# PSY 201: Statistics in Psychology

## Lecture 03
## Plots
### *Why the space shuttle blew up.*
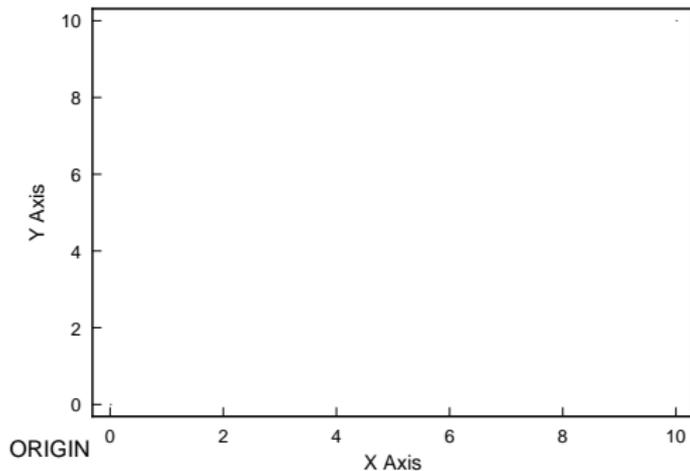
Greg Francis

Purdue University

Fall 2023

# DATA

GOAL:

- organize data in a way that helps us understand it
- often take advantage of visual interpretations
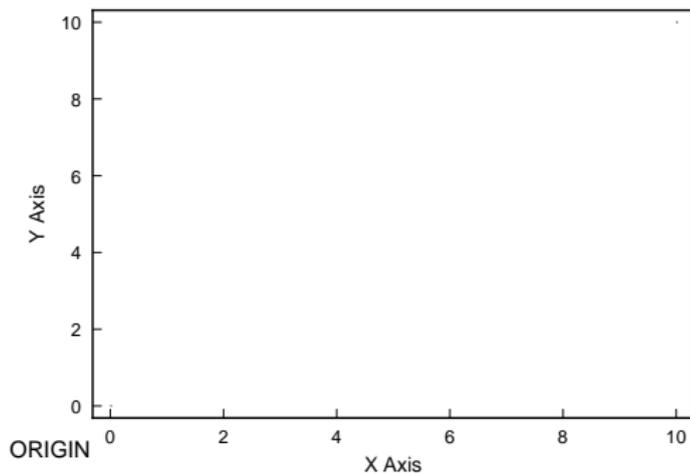- particularly important for very large sets of data

# GRAPHS



- plot one variable against another

# PLOTTING

- you make a graph to convey information
- place the dependent variable on the *y*-axis and the independent variable on the *x*-axis



- avoid everything else that might get in the way!

# SPACE SHUTTLE

- January 28, 1986
- O-ring leaked
- the Challenger exploded 59 seconds after liftoff

# SPACE SHUTTLE

- January 28, 1986
- O-ring leaked
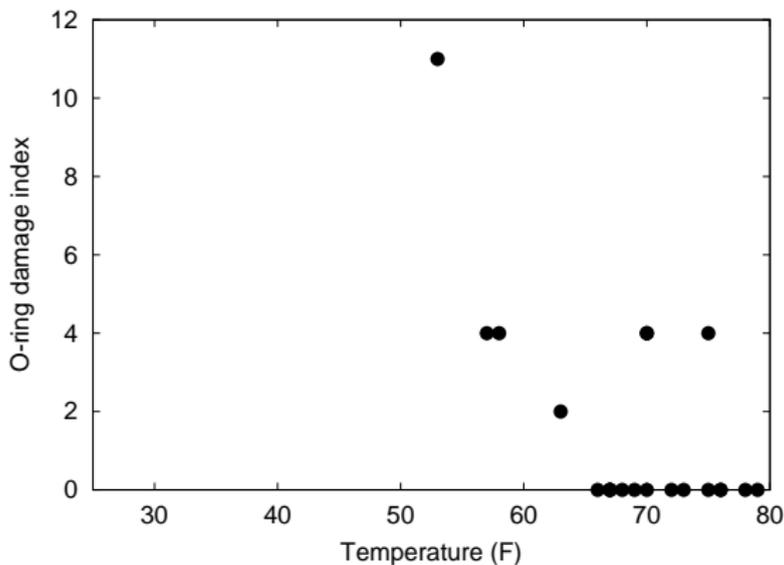- the Challenger exploded 59 seconds after liftoff

# SPACE SHUTTLE

- the night before engineers warned O-rings would leak in cold ($29°$) weather
- the engineers failed to make their case, and the shuttle blew up
- they failed to *present* their data in a way to convince others

# THE DATA

- previous launches showed damage to the O-rings increased as temperature got colder

# THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
  - used tables (not bad by itself, but a graph is often more convincing)



HISTORY OF O-RING TEMPERATURES
(DEGREES - F)

| MOTOR | MBT | AMB | O-RING | WIND |
|-------|-----|-----|--------|------|
| DM-4 | 68 | 36 | 47 | 10 MPH |
| DM-2 | 76 | 45 | 52 | 10 MPH |
| QM-3 | 72.5 | 40 | 48 | 10 MPH |
| QM-4 | 76 | 48 | 51 | 10 MPH |
| SRM-15 | 52 | 64 | 53 | 10 MPH |
| SRM-22 | 77 | 78 | 75 | 10 MPH |
| SRM-25 | 55 | 26 | 29 | 10 MPH |
|  |  |  | 27 | 25 MPH |

  - distributed information across several tables

# THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
    - cluttered graphics with irrelevant information (motor type, date of launch,...)



History of O-Ring Damage in Field Joints (Cont)
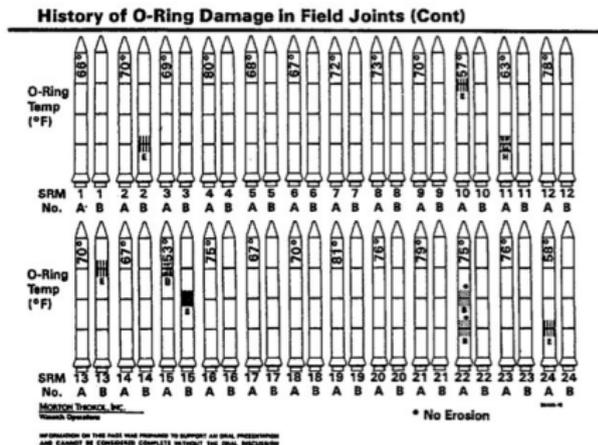
# THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
  - ▶ failed to point out that all good launches were in warm temperatures
  - ▶ failed to point out that the forecasted temperature ($29°$) was *much* colder than for any other launch (good or bad)
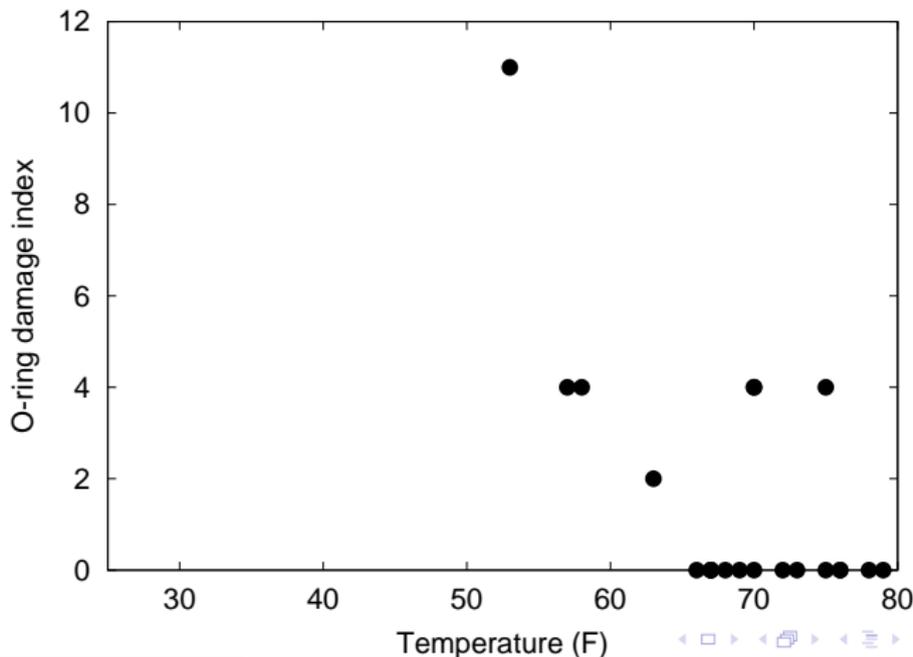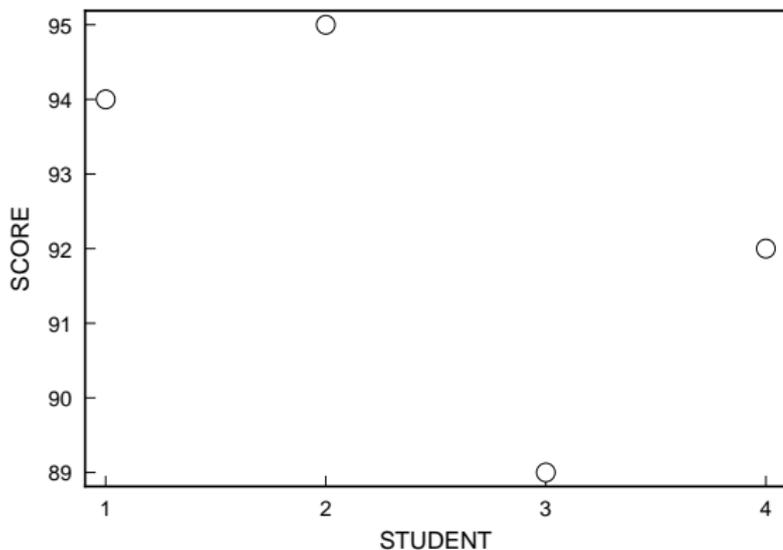
# THE LESSON

- when trying to convince someone of something, you must present it properly
- avoid fancy graphics and 3D perspectives
- keep it simple
- present the right information
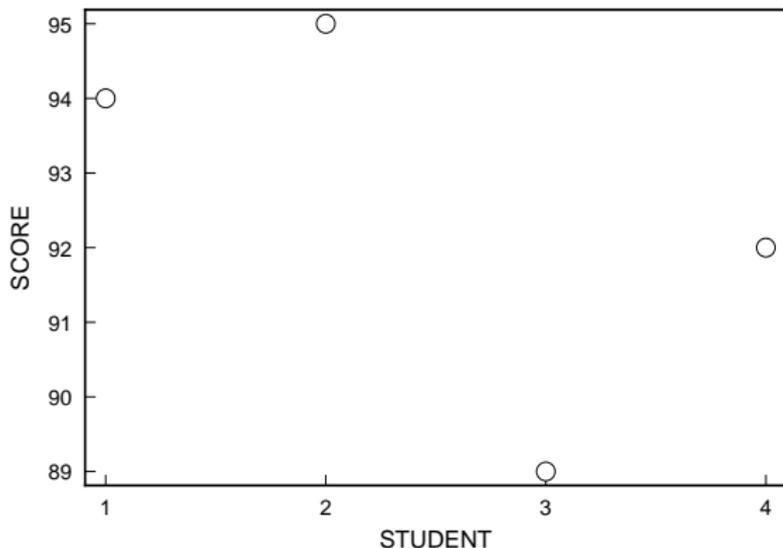- we will go over some basics of graphing...

# GRAPH

- using a small data set of four student's grades

# GRAPH

- using a small data set of four student's grades



- what measurement scale is the student variable?
- what measurement scale is the score variable?

# DATA CURVE

- it sometimes helps to connect the points
- How well did the third student do?
- changing the axis' scale makes the information look different, even though it isn't
- what matters is whether the graph conveys the intended information!
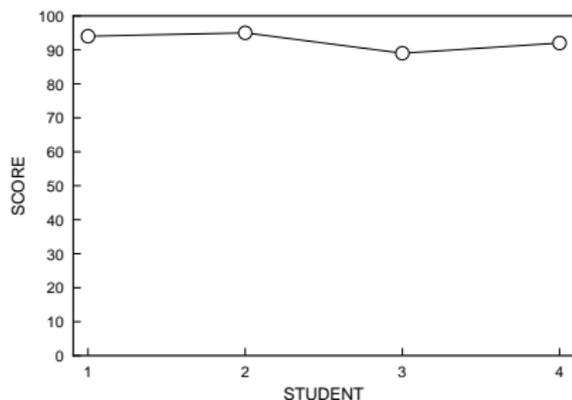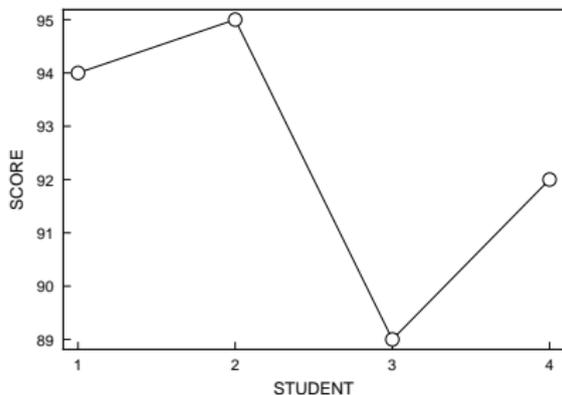
# GRAPH TYPE

- type of data often determines what type of graph to draw
- previous graph plotted ratio (or interval) data against nominal data
- consider the following data

| Make of Automobile | Repair Rate (per 1000 sold) |
|:---:|:---:|
| A | 4.2 |
| B | 6.8 |
| C | 3.3 |
| D | 0.4 |
| E | 1.2 |

- the graph should **not** suggest continuity of automobile make

# WHICH IS BETTER?
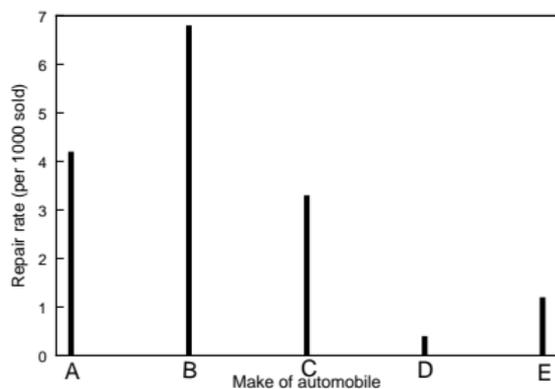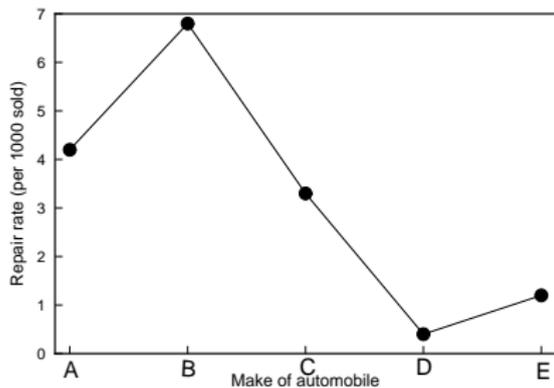
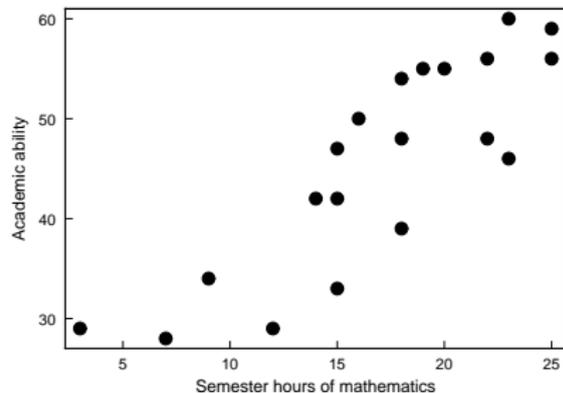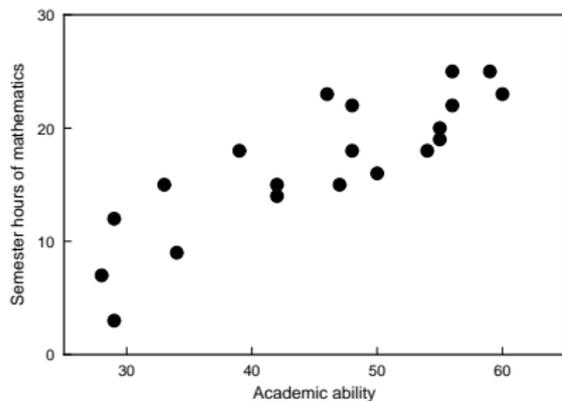# SCATTERGRAMS

- sometimes you want to look at co-occurrences of data

| Student | Academic Ability Score | Hours of Mathematics |
|---------|------------------------|----------------------|
| 1 | 54 | 18 |
| 2 | 29 | 3 |
| 3 | 42 | 14 |
| 4 | 60 | 23 |
| 5 | 33 | 15 |
| 6 | 28 | 7 |
| 7 | 56 | 22 |
| 8 | 48 | 18 |
| ... | ... | ... |

# SCATTERGRAMS

# GRAPHS

- Very useful for giving an overview of many types of data sets
- Useful for identifying trends in the data and relationships between variables
- Limited in that they depend on the viewer's interpretive abilities and sometimes graphs breakdown for really big or really small data sets
- We prefer more quantitative approaches

# FREQUENCY

- for large data sets we cannot present all the scores
- we often look at the *number* or *frequency* of scores within certain limits
- we look at how scores are spread out across different values
- this reduces the number of *presented* scores and improves understanding

# CLASS INTERVAL

Terminology

- width: exact upper limit - exact lower limit
- midpoint: value halfway between upper limit and lower limit
- exact limits: exact boundaries of interval
  - matter when we start to work with frequency distributions!
- score limits: highest and lowest possible scores that fall in the interval
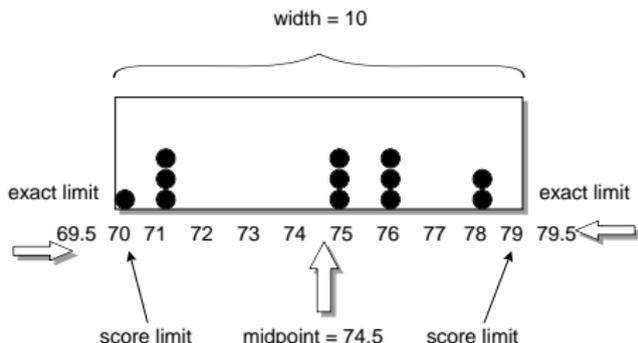
# FREQUENCIES

- compare a set of scores
  95, 22, 45, 45, 12, 79, 83, 46, 89, 96, 75, 33, 86, 57, 69, 94, 83, 75,
  77, 88, 92, 85, 31, 69

- to frequencies

  | Class Interval | $f$ |
  |---|---|
  | 10–19 | 1 |
  | 20–29 | 1 |
  | 30–39 | 2 |
  | 40–49 | 3 |
  | 50–59 | 1 |
  | 60–69 | 2 |
  | 70–79 | 4 |
  | 80–89 | 6 |
  | 90–99 | 4 |

# FREQUENCIES

- ADVANTAGES
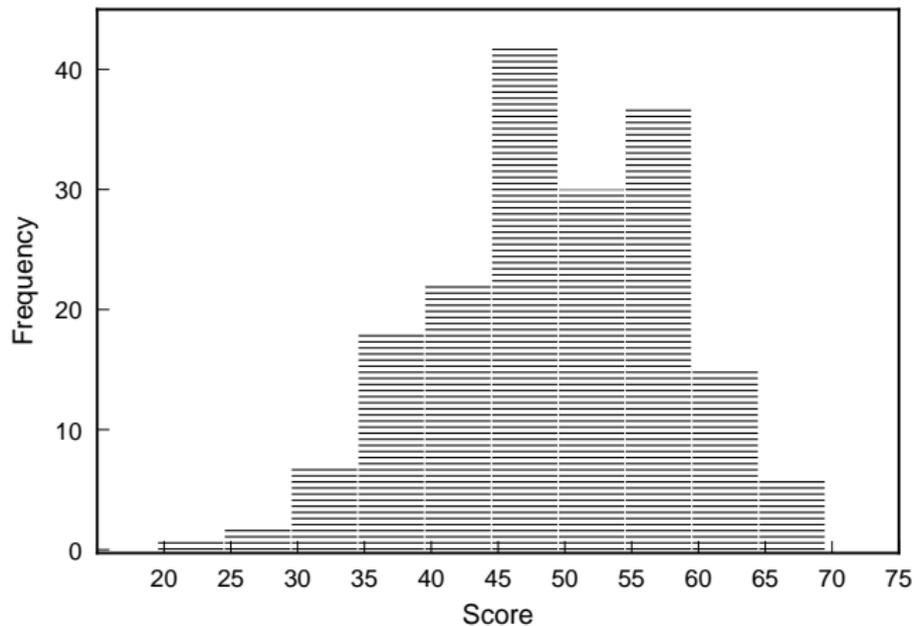  - easier to see distribution of scores
  - easier to interpret data
- DISADVANTAGES
  - loss of information
  - individual scores are missing
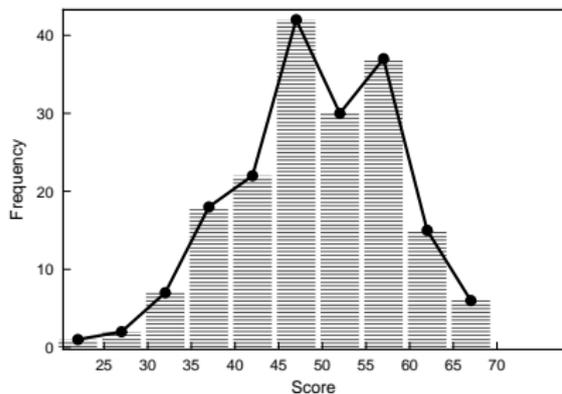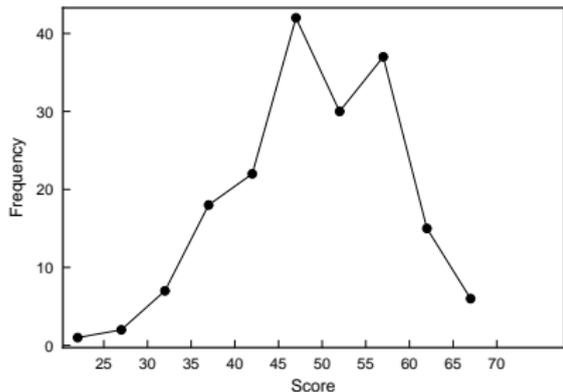  - midpoint score is often best guess
- often use frequency information to **supplement** other information (depends on your needs)

# HISTOGRAMS

frequency versus score class interval
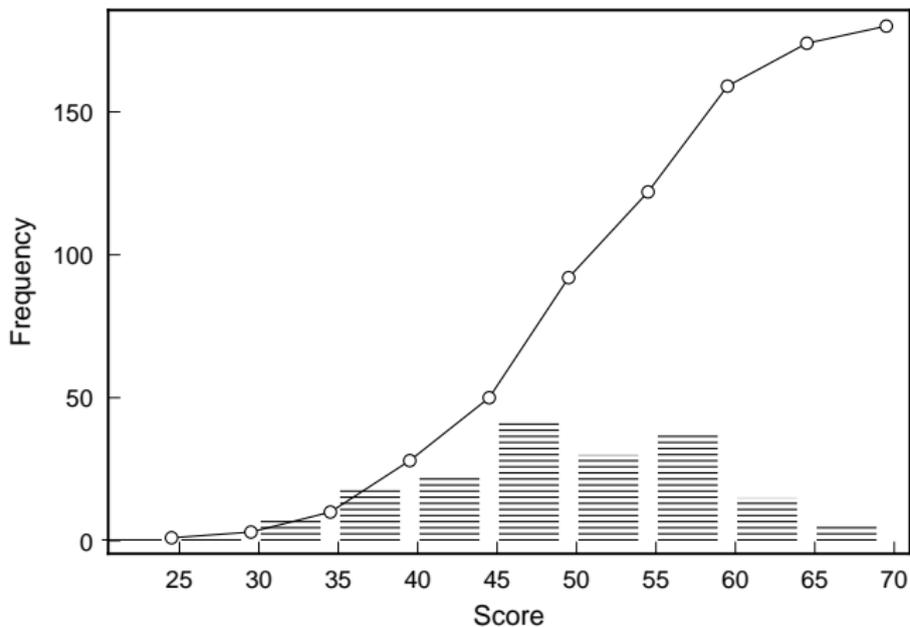
# FREQUENCY POLYGON

# CUMULATIVE FREQUENCIES

- frequency distribution tells us how many scores in each class interval
- cumulative frequency distribution tells us how many scores in all class intervals below a specific score

| Midpoint | f | cf |
|---|---|---|
| 67 | 6 | 180 |
| 62 | 15 | 174 |
| 57 | 37 | 159 |
| 52 | 30 | 122 |
| 47 | 42 | 92 |
| 42 | 22 | 50 |
| 37 | 18 | 28 |
| 32 | 7 | 10 |
| 27 | 2 | 3 |
| 22 | 1 | 1 |

# CUMULATIVE FREQUENCY DISTRIBUTION



Note: the point on the polygon has it's x-coordinate at the upper limit of the corresponding class interval

# PERCENTAGES

$$\% = \frac{\text{frequency}}{\text{total number of scores}}$$

$$c\% = \frac{\text{cumulative frequency}}{\text{total number of scores}}$$

| Midpoint | f | cf | % | c% |
|----------|-----|-----|-------|-------|
| 67 | 6 | 180 | 3.33 | 100 |
| 62 | 15 | 174 | 8.33 | 96.67 |
| 57 | 37 | 159 | 20.56 | 88.34 |
| 52 | 30 | 122 | 16.67 | 67.78 |
| 47 | 42 | 92 | 23.33 | 51.11 |
| 42 | 22 | 50 | 12.22 | 27.78 |
| 37 | 18 | 28 | 10.00 | 15.56 |
| 32 | 7 | 10 | 3.89 | 5.56 |
| 27 | 2 | 3 | 1.11 | 1.67 |
| 22 | 1 | 1 | 0.56 | 0.56 |

# OGIVE

- plot cumulative frequency percentage against upper score class interval
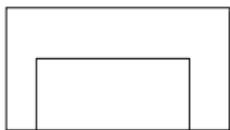- gives percentile points (next time)
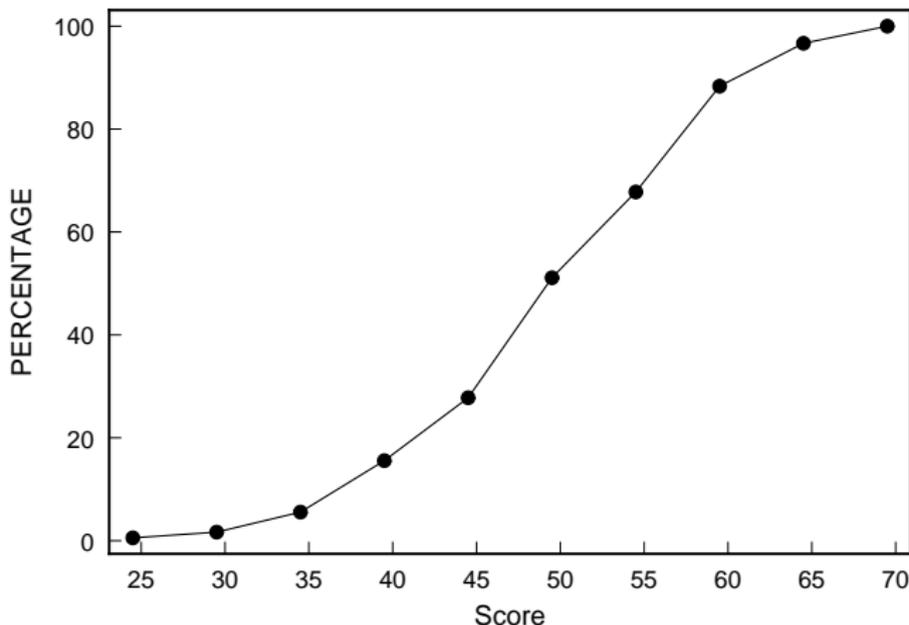
# FREQUENCY DISTRIBUTIONS

- useful to compare shapes
- any shape is possible
- some shapes are particularly important
  - uniform distribution
  - skewed distribution (long tail)
  - symmetric distribution
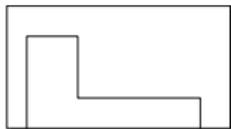  - normal distribution
  - kurtosis (peakedness)

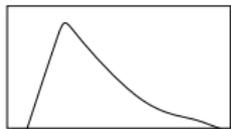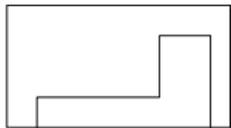# DISTRIBUTIONS



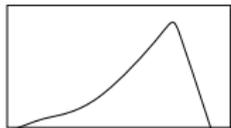UNIFORM DISTRIBUTION

SYMMETRIC
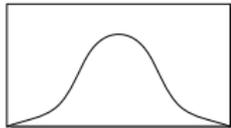
POSITIVE SKEW (RIGHT)
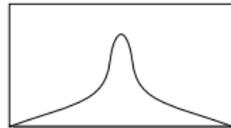
POSITIVE SKEW (RIGHT)

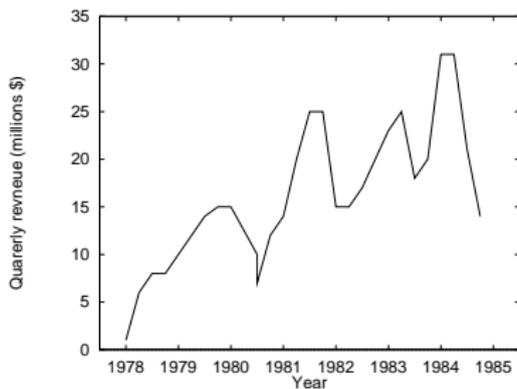NEGATIVE SKEW (LEFT)

NEGATIVE SKEW (LEFT)

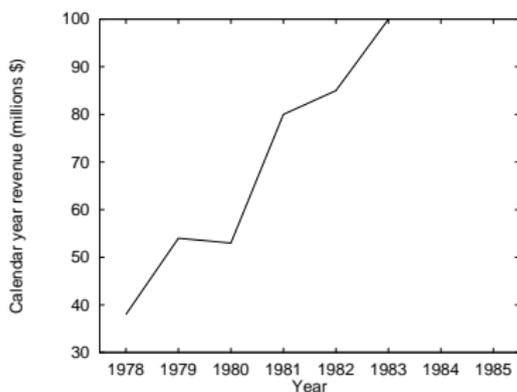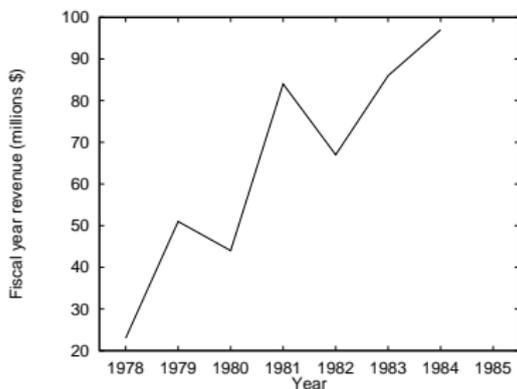NORMAL DISTRIBUTION

HIGH KURTOSIS

SYMMETRIC

# DISTRIBUTIONS

- with large data sets you *have* to group data together to make it manageable
- how you do it can sometimes have a profound effect on what people conclude
- consider revenue from a company: grouped by *quarterly* revenue

# DISTRIBUTIONS

- now look at the data when grouped by fiscal or calendar year

# DISTRIBUTIONS

- with computers people can now sift through huge amounts of data and present only those graphs that support what they want you to think
- a suspicious person might presume that the graphs you *do* see are the *best possible* for advancing the presenter's view
- the only way out of this is to either trust the presenter, or have access to the data and and knowledge to understand it

# HONESTY

- so how you define class intervals can determine how you (or someone else) will interpret the data
- statistics don't lie (they are just numbers)
- but you could (and some people do) select certain statistics to make people believe one thing versus another
- the only thing you can do about this effect is to be aware that it exists
- you need to be aware of the limitations of the data and be on guard against things that might influence you

# CONCLUSIONS

- graphing
- frequencies
- distributions
- remember: the goal is to correctly present information

# NEXT TIME

- percentiles
- percentile ranks

*How to score the SAT.*