Lecture 2

Frequency tables

Frequency tables……..

- show
  - the summary of a single variable
  - how common (frequent) different values are

- are slightly different for nominal

vs. ordinal/interval variables

The problem with data sets

- They're (often) too big!

Name	Gender	Graduating?	Class	Major	Age	Job hours	Children?
Theda Skocpol	F	no	sr	criminology	21	10	no
Jane Addams	F	yes	jr	sociology	23	15	no
Andrew Greeley	M	no	sr	criminology	23	25	yes
Karl Marx	M	no	sr	criminology	24	35	no
Georg Simmel	M	no	sr	criminology	21	34-40	no
…							
40 more cases							

- Need to be summarized.

Frequency table: Nominal variable

- Compact summary
- Only for one variable at a time
  - Here we tabulate the Major variable and evaluate proportion and percentage

Major	Frequency (f)	Proportion (P)	Percentage
criminology	22	.489	48.9%
sociology	16	.356	35.6%
no information	3	.067	6.7%
education	1	.022	2.2%
env science	1	.022	2.2%
history	1	.022	2.2%
political science	1	.022	2.2%
Total	**45**	**1**	**100.00%**

–*Interpretation*: The data set (class roster) has 22 crim majors, etc.
–which is 22/45=.489=48.9% of the total, etc.

Exercise

Construct and interpret a frequency/percentage table for Religion of European countries.

Country	Working women	GDP per person	Urban	Religion
France	44%	$19.510	73%	Catholic
Britain	46%	$17.160	89%	Protestant
W. Germany	39%	$14.730	86%	Protestant
Italy	30%	$18.090	67%	Catholic
Netherlands	31%	$17.780	89%	Protestant
Spain	22%	$13.400	76%	Catholic
Ireland	31%	$12.830	57%	Catholic

Answer

The data set (sample) has 4 Catholic countries and 3 Protestant countries.

Religion	Frequency	Percentage
Catholic	4	57,14%
Protestant	3	42,86%
TOTAL	**7**	**100,00%**

These are 57.14% and 42.86% of the total, respectively.

Frequency table: ordinal/interval variables

Ordinal/Interval variables have order, so you can also report cumulative information.

Frequency table: Ordinal variable

- Major was a nominal variable
- Let's try an ordinal variable like Class

Class	Freq.	Cum. Freq.	%	Cum. %
Junior	10	10	22.73%	22.73%
Senior	31	41	70.45%	93.18%
Graduate	3	44	6.82%	100.00%
TOTAL	**44**		**100.00%**	

- New features:
  - Values in order: jr, sr, grad
  - Cumulative frequency, cumulative %
    - How many, what % have this value *or less*?
      - e.g., 41 are undergraduates (sr or less)

Cumulative frequencies & percentages: Calculation

Class	Freq.	Cum. Freq.	%	Cum. %
Junior	10	10	23%	23%
Senior	31	41	70%	93%
Graduate	3	44	7%	100%
TOTAL	**44**		**100%**	

41/44 = 93.18%

Answer

Losses	f	cf	%	c%
0	2	2	13.33%	13.33%
1	2	4	13.33%	26.67%
2	6	10	40.00%	66.67%
3	3	13	20.00%	86.67%
4	2	15	13.33%	100.00%
TOTAL	15		100.00%	

6 teams, or 40% of the top 15, had 2 losses each.
10, or 66.67% of the top 15, had 2 losses *or fewer*.

Example

Construct & interpret cumulative freq and % for the interval variable loess

Rank	Team	Computer Avg.	Schedule Strength	Schedule Rank	Losses	Quality Wins	Total
1	Miami (Fla.)	1.17	19	0.76	0	0	2.93
2	Ohio State	1.67	20	0.8	0	-0.5	3.97
3	Georgia	3.17	5	0.2	1	0	8.37
4	USC	3.67	1	0.04	2	-0.2	10.51
5	Iowa	4.83	49	1.96	1	0	10.79
6	Washington St.	7	21	0.84	2	-0.7	16.14
7	Oklahoma	6.33	14	0.56	2	-0.1	16.79
8	Kansas State	10.67	54	2.16	2	-0.7	20.13
9	Notre Dame	6.83	15	0.6	2	0	20.93
10	Texas	9.5	22	0.88	2	-0.3	21.08
11	Michigan	9.33	2	0.08	3	0	23.91
12	Penn State	13.33	16	0.64	3	0	26.97
13	Colorado	15.17	10	0.4	4	-0.3	33.27
14	Florida State	13.83	3	0.12	4	0	33.95
15	West Virginia	17.33	41	1.64	3	0	35.97

Percentiles

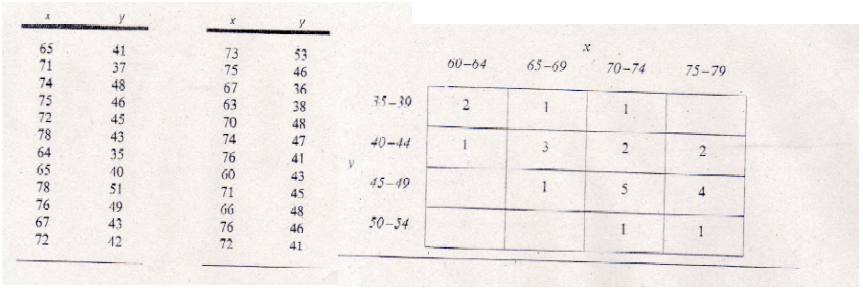
- Cumulative percent (C%) is also called *percentile.*
- Percentiles split a set of ordered data into hundredths. (Deciles split ordered data into tenths).
- The *p*th percentile is a value on a scale of 100 such that
  - at most (100*p*)% of the measurements are less than this value and at most 100(1- *p*)% are greater.
  - For example, 70% of the data should fall below the 70th percentile.
- Computation:
  - order the values in increasing order of magnitude
  - Compute the cumulative percent

Percentiles

Age	f	cf	%	c%
20	5	5	11.11%	11.11%
21	10	15	22.22%	33.33%
22	10	25	22.22%	55.56%
23	10	35	22.22%	77.78%
24	4	39	8.89%	86.67%
25	1	40	2.22%	88.89%
26	2	42	4.44%	93.33%
28	1	43	2.22%	95.56%
30	1	44	2.22%	97.78%
31	1	45	2.22%	100.00%
N	45			

- If you're 24 years old, you're in the 86.67th *percentile* for this class.
- 86.67% of the class is as young as you, or younger.

Class intervals or bins (example)



x	y
65	41
71	37
74	48
75	46
72	45
78	43
64	35
65	40
78	51
76	49
67	43
72	42

x	y
73	53
75	46
67	36
63	38
70	48
74	47
76	41
60	43
71	45
66	48
76	46
72	41

y \ x	60–64	65–69	70–74	75–79
35–39	2	1	1	
40–44	1	3	2	2
45–49		1	5	4
50–54			1	1

Class intervals or bins

- To reduce information, break age into
  - "Class intervals" (or bins)

Age	f	cf	%	c%
20-22	25	25	55.56%	55.56%
23-25	15	40	33.33%	88.89%
26-28	3	43	6.67%	95.56%
29-31	2	45	4.44%	100.00%
N	45			

Notice: All bins are the same width: important when you draw histograms!

Exercise

You definitely need bins if each case has a different value.

Country	Working women	GDP per person	Urban		
Austria	45%	$18,710	55%		
Belarus	59%	$6,440	68%		
Britain	46%	$17,160	89%		
Czech-Sloval	62%	$7,190	61%		
E. Germany	64%	$8,000	78%		
France	44%	$19,510	73%		
Hungary	48%	$6,580	63%		
Ireland	31%	$12,830	57%		
Italy	30%	$18,090	67%		
Latvia	58%	$6,060	72%		
Lithuania	56%	$3,700	70%		
Netherlands	31%	$17,780	89%		
Poland	57%	$4,830	63%		
Portugal	39%	$9,850	34%		
Romania	54%	$2,840	54%		
Slovenia	45%	$10,404	50%		
Spain	22%	$13,400	76%		
Sweden	55%	$18,320	83%		
Switzerland	43%	$22,580	60%		
W. Germany	39%	$14,730	86%		

Construct a frequency table that puts GDP in bins of $0-$5K, $5-10K etc.

Answer

Bin	*f*	cf	%	c%
$0-$5,000	3	3	15%	15%
$5,001-$10,000	6	9	30%	45%
$10,001-$15,000	4	13	20%	65%
$15,001-$20,000	6	19	30%	95%
More	1	20	5%	100%
N	20			

Distribution of a variable

Illustrates what values the variable takes, and how often it takes these values.

Frequency distribution:

- Distribution for categorical variables
- It is a Frequency table
- Lists the categories and gives the frequencies (or percent) of cases which fall in each category.
- Often quantitative variables need to be collapsed into classes or intervals

Summary

Variable	f & %	cf & c%	Class intervals or bins
Nominal			
Ordinal			
Interval			

Black—always appropriate. Blue—sometimes appropriate.
Continuous variables must use bins.