Statistical Methods

Lecture 3 – Using statistics to summarise data

Luke Dickens

Autumn 2020

Central Tendency

Variability and Normality

Sample Correlation Coefficient

Statistics

Statistics are numerical quantities computed from data:

- These summarize certain features of the data
- Central tendency: mean, median and mode
- Variation/dispersion/spread: variance, standard deviation (and percentile ranges)
- Correlation: correlation coefficient
- There are others: skew, kurtosis...

We will look first at measures of central tendency.

Sample mean

For n data-points $x_1, x_2, \ldots, x_n$ sample mean defined as:

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

- For data $y_i = x_i + c$, then $\bar{y} = \bar{x} + c$
- For data $y_i = cx_i$, then $\bar{y} = c\bar{x}$
- Weighted average:

$$\bar{x} = \sum_{i=1}^{n} w_i x_i \quad \text{where} \quad \sum_i w_i = 1$$

- Calculate mean from frequency table using weighted average:

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

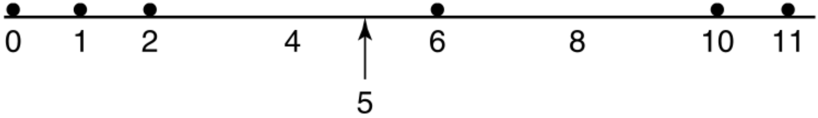
Compare severity of motorcycle accidents, for drivers with and without helmets [?] analysed in [?, Sec. 3.2].

Classification	with helmet	without helmet
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
Totals	331	439
Mean Severity	0.432	0.902

Deviations

Deviations: differences between data values and mean:

- ith deviation is $x_i - \bar{x}$
- Deviations sum to 0: $\sum_i x_i - \bar{x} = 0$
- Physical analogy: n weights of equal mass placed on (weightless) rod at positions xi, $i = 1, ..., n$
  - $\bar{x}$ is centre of mass (point at which rod balances)
  - $(x_i - \bar{x})$ is x_i's signed distance from centre



Centre of mass for 0, 1, 2, 6, 10 *and* 11 *[*?*, Sec. 3.2]*

Median

For n ordered data-points $x_1, x_2, \ldots, x_n$ (such that $x_i \leq x_{i+1}$), sample median defined as:

$$m = \begin{cases} x_i & \text{for } i = \frac{n+1}{2}, \quad \text{if } n \text{ is odd} \\ (x_i + x_{i+1})/2 & \text{for } i = \frac{n}{2}, \quad \text{if } n \text{ is even} \end{cases}$$

- If data is symmetric, mean $\approx$ median
- Median less affected by extreme values
- Median may be more relevant for some questions

Sample percentiles

Sample $100p$ percentile defined as data value such that at least: $100p\%$ of data are less than or equal and $100(1-p)\%$ are greater than or equal. If two data values satisfy condition, then it is the arithmetic average of these values.

For n ordered data-points $x_1, x_2, \ldots, x_n$ (such that $x_i \leq x_{i+1}$), then the sample $100p$ percentile

$$= \begin{cases} x_i & \text{for } i = \lceil np \rceil, \quad \text{if } np \text{ not integer} \\ (x_i + x_{i+1})/2 & \text{for } i = np, \quad \text{if } np \text{ is integer} \end{cases}$$

The 25^{th}, 50^{th} and 75^{th} percentiles respectively called the 1^{st}, 2^{nd} and 3^{rd} quartiles.

Sample mode and Benford's law

Sample mode defined as the data value that occurs most frequently in the data set.

- If no single value occurs most frequently, then all highest frequency values are called modal values...
- ...and we say there is no unique sample mode.

Benford's law (First-Digit Law) is observation that first digits in many real-life data sets are biased towards the smaller numbers.

Central Tendency

Variability and Normality

Sample Correlation Coefficient

Sample Variance and Standard Deviation

For n data-points $x_1, x_2, \ldots, x_n$ sample variance defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- A measure of the variability of data
- Almost, the average squared deviation (from the mean)
- Important algebraic identity:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

- Sample standard deviation (SD) is square root of the variance.

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

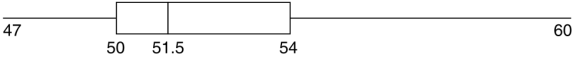
Properties

- Sample SD measured in same units as original data
- Sample variance measured in squared units of original data
- Variance (and SD) unaffected by constant shift, i.e. for data $y_i = x_i + c$, then $s_y^2 = s_x^2$
- Scaling has squared influence on variance, i.e. for data $y_i = cx_i$, then $s_y^2 = c^2 s_x^2$
- ... and for SD, $s_y = |c| s_x$

Other measures of variation/dispersion/s

Variance and SD are not the only statistical measures of dispersion. Of these, the most common is probably interquartile range:

- difference between sample 75$^{\text{th}}$ and 25$^{\text{th}}$ percentiles
- used for box-plots, e.g. starting salary data [**?**, p. 95]



Other measures are much rarer:

- Mean absolute deviation (about the mean):

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

  may be simpler to understand [**?**]

- Median absolute deviation about the median (also MAD) used in robust statistics (resistant to outliers) [**?**]

Normality and the empirical rule

A data set is said to be normal if a histogram describing it has properties:

1. Highest at the middle interval.
2. Height decreases in both directions in a bell-shape
3. Symmetric about its middle interval

- From symmetry, mean and median approx. equal
- **Empirical rule** specifies approx. proportions of data within s, $2s$, and $3s$ of $\bar{x}$ as 68%, 95% and 99.7% respectively.

Bimodal data (with 2 peaks) is not normal, and may represent a mixture of sub-populations.

Central Tendency

Variability and Normality

Sample Correlation Coefficient

Sample Correlation Coefficient

The sample correlation coefficient for paired data:

$$\{(x_i, y_i)\}_{i=1}^{n} = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

measures degree to which larger x values go with larger y values and smaller x values go with smaller y values. Given by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are means of x and y data respectively and s_x and s_y are their sample standard deviations.

Sometimes called Pearson's product-moment correlation coefficient
OR Pearson's r
OR just the r-value.

Properties of Pearson's r

Some properties of Pearson's r:

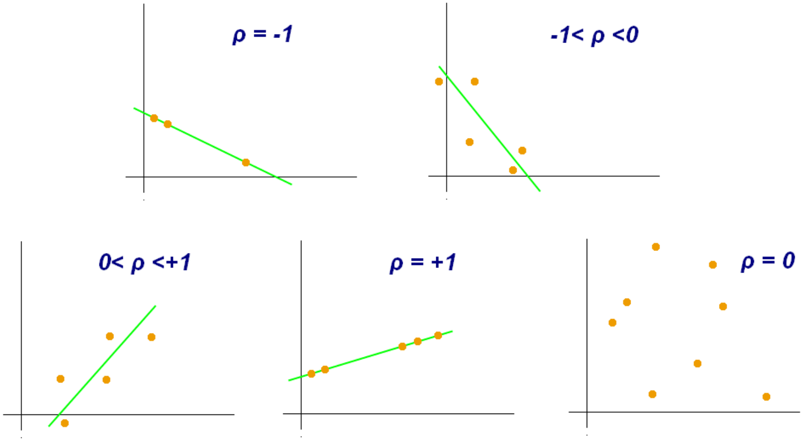
1. $-1 \leq r \leq +1$
2. if $y_i = a + bx_i$ for $b > 0$, $i = 1, \ldots, n$ then $r_{xy} = +1$ perfectly positively correlated
3. if $y_i = a + bx_i$ for $b < 0$, $i = 1, \ldots, n$ then $r_{xy} = -1$ perfectly negatively correlated
4. For $\{(u_i, v_i)\}_{i=1}^{n}$ where $u_i = a + bx_i$ and $v_i = c + dy_i$ for $i = 1, \ldots, n$ and $bd > 0$ then

$$r_{uv} = r_{xy}$$

(r unaffected by rescaling.)

Examples

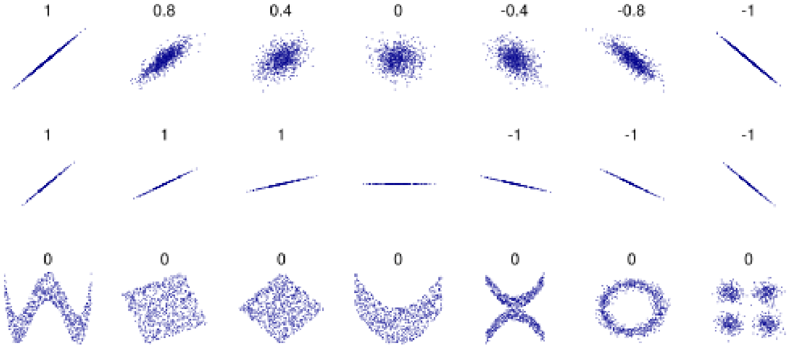
Here ρ (Greek letter rho) rather than r:



By Kiatdd - Own work, CC BY-SA 3.0 [link].

More Examples

r values shown above each image:



By DenisBoigelot, original uploader was Imagecreator - Own work, original uploader was Imagecreator, CC0, [link].

References I