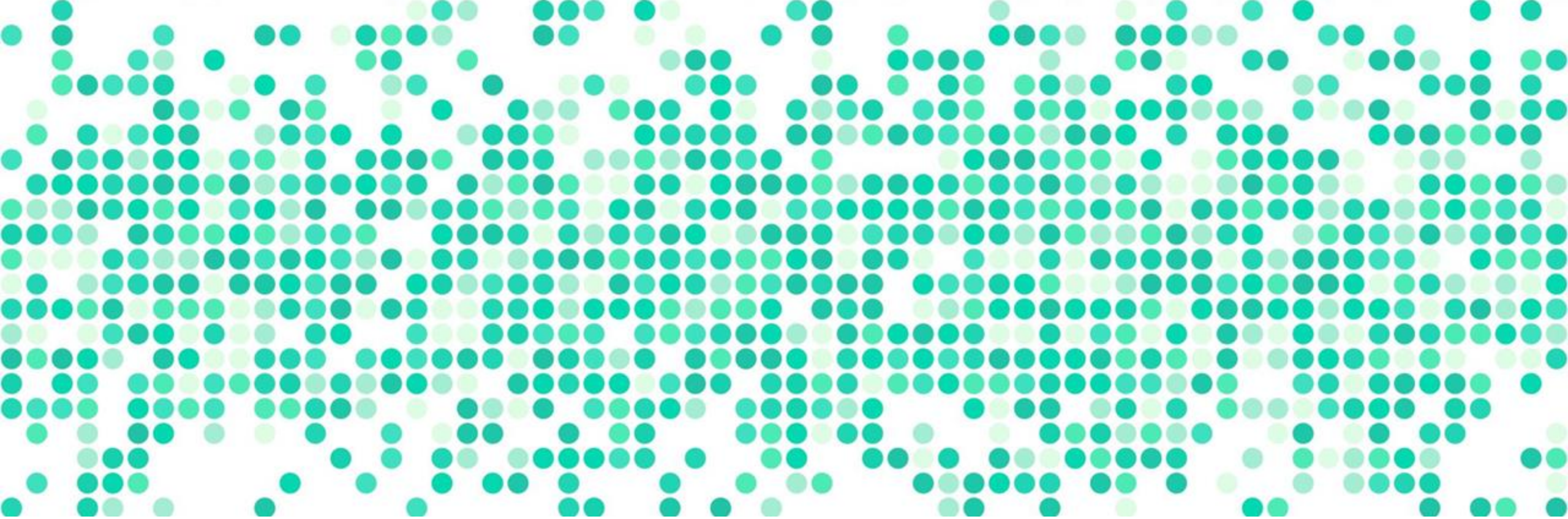


DATA ANALYSIS

Week 2: Summarizing and visualizing data

video watching grades

- Panopto tracks what percentage of the video you watched and assigns grades
- going forward, your viewing needs to be at least 50% for getting full credit, else you get half credit (partial watching) or no credit (no watching)
- ideally: watch in one sitting – try to create your own copy and work through the analyses while watching

  - Grades are not sent to Canvas until three hours after the user has started watching the video. As such, there is a delay between the time a user completes watching the video and when grades are sent.
    - The grade is sent regardless of whether or not the Viewer has finished watching the video. If a Viewer wants to stop and finish watching a video at a later time, they will need to refresh the page and resume watching the video. For this to occur, the 'assignment attempts' must be set to unlimited.
  - The reported percentage will always be the highest percentage of the video the student has watched, even if they rewind the video or watch it more than once and have a subsequent, partial viewing.

logistics

- PS1 / opt-out deadline: Feb 3
- recommended: odd problems have solutions at the back of the textbook
- Feb 11 class canceled
- Midterm 1 has been moved to February 27

2	T: January 28, 2025	W2: Summarizing & Fitting models to data
2	Th: January 30, 2025	W2 continued…
2	Su: February 2, 2025	**Week 2 Quiz due**
3	M: February 3, 2025	**PS1 due / Opt-out Deadline 1**
3	T: February 4, 2025	W3: Variability & z-scores
3	Th: February 6, 2025	W3 continued…
3	Su: February 9, 2025	**Week 3 Quiz due**
4	M: February 10, 2025	**PS2 + PS1 revision due**
4	T: February 11, 2025	W4: Correlation & Regression | No Class!
4	Th: February 13, 2025	W4 continued…
5	M: February 17, 2025	**PS2 revision due**
5	T: February 18, 2025	W5: More Correlation & Regression
5	Th: February 20, 2025	W6 continued…
5	Su: February 23, 2025	**Week 5 Quiz due**
5	M: February 24, 2025	**PS3 due**
6	T: February 25, 2025	W6: Loose Ends / Exam 1 review
6	Th: February 27, 2025	**Exam (Midterm) 1**
7	M: March 3, 2025	**PS3 revision due**

recap

- statistical thinking
- research design
- scales of measurement

lingering question

- We measured the sampling error in class by comparing the parameter to the statistic. How can this be applied when there is no viable way to measure the parameter because the population is too large (like how do we measure the sampling error in large populations)

5	Su: February 23, 2025	**Week 5 Quiz due**
5	M: February 24, 2025	**PS3 due**
6	T: February 25, 2025	W6: Loose Ends / Exam 1 review
6	Th: February 27, 2025	**Exam (Midterm) 1**
7	M: March 3, 2025	**PS3 revision due**
7	T: March 4, 2025	W7: Sampling and Hypothesis Testing
7	Th: March 6, 2025	W7 continued…
7	Su: March 9, 2025	**Week 7 Quiz due**

today's agenda

why summarize?

summarization methods

data visualization

why summarize?

- a researcher gives 25 participants a list of 10 words. After 10 minutes, they are asked to write down the words they remember - these words are then counted. The list of scores is:

  7, 8, 5, 8, 5, 6, 9, 6, 5, 7, 7, 8, 3, 8, 7, 9, 3, 7, 8, 7, 6, 8, 5, 10, 7

- scale of measurement (NOIR)?
- why would the researcher summarize these scores? what purpose does summarization serve over and above simply presenting the raw scores / data points?

why summarize?

- a researcher gives 25 participants a list of 10 words. After 10 minutes, they are asked to write down the words they remember - these words are then counted. The list of scores is:

7, 8, 5, 8, 5, 6, 9, 6, 5, 7, 7, 8, 3, 8, 7, 9, 3, 7, 8, 7, 6, 8, 5, 10, 7

- the researcher would like to present a summary of her data.

data and tables

- tables/spreadsheets allow us to view data in a sequential and ordered manner
- raw data: when each participant's observation is a different row of data
- easy calculations from raw data
  - min/max, range, sum
- is it easy to visually tell which is the most common value in the dataset?

participant	words_recalled (X)
A	7
B	8
C	5
D	8
E	5
F	6
G	9
H	6
I	5
J	7
K	7
L	8
M	3
N	8
O	7
P	9
Q	3
R	7
S	8
T	7
U	6
V	8
W	5
X	10
Y	7

Minimum	Maximum	Range	Sum
3	10	7	169

frequency table

- an organized tabulation of the number of scores at each value of the measurement scale
- gives a picture of how the scores are distributed on the scale
- each row is a possible value on the scale of measurement
- frequency (f) records how often a particular score was observed, i.e., how many people had that score?
  - adding up the frequencies will give you the total number of people whose scores were measured, i.e., sample size
- fX = product of a score (X) and number of people with that score (f)
  - adding up fX will give you the total SUM of ALL scores ($\sum fX$)

X	Frequency(f)	fX
0	0	0
1	0	0
2	0	0
3	2	6
4	0	0
5	4	20
6	3	18
7	7	49
8	6	48
9	2	18
10	1	10
	25	169

relative frequency

- relative frequency refers to the proportion and the percentage of the total group that is associated with each score, i.e., what proportion/percentage of people had this score?
- proportion (0 to 1)= $\frac{f}{N}$ = probability
- percentage (0 to 100) = $\frac{f}{N}$ x 100

X	Frequency(f)	fX
0	0	0
1	0	0
2	0	0
3	2	6
4	0	0
5	4	20
6	3	18
7	7	49
8	6	48
9	2	18
10	1	10

your survey responses!

- most of you filled out a Survey of Attitudes Towards Statistics (SATS)
- the questions spanned six different domains
  - affect: how you feel about statistics
  - cognitive competence: how you assess your intellectual abilities towards statistics
  - difficulty: how you assess the difficulty of statistics as a subject
  - effort: how much effort you expect to put into the course
  - interest: your level of interest in the course
  - value: your assessment of how relevant or useful statistics will be in your life

affect

I will like statistics
I will feel insecure when I have to do statistics problems
I will get frustrated going over statistics tests in class.
I will be under stress during statistics class.
I will enjoy taking statistics courses.
I am scared by statistics.

cognitive competence

I will have trouble understanding statistics because of how I think
I will have no idea of what's going on in this statistics course.
I will make a lot of math errors in statistics.
I can learn statistics.
I will understand statistics equations.
I will find it difficult to understand statistical concepts.

effort

I plan to complete all of my statistics assignments
I plan to work hard in my statistics course
I plan to study hard for every statistics test.
I plan to attend every statistics class session.

difficulty

Statistics formulas are easy to understand
Statistics is a complicated subject.
Statistics is a subject quickly learned by most people.
Learning statistics requires a great deal of discipline.
Statistics involves massive computations.
Statistics is highly technical.
Most people have to learn a new way of thinking to do statistics.

value

Statistics is worthless.
Statistics should be a required part of my professional training.
Statistical skills will make me more employable.
Statistics is not useful to the typical professional.
Statistical thinking is not applicable in my life outside my job.
I use statistics in my everyday life.
Statistics conclusions are rarely presented in everyday life.
I will have no application for statistics in my profession.
Statistics is irrelevant in my life.

interest

I am interested in being able to communicate statistical information to others.
I am interested in using statistics.
I am interested in understanding statistical information.
I am interested in learning statistics.

survey data exploration

- responses ranged from 1 (strongly disagree) to 7 (strongly agree)
- what type of data (NOIR)?
  - scores on individual items
  - components?
- minimum/maximum score in the dataset? range?
- high scores indicate positive attitudes
- low scores indicate negative attitudes

W2 Activity 1 (part 1)

- go to activity link under W2
- will need to “make a copy” to edit doc
- construct a frequency table
- compute f, fX, proportion, and percentage
- **answer questions on your own (no peers, no LAs!)**

W2 Activity 1 (part 2)

- talk to a peer about your analyses and answers

W2 Activity 1 (part 3)

- go to activity link under W2
- re-answer questions

W2 Activity 1 debrief

"I will like statistics"
6
6
6
4
6
6
5

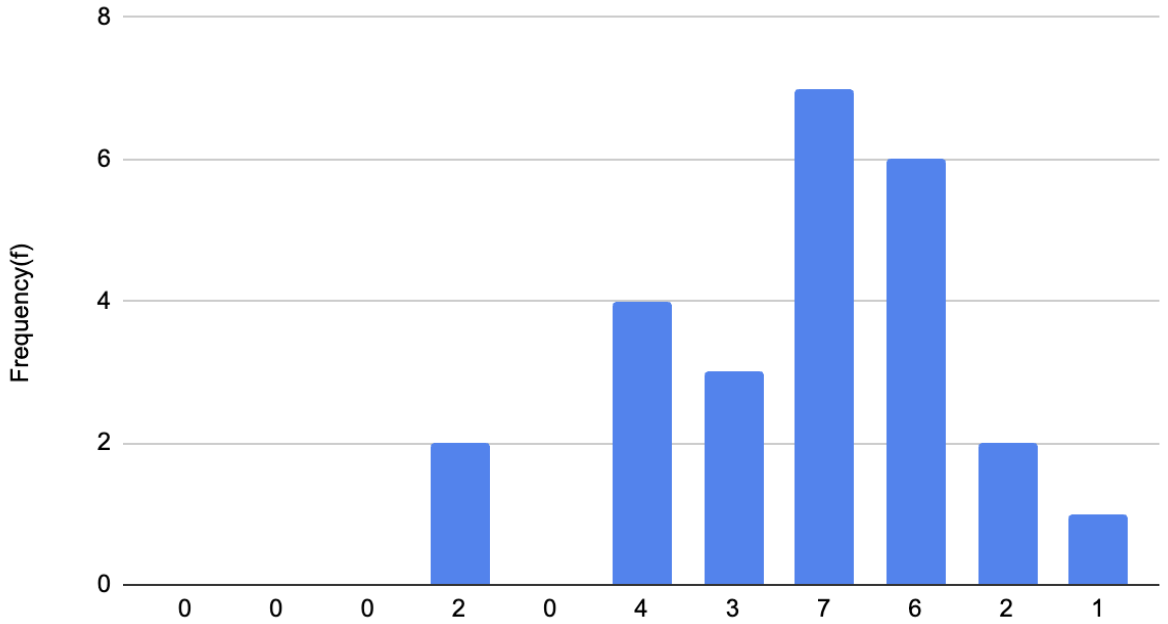
- How many students provided a rating of 4 for the question "I will like statistics"?
- What is the sample size, i.e., how many students responded to this survey item?
- What is $\sum fX$ for these data and what does it represent?
- If all calculations are correct, the sum of the proportions in a frequency table should always be 1.
- Percentages and proportions obtained from a frequency table provide the same information in different mathematical formats.
- Based on these data, what can you conclude?

X	Frequency (f)	fX	proportion	percentage	n (count)
1	0	0	0	0	32
2	0	0	0	0	
3	2	6	0.0625	6.25	
4	10	40	0.3125	31.25	
5	10	50	0.3125	31.25	
6	9	54	0.28125	28.125	
7	1	7	0.03125	3.125	
	sum	sum(fX)	sum (proportions	sum (percent)	
	32	157	1	100	

from tables to graphs: histograms

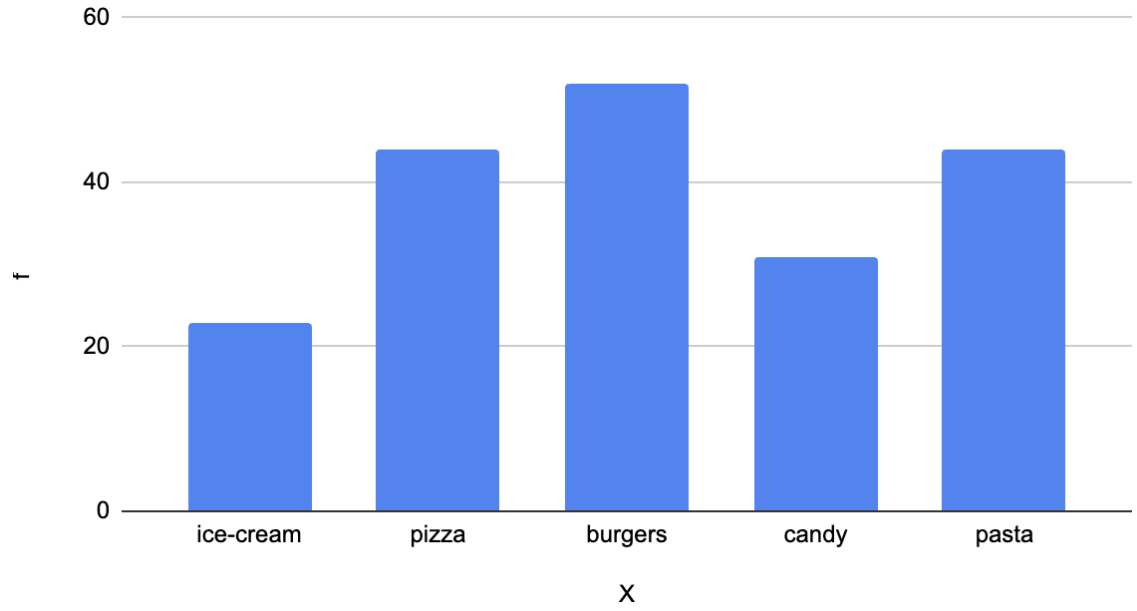
- visualizing the frequency of scores is helpful

X	Frequency(f)	
0	0	
1	0	
2	0	
3	2	xx
4	0	
5	4	xxxx
6	3	xxx
7	7	xxxxxxx
8	6	xxxxxx
9	2	x
10	1	x



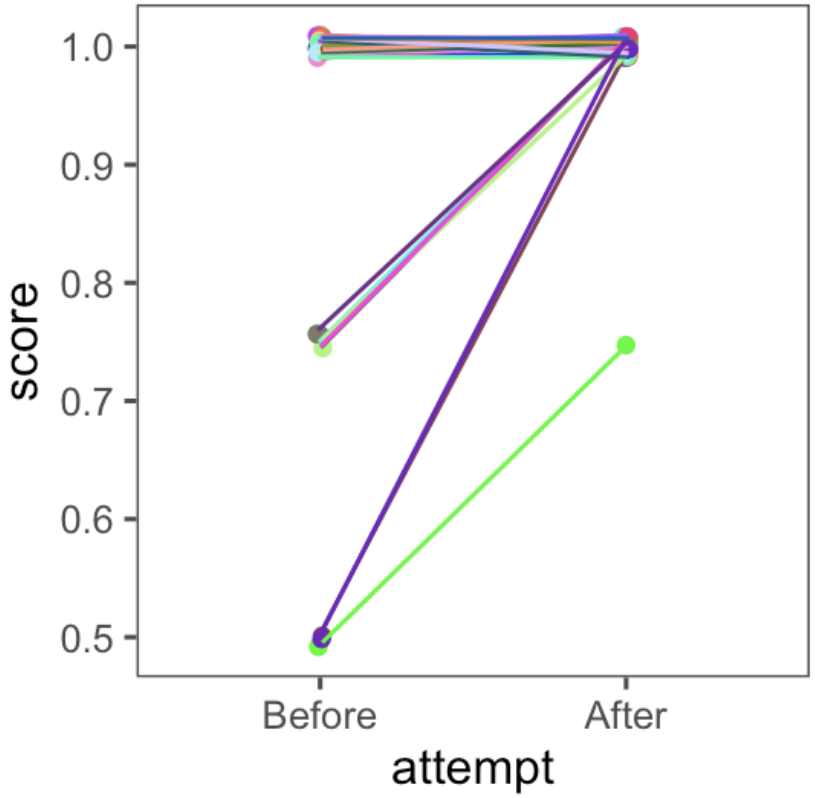
histograms vs bar graphs

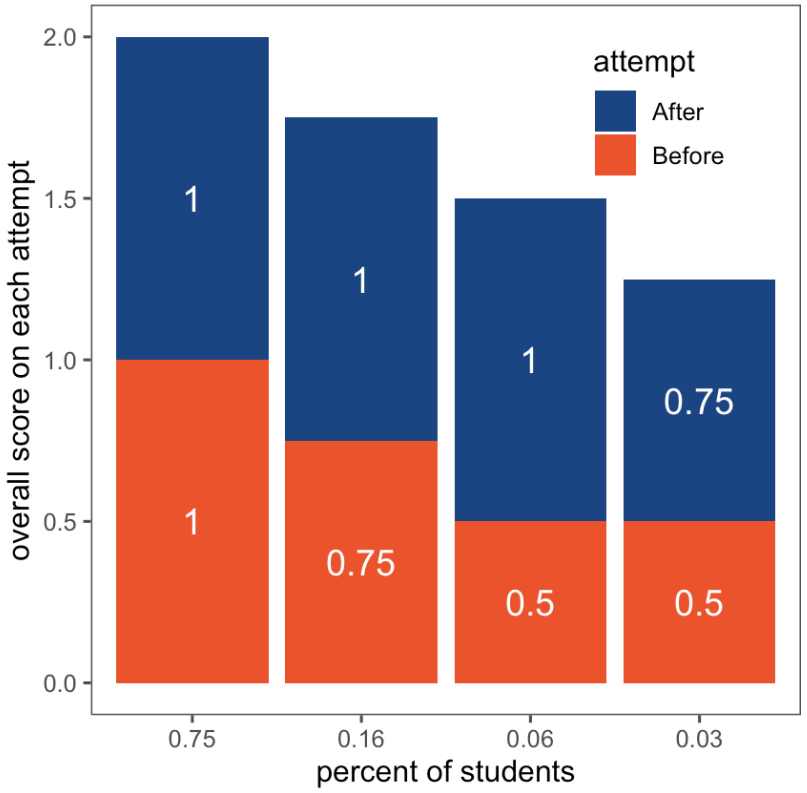
- what kind of variable (NOIR)?
- how many participants?
- histograms are for continuous variables
- bar graphs are for discrete variables (nominal / ordinal)



W2 activity summary

peer learning in action!





ranks and percentiles

- sometimes, we want to know about the position of a specific score / individual within a distribution of scores
- examples?
- rank/percentile rank: percentage of individuals with scores at or below the particular value

cumulative frequency

- cumulative frequency = cf

  = frequency of scores up until that point

- cumulative percentage = c%

  $= \frac{cf}{N} * 100$

  = percentile

X	Frequency(f)
0	0
1	0
2	0
3	2
4	0
5	4
6	3
7	7
8	6
9	2
10	1

percentile

- always use real limits (“8” is really 7.5 to 8.5)
- which score corresponds to the 88^{th} percentile?
- which score corresponds to the 36^{th} percentile?

X	Frequency(f)	cumulative frequency (cf)	c%
0	0	0	0
1	0	0	0
2	0	0	0
3	2	2	8
4	0	2	8
5	4	6	24
6	3	9	36
7	7	16	64
8	6	22	88
9	2	24	96
10	1	25	100

grouped frequency tables

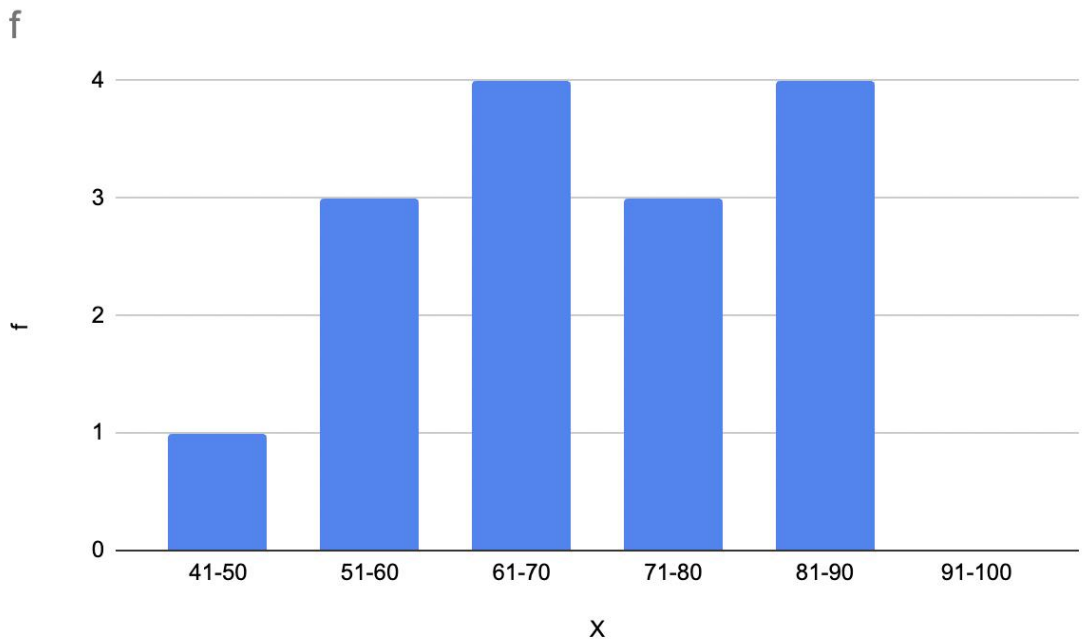
- when scores cover a wide range of possible values, it is useful to bin the data together into groups
- general guidelines
  - aim for approximately 10 bins/class intervals
  - the interval width should be a “simple” number (e.g., 5s, 10s, etc.)
  - lowest score should be multiple of class interval (e.g., starting from 5)
  - all intervals should have the same width
- real limits (continuous data): an interval of 5-10 really is an interval from 4.5 to 10.5

example

participant	X
A	61
B	63
C	73
D	53
E	66
F	52
G	86
H	82
I	50
J	65
K	55
L	75
M	88
N	90
O	80

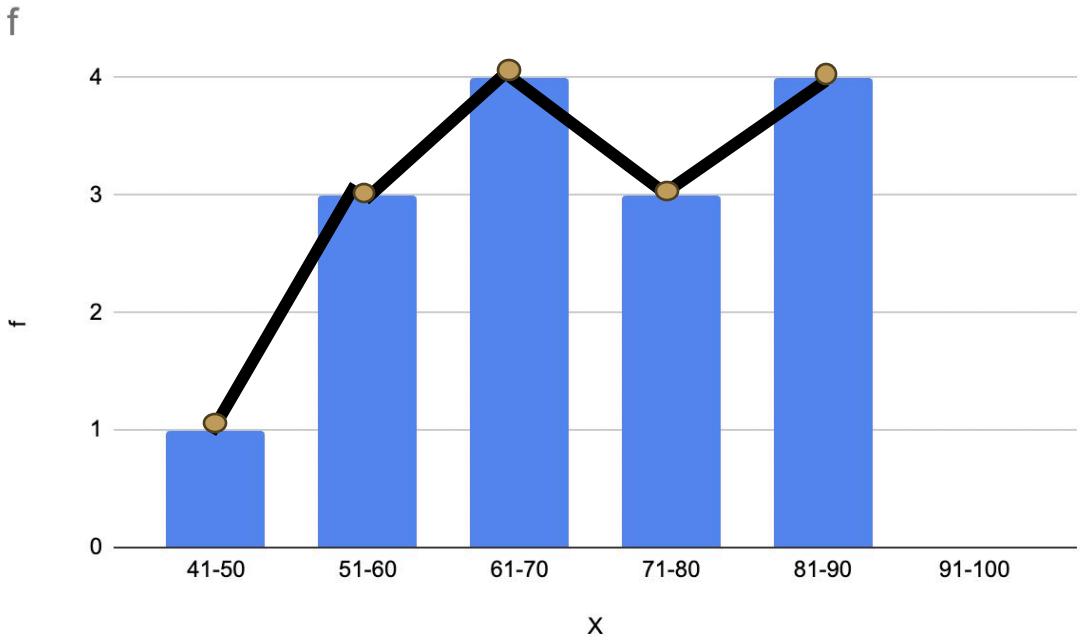
X	f
41-50	1
51-60	3
61-70	4
71-80	3
81-90	4
91-100	0

what is the width?



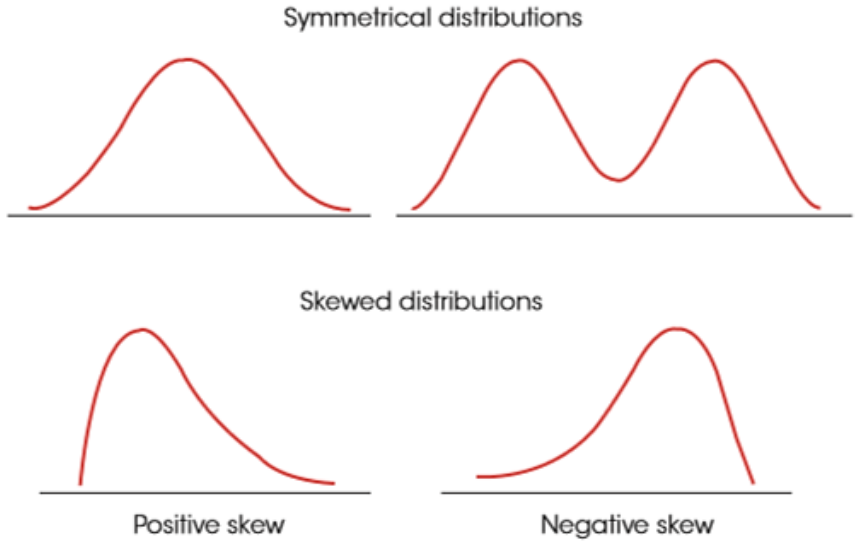
frequency polygons

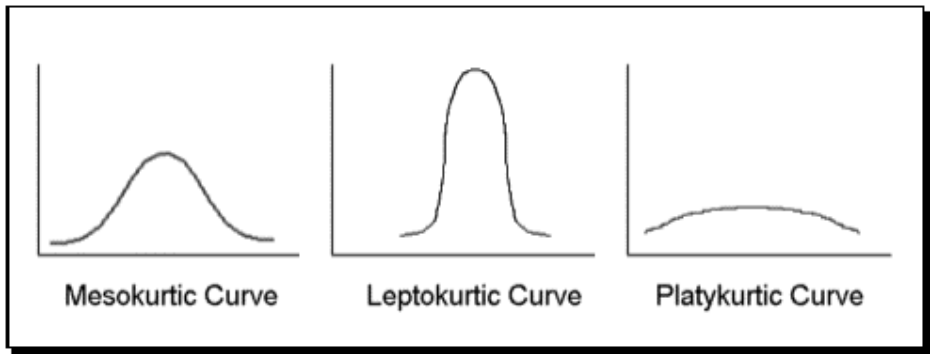
- contains the same data as a frequency histogram or table



shapes of distributions

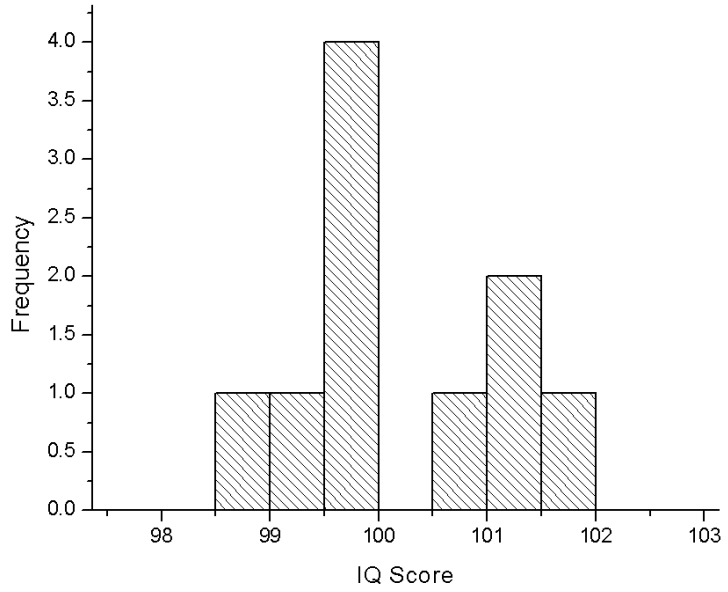
- populations are often displayed using smooth curves
- distributions are typically described along three dimensions
  - shape (symmetric, skewed, etc.)
  - central tendency (unimodal, bimodal, etc.)
  - variation/tailedness (kurtosis)





shapes of distributions

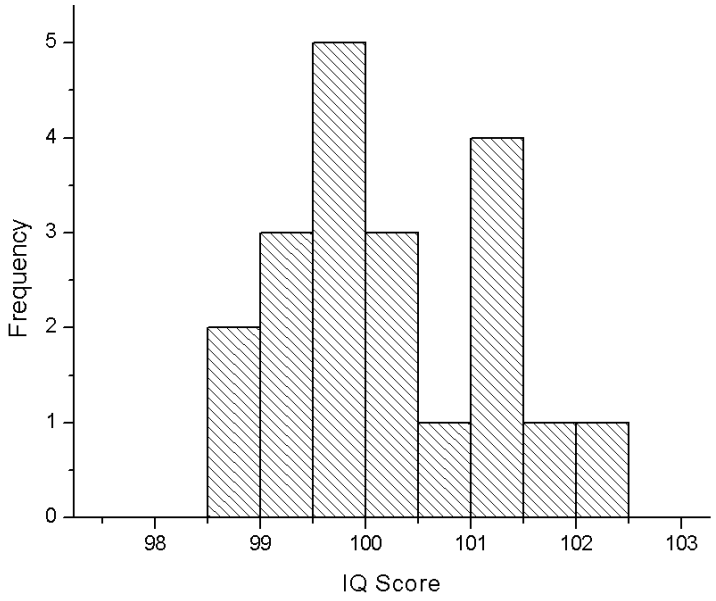
- distributions of larger samples tend to be smoother



n = 10

shapes of distributions

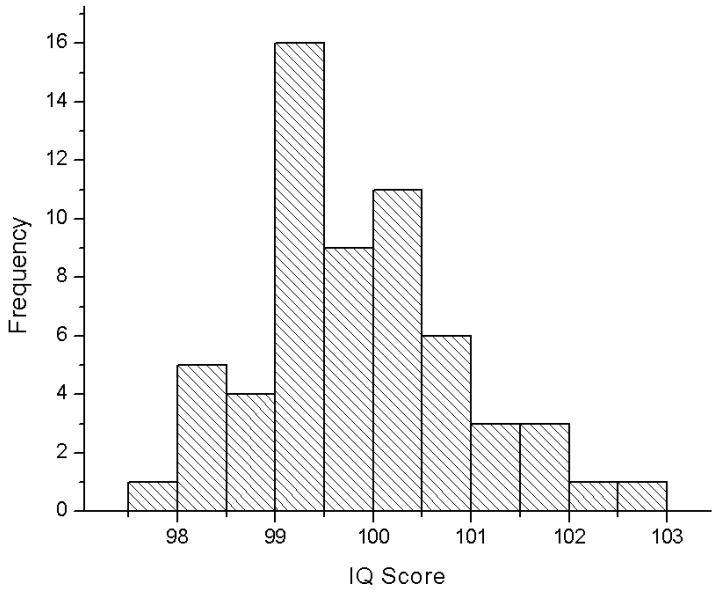
- distributions of larger samples tend to be smoother



n = 20

shapes of distributions

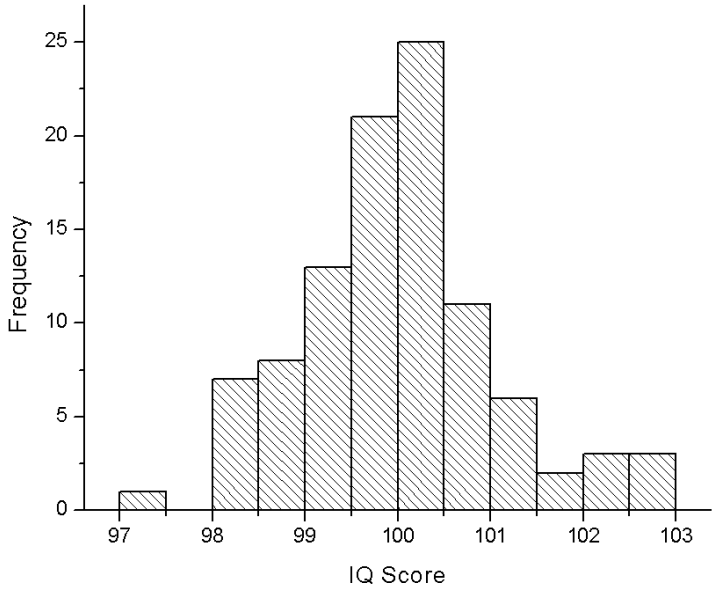
- distributions of larger samples tend to be smoother



n = 60

shapes of distributions

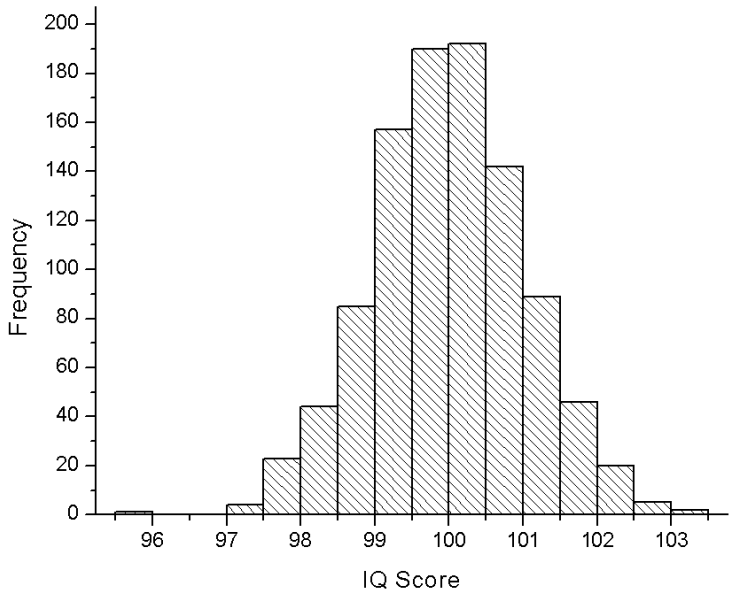
- distributions of larger samples tend to be smoother



n = 100

shapes of distributions

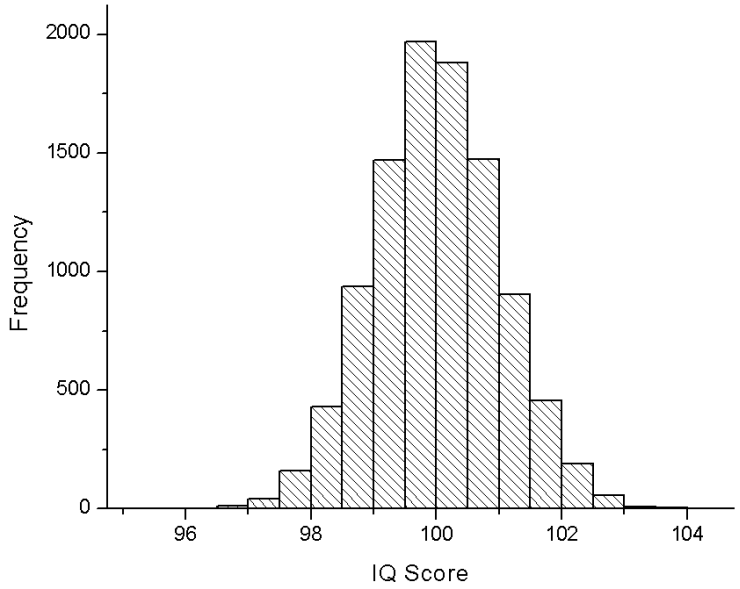
- distributions of larger samples tend to be smoother



n = 1000

shapes of distributions

- distributions of larger samples tend to be smoother

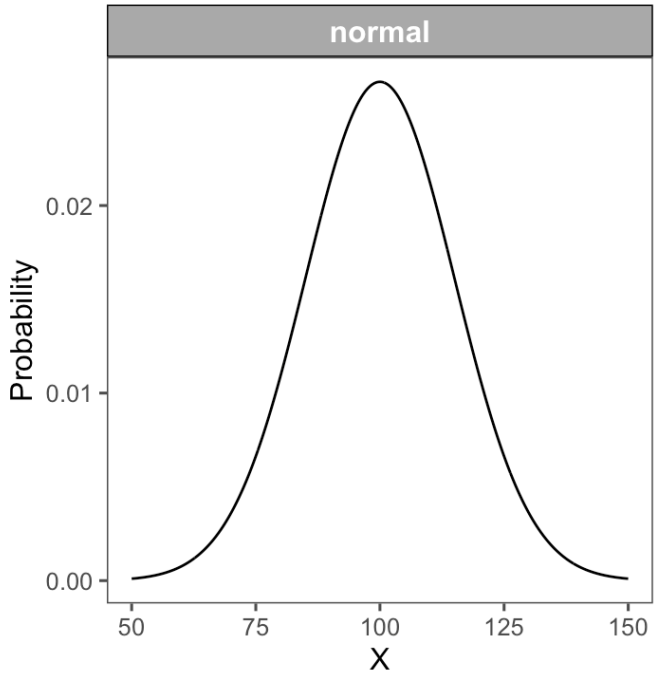


n = 10000

normal distribution

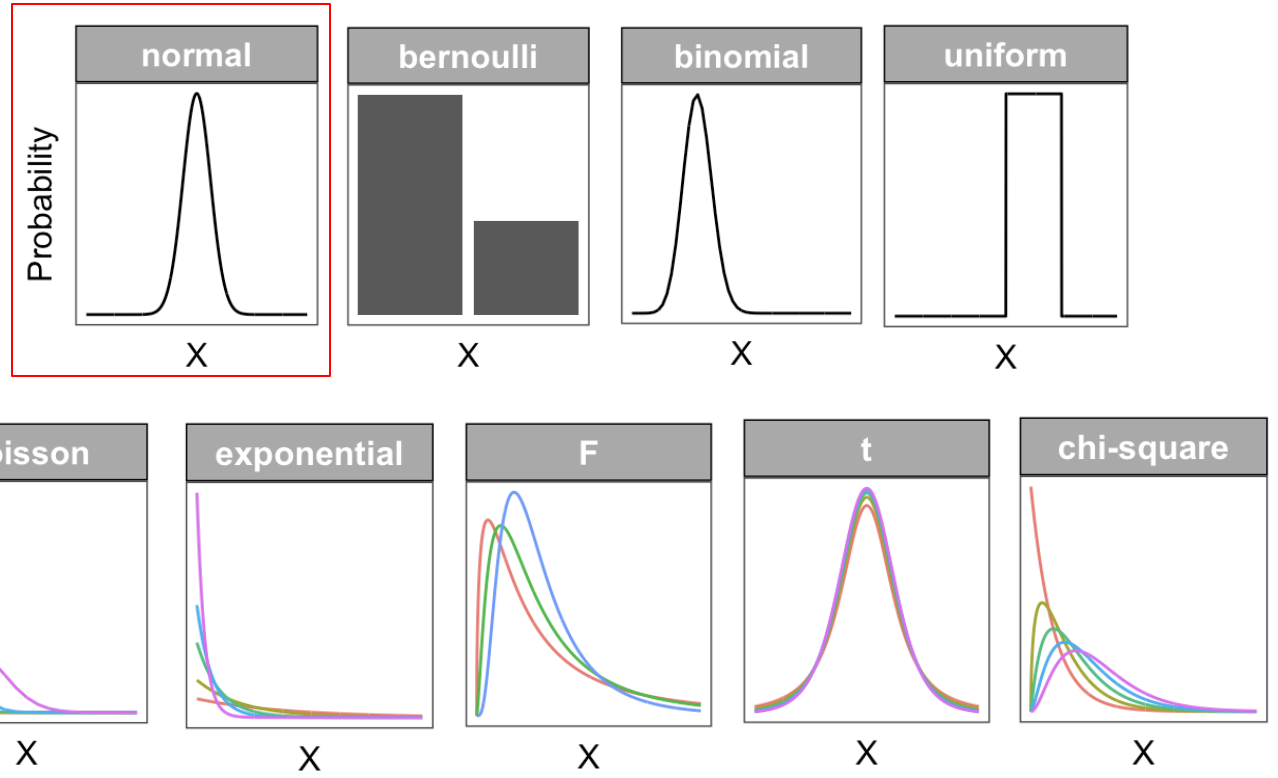
$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- the normal distribution is commonly observed for large numbers of scores
- normal ≈ typical, i.e., observed quite often
- real-life normal distributions: human heights/weights, test scores, etc.
- has a precise mathematical form that depends on two parameters (mean and standard deviation), which determine how frequent a given observation is



other distributions

- bernoulli (only 2 possible outcomes)
- binomial (bernoulli many times)
- uniform (same frequency for all)
- poisson (high frequency for low X)
- exponential
- F distribution(s)
- t distribution(s)
- chi-square distribution(s)



next time

- what is a model?
- a framework for understanding data

Before Tuesday

- Watch: Summarizing Data. (See Google Sheets Solution Here)
- Read Chapter 2 from the Gravetter & Wallnau (2017) textbook.

Before Thursday

- Watch: Central Tendencies. (See Google Sheets Solution Here)
- Read Chapter 3 from the Gravetter & Wallnau (2017) textbook.

Here are the to-do's for this week:

- Submit Week 2 Quiz
- Submit Problem Set 1 OR Opt out of Problem Sets
- Submit any lingering questions here!
- Extra credit opportunities:
  - Submit Exra Credit Questions
  - Submit Optional Meme Submission

optional: interpolation

- which score corresponds to the 50th percentile?
- if we don't have the percentile in the table, then we use interpolation
- percentile between 36 and 64 (interval width = 28 points) and scores between 6.5 and 7.5 (interval width = 1 point)
- 50% is 14 points away from 64%, i.e., 14/28 = ½ of the total interval width, i.e., ½ *(1) = 0.5 points
- so, we go 0.5 points down from the top score of 7.5, i.e., 7 points is the 50th percentile

X	Frequency(f)	cumulative frequency (cf)	c%
0	0	0	0
1	0	0	0
2	0	0	0
3	2	2	8
4	0	2	8
5	4	6	24
6	3	9	36
7	7	16	64
8	6	22	88
9	2	24	96
10	1	25	100