Last updated: 2023-12-16 20:47:16-05:00

Contents

The following notes are from ***Mathematical Statistics with Applications, 7^{th} ed.*** by Wackerly, Mendenhall & Scheaffer.

1 What Is Statistics?

1.1 Introduction

1.1 Definition (statistics).
Statistics is a theory of information with inference-making as its objective.

1.1 Remark (The objective of statistics).
The objective of statistics is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference.

1.2 Definition (population).
A ***population*** is the large body of data that is the target of our interest.

1.3 Definition (sample).
A ***sample*** is a subset selected from a population.

1.2 Characterizing a Set of Measurements: Graphical Methods

1.1 Note (relative frequency histogram).
Given data set $[2.1, 2.4, 2.2, 2.3, 2.7, 2.5, 2.4, 2.6.2.6, 2.9]$ we can construct the ***relative frequency histogram***:

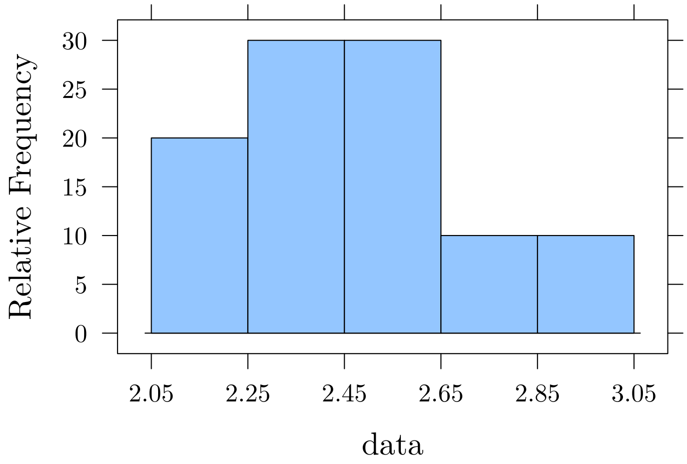

Figure 1: Relative Frequency Histogram

- 2 out of the 10 values, $[2.1, 2.2]$, lie between 2.05 and 2.25.
- 3 out of the 10 values, $[2.4, 2.3, 2.4]$, lie between 2.25 and 2.45.
- 3 out of the 10 values, $[2.5, 2.6, 2.6]$, lie between 2.45 and 2.65.
- 1 out of the 10 values, $[2.7]$, lies between 2.65 and 2.85.
- 1 out of the 10 values, $[2.9]$, lies between 2.85 and 3.05.

1.2 Remark.

When choosing the intervals of a histogram:

1. Choose so that it is impossible for a measurement to fall on a point of division. The idea is to separate the data into distinct buckets and look at what perecentage of the total data set is contained in each bucket. The 'buckets' correspond to the chosen intervals. If intervals are chosen such that a data point falls exactly between two intervals, then it's ambiguous as to which interval the data point lies in.
2. You could choose an interval that contains every data point but this defeats the purpose of using the histogram to visualize the relative frequency of data in the data set. You could also choose many small intervals which each contain at most one data point but this would defeat the purpose of broadly visualizing the relative frequency of data in the data set. The number of intervals should scale appropriately with the size of the data set.

1.2 Note.

The probability that an arbitrary value from a data set will lie in a given interval of a relative frequency histogram for that data set is equal to the percentage of the total area of the histogram lying over that interval.

1.1 Example.

The total area of the histogram of Fig.1 is:

$$(2.25 - 2.05) \cdot 0.2 + (2.45 - 2.05) \cdot 0.3 + (2.65 - 2.45) \cdot 0.3 + (2.85 - 2.65) \cdot 0.1 + (3.05 - 2.85) \cdot 0.1 = 0.2$$

The area of the histogram in the interval $[2.05, 2.45]$ is:

$$(2.25 - 2.05) \cdot 0.2 + (2.45 - 2.05) \cdot 0.3 = 0.1$$

So, based on this histogram, the probability that an arbitrary data point lies between 2.05 and 2.45 is:

$$0.1/0.2 = 0.5 = 50\%$$

1.3 Characterizing a Set of Measurements: Numerical Methods

1.4 Definition (mean).

Given a data set $[y_1, y_2, \ldots, y_n]$, the ***mean*** of the data set is given by:

$$\mu = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

1.1 Notation (mean: μ, $\bar{y}$).

μ for population, $\bar{y}$ for sample. $\bar{y}$ is read "y bar".

1.5 Definition (population variance).

Given a population $[y_1, y_2, \ldots, y_n]$, the ***population variance*** of the population is given by:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu)^2$$

1.2 Notation (population variance: σ^2).

1.6 Definition (sample variance).

Given a sample $[y_1, y_2, \dots, y_n]$, the ***sample variance*** of the population is given by:

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

1.3 Remark.

It isn't immediately obvious why the denominator of the sample variance should be $n-1$ instead of n like in the popluation variance. The reason for this will be explained later after developing some concepts.

1.3 Notation (sample variance: s^2).

1.7 Definition (standard deviation).

- Given a population, a ***standard deviation*** of the population is given by $\sigma = \sqrt{\sigma^2}$ where σ^2 is the population variance.
- Given a sample, the ***standard deviation*** of the sample is given by $s = \sqrt{s^2}$ where s^2 is the sample variance.

1.4 Notation (standard deviation: σ, s).

σ for population, s for sample.

1.3 Note (Empirical Rule).

For a distribution of measurements that is approximately normal (bell-shaped):

- the interval $\mu \pm \sigma$ contains approx. 68% of measurements
- the interval $\mu \pm 2\sigma$ contains approx. 95% of measurements
- the interval $\mu \pm 3\sigma$ contains approx. almost all measurements

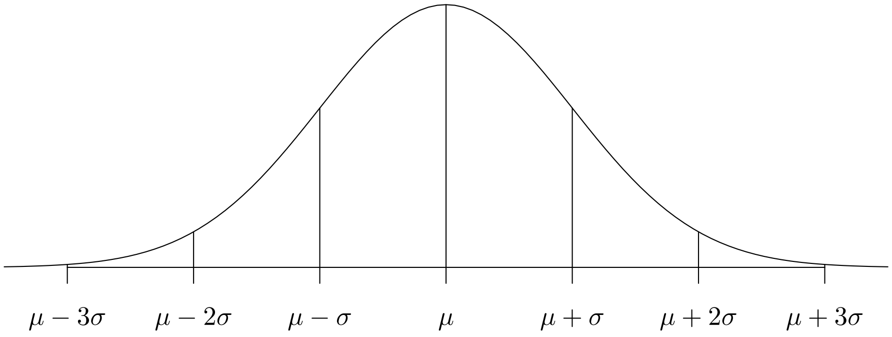

Figure 2: Empirical Rule

1.1 Lemma

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

Proof.

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(y_i^2-2\bar{y}y_i+\bar{y}^2) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-2\bar{y}\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\bar{y}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-2n\bar{y}^2+n\bar{y}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-n\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\right]
\end{aligned}$$

□

1.8 Definition (range).
Given a data set, the ***range*** of the data set is the difference between the largest and smallest values of the data set.

1.4 Note (Quarter Range Rule).
Since ∼ 95% – i.e. most – of all normally-distributed measurements lie within two standard deviations from the mean, standard deviation can be estimated as 1/4 the range of the measurements.

1.3.1 Exercises

Appendix

A List of Notation

B List of Definitions

C List of Lemmas, Theorems and Corollaries

Index