INTRODUCTION TO STATISTICS AND COMPUTATION WITH DATA

Rituparna Sen, notes by Ramdas Singh

Second Semester

List of Symbols

$\hat{\theta}$, the estimate of a variable θ.

$\overline{x}$, the mean of the observations x_i.

x_M, the mediam of the observations x_M.

Contents

Chapter 1

AN INTRODUCTION TO STATISTICS

January 2nd.

Commonly referred to as the science of data, *statistics* involves collecting, summarizing, presenting, and interpreting data. Randomness and variability in data neccessitate the use of statistics. If a process is deterministic, there is no real need of statistics.

Often in probability, we are given the probability of getting heads in a single coin toss and we may be tasked to find the probability of 4 heads appearing in 10 tosses. In contrast, statistics starts with observing 4 heads appearing in 10 tosses, and utilises this data to determine the probability of a head. If we are to determine the price of a house in city, factors to look for may include the city itself, the specific location and area, the kind of house, the square footage, the age of the house, and the change with time. Despite all this, there is still an element of randomness; it is not a true deterministic quantity.

1.1 Fundamental Elements of Statistics

We first discuss some fundamental elements.

- An *experimental unit*. It may be a singular item such a coin toss, or a single house as in the previous example(s).
- The *population*. The set of all experimental units. We may note that studying all experimental units is a population may not be possible.
- A *census* studies all the units in a population.
- A *sample* of the population. A subset of the population. It is ideally chosen in a way that represents the entire population. A true representative sample may not always be possible without the the subset being the entire population.
- A *variable*. For each experimental unit in the sample, we record the data on several variables. In the case of the prices of houses, the factors discussed are the variables.
- *Univariate* and *multivariate* samples. As expected, a univariate sample has only one variable per unit, and a multivariate one has multiple variables per unit. Two variables per unit in the sample is also often referred to as a *bivariate* sample.

1.2 Types of Statistics and Data

1.2.1 Descriptive Statistics

Often, in statistics, we use pictures, tables, and summary numbers to describe the data. R Studio (or just R) may be used to handle descriptive statistics.

1.2.2 Inferential Statistics

Here, we make statements about the population based on our sample observations. In the case of coin tossing, let us look at the population of infinite coin tosses, where the probability of a heads is an unknown p. A census of this population is not possible, so we may take a sample of 10 tosses. Suppose we get X heads in this sample. We know that X follows a binomial distribution as $X \sim \text{Bin}(10, p)$. Here, we describe a new varaible called the *estimate*, $\hat{p}$. In this case, we find that $\hat{p} = X/10$. This is how we work in statistics; we deal with an unknown variable of the population, say θ, by looking at a sample and describing an estimate $\hat{\theta}$.

We may also be tasked to find the *measure of reliability*; how reliable our estimate is. One such measure may be $\left|\hat{\theta} - \theta\right| < \delta$ where we target to make δ as small as possible.

Example 1.1. Suppose 1000 cola consumers participate in a blind taste test among 2 brands, A and B, and are asked their preference. To know which kind is preferred universally, let us begin by asking the following:

- Describing the population. In this case, it is all the cola consumers.
- Describing the sample. In this case, it is our chosen 1000 cola consumers.
- The variable of interest. Whether people prefer brand A or brand B.
- Our inference. The preference in the sample is extended to all the cola consumers.

1.2.3 Type of Data

Qualitative data, or categorical or nominal or ordinal data, is data with no numerical value representation of it. In reference to the previous example, preference between A and B is a qualitative piece of data. Other such examples include choice of elective courses of a student, the gender of a person, a preference of a cricket team, etc..

Quantitative data on the other hand has a numerical value which interests us in statistics. Examples include the age of a person, the semestral marks of a student, the salary of a worker, the cost of books, etc.. Quantitative data is also divided into two parts, discrete and continuous.

1.3 Collecting Data

January 7th.

To collect data to perform statistics on, one may choose of the following ways to do so; a most basic source for sampling is a *published source*. In a *designed experiment*, we select experimental units and administer some treatment on each one. In the medical field, these medical experiments are called clinical trials. An *observational study* may also be conducted.

1.3.1 Sampling

Two kinds of simple random sampling exist; one is *simple random sampling without replacement* (SRSWOR), and the other is *simple random sampling with replacement* (SRSWR). Here, the experimental units are chosen sequentially at random with or without replacement. In *cluster sampling*, the population is divided into smaller groups known as clusters. Experimental units are then randomly selected among these clusters to form a sample. In *convenience sampling*, which is a non-probability sampling, the sample is drawn from that part of the population which is close to hand.

1.3.2 Sources of Error due to Flawed Sampling

A major source of error is *selection bias*; some parts of the population are deliberately left out when choosing the sample. It is an error of bias and not of randomness. For example, online surveys leave out people without access to internet. While this may be at fault of the person sampling the population, a *non-response bias* occurs when the population does not respond. A *reponse bias* also exists where the population does not reflect the true value. An error is noting down values or measuring samples may also occure, known as a *measurement error*.

Chapter 2

REPRESENTING QUANTITATIVE DATA

The most one can possible do with qualitative data is to pictorially represent it via tables and charts. *January 9th.*

For qualitative data, frequency tables, bar charts, pie charts, and ogives are the most approprite way to pictorially represent them. Quantitative data, on the other hand, have more variety in terms in descriptive statistics. They can be represented in graphs such as a dot plot, stem and leaf plot, histogram, adn even a box and whisker plot.

Plots can also be categorized as skewed and symmetric. A *symmetric plot* is (roughly) the same as its image under reflection about some vertical line. We typically look for symmetry of the population from which the data is a sample. A *skewed plot* is one where a peak in the plot occurs. A peak towards higher values is termed a left skewed plot, while a peak towards lower values is termd a right skewed plot. We can also have a *bimodal distribution* where 2 peaks occur, or even a *multimodal distribution*. More peaks suggests a mixture in the population.

Another thing to look out for are *outliers*. These are a few data points that are very different from the rest of the data. When such an outlier occurs, we are generally provoked to investigate it. A reason for an outlier could be a mistake in recording of data, which can be removed by fixing it. Another possible reason for a outlier is that the data point(s) come from a different distribution. In such a case, it is best to drop such outliers. Outliers may also occur purely due to the population being highly skewed; in this case, such outliers are expected by one.

2.1 Graphical Measures

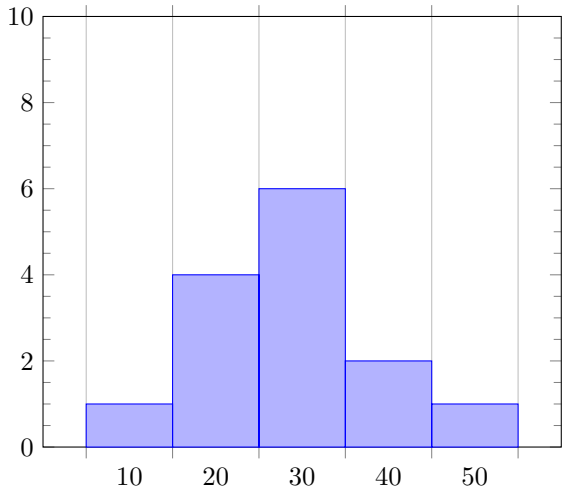
A *stem and leaf plot*. It is typically used when there are too many data points to draw a meaningful dot plot. It also has the added advantage of no loss of data. One may even look at it as a horizontal

Table 2.1: Key: $1|1 = 11$

Stem	Leaf
1	3
2	2 4 8 9
3	1 2 6 6 7 8
4	3 7
5	2

histogram, with the number of values per row representing the 'height' of the bars.

A *histogram*. For the same data as the previous example, we use bars instead to categorize classes. However, their is a loss of individual data points here.



A histogram.

A general rule of thumb exists for determining the number of classes. When there are less than 25 observations, we choose 5–6 classes; for 25 to 50 observations, 7–14 classes are enough. For more than 50 observations, one may even choose to use about 15–20 classes. Another thing to note about the histogram is that the area under the curve for a particular region represents the proportion of that region in the sample

A *box and whisker plot*. Again using the same data, a box and whisker plot utilises other aspects of the data such as the maximum value, the minimum value, the median value, the first quartile end value, and the third quartile end value. In our case, these are 52, 12, 24, 28, and 38 respectively (*A box and whisker plot is to be added here later*).

A *scatter plot*. The plots discussed before were for univariate data. For bivariate data, a scatter plot is helpful. Here, we plot one variable against the other by representing them as 2-dimensional points. We also discuss the linear relationship in this case.

2.2 Numerical Measures

We first discuss some central values.

- The *mean*. It is the average value (the arithmetic mean) of the sample.
- The *median*. It is that value the splits the data in half; half the data points are below this value, while the other half are above it.
- The *mode*. It is that value which occurs with the highest frequency. Often, for larger samples, a modal class is more meaningful.

Note that the mean and median do not make sense for qualitative data. In contrast to this, the mode can be used for both qualitative and quantitative data. The median is sometimes used in qualitative data when it is ordinal.

Example 2.1. For a set of data points $x_1, \ldots, x_n$, the mean is defined as $\overline{x} = \left(\sum x_i\right)/n$. Show that, for $a = \overline{x}$, the sum of squared deviations of the data points from the value a is minimized, that is, the value $\sum (x_i - a)^2$ is minimized.

Proof. We have

$$\begin{aligned}\sum (x_i - a)^2 &= \sum (x_i - \overline{x} + \overline{x} - a)^2 = \sum (x_i - \overline{x})^2 + \sum (\overline{x} - a)^2 + 2\sum (x_i - \overline{x})(\overline{x} - a) \\ &= \sum (x_i - \overline{x})^2 + n(\overline{x} - a)^2 + 2(\overline{x} - a)\sum (x_i - \overline{x}) \\ &= \sum (x_i - \overline{x})^2 + n(\overline{x} - a)^2.\end{aligned}$$

Now $(\overline{x} - a)^2 \geq 0$; it is 0 if and only if $a = \overline{x}$. ■

January 16th.

We also have another result that says that $\sum |x_i - a|$ is minimized when the value of $a = \text{median}(x_1, \ldots, x_n)$.

2.2.1 Variability or Spread

We also are interested in the *variability* or the *spread*; how far the observations or data points are from each other or from the centre.

- One is such measure is the *range* defined as $\max - \min$. The range is a simple and intuitive measure of spread. Between different samples from the same population, the value of the range can be very different. It depends on only two values.
- Another measure if the *variance*. For n x_i observations, we have

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.1}$$

  as the variance; it is the average squared distance of eah point from the centre. Dividing by $n-1$ makes the measure 'good' as it is unbiased for the population variance.
- If we define x_M to be the median of the x_i's, we can define another measure: *MAD from the median*.

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x_M|\,. \tag{2.2}$$

  MAD is short for *mean absolute deviation*.
- The *standard deviation* may be defined as the square root of the variance.
- The *interquartile range* or IQR. It is a robust measure; applicable if there are a few very high or very low values. It is defined as

$$\text{IQR} = Q_3 - Q_1 \tag{2.3}$$

  where Q_3 is the third quartile, the value for which 75% of the data resides below it, and Q_1 is the first quartile, the value for which 25% of the data resides below it.

Example 2.2. Let us work with the following dataset—

$$40, 60, 65, 65, 65, 68, 68, 70, 70, 70, 70, 70, 70, 74, 75, 75, 90, 95$$

For this dataset, the median is the $\frac{1}{2}(n+1)$th observation, or the 9.5th observation: $70+0.5(70-70) = 0$.
Q_1 will be the $\frac{1}{4}(n+1)$th observation, or the 4.75th observation: $65 + 0.75(65-65) = 65$.
Q_3 will be the $\frac{3}{4}(n+1)$th observation, or the 14.25th observation: $74 + 0.25(75-74) = 74.25$.

The Chebyshev Inequality

The *Chebyshev inequality* can be summarised as

$$P\left(\left|\frac{X-\mu}{\sigma}\right|\right) \leq \frac{1}{k^2} \tag{2.4}$$

where μ is the population mean and σ is the population standard deviation. Along with this, we also adopt the convention $\overline{X}$ for the sample mean, and s for the sample standard deviation. We can also infer that 2 standard deviations worth of values around the mean contain, at the very least, three-fourths of the values.

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4}. \tag{2.5}$$

Note that this is true for all distributions. Similarly, we have $P(\mu - 3\sigma < X < \mu + 3\sigma) \geq \frac{8}{9}$. If we have $\overline{X}$ and s, then we can say that $\frac{8}{9}$ths of the population values will lie between $\overline{X} - 3s$ and $\overline{X} + 3s$, approximately. In practice, the coverage is more. For example, if have the normal distribution, 95% of the values will lie in between $\mu \pm 2\sigma$. Chebyshev guarantees that 75% of the values will lie here for any distribution.

2.2.2 Measures of Relative Standing

These measure look at the 'position' of a particular value relative to other, or in comparison to others.

- A *rank* is simply ordering of the population, and numbering them in order. The number assigned is the rank.
- A *percentile* or *quartile* also exists; it is defined as

$$\text{Quartile of value } a = \frac{\#\{x_i \leq a\}}{n}. \tag{2.6}$$

- The *z-score* measures how many standard deviations above the mean is the value a—

$$\text{z-score of } a = \frac{a - \overline{x}}{s}. \tag{2.7}$$

Detecting Outliers

The z-score is helpful it detecting outliers; if the absolute value of the z-score of a value a is above 3, there is a high likelihood it is an outlier.

For a boxplot, a good detection system is to check for those values greater than $Q_3 + 1.5\text{IQR}$ or less than $Q_1 - 1.5\text{IQR}$.

Appendices

Chapter A

Appendix

Extra content goes here.

Appendix

Index