Quantitative Sociological Analysis

Data, Descriptive Statistics, and Central Tendency

Part 4

February 6, 2025

In-Class Group Exercises

Data Collection

I've turned the concepts y'all operationalized into variables and designed a survey to generate a dataset we can use for future in-class exercises.

Many of the concepts are somewhat sensitive in nature, and survey respondents' (your) identity should be protected. Unfortunately, the class is too small to ensure anonymity if you submit responses as yourselves.

Thus, complete the survey as a hypothetical respondent.

https://uky.az1.qualtrics.com/jfe/form/SV_bIx3tfhTD55tMI6

Science: a **process** of organizing, and acquiring new, knowledge

Steps in the process

1. Start with a perspective
2. Select a theory
3. Derive a research proposition
4. Derive a research question
5. Derive a hypothesis
6. Find or collect data
7. Analyze data
8. Report results & Answer question
9. Interpret results in terms of theory
10. Draw implications for theory

How does all this lead up to the data?

Without knowing above, all this is worse than lies or even damn lies

Part 4

Learning objective: begin to understand why perspective, theory, proposition, question, hypothesis, data, and methods are ideally intricately interwoven, opposed to loosely interdependent steps, within the scientific process

recognize how:

earlier steps in the scientific process determine data requirements

methods are tools we use to help make sense of the data

level of measurement determines which methods may be appropriate

Takeaway: descriptive statistics are foundational methods to begin making sense of data in important ways, which will be useful later for determining whether the data are appropriate for addressing the research question

What are data?

- Recall: scientific knowledge is the accumulation of theories recognized to provide an understanding of natural or social phenomena
  - observable events/occurrences in the natural or social world

- Data represents phenomena that have been systematically documented
  - information intended to reflect a representation of reality
    - the bridge between phenomena and our ability to study and understand them
      - the material from which empirical evidence is constructed

Two basic types of data

- Qualitative data: information that is stored and understood descriptively

- Quantitative data: information that is stored and understood numerically

Feature	Qualitative Data	Quantitative Data
Definition	Descriptive, non-numerical information.	Numerical, measurable data.
Purpose	Explores meanings, concepts, and experiences.	Measures quantities, amounts, or frequencies.
Examples	Interview transcripts, open-ended survey responses, photos, social media posts.	Age, income, test scores, height, temperature.
Data Type	Words, images, symbols, or categories.	Numbers and statistical values.
Common Methods	Interviews, focus groups, observations, content analysis.	Surveys, experiments, census data, statistical analysis.
Analysis	Thematic coding, content analysis, pattern identification.	Statistical tests, mathematical modeling, data visualization.
Outcome	Provides depth, context, and insight.	Produces numerical trends and comparisons.

What data are needed?

- Perspective: data must align with assumed reality
  - reflects how defined portion of the empirical world is understood
- Theory: data must contain key elements
  - includes information required to explain the phenomena of interest
- Research question: data must match specific reframing
  - documented within, or aggregated to, respective unit of analysis
  - recorded in appropriate measurable terms

Unit of analysis

- the main entity being studied
  - what or who is being analyzed to answer a research question

Unit of Analysis	Definition	Example Research Question
Individuals	Single persons being studied	How does social media usage affect mental health?
Groups	Small or large social groups	How do book clubs vary in their discussion styles and group dynamics?
Organizations	Formal entities such as businesses, schools, or governments	How does leadership structure impact the financial stability of Fortune 500 companies?
Communities	Geographic or social communities	How does neighborhood poverty relate to crime rates?
Events	Specific occurrences over time	How do protests influence government policy changes?
Texts/Media	Written, spoken, or visual communication	How are women represented in political campaign speeches?
Interactions	Social exchanges between individuals or groups	How do doctors and patients communicate about chronic illness?
Artifacts/Objects	Material culture, objects, or technology	How have smartphone designs evolved over time?

Recall….

- Ecological Fallacy: individual-level claims drawn from group-level
- Individualistic Fallacy: group-level claims drawn from individual-level

Measurable terms

- Recall: concepts in the research question must be operationalized,
  - process of defining concepts into measurable terms
- because this enables a concept to become a variable
  - representation of a characteristic that can take different values
- Thus, a variable is the numeric representation of a characteristic that can vary across the units of analysis
  - the basic building blocks of quantitative data

Direct vs indirect measurement

- many elements we, as social scientists, are concerned with conceptualizing, operationalizing, and turning into variables are not directly measurable
  - abstract/latent concept inferred by indicator or proxy variable(s)

Latent Concept	Indicator Variables (Observed Measures)
Happiness	Self-reported life satisfaction, frequency of smiling, stress levels
Social Capital	Number of friendships, trust in institutions, community participation
Political Ideology	Voting behavior, policy preferences, party affiliation
Intelligence	IQ test scores, problem-solving ability, memory recall

- Reliability: the consistency of a measure
  - ask same person again and again and get same response

- Validity: the accuracy of a measurement
  - variable measures what it is intended to

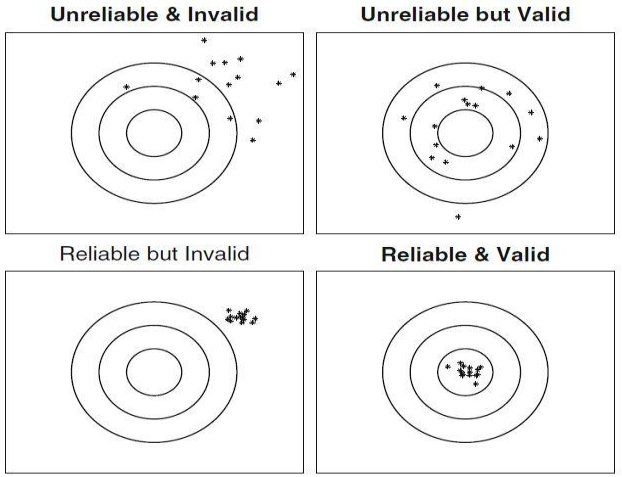
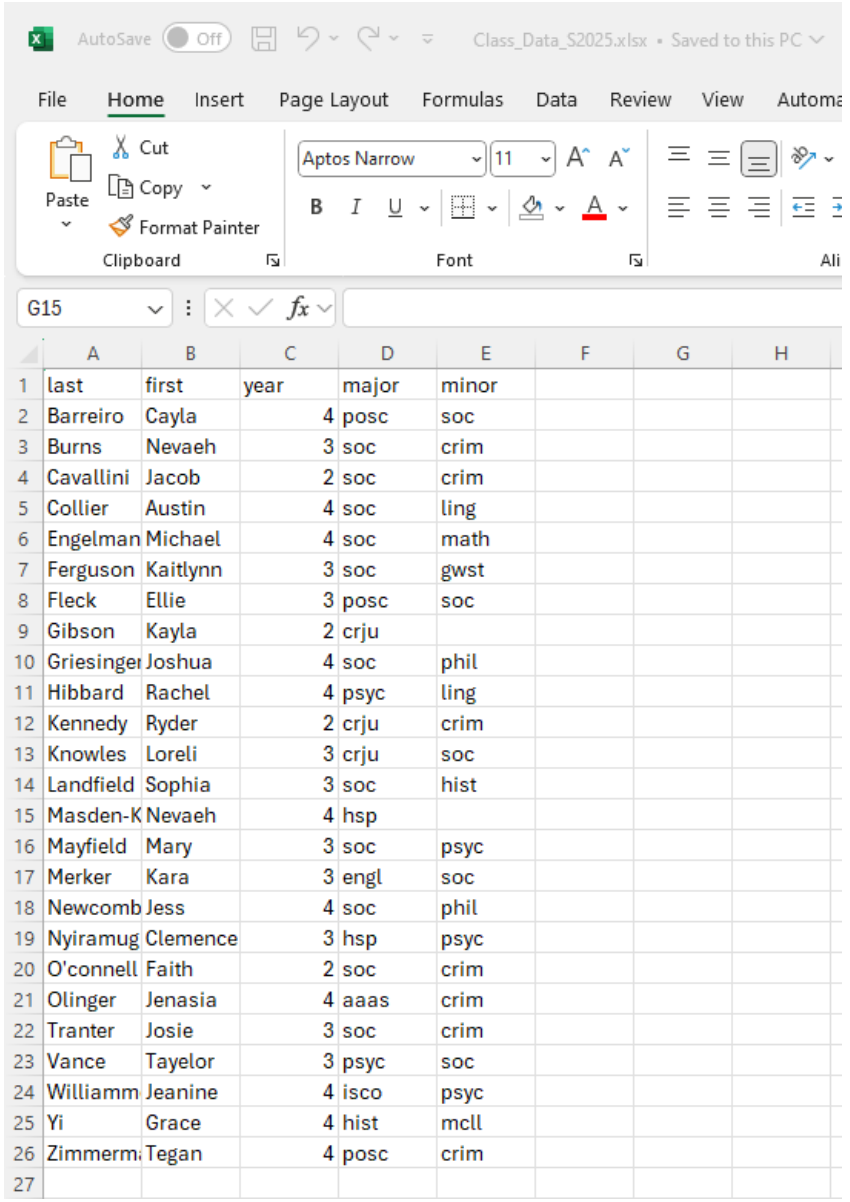


Fig. 3.1. Illustration of concepts of reliability and validity of measurement

Data frame

- a structured set of data organized into rows and columns
  - common format used to store and analyze data
- units of analysis often demarcated by rows
- variables often demarcated by columns
  - not always, but let's avoid unnecessary complexity



	A	B	C	D	E	F	G	H
1	last	first	year	major	minor			
2	Barreiro	Cayla	4	posc	soc			
3	Burns	Nevaeh	3	soc	crim			
4	Cavallini	Jacob	2	soc	crim			
5	Collier	Austin	4	soc	ling			
6	Engelman	Michael	4	soc	math			
7	Ferguson	Kaitlynn	3	soc	gwst			
8	Fleck	Ellie	3	posc	soc			
9	Gibson	Kayla	2	crju				
10	Griesinger	Joshua	4	soc	phil			
11	Hibbard	Rachel	4	psyc	ling			
12	Kennedy	Ryder	2	crju	crim			
13	Knowles	Loreli	3	crju	soc			
14	Landfield	Sophia	3	soc	hist			
15	Masden-K	Nevaeh	4	hsp				
16	Mayfield	Mary	3	soc	psyc			
17	Merker	Kara	3	engl	soc			
18	Newcomb	Jess	4	soc	phil			
19	Nyiramug	Clemence	3	hsp	psyc			
20	O'connell	Faith	2	soc	crim			
21	Olinger	Jenasia	4	aaas	crim			
22	Tranter	Josie	3	soc	crim			
23	Vance	Tayelor	3	psyc	soc			
24	Williamm	Jeanine	4	isco	psyc			
25	Yi	Grace	4	hist	mcll			
26	Zimmerm	Tegan	4	posc	crim			
27								

Levels of measurement

- a variable takes different values that represent a characteristic of the phenomenon it was intended to reflect
  - differs by level of measurement

- nominal and ordinal variables are inherently qualitative
  - categories assigned numeric values

0 = male, 1 = female

0 = <HS, 1 = HS, 2 = >HS

Level of Measurement	Definition	Key Characteristics	Examples
Nominal	Categorizes data without a meaningful order.	- Categories are **mutually exclusive** and **unordered.** - No mathematical operations (other than counting frequency).	Gender (Male, Female, Nonbinary), Political Party (Democrat, Republican), Eye Color (Blue, Brown, Green)
Ordinal	Categorizes data with a meaningful order but without equal intervals.	- **Ranked categories** (higher/lower has meaning). - Differences between ranks are **not necessarily equal.**	Education Level (High School, Bachelor's, Master's, PhD), Satisfaction Rating (Satisfied, Neutral, Dissatisfied), Military Ranks (Private, Sergeant, Captain)
Interval	Numeric data with equal intervals but **no true zero.**	- Can add/subtract values. - No meaningful ratio (e.g., 40°F is not "twice as warm" as 20°F).	Temperature in Fahrenheit/Celsius, IQ Scores, SAT Scores
Ratio	Numeric data with equal intervals and a **true zero.**	- Can multiply/divide values (e.g., 10 lbs is **twice** as heavy as 5 lbs). - True zero represents an absence of the quantity.	Height, Weight, Age, Income, Number of Children

Quantitative analysis

- an attempt to make sense of numerical data
  - includes many different methods
    - consider like tools

- foundational step involves summarizing data, so they're easily understood
  - variables reduced to one or a few informative quantities

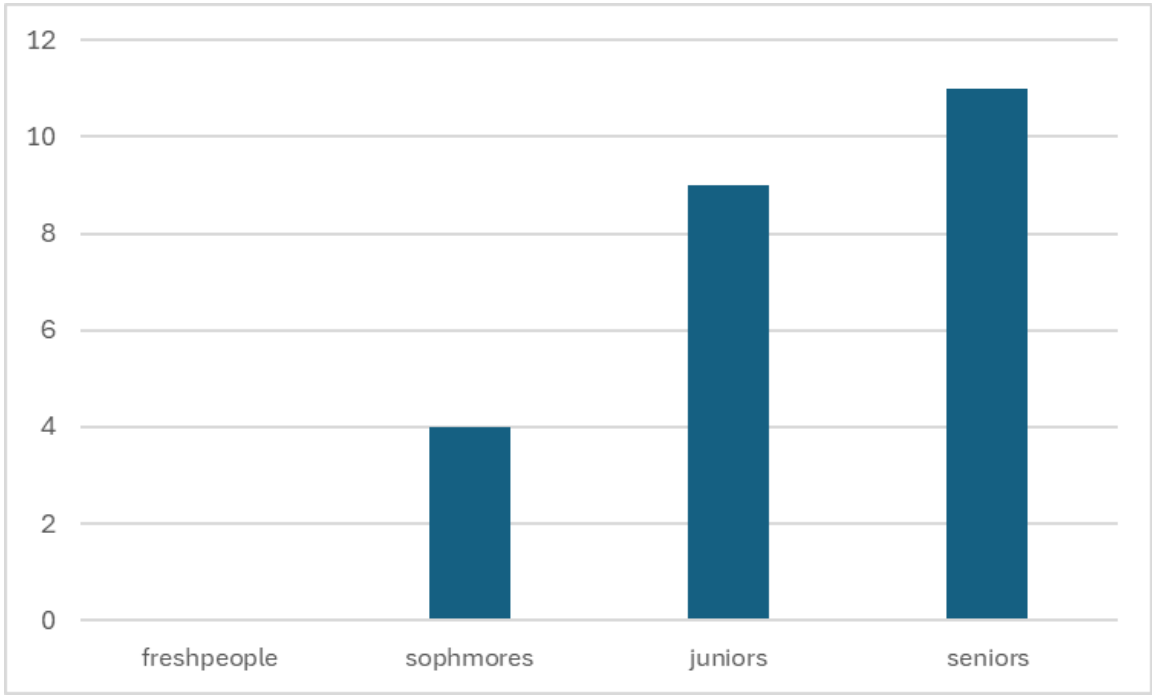
- Descriptive statistics: methods used to summarize data in meaningful ways
  - determined by the variable(s) level(s) of measurement
    - univariate methods: one variable
    - bivariate methods: two variables
    - multivariate methods: more than two variables

Frequency distribution

- count of how often a variable's unique values appear in the data frame
  - this is the most basic statistic, but it is helpful for organizing data

	A	B	C	D	E	F
1	last	first	year	count	major	minor
2	Cavallini	Jacob	2	1	soc	crim
3	Gibson	Kayla	2	2	crju	
4	Kennedy	Ryder	2	3	crju	crim
5	O'connell	Faith	2	4	soc	crim
6	Burns	Nevaeh	3	1	soc	crim
7	Ferguson	Kaitlynn	3	2	soc	gwst
8	Knowles	Loreli	3	3	crju	soc
9	Landfield	Sophia	3	4	soc	hist
10	Mayfield	Mary	3	5	soc	psyc
11	Merker	Kara	3	6	engl	soc
12	Nyiramug	Clemence	3	7	hsp	psyc
13	Tranter	Josie	3	8	soc	crim
14	Vance	Tayelor	3	9	psyc	soc
15	Barreiro	Cayla	4	1	posc	soc
16	Collier	Austin	4	2	soc	ling
17	Engelman	Michael	4	3	soc	math
18	Griesinger	Joshua	4	4	soc	phil
19	Hibbard	Rachel	4	5	psyc	ling
20	Masden-K	Nevaeh	4	6	hsp	
21	Newcomb	Jess	4	7	soc	phil
22	Olinger	Jenasia	4	8	aaas	crim
23	Williamm	Jeanine	4	9	isco	psyc
24	Yi	Grace	4	10	hist	mcll
25	Zimmerm	Tegan	4	11	posc	crim

- bar charts can be useful



Frequency proportions or percentages

$$\text{proportion} = \left(\frac{f}{N}\right)$$

$$\text{percentage} = \left(\frac{f}{N}\right) * 100$$

f = frequency, or count of unique values observed

N = the number of all observations

	A	B	C	D	E	F
1	last	first	year	count	major	minor
2	Cavallini	Jacob	2	1	soc	crim
3	Gibson	Kayla	2	2	crju	
4	Kennedy	Ryder	2	3	crju	crim
5	O'connell	Faith	2	4	soc	crim
6	Burns	Nevaeh	3	1	soc	crim
7	Ferguson	Kaitlynn	3	2	soc	gwst
8	Knowles	Loreli	3	3	crju	soc
9	Landfield	Sophia	3	4	soc	hist
10	Mayfield	Mary	3	5	soc	psyc
11	Merker	Kara	3	6	engl	soc
12	Nyiramug	Clemence	3	7	hsp	psyc
13	Tranter	Josie	3	8	soc	crim
14	Vance	Tayelor	3	9	psyc	soc
15	Barreiro	Cayla	4	1	posc	soc
16	Collier	Austin	4	2	soc	ling
17	Engelman	Michael	4	3	soc	math
18	Griesinger	Joshua	4	4	soc	phil
19	Hibbard	Rachel	4	5	psyc	ling
20	Masden-K	Nevaeh	4	6	hsp	
21	Newcomb	Jess	4	7	soc	phil
22	Olinger	Jenasia	4	8	aaas	crim
23	Williamm	Jeanine	4	9	isco	psyc
24	Yi	Grace	4	10	hist	mcll
25	Zimmerm	Tegan	4	11	posc	crim

- frequency tables are common

Year in college	Frequency	Proportion	Percent
freshpeople	0	0.00	0.00
sophmores	4	0.17	16.67
juniors	9	0.38	37.50
seniors	11	0.46	45.83
Total	24	1	100

Cumulative frequency proportions or percentages

sometimes useful to show how many cases fall at or below a given classification

	A	B	C	D	E	F
1	last	first	year	count	major	minor
2	Cavallini	Jacob	2	1	soc	crim
3	Gibson	Kayla	2	2	crju	
4	Kennedy	Ryder	2	3	crju	crim
5	O'connell	Faith	2	4	soc	crim
6	Burns	Nevaeh	3	1	soc	crim
7	Ferguson	Kaitlynn	3	2	soc	gwst
8	Knowles	Loreli	3	3	crju	soc
9	Landfield	Sophia	3	4	soc	hist
10	Mayfield	Mary	3	5	soc	psyc
11	Merker	Kara	3	6	engl	soc
12	Nyiramug	Clemence	3	7	hsp	psyc
13	Tranter	Josie	3	8	soc	crim
14	Vance	Tayelor	3	9	psyc	soc
15	Barreiro	Cayla	4	1	posc	soc
16	Collier	Austin	4	2	soc	ling
17	Engelman	Michael	4	3	soc	math
18	Griesinger	Joshua	4	4	soc	phil
19	Hibbard	Rachel	4	5	psyc	ling
20	Masden-K	Nevaeh	4	6	hsp	
21	Newcomb	Jess	4	7	soc	phil
22	Olinger	Jenasia	4	8	aaas	crim
23	Williamm	Jeanine	4	9	isco	psyc
24	Yi	Grace	4	10	hist	mcll
25	Zimmerm	Tegan	4	11	posc	crim

- frequency tables should include features that are most useful for making sense of the data

Year in college	Frequency	Cum. Freq.	Proportion	Cum. Prop.	Percent	Cum. Perc.
freshpeople	0	0	0.00	0.00	0.00	0.00
sophmores	4	4	0.17	0.17	16.67	16.67
juniors	9	13	0.38	0.54	37.50	54.17
seniors	11	24	0.46	1.00	45.83	100
Total	24		1		100	

Summarizing nominal and ordinal variables

generally limited in how they can be meaningfully summarized

- ratios: compare the relative sizes of categories

	A	B	C	D	E	F	G
1	last	first	year	count	major	major code	minor
2	Olinger	Jenasia	4	8	aaas	1	crim
3	Merker	Kara	3	6	engl	2	soc
4	Yi	Grace	4	10	hist	3	mcll
5	Williamm	Jeanine	4	9	isco	4	psyc
6	Barreiro	Cayla	4	1	posc	5	soc
7	Zimmerm	Tegan	4	11	posc	5	crim
8	Nyiramug	Clemence	3	7	hsp	6	psyc
9	Masden-K	Nevaeh	4	6	hsp	6	
10	Vance	Tayelor	3	9	psyc	7	soc
11	Hibbard	Rachel	4	5	psyc	7	ling
12	Gibson	Kayla	2	2	crju	8	
13	Kennedy	Ryder	2	3	crju	8	crim
14	Knowles	Loreli	3	3	crju	8	soc
15	Cavallini	Jacob	2	1	soc	9	crim
16	O'connell	Faith	2	4	soc	9	crim
17	Burns	Nevaeh	3	1	soc	9	crim
18	Ferguson	Kaitlynn	3	2	soc	9	gwst
19	Landfield	Sophia	3	4	soc	9	hist
20	Mayfield	Mary	3	5	soc	9	psyc
21	Tranter	Josie	3	8	soc	9	crim
22	Collier	Austin	4	2	soc	9	ling
23	Engelman	Michael	4	3	soc	9	math
24	Griesinger	Joshua	4	4	soc	9	phil
25	Newcomb	Jess	4	7	soc	9	phil

- one to total ratio = $\frac{\text{Frequency of One Category}}{\text{Total Count of All Categories}}$

aaas	$\frac{1}{24} = 0.042$	0.042
engl	$\frac{1}{24} = 0.042$	0.126
hist	$\frac{1}{24} = 0.042$	0.168
isco	$\frac{1}{24} = 0.042$	0.251
posc	$\frac{2}{24} = 0.083$	0.334
hsp	$\frac{2}{24} = 0.083$	0.417
crju	$\frac{3}{24} = 0.125$	0.542
soc	$\frac{11}{24} = 0.458$	1.000

Summarizing nominal and ordinal variables

sometimes useful to use

- central tendency: methods to describe the center value or typical case

	A	B	C	D	E
1	last	first	year	major	minor
2	Gibson	Kayla	2	crju	
3	Kennedy	Ryder	2	crju	crim
4	Cavallini	Jacob	2	soc	crim
5	O'connell	Faith	2	soc	crim
6	Merker	Kara	3	engl	soc
7	Nyiramug	Clemence	3	hsp	psyc
8	Vance	Tayelor	3	psyc	soc
9	Knowles	Loreli	3	crju	soc
10	Burns	Nevaeh	3	soc	crim
11	Ferguson	Kaitlynn	3	soc	gwst
12	Landfield	Sophia	3	soc	hist
13	Mayfield	Mary	3	soc	psyc
14	Tranter	Josie	3	soc	crim
15	Olinger	Jenasia	4	aaas	crim
16	Yi	Grace	4	hist	mcll
17	Williamm	Jeanine	4	isco	psyc
18	Barreiro	Cayla	4	posc	soc
19	Zimmerm	Tegan	4	posc	crim
20	Masden-K	Nevaeh	4	hsp	
21	Hibbard	Rachel	4	psyc	ling
22	Collier	Austin	4	soc	ling
23	Engelman	Michael	4	soc	math
24	Griesinger	Joshua	4	soc	phil
25	Newcomb	Jess	4	soc	phil

- mode: most commonly-occurring value

  = 4 (senior)

- median: the middle of the distribution when all cases are sorted by rank order

  = 3 (sophomore)

  - 50% of cases have lower and 50% of cases have higher

- mean: the average value

  = 3.29 (not exactly meaningful)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$x\ bar$ (mean) = $\sum_{i=1}^{n}$ (summation of) x_i (all values of the variable indexed by each case), divided by number of cases (n)

Frequency distribution: RStudio

- like the year in college example from our class data in Excel
  - let's explore educational attainment in the GSS data

- frequency tables

```
53  # let's run the table command again but for edu_cat
54  table(GSS$educ_cat)
```

```
  <HS              HS    some college              BA Graduate Degree
13925           19307           15364           10840            5119
```

```
60  # table proportions
61  prop.table(table(GSS$educ_cat))
```

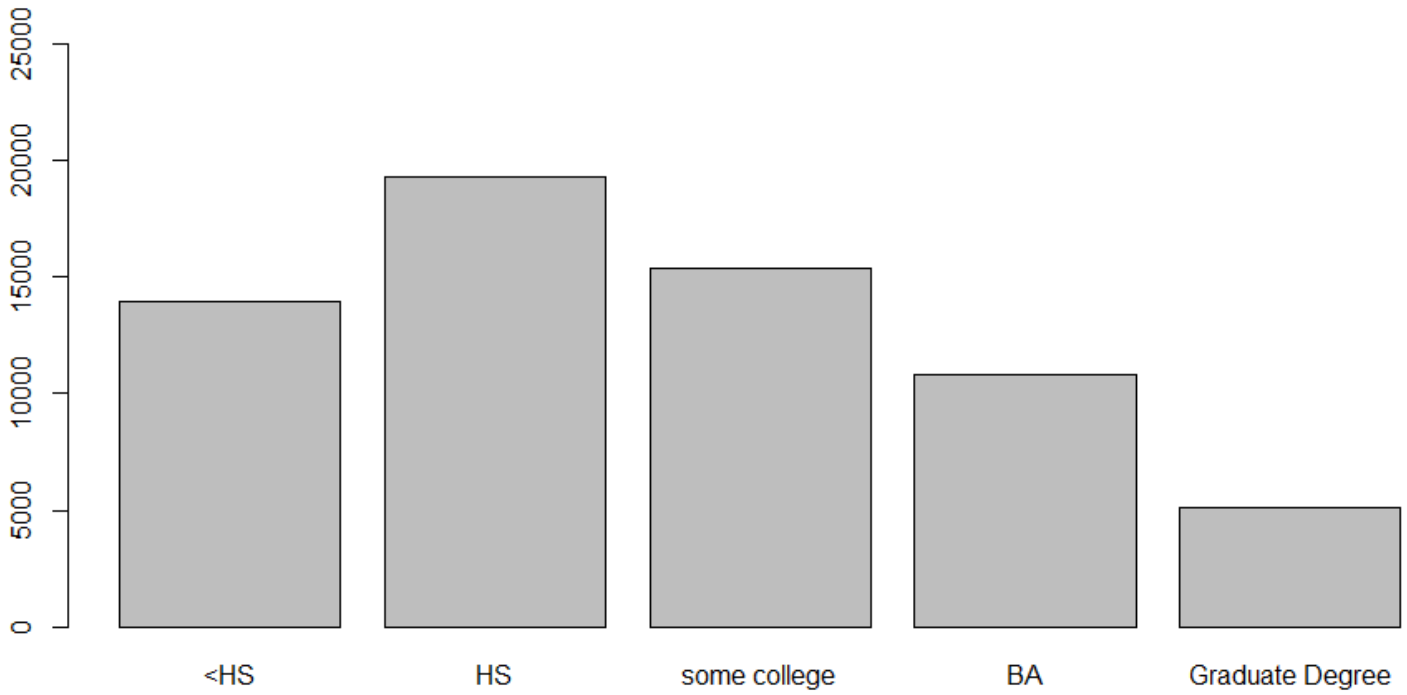
```
      <HS              HS    some college              BA Graduate Degree
0.21570754      0.29907831      0.23799861      0.16791883      0.07929672
```

```
63  # table percentages by converting proportions
64  prop.table(table(GSS$educ_cat))*100
```

```
      <HS              HS    some college              BA Graduate Degree
21.570754       29.907831       23.799861       16.791883        7.929672
```

- bar charts can be useful

```
72  # let's change the y axis dimensions to better fit the data
73  barplot(table(GSS$educ_cat),ylim=c(0,25000))
```



Summarizing nominal and ordinal variables: RStudio

- mode: no built-in function, command
  - that's okay, because this is just the most frequent value
    - or values if more than one mode

- mean: will not compute for non-numeric variables
  - knows this does not make very much sense
    - What is 1.596 edu_deg?

```
75  # means
76  mean(GSS$educ_cat)
Warning message:
In mean.default(GSS$educ_cat) :
  argument is not numeric or logical: returning NA
```

```
80  #let's use the original edu_deg variable that has numeric values
81  mean(GSS$educ_deg)
```

- median: will not compute for non-numeric variables
  - makes more sense, but still same issue

= 1

```
83  # median: same issue as above
84  median(GSS$educ_deg)
```

Summarizing interval-ratio variables: RStudio

- frequency tables are often not ideal for variables with many different values

```
86  ### let's work with an interval-ratio variable, age in years
87  table(GSS$age)
```

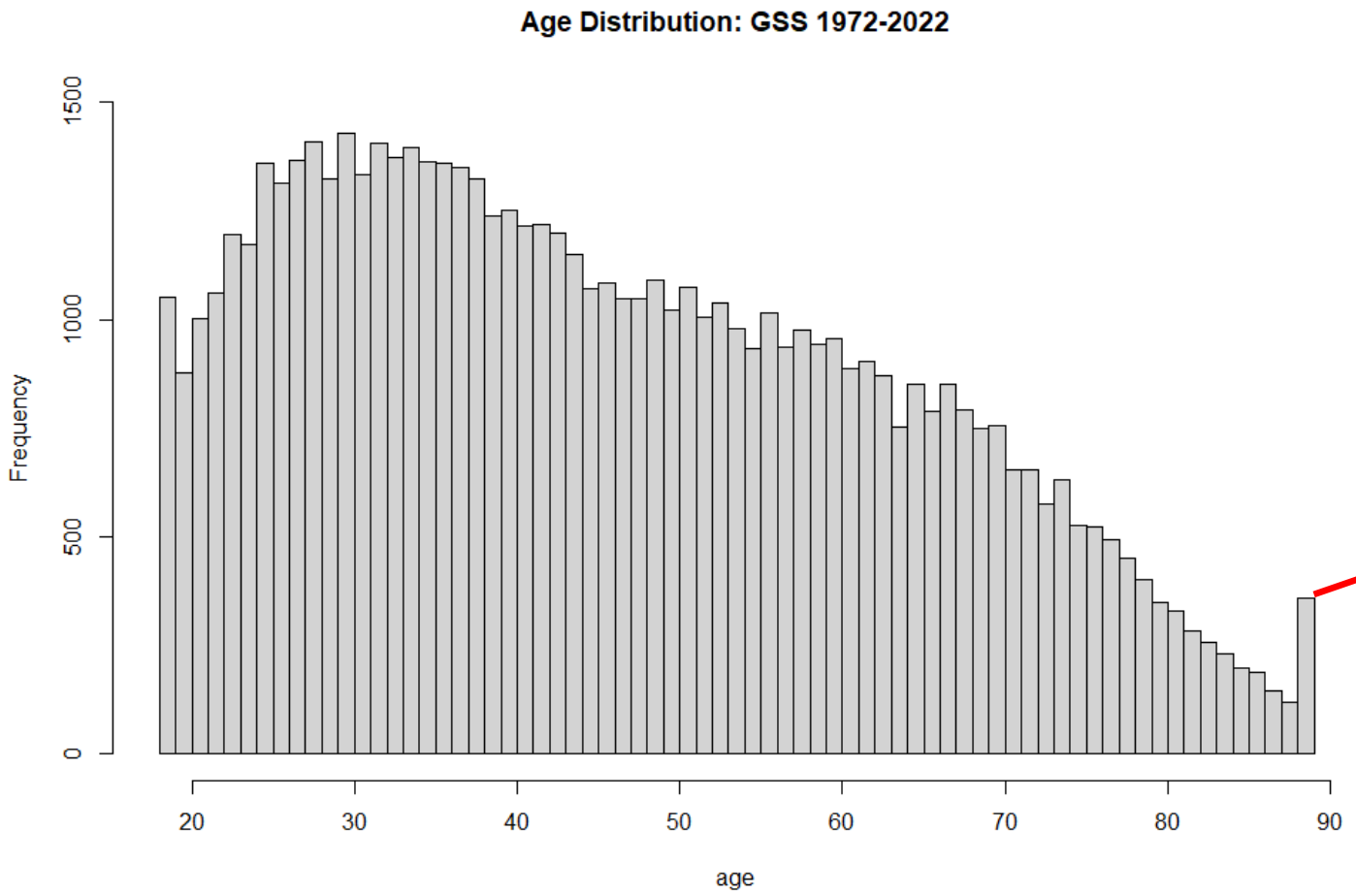
```
 18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39
230  823  878 1003 1063 1196 1174 1361 1315 1368 1410 1323 1429 1334 1407 1374 1395 1363 1359 1350 1324 1238
 40   41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60   61
1251 1216 1220 1201 1149 1072 1084 1050 1049 1091 1021 1074 1005 1037  979  934 1017  938  975  945  955  889
 62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81   82   83
903  870  752  850  788  852  792  750  756  655  654  576  633  528  523  492  450  400  349  331  282  256
 84   85   86   87   88   89
232  198  189  147  119  359
```

```
89  prop.table(table(GSS$age))
```

```
         18          19          20          21          22          23          24          25          26
0.003562853 0.012748819 0.013600806 0.015537139 0.016466579 0.018526838 0.018186043 0.021082798 0.020370227
         27          28          29          30          31          32          33          34          35
0.021191232 0.021841840 0.020494152 0.022136163 0.020664550 0.021795368 0.021284176 0.021609480 0.021113779
         36          37          38          39          40          41          42          43          44
0.021051816 0.020912400 0.020509643 0.019177446 0.019378824 0.018836651 0.018898614 0.018604291 0.017798776
         45          46          47          48          49          50          51          52          53
0.016605995 0.016791883 0.016265200 0.016249710 0.016900318 0.015815971 0.016636976 0.015568120 0.016063822
         54          55          56          57          58          59          60          61          62
0.015165363 0.014468283 0.015754008 0.014530246 0.015103400 0.014638680 0.014793587 0.013771203 0.013988072
         63          64          65          66          67          68          69          70          71
0.013476880 0.011648981 0.013167067 0.012206645 0.013198048 0.012268608 0.011618000 0.011710944 0.010146387
         72          73          74          75          76          77          78          79          80
0.010130896 0.008922624 0.009805592 0.008179072 0.008101619 0.007621408 0.006970800 0.006196267 0.005406243
         81          82          83          84          85          86          87          88          89
0.005127411 0.004368368 0.003965611 0.003593835 0.003067152 0.002927736 0.002277128 0.001843389 0.005561149
```

Summarizing interval-ratio variables: RStudio

- histograms: no gap between bars to indicate *continuous* flow of data
  - gaps in bar charts indicate distinct categories




```
94  # let's customize our histogram
95  hist((GSS$age),ylim=c(0,1500),breaks=72,
96  main="Age Distribution: GSS 1972-2022",xlab="age")
```

Aside from developing a basic understanding of your data, descriptive statistics are useful for identifying abnormalities.

Any guesses why this lumping at age 89?

Hint: the GSS started surveying Americans over fifty years ago.

Summarizing interval-ratio variables: RStudio

- measures of central tendency tend to make a lot more sense

- mode age = 29 years old
- mean age = 46.45 years old
- median age = 44 years old

see how these measures of central tendency for the same variable can each tell something different about the data

we will make more sense of this next week when we learn about dispersion