Outline

- **Data Summary**
  - Levels of Measurement
- **Frequency Distributions**
  - Reduction & Parts
  - 3 examples – Nominal, Interval, Ordinal
- **Histograms (!)**
- **Shape**
  - Choices & Cautions
- **Lab** (2 parts!)

Where we are…

#	Date	Read (5th)	Due	Area	Lecture Topic	Lab #	Lab Assignment	T Lab	R Lab
1	Tue Aug 26			*Orientation*	Welcome & Orientation	-			
2	Thu Aug 28	1.1 to 1.4			Basic Terms	-		-	-
3	Tue Sep 2	[illegible]			Measurement Issues	[illegible]	[illegible]	1	
4	**Thu Sep 4**	3.1			**Data Reduction**	2	Heaven & Hell		1, 2
5	**Tue Sep 9**	-	**HW1**	*Description*	**Display & Analysis** (Shapes & SPSS)	3	Total Miles	2, 3	
6	**Thu Sep 11**	3.2 & 3.5			**Central Tendency**	4	CT		3, 4
7	**Tue Sep 16**	3.3 & 3.4			**Dispersion**	5	Dispersion	4, 5	
8	**Thu Sep 18**	3.7			**Indices & Data Cleaning**	6	Music index		5, 6
9	**Tue Sep 23**	4.2		*Inference*	**Probability & Z Scores**	7	Standardizing Scores	6, 7	
10	**Thu Sep 25**	4.3	**HW2**		**Zs & Ps**	8	Table A		7, 8
11	**Tue Sep 30**	3.6 & 5.1			**Parameters & Pt Estimation**	9	Distances	8, 9	
12	**Thu Oct 2**	2.2 to 2.4 & 4.3			**Sampling (Issues, Methods, Effects)**	10	Sampling		9, 10
13	**Tue Oct 7**	4.4 to 4.6			**The Central Limit Theorem**	11-EC	CLT/World (EC)	10, 11*	
14	**Thu Oct 9**	5.3		*Estimation*	**Confidence Intervals**	12	CI for Intervals		11*, 12
15	**Tue Oct 14**		**HW3**		**CIs for Proportions**	13	**CI for Proportions**	12, 13	
16	**Thu Oct 16**	6.1 & 6.4			**Hypotesting & Zs**	14	Writing Hypotheses		13, 14
17	**Tue Oct 21**				**Hypotesting for Large ns**	15	Two Tests	14, 15	
18	**Thu Oct 23**	6.3 & 6.8	**HW4**		**The "t" test, for small ns**	16	CI & Test Ages		15, 16
19	**Tue Oct 28**	5.4			**Sample Size Estimation**	17	Estimating n Needed	16, 17	
20	**Thu Oct 30**	7.1, 7.3, & 10.1		*Covariation*	**Differences in Means**	18	Comparing Means		17, 18
21	**Tue Nov 4**	7.2	**HW5**		**Differences in Proportions**	19	Comparing Proportions	18, 19	
22	**Thu Nov 6**	12.1			**Analysis of Variance**	20, 21-EC	ANOVA (+ MODELS EC)		19, 20, 21*
	Tue Nov 11				(Veteran's Day)				
23	**Thu Nov 13**	9.4 & 9.5			**Scatterplots & Correlation**	22	Grade Correlations		22
24	**Tue Nov 18**	9.1 to 9.3	**HW6**		**Regression**	23	Regression Lab	20,21*,22,23	
25	**Thu Nov 20**	10.2 & 11.1			**Multiple Regression**	24-EC	Multiple Reg (EC)		23, 24*
26	**Tue Nov 25**	8.1	**HW7**	*Association*	**Crosstabulations**	25	TBA (any)	24*, 25	
	Thu Nov 27				(Thanksgiving)				
27	**Tue Dec 2**	8.2 & p.233			**Dependence**	26, 27-EC	TBA (SCU) (& 27-EC)	26, 27*	
28	**Thu Dec 4**	pp.238 to 243	**HW8**		**Association**	28-EC	Measures of Assoc (EC)		25,26,27*,28
x	Tue Dec 9				(no lecture - work session only)			28*	
x	Thu Dec 11				(no lecture - work session only)				
x	Thu Dec 18	-	HW9		(no meetings - deadline only - 10am, firm!)				

Seeking Data Summarizies

- **Can't do much w/ raw, original, unsorted data**
  - Examples coming – and lists usually much longer
- **Need ways of summarizing patterns in data**
  - What's typical about a set of cases?
  - How much do other cases differ from what's typical?
  - How does one set of cases differ from another?
- **Some statistical (inc. those 3), others graphical**
  - Bell-curves and other "shapes"
- **But everything starts w/ how the data's sorted**
  - Especially, what level of measurement…

SOC424 – Statistics w/ Dr. Ellis Godard

DATA SUMMARY:
Reduction, Tables, & Shapes

Levels of Measurement

- **Lots of ways to measure *anything***
  - Gender, Race, Income, Crime, Education
  - Length, Weight, Time – all measures are constructs
- **But only 3 choices** (in this class)
  - Nominal, Ordinal, or Interval
  - Vary in precision: N groups, O ranks, or I counts
- **And only 2 questions**
  - Can the values be *out* of order?
    - If not, it's nominal – nominal values have no order
  - Can the values be subtracted from each other?
    - If not, it's ordinal – including ranges of values
    - If they can be in order *and* can be subtracted, it's interval

Raw Nominal Data: Birth State

States could be in any order (alphabetical? length? random?) – but the *values* have no *meaningful* order, *as states*

President	State of Birth	President	State of Birth
Washington	Virginia	Harrison	Ohio
J. Adams	Massachusetts	McKinley	Ohio
Jefferson	Virginia	T. Roosevelt	New York
Madison	Virginia	Taft	Ohio
Monroe	Virginia	Wilson	Virginia
J.Q. Adams	Massachusetts	Harding	Ohio
Jackson	South Carolina	Coolidge	Vermont
Van Buren	New York	Hoover	Iowa
W.H. Harrison	Virginia	F.D. Roosevelt	New York
Tyler	Virginia	Truman	Missouri
Polk	North Carolina	Eisenhower	Texas
Taylor	Virginia	Kennedy	Massachusetts
Fillmore	New York	L.B. Johnson	Texas
Pierce	New Hampshire	Nixon	California
Buchanan	Pennsylvania	Ford	Nebraska
Lincoln	Kentucky	Carter	Georgia
A. Johnson	North Carolina	Reagan	Illinois
Grant	Ohio	Bush (George H.W.)	Massachusetts
Hayes	Ohio	Clinton	Arkansas
Garfield	Ohio	Bush (George W.)	Texas
Arthur	Vermont	Obama	Hawaii
Cleveland	New Jersey	Trump	New York
		Biden	Pennsylvania

Which's most common?

Percents vs. Percentiles

- **Percent literally *means* "out of 100"**
  - 5% = 5 "per cent" = 5 "out of 100" = 5/100 = 0.05
  - Not all numbers are percents (or rates/ratios, etc.)
    - Don't move decimals *unless* you're adding/removing "percent"
    - An average of 12.4 years is not .124, or 12.4%, or .124%, or…
- **Percentile is the percent of cases that below a value**
  - Provides a measure of relative standing
  - Imposes 100-pt scale on original scores
  - Number of points separating each percentile not necessarily equal
    - Ordinal in sense of original measure
    - Interval only with respect to number of cases

Frequency Distribution of that Data

- Cases grouped by (nominal) values
- One unique value per row
- First column is the ***value labels***
  - Could differ from *values*
  - Could be 1 for Arkansas, 2 for California, etc.
  - Numbered *values* don't mean interval!
- Other cols give Frequency & Rel. Freq
  - ***Frequency*** = *count*
  - ***Relative frequency*** = *row count / total*
    - *If multiply it by 100, you'd get the percent*
    - *Rel. Freq. = Percent / 100*

Which's most common?

State	Freq.	RelFreq
Arkansas	1	0.02
California	1	0.02
Georgia	1	0.02
Hawaii	1	0.02
Illinois	1	0.02
Iowa	1	0.02
Kentucky	1	0.02
Massachusetts	4	0.09
Missouri	1	0.02
Nebraska	1	0.02
NewHampshire	1	0.02
NewJersey	1	0.02
NewYork	5	0.11
NorthCarolina	2	0.04
Ohio	7	0.16
Pennsylvania	2	0.04
SouthCarolina	1	0.02
Texas	3	0.07
Vermont	2	0.04
Virginia	8	0.18

Raw Interval Data: Homework Scores

These values have an order (higher/lower scores) *and* can be subtracted from each other

Student 1	75	Student 14	83
Student 2	68	Student 15	78
Student 3	91	Student 16	79
Student 4	83	Student 17	82
Student 5	85	Student 18	92
Student 6	98	Student 19	67
Student 7	66	Student 20	75
Student 8	82	Student 21	72
Student 9	78	Student 22	68
Student 10	89	Student 23	90
Student 11	88	Student 24	63
Student 12	90	Student 25	80
Student 13	85		

Use of a Frequency Distribution

- Consequences
  - Easier to extract information
    - Identify mode (OH)
    - Observe patterns (VA, MA, NY, TX)
  - Some information lost
    - Location of specific cases
    - Which president born where?

State	Freq.	RelFreq
Arkansas	1	0.02
California	1	0.02
Georgia	1	0.02
Hawaii	1	0.02
Illinois	1	0.02
Iowa	1	0.02
Kentucky	1	0.02
Massachusetts	4	0.09
Missouri	1	0.02
Nebraska	1	0.02
NewHampshire	1	0.02
NewJersey	1	0.02
NewYork	5	0.11
NorthCarolina	2	0.04
Ohio	7	0.16
Pennsylvania	2	0.04
SouthCarolina	1	0.02
Texas	3	0.07
Vermont	2	0.04
Virginia	8	0.18

Cumulative Raw Frequencies of that Data

- First two columns like before
  - Values
  - Frequencies
- Third is cumulative
  - Total in that or earlier rows
  - E.g. 5 is 1 + 1 + 1 + 2

Test Score	Freq.	Cum. Frequency
98	1	1
92	1	2
91	1	3
90	2	5
89	1	6
88	1	7
85	2	9
83	2	11
82	2	13
80	1	14
79	1	15
78	3	18
75	2	20
72	1	21
68	1	22
67	1	23
66	1	24
63	1	25

Cumulative *Relative* Frequencies

- Similar but *relative*
  - i.e. percents, not raw counts
  - % in that row or any earlier/higher row
- This is what's in SPSS
  - Cumulative Percentages
    - *not* Cum. Raw Frequencies
  - You'll use it, in labs & more!

Test Score	Freq.	Relative Freq.	Cum Rel Freq.
98	1	.04	.04
92	1	.04	.08
91	1	.04	.12
90	2	.08	.20
89	1	.04	.24
88	1	.04	.28
85	2	.08	.36
83	2	.08	.44
82	2	.08	.52
80	1	.04	.56
79	1	.04	.60
78	3	.12	.72
75	2	.08	.80
72	1	.04	.84
68	1	.04	.88
67	1	.04	.92
66	1	.04	.96
63	1	.04	1.00

15 SOC424 w/ Dr. Ellis Godard

That Data Grouped – Now *ordinal*!

Games Won

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	60-69	2	7.7	7.7	7.7
	70-79	12	46.2	46.2	53.8
	80-89	7	26.9	26.9	80.8
	90-99	4	15.4	15.4	96.2
	100+	1	3.8	3.8	100.0
	Total	26	100.0	100.0	

Is "70-79" a *lot* more wins than "60-69"? The difference is unclear.
It could be one game (from 69 to 70), or 19 games (from 60 to 79).
The intervals *between* values is not equal, precise, consistent, etc.
You can't subtract these values from each other. (70-79 minus 60-69 ain't a thing.)
The values are grouped ranges (don't say "intervals"), so this is ordinal.

18 SOC424 w/ Dr. Ellis Godard

Ordinal Example: Sports Ball

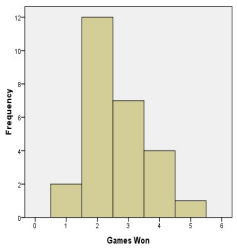
- # Games Won
- That's interval, as is
- But not if it's grouped…

-N	E	1	N.Y.	103	54
-N	E	2	Phi.	86	75
-N	E	3	St.L.	79	82
-N	E	4	Mon.	78	83
-N	E	5	Chi.	70	90
-N	E	6	Pit.	64	98
-N	W	1	Hou.	96	66
-N	W	2	Cin.	86	76
-N	W	3	S.F.	83	79
-N	W	4	S.D.	74	88
-N	W	5	L.A.	73	89
-N	W	6	Atl.	72	89
-A	E	1	Bos.	95	66
-A	E	2	N.Y.	90	72
-A	E	3	Det.	87	75
-A	E	4	Tor.	86	76
-A	E	5	Cle.	84	78
-A	E	6	Mil.	77	84
-A	E	7	Bal.	73	89
-A	W	1	Cal.	92	70
-A	W	2	Tex.	87	75
-A	W	3	K.C.	76	86
-A	W	4	Oak.	76	86
-A	W	5	Chi.	72	90
-A	W	6	Min.	71	91
-A	W	7	Sea.	67	95

16 SOC424 w/ Dr. Ellis Godard

Histogram

- Vertical bars, one per value
- Relative heights show relative freq's
  - Represents number (or %) of cases in that category
    - x-axis = categories
    - y-axis = (relative) freq in each category



19 SOC424 w/ Dr. Ellis Godard

Ungrouped Values – Still Interval

Games Won

Value	Freq	Value	Freq	Value	Freq
64.00	1	76.00	2	87.00	2
67.00	1	77.00	1	90.00	1
70.00	1	78.00	1	92.00	1
71.00	1	79.00	1	95.00	1
72.00	2	83.00	1	96.00	1
73.00	2	84.00	1	108.00	1
74.00	1	86.00	3		

26 (Total)

17 SOC424 w/ Dr. Ellis Godard

Ordinal Example

- Ranks w/ inconsistent/unequal differences

Variable Name & Label

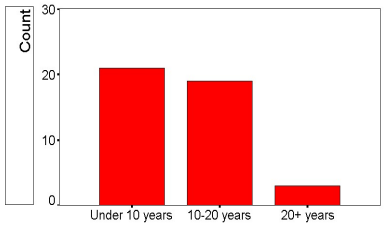
YEAR
How long the senator has been in congress

Value Label	Value	Freq.	Percent	Valid Percent	Cum Percent
Under ten years	1.00	21	48.8	48.8	48.8
Ten to twenty years	2.00	19	44.2	44.2	93.0
Twenty + years	3.00	3	7.0	7.0	100.0
Total		43	100.0	100.0	

Valid cases 43 Missing cases 0

20 SOC424 w/ Dr. Ellis Godard

Tend to be new, not longterm



How long the senator has been in congress

SPSS Processor is ready

21 SOC424 w/ Dr. Ellis Godard

For each senator in 1998, the % of their constituents that voted for Clinton in '92

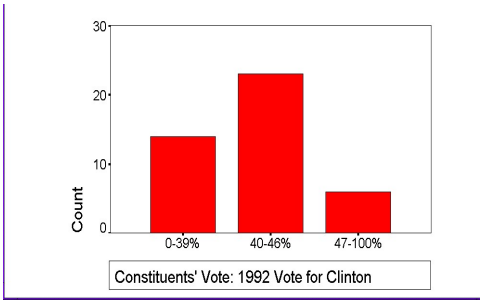
Value Label	Value	Freq.	Valid Percent	Cum Percent	Percent
0-39%	1.00	14	32.6	32.6	32.6
40-46%	2.00	23	53.5	53.5	86.0
47-100%	3.00	6	14.0	14.0	100.0
	Total	43	100.0	100.0	

Valid cases 43 Missing cases 0

-

22 SOC424 w/ Dr. Ellis Godard

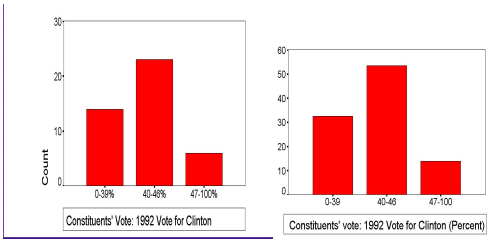
Tend to middle, not extremes



23 SOC424 w/ Dr. Ellis Godard

Relative Frequency Histogram

- **Y axis is relative freq. (not count)**
  - i.e. percentages on the left side
  - But same shape – dividing values by 100 doesn't change relative heights



24 SOC424 w/ Dr. Ellis Godard

Where's the histogram centered?

- **Central Tendency**
  - Where's the "middle"?
    - Does it have 1 (unimodal)? 2 (bimodal)? more (multi-modal)?
  - Where are values *concentrated?*
  - Which attribute do cases tend have, more often than the other attributes?
- **That's 2 lectures away**
- **Then Dispersion –**
  - How far values *differ* from C.T.
  - Ranges, Deviations, Standard Deviation
- **But first…**

25 SOC424 w/ Dr. Ellis Godard

What is the *shape of a* histogram?

- **Is it "bell-shaped"?**
  - Concentrated in middle, & symmetric
- ***Not* the same as "normal"**
  - And don't say "sorta normal"
  - Like "pregnant" or "dead", it either *is* or *isn't* – and it **isn't**! Ever.
  - *No* data output, in SPSS or anywhere else, will *ever* be normal. In this class or in your life. Ever.
  - Again: Do not ever say that any distribution is normal. It won't be, your answer will be wrong, and I will take off *extra* points. This matters.

26 SOC424 w/ Dr. Ellis Godard

Pre Reduction Parts Histograms Shape Next

Seriously. Reality is *never* Normal!

- Some people & disciplines, *pretend* normality happens in nature

The bell-shaped curve is a common feature of nature and psychology

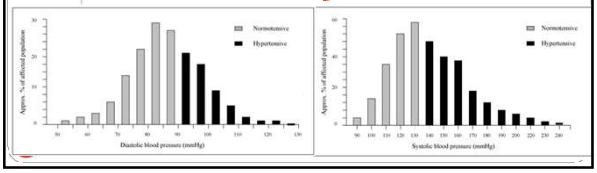
http://www.simplypsychology.org/normal-distribution.html

The normal distribution is the most important probability distribution in statistics because many continuous data in nature and psychology displays this bell-shaped curve when compiled and graphed.

For example, if we randomly sampled 100 individuals we would expect to see a normal distribution frequency curve for many continuous variables, such as IQ, height, weight and blood pressure.

but "normal" and "bell-shaped" are NOT THE SAME THING! ☹

But those *are't* normal –bell-shaped, but skewed so *not* normal…



Pre Reduction Parts Histograms Shape Next

Shape is more than 1 word ☺

- ***Could* be measured w/ 2 statistics *we won't use***
  - Skewness?
    - Skewed if a "straggly tail" to one side
    - Left or right? (which direction is the tail?)
    - How much? More skewed -> higher "skewness" stat
  - Kurtosis?
    - Tall and skinny, vs. short and squat
    - Higher kurtosis = more concentrated (tall/skinny; "peaked")
- ***Could* be assessed w/ unclear ideas**, e.g. Outliers
  - Extreme values, far removed from the rest
  - *Several* ways to measure
    - You'll measure it *two different* ways
    - Read instructions carefully!

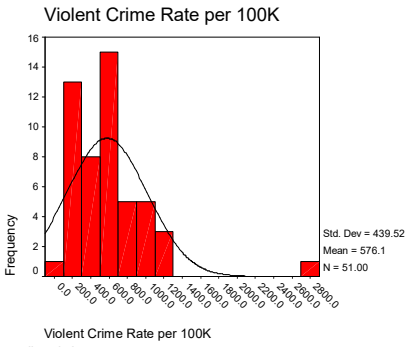
Pre Reduction Parts Histograms Shape Next

What is the *shape of a* histogram?

- **NOT. NORMAL.** *Anything* else, maybe.
- ***Maybe* bell-shaped? But probably skewed**
- **Could be something else**
  - U-shaped (upside-down bell, highest on ends)
  - S-curve, sine-wave, etc.
  - Uniform? (flat/equal distribution among values)
- **Could be nothing in particular**
  - No general shape? (no clear pattern, "noisy")

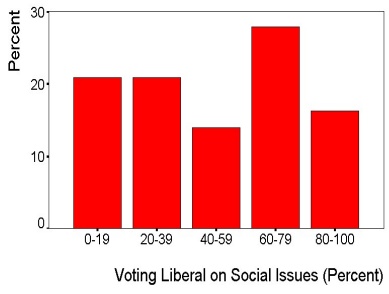
Pre Reduction Parts Histograms Shape Next

No "bell", but skew/outliers

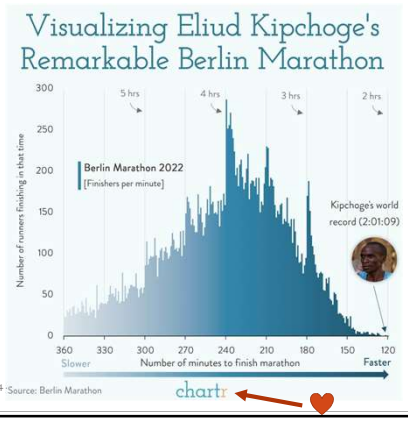


Pre Reduction Parts Histograms Shape Next

An example with *no* particular shape



Pre Reduction Parts Histograms Shape Next



Limits on "Shapes" Search

- **Interpretation matters**
  - Not obvious – judgment call
    - Can be wrong, but may not be one "right"
  - Can/should emphasize more than 1 element
    - *Never* "normal"
    - Not just bell – how close to normal? Any skew?
- **Ordinals & Intervals only**
  - For nominal, order is arbitrary
    - Numbers don't represent any meaningful order
    - No "left" or "right"; no "positive" or "negative"; no "sides"
  - No "shape" in the sense that ordinal or intervals have
    - Can say "concentrated"?
    - Can say more or less evenly distributed?

Limits on Use of "Shape"

- **Problem w/ graphs**
  - Can't easily compare two in a consistent way
  - Little flatter, little more to right
  - What do those mean?
- **Numerical descriptive methods help**
  - Univariate descriptive statistics
  - Start that next lecture

Data Displays in SPSS

- **Frequency distributions**
  - Just text tables
- **Other common data displays**
  - Stem and Leaf plots – also in text, but limited use
  - Histograms = key! Get for any variable!
  - Bar charts – avoid! (confusing & misleading)
  - Pie charts = silly (see Wired article)
  - Scatter plots – later
- **Other output in this class** (all near the end)
  - Cross-tabulations
  - Output from four statistical tests
    - chi-square, t-tests, ANOVA, and regression

Preview of next three lectures

- **Other ways of presenting data**
  - Conventional options
  - Innovative options
  - Problems/concerns
  - Intro to SPSS (!!)
- **Measures of Central Tendency**
  - What's the typical value of a variable?
  - Where on a variable is a sample centered?
- **Measures of Dispersion**
  - How are the values distributed across the values?
  - What is their range, & how concentrated are they?

For Your Next Lab…

- **See 3 questions in problem 1.7 in text** (p.8 4th edition, 9? in 5th?)
  - Follow the instructions there to use a website to access GSS data **BUT**….
    - Choose "New SDA 3.5" (with weights) – might be updated again ☹
  1. For (a), "one of the yes responses" means *both* (don't pick one; sum them!)
  2. For (b), use 1991 (just 1991, not 2008)
     - Otherwise, the answers are in the back of the book ☺
  3. For (c), compare to 1991 (not 2008)
- **3 more questions:** What kind of measures are these:
  4. Direct or indirect?
  5. Categorical or quantitative?
  6. Nominal, ordinal, or interval?
- **Cautions:**
  - **This lab, like many, has *TWO* PARTS** (3-part text prob, & those "3 more questions")
  - Do ALL the lab for credit. I don't have a way to give partial credit for doing *some* of a lab.
  - **Remember to designate a Secretary at submission!!**