Panel 1

Last Time

frequency tables (categ vars)

histograms (numeric vars)

bar charts

pie charts

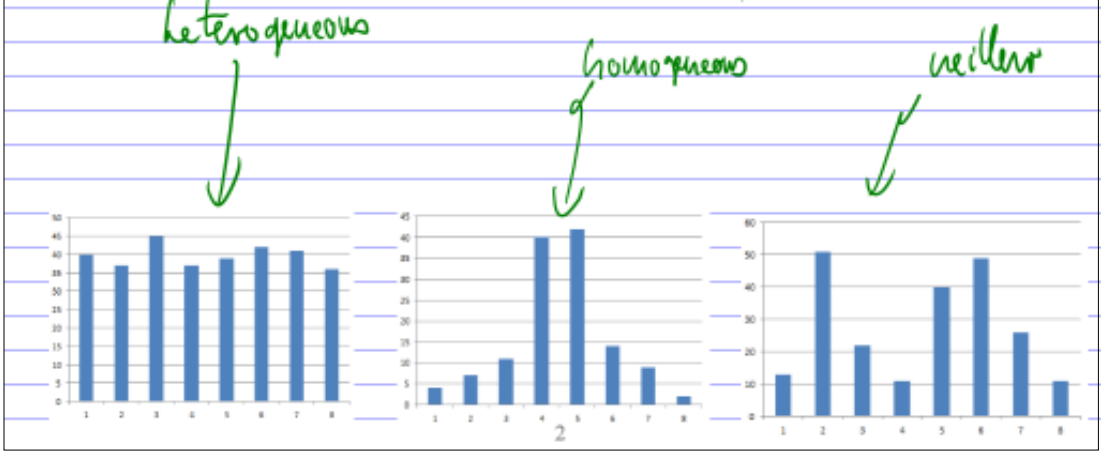
1

Panel 2

Homogeneous vs. Heterogeneous Distribution

Homogeneous distr: values cluster around one value

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



2

Panel 3

Numeric Data Representation: Measures of Central Tendency

① The mean (average):

add up all #'s,

divide by number of #'s

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

$\bar{x} = \frac{1}{n}\sum x_i$ ← sigma, reads as "sum of"

or

μ (mu)

$\bar{x}$ 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 $\frac{5+2}{2} = 3.5$

4

Panel 5

③ Mode

Data value that occurs most often.

Ex: 1, 3, 7, 2, 1, 5, 8 ⇒ mode = 1

3, 7, 5, 3, 5, 2, 1 ⇒ mode = 3 and 5

1, 2, 3, 4 ⇒ mode = 1, 2, 3, 4

Why 3 measures of central tendency

	mean	median	mode
numeric 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.

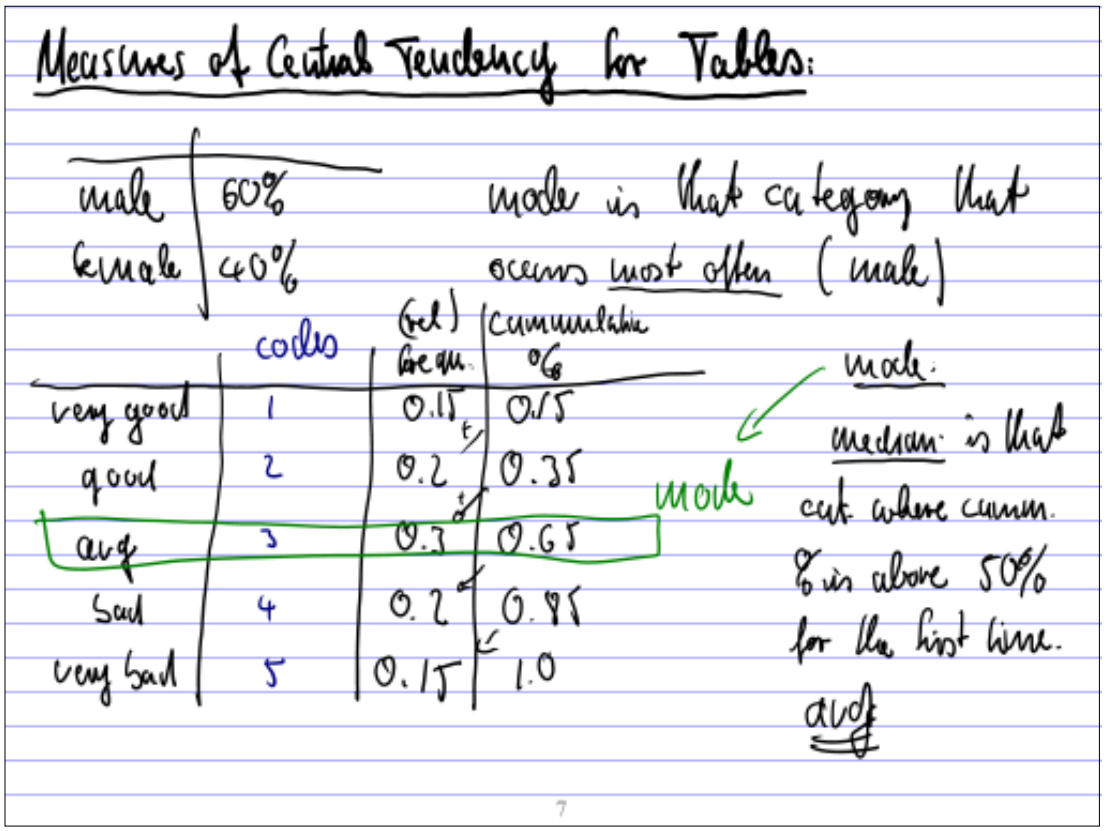
{$20,000, $60,000, $1,200,000}

mode: all
median: 60K
mean: $426,000

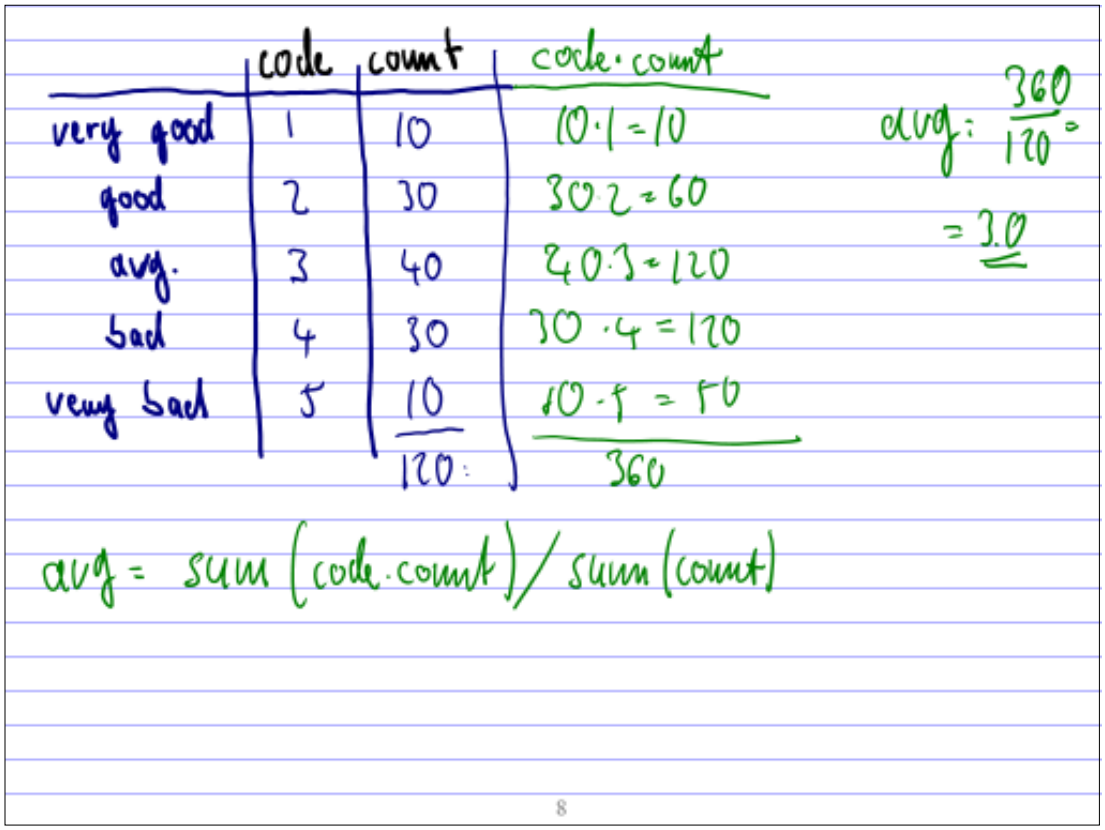
Median is more robust, mean is impacted by extreme outliers ⇒ Median is preferred

6

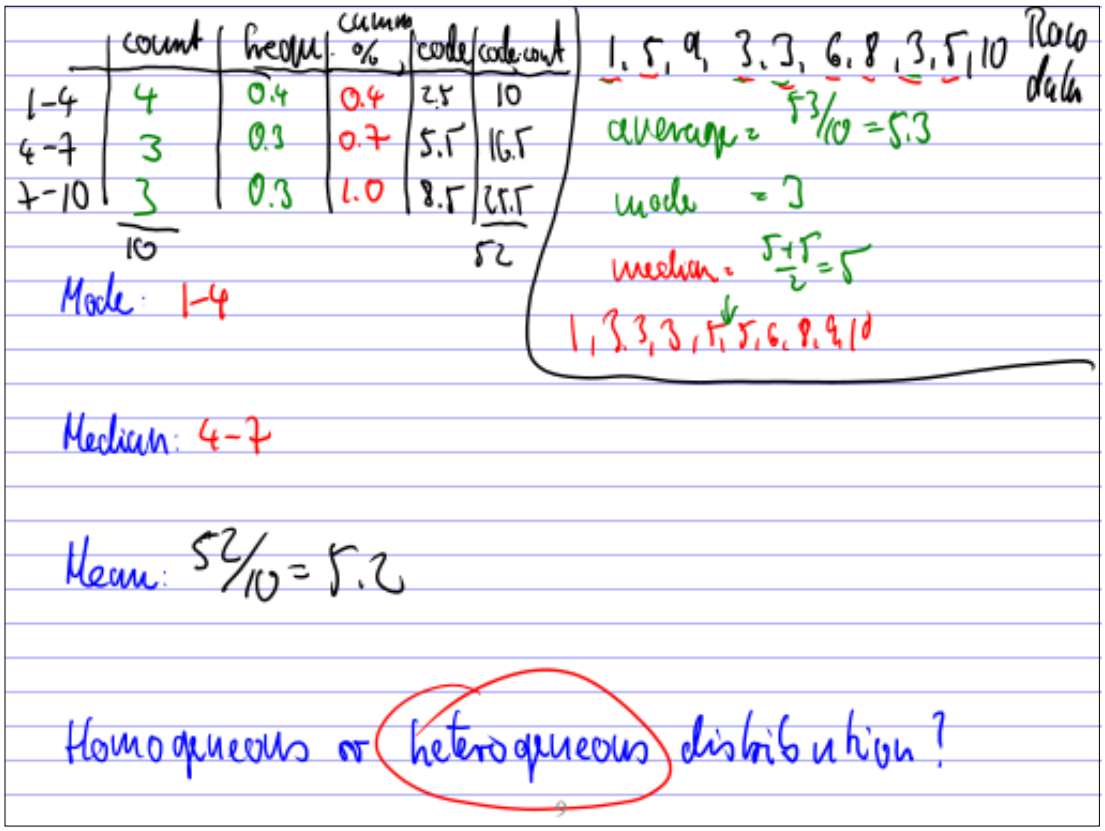
Panel 7



Panel 8



Panel 9



Panel 10