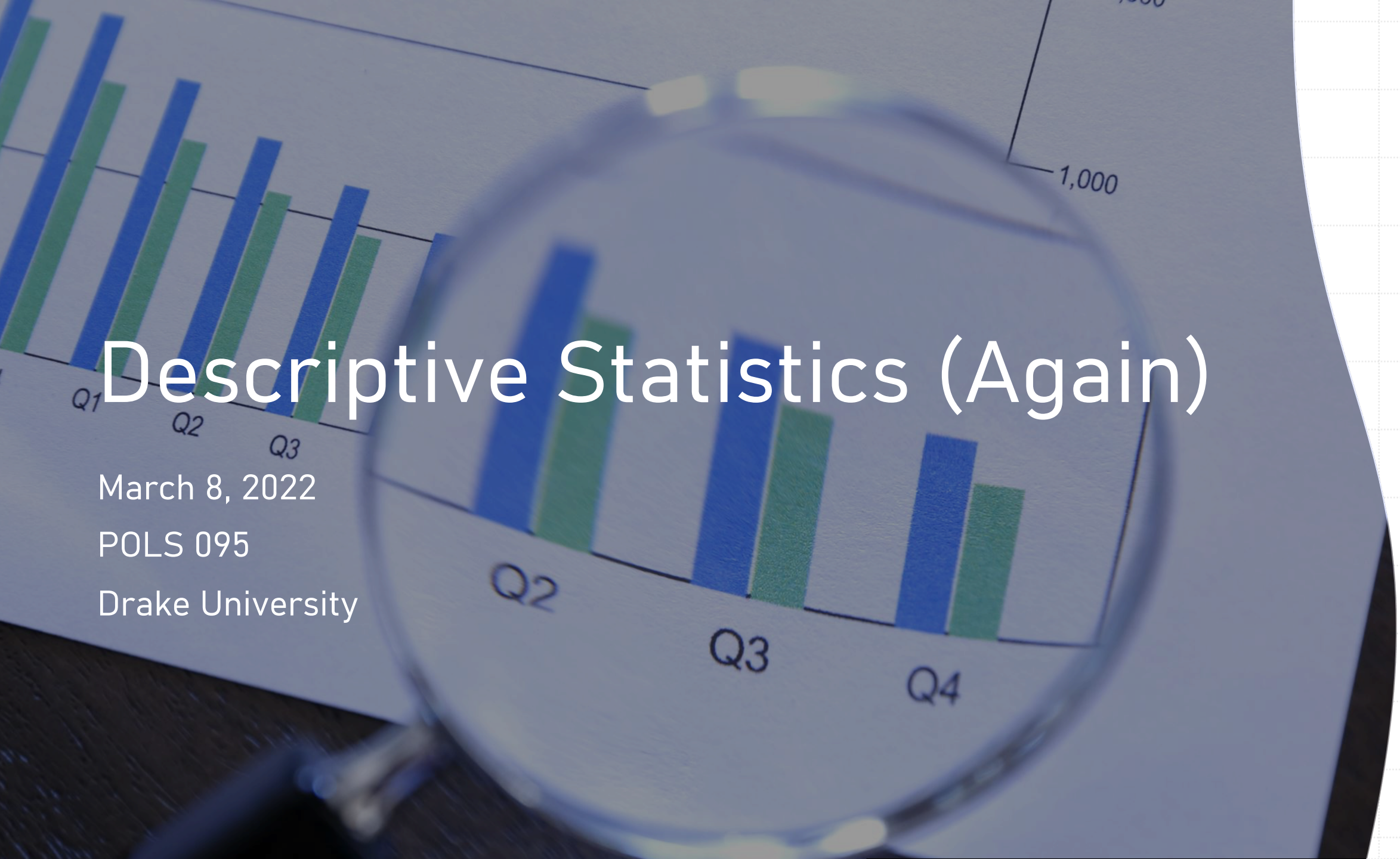


Descriptive Statistics (Again)

March 8, 2022

POLS 095

Drake University

Review: Variables and Measurement

- Variable: measurement of a concept that varies across subjects in a population of subjects
- Different levels of measurement:
  - Nominal
    - (male/female, single/married/divorced)
  - Ordinal
    - (Strongly agree, agree, disagree, strongly disagree)
    - Distance between categories is unknown
  - Interval
    - (Income, # of pupils per teacher, age, weight, etc.)
    - Specific distance between each level
    - My sister is not only younger than I am, but she is 2 years younger
    - My parents have 4 children. Number of children is a discrete variable, you cannot subdivide children, you have 1, or 1, or 3. You can't have 2.5 children.

Nominal level measures

- Can quantify these data by tabulating them.
- Normally represent nominal data in a simple table with percentages.
- Take the marital status of all of my 25 friends (*i.e.* the population we are looking at is "all Prof. Wolf's friends").

Marital status	Number	%
Single	18	72%
Married	6	24%
Divorced	1	4%
Total	25	100%

Descriptive Statistics

- Describe a large amount of data in summary form
- Interested in what a typical person, country, school, legislature, etc. looks like

Measuring the central tendency

- Want to simplify interval level measurements to a few numbers.
- The salaries of all of my best friends (the population is Prof. Wolf's best friends).
- What is the *typical* annual salary of a best friend of mine?

Name	Salary
Derek	$75,000
Drew	$13,000
Amanda	$31,000
Ryan	$26,000
Maddie	$15,000

The mean

- The mean (or average) is the most usual way to *measure central tendency* of an interval variable
  - Sum of the measurements divided by the number of observations.
- For our salaried people:

$$\frac{75{,}000 + 13{,}000 + 31{,}000 + 26{,}000 + 15{,}000}{5}$$

  - Mean = $32,000

The mean's properties

- Shift of origin of measurement.
  - If everyone earns $2000 more, then the new mean salary is just the old mean salary ($32,000) PLUS $2000.
- Change of scale.
  - If we calculate salary in pounds sterling (say $1 = £0.5), then the new mean salary is simply half the old mean salary.
- Sum of two variables.
  - Imagine that income = salary + savings interest.
  - Mean income = mean salary + mean savings interest.

The median

- Imagine we ranked all observations, the median is simply the observation in the middle (½ of observations above and ½ below).
  - In ascending order the salaries are:

    13,000; 15,000; 26,000; 31,000; 75,000.
    - Median = $26,000.
  - 13,000;  26,000; 31,000; 75,000.

    Median = ½(26,000+31,000) = 28,500.

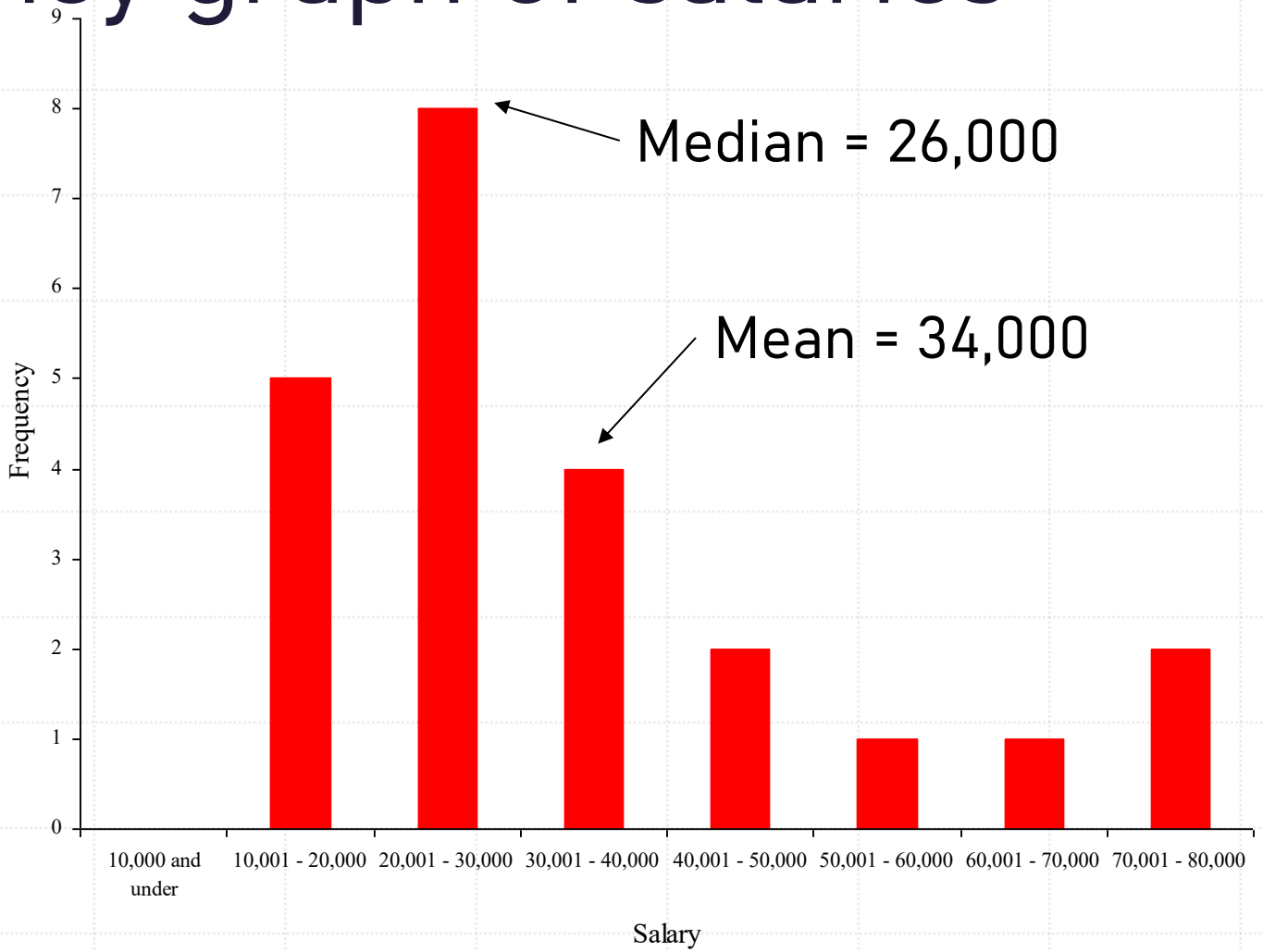
The median's properties (1)

- Shift of origin of measurement. YES
- Change of scale. YES
- Sum of two variables. NO
  - The lack of this property is somewhat important (which will become apparent in the following weeks), and is related to one of the reasons why we generally use the mean in most statistical analysis.
  - Nonetheless, the median does have some advantages over the mean in describing some types of data.

The median's properties (2)

- For our salary example, the mean of my best friends' salaries gives a substantially higher value than the median ($6000 more).
- This is due to the distribution of the observations. For the mean and median to be the same the distribution of observations needs to be *symmetrical*.
- Imagine we now look at all my friends and acquaintances (the population of 25 people as before), and plot the frequency of each salary for all 25.

Frequency graph of salaries



Positions of the median and mean

- For distributions with a long *tail* to the right, the mean will take a higher value than the median.
- This is generally true across the world for income distributions, and is captured by Pen's (1971) "parade of dwarfs and a few giants".
  - If such a parade were organised today, then the person of mean height (and income) would be taller (and richer) than 65% of the population and so would pass by after 40 minutes had elapsed.
  - Mean income is ~$54,000, median income is ~$36,000.
- For data with 'outliers' the median can give a better idea of what the "typical" observation is like.

Ordinal level data

- The median can be used for ordinal level data.
- Imagine we had asked my 5 best friends about their position on the joining the war in Ukraine; 2 strongly agreed with sending American troops, one agreed, one disagreed and one strongly disagreed.
  - We can rank these *answers* and then find the median.

  Strongly agree; strongly agree; agree; disagree; strongly disagree.

  - Thus the median answer is agree. However, we cannot find the mean position, it's just not possible with ordinal (or any categorical) data.

Nominal level data

- In general, we can't use the median or mean for nominal data.
  - Normally use the **mode**. This is the most commonly occurring value.
    - *e.g.*, if 53 people here are politics students, 40 sociology students, and 46 are other subjects, then the *modal value* is politics.
  - There is one special case in which we can use the mean for nominal data however…

Nominal binary data

- …binary data is an exception as we can use the mean. Binary data (*e.g.* Yes/No, Male/Female) can be coded as 0 or 1.
  - A variable measuring sex, men are coded 1 and women coded 0.
  - The mean score for those 0s and 1s is the proportion of men. There were 2 women and 3 men amongst my best friends.
  - $$\text{Mean} = \frac{0+0+1+1+1}{5} = 0.6 = 60\%$$
  - The median does NOT make sense for binary data. It just tells us what the majority of the population is.

Exercise

- Population is all countries with nuclear capability, and variable is approximate number of nuclear weapons.
- What's the mean, mode and median for no. of nukes?
- How good is each of these at summarizing the data, do we need more information than just a measure of central tendency?

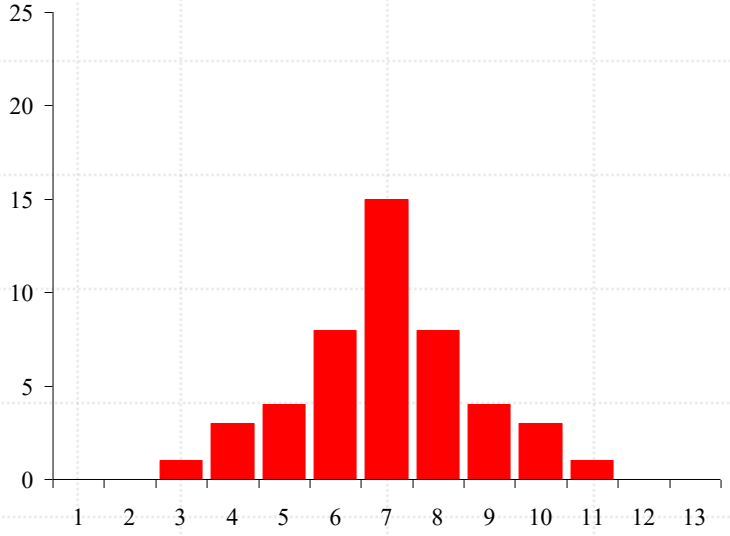
Country	No. of nukes
USA	10,000
India	75
China	400
France	400
Britain	200
Russia	12,800
Israel	100
Pakistan	25

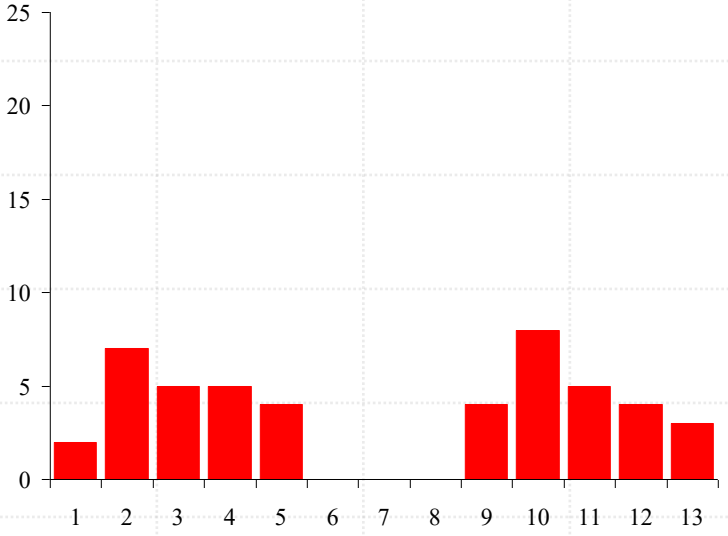
Some answers

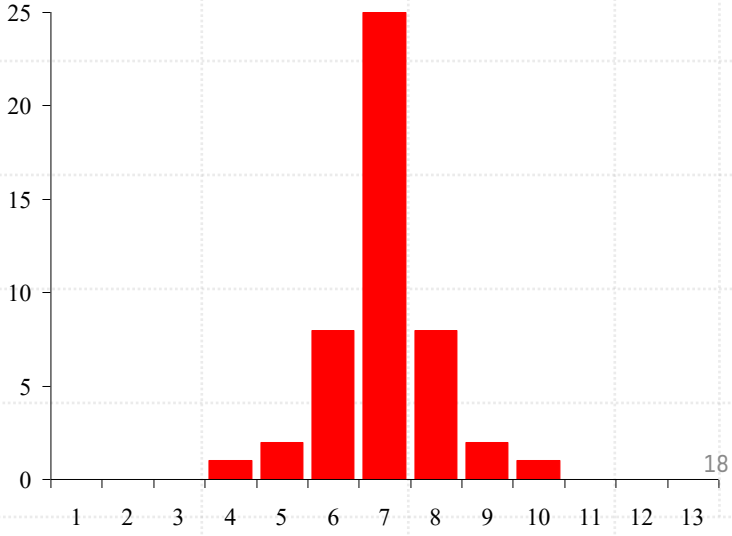
- Mean = 24,000/8 = 3,000
- Median = (400+200)/2 = 300
- Mode = 400
- These summary measures are useful, but we also need to know something about the distribution, because two countries account for virtually all the nuclear weapons in the world.

Measures of dispersion

- The mean (or median) tells us something about the center of the distribution, but what about its dispersion?
- The means/medians of the below distributions of children's scores on a math test in three different classes are all the same (48 observations, mean of 7, median of 7), but each tells a quite different story.







The range

- The range is simply a measure of the distance between the largest and smallest observations.
- The range for our salary example is therefore:

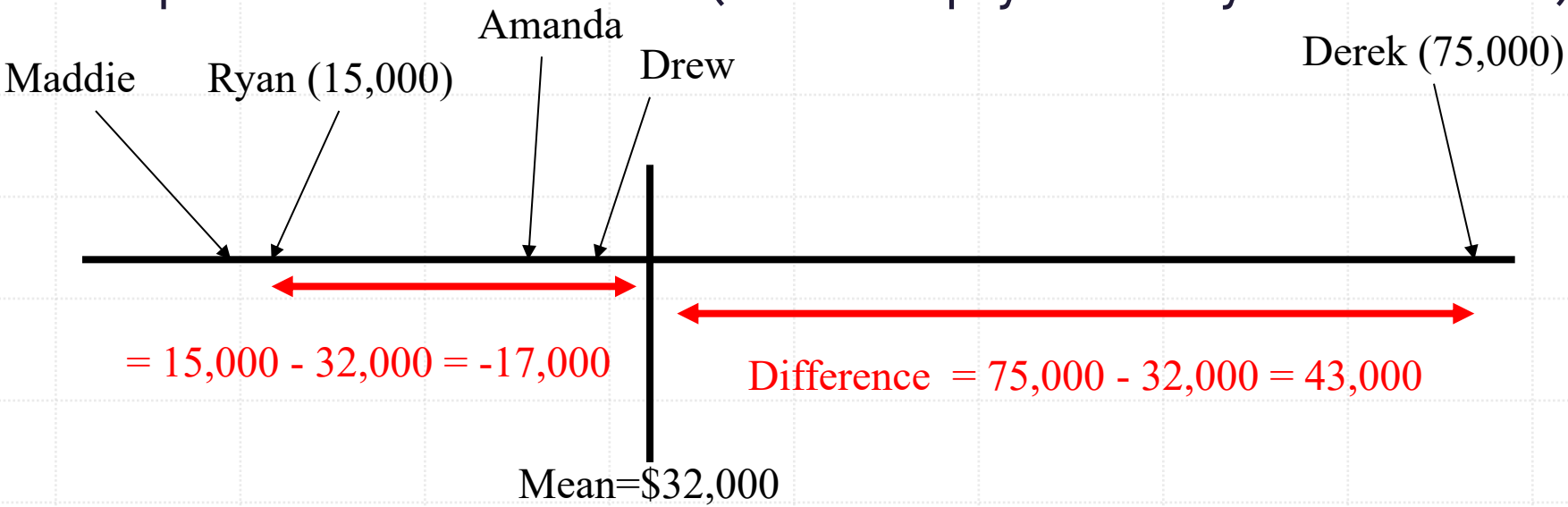
  75,000 – 13,000 = 62,000.
- Clearly this is not ideal as it relies on only two observations.
  - Say we have 1000 poker players. 999 win nothing, and 1 wins $1million. The range indicates lots of variation, when most people are actually identical.

The variance and standard deviation

- A better way of assessing how much values of a variable vary around the mean is to use the standard deviation or variance.
- Basic idea is to measure how different individual values are from the mean value.
- Some of these *deviations* from the mean will be positive and some negative, so we square each deviation.

The variance

- Take my 5 best friends. The mean salary was $32,000.
- If we added up all the differences then we would get zero, so we need to square the differences (*i.e.* multiply them by themselves).



Calculating variance

- Salary example, with 5 observations, and mean of 32,000.

Salary (000s)	Deviation from mean	Squared deviation
75	75 – 32 = 43	43 * 43 = 1849
13	13 – 32 = -19	-20 * -20 = 361
31	31 – 32 = -1	-1 * -1 = 1
26	26 – 32 = -6	-6 * -6 = 36
15	15 – 32 = -17	-17 * -17 = 289

$$\text{Total squared deviations} = 1849 + 361 + 1 + 36 + 289 = 2536$$

$$\text{Variance} = \frac{\text{Total squared deviations}}{n} = \frac{2536}{5} = 507.2$$

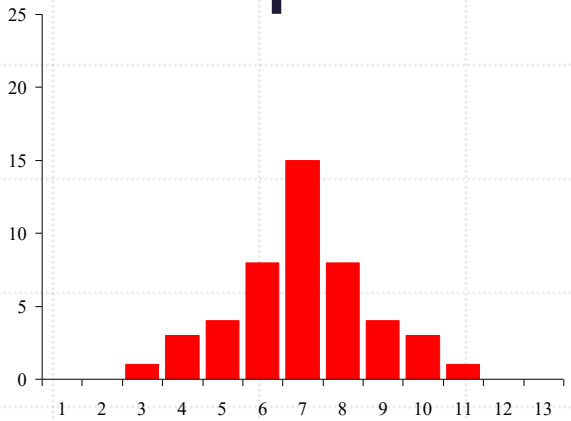
Calculating standard deviation

- The standard deviation is the most common way to measure deviation from the mean and is simply the square root of the variance.
  - We normally call the variance s^2 and the standard deviation s. Thus for our example, s^2 = 507.2, and s = 22.5.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n}}$$

**we usually use n-1 in the denominator

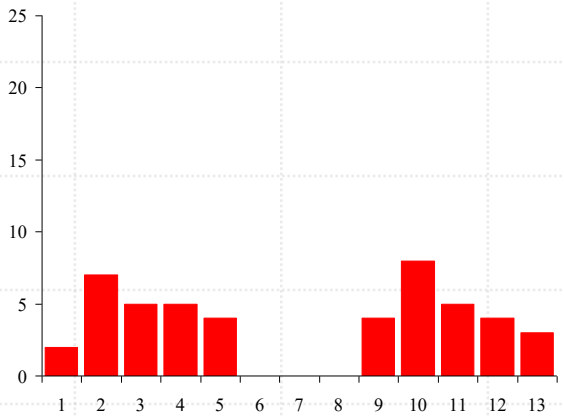
Examples of standard deviation



s = 1.67

Clustered distribution

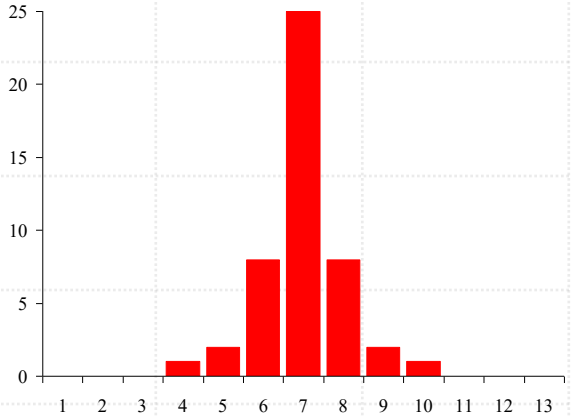
(Most children perform to a similar level, with some variation)



s = 4.01

Dispersed distribution

(One group of geniuses, one group of far from genius)



s = 1.02

Tight distribution

(All children perform similarly)

But what does it mean…?

- Our salary example had a standard deviation of 22.5, but for the distributions above the *s* varied between 1 and 4, what does this tell us?
- Best way to think of it is as a kind of rough average distance of an observation to the mean.
- Thus the standard deviation depends on the units we are measuring in.