

3190 Week 1

Statistics, Science, Hypotheses,
Measurement

Statistics

- A quantitative observation of some characteristic of a sample
 - For example, a “count”
- A *sample* is a set of observations on a sub-group of a population (e.g., 3190 students)
- A *population* is the whole body of phenomena of interest (e.g., the UNT student body)
- A quantitative observation of some characteristic of a population is not a statistic, but a parameter

Science

- A sense making system created through observation and generalization that is used to *disconfirm* hypotheses
- A hypothesis = a proposed explanation for an event's causes (a potential answer to a question)
 - Scientists use multiple working hypotheses (MWH)
- In sum, the *process of science* is to use general statements assumed to be true to disconfirm MWH

Example

- Question: why are basketball players seemingly taller on average than members of the general population?
 - H_1 = They are selected because they are better at putting balls through 10 foot hoops
 - H_2 = they are more accommodating to coaches than “short people”
 - H_0 = the pattern is random (null hypothesis)
- The *null hypothesis* is the one tested against alternatives that must be rejected to accept another

Aka...

- The scientific method...
 - Which is counter-intuitive...
 - Why?
-
- Because we typically rely on a different sense making system other than science...

Common sense

- This is our native sense making system that we inherit from our culture
- It is largely confirmatory...
- “Why did he break up with me..?”
 - “Because he is an a...!”
 - And then we go look for evidence to confirm our story
 - We tend not to use MWH

Truth

- Can hypotheses be true?
- No they cannot, because any one of them can be rejected with new data
- We simply accept the best *current* answer to a question
- Science is, in a sense, a willingness to be wrong...

A role for stats

- Hypotheses, science, math....
 - Perhaps a bit intimidating
- But really quite simple...
 - In stats we have two hypotheses
 - The one we think is correct... = alternative H_a
 - That we could be wrong due to chance = null $H (H_0)$
- We simply use stats to help us describe our data and infer whether or not we can reject the H_0
 - Hence *descriptive statistics & inferential statistics*

Descriptive Statistics

- Purpose is to describe a set of data
- Data = observations of phenomena (here, usually quantitative observations)
- Description allows communication
 - E.g., UNT basketball players are “on average” taller than UNT professors
 - We can calculate an average and it is a description

Inferential Statistics

- Purpose is to analyze data from a sample(s) to learn about a population and to answer questions (test hypotheses)
 - E.g., Are two samples of people different enough in size that it is likely they come from different populations?
- Inferential stats commonly answer two questions (among many other ones)
 - 1 How similar or different are samples?
 - 2 How closely related are two or more variables?
- We will discuss elementary inferential stats later this semester

Measurement

- Four scales of measurement...
 - Nominal
 - Ordinal
 - Interval
 - Ratio

Nominal Scale

- “nominal” means “name”
- A scale that uses qualitative or categorical classes
- “Texan” is a nominal scale category, so is “New Yorker”
 - The categories convey difference
 - Do not convey *greater than* and *less than* differences

Ordinal Scale

- Measurement that places phenomena in relative order
 - Provides information on *greater than, less than* relationships, but not *how much so*
- A good example is stratigraphy in geology
 - Deeper tends to be older and shallower tends to be younger, but the position will not tell you how much so w/o independent information
- We might order several people by height, but without an independent measure we would not know *magnitude of difference*

Interval/Ratio Scales

- Measurement of variables on scales that can determine magnitude of difference
 - Measurement units are all equal (e.g., meters or grams)
- Ratio scales have a true zero point
 - E.g., weight in lbs (0 = no weight) 40 lbs is 2X heavier than 20 lbs
- Interval scale have not true zero point
 - E.g., temperature in Celsius; 40° is not 2X as warm as 20°

Measurement Error

- We use four concepts to gauge error
 - Precision
 - Accuracy
 - Validity
 - Reliability

Precision

- Level of exactness of measurement
- The finer the measurement the more precise
 - E.g., Length measured to the millimeter is more precise than that to the centimeter
 - A rain gauge that measures to the inch is less precise than one that measures to the 1/16 inch
- Precision is irrelevant if measurement is inaccurate

Accuracy

- Extent of systematic bias in measurement
 - E.g., shooting to the same spot always to the left of the target is precise but inaccurate
 - E.g., a rain gauge with a wad of paper towel in the bottom of it may still be precise (e.g., to the 1/6 inch) but is *inaccurate*
 - Deer astragalus/caliper example...

Validity

- Consideration of whether or not the most appropriate variable is being examined to answer the research question at hand.
 - If we are interested in height, it does not matter how precisely and accurately we measure weight
 - Deer weight vs astragalus size

Reliability

- An ability to collect and re-collect data in a repeatable, precise, accurate, and valid manner
- A big concern with long-term studies that collect data multiple times
- The caliper example is a good one here...
 - The only way to assess reliability would be to go back and re-measure the same specimens

Grouping Data

- Often we want to group data so that we can more easily comprehend it's distribution in a table or chart
- Natural Breaks
- Equal Intervals (based/not based on the range)
- Quantile Breaks (here quartiles)

Natural Breaks

- Relies on gaps in the distribution of the values of data
 - You can see many gaps in Table 2.5 (3.3, 3.9, 4.7)
 - So make these the boundaries of groups
 - 2.9 to 3.3 is group 1
 - 3.4 to 3.9 is group 2
 - 4.0 to 4.7 is group 3
 - and so forth...
- Advantage = easy to understand
- Disadvantage = very subjective (are groups meaningful?)

Equal Interval Based on the Range

- Uses intervals of the data values to create groups
- Subtract minimum value from maximum = range
 - Divide the range into the desired # of categories
 - Here max = 8.1, min = 2.9, range = 5.2
 - If four groups are desired, what will the interval be?
 - What are the boundaries for groups 1-4?
 - You should have a question about group 4.

Equal Intervals *not* Based on Range

- Here the interval is based on a set of equal interval classes that encompass the range of value, but are not determined from it.
 - Pick the whole number below the minimum
 - Pick the whole number above the maximum
 - Subtract the former from the latter
 - Divide by the desired number of groups
 - Create **mutually exclusive** groups
 - So do this for Table 2.5

Quantile Breaks

- Divide # of cases (states here) as equally as possible into a desired # of groups
 - Quartiles = 4 groups
 - Quintiles = 5 groups
 - Could be any number
 - For quartiles, you want 4 groups, each with an equal number of states
 - The lowest $\frac{1}{4}$ of states, the 2nd lowest, the 3rd lowest, and the highest $\frac{1}{4}$.

Calculating Quartiles

- For $Q1 = n+1/4$
 - Gives you the position of the break
 - Here equals $51/4 = 12.75$
 - Round up & states 1 – 13 are in Q1
- For $Q2 = 2(n+1)/4 = 102/4 = 25.5$
 - Round up, states 14 – 26 are in Q2
- For $Q3 = 3(n+1)/4 = 153/4 = 38.25$
 - Round down, states 27 – 38 are in Q3
- The remaining states, 39 – 50 are in Q4

If DC & PR are included

- For $Q1 = n+1/4$
 - Gives you the position of the break
 - Here equals $53/4 = 13.25$
 - Round down & states 1 – 13 are in Q1
- For $Q2 = 2(n+1)/4 = 106/4 = 26.5$
 - Round up, states 14 – 27 are in Q2
- For $Q3 = 3(n+1)/4 = 159/4 = 39.75$
 - Round up, states 28 – 40 are in Q3
- The remaining states, 41 – 52 are in Q4

Practice with
DC, but not PR

Rules, shmooles...

- There are numerous ways to group quartiles
- Some recommend always bending into the lower group (not rounding)
 - If the Q1 is at 13.25 include 14 with the Q1 group
 - Others, such as the one we use recommend staying true to rounding
 - You will find different methods in different books & classes
 - Here use the **rounding** method, **not** the **bending** method
 - *For future reference, what matters most is that you choose a rule and follow it!*

Ordered arrays

- The simplest organizational tool for working with data is to order it
- An ordered array is a list of numerical values associated with a variable in rank order from the smallest value to the largest value
- So, are the unemployment data an ordered array?

So practice...

- Make a frequency distribution for the Table 2.5 data
 - First with the equal interval groups not based on the range
 - Then with quartile groups

Introduction to SPSS

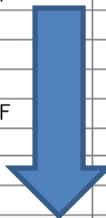
Data files (.sav)

Lab1.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : VAR00001 B

	VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	var	var	var	var
1	B	4.50	99.00	8.00	16.50	2.00				
2	B	35.00	95.00	8.00	15.50	2.00				
3	D	.50	43.00	.	.	2.00				
4	B	2.50	90.00	6.00	11.50	5.00				
5	B	1.50	78.00	6.00	9.25	8.00				
6	D	1.50	59.00	.	.	6.00				
7	B	2.50	93.00	7.00	13.50	6.00				
8	D	1.50	55.00	.	.	6.00				
9	B	4.50	101.00	16.00	16.00	4.00				
10	DF	.50	34.00	.	.	4.00				
11	D	1.50	56.00	.	.	4.00				
12	B	1.50	69.00	4.00	11.00	4.00				
13	B	1.50	57.00	4.00	5.75	4.00				
14	D	2.50	63.00	.	.	14.00				
15	S	1.50	73.00	2.00	7.25	10.00				
16	DF	.50	33.00	.	.	7.00				
17	S	1.50	66.00	2.00	8.50	6.00				
18	D	.50	41.00	.	.	6.00				
19	B	3.50	88.00	10.00	14.50	20.00				
20	B	3.50	106.00	8.00	17.00	7.00				
21	D	1.50	59.00	.	.	7.00				
22	DF	.50	29.00	.	.	6.00				
23	B	1.50	66.00	4.00	11.75	9.00				
24	B	2.50	92.00	12.00	12.50	3.00				
25	D	2.50	61.00	.	.	26.00				
26	DF	.50	27.00	.	.	12.00				
27	D	2.50	57.00	.	.	9.00				
28	B	1.50	65.00	8.00	11.00	18.00				
29	B	2.50	76.00	8.00	12.00	9.00				
30	BB	.50	39.00	2.00	3.25	18.00				



Variable View, Data View

28	B	1.50	65.00	8.00	11.00	18.00													
29	B	2.50	76.00	8.00	12.00	9.00													
30	BB	.50	39.00	2.00	3.25	18.00													
31	D	2.50	69.00	.	.	18.00													
32	B	1.50	75.00	3.00	7.00	7.00													
33	D	1.50	57.00	.	.	7.00													
34	D	2.50	55.00	.	.	7.00													
35	S	1.50	72.00	2.00	3.50	5.00													
36	B	1.50	54.00	4.00	5.75	5.00													
37	D	2.50	67.00	.	.	17.00													
38	S	1.50	50.00	2.00	5.25	17.00													
39	D	4.50	72.00	.	.	17.00													
40	B	1.50	69.00	4.00	6.50	17.00													
41	B	4.50	119.00	12.00	17.00	17.00													
42	B	2.50	92.00	7.00	12.25	17.00													
43	B	1.50	72.00	6.00	9.50	17.00													
44	S	1.50	56.00	2.00	4.00	17.00													
45	D	2.50	64.00	.	.	17.00													

SPSS Processor is ready

Variable View

Lab1.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VAR00001	String	8	0		None	None	8	Left	Nominal
2	VAR00002	Numeric	8	2		None	None	8	Right	Scale
3	VAR00003	Numeric	8	2		None	None	8	Right	Scale
4	VAR00004	Numeric	8	2		None	None	8	Right	Scale
5	VAR00005	Numeric	8	2		None	None	8	Right	Scale
6	VAR00006	Numeric	8	2		None	None	8	Right	Scale
7										
8										

Lab1.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Ali
1	VAR00001	String	8	0		None	None	8	Left
2	VAR00002	Numeric	8	2		None	None	8	Right
3	VAR00003	Numeric	8	2		None	None	8	Right
4	VAR00004	Numeric	8	2		None	None	8	Right
5	VAR00005	Numeric	8	2		None	None	8	Right
6	VAR00006	Numeric	8	2		None	None	8	Right
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									

Variable Type [?] [X]

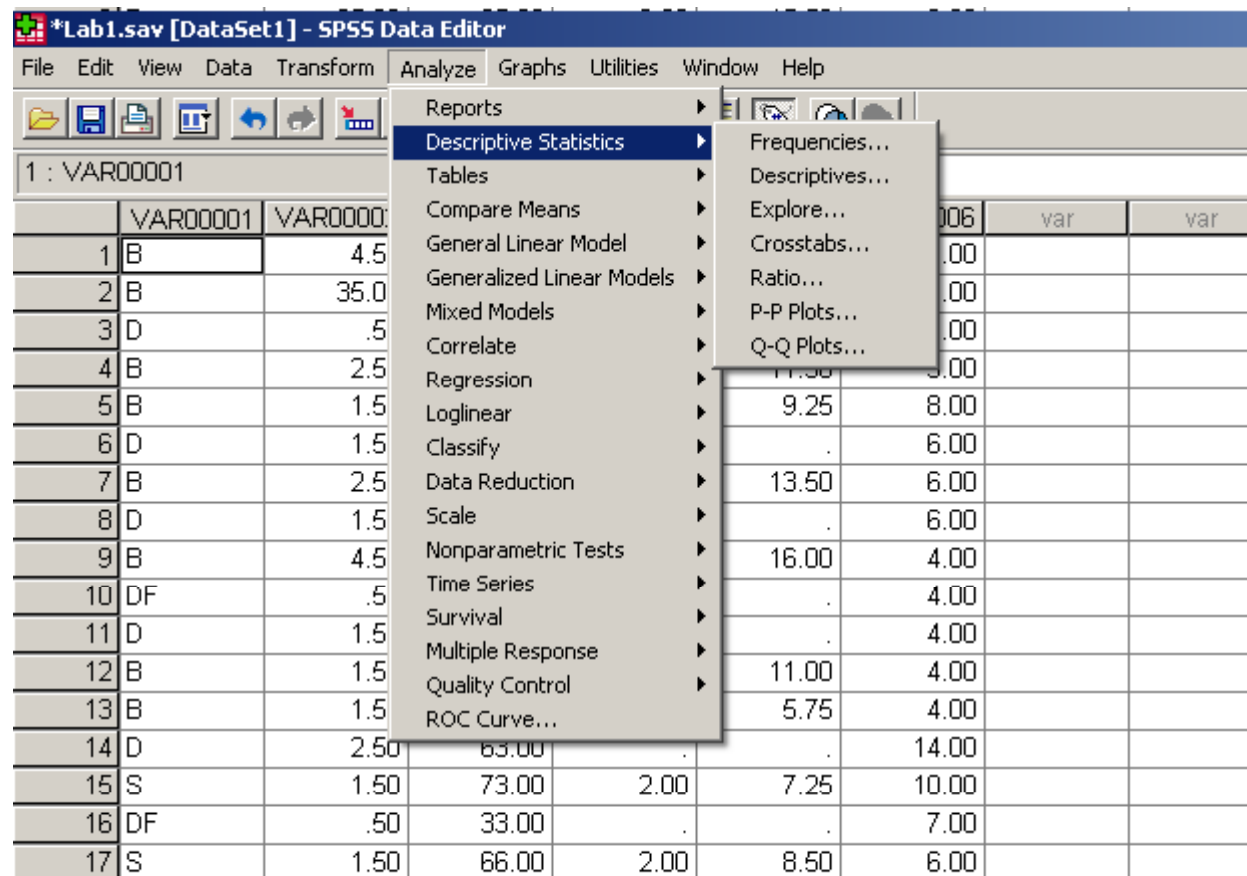
Numeric
 Comma
 Dot
 Scientific notation Characters: 8
 Date
 Dollar
 Custom currency
 String

Enter Variable Labels

- Labels must begin with a letter and can have no spaces
- VAR00001 = “Sex” Sex of the animal harvested
- VAR00002 = “Age” Age of the animal harvested
- VAR00003 = “Weight” Dressed weight of the animal
- VAR00004 = “TOTpts” Total antler points for bucks
- VAR00005 = “Spread” Spread of antlers for bucks
- VAR00006 = “tAREA” Training area of harvest

Analysis

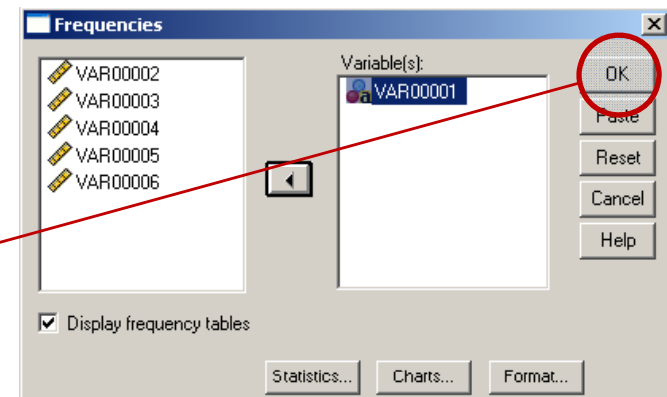
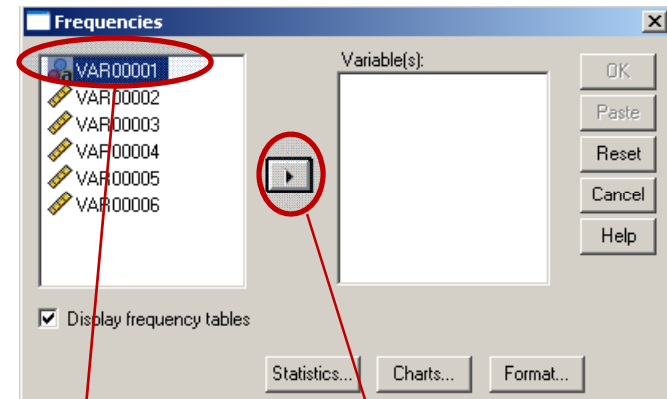
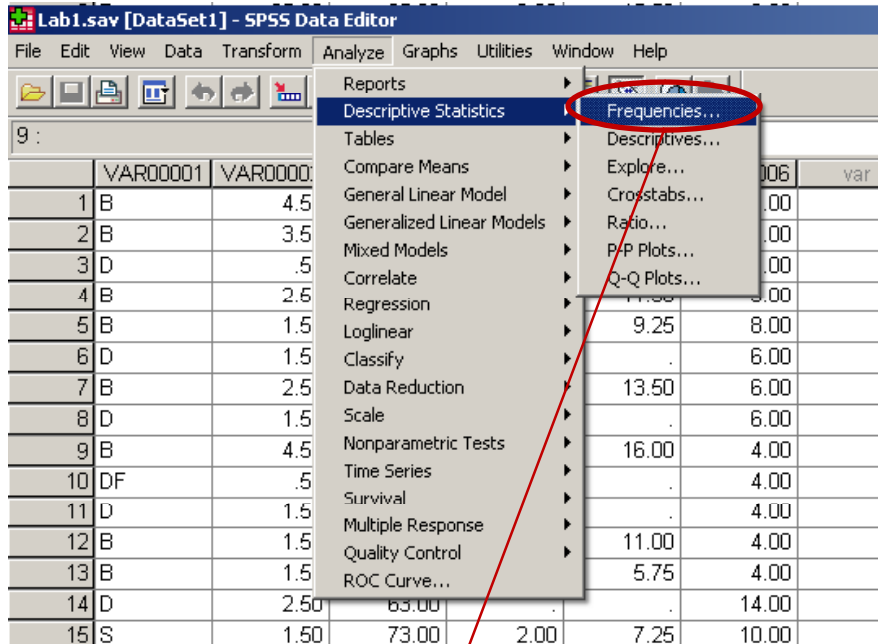
- All of the descriptive & inferential statistics that we use in this course are in the “analysis” column



Frequency Distribution

- Almost all of the descriptive statistics you will need can be found under “descriptives-statistics, frequencies”
- One of the most important tools under “frequencies” is the ability to derive a **frequency distribution** of grouped data
- In this dataset, each deer is labeled by sex, b, bb, bf, s = buck, button buck, buck fawn, or spike; d, df = doe, doe fawn
- A frequency distribution is a table that portrays counts of individuals in groups in a dataset

Creating a Frequency Distribution



Your output will look like this

Output5 [Document5] - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output
Log
Frequencies
Title
Notes
Active Dataset
Statistics
VAR00001

FREQUENCIES
VARIABLES=VAR00001
/ORDER= ANALYSIS .

→ **Frequencies**

[DataSet1] C:\Documents and Settings\wolverton\Desktop\QM2009\Week 1\Lab1.s

Statistics

VAR00001

N	Valid	2712
	Missing	0

VAR00001

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid B	890	32.8	32.8	32.8
BB	294	10.8	10.8	43.7
BF	2	.1	.1	43.7
D	1062	39.2	39.2	82.9
DF	268	9.9	9.9	92.8
S	196	7.2	7.2	100.0
Total	2712	100.0	100.0	

SPSS Processor is ready.