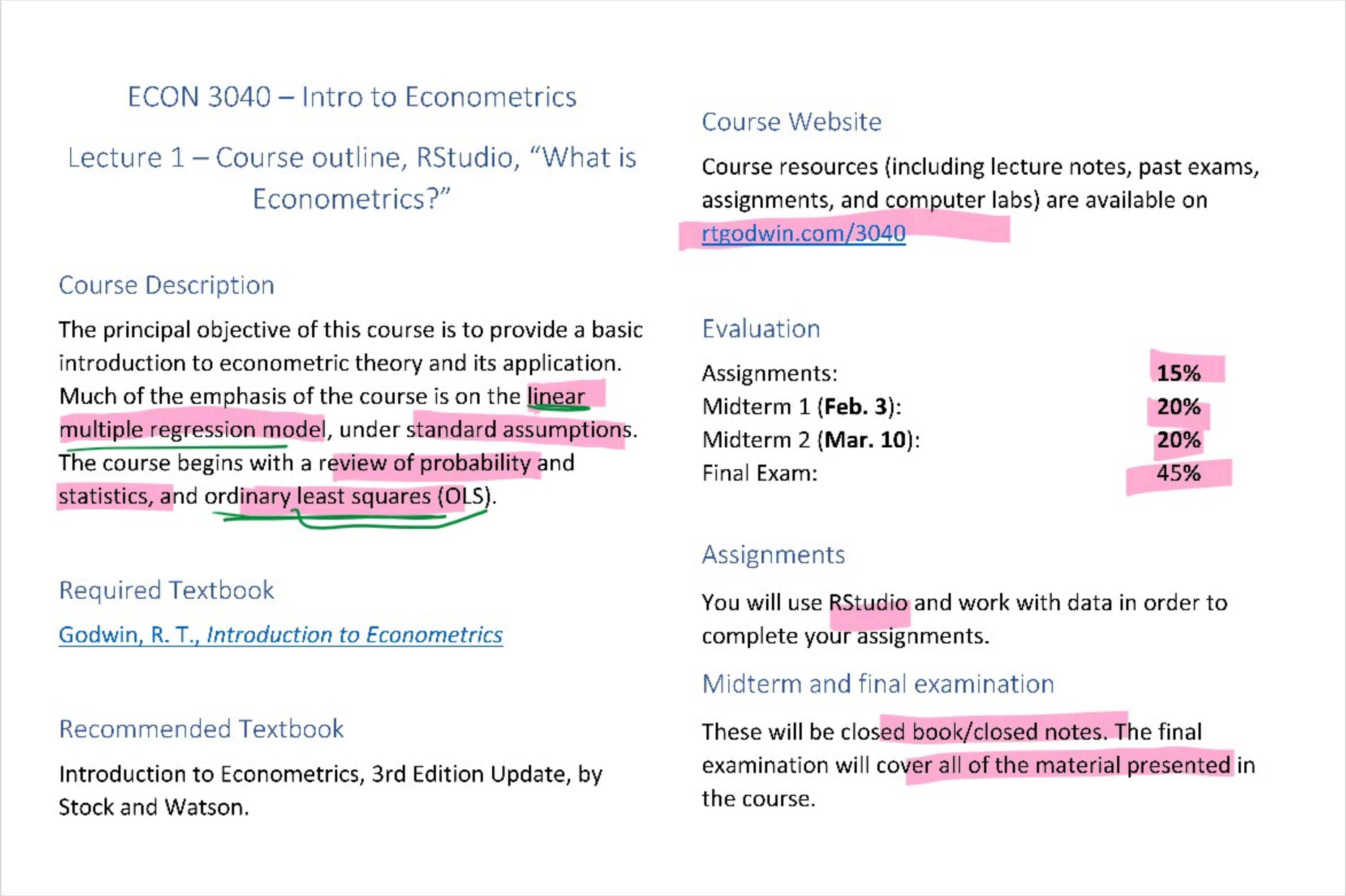
Chapter 1 intro

Monday, January 6, 2025 10:29 AM

slides1



ECON 3040 – Intro to Econometrics

Lecture 1 – Course outline, RStudio, "What is Econometrics?"

Course Description

The principal objective of this course is to provide a basic introduction to econometric theory and its application. Much of the emphasis of the course is on the linear multiple regression model, under standard assumptions. The course begins with a review of probability and statistics, and ordinary least squares (OLS).

Required Textbook

Godwin, R. T., *Introduction to Econometrics*

Recommended Textbook

Introduction to Econometrics, 3rd Edition Update, by Stock and Watson.

Course Website

Course resources (including lecture notes, past exams, assignments, and computer labs) are available on rtgodwin.com/3040

Evaluation

Assignments:	**15%**
Midterm 1 (**Feb. 3**):	**20%**
Midterm 2 (**Mar. 10**):	**20%**
Final Exam:	**45%**

Assignments

You will use RStudio and work with data in order to complete your assignments.

Midterm and final examination

These will be closed book/closed notes. The final examination will cover all of the material presented in the course.

Grading scale

A+	93 – 100
A	87 – 93
B+	80 – 87
B	72 – 80
C+	64 – 72
C	57 – 64
D	50 – 57
F	0 – 50

- A missed assessment will result in make-up work, or reweighting of your grade.
- Mar. 19 is the last day for Voluntary Withdrawal from courses.

Academic Integrity

- All assignments and exams must be completed independently.
- Do not engage in "contract" cheating.
- Do not provide your UM Learn login information to anyone else. This is "personation", a serious form of academic misconduct.

Ignorance is not a defense. Familiarize yourself with section 2.5 of Academic Misconduct Procedures.

I own the copyright to all course content. Sharing my content (e.g. on Course Hero) is illegal!

All course material is copyrighted by Ryan Godwin, 2025. No audio or video recording of this material, lectures, or presentations is allowed in any format, openly or surreptitiously, in whole or in part without permission of Ryan Godwin. Course materials are for the participant's private study and research, and must not be shared. Violation of these and other Academic Integrity principles, will lead to serious disciplinary action.

Tentative Course Topics

- Review of Probability
- Review of Statistics
- Linear Regression with One Regressor
- Hypothesis Tests
- Linear Regression with Multiple Regressors
- Hypothesis Tests in Multiple Regression
- Nonlinear Regression Functions
- Instrumental Variables ← DiD
- Heteroskedasticity

Student Accessibility Services

Students with disabilities should contact Student Accessibility Services to facilitate the implementation of accommodations, and meet with me to discuss the accommodations recommended by Student Accessibility Services.

Academic Supports

Sample Lecture

What is Econometrics?

- Econometrics is a subset of statistics
- Science of testing economic theories
- Used to estimate causal effects
- Used to forecast or predict (not covered in this course)
- Often characterized by "observational data"

Causal Effects

Economic models often suggest that one variable causes another. This often has *policy implications*. The economic models, however, do not provide *quantitative magnitudes* of the causal effects.

---

tax ✓ $1 tax → ↓ #

For example:

- How would a change in the *price* of alcohol or cigarettes effect the *quantity* consumed?
- If *income* increases, how much of the increase will be *consumed*?
- If an additional *fireplace* is added to a house, how much will the *price* of the house increase?

Chapter 6

- How does another year of *education* change *earnings*?

Using data to estimate causal effects

An experiment would be best.

- ~~How would you determine the effect of fertilizer on crop yield?~~
- ~~How would you use an experiment to determine the above four causal effects (on the previous slide)?~~
- What is the advantage of experiments? "gold standard"

the previous slide)?
- What is the advantage of experiments? "gold standard"

→ randomly assign individuals

→ takes care of confounding factors

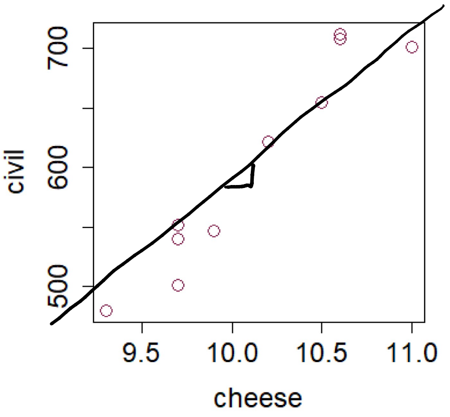
Economic experiments are usually unethical and/or too expensive.

We usually don't have *experimental* data in econometrics – we have *observational* data.

There are issues when dealing with observational data:

- Omitted variables
- Simultaneous causality
- Correlation vs. causation

Civil engineering PhDs awarded, and per-capita consumption of cheese, from 2000-2009 in the U.S. (Spurious correlations, Tyler Vigen)



What is wrong with the above picture?

Objectives of this course

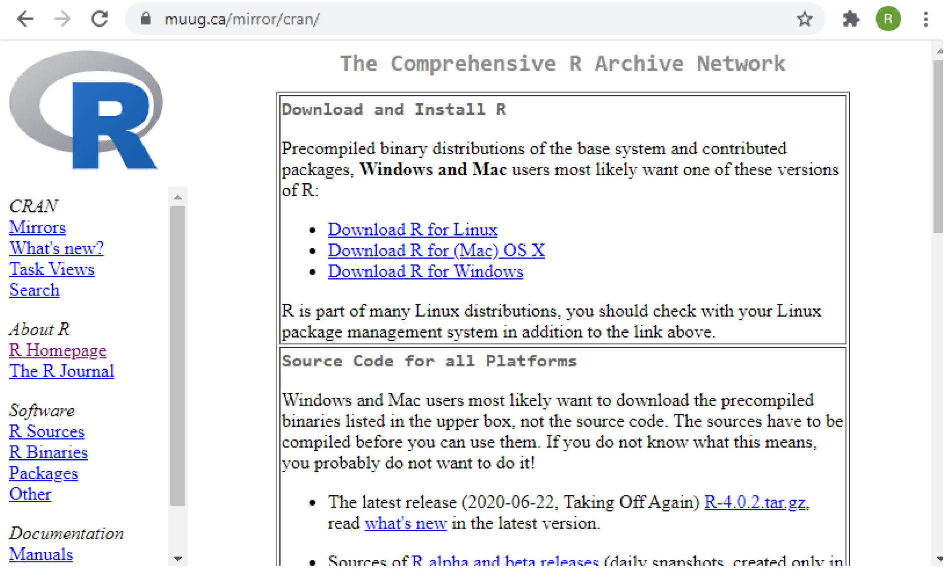
- Learn a method for estimating causal effects (least squares, “LS”)
- Understand some theoretical properties of LS
- Learn about hypothesis testing
- Practice LS using data sets

R and RStudio

The theory and concepts presented in this course will be illustrated by analysing several data sets. Data analysis will be accomplished through the R Statistical Environment and RStudio. Both are free, and R is fast becoming the best and most widely used statistical software.

First, install R

- Go to https://muug.ca/mirror/cran/
- Choose Windows or Mac



The Comprehensive R Archive Network

`Download and Install R`

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

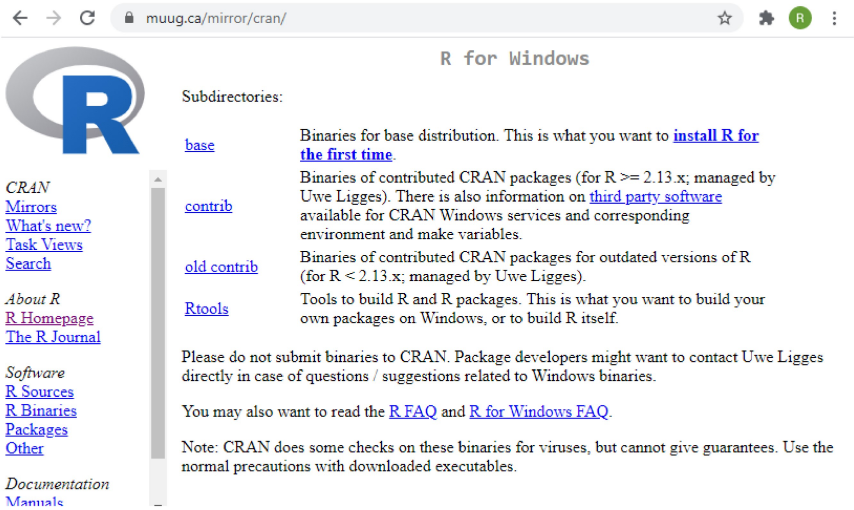
`Source Code for all Platforms`

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-06-22, Taking Off Again) R-4.0.2.tar.gz, read what's new in the latest version.

- Sources of R alpha and beta releases (daily snapshots, created only in

- Click “install R for the first time”



R for Windows

Subdirectories:

base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

- Click “Download R 4.4.1 for Windows” (or Mac)
- Run the “.exe” file
- Click “Next” a bunch of times
- Don’t download RTools!

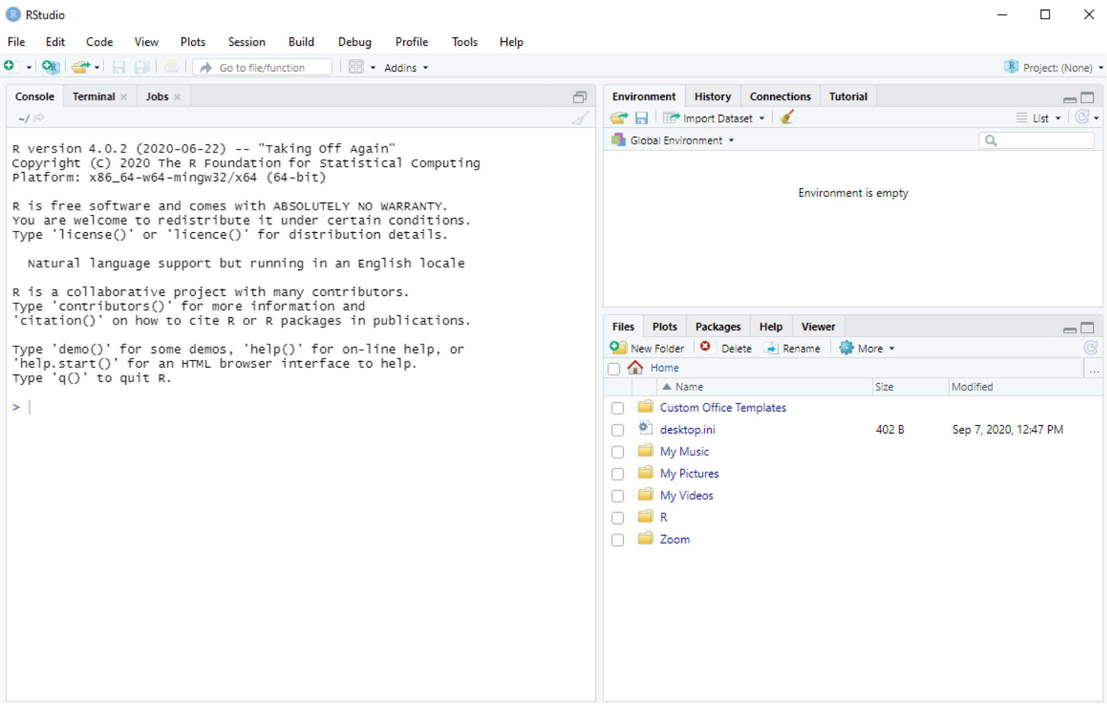
Second, install RStudio

- Go to https://rstudio.com/products/rstudio/download/
- Scroll down until you see the download button "Download RStudio Desktop for Windows (Mac)". Click it.

Step 2:
Install RStudio Desktop

Size: 202.76MB | SHA-256: FD8EA4B4 | Version: 2022.12.0+353 |
Released: 2022-12-15

- Run the ".exe"
- Keep clicking "next" / "install"
- Find RStudio on your computer and open it. It should look something like this:

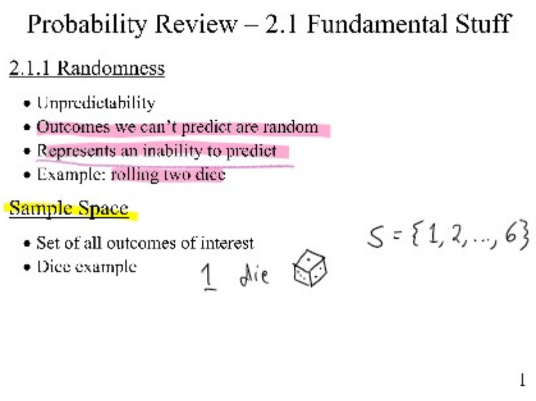


```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Environment is empty

Name	Size	Modified
☐ Custom Office Templates		
☐ desktop.ini	402 B	Sep 7, 2020, 12:47 PM
☐ My Music		
☐ My Pictures		
☐ My Videos		
☐ R		
☐ Zoom		

slides 2



Probability Review – 2.1 Fundamental Stuff

2.1.1 Randomness

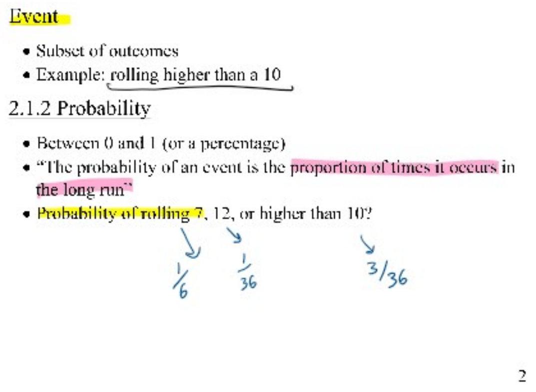
- Unpredictability
- Outcomes we can't predict are random
- Represents an inability to predict
- Example: rolling two dice

Sample Space

- Set of all outcomes of interest
- Dice example

1 die

$S = \{1, 2, \ldots, 6\}$



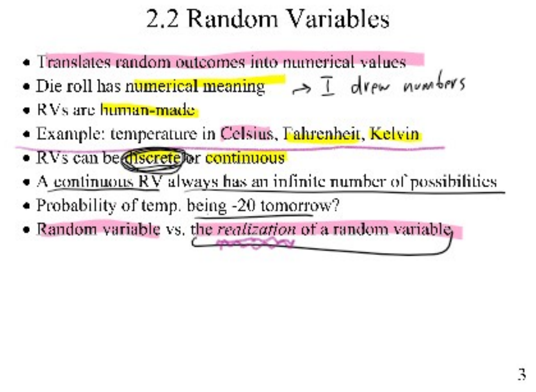
Event

- Subset of outcomes
- Example: rolling higher than a 10

2.1.2 Probability

- Between 0 and 1 (or a percentage)
- "The probability of an event is the proportion of times it occurs in the long run"
- Probability of rolling 7, 12, or higher than 10?

$\frac{1}{6}$ $\frac{1}{36}$ $3/36$



2.2 Random Variables

- Translates random outcomes into numerical values
- Die roll has numerical meaning → I drew numbers
- RVs are human-made
- Example: temperature in Celsius, Fahrenheit, Kelvin
- RVs can be discrete or continuous
- A continuous RV always has an infinite number of possibilities
- Probability of temp. being -20 tomorrow?
- Random variable vs. the *realization* of a random variable

2.3 Probability function

Probability function = probability distribution = probability distribution function (PDF) = probability mass function (PMF) = **probability function**

list sample space

- Usually an equation
- Probability function: (i) lists all possible numerical values the RV can take; (ii) assigns a probability to each value.
- Prob. function contains all possible knowledge we can have about an RV
- 2.3.1 Example: die roll

$$Pr(Y = y) = \frac{1}{6};\ y = 1, \ldots, 6 \tag{2.2}$$

parameters (unknown numbers)

- 2.3.2 Example: a normal RV

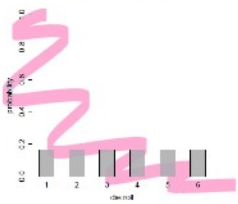
$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y - \mu)^2}{2\sigma^2} \tag{2.3}$$

- Probability function for die roll in a picture:

mean
"2"

variance
"1.5"

Figure 2.1: Probability function for the result of a die roll



2.3.3 Probabilities of events

Probability function can be used to calculate the probability of events occurring.

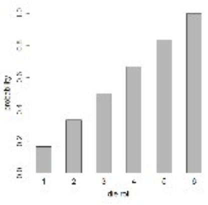
Example. Let Y be the result of a die roll. What is the probability of rolling higher than 3?

$$Pr(Y > 3) = Pr(Y = 4) + Pr(Y = 5) + Pr(Y = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

2.3.4 Cumulative distribution function (CDF)

- CDF is related to the probability function
- It's the prob. that the RV is *less than or equal to* a particular value
- In a picture:

Figure 2.2: Cumulative density function for the result of a die roll



2.4 Moments of a random variable

- "Moment" refers to a concept in physics
- 1st moment is the mean
- 2nd (central) moment is the variance
- 3rd is skewness
- 4th is kurtosis
- Covariance and correlation is a mixed moment

Moments summarize information about the RV. Moments are obtained from the probability function

2.4.1 Mean (expected value)

- Value that is expected
- Average through repeated realizations of the RV
- Determined from the probability function (do some math to it)
- Mean is summarized info that is already contained in the prob. function

- Let Y be the RV
- Mean of Y = expected value of Y = μ_Y = $E[Y]$
- If Y is discrete:

The mean is the weighted average of all possible outcomes, where the weights are the probabilities of each outcome.

The equation for the mean of Y (Y is discrete):

$$E[Y] = \sum_{i=1}^{K} p_i Y_i \tag{2.5}$$

where p_i is the probability of the ith event, Y_i is the value of the ith outcome, and K is the total number of outcomes (K can be infinite). Study this equation. It is a good way of understanding what the mean is.

Exercise: calculate the mean die roll. $E[Y]=3.5$

What are the *properties* of the mean?

Let Y be result of a die roll.

$$E[Y] = \tfrac{1}{6}(1) + \tfrac{1}{6}(2) + \tfrac{1}{6}(3) + \ldots + \tfrac{1}{6}(6)$$

$$= 3.5$$

"constant"

Properties of Expected Value

$E[cY] = c\,E[Y]$ → Example: Let $Z = 2Y$

$E[Z] = 7$

$E[c+Y] = c + E[Y]$ → $W = Y + 1$

$E[W] = 4.5$

$E[c] = c$

$E[X+Y] = E[X] + E[Y]$ → X is another "regular" die

another r.v.

$E[X+Y] = 7$

The equation for the mean of y (y is continuous):

Let y be a random variable. The mean of y is

$$E[y] = \int y f(y)\, dy$$

If y is normally distributed, then $f(y)$ is equation (2.3), and the mean of y turns out to by μ. You do not need to integrate for this course, but you should have some idea about how the mean of a continuous random variable is determined from its probability function.

The *mean* is different from the *median* and the *mode*, although all are measures of central tendency.

The mean is different from the sample mean or sample average. The mean comes from the probability function. The sample mean/average comes from a sample of data.

2.4.3 Variance

- Measure of the *spread* or *dispersion* of a RV
- Denoted by σ^2. The variance of y would be σ_y^2 and the variance of X would be σ_X^2
- Variance is the expected squared difference of a variable from its mean
- Equation:

$$E\left[(Y - E[Y])^2\right] = E[Y - \mu_Y]^2$$

2.4.3 Variance

- Measure of the *spread* or *dispersion* of a RV
- Denoted by σ^2. The variance of y would be σ_y^2 and the variance of X would be σ_X^2
- Variance is the expected squared difference of a variable from its mean
- Equation:

$$\mathrm{Var}(Y) = \mathrm{E}[(Y - \mathrm{E}[Y])^2] \tag{2.6}$$

When Y is a discrete random variable, then equation (2.6) becomes

$$\mathrm{Var}(Y) = \sum_{i=1}^{K} p_i \times (Y_i - \mathrm{E}[Y_i])^2 \tag{2.7}$$

- For variance (the 2nd moment), we are taking the expectation of a squared term
- For skewness (the 3rd moment), we would take the expectation of a cubed term, etc.

Exercise: calculate the variance of a die roll

$$var(Y) = \tfrac{1}{6}(1-3.5)^2 + \tfrac{1}{6}(2-3.5)^2 + \ldots + \tfrac{1}{6}(6-3.5)^2 \approx 2.92$$

What are the *properties* of the variance?

$$var[cY] = c^2 var[Y] \mid var[c+Y] = var[Y] \mid var[c] = 0$$

Exercise: I change the sides of the die to equal 2,4,6.8,10,12. What is the mean and variance of the die roll?

Exercise: What is the mean and variance of the sum of two dice?

$$var[X+Y] = var[X] + var[Y] + 2cov(X,Y)$$

$$P(Y=y) = \frac{1}{12} \; ; \; y = 1,3,4,5,6$$

$$P(Y=2) = \frac{7}{12}$$

$$E[Y] = \frac{1}{12}(1) + \frac{7}{12}(2) + \ldots + \frac{1}{12}(6) \approx 26$$

$$var(Y) = \frac{1}{12}(1-2.6)^2 + \frac{7}{12}(2-2.6)^2 + \ldots$$

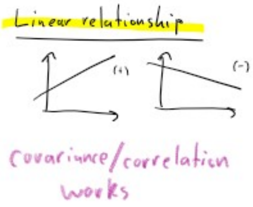
2.4.5 Covariance

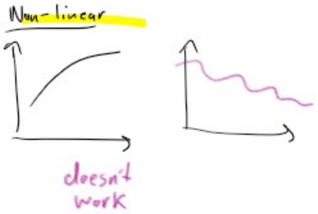
- Measures the relationship between two random variables
- Random variables Y and X have a *joint* probability function
- Joint prob. func.: (i) lists all possible combos of Y and X; (ii) assign a probability to each combination
- A useful summary of a joint probability function is the *covariance*
- The covariance between Y and X is the expected difference of Y from its mean, multiplied by the expected difference of X from its mean
- Covariance tells us something about how two variables are *related*, or how they *move together*
- Tells us about the direction and strength of the relationship between two variables

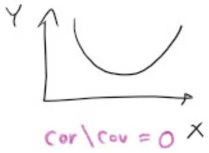
$$\boxed{\mathrm{Cov}(Y, X) = \mathrm{E}[(Y - \mu_Y)(X - \mu_X)]} \tag{2.8}$$

The covariance between Y and X is often denoted as σ_{YX}. Note the following properties of σ_{YX}:

- σ_{YX} is a measure of the *linear* relationship between Y and X. Non-linear relationships will be discussed later.
- $\sigma_{YX} = 0$ means that Y and X are linearly independent.
- If Y and X are independent (neither variable causes the other), then $\sigma_{YX} = 0$. The converse is not necessarily true (because of non-linear relationships).
- The $\mathrm{Cov}(Y, Y)$ is the $\mathrm{Var}(Y)$. $Cov(Y,Y) = E[(Y-\mu_Y)(Y-\mu_Y)] = E[(Y-\mu_Y)^2] = var(Y)$
- A positive covariance means that the two variables tend to differ from their mean in the *same* direction.
- A negative covariance means that the two variables tend to differ from their mean in the *opposite* direction.







2.4.6 Correlation

- Correlation usually denoted by ρ "rho"
- Similar to covariance, but is easier to interpret

$$\rho_{YX} = \frac{\mathrm{Cov}(Y, X)}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(X)}} = \frac{\sigma_{YX}}{\sigma_Y \sigma_X} \tag{2.9}$$

The difficulty in interpreting the value of covariance is because $-\infty < \sigma_{YX} < \infty$. Correlation transforms covariance so that it is bound between -1 and 1. That is, $-1 \leq \rho_{YX} \leq 1$.

- $\rho_{YX} = 1$ means perfect positive linear association between Y and X.
- $\rho_{YX} = -1$ means perfect negative linear association between Y and X.
- $\rho_{YX} = 0$ means no linear association between Y and X (linear independence).

2.4.7 Conditional distribution

- Joint distribution 2 RVs
- Conditional distribution – fix (condition on) one of those RVs
- Condition expectation "mean" – the mean of one RV after the other RV has been "fixed"

Let Y be a discrete random variable. Then, the conditional mean of Y given some value for X is

$$\mathrm{E}(Y|X = x) = \sum_{i=1}^{K} (p_i|X = x) Y_i \tag{2.10}$$

- If the two RVs are independent, the conditional distribution is the same as the *marginal* distribution

Example: Blizzard and cancelled midterm

Suppose that you have a midterm tomorrow, but there is a possibility of a blizzard. You are wondering if the midterm might be cancelled.

Table 2.1: Joint distribution for snow and a canceled midterm

	Midterm ($Y = 1$)	No Midterm ($Y = 0$)
Blizzard ($X = 1$)	0.05	0.20
~~No Blizzard ($X = 0$)~~	~~0.72~~	~~0.03~~
	.77	.23

add to 1

in morning you see a blizzard

- ~~What are the *marginal* probability distributions?~~
- What is $E[Y]$? What is $E[Y\ X - 1]$?
- What is the covariance and correlation between X and Y?
- More exercises in the "Review Questions"

$E[Y] = (.77)1 + (.23)0 = 0.77$

$(.2)1 + (.8)0 = .2$

$\frac{.05}{.25}$ $\frac{.20}{.25}$

sum to 1

2.5 Some special probability functions

↳ (i) lists all possibilities
(ii) prob. assigned to possibilities

2.5.1 The normal distribution

- Common because of the "central limit theorem" (in a few slides)

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\cancel{\sigma^2}\to 1}} \exp -\frac{(y-\overset{0}{\cancel{\mu}})^2}{2\cancel{\sigma^2}\to 1} \qquad (2.3)$$

- Mean of y is μ
- Variance of y is σ^2

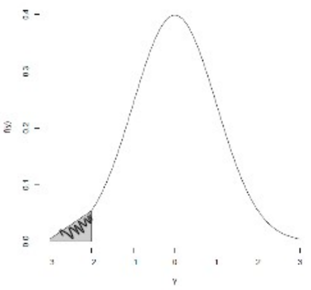
2.5.2 The standard normal distribution

- Special case of a normal distribution, where $\mu = 0$ and $\sigma^2 = 1$
- Equation 2.3 becomes:

$$\boxed{f(y) = \frac{1}{\sqrt{2\pi}} \exp \frac{-y^2}{2}} \qquad (2.11)$$

- Any normal random variable can be "standardized"
- How to standardize? subtract μ, divide by σ
- Standardizing has long been used in hypothesis testing (as we shall see)

Figure 2.3: Probability function for a standard normal variable, $p_{y<-2}$ in gray



2.5.3 The central limit theorem

- There are hundreds of different probability functions
- Examples: Poisson, Binomial, Generalized Pareto, Nakagami, Uniform
- So why is the normal distribution so important? Why are so many RVs normal?
- Answer: CLT
- CLT (loosely speaking) – if we add up enough RVs, the resulting sum tends to be normal

Exercise: draw the probability function for one die roll, then for the sum of two dice.

Figure 2.1: Probability function for the result of a die roll

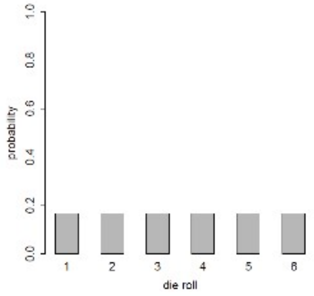


Figure 2.4: Probability function for the sum of two dice

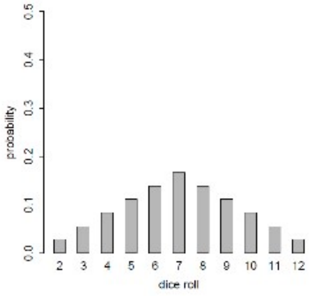


Figure 2.5: Probability function for three dice, and normal distribution

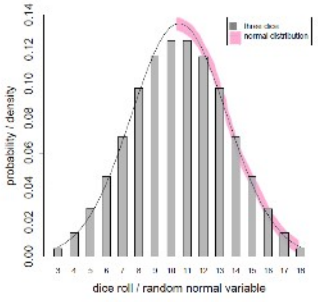
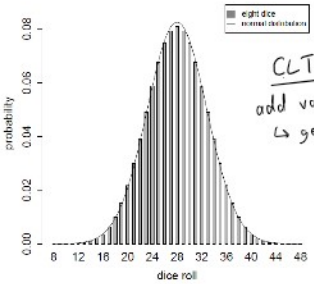


Figure 2.6: Probability function for eight dice, and normal distribution



CLT

add variable
↳ get N

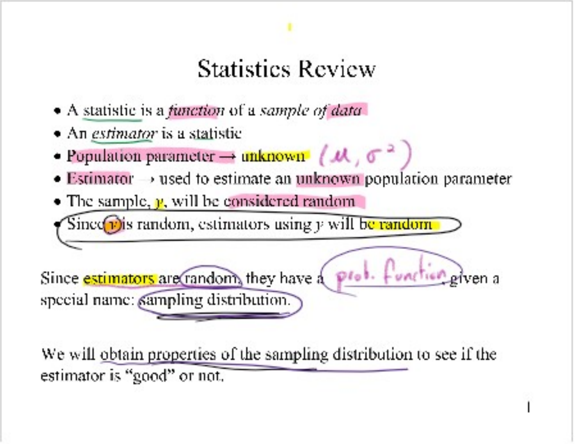
2.5.4 The chi-square distribution

- Add to a normal RV still normal
- Multiply a normal RV – still normal
- Square a normal RV – now it is chi-square distributed
- We will use the chi-square distribution for the F-test in a later chapter

Chapter 3

Monday, January 13, 2025 10:00 AM

slides 3



Statistics Review

- A statistic is a *function* of a *sample of data*
- An *estimator* is a statistic
- Population parameter → unknown (μ, σ^2)
- Estimator → used to estimate an unknown population parameter
- The sample, y, will be considered random
- Since y is random, estimators using y will be random

Since estimators are random, they have a prob. function, given a special name: sampling distribution.

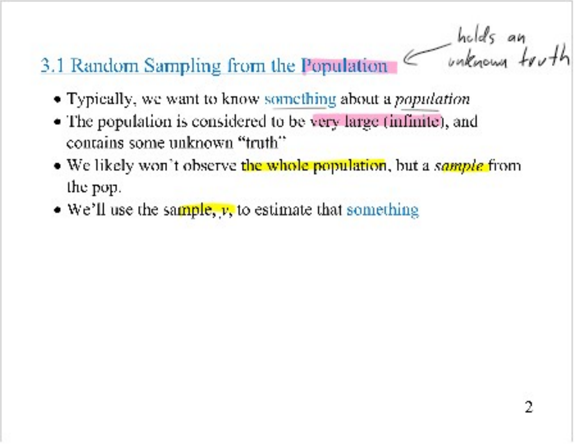
We will obtain properties of the sampling distribution to see if the estimator is "good" or not.

→ is random!

y is a sample of values

↳ like in Assign 1 die rolls

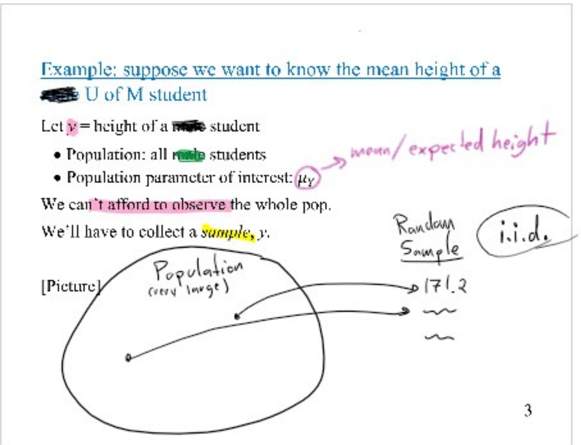
↳ anything I calculate using y is also random!



3.1 Random Sampling from the Population

← holds an unknown truth

- Typically, we want to know something about a *population*
- The population is considered to be very large (infinite), and contains some unknown "truth"
- We likely won't observe the whole population, but a *sample* from the pop.
- We'll use the sample, y, to estimate that something



Example: suppose we want to know the mean height of a [redacted] U of M student

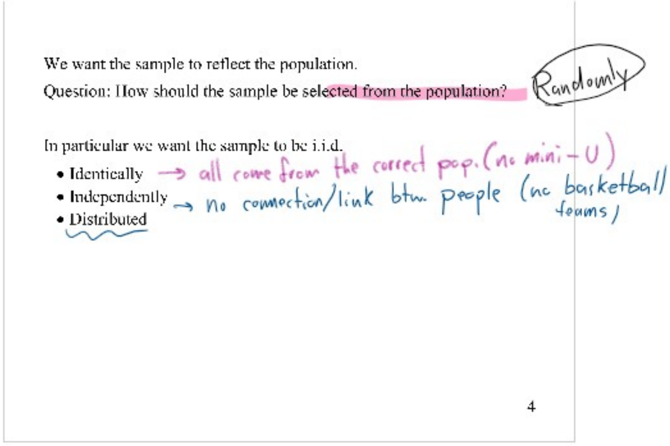
Let y = height of a [redacted] student

- Population: all [redacted] students
- Population parameter of interest: μ_Y → mean/expected height

We can't afford to observe the whole pop.

We'll have to collect a *sample*, y.

[Picture]



We want the sample to reflect the population.

Question: How should the sample be selected from the population? Randomly

In particular we want the sample to be i.i.d.

- Identically → all come from the correct pop. (no mini-U)
- Independently → no connection/link btw. people (no basketball teams)
- Distributed

---

anything that follows is also random!

So, the sample y is random!!

- Could have gotten a different y
- Parallel universe

Table 3.1: Entire population of heights (in cm). The true (unobservable) population mean and variance are $\mu_y = 176.8$ and $\sigma_y^2 = 39.7$.

177.3	170.2	187.2	178.3	170.3	179.4	181.2	180.0	173.9
178.7	171.7	160.5	183.9	175.7	175.9	182.6	181.7	180.2
181.5	176.5	162.1	180.3	175.6	174.9	165.7	172.7	178.9
175.3	178.7	175.6	166.4	173.1	173.2	175.6	183.7	181.3
174.2	180.9	179.9	171.2	171.0	178.6	181.4	175.2	182.2
171.7	178.4	168.1	186.0	189.9	173.4	168.7	180.0	175.1
175.7	180.8	176.2	170.8	177.3	163.4	186.3	177.1	191.2
171.0	180.3	169.5	167.2	178.0	172.9	176.0	176.5	171.9
175.1	184.2	165.3	180.2	178.3	183.4	173.9	178.6	177.9
184.5	184.1	180.9	187.1	179.9	167.1	172.0	167.4	172.7
171.6	186.6	182.4	185.5	174.8	178.8	192.8	179.3	172.0

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

How could i.i.d. be violated in the heights example? basketball mini-U

Example: mean income of people in Canada ~~Canadians~~. How could i.i.d. be violated? using phone not everyone has one

How should we estimate the mean height? sample mean sample average average

We want μ_y. Use $\bar{y}$ to estimate μ_y.

3.2 Estimators and Sampling Distributions

An estimator uses the sample y to "guess" something about the pop.

We collect our sample, $y = \{173.9, 171.7, 182.6, 181.5, 162.1, 174.9, 165.7, 182.2, 171.7, 168.1, 189.9, 175.7, 163.4, 186.3, 169.5, 171.9, 173.9, 172.0, 172.7, 172.0\}$. How should we use this sample to *estimate* the mean height?

3.2.1 Sample mean

A popular choice for estimating a population mean is by using a *sample mean* (or *sample average* or just *average*)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.1)$$

in reality don't know this

sample variance to estimate σ_Y^2

- From heights example: $\bar{y} = 174.1$, $\mu_y = 176.8$
- There are many ways to estimate μ_y. Examples? mode/median/geometric average/harmonic average/ $\frac{\min(y)+\max(y)}{2}$
- Why is (3.1) so popular?
- How good is $\bar{y}$ at estimating μ_y in general?
- To answer these questions: idea of a *sampling distribution*

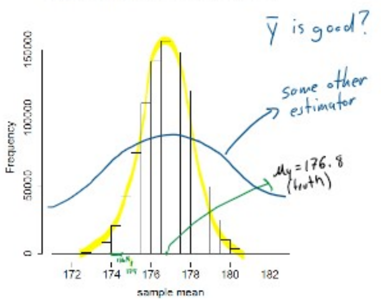
Recall that the sample, y, is random. Each element of y was selected randomly from the population. We could have selected a different sample of size $n = 20$. For example, in a parallel universe, we could have gotten y^* $= \{175.9, 175.3, 182.2, 178.6, 175.2, 180.3, 178.3, 183.7, 176.0, 167.4, 178.7, 178.7, 186.0, 175.6, 180.0, 168.7, 178.6, 173.1, 173.2, 187.1\}$, where the $*$ in y^* denotes that we are in the parallel universe. In this parallel universe, we got $\bar{y}^* = 177.6$. But in every universe, the population (table 3.1), is the same. $\bar{y} = 174.1$

- Randomly sample from the population → get y
  - ○ y is random
- Use y to calculate $\bar{y}$
  - ○ $\bar{y}$ is random
  - ○ could have gotten a different sample → could have gotten a different $\bar{y}$
  - ○ population is always the same (μ_y)

3.2.2 Sampling distribution of the sample mean

- $\bar{y}$ is random variable (it's an estimator, all estimators are random)
- random variables usually have probability functions
- $\bar{y}$ has a *sampling distribution* (probability function for an estimator)
- *sampling distribution* – imagine all possible values for $\bar{y}$ that you could get – plot a histogram
- Using a computer, I drew 1 mil. different random samples of $n = 20$ from table 3.1. Calculate $\bar{y}$ each time. Plot histogram:

Figure 3.1: Histogram for 1 million $\bar{y}$s



$\bar{y} = \frac{\Sigma y_i}{n}$ ← CLT

Normal

Which probability function is right for $\bar{y}$? Why?

- Look at figure 3.1
- Notice the summation operator in equation 3.1
- Answer: Normal Reason: CLT

$\bar{y}$ is random. We'll derive its:

- mean
- variance

Use these to determine if it's a "good" estimator via three statistical properties:

- Bias
- Efficiency
- Consistency

3.2.3 Bias

$\bar{y}$, ~~mode~~, median, max, etc.

An estimator is unbiased if its expected value ("mean") is equal to the population parameter it's estimating.

That is, $\bar{y}$ is unbiased if $E[\bar{y}] = \mu_y$.

Unbiased if it gives "the right answer on average".

Biased if it gives the wrong answer on average.

$$\mathrm{E}[y] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right]$$

$\bar{y} = \frac{1}{n}\Sigma y_i$

Rules of the mean

(i) $E[cY] = cE[Y]$

(2) $E[X+Y] = E[X] + E[Y]$

$$
\begin{aligned}
\mathrm{E}[\bar{y}] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}[y_1 + y_2 + \cdots + y_n] \\
&= \frac{1}{n}(\mathrm{E}[y_1] + \mathrm{E}[y_2] + \cdots + \mathrm{E}[y_n]) \\
&= \frac{1}{n}(\mu_y + \mu_y + \cdots + \mu_y) \\
&= \frac{n\mu_y}{n} = \mu_y
\end{aligned}
\qquad (3.2)
$$

$\bar{y} = \frac{1}{n}\sum y_i$

$E[\bar{y}] = \mu_y$ if unbiased

Rules of the mean

(i) $E[cY] = cE[Y]$

(ii) $E[X+Y] = E[X] + E[Y]$

$E\left[\frac{1}{n}\sum y_i\right] = \frac{1}{n}E\left[\sum y_i\right] = \frac{1}{n}E[y_1 + y_2 + \ldots + y_n]$

$= \frac{1}{n}\{E[y_1] + E[y_2] + \ldots + E[y_n]\}$ (iii) sample is i.i.d. → "identical"

$= \frac{1}{n}\{\mu_y + \mu_y + \ldots + \mu_y\} = \frac{1}{n}n\mu_y = \mu_y$

3.2.4 Efficiency

An estimator is efficient if it has the smallest variance among all other potential estimators (for us, potential = linear, unbiased)

Need to get the variance of $\bar{y}$.

$Var(\bar{y}) = var\left(\frac{1}{n}\sum y_i\right)$

$= \frac{1}{n^2}var\left(\sum y_i\right) = \frac{1}{n^2}var(y_1 + y_2 + \ldots y_n)$

$= \frac{1}{n^2}\{var(y_1) + var(y_2) + \ldots + var(y_n)\}$ i.i.d. ↳ independent ↳ if we random sample

$= \frac{1}{n^2}\{\sigma_Y^2 + \sigma_Y^2 + \ldots + \sigma_Y^2\} = \frac{n\sigma_Y^2}{n^2}$

$= \frac{\sigma_Y^2}{n}$

Rules of var

$var(cY) = c^2 var(Y)$

$Var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$

if 2 variables are independent: then: 0 cov, 0 corr.

$$
\begin{aligned}
\mathrm{Var}[\bar{y}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}[y_1 + y_2 + \cdots + y_n] \\
&= \frac{1}{n^2}(\mathrm{Var}[y_1] + \mathrm{Var}[y_2] + \cdots + \mathrm{Var}[y_n]) \\
&= \frac{1}{n}\left(\sigma_y^2 + \sigma_y^2 + \cdots + \sigma_y^2\right) \\
&= \frac{n\sigma_y^2}{n^2} = \frac{\sigma_y^2}{n}
\end{aligned}
\qquad (3.3)
$$

$var(\bar{y}) < var(\hat{u})$?

- Gauss-Markov theorem proves this is minimum variance
- We'll also need this to prove consistency, and for hyp. testing

→ super important → explains why $\bar{y}$ is so popular

$\bar{y} = \frac{1}{n}\sum y_i$

also random

random

properties? expect / variance → $E[\bar{y}] = \mu_y$ unbiased

$var[\bar{y}] = \frac{\sigma_y^2}{n}$ → efficient (min. variance) → consistent

3.2.5 Consistency

Suppose we had a lot of information. ($n \rightarrow \infty$)

What value should we get for our estimator? → truth w/ prob. 1

How would state this mathematically?

$\lim_{n\to\infty} var(\bar{y}) \rightarrow 0$ and $\lim_{n\to\infty} E[\bar{y}] \rightarrow \mu_Y$

Q) Prove that the sample mean is a consistent estimator for the population mean.

✓

Q) Define the terms unbiasedness, efficiency, and consistency.

reject / fail to reject

very unrealistic assumption

3.3 Hypothesis tests (known σ_y^2)

"null"

$H_0 : \mu_y = \mu_{y,0}$ ← a # we pick

almost all tests in Econ

(3.4)

"null"

3.3 Hypothesis tests (known σ_y^2)

almost all tests in Econ

$$H_0 : \mu_y = \mu_{y,0} \quad \text{a \# we pick}$$
$$H_A : \mu_y \neq \mu_{y,0} \text{ (2-sided alternative)} \qquad (3.4)$$

- Estimate μ_y (using $\bar{y}$ for example)
- See if $\bar{y}$ appears "close" to $\mu_{y,0}$
  - Remember, $\bar{y}$ is random! (and Normal)
- If it's close → fail to reject
- If it's far → reject

In assign #1: $H_0 : \mu_y = 3.5$

Example:

- Hypothesize that mean height of a U of M student is 173cm

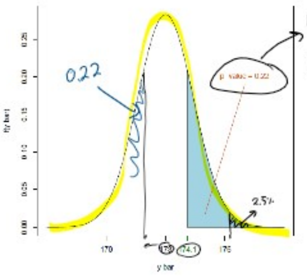
$$H_0 : \mu_y = 173$$
$$H_A : \mu_y \neq 173 \qquad (3.5)$$

- Collect a sample: y = {173.9, 171.7, ..., 172.0}
- Calculate $\bar{y} = 174.1$
- Suppose (very unrealistically that we know that $\sigma_y^2 = 39.7$
- What now?

$$\bar{y} - \mu_{y,0} = 174.1 - 173 = 1.1$$

$var(\bar{y}) = \frac{\sigma_y^2}{n}$

Figure 3.2: Normal distribution with $\mu = 173$ and $\sigma^2 = 39.7/20$. Shaded area is the probability that the normal variable is greater than 174.1.



prob. of getting higher than 1.1 away from the hyp.

p-value = 0.44

44%

↳ big prob.

↳ fail to reject

Significance level

Pre-determined p-value that decided if you reject/fail to reject

10% / 5% / 1%

The p-value for the above test is 0.44. How to interpret this?

↳ prob. of getting a $\bar{y}$ that is more adverse to H_0, compared to what we just observed.

3.3.1 Significance of a test

$\alpha = 10\% / 5\% / 1\%$ $\quad Pr(H_0 \text{ is true}) = 0 \text{ or } 1$

3.3.2 Type I error

$$pr(\text{reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

3.3.3 Type II error (and power)

$$pr(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = \beta = ?$$

$$\text{power} = pr(1 - \text{type II}) = pr(\text{reject } H_0 \mid H_0 \text{ is false}) = ?$$

How big is $\underset{\text{me.}}{\mu_{y,0}} - \underset{\text{truth?}}{\mu_y}$

3.3.4 Test statistics

↗ Z-test statistic
t-test stat.

- Just a more convenient way of getting the p-value for the test
- Each hypothesis test would present us with a new normal curve

$H_0 : \mu_y = 1000$

$\bar{y} = 1022.3$

3.3.4 Test statistics

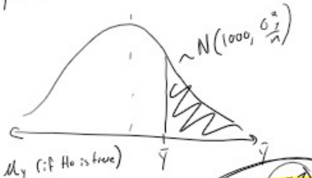
z-test statis…
t-test stat.

- Just a more convenient way of getting the p-value for the test
- Each hypothesis test would present us with a new normal curve that we would have to draw, and calculate a new area (see fig. 3.2)
- Instead: *standardize*
- This gives us *one curve for all testing problems* (the standard normal curve)
- Calculate a bunch of areas under the curve, and tabulate them
- Not an issue with modern computers, but this is still the way we do things
- How to get a z test statistic?
- Do a z test for our heights example.

$$Z = \frac{\text{estimate} - \text{hyp.}}{\sqrt{\text{var(estimator)}}} = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\frac{\sigma_y^2}{n}}} = \frac{\overset{\bar{y}}{\boxed{174.1}} - \overset{\text{true } \mu_y \text{ (if Ho is true)}}{\boxed{173}}}{\sqrt{\underset{\text{var}(\bar{y})}{\boxed{39.7/20}}}} = \boxed{0.78}$$

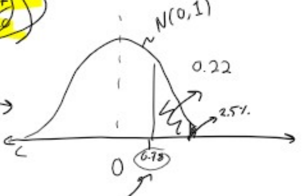
$H_0: \mu_y = 1000$

$\bar{y} = 1022.3$



$\bar{y} \sim N\left(173, \frac{39.7}{20}\right)$

$Z \sim N(0,1)$



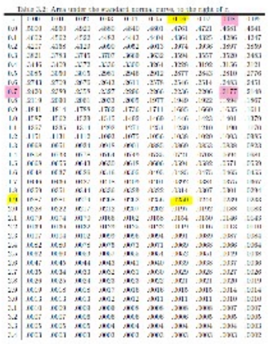


Table 3.2: Area under the standard normal curve, to the right of z.

3.3.5 Critical values

→ max z-stat before you reject (t-stat)
→ 1.96 for 5% significance → if $|z| > 1.96$ ↳ we reject Ho @ 5% level

In R, to do a ttest

3.3.6 Confidence intervals

What is the probability that our z statistic will be within a certain interval, if the null hypothesis is true? For example, what is the following probability?

$$\Pr(-1.96 \le z \le 1.96)? = 95\% \tag{3.12}$$

$$\Pr\left(-1.96 \le \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma^2_{\bar{y}}/n}} \le 1.96\right) = 0.95 \tag{3.13}$$

Finally, we solve equation 3.13 so that the null hypothesis $\mu_{y,0}$ is in the middle of the probability statement:

$$\Pr\left(\bar{y} - 1.96 \times \sqrt{\frac{\sigma_y^2}{n}} \le \mu_{y,0} \le \bar{y} + 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}\right) = 0.95 \tag{3.14}$$

$\bar{y} \pm 1.96 \times \text{s.e.}(\bar{y})$
↓
95%

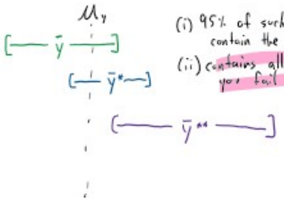
- unbiased
- efficient
- consistent

} properties that are desirable
$\bar{y}$ is an example

3 ways to decide on Ho

(i) compare p-value to α
(ii) compare z-stat to crit. value (1.96)
(iii) see if Ho is inside a confidence interval
↳ [— ȳ —]
↑ if Ho is in interval
↳ fail to reject

Interpretation



(i) 95% of such intervals contain the truth
(ii) contains all null hypotheses you fail to reject

Ch. 3

$H_0: \mu_y = 173$
$H_A: \mu_y \neq 173$

Assign #1

$H_0: \mu_y = 3.5$ ← $\mu_{y,0}$
$H_A: \mu_y \neq 3.5$

$$Z = \frac{\bar{y} - \mu_{y,0}}{\underbrace{\sqrt{\text{var}(\bar{y})}}_{\text{s.e.}(\bar{y}) \text{ “standard error”}}}$$

$\text{var}(\bar{y}) = \frac{\sigma_y^2}{n}$ → sample size

3.4 Hypothesis Tests (unknown σ_y^2)

- Much more realistically, σ_y^2 (variance of y) will be unknown.
- Recall that: $Var[\bar{y}] = \sigma_y^2/_n$
- $z = \frac{\bar{y} - \mu_{y,0}}{s.e.(\bar{y})} = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/_n}}$
- So, we need to estimate σ_y^2 in order to perform hypothesis tests.

$\bar{y} = \frac{1}{n}\sum y_i$

$\text{mean}(Y) = E(Y)$

$\text{var}(Y) = E[(Y - \mu_Y)^2]$
↑ $\frac{1}{n}\sum(y_i - \bar{y})^2$

3.4.1 Estimating σ_y^2

- A “natural” estimator:

“hat”

$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$ BIASED

$$E\left[\underbrace{\frac{n}{n-1}\frac{1}{n}\sum(y_i - \bar{y})^2}_{\frac{\cancel{n}}{n-1}\frac{1}{\cancel{n}}\sum(y_i - \bar{y})^2}\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2$$

3.4.1 Estimating σ_y^2

- A "natural" estimator:

"hat" →

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \quad (3.15)$$

BIASED

↑ $\frac{1}{n}\sum(y_i - \bar{y})^2$

$E\left[\frac{n}{n-1}\frac{1}{n}\sum(y_i-\bar{y})^2\right]$

$\frac{\cancel{n}}{n-1}\frac{1}{\cancel{n}}\sum(y_i-\bar{y})^2$

↓

$\frac{1}{n-1}\sum(y_i-\bar{y})^2 = s^2$

- Is this a good estimator? → NO Why or why not? → it's biased
- A better estimator:

→ unbiased

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.17)$$

- Degrees-of-freedom correction → after $\bar{y}$, there are n-1 d.o.f.
  ↳ lose 1 piece of information

$y = \{1, 3, ?\}$ $\quad \bar{y} = 3$

↓

5

So:

$$\text{Estimated variance of } \bar{y} = \frac{s_y^2}{n}$$

We can implement hypothesis testing by replacing the unknown σ_y^2 with its estimator s_y^2. The z test statistic now becomes:

$$\frac{\bar{y} - \mu_{y,0}}{\sqrt{s_y^2/n}} = t$$

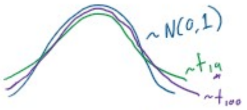
replace σ^2 w/ s^2
↳ z becomes t

$Z \sim N(0,1)$

$t \sim t_{n-1}$

↓ determines "shape" of curve

Note: for large n, the t test is equivalent to the z test



↓

t converges to N(0,1)