

Introduction

1

Statistics - is the science of *learning from data*. It uses the theory of probability to make inferences about populations or processes using data – data that are both limited in numbers as well as *variable*.

* * * Variability * * *

Typically, no two objects from the same population are identical, regardless of whether they are man made or naturally occurring.

Examples of objects or characteristics of objects that vary:

1. Humans, dogs of same breed, fish of same species.
2. Lifetime of light bulbs or diameter of ball bearings.
3. Potency of drugs available in bottles.

For example, we would like to know the mean of a characteristic of a set of elements of a population, but we can't observe or measure each one because there are too many elements or not all are available.

Examples:

1. Mean blood pressure of 40 year old pregnant females.
2. Mean lifetime of 60-watt bulbs being made at the G.E. plant in Cleveland.
3. Mean potency of an antibiotic after storage for 20 months on store shelves.

Since we can't obtain the average exactly, we do the next best thing and approximate it by finding the average for a selected subset of the elements, i.e., we do a statistical study. Usually, there are many objectives of a statistical study, than estimating the mean of the population.

A Statistical Study – consists of:

1. **Collecting data** - by making observations or taking measurements on some sample or process of interest, possibly as a part of a statistically designed experiment or a survey.

2. **Summarizing the data** - using statistical summary methods (calculating means or constructing a histogram, for example).

3. **Analyzing the data and making inferences** - using models and statistical methods for drawing conclusions about the situation being considered.

Definitions:

Population - The set of all objects (or measurements) of interest; they are called elements of the population.

Sample - Any subset of elements from a population.

A statistical study is made using a sample from a population.

How to Select a Sample?

3

Theoretically we get a random sample of n elements by:

- making n draws
- on each draw, each remaining element in the population is equally likely to be the one drawn.

(The draws are supposed to be done without replacement of the elements already drawn - but it is immaterial from practical standpoint for very large populations).

Practically we can seldom satisfy the theoretical requirements, so we do the best we can to introduce randomness into our sample selection process.

Can the result of a statistical study of population mean (average) be far away from the true mean? - sure, sometimes - a statistical study provides evidence. It doesn't prove anything.

Example:

Consider an example where 50 different samples of size 10 are drawn from the same population (this is equivalent to looking at the data from several different statistical studies, all having an objective to estimate the population mean.)

7.71 13.53 5.72 6.06 9.85
6.56 13.35 13.52 11.58 10.02
11.91 13.39 13.01 10.06 13.94
3.95 10.27 8.30 10.72 8.08
8.56 13.75 9.27 9.88 8.07
9.86 9.65 6.65 9.23 9.15
6.27 8.34 9.74 8.23 10.65
11.67 9.75 12.00 10.04 7.34
8.72 10.80 9.30 7.28 9.68
13.43 11.97 8.15 15.28 7.32

4

Each sample mean is an estimate of the mean of the population elements from which the sample was taken (call it the sampled population.) A summary is:

(3, 4] |
(4, 5] |
(5, 6] |
(6, 7] ||||
(7, 8] ||||
(8, 9] |||||
(9, 10] |||||
(10, 11] |||||
(11, 12] |||||
(12, 13] |||||
(13, 13] |||||
(14, 15] |||||
(15, 16] |

Looking at the summary table we see that the middle of the set of sample means is in the interval (9, 10]. The mean (average) of the 50 sample means is 9.83. This is our best estimate of the population mean (which is known to be equal to 10).

The above graphic shows the **sampling distribution** of the sample mean. The variability or spread of the means reflects the variation among the sampled population elements. It is measured by the variance of the sample means around the central value above.

The procedure of selecting a random sample described earlier ensures that each sample of size n drawn from a population has the same chance (or probability) of being selected. Methods that ensure this property are called **simple random sampling**.

Data Description

6

Two broad applications of Statistics

- descriptive statistics
- inferential statistics

When measurements from an entire population is available to us data description will be a major objective.

In this case, methods for organizing, summarizing, and describing data are needed. Good descriptive statistics enable us to make sense of the data by reducing large amounts of data to a few summary measures and graphical displays.

Example: Monthly, quarterly, yearly data on medical costs collected by HMO's on many variables such as type of illness, age of patient, prescription costs, physician charges, inpatient or out-patient care etc. This type of data requires extensive use of descriptive statistics to be useful to consumers of such data.

When only a random sample is available from a large population (or a process), inferential statistics becomes the major focus. However, even in this case, descriptive statistics of the sample data is still important since they are used to draw conclusions about the population from which the sample was taken.

Graphical Methods

7

- Bar Chart (Section 3.3)
- Histogram (Section 3.3)
- Stem-and-Leaf plot (Section 3.3)
- Time Series plot (Section 3.3)
- Boxplot (Section 3.6)
- Normal Probability Plot(Notes)

Numerical Summaries

Measures of Central Tendency (Section 3.4)

- Mode
- Median
- Mean

Measures of Variability (Section 3.5)

- Range
- Percentile (Quantile, Quartile)
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Histograms

8

To plot a histogram, a **frequency** table must first be constructed. The range of the data is divided into **class intervals**, the number of such intervals determined by the number of observations in the data set.

Example:

Weight gains of chicks fed an antibiotic(p.47)

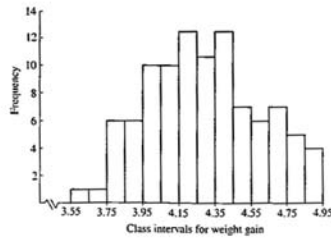
3.7	4.2	4.4	4.4	4.3	4.2	4.4	4.8	4.9	4.4
4.2	3.8	4.2	4.4	4.6	3.9	4.1	4.5	4.8	3.9
4.7	4.2	4.2	4.8	4.5	3.6	4.1	4.3	3.9	4.2
4.0	4.2	4.0	4.5	4.4	4.1	4.0	4.0	3.8	4.6
4.9	3.8	4.3	4.3	3.9	3.8	4.7	3.9	4.0	4.2
4.3	4.7	4.1	4.0	4.6	4.4	4.6	4.4	4.9	4.4
4.0	3.9	4.5	4.3	3.8	4.1	4.3	4.2	4.5	4.4
4.2	4.7	3.8	4.5	4.0	4.2	4.1	4.0	4.7	4.1
4.7	4.1	4.8	4.1	4.3	4.7	4.2	4.1	4.4	4.8
4.1	4.9	4.3	4.4	4.4	4.3	4.6	4.5	4.6	4.0

It was decided to have 10 class intervals. Since the range is 1.3, a class interval width of $1.3/10 \approx .13$ was used. The end-points of the intervals were selected to be at one-half the unit of measurements (.1 gram) so no observation falls on any end-point. Since the smallest observation is 3.6 we start with the interval 3.55 – 3.65. Once the class intervals have been determined observations in the data set falling into each of these classes are tallied:

TABLE 3.3
Frequency table for the chick data

Class	Class Interval	Frequency f_i	Relative frequency f_i/n
1	3.55-3.65	1	.01
2	3.65-3.75	1	.01
3	3.75-3.85	6	.06
4	3.85-3.95	6	.06
5	3.95-4.05	10	.10
6	4.05-4.15	10	.10
7	4.15-4.25	13	.13
8	4.25-4.35	11	.11
9	4.35-4.45	13	.13
10	4.45-4.55	7	.07
11	4.55-4.65	6	.06
12	4.65-4.75	7	.07
13	4.75-4.85	5	.05
14	4.85-4.95	4	.04
Totals		$n = 100$	1.00

FIGURE 3.7(a)
Frequency histogram for the chick data of Table 3.3



Stem-and-Leaf Plots

These are useful as an alternative to histograms or frequency tables. The idea is to substitute digits for frequency counts used in constructing histograms and in the process convey more information. Its profile is much like that of a histogram but it shows the numerical values of all observations.

To construct a stem-and-leaf diagram, group each data value into two parts: a set of **leading digits** and a set of **trailing digits**. Then first write down, to the left of a vertical line, all possible values of the leading digits. Then we represent each data value by writing its trailing digits in the appropriate row to the right of the line.

The column of digits on the left of the vertical line represents the 'stem' and the digits on the right the 'leaves' of the stem-and-leaf plot. Thus each data value can be reconstructed by combining its stem part and its leaf part. The leaves in each stem are usually written successively (and usually contiguously) in the increasing order of magnitude.

In those cases where the data values may not be adequately represented by just taking the last digits as leaves the data values must be appropriately rounded before a stem-and-leaf diagram can be constructed.

Example

Below are the measured manganese content (in points or .01%'s) for 20 heats of 1045 steel taken from a text by Burr.

74 79 77 81
68 79 81 76
81 80 80 78
88 83 79 91
79 75 74 73

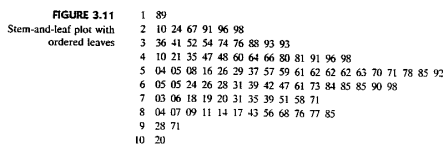
A stem and leaf diagram for these data is:

```

6 | 8
7 | 344
7 | 56789999
8 | 001113
8 | 8
9 | 1
    
```

Example:

Stem-and-leaf plot of crime data (Table 3.4) from text.



Some sets of observations have a very narrow range that the stem and leaf has only 2 or 3 stems. Such a plot is not useful for studying the distribution of the data. One solution is to attempt to increase the number of stems, by dividing the data into subgroups within each original stems i.e., to use each stem value twice or more. As an example, consider the following stem a leaf plot of height in inches of students in a statistics class taken from Koopmans (1981):

```

5 | 47
6 | 02333444555556666677777889
7 | 00112223
    
```

Obviously, this does not represent a satisfactory visual display of the distribution of the data. Suppose each stem is divided into two parts: one to represent the leaves from 0 through 4 and the other for the leaves from 5 through 9. Thus we have the following stem-and-leaf plot:

```

5a | 4
5b | 7
6a | 02333444
6b | 555556666677777889
7a | 00112223
    
```

This may even be further improved by repeating each stem five times to represent two successive leading digits:

```

5(01) |
(23)
(45) 4
(67) 7
(89)
6(01) 0
(23) 2333
(45) 44455555
(67) 6666677777
(89) 889
7(01) 0011
7(23) 2223

```

The JMP **Analyze**→**Distribution** produced the following stem-and-leaf plot for the crime data:

```

10 | 2          1
9   | 37        2
8   | 011112467889 12
7   | 001222344567 12
6   | 1123334445678999 15
5   | 000112334666666677899 21
4   | 12455667889   11
3   | 00445578999   11
2   | 1279          4
1   | 9            1

```

1 | 9 represents 190

Quantiles and Percentiles

As you know, the 80th percentile is a data value such that approximately 80% of the data values in the data set are less, and approximately 20% are larger, than that value.

For a theoretical distribution (e.g. normal distribution) the .8 quantile is a value such that 80% of the probability mass of the distribution lies to the left and 20% lies to the right of the value of the quantile.

For an empirical distribution (i.e., the distribution of a data set) we need another definition. Suppose we have only 10 values in a dataset:

52, 63, 67, 71, 75, 76, 76, 82, 87, 94

What is the 80th percentile? i.e., what value is such that approximately 80% are less and 20% are greater? Since there are 10 values, need to find a number such that 8 values are less, and 2 values greater than that number. There isn't a value in the dataset that satisfies both conditions. A number half way between 82 and 87, i.e., 84.5, might be reasonably chosen as the 80th percentile. It satisfies the definition approximately.

A definition that will give unambiguous percentiles for any set of data is needed. Let p ($0 < p < 100$) denote the desired percentile, n the number of data values, and let $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ denote the ordered observations so that

$$y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$$

Definition: Percentiles and Quantiles

1. If p is one of the numbers $100(i - .5)/n$ then $y_{(i)}$ is the p^{th} percentile, for $i = 1, 2, 3, \dots, n-1$ or n .
2. If p is between $100(0.5/n)$ and $100(n - .5)/n$, and p is not one of the numbers $100(i - .5)/n$ then the percentile is obtained by linear interpolation between the two percentiles that bracket p .

Quantiles are a generalization of percentiles. The p^{th} quantile of a dataset is the $100p^{\text{th}}$ percentile. Thus the definition of p^{th} quantile of the above dataset with ($0 < p < 1$) is the following:

1. If p is one of the numbers $(i - .5)/n$ then $y_{(i)}$ is the p^{th} quantile, for $i = 1, 2, 3, \dots, n-1$ or n .
2. If p is between $(0.5/n)$ and $(n - .5)/n$, and p is not one of the numbers $(i - .5)/n$ then the quantile is obtained by linear interpolation between the two quantiles that bracket p .

The p^{th} quantile of a dataset is denoted by $Q(p_i)$

Example:

Suppose we have a dataset consisting of the 10 observations:

9.614, 9.614, 10.688, 7.583, 8.572, 8.527, 8.577, 9.471, 9.165, 9.011

The following table shows the ordered data values as quantiles:

Quantiles of the above data set

i	$p_i = (i - .5)/10$	$Q(p_i)$
1	0.05	7.583
2	0.15	8.527
3	0.25	8.572
4	0.35	8.577
5	0.45	9.011
6	0.55	9.165
7	0.65	9.471
8	0.75	9.614
9	0.85	9.614
10	0.95	10.688

$Q(p)$ for values of $p = (i - .5)/n$ correspond to the ordered observations for $i = 1, 2, \dots, n$. For example $Q(.55) = 9.165$. How is, say, $Q(.525)$ determined? This is obtained by linear interpolation between the two quantiles that bracket .525 i.e. $Q(.45)$ and $Q(.55)$.

For a p such that $p_i < p < p_{i+1}$, $Q(p)$ is computed as the weighted average $Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1})$ where $f = (p - p_i)/(p_{i+1} - p_i) = n(p - p_i)$

Example

Suppose $n = 10$ and one needs to compute $Q(.525)$. Since $.45 < p < .55$, $f = 10(.525 - .45) = .75$. Thus $Q(.525) = .25Q(.45) + .75Q(.55)$.

In the above example, therefore

$$Q(.525) = .25(9.011) + .75(9.165) = 9.1265$$

Box Plots (Tukey, 1977)

First a few more definitions of terms are needed

Upper Quartile = 75th percentile = $Q(.75) = Q_3$

Median = 50th percentile = $Q(.50) = M$

Lower Quartile = 25th percentile = $Q(.25) = Q_1$

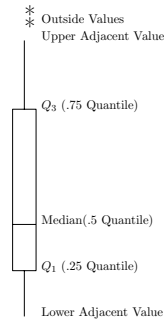
Interquartile Range (IQR) =

$$\text{Upper Quartile} - \text{Lower Quartile} = Q_3 - Q_1$$

The **boxplot** is a summary display of a set of data, that is useful for studying the shape of the distribution including its symmetry or asymmetry around the central location, based on the quantiles Q_1 , M , and Q_3 . The following quantities also need to be computed for constructing a box plot.

1. Upper Adjacent Value: Largest observation that is less than or equal to $Q_3 + 1.5 \text{ IQR}$.
2. Lower Adjacent Value: Smallest observation that is greater than or equal to $Q_1 - 1.5 \text{ IQR}$.
3. Outside Values: Observations that fall outside adjacent values.

These quantities are computed then used to construct the diagram shown below:



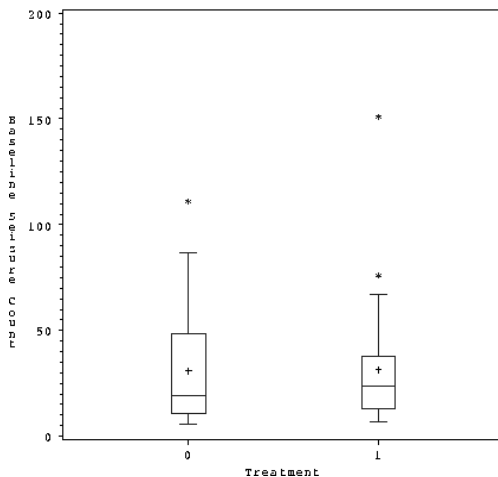
The SAS procedure UNIVARIATE produced the following boxplot for the crime data in Table 3.4 of your text.:



Note that $Q_1 = 464$, $Q_3 = 719$ and the median = 574.5 are calculated in your text.

Side-by-side boxplots are useful comparing distributions of variables across subsamples defined by values of a categorical variable. Table 3.10 of the text shows data for patients from 59 epileptics that were randomly assigned to receive the anti-epileptic drug Progabide or a placebo. The side-by-side boxplots of baseline seizure rates for each of the two groups are shown below:

Box Plots of Baseline Seizure Count by Treatment



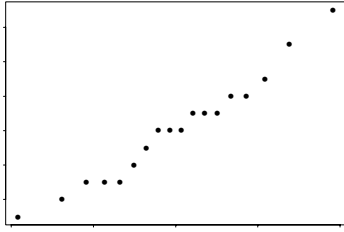
Normal Probability Plot

Consider a sample y_1, y_2, \dots, y_n of size n . The data are first ordered in increasing order of magnitude then a value p is assigned to each data value using the formula $p = (i - 0.5)/n$. Thus the ordered data values will be the p_i^{th} quantile $Q(p_i)$ for $i = 1, 2, \dots, n$.

Example n = 18

i	y_i	$p_i = (i - 0.5)/18$	z_p
1	57	1/36 = 0.0278	-1.9145
2	58	3/36 = 0.0833	-1.3830
3	59	5/36 = 0.1389	-1.0853
4	59	7/36 = 0.1944	-0.8616
5	59	9/36 = 0.2500	-0.6745
6	60	11/36 = 0.3055	-0.5085
7	61	13/36 = 0.3611	-0.3555
8	62	15/36 = 0.4167	-0.2104
9	62	17/36 = 0.4722	-0.06968
10	62	19/36 = 0.5278	0.06968
11	63	21/36 = 0.5833	0.2104
12	63	23/36 = 0.6389	0.3555
13	63	25/36 = 0.6944	0.5085
14	64	27/36 = 0.7500	0.6745
15	64	29/36 = 0.8056	0.8616
16	65	31/36 = 0.8611	1.0853
17	67	33/36 = 0.9167	1.3830
18	69	35/36 = 0.9722	1.9145

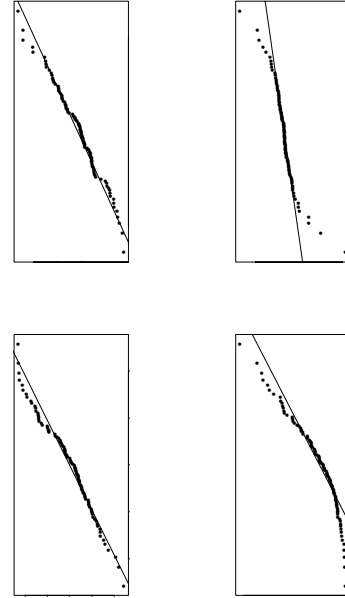
The quantile of the standard normal distribution, z_p , corresponding to each p value, is then computed using the table of Normal percentiles (i.e. the z -table). The Normal Probability plot is the plot of the quantiles $Q(p)$ of the data vs. z_p , the corresponding standard normal quantiles. (Thus this plot is also called the quantile-quantile plot or the Q-Q plot).



If the data is a random sample from a normal population the plotted points will lie approximately in a straight-line. Departures from a straight-line may indicate that the population distribution is different from a Normal distribution. Specific types of departures can be used to identify how the population distribution differs from a Normal distribution.

The following page contains normal probability plots of computer generated data from populations that resemble distributions that are as described. The top set of plots are from normal populations with different mean and variance parameters while the bottom two are from populations that differ from the normal distribution. Notice that the bottom two plots show specific patterns (the left graph shows a bowl-shape while the other shows an inverted S-shaped pattern). These are markedly dif-

ferent from the deviation from a straight line shown in the top graphs which are only due to random variation.



Sample Statistics vs. Parameters

A **parameter** is a descriptive statistic of a population.

A **sample statistic** is a descriptive statistic of a sample.

Take a *census* of a population recording values of a variable as x_1, x_2, \dots, x_N , where N is the population size, then

- Population mean, $\mu = \frac{\sum x}{N}$
- Population variance, $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
- Population standard deviation, $\sigma = \sqrt{\sigma^2}$

Since censuses are not taken for every population we wish to study, parameters cannot be exactly calculated and thus usually unknown.

Take a *sample* of a population recording a values of a variable as x_1, x_2, \dots, x_n , where n is the sample size, recording a values as

- sample mean, $\bar{x} = \frac{\sum x}{n}$
- sample variance, $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
- sample standard deviation, $s = \sqrt{s^2}$

μ , σ^2 , and σ are **parameters**.

\bar{x} , s^2 , and s are **sample statistics**.

An Example

Calculate the sample mean, variance, and standard deviation for data: 2, 3, 3, 4, 3

i	x	x^2
1	2	4
2	3	9
3	3	9
4	4	16
5	3	9
$\sum x = 15$		$\sum x^2 = 47$

Compute sample variance s^2 and sample standard deviation s .

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{47 - \frac{(15)^2}{5}}{5-1} = 2/4 = 0.5$$

$$s = \sqrt{(s^2)} = \sqrt{0.5} = 0.71$$

The sample mean and the sample median are examples of **numerical descriptive measures of the central tendency** of a sample and describes the center or location around which the data are distributed.

The sample variance and standard deviation are examples of **numerical descriptive measures of the variability** or the spread of the sample around the center.

These numerical descriptive measures are used both to describe a population, if the entire population of measurements is available, (e.g., census) or to draw inferences about a population from the statistics calculated on a random sample drawn from the population.