Simple statistics I
Statistics

Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: “There are three kinds of lies: lies, damned lies, and statistics.”

Autobiography of Mark Twain
The goal of statistics is to

- Report data in meaningful ways
- Make predictions about future events

![Bar chart showing weight distribution for Course 15 and Course 6.](chart.png)
Statistics has 3+ components

- **Data analysis**
- **Descriptive statistics**
  - Probability calculations
- **Statistical inference**
  - Inferential statistics
- **Models** ....
Describing a state

- *Descriptive statistics*
  - *Capturing a picture of the data*)
  - *This was the origin of statistics*
    - *Started for gambling*
First some descriptive statistics

- 15.301 is the “best class ever”?
Central tendencies

- Representing central tendencies of distributions is a very efficient way to understand something about it.
  - Mode
  - Median
  - Mean
The Mode

• The most “popular” frequent occurring instance in the sample.
  – This is the only central tendency that can be used with a nominal scale
• The mode is sensitive to aggregation of categories
  – Age 18 vs age 18-21
• Sometimes there are multiple modes
  – Bimodal distributions
The Median

- The median is a value which 1/2 of the values are above it and 1/2 below
- After sorting the values by magnitude, the mode is at the \((n+1)/2\) location
- 123, 85, 34, 20, 18, 15, 14 → 20
- 123, 85, 34, 20, 18, 15 → \((20 + 34)/2 = 27\)
- When data is grouped, calculating the mode is a bit more complex
The Mean

- \( \text{Mean} = \frac{\sum X_i}{n} \)
- The most important statistic
- Used for many other computations
- Stable
  - Smallest mean square deviations from it
- Sensitive to extreme values
- Not “well behaved’ in non-standard distributions
Location of central tendencies

Normal
Mean
Mode
Median
Location of central tendencies

Bimodal
Mean
Mode
Median
Location of central tendencies

*Skew to right*

- Mean
- Mode
- Median
Location of central tendencies

*Skew to left*

Mean
Mode
Median
Distribution descriptors

- **The Range**
  - The range is *(Max - Min)*

- **Interquartile range**
  - Calculating is similar to median
  - *(Q3 - Q1) (1/2 of the observations)*
Variation I

- **Variance** \((\sigma^2)\)
  - \(\sum (X_i - \mu)^2 / (n)\)
  - \(\sum (X_i - \mu)^2 / (n - 1)\)

- **Standard deviation** \((\sigma)\)
  - *Square root of variance*
  - *Standard deviation is in the same units as the distribution*
Variation II

• Variance ($\sigma^2$) is:
  • insensitive to transformations consisting of adding a constant.
  • sensitive to transformations consisting of multiplying by a constant.
Describing scores:

- **Z scores**
  
  \[ z = \frac{r - \mu}{\sigma} \]
  
  \[ \mu = 0, \sigma = 1 \]

- **T scores**
  
  \[ \mu = 50, \sigma = 10 \]

  SAT, GRE etc.
Confidence in estimates?

• How sure can we be that we know the mean of the distribution, for example?
• Standard error of the mean
  \[ \mu^2 / \text{Square root of } N \]
The Correlation

- The relationship between 2 variables does not have to be linear
  - But in many cases they are
- Positive and negative correlations
Estimating correlations in scatter grams

• *What is the correlation here?*
Estimating correlations in scatter grams

- What is the correlation here?
Estimating correlations in scatter grams

- *What is the correlation here?*
Estimating correlations in scatter grams

• What is the correlation here?
Estimating correlations in scatter grams

• **What is the correlation here?**
Estimating correlations in scatter grams

• **What is the correlation here?**
Estimating correlations in scatter grams

- What is the correlation here?
The correlations were:

- $1 \leftrightarrow 0.1$
- $2 \leftrightarrow 0.3$
- $3 \leftrightarrow 0.5$
- $4 \leftrightarrow 0.7$
- $5 \leftrightarrow 0.9$
- $6 \leftrightarrow 0.99$
- $7 \leftrightarrow 0.1$
What is a correlation?

• What line to pick?
  – Sum of all deviations from the line is 0
  – The sum of square deviations of the points from the line is minimal.

• \( R = \frac{S_{xy}}{S_x \times S_y} \)
  – The relationship of their joint standard deviation to their individual standard deviation

• \( R^2 \) is the amount of explained variance
One of the main usages of statistics is to describe data

- Central tendencies: Mean, Mode, Median
- Distribution tendencies: Variance, IQR, Correlations