Quantitative Sociological Analysis

Descriptive Statistics
Dispersion

Part 4

February 11, 2025

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

Summarizing interval-ratio variables

this is where we left off last week

- 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

The spread around the center value also contains useful information that can be used to help make sense of the data

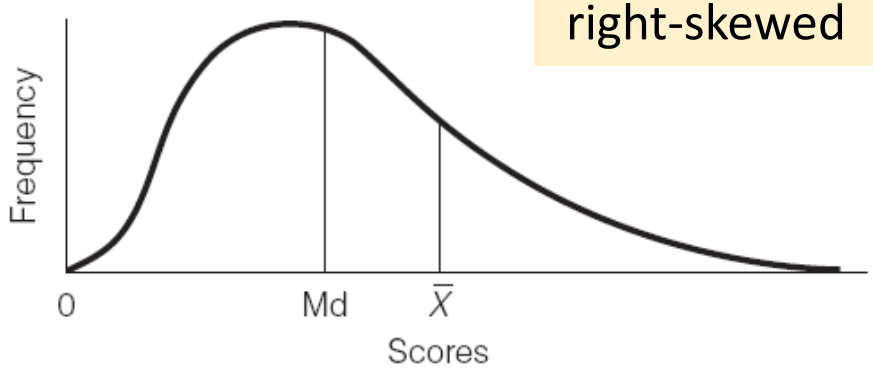
Note how the mean is slightly greater than the median in this example…

Summarizing interval-ratio variables

```
# Note: this example code is to produce output for learning outcome
# let's overlay a density curve to the age histogram from above to consider the skew
#same plot as above, but switch y axis from frequency count to probability
hist(GSS$age),probability=TRUE,ylim=c(0,0.025),breaks=72,
  main="Age Distribution: GSS 1972-2022",xlab="age")
# specifying normal density curve to fit the age distribution
x<-seq(min(GSS$age),max(GSS$age),length=100)
y<-dnorm(x,mean=mean(GSS$age),sd=sd(GSS$age))
# add the curve to the histogram
lines(x,y,lwd=2)
# add mean and median lines to consider skew in relation to central tendency
abline(v = mean(GSS$age),lwd=2)
abline(v = median(GSS$age),lwd=2,lty=2)
```

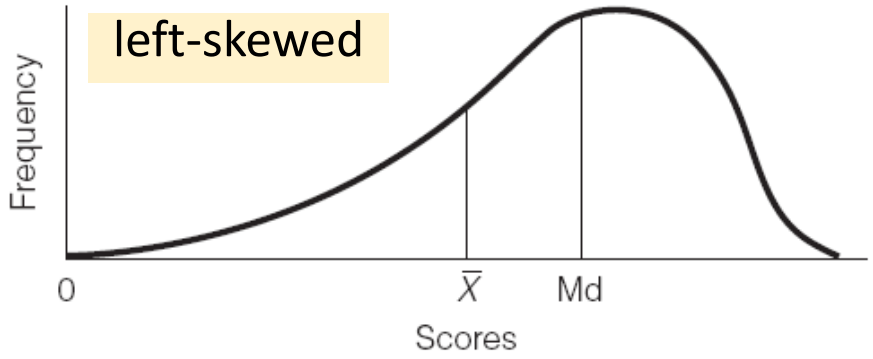
A Positively Skewed Distribution (The mean is greater in value than the median)

right-skewed



A Negatively Skewed Distribution (The mean is less than the median)

left-skewed



Is the age distribution positively or negatively skewed?

Choosing measures of central tendency

Use the mode when:	1. The variable is measured at the nominal level. 2. You want a quick and easy measure for ordinal and interval-ratio variables. 3. You want to report the most common score.
Use the median when:	1. The variable is measured at the ordinal level. 2. A variables measured at the interval-ratio level has a highly skewed distribution. 3. You want to report the central score. The median always lies at the exact center of a distribution.
Use the mean when:	1. The variable is measured at the interval-ratio level (except when the variable is highly skewed). 2. You want to report the typical score. The mean is "the fulcrum that exactly balances all of the scores." 3. You anticipate additional statistical analysis.

Measures of dispersion

- methods to describe the spread around the center value of a variable

- Range: difference between the max and min observed values
    - the age range was (89 - 18) 71 years

- the range is useful but limited
    - no information about spread between min and max
    - very sensitive to extreme scores

Let's consider the household size variable in the GSS

Range hhs = (16-1) 15

```
124 range(GSS$hhsize)
```

mean hhs = 2.64

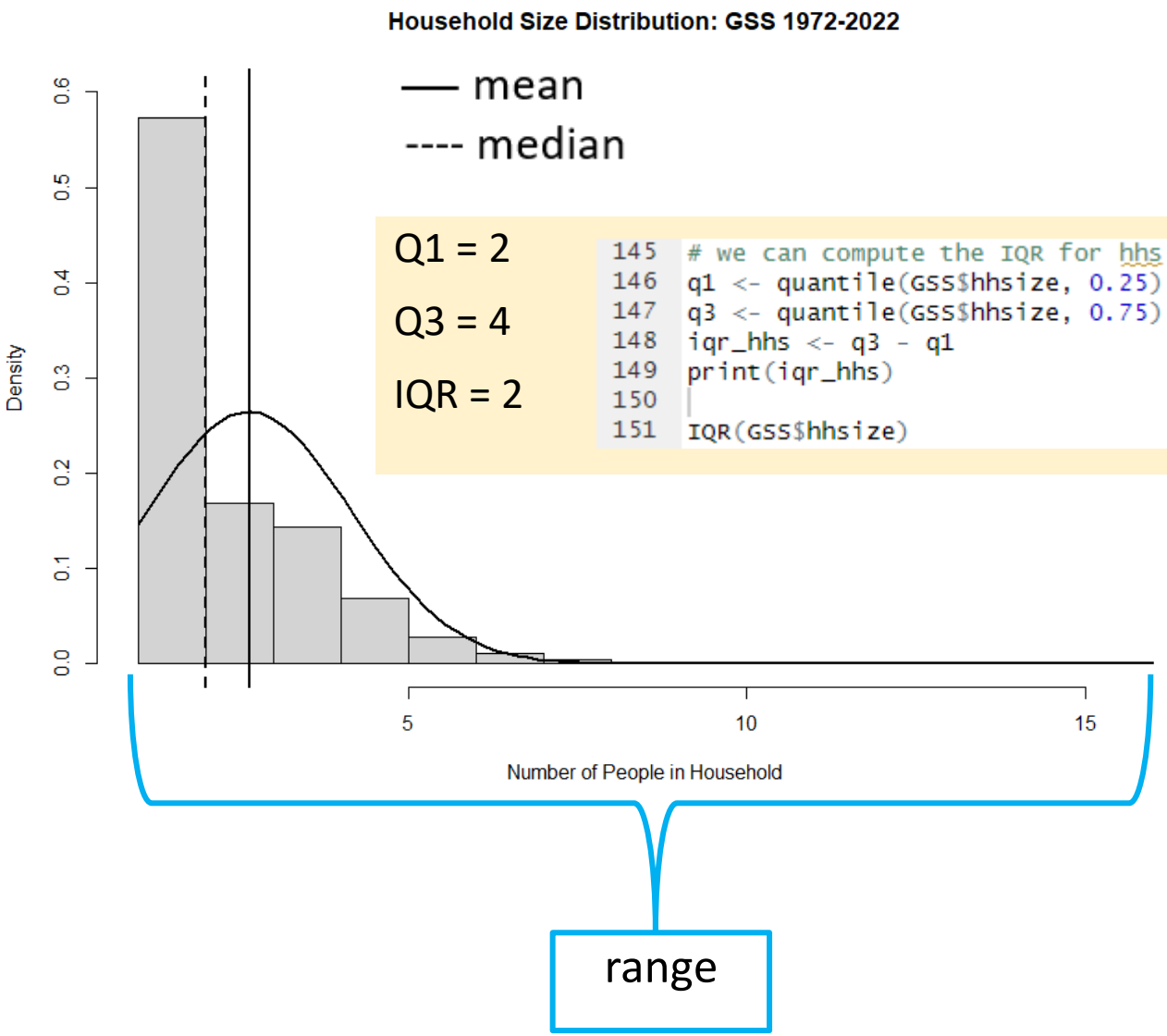
median hhs = 2.00

```
127 table(GSS$hhsize)
128 mean(GSS$hhsize)
129 median(GSS$hhsize)
```

Is the distribution positively or negatively skewed?

Which measure best reflects the center value?

Measures of dispersion: range



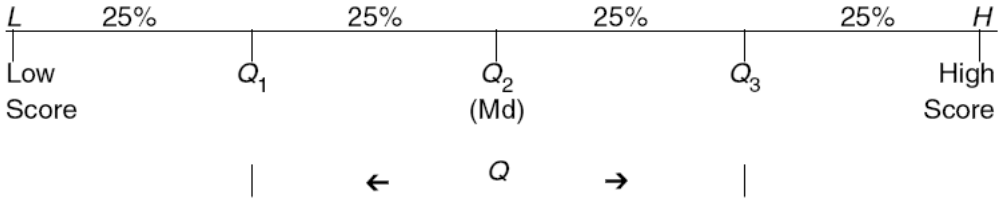
Q1 = 2

Q3 = 4

IQR = 2

```
145  # we can compute the IQR for hhs
146  q1 <- quantile(GSS$hhsize, 0.25)
147  q3 <- quantile(GSS$hhsize, 0.75)
148  iqr_hhs <- q3 - q1
149  print(iqr_hhs)
150
151  IQR(GSS$hhsize)
```

- Interquartile Range (IQR)
  - represents range within which 50% of the scores in a distribution fall
    - $IQR = Q3 - Q1$

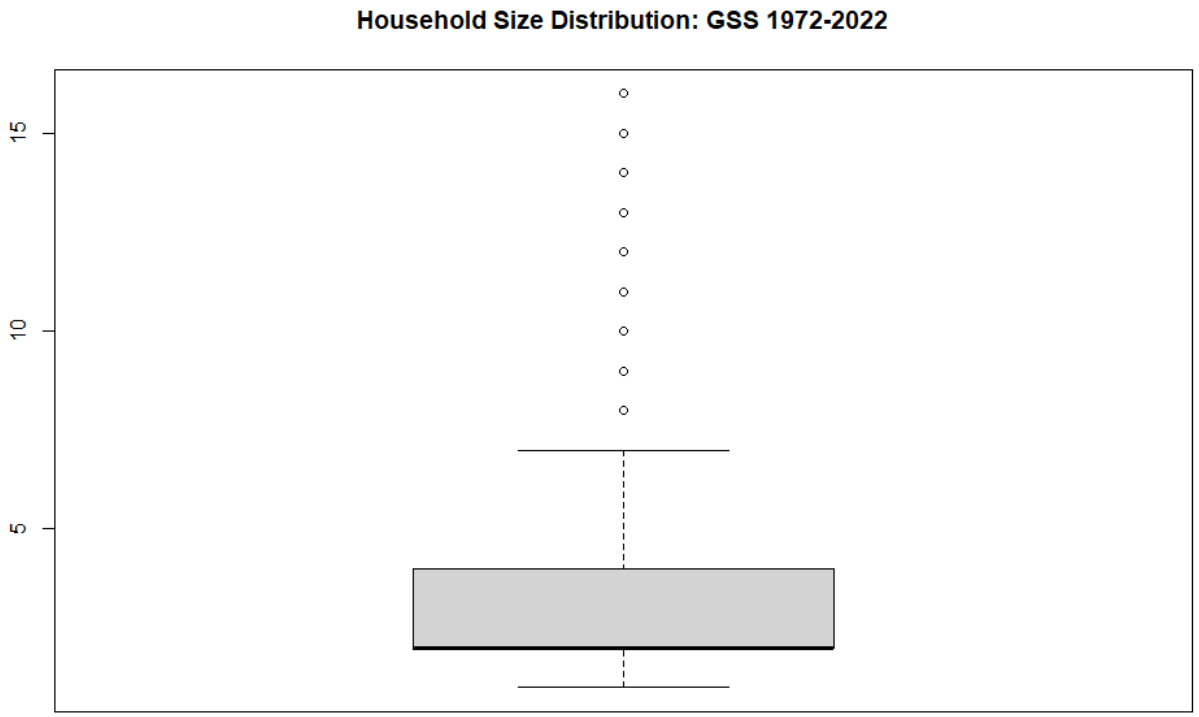


- useful for making sense of variables with highly skewed distributions
  - often due to outliers, data points substantially different than the rest

Measures of dispersion: box plot

- sometimes called a bar and whiskers plot

```
154  boxplot(GSS$hhsize,main="Household Size Distribution: GSS 1972-2022")
```



- box: depicts IQR
  - contains 50% of the data points
- median: line inside box
  - Q2
- whiskers: closest observed data point within 1.5 * IQR from Q1 and Q3
- outliers: any points outside whiskers

Measures of dispersion: variance

- average of the squared deviations of the data around the mean
  - useful measure for how much the observed values of a variable differ from the mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

```
160  # we can compute the variance of hhs
161  varinace_hhs<-sum((GSS$hhsize - mean(GSS$hhsize))^2)/(length(GSS$hhsize)-1)
162  print(varinace_hhs)
163  # or just use the var command
164  var(GSS$hhsize)
```

The variance of hhs = 2.28

- consider like the squared average distance between each score and the mean
  - becomes more meaningful when we get rid of the squared part…

Measures of dispersion: standard deviation

- average distance between any given observed value of a variable and its mean
  - provides interpretable units

informal definition

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{or} \quad s = \sqrt{s^2}$$

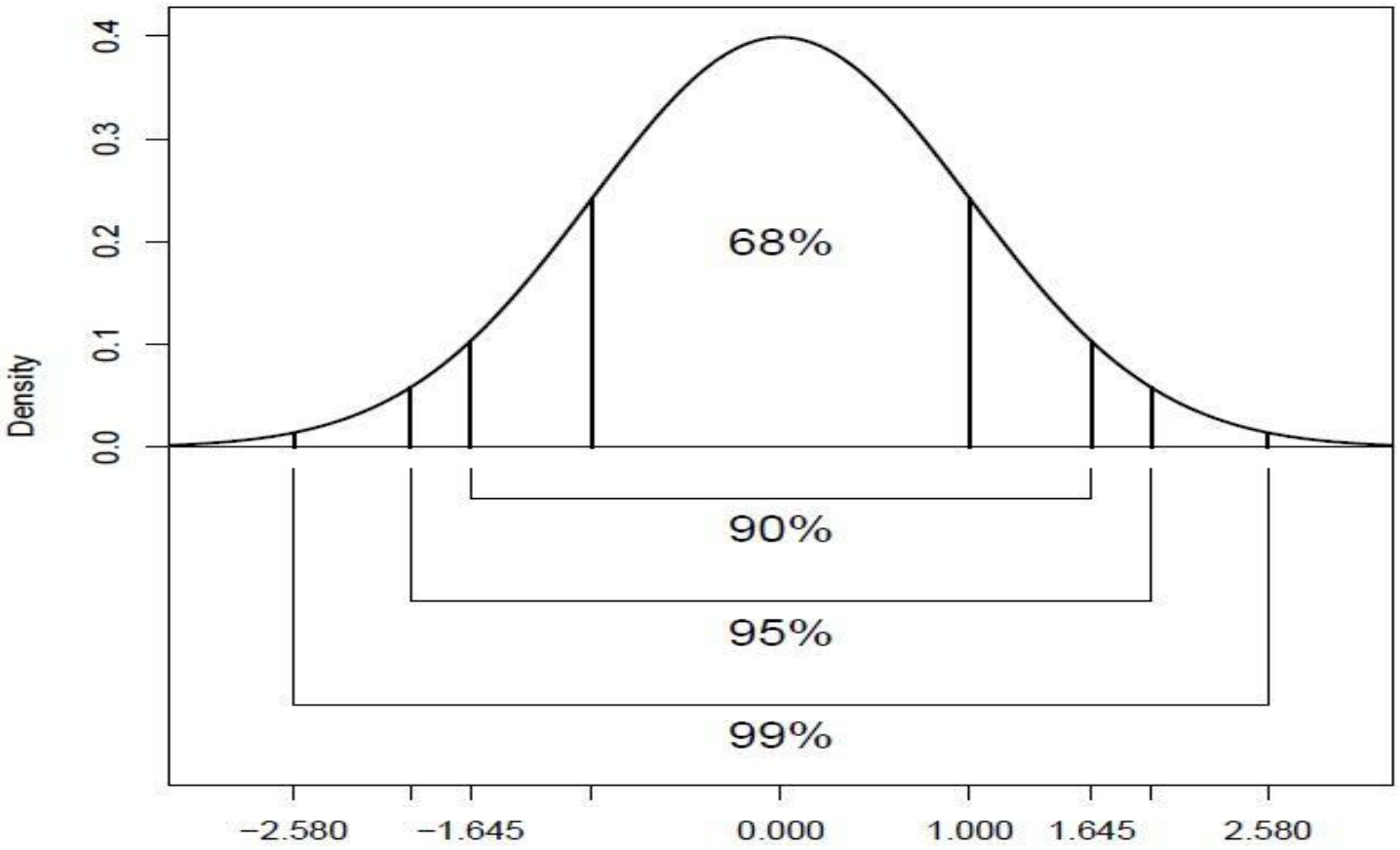
```
170 # we can compute the standard deviation for hhs
171 std_dev_hhs<-sqrt(sum((GSS$hhsize - mean(GSS$hhsize))^2)/(length(GSS$hhsize)-1))
172 print(std_dev_hhs)
173 # or just use the sd command
174 sd(GSS$hhsize)
```

The standard deviation for $hhs = \sqrt{2.28}$ = 1.509

- consider that most data points fall within $\pm$ 1 SD of the mean
  - this will become more meaningful as we get further into the material…

Important regions of the z distribution

- 68% of data points fall within $\pm$ 1 SD of the mean
  - in a normal distribution



- Future material with details no needed to be concerned with at the moment
- For now, let's develop a strong foundation in descriptive statistics…

Summarizing descriptive statistics: GSS

- descriptive statistics table
  - includes summary for all measures of interest

Descriptive Statistics Table: General Social Survey 1972-2022 (N = 64,555)

Variable	Mean (SD)	Median	Min.	Max.	Level of Measurement
Happiness			1	3	ordinal
Not too Happy	0.14				
Pretty Happy	0.56				
Very Happy	0.30				
Age	46.45 (17.63)	44.00	18	89+	interval-ratio
Female	0.56		0	1	nominal
White	0.80		0	1	nominal
Educational Attainment			0	4	ordinal
Less than HS	0.21				
HS	0.30				
Some College	0.24				
BA	0.17				
Graduate Deg.	0.08				
Married	0.53		0	1	nominal
Household Size	2.64 (1.51)	2.00	1	16	interval-ratio
Political Party Affiliation			1	3	nominal
Democrat	0.49				
Indep./Other	0.17				
Republican	0.34				

Note: table made in MS Word based on output from R

```
176 ###
177 # Computing Descriptive Statistics for Summary Table
178 ##
179
180 # Happiness: ordinal variable
181 prop.table(table(GSS$happy))
182
183 # Age: interval-ratio variable
184 # see how the summary command provides many different statistics
185 summary(GSS$age)
186 sd(GSS$age)
187
188 # Female: nominal variable (binary, means only two categories)
189 # this is a special case where only need to report one category,
190 # because remainder is intuitive (sums to 100%, see Descriptive Table)
191 mean(GSS$female)
192
193 # White: nominal variable (binary)
194 mean(GSS$white)
195
196 # Educational Attainment: ordinal variable
197 prop.table(table(GSS$educ_deg))
198
199 # Married: nominal variable (binary)
200 mean(GSS$married)
201
202 # Household Size: interval-ratio variable
203 summary(GSS$hhsize)
204 sd(GSS$hhsize)
205
206 # Political Party Affiliation: nominal variable
207 prop.table(table(GSS$polit_party))
208
209 ### End Descriptive Example for Summary Table ###
```

Summarizing descriptive statistics: Netflix

- Let's start to make sense of our Netflix survey data
  - and work toward making a descriptive statistics table

- First, we need to consider each variable's level of measurement
  - so we know how to appropriately summarize the data

Exercise 3

Review the survey and try to determine how responses to each respective question should be converted into a variable to make sense of these data.

See Ex3_PPT_SOC303