# PSY 201: Statistics in Psychology

## Lecture 04
## Describing distributions
## *How to score the SAT.*

Greg Francis

Purdue University

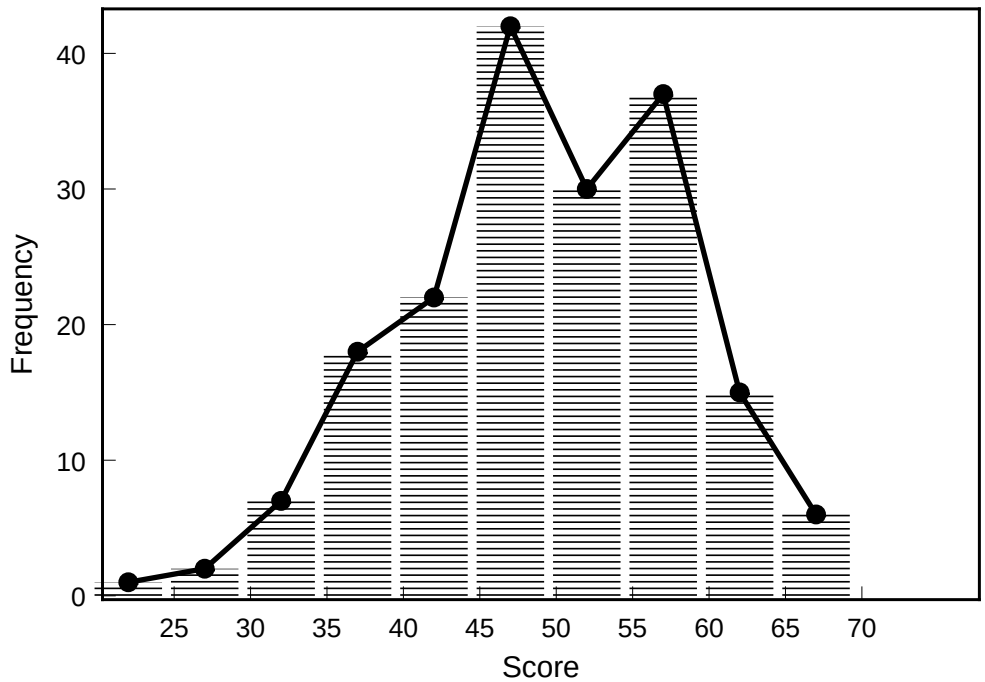
Fall 2019

# DISTRIBUTIONS

- As we saw last time, a well-drawn graph conveys a lot of useful information...
- but a poorly drawn graph can mislead and confuse.
- We would like a **quantitative** method of describing distributions
- may not entirely avoid misinformation, but at least the limitations will be identifiable

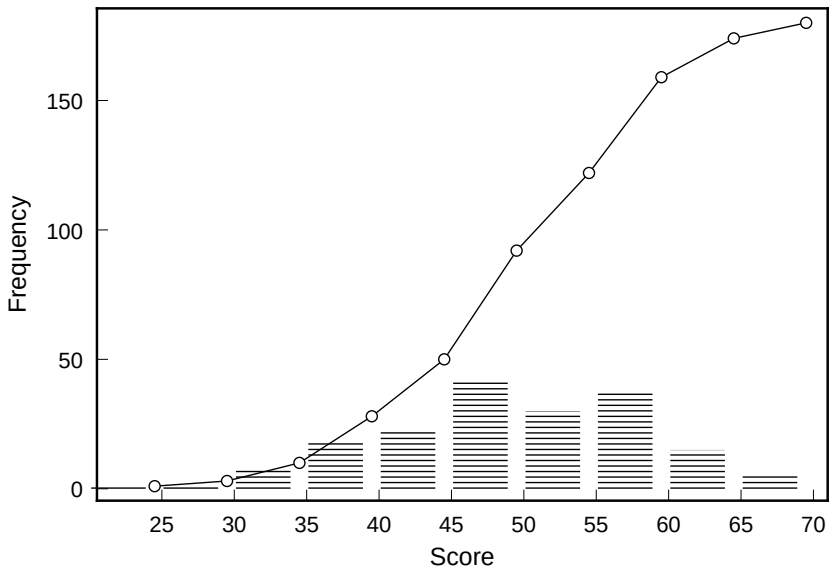
# FREQUENCY DISTRIBUTIONS

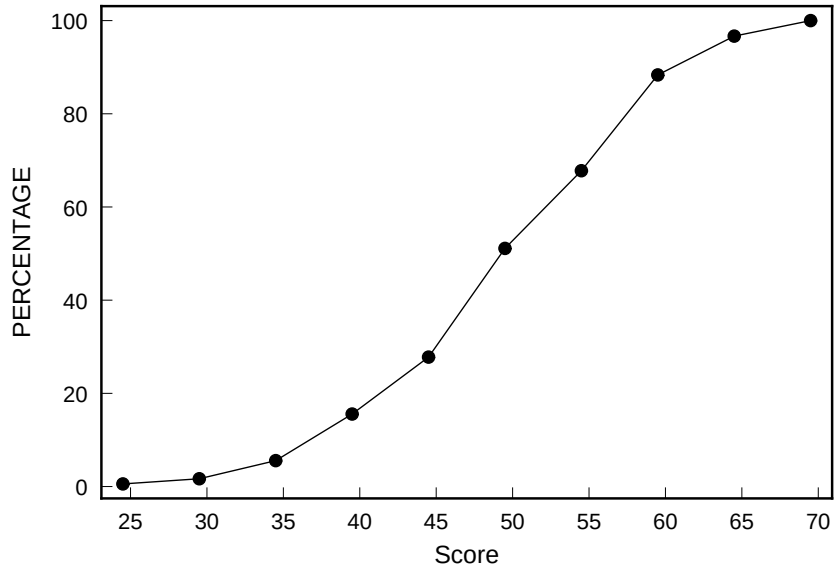
- A data set of exam scores can be described in many ways
  - frequency versus score class interval



# CUMULATIVE

- A data set of exam scores can be described in many ways
  - cumulative distributions



# TABLE FORMAT

- A data set of exam scores can be described in many ways
  - frequency table

| Exact Limits | Midpoint | f | cf | % | c% |
|---|---|---|---|---|---|
| 64.5–69.5 | 67 | 6 | 180 | 3.33 | 100 |
| 59.5–64.5 | 62 | 15 | 174 | 8.33 | 96.67 |
| 54.5–59.5 | 57 | 37 | 159 | 20.56 | 88.34 |
| 49.5–54.5 | 52 | 30 | 122 | 16.67 | 67.78 |
| 44.5–49.5 | 47 | 42 | 92 | 23.33 | 51.11 |
| 39.5–44.5 | 42 | 22 | 50 | 12.22 | 27.78 |
| 34.5–39.5 | 37 | 18 | 28 | 10.00 | 15.56 |
| 29.5–34.5 | 32 | 7 | 10 | 3.89 | 5.56 |
| 24.5–29.5 | 27 | 2 | 3 | 1.11 | 1.67 |
| 19.5–24.5 | 22 | 1 | 1 | 0.56 | 0.56 |

# DISTRIBUTION USES

- summarize data
- indicate most frequent data values
- indicate amount of variation across data values
- allows us to interpret a single score in the context of other scores
- we will explore quantitative methods to describe distributions

# PERCENTILE

- point in a distribution at (or below) which a given percentage of scores is found
- written as

$$P_{\text{percentage}}$$

- 28th percentile is written as $P_{28}$
- 99th percentile is written as $P_{99}$
- ...

# PERCENTILE

- what are the data values for the lowest 60% of the population?
- several steps
  1. Find out how many data values make up 60% of the population.
  2. Find the lowest class interval in the cumulative frequency distribution that includes at least that many data values.
  3. Estimate how far into the class interval you must go to reach exactly the percentile.
- works for any percentage!

# CALCULATIONS

- find $P_{60}$ using the above data set of scores
  (1) number of scores making up 60% of student scores is

$$(180)(0.60) = 108$$

  In general, calculate

$$(n)(p)$$

  where $n$ is the size of the population (number of scores) and $p$ is the percentage in decimal form

# CALCULATIONS

(2) lowest class interval in the *cf* including 108 scores is with midpoint 52

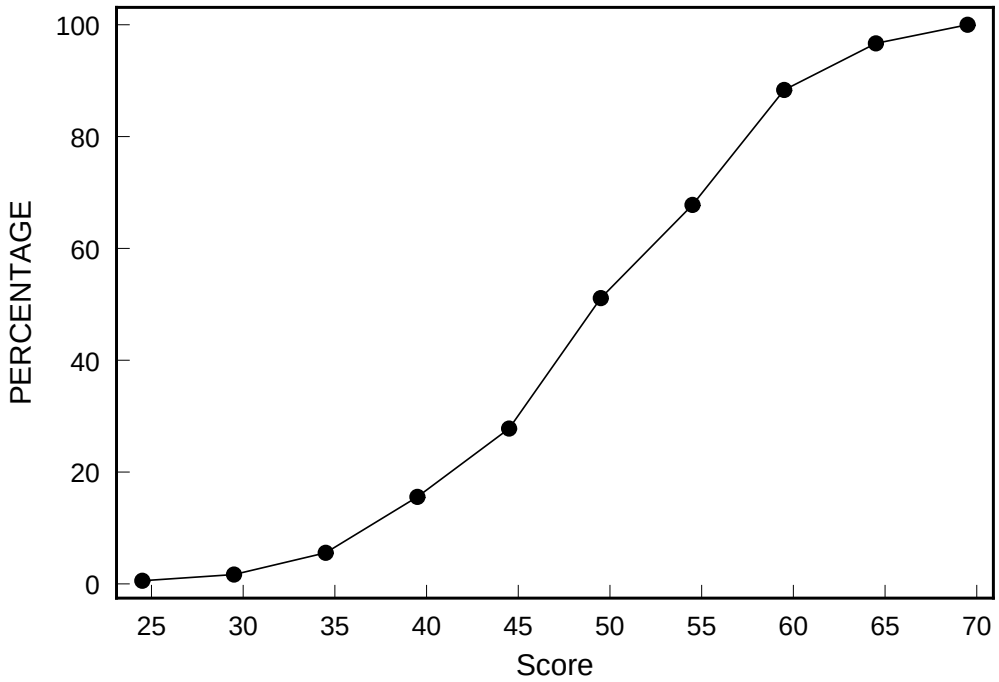
| Exact Limits | Midpoint | f | cf | % | c% |
|---|---|---|---|---|---|
| 64.5–69.5 | 67 | 6 | 180 | 3.33 | 100 |
| 59.5–64.5 | 62 | 15 | 174 | 8.33 | 96.67 |
| 54.5–59.5 | 57 | 37 | 159 | 20.56 | 88.34 |
| **49.5–54.5** | **52** | **30** | **122** | **16.67** | **67.78** |
| 44.5–49.5 | 47 | 42 | 92 | 23.33 | 51.11 |
| 39.5–44.5 | 42 | 22 | 50 | 12.22 | 27.78 |
| 34.5–39.5 | 37 | 18 | 28 | 10.00 | 15.56 |
| 29.5–34.5 | 32 | 7 | 10 | 3.89 | 5.56 |
| 24.5–29.5 | 27 | 2 | 3 | 1.11 | 1.67 |
| 19.5–24.5 | 22 | 1 | 1 | 0.56 | 0.56 |

# CALCULATIONS

- so we know that the percentile is somewhere between 49.5 and 54.5. We want a more precise **estimate**
- we need to know
  - width of class interval (5)
  - frequency of scores in the class interval containing the percentile point (30)
  - exact lower limit of class interval containing the percentile point (49.5)
  - *cf* of scores **below** the class interval containing the percentile point (92)
  - remaining number of scores in class interval containing the percentile point (108 - 92 = 16)

# CALCULATIONS

- estimate of percentile point
- go into the interval the remaining (unaccounted for) percentage



# CALCULATIONS

$$P_X = ll + \left( \frac{np - cf}{f_i} \right) (w)$$

- $ll$ = exact lower limit of the interval containing the percentile point
- $n$ = total number of scores
- $p$ = X/100, proportion corresponding to percentile (decimal form)
- $cf$ = cumulative frequency of scores **below** the interval containing the percentile point
- $f_i$ = frequency of scores **in** the interval containing the percentile point
- $w$ = width of class interval

# PERCENTILE RANK

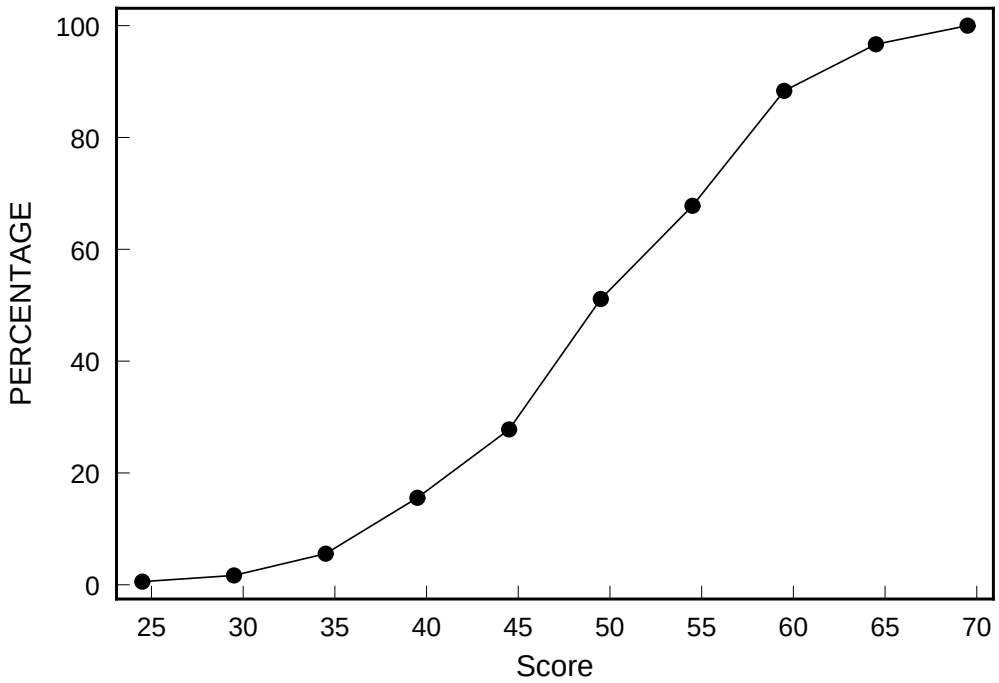
- given a particular data value, what percentage of data values are smaller?
- e.g. given a score on a test, what percentage of scores were lower?
- sort of the reverse of percentile
- for a data value of 39, we write the percentile rank as

$$PR_{39}$$

- (Used on achievement tests!)

# OGIVE

- plot cumulative frequency percentage against score class interval (gives percentile rank)



# CALCULATIONS

$$PR_X = \left\{ \frac{cf + (f_i)(X - ll)/w}{n} \right\} (100)$$

- $X$ = score for which percentile rank is to be determined
- $cf$ = cumulative frequency of scores **below** the interval containing the score $X$
- $ll$ = exact lower limit of the interval containing $X$
- $w$ = width of class interval containing $X$
- $f_i$ = frequency of scores in the interval containing $X$
- $n$ = total number of scores

# CALCULATIONS

$$PR_X = \left\{ \frac{cf + (f_i)(X - ll)/w}{n} \right\} (100)$$

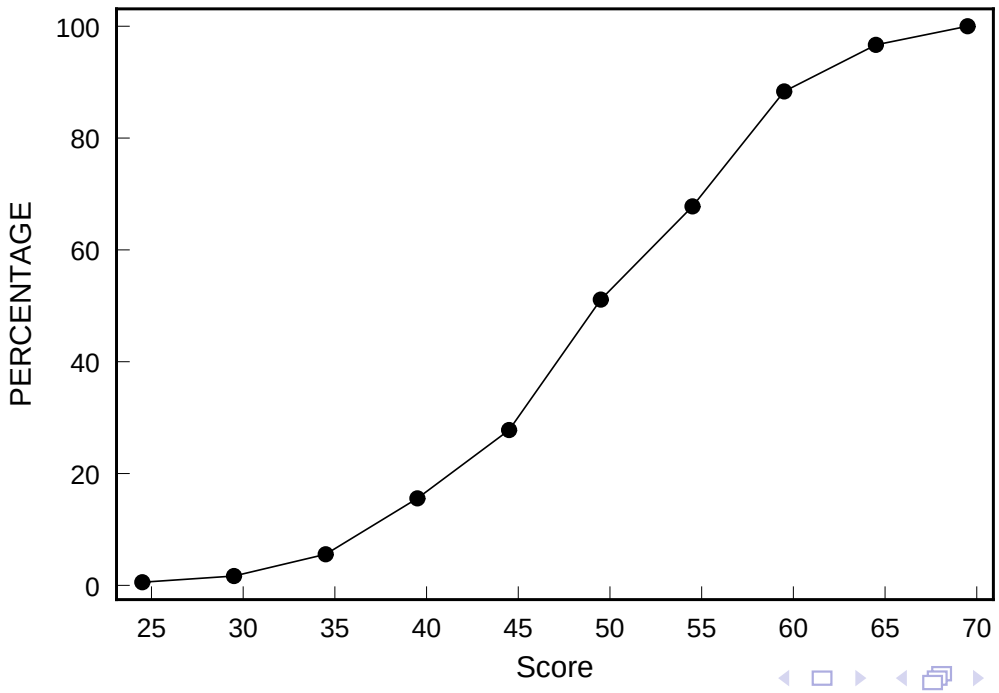
$$PR_{39} = \left\{ \frac{10 + (18)(39 - 34.5)/5}{180} \right\} (100)$$

$$PR_{39} = 14.556$$

| Exact Limits | Midpoint | f | cf | % | c% |
|---|---|---|---|---|---|
| 64.5–69.5 | 67 | 6 | 180 | 3.33 | 100 |
| 59.5–64.5 | 62 | 15 | 174 | 8.33 | 96.67 |
| 54.5–59.5 | 57 | 37 | 159 | 20.56 | 88.34 |
| 49.5–54.5 | 52 | 30 | 122 | 16.67 | 67.78 |
| 44.5–49.5 | 47 | 42 | 92 | 23.33 | 51.11 |
| 39.5–44.5 | 42 | 22 | 50 | 12.22 | 27.78 |
| **34.5–39.5** | **37** | **18** | **28** | **10.00** | **15.56** |
| 29.5–34.5 | 32 | 7 | *10* | 3.89 | 5.56 |
| 24.5–29.5 | 27 | 2 | 3 | 1.11 | 1.67 |
| 19.5–24.5 | 22 | 1 | 1 | 0.56 | 0.56 |

# LIMITATIONS

- percentiles help *describe* a data value relative to its frequency distribution
- but they have some drawbacks
  - percentiles use an ordinal scale
  - equal differences in percentiles do not indicate equal differences in raw scores!
  - class intervals with higher frequency cover a broader range of percentiles (steeper part of ogive)



# LIMITATIONS

- percentiles exaggerate differences in scores when lots of people have similar scores
- underestimate actual differences when lots of people have very different scores
- differences in percentiles should **not** be compared across different distributions!!!
  - only provide information on relative ranking of scores: ordinal scale!
  - cannot be meaningfully averaged, summed, multiplied,...
- fixing these problems requires additional terms for describing distributions (central tendency)

# CONCLUSIONS

- percentiles
- percentile ranks

# NEXT TIME

- central tendency
  - mode
  - median
  - mean

*Does a company deserve a tax break?*