ECO220Y1Y: Course Overview, Sampling, Data, and Describing Categorical Data

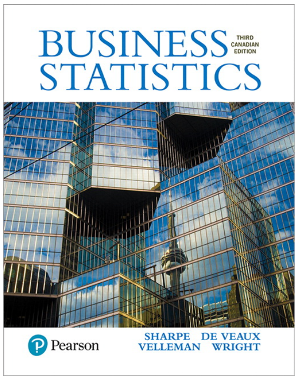
Lecture 1

Reading: Chapters 1 – 4

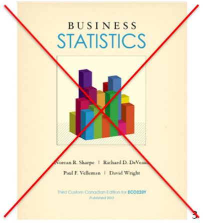
Includes **ALL** margin notes and boxes: “For Example,” “Guided Example,” “Notation Alert,” “Just Checking,” “Optional Math Boxes”, “What Can Go Wrong?” and “Ethics in Action”

Intro to Data Analysis & Applied Econometrics: Some syllabus highlights

- Quercus course site (slides up before class)
- Sections L0201, L0301, L0401
  - What happens on Fridays from 9 – 11am?
    - Usually, 65-minute interactive DACM TA tutorials
    - Five 110-minute term tests
    - Some 50-minute TA review tutorials (e.g. Friday, Sept 6)
  - Use Piazza, not e-mail, for your questions
    - Sign up: https://piazza.com/utoronto.ca/fall2019/eco220y1y
  - iClicker: Register your remote via Quercus
  - Quercus quizzes



The 2017 edition is most convenient for you.


BUSINESS
STATISTICS
Norean R. Sharpe | Richard D. DeVeaux
Paul F. Velleman | David Wright

Required Supplements to Textbook

- *Readings page* on Quercus gives you easy access
  - *The DACM Handbook for ECO220Y1Y*
    - Accept your invitation to the DACM Quercus site
    - Your first DACM tutorial is this Friday, Sept. 13th
    - Read pages 1 to 20 of the handbook and take the necessary actions *before* arriving at your tutorial section
  - *Quiz and Prerequisite Review for ECO220Y1Y*
    - Study for Quercus Quiz 1 and Test #1 (and beyond)
  - Aid sheets (formulas and statistical tables)
  - *Logarithms in Regression Analysis with Asiaphoria*
  - *The Normal Table: Read It, Use It*

U of T 2017 Alumni Impact Survey

- To kick off key material (today!), a real survey
  - *Background & Methodology:* "The total living U of T alumni population is about 545,000. The overall response rate was approximately 8% or just over 21,000 respondents."

SURVEY RESPONDENTS COMPARED TO ALUMNI POPULATION

SEGMENT	# COMPLETES	% COMPLETES	% OF ALUMNI POPULATION
Male	8,862	45.8	47.1
Female	10,472	54.1	52.8

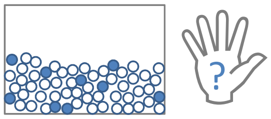
Source: University of Toronto website, Alumni page, Alumni Impact Survey, retrieved from https://alumni.utoronto.ca/alumni-impact-survey on September 7, 2018.

Key Terms

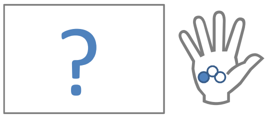
- Population: set of all items of interest
- Sample: subset of the population
- Parameter: number describing a population
- Statistic: number describing a sample

- Descriptive statistics: describe a sample (data)
- Inferential statistics: make inference about a population and its parameters using data

Probability versus Statistics



Probability asks: If I shake up the box (i.e. population) and randomly select 3 balls, what is in my hand (i.e. sample)?



Descriptive statistics says: I have 2 white and 1 solid balls

Inferential statistics asks: Given what is in my hand (i.e. sample) what is in the box (i.e. population)?

Parameters: Usually Greek Letters

α	alpha	ι	iota	ρ	rho
β	beta	κ	kappa	σ	sigma
χ	chi	λ	lamda	τ	tao
δ	delta	μ	mu	θ	theta
ε	epsilon	ν	nu	υ	upsilon
η	eta	o	omikron	ϖ	omega
ϕ	phi	π	pi	ξ	xi
γ	gamma	ψ	psi	ζ	zeta

http://platonicrealms.com/encyclopedia/Greek-alphabet

Sampling Error

- Sampling Error: The purely random difference between a sample and the population that arises because the sample is a random subset of the population
  - Does “error” mean that a mistake was made?
  - Also called: “sampling noise,” “white noise,” “sampling variability” (textbook)
  - As sample size gets larger the sampling error tends to get smaller

Do “Just Checking” as you read

JUST CHECKING (p. 32)

1. Why is each of the following claims not correct?
   a) It is always …
   b) Stopping customers …
   c) We drew a sample of 100 from the 3,000 students in a school. To get the same level of precision for a school of 30,000 students, we’ll need a sample of 1,000.
   d) A poll taken …
   e) The true percentage of all people who enjoy statistics is called a “population statistic.”

Answers are found in Appendix A.

For c), see “The Sample Size is What Matters” on pp. 30 – 31. We do *large population* statistics in ECO220Y. When Gallup surveys people about global warming, it uses a sample size of around 2,000 each in China and Iceland.

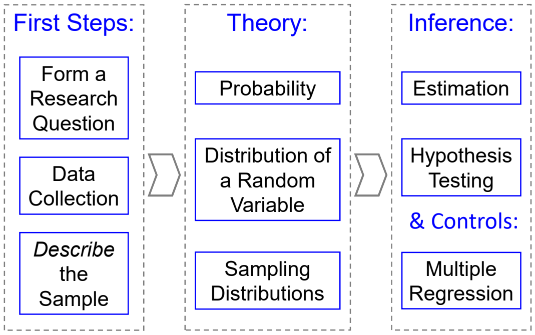
Sampling versus Non-Sampling Errors

- Non-sampling errors: Systematic (*not* random) differences between sample and population
  - Biased estimate: Sample statistic is *systematically* higher or lower than the population parameter
  - Systematic lying (e.g. overstate income); Poor survey design (e.g. misinterpreted question)
  - Low response rate where non-responders differ from responders: non-response bias
  - Sampling frame differs from target population

Note: All real data have some non-sampling errors.

What is the point of the table for the U of T alumni?

ECO220Y: Overview



Example from today's readings

- Divide the labor force into two groups by education:
  - Those with a high school degree or less
  - Those with more than a high school degree

If from 1968 to 2018 the unemployment rate for each group *rose*, is it possible the overall unemployment rate *dropped*?

	1968			**2018**		
Highest Ed.	# UnE	# in LF	Rate	# UnE	# in LF	Rate
HS or less	2,000	50,000	0.04	1,500	30,000	0.05
More than HS	200	10,000	0.02	2,700	90,000	0.03

See Section 4.5 "Simpson's Paradox," which also called composition effects.

According to Merriam-Webster, a paradox is "a statement that is seemingly contradictory or opposed to common sense and yet is perhaps true."

Data: Variables & Observations

resp_id	**seconds**	**age**	**male**	**highest_ed**	**hh_inc**
R_06C1	118	34	0	4-year Degree (BA/BS)	\$50,000 - \$74,999
R_06RY	23	25	1	4-year Degree (BA/BS)	\$25,000 - \$49,999
R_06wD	225	52	0	Master's Degree (MA/MS)	\$50,000 - \$74,999
R_08Ps	111	36	1	High School/GED	\$100,000 - \$149,999
…					
R_zZmw	26	21	1	Some College	Under \$25,000

There are six variables: **resp_id**, is an *identifier variable*; **seconds** and **age** are *quantitative variables*; and **male**, **highest_ed** and **hh_inc** are *qualitative (categorical/nominal) variables*. (Further, **male** is a *dummy variable*.)

There are 1,603 observations in these data, one per respondent, which is the *unit of observation*, $n = 1{,}603$. Only five observations are shown above.

This in an excerpt of real data you will see in Module A.1 of DACM.

2 Types of Information; 2 Types of Variables

- Quantitative info.: tells a numeric measure
  - e.g. Elasticity of demand for green apples is -2.2
- Interval (quantitative) variables: contain numerical measures
  - e.g. The debt-to-GDP ratio for each of 30 countries in 2014

- Qualitative info.: gives an assessment of kind
  - e.g. If the price of green apples rises, Q_D declines
- Nominal (categorical) variables: record which category
  - e.g. Is each country in the OECD?
  - Dummy variable: =1 if in a category, =0 otherwise

Three Types of Data Sets

- Cross-sectional: Same variable(s) in same time period measured for different units
  - E.g. Annual GDP in 2010 for 20 different countries
    - Unit of observation: a country
- Time series: Same variable(s) for same unit measured at different time periods
  - E.g. Annual Canadian GDP in 2000,…,2010
    - Unit of observation: a year
- Panel (Longitudinal): Same variable(s) measured for a range of units & time periods
  - E.g. Annual GDP for 20 different countries in 2000,…,2010
    - Unit of observation: a country-year pair

Also, see DACM Handbook, including page 15.

Panel Data in “Wide” Format

countryname	code	oecd	pctdbt95	pctdbt00	pctdbt05	pctdbt10
Albania	ALB	0	35.29	.	.	.
Australia	AUS	1	.	29.55	22.58	29.32
Austria	AUT	1	61.80	65.59	67.50	74.04
…						
Canada	CAN	1	79.37	60.07	47.15	52.68
…						
Italy	ITA	1	121.12	118.67	112.80	117.59
…						
United Kingdom	GBR	1	50.92	45.45	46.14	86.65
United States	USA	1	.	.	47.34	76.84
Uruguay	URY	0	.	.	76.11	45.20
Zimbabwe	ZWE	0	77.09	.	.	.

http://data.worldbank.org/indicator/GC.DOD.TOTL.GD.ZS

Example of Cross-Sectional Data

countryname	code	oecd	pctdbt10
Albania	ALB	0	.
Australia	AUS	1	29.32
Austria	AUT	1	74.04
…			
Canada	CAN	1	52.68
…			
Italy	ITA	1	117.59
…			
United Kingdom	GBR	1	86.65
United States	USA	1	76.84
Uruguay	URY	0	45.20
Zimbabwe	ZWE	0	.

Example of Time-Series Data

countryname	code	oecd	year	pctdbt
Italy	ITA	1	1995	121.12
Italy	ITA	1	2000	118.67
Italy	ITA	1	2005	112.80
Italy	ITA	1	2010	117.59

Panel Data in “Long” Format

countryname	code	oecd	year	pctdbt
Albania	ALB	0	1995	35.29
Albania	ALB	0	2000	.
Albania	ALB	0	2005	.
Albania	ALB	0	2010	.
Australia	AUS	1	1995	.
Australia	AUS	1	2000	29.55
Australia	AUS	1	2005	22.58
Australia	AUS	1	2010	29.32
…				
Zimbabwe	ZWE	0	1995	77.09
Zimbabwe	ZWE	0	2000	.
Zimbabwe	ZWE	0	2005	.
Zimbabwe	ZWE	0	2010	.

More Examples

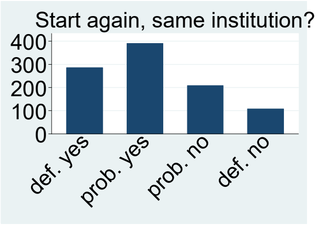
Consider data on 257 people who tasted a new snack product at Loblaws. Each was asked: how likely is it that you will purchase this product in the future? (definitely, probably, not sure, probably not, definitely not). Which kind of data are these?

Consider data on ECO220Y1Y recording the price of the textbook each year for the past decade and the percent of students that had a copy. Which kind of data are these?

Describing one variable
(with not too many unique values)

- Tabulation: Lists all unique values in data & relative frequency
  - Interval or nominal data
  - Basis of bar/pie chart
  - aka: frequency table, relative frequency table

```
    x |  Freq.   Percent     Cum.
------+---------------------------
    1 |    286     28.83    28.83
    2 |    390     39.31    68.15
    3 |    208     20.97    89.11
    4 |    108     10.89   100.00
------+---------------------------
Total |    992    100.00
```



NSSE 2006, U of T seniors: *If you could start again, would you go to the same institution you are now attending?*

1= definitely yes
2= probably yes
3= probably no
4= definitely no

http://www.utoronto.ca/about-uoft/measuring-our-performance.htm

Describing Two Variables
(with not too many unique pairs)

- Cross tabulation: Measures frequency that two variables take each possible pair of values
  - Shows *relationship* between two variables
  - Interval or nominal data
  - Basis of bar/pie chart
  - aka: contingency table or two-way table

Example: Advertising

- Goal: compare ads
  - “Power”: Emphasizes horsepower
  - “Safety”: Emphasizes safety features
  - Participants watch TV alone for 2 hours
  - Insert one ad at random with other ads
  - Give “quiz” about car

id	sex	ad	recall
1	Female	Safety	Yes
2	Female	Power	No
3	Male	Safety	No
4	Female	Safety	No
…	…	…	…
251	Female	Power	No
252	Female	Safety	Yes
253	Male	Safety	No
254	Male	Safety	No

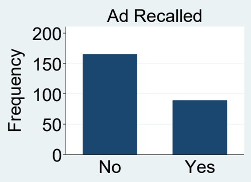
Summary: Experimental Design

```
         ad |      Freq.     Percent        Cum.
------------+-----------------------------------
      Power |        129       50.79       50.79
     Safety |        125       49.21      100.00
------------+-----------------------------------
      Total |        254      100.00

        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
     Female |        131       51.57       51.57
       Male |        123       48.43      100.00
------------+-----------------------------------
      Total |        254      100.00
```

Are these tabulations or cross-tabulations?

Summary: Results (Incomplete)



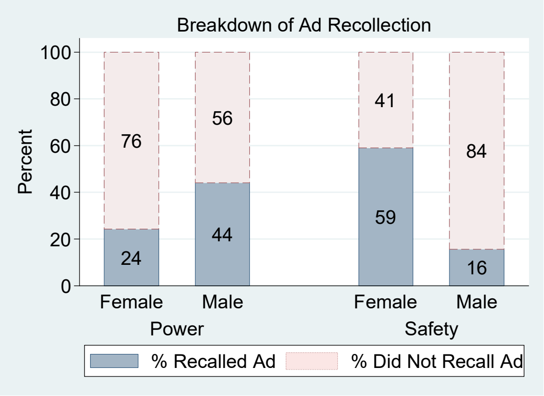
```
     recall |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |        165       64.96       64.96
        Yes |         89       35.04      100.00
------------+-----------------------------------
      Total |        254      100.00
```

Two Cross-tabs

```
-> ad = Power
           |         sex
    recall |    Female       Male |     Total
-----------+----------------------+----------
        No |        53         33 |        86
       Yes |        17         26 |        43
-----------+----------------------+----------
     Total |        70         59 |       129

-> ad = Safety
           |        sex
    recall |    Female       Male |     Total
-----------+----------------------+----------
        No |        25         54 |        79
       Yes |        36         10 |        46
-----------+----------------------+----------
     Total |        61         64 |       125
```

How to read these tables?

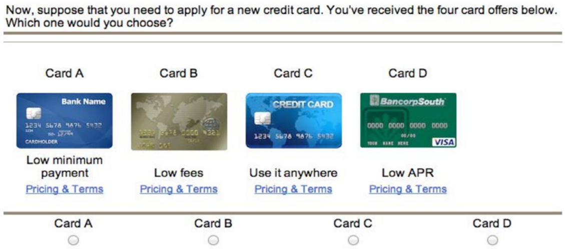


See conditional distributions in textbook (p. 64)

Academic Article: Working Paper

Abstract: The growing use of on-line educational content and related video services has changed the way people access education, share knowledge, and possibly make life decisions. In this paper, we characterize how video content affects individual decision-making and willingness to share in the context of a personal financial decision. Content geared toward giving better instructions leads to better financial decisions, but less information sharing. Misleading advertising not only causes worse decisions, but makes it less likely that videos with useful content get shared in the market. This implies that the effects of deception have externalities on other peoples' literacy and decision-making. Our work has important implications for policies guiding financial literacy training, and also has broader impact for education in the information age. http://www.nber.org/papers/w20268.pdf [Carlin, Li, and Spiller (2014)]

"Learning Millennial-Style" *NBER*, 2014



Now, suppose that you need to apply for a new credit card. You've received the four card offers below. Which one would you choose?

Card A	Card B	Card C	Card D
Low minimum payment	Low fees	Use it anywhere	Low APR
Pricing & Terms	Pricing & Terms	Pricing & Terms	Pricing & Terms

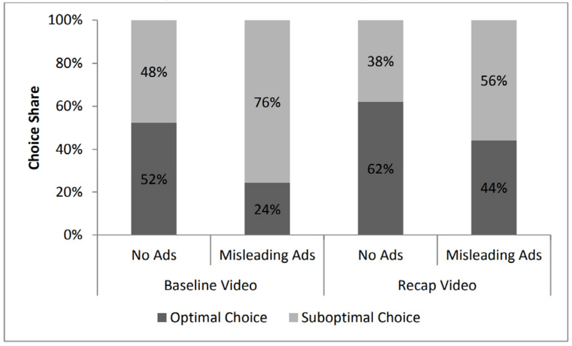
Card A ○ Card B ○ Card C ○ Card D ○

Video: https://player.vimeo.com/video/70597491

Before we see the Carlin et. al (2014) results … Some Important Reminders

- Read our course syllabus
- DACM Module A.1 this Friday morning
  - Read pages 1 – 20 of the DACM Handbook *before*
- Catch-up: read Chap. 1 – 4 (today's readings)
- Visit our course site and complete HW 1 and readings for Lecture 2 (Sections 5.1 – 5.6)
- First Quercus quiz <u>due by Mon, Sept 16, 6pm</u>
- Bring your iClicker remote to Lecture 2

Figure 4. Choice share per condition. Darker areas represent choice of the best card.



The total sample size (number of observations) is $n = 401$. In Module A.1 of DACM, you will work with the 2017 published version of this paper and see a similar figure but for a fresh sample of $n = 1{,}603$.