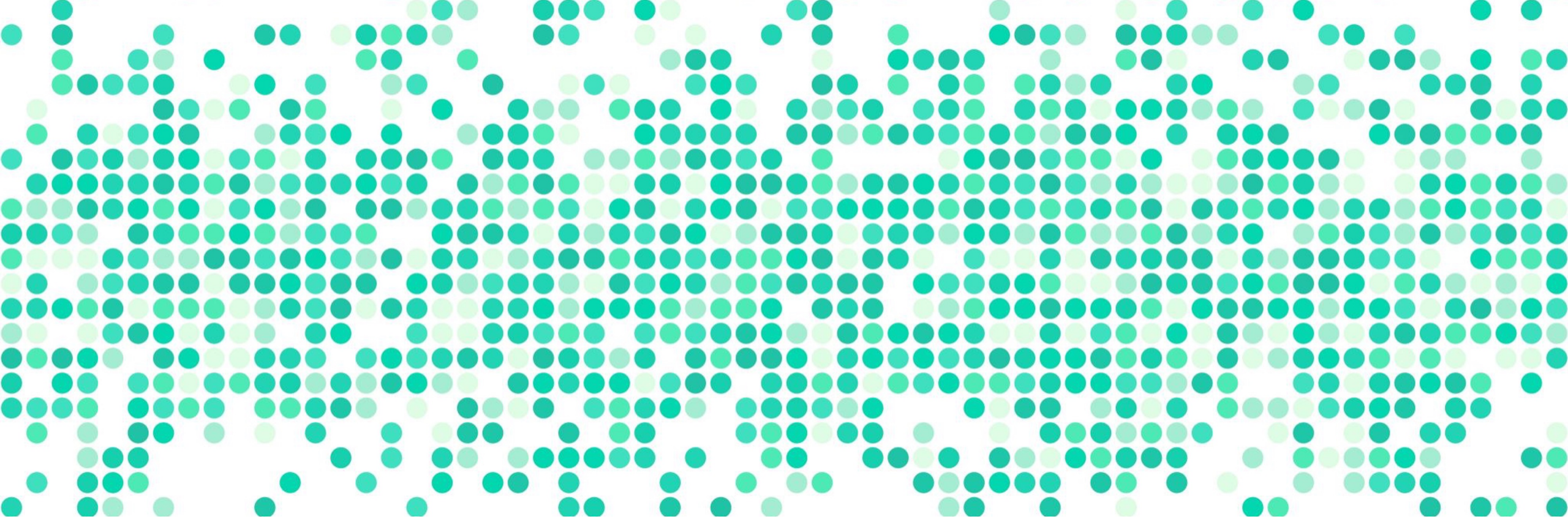


DATA ANALYSIS

Week 2: Summarizing and visualizing data

logistics

- problem set 1 / opt-out deadline: Feb 5
- recommended: odd problems have solutions at the back of the textbook
- *usual* office hours
  - Prof. Kumar: Wed 2-4 (in-person) and Thurs 2-4 (zoom)
  - Yanevith: Sun, 3.30 pm - 5 pm
  - Whitt: Mon, 7 pm - 8.30 pm
- *additional* office hours:
  - Prof. Kumar: Feb 5 (Mon), 10-11.30 and 4.30-5.30 [in-person]

recap

- what we covered:
  - course overview
  - statistical thinking / scales of measurement / research methods
- your to-dos:
  - *try*: week 1 quiz
  - *apply*: problem set #1 (chapter 1 problems)
  - *apply*: optional meme / discussion board post
  - *prep*: Chapter 2 from textbook

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?
- statistical thinking is about explaining the complex world in simple terms
- summarization helps us simplify the complexity in data

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.
- what would be some helpful summary information to present to an audience?
  - minimum/maximum value in the dataset?
    - range = maximum – minimum
  - most common score in the dataset? average score?

data and tables

- tables/spreadsheets allow us to view data in a sequential and ordered manner
- view the data
- 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
- go to second sheet
- 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 and number of people with that score
  - adding up fX will give you the total SUM of ALL scores

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)= $\frac{f}{N}$

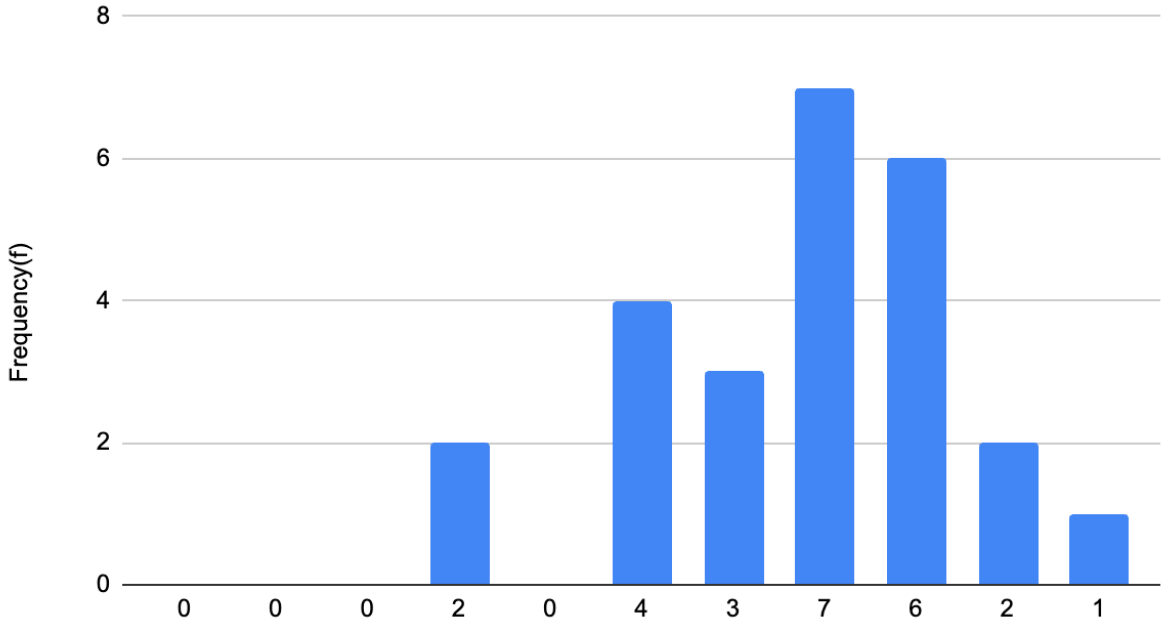
- percentage (0 to 100) = $\frac{f}{N}$ x 100

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

from tables to graphs: histograms

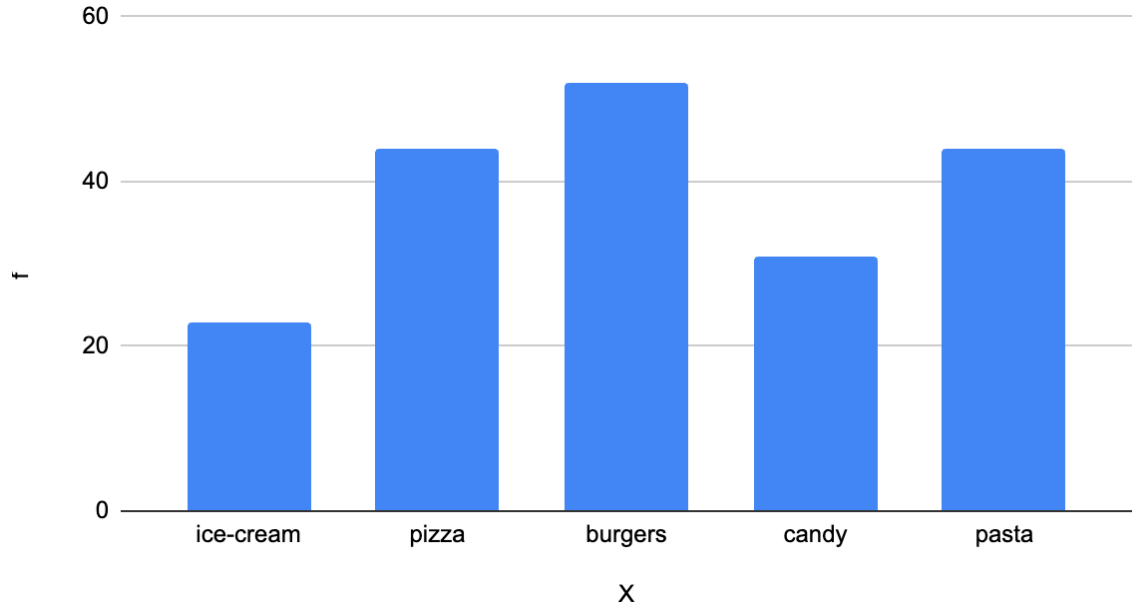
- visualizing the frequency of scores is helpful [video tutorials on course website]

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

- view the “bar graph” sheet
- what kind of variable (NOIR)?
- how many participants?
- histograms are for continuous variables
- bar graphs are for discrete variables (nominal / ordinal)



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

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

activity

- data: your responses on the item “I will like statistics”
- construct a frequency table
- compute f, fX, proportion, and percentage
- will need to “make a copy” to edit doc

"I will like statistics"
6
6
5
3
7
5
2

X	f	fX	proportion	percent
1	0	0	0	0
2	2	4	0.0625	6.25
3	1	3	0.03125	3.125
4	9	36	0.28125	28.125
5	9	45	0.28125	28.125
6	8	48	0.25	25
7	3	21	0.09375	9.375
	sum (f)	**sum (fX)**	**sum(proportions)**	**sum(percent)**
	32	157	1	100

WOMEN IN SCIENCE

Expectations of brilliance underlie gender distributions across academic disciplines

Sarah-Jane Leslie,[1*†] Andrei Cimpian,[2*†] Meredith Meyer,[3] Edward Freeland[4]

The gender imbalance in STEM subjects dominates current debates about women's underrepresentation in academia. However, women are well represented at the Ph.D. level in some sciences and poorly represented in some humanities (e.g., in 2011, 54% of U.S. Ph.D.'s in molecular biology were women versus only 31% in philosophy). We hypothesize that, across the academic spectrum, women are underrepresented in fields whose practitioners believe that raw, innate talent is the main requirement for success, because women are stereotyped as not possessing such talent. This hypothesis extends to African Americans' underrepresentation as well, as this group is subject to similar stereotypes. Results from a nationwide survey of academics support our hypothesis (termed the field-specific ability beliefs hypothesis) over three competing hypotheses.

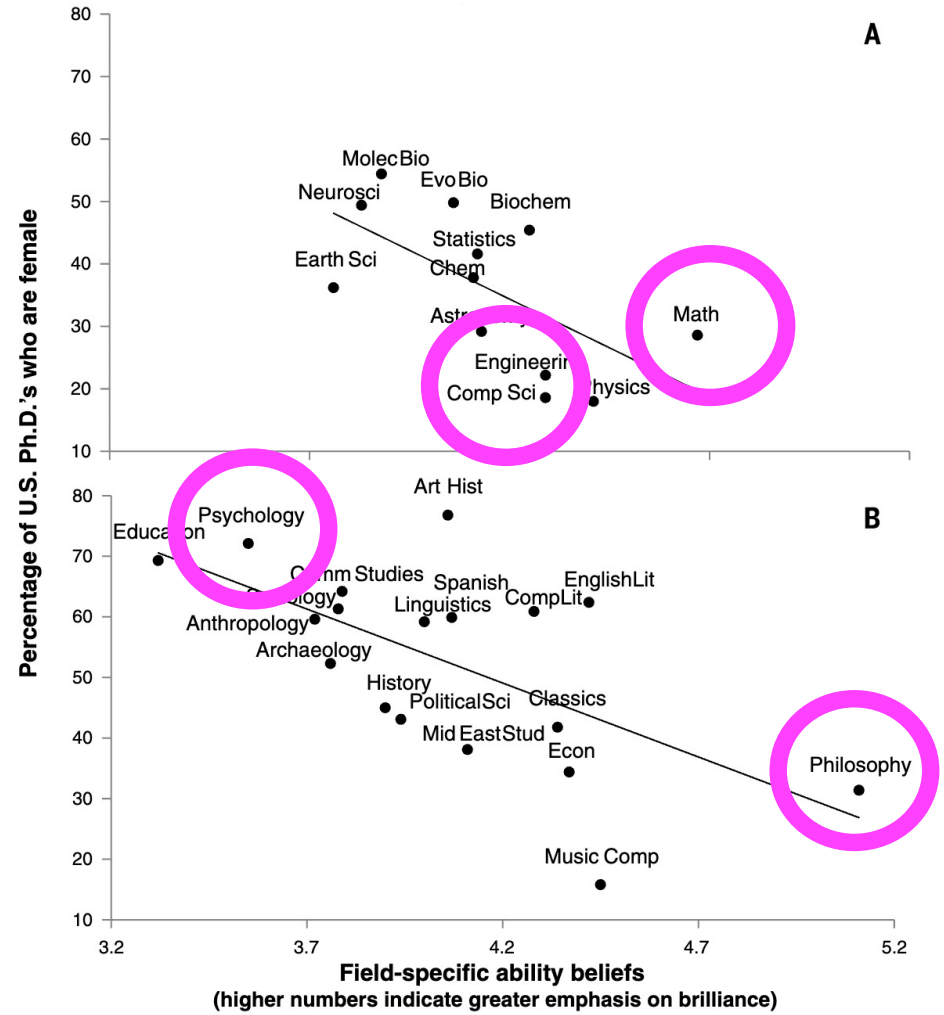


Fig. 1. Field-specific ability beliefs and the percentage of female 2011 U.S. Ph.D.'s in (A) STEM and (B) Social Science and Humanities.

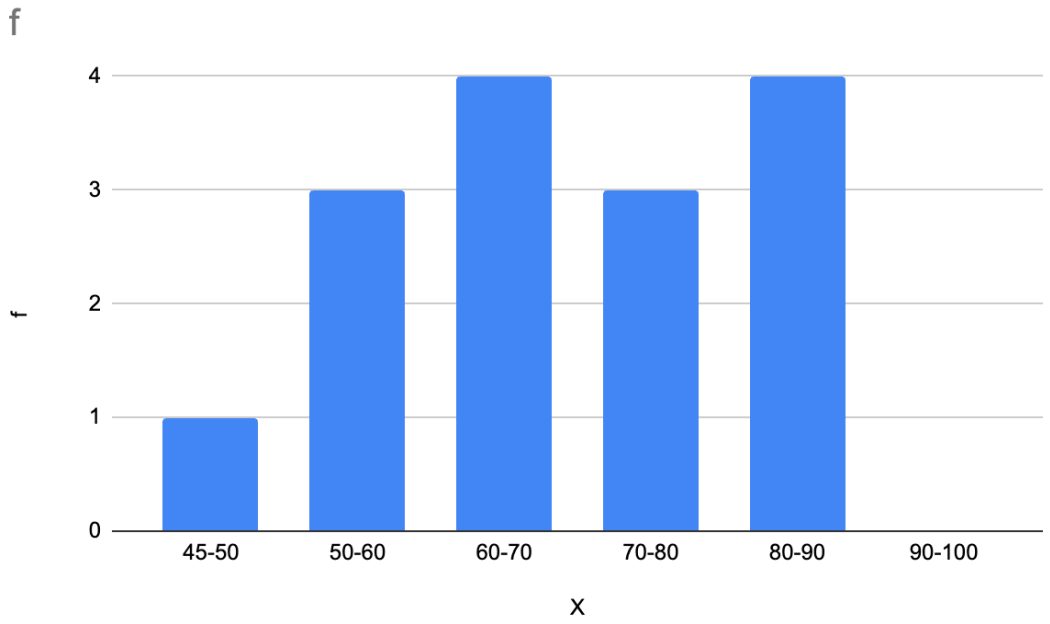
grouped frequency tables

- 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

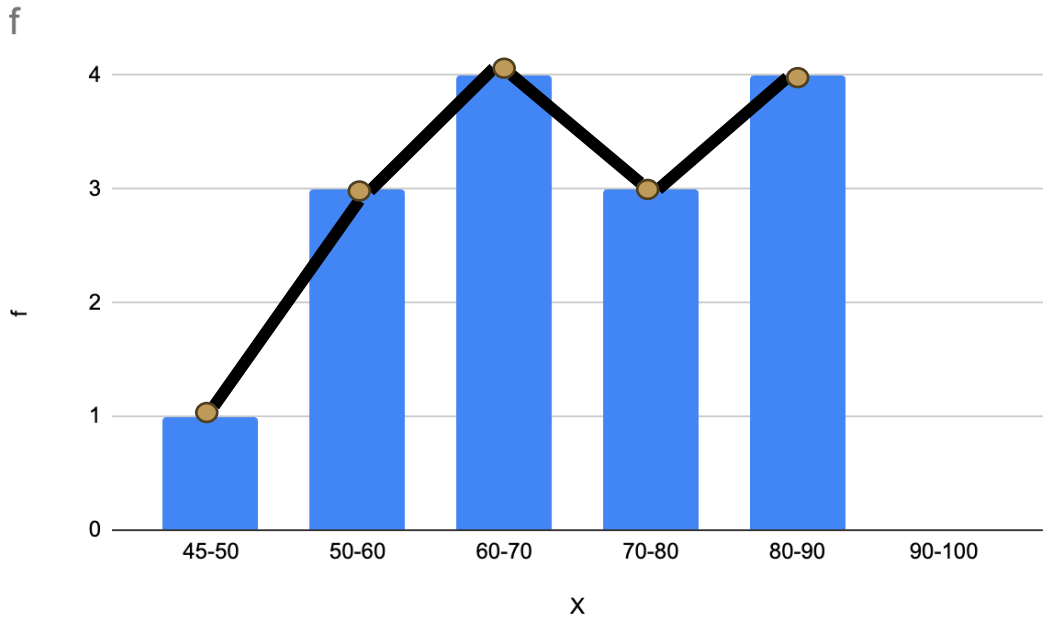
participant	X
A	61
B	63
C	73
D	53
E	66
F	52
G	86
H	82
I	50
J	65
K	55
L	75
M	88
N	90
O	80

X	f
45-50	1
50-60	3
60-70	4
70-80	3
80-90	4
90-100	0



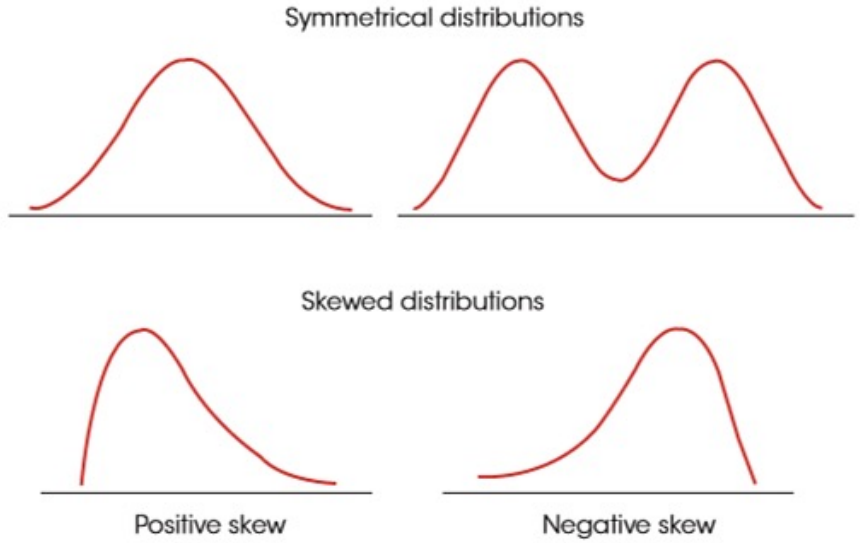
frequency polygons

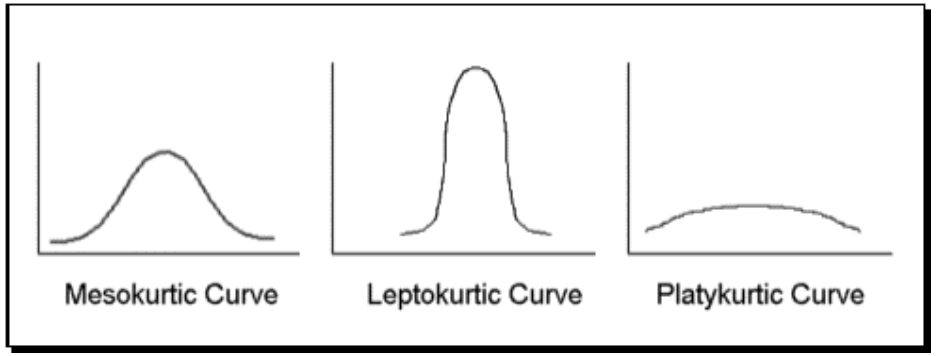
- contains the same data as a frequency histogram or table



shapes of distributions

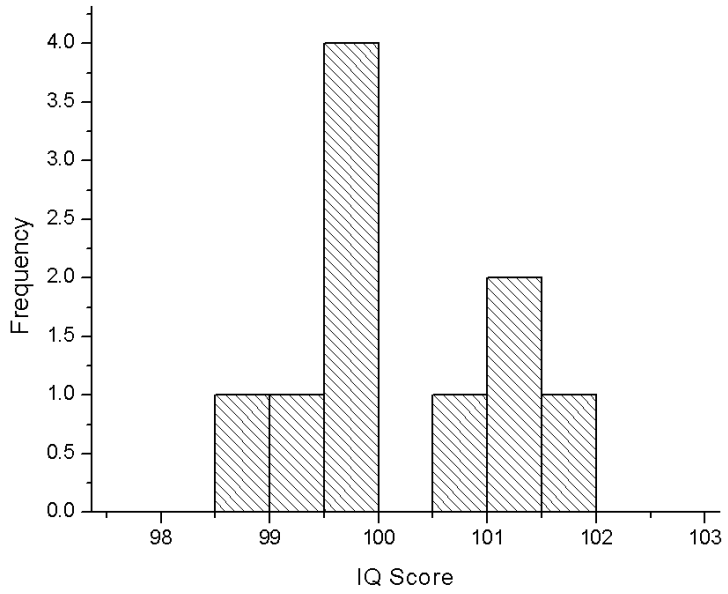
- populations are often displayed using smooth curves
- distributions are typically described along three dimensions
  - shape (symmetric, skewed, etc.)
  - central tendency (unimodal, bimodal, etc.)
  - variation/tailedness (kurtosis)





shapes of distributions

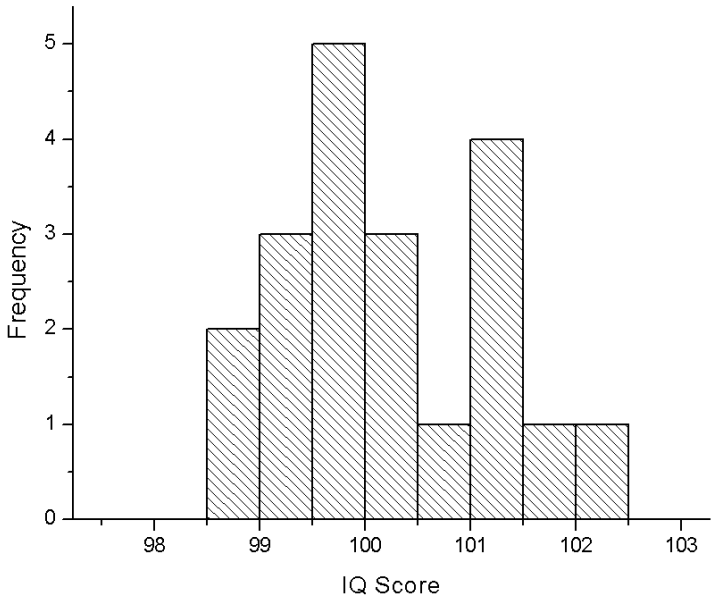
- distributions of larger samples tend to be smoother



n = 10

shapes of distributions

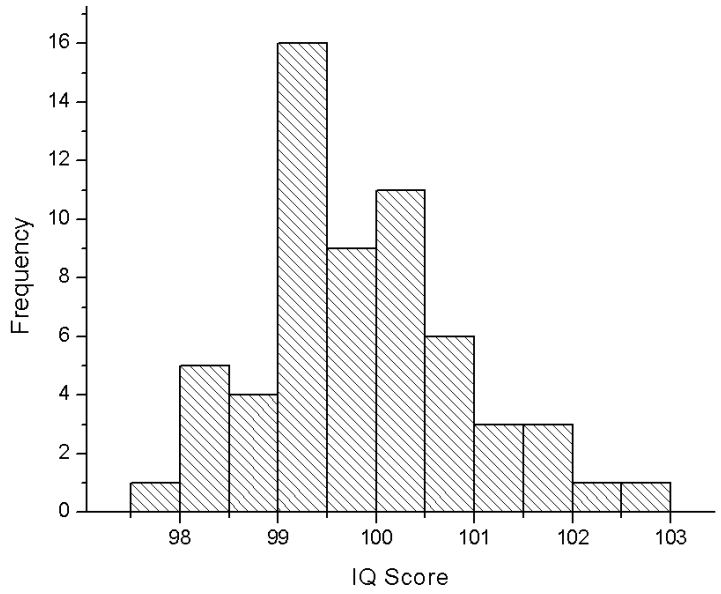
- distributions of larger samples tend to be smoother



n = 20

shapes of distributions

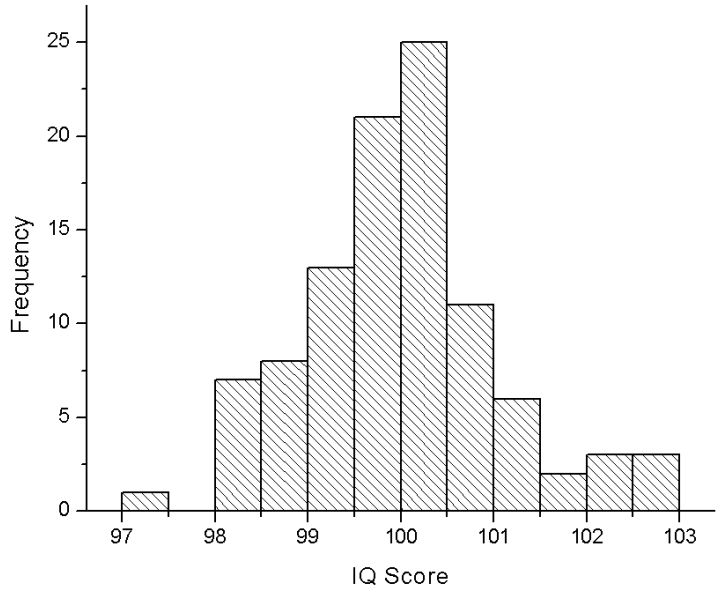
- distributions of larger samples tend to be smoother



n = 60

shapes of distributions

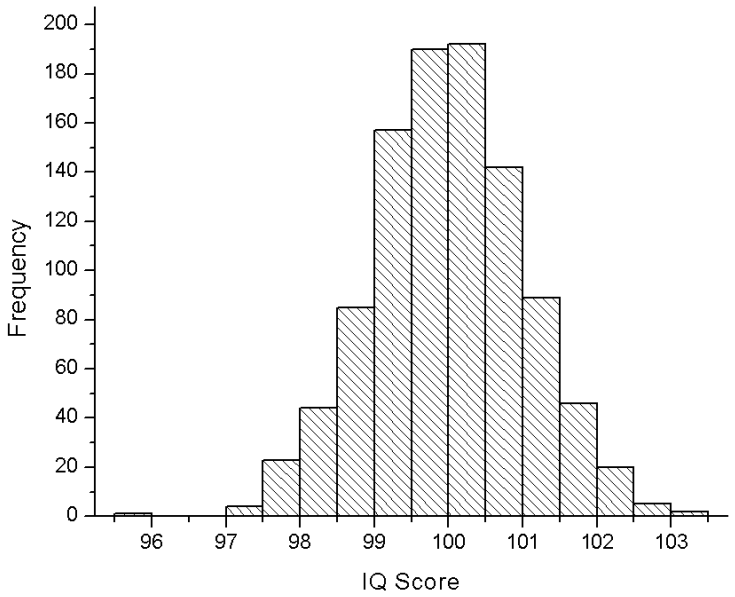
- distributions of larger samples tend to be smoother



n = 100

shapes of distributions

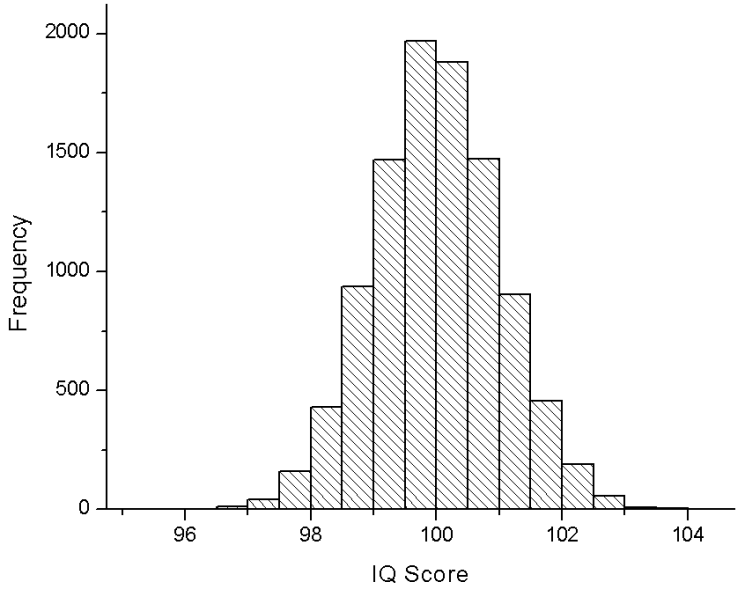
- distributions of larger samples tend to be smoother



n = 1000

shapes of distributions

- distributions of larger samples tend to be smoother

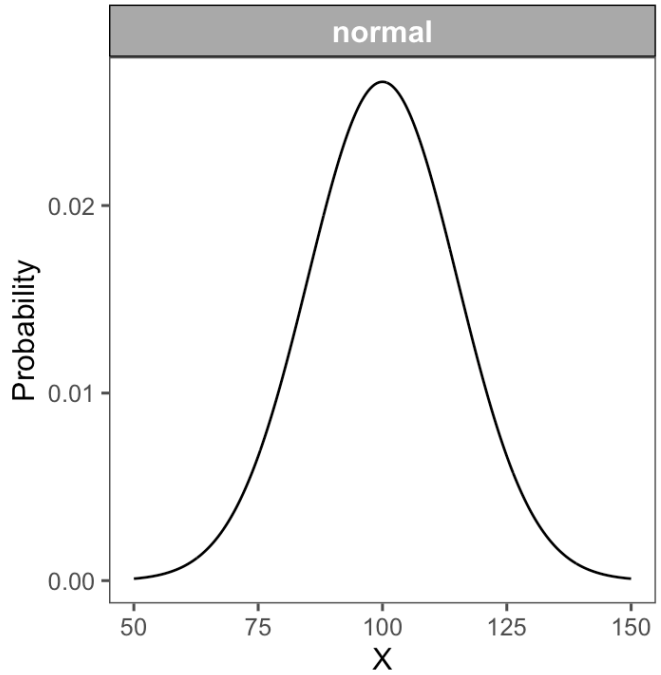


n = 10000

normal distribution

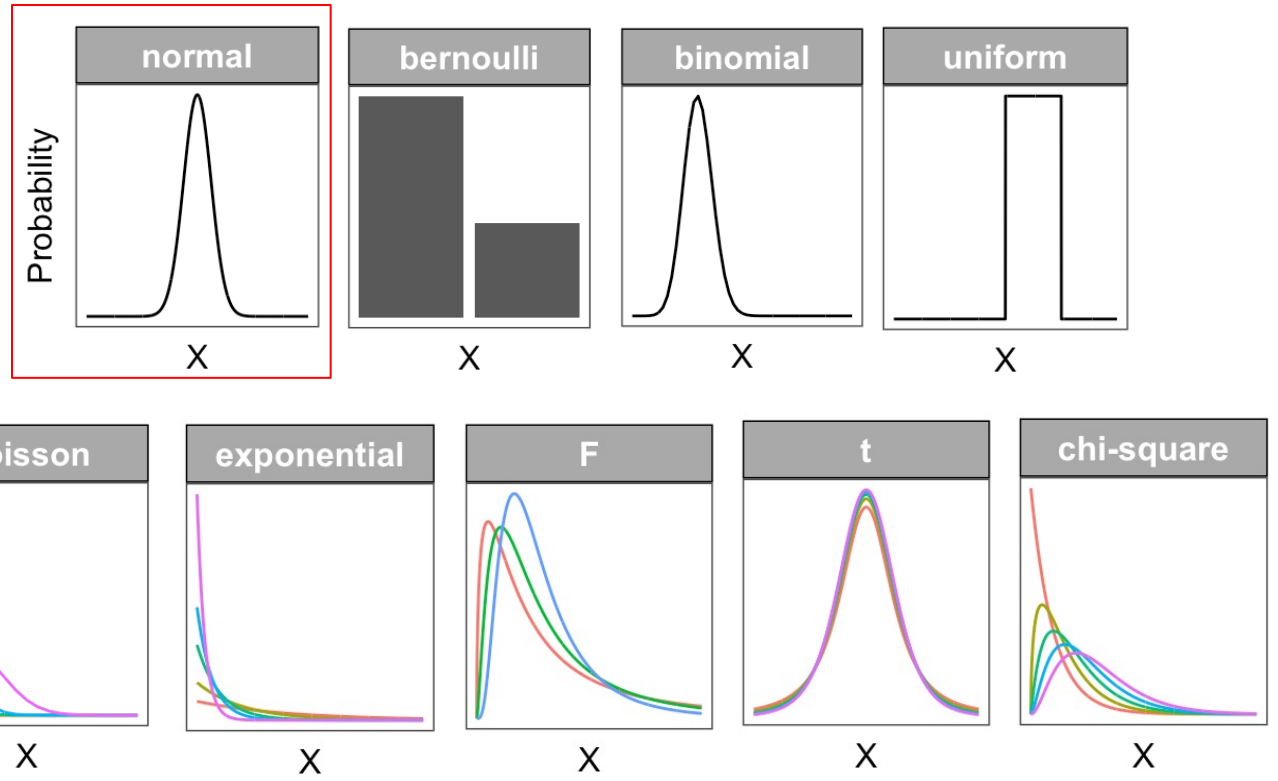
$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- 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)



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

- which score corresponds to the 88th percentile?
- which score corresponds to the 36th percentile?
- always use real limits

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

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

big takeaways from today

- jot down the key takeaways from today **without** looking at the slides/notes someplace you can revisit
- retrieval practice + elaborative encoding
- FYI: we are NOT covering stem and leaf plots

next time

- **before** class
  - *prep*: video tutorial: Summarizing data
  - *apply*: problem set 1 (chapter 2 problems)
  - *prep*: read Chapter 3 from the Gravetter & Wallnau (2017) textbook.
- **during** class
  - what is a model?
  - a framework for understanding data