MICHIGAN STATE UNIVERSITY

AFRE 802
Statistical Methods for Agricultural, Food, & Resource Economists

Course overview & introduction to statistics (WMS Ch. 1)

August 31, 2017

Nicole Mason
Michigan State University
Fall 2017

MICHIGAN STATE UNIVERSITY

GAME PLAN

1. Introductions
2. Syllabus & schedule
3. Intro to statistics (WMS Ch. 1)
   a. What is it & why study it?
   b. Summarizing data
   c. Sample means, variances, and std. dev.
   d. The empirical rule
4. (Time-permitting) Basic summary stats in Stata

1

MICHIGAN STATE UNIVERSITY

Introductions

- Name
- Country of origin
- Where you did your undergrad/MS & major
- Grad program at MSU
- Current level of study (MS/PhD)
- Previous statistics/probability courses?
- Research interests

2

MICHIGAN STATE UNIVERSITY

Syllabus

- Main topics:
  - What is statistics and why study it?
  - Probability
  - Random variables and their probability distributions
  - Sampling distributions and the Central Limit Theorem
  - Estimation
  - Hypothesis testing
  - Ordinary least squares (OLS) / regression analysis
- Syllabus and quick tour of D2L

3

MICHIGAN STATE UNIVERSITY

What is statistics?

- "Statistics is about data. … Statistics uses data to gain insight and draw conclusions." (Source: CalPoly)
- "The practice or science of collecting and analyzing [and interpreting] numerical data in large quantities, especially for the purpose of inferring proportions in a whole [population] from those in a representative sample." (Source: CalPoly)
- Also "effective communication and presentation of results relying on data" (Source: BU)

4

MICHIGAN STATE UNIVERSITY

Why study it?

- To learn how to **transform data into information**
- So that you can
  - Do **good quantitative research**
  - Produce a **high quality thesis**
  - **Interpret and evaluate** others' work
  - **Contribute to knowledge, public policy**, etc.
  - …and so that you are ready **for AFRE 835**!

5

MICHIGAN STATE UNIVERSITY

The objectives of statistics

1. To make an **inference about a population based on info in a sample** from that population

2. To provide a **measure of the 'goodness'** of that inference

- So defining your population of interest is key. *Examples from your research?*

6

MICHIGAN STATE UNIVERSITY

Put another way, the point of statistics is:

1. To **summarize** huge quantities of **data**.
2. To make **better decisions**.
3. To **answer** important **social questions**
4. To **recognize patterns**
5. To **evaluate the effectiveness** of policies, programs, innovations, etc.
6. To be able to **distinguish good statistical work from not so good**

Charles Whelan, Naked Statistics (2013, p. 14)

7

MICHIGAN STATE UNIVERSITY

Summarizing data

- In order to make inferences, we need to characterize/summarize our data
- Say we have data from nearly 9,000 smallholder farm households in Zambia on their:
  - Area cultivated
  - Education level of the HH head
- *How might we summarize these data in a useful way so that we don't have to look at 9,000 individual data points?*

8

Summarizing a set of measurements: **frequency table**

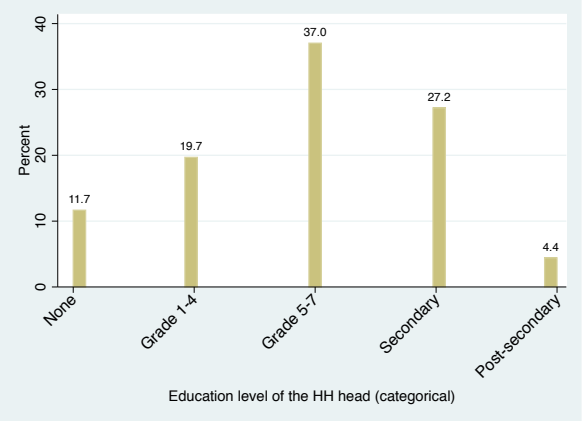
- Table showing the proportion or % of observations at each value (or range of values) in the dataset
- Stata: <tab *varname*>

Education level of the HH head (categorical)	Freq.	Percent	Cum.
None	1,030	11.66	11.66
Grade 1-4	1,739	19.69	31.34
Grade 5-7	3,271	37.03	68.37
Secondary	2,403	27.20	95.57
Post-secondary	391	4.43	100.00
Total	8,834	100.00	

9

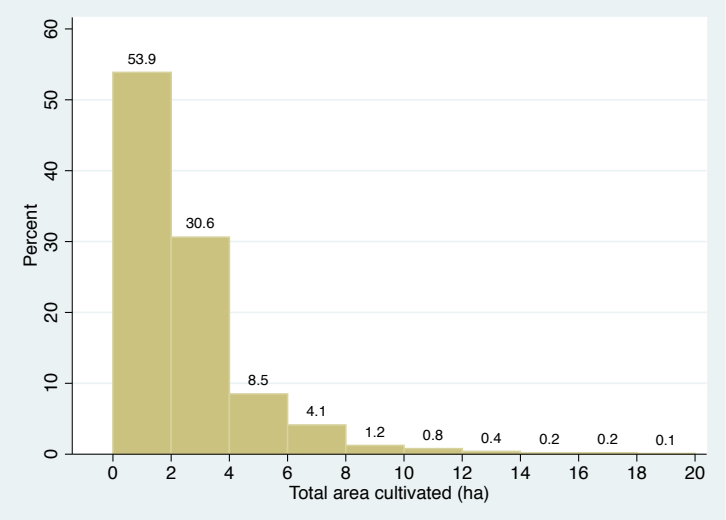
Summarizing a set of measurements: **relative frequency histogram**

- Graph showing the proportion or % of observations at each value (or range of values)
- Stata: <histogram *varname*, percent>



10

Summarizing a set of measurements: **relative frequency histogram**



Note: values falling on boundary are included in the bar for the upper category (e.g., 2.00 ha is in the 2-4 ha bar)

11

Summarizing a set of measurements: **mean, median, mode**

- *Definitions?*

Sample mean: $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ Population mean: μ

- In Stata:
  - <summarize *varname*> (gives mean and other stats)

```
summarize hect_cult
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hect_cult	8834	2.305795	2.269354	0	20

12

Summarizing a set of measurements: **mean, median, mode**

- In Stata:
  - <summarize *varname*, detail> (gives mean, median, other percentiles, 4 smallest & 4 largest values, etc.)

```
. summarize hect_cult, detail
```

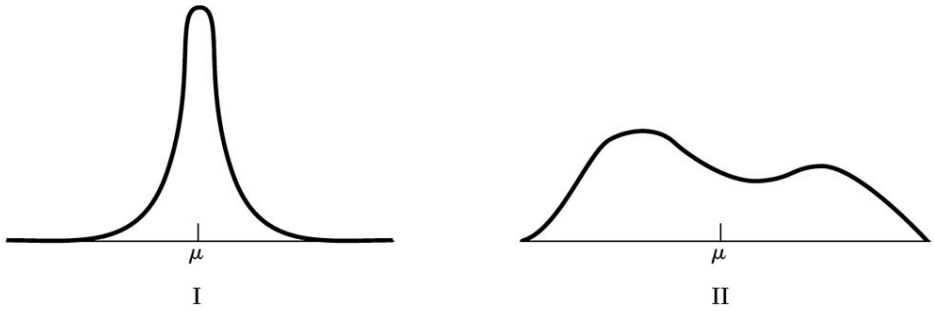
Total area cultivated (ha)

	Percentiles	Smallest		
1%	0	0		
5%	.25	0		
10%	.405	0	Obs	8834
25%	.81	0	Sum of Wgt.	8834
50%	1.75		Mean	2.305795
		Largest	Std. Dev.	2.269354
75%	2.875	19		
90%	5.12	19	Variance	5.149966
95%	6.5	19.105	Skewness	2.557384
99%	11.25	20	Kurtosis	12.86027

13

Summarizing a set of measurements: **variance & standard deviation**

- Measures of dispersion, variability



14

Summarizing a set of measurements: **variance & standard deviation**

Sample variance: $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

Population variance: σ^2

Sample standard deviation: $s = \sqrt{s^2}$

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

15

Summarizing a set of measurements: **variance & standard deviation**

```
. summarize hect_cult, detail

                Total area cultivated (ha)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%         .25              0
10%        .405              0       Obs                8834
25%         .81              0       Sum of Wgt.        8834

50%        1.75                     Mean               2.305795
                          Largest    Std. Dev.          2.269354
75%       2.875             19
90%        5.12             19      Variance           5.149966
95%         6.5          19.105      Skewness           2.557384
99%       11.25             20      Kurtosis           12.86027
```

16

EX) Sample mean, variance & std. dev.

Obs. #	Ha cultivated
1	2
2	0
3	2
4	1.1
5	2.5
6	0.5
7	5.5
8	1.1
9	7
10	1
N = 10	Sum = 22.7

Sample mean

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

Sample variance

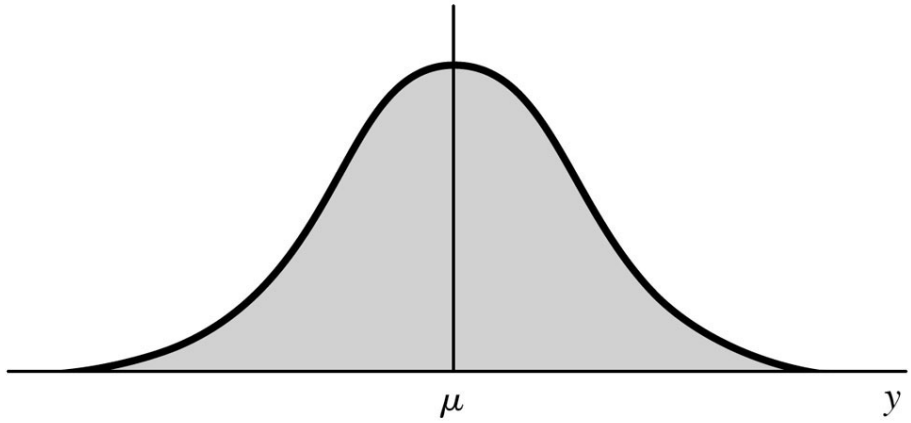
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Sample std. dev.

17

Normal distribution

- Bell-shaped curve, symmetric

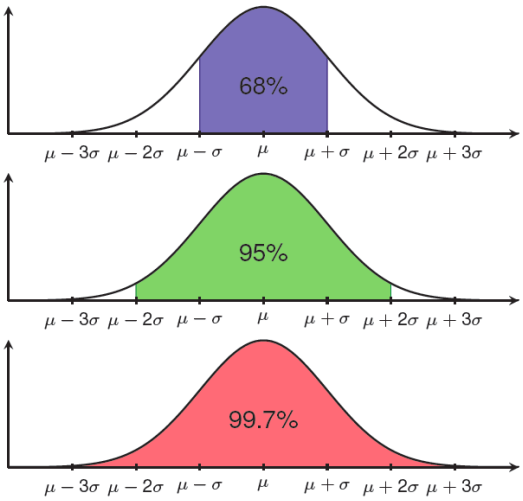


18

The "empirical rule"

For a distribution that is approximately normal:

- 68% of obs. are w/in σ of μ
- 95% of obs. are w/in 2σ of μ
- Almost all obs. are w/in 3σ of μ



19

The "empirical rule"

For a distribution that is approximately normal:

- 68% of obs. are w/in σ of μ
- 95% of obs. are w/in 2σ of μ
- Almost all obs. are w/in 3σ of μ

EXAMPLE

Scores on the quantitative portion of the GRE in 2011 were approximately normally distributed and averaged 151 with a standard deviation of 9.

a. Approximately what percentage of test-takers got scores between 142 and 160?
b. An elite grad school only considers applicants with quantitative GRE scores in the top 2.5%. What minimum score would this be?

20

MICHIGAN STATE UNIVERSITY

(Time-permitting) Basic summary stats in Stata

- The syntax file (a.k.a. "do-file") and data file are available on D2L

21

MICHIGAN STATE UNIVERSITY

Homework:

- WMS Ch. 1
  - Populations: 1.1 (a-d; identify the population only)
  - Relative frequencies & histograms: 1.2 (a, c, d; feel free to use Excel, Stata, or other software for a), 1.6
  - Sample means, var., std. dev., empirical rule: 1.9, 1.11, 1.12, 1.21
- Due at the beginning of class on Thursday, Sep. 7

Next class: Introduction to probability (Part 1 of 2)

Reading for next class:

- WMS Ch. 2: 2.1 through 2.6

Application to look into for next class:

- What are permutations and combinations? Then pick one and find at least one example of how it it used in your field.

22