Statistical Data Analysis Notes

Şafak Bilici

1 Introduction

Categories of data analysis are

- Narrative (laws, arts etc.)
- Descriptive (social sciences)
- Statistical/mathematical
- Audio-optical
- Others

Descriptive Statistics: Uses sample information to explain/make abstraction of population. (non-parametrics)
Inferential Statistics: Uses sample statistics to infer some "phenomena" of population parameters. (parametric)
Data: unprocessed facts and figures without any added interpretation or analysis.
Information: data that has been interpreted so that it has meaning for the user.
Knowledge: a combination of information, experience and insight that may benefit the individual or the organisation.
Analysis: Analysis is defined as the procedure by which we break down an intellectual or substantial whole into parts.
Synthesis: Synthesis is defined as the procedure by which we combine separate elements or components in order to form a coherent whole.

2 Median

- the median is more robust against outliers.
- The median may be given with its interquartile range (IQR).
- The 1st quartile point has the $\frac{1}{4}$ of the data below it
- The 3rd quartile point has the $\frac{3}{4}$ of the sample below it.
- The IQR contains the middle $\frac{1}{2}$ of the sample.

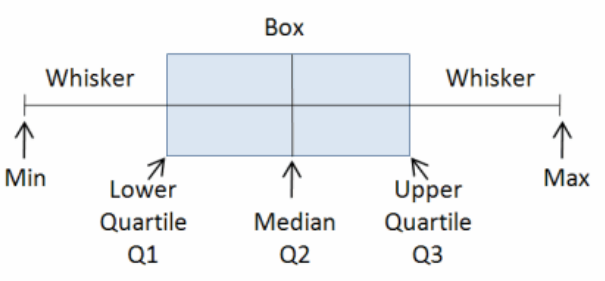

Figure 1: box and whisker plot

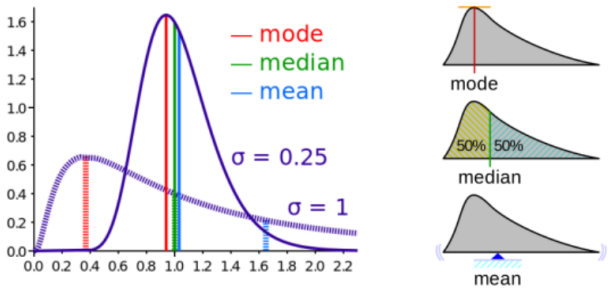

Figure 2: Geometric visualisation of the mode, median and mean of an arbitrary probability density function.

3 Moving Average Filter

$$y_n = \sum_{j=-k}^{k} w_j \cdot x_{n+j}$$

k is the order.

4 Moving Difference Filter

First order:

$$Dy_t = y_t - y_{t-1}$$

Higher order differences (2nd order):

$$D^2 y_t = D(Dy_t) = Dy_t - Dy_{t-1} = y_t - 2y_{t-1} + y_{t-2}$$

5 Variance and Standard Deviation

The sample variance is a common measure of dispersion based on the squared deviations:

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

The square root of the variance is called the sample standard deviation:

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

- SD indicates how much a set of values is spread around the average:

- A range of one SD above and below the mean (abbreviated to ±1 SD) includes 68.2% of the values.
- ±2 SD includes 95.4% of the data.
- ± includes 99.7%.

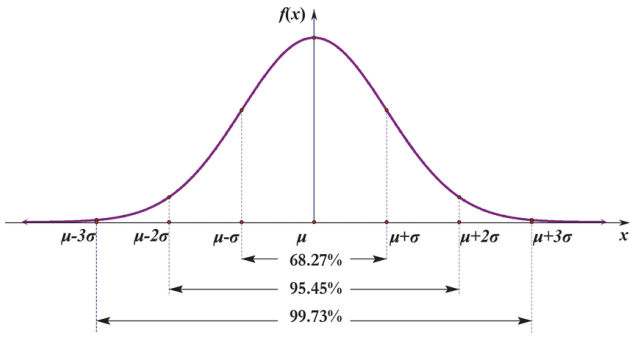

Figure 3: 3 sigma rule

6 Percentiles

We have n sample size and we want to find kth percentile.

1. $j \leftarrow \frac{k}{100} \cdot n$
2. if j is not whole number, round up j and select the jth item.
3. else if j is whole number, average jth and $(j+1)$th item.

7 Quartile

For sampled data, the median is also known as the 2nd quartile, Q2. Given Q2, we can find the 1st quartile, Q1, by simply taking the median value of those samples that lie below the 2nd quartile. We can find the 3d quartile, Q3, by taking the median value of those samples that lie above the 2nd quartile.
Quartiles can also be found in terms of percentiles:

- 1st quartile is 25th percentile.
- 2nd quartile is 50th percentile.
- 3rd quartile is 75th percentile.

8 Five-Number Summary

min, 1st quartile, median, 3rd quartile, max. This way, the five-number summary provides 0, 0.25, 0.50, 0.75, and 1 quantiles.

9 Interquartile Range (IQR)

$$IQR = Q3 - Q1$$

- If a sample is higher than $Q3 + 1.5 \cdot IQR$, then it is said to be outlier.
- If a sample is less than $Q1 - 1.5 \cdot IQR$, then it is said to be outlier.

- Considering the following (25) test scores
  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
- Q1 (25^{th} percentile)
  0.25 * 25 = 6.25 ➔ (round up) ➔ 7 Q1 = 68
- Q2 (50^{th} percentile)
  0.50 * 25 = 12.5 ➔ (round up) ➔ 13 Q2 = 77
- Q1 (75^{th} percentile)
  0.75 * 25 = 18.75 ➔ (round up) ➔ 19 Q3 = 89

Figure 4: quartile example

10 Data Transformation

- to make the distribution of the data normal. This increases the sensitivity of statistical tests.
- logarithm or sqrt: to reduce the influence of extreme values in our analysis.
- We can create a new variable based on two or more existing variables (for example BMI).

If your data distribution is…	Try this transformation method
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

- C = a constant added to each score so that the smallest score is 1.
- K = a constant from which each score is subtracted

Figure 5: data transformation functions

11 Coefficient of Variation

In general, the coefficient of variation is used to compare variables in terms of their dispersion when the means are substantially different. To quantify dispersion independently from units, we use the coefficient of variation

$$CV = \frac{s}{\mu}$$

In general, when we multiply the observed values of a variable by a constant a, its mean, standard deviation, and variance are multiplied by a, $\mid a \mid$, and a^2 , respectively. The coefficient of variation is not affected.

12 Scaling and Shifting Variables

If we shift the observed values by b, $y = x + b$, then

- $\mu(y) = \mu(x) + b$

- $s_y = s_x$
- $s_y^2 = s_x^2$

If we multiply the observed values by the constant a and then add the constant b to the result, $y = ax + b$, then

- $\mu(y) = a\mu(x) + b$
- $s_y = \mid a \mid s_x$
- $s_y^2 = a^2 s_x^2$

13 Correlation

Consider a set of observed pairs of values, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. For these observed pairs of values, Pearson's correlation coefficient is calculated as follows:

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

For the two variable, s_x and s_y denote the sample standard deviations.

- The values of r are always between ± 1 and ± 1.

14 Sample Covariance

If the standard deviations are removed from the denominator in Pearson's correlation coefficient, the statistic is called the sample covariance,

$$v_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu(x)) \cdot (y_i - \mu(y))}{n-1}$$

15 Two Categorical Variables

- **difference of proportions**:
  - $p_2 - p_1$
  - We can present this difference as a percentage using the sample proportion (risk) in the p_1 group as the baseline: $\frac{p_2 - p_1}{p_1} \cdot 100\%$
- **relative proportion**
  - $\frac{p_2}{p_1}$
- It is more common to compare the sample odds:
  - $o = \frac{p}{1-p}$
- We usually compare the sample odds using the sample odds ratio
  - $OR_{21} \frac{o_2}{o_1}$
  - An odds ratio equal to 1 means that the odds are equal in both groups and is interpreted as no relationship between the two categorical variables.
  - Values of the odds ratio away from 1 (either greater than or less than 1) indicate that the relationship is strong.

16 Joint vs. Marginal Probability

- We refer to the probability of the intersection of two events, $p(E_1 \cap E_2)$ as their joint probability.
- In contrast, we refer to probabilities $p(E_1)$ and $P(E_2)$ the marginal probabilities of events:

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

17 Disjoint Events

Two events are called disjoint or mutually exclusive if they never occur together. Disjoint events have no elements (outcomes) in common, and their intersection is the empty set.

18 Conditional Probability

$$p(A \mid B) = \frac{p(A \cup B)}{p(B)}$$

$$p(A' \mid B) = 1 - p(A \mid B)$$

$$p(A \cap B \mid C) = p(A \mid C) + p(B \mid C) - p(A \cup B \mid C)$$

19 The law of total probability

$$p(A) = p(A \mid B_1)p(B1) + ... + p(A \mid B_k)p(B_k)$$

20 Independent Events

$$p(A \cap B) = p(A) \cdot p(B)$$

$$p(A \cup B) = p(A) + p(B) - p(A) \cdot p(B)$$

21 Bayes' Theorem

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

$$p(A_i \mid B) = \frac{p(B \mid A_i) \cdot p(A_i)}{\sum_{i=0}^{k} p(A \mid B_k) \cdot p(B_k)}$$

22 Discrete probability distributions

The probability distribution of a discrete random variable is fully defined by the probability mass function (pmf). This is a function that specifies the probability of each possible value within range of random variable.

23 Bernoulli Distribution

$$X \sim Bernoulli(\theta)$$

$$p(X = x) = \begin{cases} 1 - \theta & \text{for } x = 0 \\ \theta, & \text{for } x = 1 \end{cases}$$

$$\mu = \theta$$

$$\sigma^2 = \theta \cdot (1 - \theta)$$

24 Binomial Distribution

A sequence of binary random variables is called Bernoulli trials.

$$X \sim Binomial(n, \theta)$$

$$p(X = x) = \binom{n}{x} p^x \cdot (1 - p)^{n-x}$$

$$\mu = n \cdot \theta$$

$$\sigma^2 = n \cdot \theta \cdot (1 - \theta)$$