DESCRIPTIVE STATISTICS I
(DA_2021)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2021 Lecture n. 4, Rome 10th march 2021

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

KEYWORDS OF LECTURE N. 4

- Descriptive Statistics (location/spread of data)
- DATA COMPRESSION
- Means (why we do means, averages?)
- Chisini's criterion
- Other statistics(empirically computed parameters)
- Medians, modes (frequencies, rankings)
- Skewness of the distribution of the data (histograms)

Study materials for Descriptive Statististics
part I (measures of location of the data)
part II (measures of spread of the data)

- Rossner [R] chapter 2
- Whitlock&Sluter [**WS**] chapter 3
- Suggestion: read [**WS**] first and then [**R**]

From:[**R**]Bernard Rosner - Fundamentals of Biostatistics-Brooks Cole (2015)

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

Summary

Of what we have discussed in the previous lect

- Statistics is the study of methods for measuring aspects of populations from samples and for quantifying the uncertainty of the measurements.
- Much of statistics is about estimation, which infers an unknown quantity of a population using sample data.
- Statistics also allows hypothesis testing, a method to determine how well hypotheses about a population parameter fit the sample data.
- Sampling error is the chance difference between an estimate describing a sample and the corresponding parameter of the whole population. Bias is a systematic discrepancy between an estimate and the population quantity.
- The goals of sampling are to increase the accuracy and precision of estimates and to ensure that it is possible to quantify precision.
- In a random sample, every individual in a population has the same chance of being selected, and the selection of individuals is independent.
- A sample of convenience is a collection of individuals easily available to a researcher, but it is not usually a random sample.
- Volunteer bias is a systematic discrepancy in a quantity between the pool of volunteers and the population.
- Variables are measurements that differ among individuals.
- Variables are either categorical or numerical. A categorical variable describes which category an individual belongs to, whereas a numerical variable is expressed as a number.
- The frequency distribution describes the number of times each value of a variable occurs in a sample. A probability distribution describes the number of times each value occurs in a population. Probability distributions in populations can often be approximated by a normal distribution.
- In studies of association between two variables, one variable is typically used to predict the value of another variable and is designated as the explanatory variable. The other variable is designated as the response variable.
- In experimental studies, the researcher is able to assign subjects randomly to different treatments or groups. In observational studies, the assignment of individuals to treatments is not controlled by the researcher.

From:[**WS**] M.C. Whitlock and D. Schluter - The Analysis of Biological Data-W. H. Freeman and Company (2015).

The problem of pseudoreplication

Pseudoreplication is probably the single most common fault in the design and analysis of ecological field experiments. It is at least equally common in many other areas of research.

—**Stuart Hurlbert (1984)**

Most statistical techniques, including almost everything in this book, assume that each data point is independent of the others. Independence is, after all, built into the definition of a random sample. When we assume that data points are independent of each other, we give each data point equal credence and weigh its information as heavily as every other point. If two data points are not independent, though, then treating them as independent makes it seem as if we have more information than we really do. We would be treating the data set as if it were larger than it really is, and as a result we would calculate confidence intervals that were too narrow and *P*-values (see Section 6.2) that were too small.

From: W&S, interleaf 2

TABLE 2.1 Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

DEFINITION 2.1 The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

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

Basic linear invariance of the a.m.

EQUATION 2.3 Let $x_1, \ldots, x_n$ be the original sample of data and let $y_i = c_1 x_i + c_2$, $i = 1, \ldots, n$ represent a transformed sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2.

If $y_i = c_1 x_i + c_2, \quad i = 1, \ldots, n$

then $\bar{y} = c_1 \bar{x} + c_2$

Chisini's means (from Oscar Chisini, *Sul concetto di media*, Periodico di Matematiche, **2**, 106-116 (1929)).

$$f(\bar{x}, \bar{x}, \ldots, \bar{x}) = f(x_1, \ldots, x_n). \qquad (1)$$

Table 1. Examples of invariance requirements (f) and the corresponding Chisini means. The weights w_i's are assumed to be nonnegative and not all equal to 0.

	f	Conditions on x_i's	Mean	Name
1)	$\sum_{i=1}^{n} w_i x_i$	– –	$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$	weighted arithmetic mean
2)	$\sum_{i=1}^{n} w_i x_i^2$	$x_i \geq 0$	$\sqrt{\frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i}}$	weighted quadratic mean
3)	$\sum_{i=1}^{n} w_i x_i^{-1}$	$x_i > 0$	$\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i x_i^{-1}}$	weighted harmonic mean
4)	$\prod_{i=1}^{n} x_i^{w_i}$	$x_i > 0$	$\left(\prod_{i=1}^{n} x_i^{w_i}\right)^{\frac{1}{\sum_{i=1}^{n} w_i}}$	weighted geometric mean
5)	$\sum_{i=1}^{n} w_i x_i^k$	$x_i > 0$, $(k \neq 0, k \in \mathbb{R})$	$\sqrt[k]{\frac{\sum_{i=1}^{n} w_i x_i^k}{\sum_{i=1}^{n} w_i}}$	weighted power mean
6)	$\sum_{i=1}^{n} w_i e^{x_i}$	$x_i \geq 0$	$\log \frac{\sum_{i=1}^{n} w_i e^{x_i}}{\sum_{i=1}^{n} w_i}$	weighted exponential mean

The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the **median** or, more precisely, the **sample median**.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

DEFINITION 2.2 The **sample median** is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. Samples with an odd sample size have a unique central point; for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger. Samples with an even sample size have no unique central point, and the middle two values must be averaged. Thus, for samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

The Mode

Another widely used measure of location is the mode.

DEFINITION 2.3 The **mode** is the most frequently occurring value among all the observations in a sample.

Symmetry/asymmetry of distributions and localisation measures

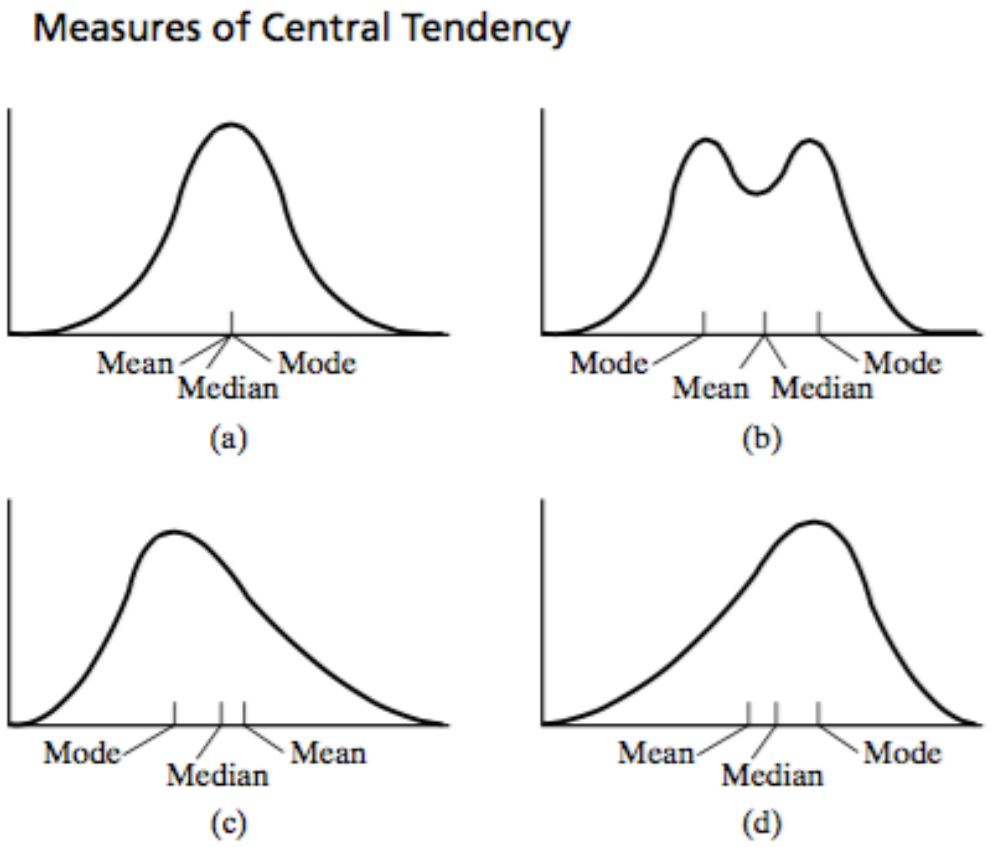

FIGURE 2: Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is asymmetrical and said to be positively skewed, and (d) is asymmetrical and said to be negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution b is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.

EXAMPLE 4 The Geometric Mean of Ratios of Change

Decade	*Population Size*	*Ratio of Change* X_i
0	10,000	
1	10,500	$\frac{10{,}500}{10{,}000} = 1.05$
2	11,550	$\frac{11{,}550}{10{,}500} = 1.10$
3	13,860	$\frac{13{,}860}{11{,}550} = 1.20$
4	18,156	$\frac{18{,}156}{13{,}860} = 1.31$

$$\overline{X} = \frac{1.05 + 1.10 + 1.20 + 1.31}{4} = \frac{4.66}{4} = 1.1650$$

and $(10{,}000)(0.1650)(1.650)(1.650)(1.650) = 18{,}421$

But,

$$\overline{X}_G = \sqrt[4]{(1.05)(1.10)(1.20)(1.31)} = \sqrt[4]{1.8157} = 1.1608$$

or

$$\overline{X}_G = \text{antilog}\left[\frac{\log(1.05) + \log(1.10) + \log(1.20) + \log(1.31)}{4}\right]$$

$$= \frac{\text{antilog}(0.0212 + 0.0414 + 0.0792 + 0.1173)}{4} = \frac{\text{antilog}(0.2591)}{4}$$

$$= \text{antilog } 0.0648 = 1.1608$$

and $(10{,}000)(1.1608)(1.1608)(1.1608)(1.1608) = 18{,}156$

EXAMPLE 5 **The Harmonic Mean of Rates**

$$X_1 = 40 \text{ km/hr}, \; X_2 = 20 \text{ km/hr}$$

$$\overline{X} = \frac{40 \text{ km/hr} + 20 \text{ km/hr}}{2} = \frac{60 \text{ km/hr}}{2} = 30 \text{ km/hr}$$

But

$$\overline{X}_H = \frac{2}{\dfrac{1}{40 \text{ km/hr}} + \dfrac{1}{20 \text{ km/hr}}} = \frac{2}{0.0250 \text{ hr/km} + 0.0500 \text{ hr/km}}$$

$$= \frac{2}{0.075 \text{ hr/km}} = 26.67 \text{ km/hr}$$

At this point go to [R] Rosner's textbook and illustrate the concepts in detail

Then connect to Di Leonardo's course Data Analysis:
lecture n. 1

2.12 SUMMARY

This chapter presented several **numeric and graphic methods for describing data.** These techniques are used to

(1) quickly summarize a data set

(2) present results to others

In general, a data set can be described numerically in terms of a **measure of location** and a **measure of spread.** Several alternatives were introduced, including the **arithmetic mean, median, mode,** and **geometric mean,** as possible choices for measures of location, and the **standard deviation, quantiles,** and **range** as possible choices for measures of spread. Criteria were discussed for choosing the appropriate measures in particular circumstances. Several graphic techniques for summarizing data, including traditional methods, such as the **bar graph**, and more modern methods characteristic of exploratory data analysis (EDA), such as the **stem-and-leaf plot** and **box plot**, were introduced.

How do the descriptive methods in this chapter fit in with the methods of statistical inference discussed later in this book? Specifically, if, based on some prespecified hypotheses, some interesting trends can be found using descriptive methods, then we need some criteria to judge how "significant" these trends are. For this purpose, several commonly used **probability models** are introduced in Chapters 3 through 5 and approaches for testing the validity of these models using the methods of **statistical inference** are explored in Chapters 6 through 14.

A nice solution to the exercise I gave to you: Debora Mazzetti's style

```
f=open("BMI.txt",'r')
lines=f.readlines()

f.close()
for row in lines:
  splitted_row = row.split(',')
  print(splitted_row[0])
  weight = int(splitted_row[0])
  height = float(splitted_row[1])
  bmi = weight/(height**2)
  print("BMI:", bmi)
```

So, now you can compute: averages, modes medians, histograms, global, and separated by the categoricqal variables M/F

SEE YOU NEXT MONDAY