

Panel 1

Least Times

frequency tables (categ vars)

histograms (numeric vars)

bar charts

pie charts

✓

1

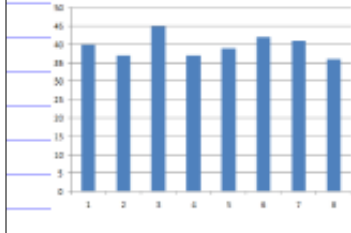
Panel 2

Homogeneous vs. Heterogeneous Distribution

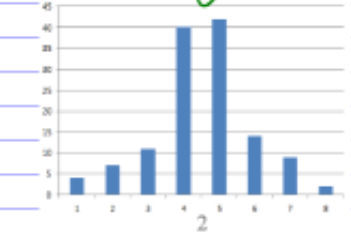
Homogeneous distr.: values cluster around one value

Heterogeneous distr.: all categories are represented equally, more or less

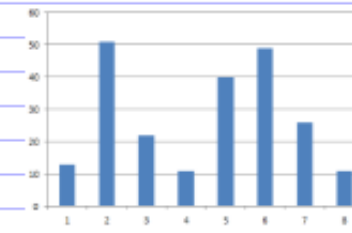
heterogeneous



homogeneous



neither



Panel 3

Numeric Data Representation: Measures of Central Tendency

① The mean (average):

add up all #'s,

divide by number of #'s

$$\frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

$$\bar{X} = \frac{1}{n} \sum x_i$$

← sigma, reads as "sum of"

or
 μ (mu)

reads as "x-bar"
 \bar{X} = sample mean,

μ = pop. mean

Ex: 1, 3, 5, 7 sample

Then the mean is $\bar{X} = \frac{1}{4}(1+3+5+7) = \frac{1}{4}16 = 4$

Panel 4

② Median = "middle"

That # s.t. 50% of data values are less and
50% are bigger than it.

Ex: 5, 3, 7 → 3, 5, 7

① Sort it ② Pick middle # or half in between

2, 6, 1, 5

→ 1, 2, 5, 6

Median $\frac{5+2}{2} = 3.5$

Panel 5

③ Mode

Data value that occurs most often.

Ex: 1, 3, 7, 2, 1, 5, 8 \Rightarrow mode = 1

3, 7, 5, 3, 5, 2, 1 \Rightarrow mode = 3 and 5

1, 2, 3, 4 \Rightarrow mode = 1, 2, 3, 4

Why 3 measures of central tendency

| | mean | median | mode |
|--------------|------|--------|------|
| interval var | ✓ | ✓ | ✓ |
| ordinal var | ✗ | ✓ | ✓ |
| nominal var | ✗ | ✗ | ✓ |

5

Panel 6

Which one is best?

Central measure of income in NYC:

Ex: { \$20,000, \$60,000, \$90,000 }

mode: all

median: 60K

mean: \$56,667

By chance you hit on a rich guy:

{ \$20,000, \$60,000, \$1,200,000 }

mode: all

median: 60K

mean: \$426,000

Median is more robust, mean is impacted by extreme outliers \Rightarrow Median is preferred

6

Panel 7

Measures of Central Tendency for Tables:

male 60%
female 40%

mode is that category that occurs most often (male)

| | codes | (rel) freq. | cumulative % |
|-----------|-------|-------------|--------------|
| very good | 1 | 0.15 | 0.15 |
| good | 2 | 0.2 | 0.35 |
| avg | 3 | 0.3 | 0.65 |
| bad | 4 | 0.2 | 0.85 |
| very bad | 5 | 0.15 | 1.0 |

mode: avg

median: is that cat. where cumm. % is above 50% for the first time.

avg

7

Panel 8

| | code | count | code · count |
|-----------|------|-------|--------------|
| very good | 1 | 10 | 10 · 1 = 10 |
| good | 2 | 30 | 30 · 2 = 60 |
| avg. | 3 | 40 | 40 · 3 = 120 |
| bad | 4 | 30 | 30 · 4 = 120 |
| very bad | 5 | 10 | 10 · 5 = 50 |
| | | 120 | 360 |

avg: $\frac{360}{120} = \underline{\underline{3.0}}$

avg = $\frac{\text{sum}(\text{code} \cdot \text{count})}{\text{sum}(\text{count})}$

8

Panel 9

| | count | freq. | cum. % | code | code count |
|------|-----------|-------|--------|-----------|------------|
| 1-4 | 4 | 0.4 | 0.4 | 2.5 | 10 |
| 4-7 | 3 | 0.3 | 0.7 | 5.5 | 16.5 |
| 7-10 | 3 | 0.3 | 1.0 | 8.5 | 25.5 |
| | <u>10</u> | | | <u>52</u> | |

1, 5, 9, 3, 3, 6, 8, 3, 5, 10

Row data

average = $\frac{53}{10} = 5.3$

mode = 3

median = $\frac{5+5}{2} = 5$

1, 3, 3, 3, 5, 5, 6, 8, 9, 10

Mode: 1-4

Median: 4-7

Mean: $\frac{52}{10} = 5.2$

Homogeneous or heterogeneous distribution?

Panel 10

| Frequency table results for HIGHEST DEGREE: | | | cum % | code = freq |
|---|-----------|--------------------|-------|----------------|
| HIGHEST DEGREE | Frequency | Relative Frequency | | |
| 0 - Less than HS | 297 | 0.14688428 | | |
| 1 - High School | 1003 | 0.49604353 | | |
| 2 - Junior College | 173 | 0.085558854 | | |
| 3 - Bachelor | 355 | 0.17556074 | | |
| 4 - Graduate | 194 | 0.095944606 | | |

Median:

Mean:

Mode: HS

homog or heterog.

HLW