ECO220Y
INTRODUCTION
Readings: Chapter 1, 2, 3, 5

Fall 2010

Lecture 1

Syllabus

- Hours and Locations
- Textbook
- Assessments
- Questions?

What is Statistics?

What makes statistics unique is its ability to *quantify* uncertainty, to make it precise. This allows statisticians to make *categorical* statements, with complete assurance - about their level of uncertainty!

Compare:

"It is likely to rain today"

and

"I am 95% confident that the chance of rain today is between 73% and 77%"

"Hal Varian, Google's chief economist, predicts that the job of statistician will become the "sexiest" around. Data, he explains, are widely available; what is scarce is the ability to extract wisdom from them."

"...The ability to take the data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a *hugely* important skill in the next decades, ..."

The Economist, Feb 27th-March 5th 2010

`http://www.youtube.com/watch?v=D4FQsYTbLoI`

Expectations (At the Beginning of the Year)

- Set up and solve problems with algebra
- Sums ($\sum$) and inequalities ($\leq$, $\geq$, $>$, $<$, $\neq$)
- Linear and non-linear functions:
  -equation of line, slope, intercept
  -powers, square roots, logs, exponentials
- Graphing equations and finding simple areas
- Common symbols:∞, exp, $\approx$, $\pm$

Expectations (At the End of the Year)

- Know when using each method is appropriate
- Interpret results correctly
- Think critically about analysis done by others

What we will study in this course?

1. Data Analysis: The gathering, display, and summary of data.
2. Probability Theory: The laws of chance.
3. Inference: The science of drawing statistical conclusions from specific data, using a knowledge of probability.

Our Methods and Tools

1. Describe data (=sample) using statistics.
2. Make inference about population using observed data (=sample)
   - Population - the group of all items of interest.
   - Parameter - descriptive measure of a population. Parameter is constant in a population.
   - Sample - a (sub)set of data drawn from the population.
   - Statistic - descriptive measure of a sample. Varies across samples from the same population $\longrightarrow$ Variable!

Parameters are usually denoted by Greek Letters

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

Statistics are denoted with English Letters: s, r, $\bar{X}$

Population vs Sample

Why uncertainty?

We do not always have the luxury to learn about all the items in the population of interest.

We use samples from populations. This brings in sampling error or noise.

Sampling Error or Noise

Definition

Sampling error is a purely random difference between a sample and population of interest that arises because the sample is a random subset of the population.

Illustration of Sampling Error

Goal: Estimate average number of children per household for a population with three households:

Household A:	1 child
Household B:	2 children
Household C:	3 children

Average based on the full population is 2 children per household.

$$(1+2+3)/3=6/3=2$$

Example Cont'd

What if we have three different samples?

Sample	Number of Children	Average Number
Households A and B:	1 and 2 children	$(1+2)/2 = 1.5$
Households B and C:	2 and 3 children	$(2+3)/2 = 2.5$
Households A and C:	1 and 3 children	$(1+3)/2 = 2$

Sampling Error: Only in one case out of three the sample average is equal to the population average!

Chapter 2

- Variables and Observations
- Types of Data
- Types of Data sets

Data

- Data (*plural*) come in a variety of forms: numbers, names, sequences of letters and digits, etc...
- Observations are items/individuals/cases we are seeking to learn about
- Variables are characteristics or attributes of each item/individual/case

Observations and Variables

id	gender	age	birth order	work hours
1	1	13	2	12
2	1	10	3	6
3	0	8	1	8
4	1	15	2	21
5	1	14	1	10
...	...	...	...	
1000	0	6	4	14
1001	1	10	3	15

Types of Variables

- Quantitative: measures the quantity; numerical; we can perform arithmetic operations.
- Categorical: sort answers into categories; numerical/words; often "yes" or "no", or ranking.
  - Ordinal: order of categories matters (5=Extremely helpful,..., 1=Not helpful at all)
  - Nominal: no particular order of categories (1=Male, 2=Female)
  - Identifier: special case when number of observations is equal to the number of categories

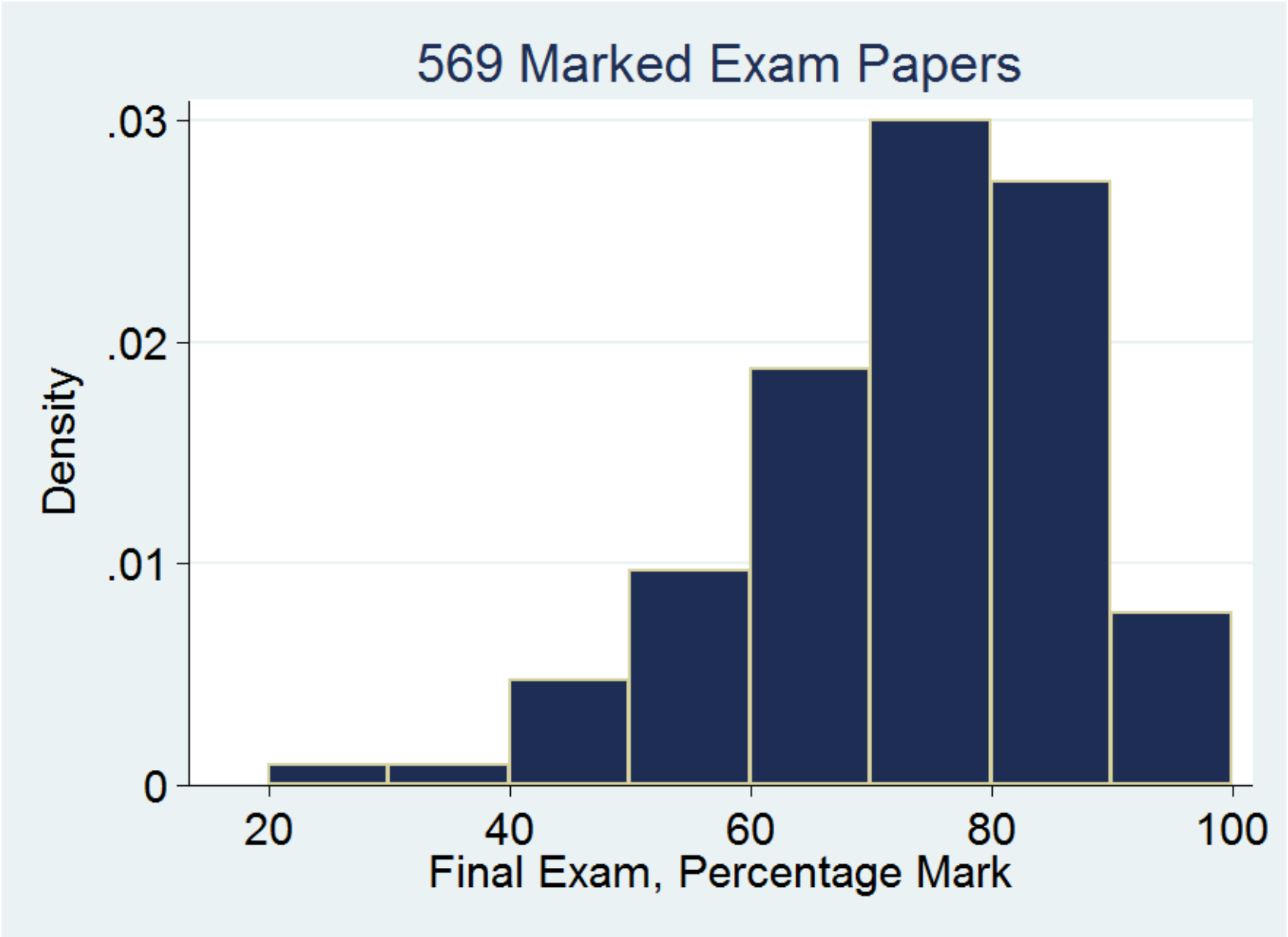
Types of Data sets

- Cross-section (Different units captured at a single point in time)
  - GDP growth rates in 2010 for OECD countries
  - Retail gas prices as of September 15th in 100 randomly selected gas stations in Toronto
- Time series (Same unit followed over time)
  - GDP growth rates for Canada in 1960-2010
  - Retail price of gas at the Finch/Dufferin gas station recorded daily in September
- Panel (Different units followed over time)
  -
  -

Describing Quantitative Variables

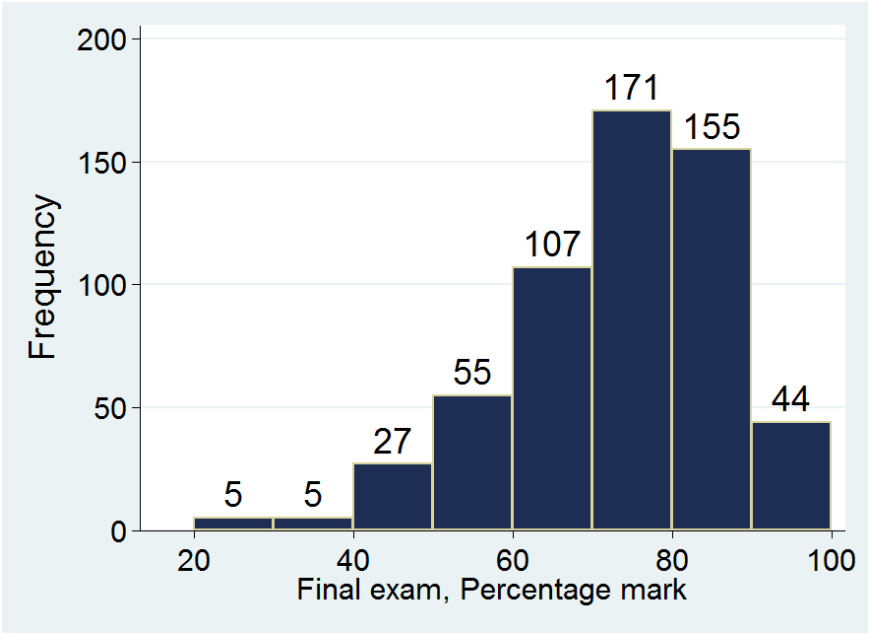
- Histograms ✓
  - Histogram is a convenient way to summarize the distribution of the data.
  - Thee types of histogram: frequency, relative frequency and density.
- Stem-and-Leaf Displays

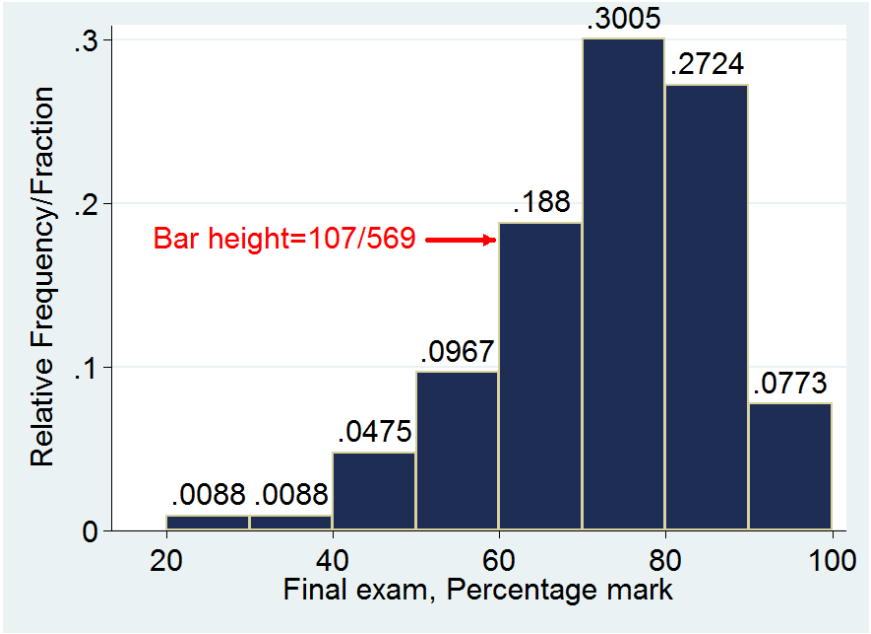
Histogram

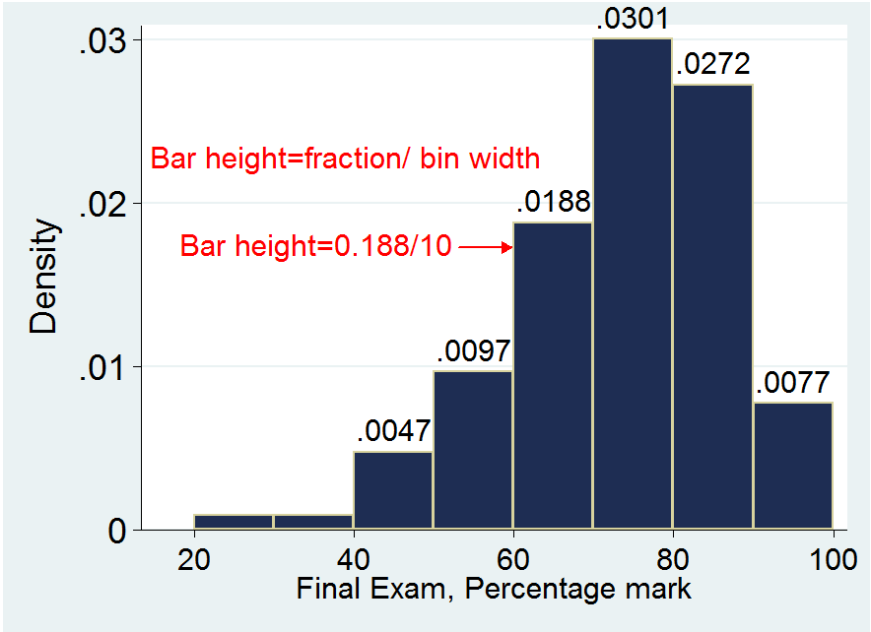

569 Marked Exam Papers
.03
.02
.01
0
Density
20
40
60
80
100
Final Exam, Percentage Mark

Frequency Table

Range	Frequency	Fraction	Density
20-29	5	0.0088	0.0008
30-39	5	0.0088	0.0008
40-49	27	0.0474	0.0047
50-59	55	0.0967	0.0097
60-69	107	0.1880	0.0188
70-79	171	0.3005	0.0301
80-89	155	0.2724	0.0272
90-100	44	0.0773	0.0077
Total	569	1	
	↑ Frequency histogram	↑ Relative frequency	↑ Density histogram


Frequency
200
150
100
50
0
5
5
27
55
107
171
155
44
20
40
60
80
100
Final exam, Percentage mark


Relative Frequency/Fraction
.3
.2
.1
0
.0088
.0088
.0475
.0967
.188
.3005
.2724
.0773
Bar height=107/569
20
40
60
80
100
Final exam, Percentage mark


Density
.03
.02
.01
0
.0047
.0097
.0188
.0301
.0272
.0077
Bar height=fraction/ bin width
Bar height=0.188/10
20
40
60
80
100
Final Exam, Percentage mark

Guidelines for forming class intervals

1. Use Sturges' formula to pick appropriate number:
   - # of bins= 1+ 3.3*log(n),
2. or use intervals of equal length with midpoints at convenient round numbers.
3. For a small dataset, use a small number of intervals.
4. For a large dataset, use more intervals!