Introduction to Statistics

Ramon C. Littell

littell@ufl.edu

What is Statistics?

The purpose of statistics: To make *inference* about unknown quantities from *samples* of data.

Basically, you have questions about a set of objects that are too numerous to observe in its entirety. The large set of objects is called a *population*. But can observe a subset of the objects, called a *sample*. You obtain a *sample* from the population and observe each object in the sample, record data, and use information in the sample to make inference about the population.

For example: You want to know something about the age distribution of undergraduate students at the University of Florida; that is, how many ages are <18, <19, <20, <21, <22, etc. Or, you might want to know the average age, or the age range.

In either case you want information about the set of ages of all UF undergraduate students. These ages would be the *population* of interest. (Note: the *ages* are the population, not the *people*.)

It is infeasible to get the ages of all UF students. You cannot observe the entire population. Instead, you get ages of a subset of the population. The subset is called a *sample*. Then, you use the data in the sample to *estimate* what you want to know about the population.

Getting a Sample of Data from a Population

There are several ways to get a sample of data from a population. In the case of the population of ages of UF graduate students, here are some examples:

1. Draw 100 names of undergraduate students at random from the UF Student Directory. Contact them and ask their ages.
2. Get the ages of the students in STA 3032 during a particular semester.
3. Go to a bar during finals week and ask the ages of all the patrons.

Each of these approaches has its own drawbacks. Probably the first approach is best and the third is worst. The second approach might be acceptable, to the extent that students who take STA 3032 represent all UF undergraduate students.

Types of Sample:

Simple Random Sample: Each subset of a given size has the same chance of being drawn.

Convenience Sample: Using data that is immediately available.

Types of Populations:

Tangible Populations: Populations whose members physically exist; e.g. ages of UF undergraduate students.

Conceptual Populations: Populations whose members exist only in our imaginations; e.g. breaking strength of pencils all of a certain type that you could possibly use this semester.

Types of research studies and sources of data

1. *Designed experiments*: Treatments are applied to experimental units according to a prescribed plan
2. *Surveys*: Data are collected on existing units selected form a population according to a prescribed plan
3. *Observational Studies*: Data are gathered on units that are available

Questions from the previous page:

What would you call the first example of getting a sample?

What would you call the second example of getting a sample?

Other Examples of Populations and Samples

Populations:

1. Numbers of cars passing an intersection in an hour
2. Serum zinc levels in dogs in Gainesville area
3. Strengths of concrete from given mix of sand, cement and gravel

Samples:

1. Counts of cars passing the intersection in a specified set of hours
2. Serum zinc levels in dogs entering UF College of Veterinary Medicine Small Animal Clinic
3. Measurements from samples of concrete with known ingredients in concrete mix

Data Summarization

It is usually difficult to learn much about a set of measurements from a list. If you wanted to report information about the ages of UF graduate students, you would probably employ some method of data summarization.

Here are some possible ways to summarize the data:

1. Report the *mean* or the *range* of the data
2. Report how many values are in various age *categories*
3. Construct a *graph* to display the data

Example of Data Summarization

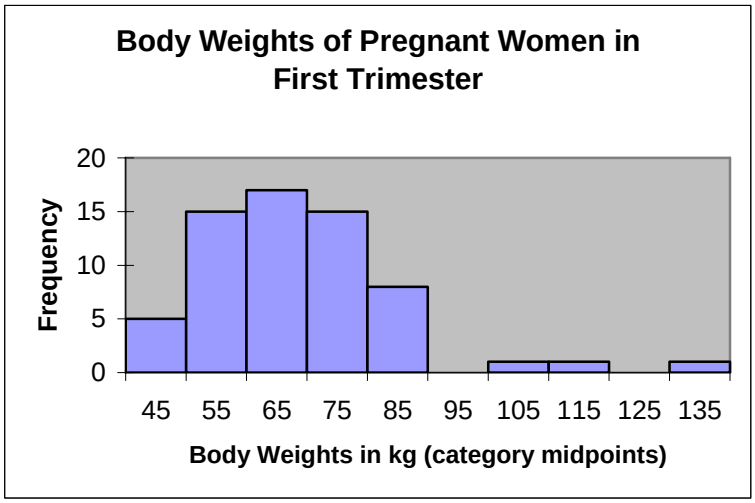
Sixty-three pregnant women participated in a nutritional intake study. As a baseline indicator, their bodyweights (in kg) were recorded at the end of the first trimester. Here are the data:

42.3	51.8	61.4	70.2	80.5		104.5
44.8	52.7	61.8	70.5	81.8		112.0
47.3	53.6	62.3	70.5	84.8		131.8
48.9	53.9	62.3	70.7	84.8		
49.5	55.0	63.0	71.4	86.4		
	55.5	63.2	72.0	86.4		
	55.9	63.4	72.7	88.2		
	56.4	64.1	73.9	89.8		
	57.0	64.3	74.5			
	57.0	64.8	74.8			
	57.0	66.6	75.0			
	57.5	66.8	75.5			
	57.5	67.3	75.7			
	59.1	68.2	75.9			
	59.3	68.2	75.9			
		68.9				
		69.8				
5	15	17	15	8	0	3

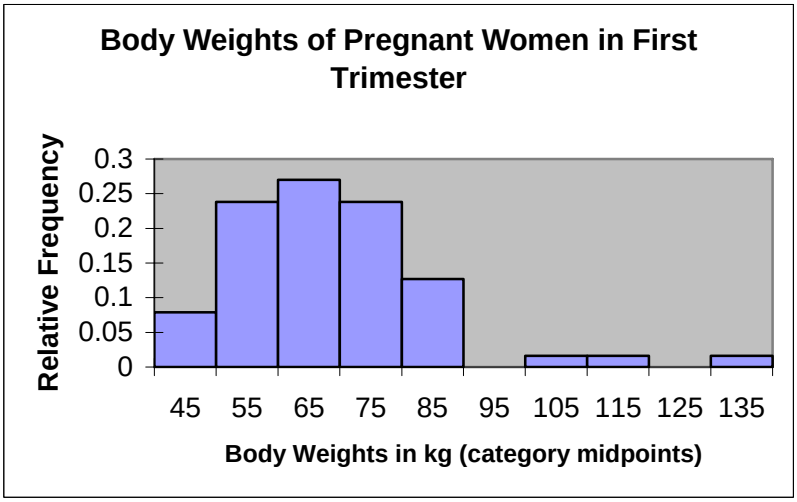
Summary Statistics:

Min:	42.3
Max:	131.8
Mean:	68.4
Range:	89.5
Standard deviation:	15.6

Frequency Histogram of Bodyweight Data



Relative Frequency Histogram of Bodyweights



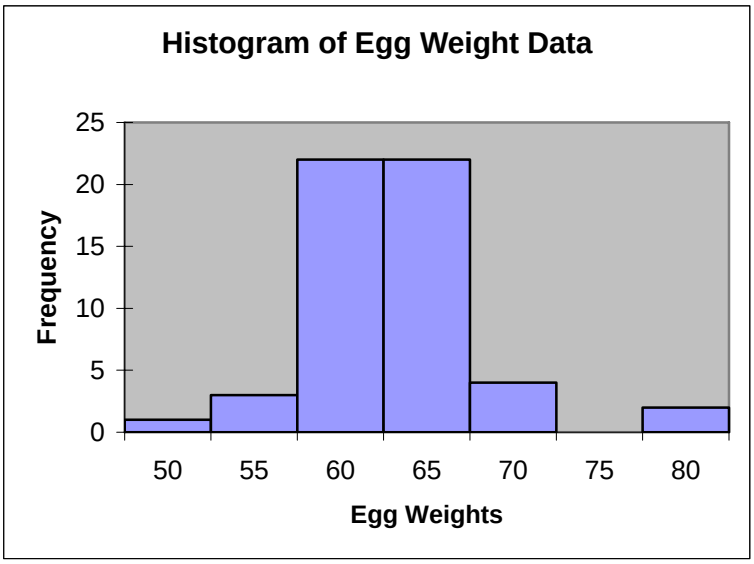
Guideline for Histogram Construction

Divide range of data into 5 to 20 intervals.
Counts number of data values in each interval.
Draw bars whose heights reflect counts.

Another Example of Data Description

Egg weights on particular date from 54 hens
53.4, 55.2,…, 80.8, 83.1 range = 83.1 – 53.4 = 29.7

intervals	50-55	55-60	60-65	65-70	70-75	75-80	80-85
freq	1	3	22	22	4	0	2
rel freq	.0185	.0556	.4074	.4074	.0741	.0	.0370



Question:

How do the body weights of pregnant women characteristically differ from the egg weights? Does this surprise you?

Sample descriptive statistics

- Data (y_i) $y_1 = 67.0, y_2 = 71.2, \ldots, y_{53} = 83.1, y_{54} = 69.7$
- Sample size n=54
- Sum $y_1 + y_2 + \cdots + y_{53} + y_{54} = \sum_{i=1}^{54} y_i = 3531.3$
- Mean $\bar{y} = \sum y_i / n = 3531.3/54 = 65.39$
- Ordered data
  $y_{(1)} = 53.4, y_{(2)} = 55.2, \ldots, y_{(53)} = 80.8, y_{(54)} = 83.1$
- Median $\left(y_{(27)} + y_{(28)}\right)/2 = \left(65.2 + 65.3\right)/2 = 65.25$
- 75^{th} Percentile $(.75)54 = 40.5$ $y_{(41)} = 68.1$
- 25^{th} Percentile $(.25)54 = 13.5$ $y_{(14)} = 62.1$

Interpretation:
No more than 25% below and no more than 75% above 62.1
No more than 75% below and no more than 25% above 68.1

Measures of Central Tendency

- Mean $65.39 = \overline{y}$
- Median $65.25 = 50^{th}$ percentile (middle value)
- Mode 62.7 = most frequently occurring observation

Measures of Dispersion

- Range $y_{max} - y_{min} = 83.1 - 53.4$

- Inter-quartile range $q_3 - q_1 = 68.1 - 62.1 = 6.0$

- Variance $$s^2 = \frac{\sum (y_i - \overline{y})^2}{n-1} = \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n-1} = 26.747$$

- Standard deviation $$s = \sqrt{s^2} = \sqrt{26.747} = 5.172$$

Empirical Rule

The *Empirical Rule* provides a practical use of the standard deviation

- If the distribution is "mound-shaped" then:

  Approx. 68% of the data are between $\bar{y} - s$ and $\bar{y} + s$
  Approx. 95% of the data are between $\bar{y} - 2s$ and $\bar{y} + 2s$
  Approx. 99% of the data are between $\bar{y} - 3s$ and $\bar{y} + 3s$

Empirical Rule for Egg Weight Data

Egg Weights

1	53.4
1	55.2
2	58.3 59.2
7	60.2 60.3 61.0 61.4 61.5 61.5 61.8
12	62.0 62.0 62.1 62.2 62.6 62.7 62.7 62.7 63.0 63.0 63.5 63.6
8	64.3 64.5 64.7 65.2 65.3 65.4 65.4 65.9
9	66.0 66.0 66.0 66.3 67.0 67.0 67.4 67.5 67.6
8	68.1 68.2 68.8 69.0 69.1 69.2 69.7 69.8
2	71.2 71.8
2	72.0 73.1
1	80.8
1	83.1

$\bar{y} = 65.39 \quad s = 5.17$

Lower	Upper	Count	%
$\bar{y} - s = 60.22$	$\bar{y} + s = 70.56$	43	79
$\bar{y} - 2s = 55.02$	$\bar{y} + 2s = 75.73$	51	95
$\bar{y} - 3s = 49.88$	$\bar{y} + 3s = 80.90$	53	98

Populations and Samples
Parameters and Statistics

There are means, standard deviations, etc. for samples and populations, but conventionally use different notation.

Sample quantities are called *statistics*. You *can compute* statistics because the sample values are available to you.

Population quantities are called *parameters*. You *cannot compute* parameters because not all of the population values are available to you.

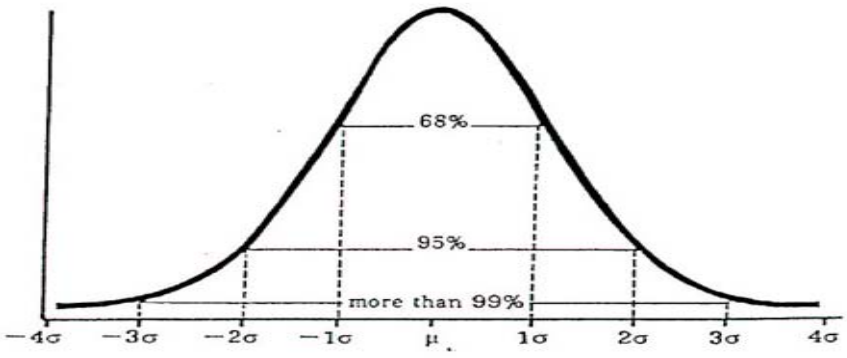
Notation for Populations and Samples

	Sample Statistic	Population Parameter
Mean	$\overline{y}$	μ
Variance	s^2	σ^2
Standard Deviation	s	σ

Sample Statistics are *estimates* of the corresponding Population Parameters

Empirical Rule for Normally Distributed Population

- 68% of measurements are between $\bar{y} - \sigma$ and $\bar{y} + \sigma$
- 95% of measurements are between $\bar{y} - 2\sigma$ and $\bar{y} + 2\sigma$
- >99% of measurements are between $\bar{y} - 3\sigma$ and $\bar{y} - 3\sigma$



Normal Distribution and the Empirical Rule

Other Graphical Procedures

- Box Plot
- Stem–and–leaf
- Distribution function
- Normal probability plot

SAS PROC UNIVARIATE Output

The UNIVARIATE Procedure
Variable: ew

Moments

N	54	Sum Weights	54
Mean	65.3944444	Sum Observations	3531.3
Std Deviation	5.17178181	Variance	26.747327
Skewness	0.93676395	Kurtosis	2.83048891
Uncorrected SS	232345.01	Corrected SS	1417.60833
Coeff Variation	7.90859506	Std Error Mean	0.70379036

Basic Statistical Measures

Location		Variability	
Mean	65.39444	Std Deviation	5.17178
Median	65.25000	Variance	26.74733
Mode	62.70000	Range	29.70000
		Interquartile Range	6.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 3.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	92.91751	Pr > |t|	<.0001
Sign	M	27	Pr >= |M|	<.0001
Signed Rank	S	742.5	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	83.10
99%	83.10
95%	73.10
90%	71.20
75% Q3	68.10
50% Median	65.25
25% Q1	62.10
10%	60.30
5%	58.30
1%	53.40
0% Min	53.40

SAS PROC UNIVARIATE Output

The UNIVARIATE Procedure
Variable: ew

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
53.4	1	71.8	50
55.2	2	72.0	51
58.3	3	73.1	52
59.2	4	80.8	53
60.2	5	83.1	54

Stem	Leaf	#	Boxplot
82	1	1	0
80	8	1	0
78			
76			
74			
72	01	2	|
70	28	2	|
68	12801278	8	+-----+
66	000300456	9	| |
64	35723449	8	*--+--*
62	001267770056	12	+-----+
60	2304558	7	|
58	32	2	|
56			|
54	2	1	|
52	4	1	|

----+----+----+----+

SAS PROC UNIVARIATE Output