Chapter 4 - Summarizing Numerical Data

15.075 Cynthia Rudin

Here are some ways we can summarize data numerically.

- **Sample Mean**:

$$\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n}.$$

  Note: in this class we will work with both the population mean μ and the sample mean $\bar{x}$. Do not confuse them! Remember, $\bar{x}$ is the mean of a sample taken from the population and μ is the mean of the whole population.

- **Sample median**: order the data values $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, so then

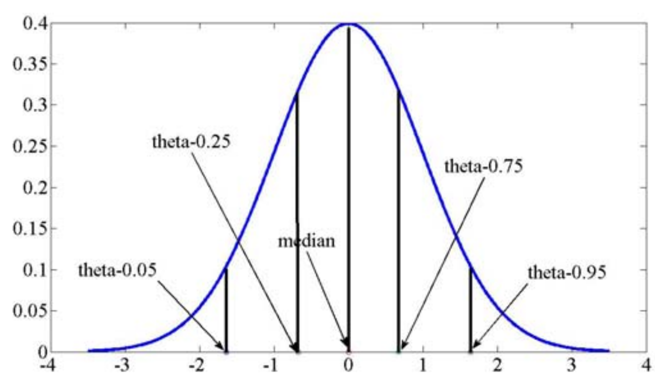
$$\text{median} := \bar{x} := \begin{Bmatrix} x_{(\frac{n+1}{2})} & \text{n odd} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] & \text{n even} \end{Bmatrix}.$$

  Mean and median can be very different: $1, 2, 3, 4, \underbrace{500}_{\text{outlier}}$.

  The median is more robust to outliers.

- **Quantiles/Percentiles**: Order the sample, then find $\tilde{x}_p$ so that it divides the data into two parts where:
  - a fraction p of the data values are less than or equal to $\tilde{x}_p$ and
  - the remaining fraction $(1-p)$ are greater than $\tilde{x}_p$.

  That value $\tilde{x}_p$ is the p^{th}-quantile, or $100 \times p^{\text{th}}$ percentile.



- **5-number summary**

$$\{x_{\min}, Q_1, Q_2, Q_3, x_{\max}\},$$

  where, $Q_1 = \theta_{.25}$, $Q_2 = \theta_{.5}$, $Q_3 = \theta_{.75}$.

- **Range:** $x_{\max} - x_{\min}$ measures dispersion

- **Interquartile Range:** IQR $:= Q_3 - Q_1$, range resistant to outliers

- **Sample Variance s^2 and Sample Standard Deviation s:**

$$s^2 := \underbrace{\frac{1}{n-1}}_{\text{see why later}} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

  Remember, for a large sample from a normal distribution, $\approx 95\%$ of the sample falls in $[\bar{x} - 2s, \bar{x} + 2s]$.

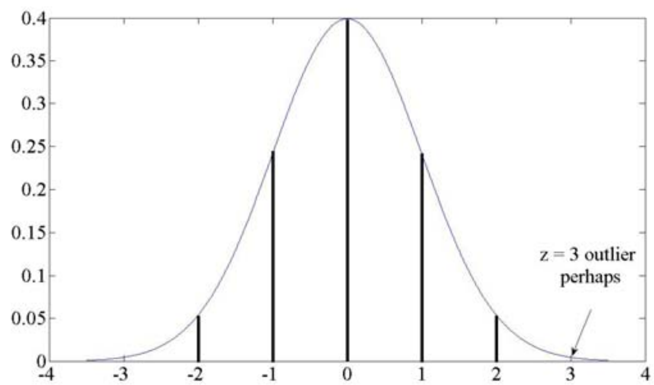
  Do not confuse s^2 with σ^2 which is the variance of the population.

- **Coefficient of variation** (CV) $:= \frac{s}{\bar{x}}$, dispersion relative to size of mean.

- **z-score**

$$z_i := \frac{x_i - \bar{x}}{s}.$$

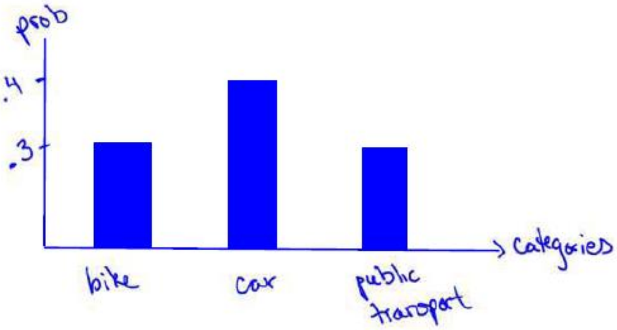
  - It tells you where a data point lies in the distribution, that is, how many standard deviations above/below the mean.
    E.g. $z_i = 3$ where the distribution is $N(0, 1)$.

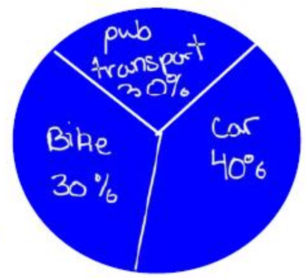


  - It allows you to compute percentiles easily using the z-scores table, or a command on the computer.

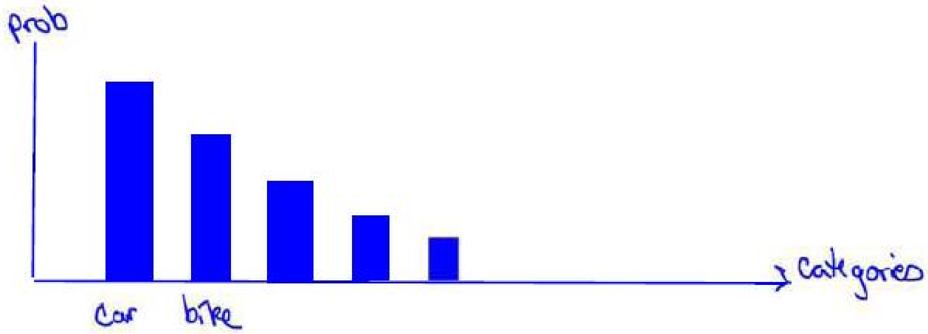
Now some graphical techniques for describing data.

- **Bar chart/Pie chart** - good for summarizing data within categories

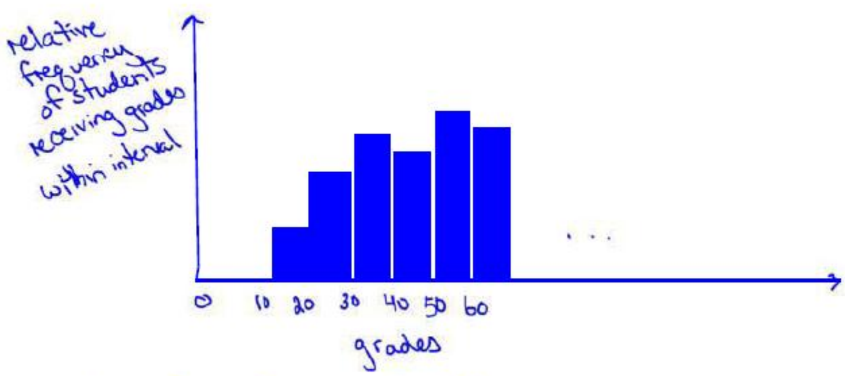




- **Pareto chart** - a bar chart where the bars are sorted.



- **Histogram**



(sample estimate of pdf)

Boxplot and normplot

Scatterplot for bivariate data

Q-Q Plot for 2 independent samples

Hans Rosling

Chapter 4.4: Summarizing bivariate data

Two Way Table

Here's an example:

Respiratory Problem?

	yes	no	row total
smokers	25	25	50
non-smokers	5	45	50
column total	30	70	100

Question: If this example is from a study with 50 smokers and 50 non-smokers, is it meaningful to conclude that in the *general population*:

a) $25/30 = 83\%$ of people with respiratory problems are smokers?

b) $25/50 = 50\%$ of smokers have respiratory problems?

Simpson's Paradox

- Deals with aggregating smaller datasets into larger ones.
- Simpson's paradox is when conclusions drawn from the smaller datasets are the *opposite* of conclusions drawn from the larger dataset.
- Occurs when there is a *lurking variable* and *uneven-sized groups* being combined

E.g. Kidney stone treatment (Source: Wikipedia)

Which treatment is more effective?

Treatment A	Treatment B
$78\%\ \frac{273}{350}$	$83\%\ \frac{289}{350}$

Including information about stone size, now which treatment is more effective?

	Treatment A	Treatment B
small stones	group 1 $93\%\ \frac{81}{87}$	group 2 $87\%\ \frac{234}{270}$
large stones	group 3 $73\%\ \frac{192}{263}$	group 4 $69\%\ \frac{55}{80}$
both	$78\%\ \frac{273}{350}$	$83\%\ \frac{289}{350}$

What happened!?

Continuing with bivariate data:

- **Correlation Coefficient**- measures the strength of a linear relationship between two variables:

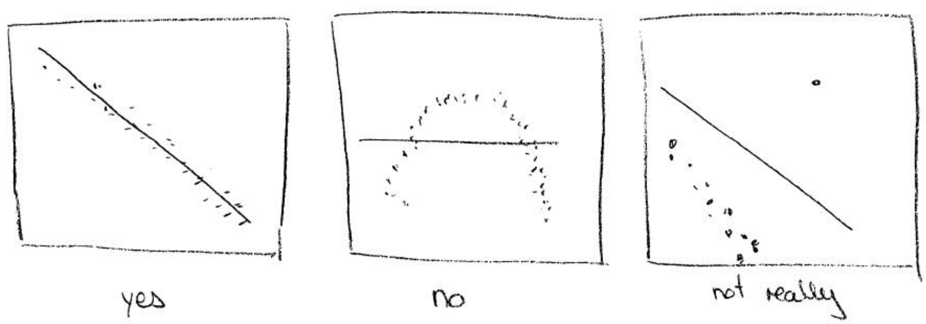
$$\text{sample correlation coefficient} = r := \frac{S_{xy}}{S_x S_y},$$

where

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

This is also called the "Pearson Correlation Coefficient."

  - If we rewrite
  $$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_x}\frac{(y_i - \bar{y})}{S_y},$$
  you can see that $\frac{(x_i-\bar{x})}{S_x}$ and $\frac{(y_i-\bar{y})}{S_y}$ are the z-scores of x_i and y_i.
  - $r \in [-1, 1]$ and is ± 1 only when data fall along a straight line
  - sign(r) indicates the slope of the line (do y_i's increase as x_i's increase?)
  - always plot the data before computing r to ensure it is meaningful

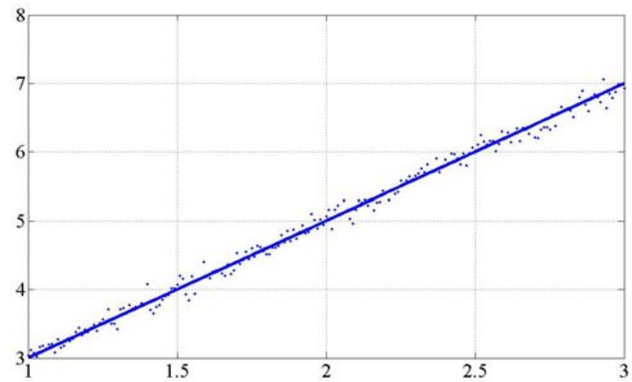


  - Correlation *does not imply* causation, it only implies *association* (there may be lurking variables that are not recognized or controlled)

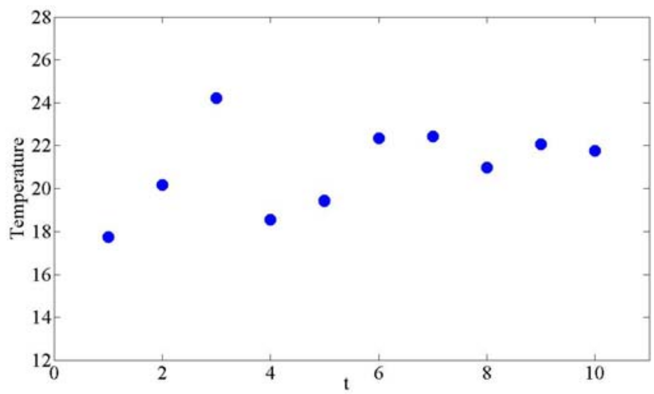
  For example: There is a correlation between declining health and increasing wealth.

- **Linear regression** (in Ch 10)

$$\frac{y - \bar{y}}{S_y} = r\frac{x - \bar{x}}{S_x}.$$



Chapter 4.5: Summarizing time-series data



- **Moving averages**. Calculate average over a window of previous timepoints

  –
  $$MA_t = \frac{x_{t-w+1} + \cdots + x_t}{w},$$
  where w is the size of the window. Note that we make window w smaller at the beginning of the time series when $t < w$.

  Example

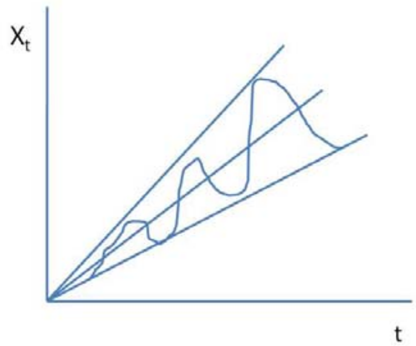
  To use moving averages for forecasting, given $x_1, \ldots, x_{t-1}$, let the predicted value at time t be $\hat{x}_t = MA_{t-1}$. Then the forecast error is:

  $$e_t = x_t - \hat{x}_t = x_t - MA_{t-1}.$$

- **The Mean Absolute Percent Error (MAPE)** is:

  $$MAPE = \frac{1}{T-1} \sum_{t=2}^{T} \left| \frac{e_t}{x_t} \right| \cdot 100\%.$$

The MAPE looks at the forecast error e_t as a fraction of the measurement value x_t. Sometimes as measurement values grow, errors, grow too, the MAPE helps to even this out.



For MAPE, x_t can't be 0.

- **Exponentially Weighted Moving Averages (EWMA)**.
  - It doesn't completely drop old values.
  $$EWMA_t = \omega x_t + (1-\omega)EWMA_{t-1},$$
  where $EWMA_0 = x_0$ and $0 < \omega < 1$ is a smoothing constant.

  Example

  - here ω controls balance of recent data to old data
  - called "exponentially" from recursive formula:
  $$EWMA_t = \omega[x_t + (1-\omega)x_{t-1} + (1-\omega)^2 x_{t-2} + \dots] + (1-\omega)^t EWMA_0$$
  - the forecast error is thus:
  $$e_t = x_t - \hat{x}_t = x_t - EWMA_{t-1}$$
  - HW? Compare MAPE for MA vs EWMA
- **Autocorrelation coefficient.** Measures correlation between the time series and a lagged version of itself. The k^{th} order autocorrelation coefficient is:

$$r_k := \frac{\sum_{t=k+1}^{T}(x_{t-k}-\bar{x})(x_t-\bar{x})}{\sum_{t=1}^{T}(x_t-\bar{x})^2}$$

Example

15.075J / ESD.07J Statistical Thinking and Data Analysis
Fall 2011