

LECTURE 1: SAMPLING AND DESCRIPTIVE STATISTICS

ECEN 321

Engineering Statistics



VICTORIA UNIVERSITY OF
WELLINGTON
TE HERENGA WAKA

Why Statistics?



- Deal with uncertainty in repeated scientific measurements
- Draw conclusions from data
- Design valid experiments and draw reliable conclusions
- Be a well-informed member of society

Example 1

Consider a machine that makes steel rods for use in optical storage devices. The specification for the diameter of the rods is 0.45 ± 0.02 cm. During the last hour, the machine has made 1000 rods. The quality engineer wants to know approximately how many of these rods meet the specification. He does not have time to measure all 1000 rods. So he draws a random sample of 50 rods, measures them, and finds that 46 of them (92%) meet the diameter specification. Now, it is unlikely that the sample of 50 rods represents the population of 1000 perfectly.

Example 1

The engineer might need to answer several questions based on the sample data. For example:

1. How large is a typical difference for this kind of sample?
2. What interval gives a good estimate of the percentage of acceptable rods in the population with reasonable certainty?
3. How certain can the engineer be that at least 90% of the rods are good?

Statistics can help us to address questions like these.

Section 1.1: Sampling

Definitions:

- A **population** is the entire collection of objects or outcomes about which information is sought.
- A **sample** is a subset of a population, containing the objects or outcomes that are actually observed.
- A **simple random sample (SRS)** of size n is a sample chosen by a method in which each collection of n population items is equally likely to comprise the sample, just as in the lottery.

Example 2



At a large university, there is a professor who is interested in the average height of students at the university. She obtains a list of all 50,000 students enrolled at the university and assigns a number to each student. A random number generator generates 100 numbers, and the students corresponding to those numbers are selected to have their height measured. This is a simple random sample.

Sampling (cont.)

Definition: A **sample of convenience** is a sample that is not drawn by a well-defined random method.

Things to consider :

- Convenience samples may differ systematically in some way from the population.
- Thus, they should only be used when it is not feasible to draw a random sample.
- If you do have a convenience sample, think carefully about the ways in which the sample might differ systematically from the population.

Example 3

A construction engineer has received a shipment of 1000 concrete blocks, each weighing approximately 50 pounds. The blocks are in a large pile. The engineer wishes to investigate the crushing strength of the blocks by measuring the strengths in a sample of 10 blocks. It may be difficult to take a SRS since that would involve getting blocks from the middle and bottom of the pile, so the engineer may just take 10 off the top. This would be a sample of convenience.

Simple Random Sampling

- A SRS is not guaranteed to reflect the population perfectly.
- SRS's always differ in some ways from each other; occasionally a sample is substantially different from the population.
- Two different samples from the same population will vary from each other as well.
- This phenomenon is known as **sampling variation**.

Example 1 cont.

- Remember the rod example?
- In the sample the engineer collected, there were 92% that met specification.
- In the population of all 1000 rods, it is unlikely that exactly 92% will meet specification.
- Because of sampling variation, it is more realistic to think that the true proportion of rods that meet specification will be close to the sample proportion, or 92%.

More on SRS

Definition: A **conceptual population** consists of all of the values that might possibly have been observed from a population.

- For example, a geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a slightly different reading.
- Here the population is conceptual. It consists of all the readings that the scale could in principle produce.

SRS (cont.)

- The items in a sample are said to be **independent** if knowing the values of some of the items does not help to predict the values of the others.
- Items in a simple random sample may be treated as independent in most cases encountered in practice. The exception occurs when the population is finite and the sample comprises a substantial fraction (more than 5%) of the population.

Other Sampling Methods

- Samples other than simple random samples can be useful in various situations.
- These include (but are not limited to):
 - ▣ Weighted sampling
 - ▣ Stratified random sampling
 - ▣ Cluster sampling

Types of Data

- When a numerical quantity is assigned to each item in the sample, the resulting set of values is **numerical or quantitative**.
 - Height (in centimeters)
 - Weight (in kilograms)
 - Age (in years)

- When sample items are placed into categories and category names are assigned to the sample items, the data are **categorical or qualitative**.
 - Hair color
 - Country of origin
 - Zip code

Controlled Experiments and Observational Studies

- A **controlled experiment** is one in which the experimenter controls the values of the factors.
 - ▣ When designed and conducted properly, controlled experiments can produce reliable information about cause-and-effect relationships between factors and response.

Controlled Experiments and Observational Studies

- An **observational study** is one where the experimenter simply observes the factors as they are, without having any control over them.
- ▣ Observational studies are not nearly as good as controlled experiments for obtaining reliable conclusions regarding cause and effect.

Section 1.2: Summary Statistics

Let X_1, \dots, X_n be a sample.

□ **Sample Mean:**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

□ **Sample Variance:**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

□ **Sample standard deviation** is the square root of the sample variance.

More on Summary Statistics

- If X_1, \dots, X_n is a sample, and $Y_i = a + bX_i$, where a and b are constants, then

$$\bar{Y} = a + b\bar{X}$$

- If X_1, \dots, X_n is a sample, and $Y_i = a + bX_i$, where a and b are constants, then

$$s_y^2 = b^2 s_x^2, \text{ and } s_y = |b| s_x.$$

Outliers

- **Outliers** are points that are much larger or smaller than the rest of the sample points.
- Outliers may be data entry errors or they may be points that really are different from the rest.
- Outliers should not be deleted without considerable thought—sometimes calculations and analyses will be done with and without outliers and then compared.

Definition of a Median

The **median** is another measure of center, like the mean.

Order the n data points from smallest to largest. Then

➤ If n is odd, the sample median is the number in

$$\text{position } \frac{n+1}{2}.$$

➤ If n is even, the sample median is the average

of the numbers in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

Quartiles

- ❖ **Quartiles** divide the data as nearly as possible into quarters.
- ❖ The **first quartile** is the median of the lower half of the data.
 - ❖ To find the first quartile, compute $0.25(n + 1)$. If this is an integer, then the sample value in that position is the first quartile. If not, take the average of the sample values on either side of this value.

Quartiles

- ❖ The **third quartile** is the median of the upper half of the data.
 - ❖ To find the first quartile, compute $0.75(n + 1)$. If this is an integer, then the sample value in that position is the first quartile. If not, take the average of the sample values on either side of this value.
- ❖ Note: The computation we used for the location of the median is equivalent to $0.5(n + 1)$. The median is the second quartile.

Definition of Percentile

- The ***p*th percentile** of a sample, for a number p between 0 and 100, divides the sample so that as nearly as possible $p\%$ of the sample values are less than the p th percentile, and $(100 - p\%)$ are greater.
- The computation of the location of the p th percentile is analogous to what we did for the quartiles.

To Find Percentiles

- Order the n sample values from smallest to largest.
- Compute the quantity $(p/100)(n + 1)$, where n is the sample size.
- If this quantity is an integer, the sample value in this position is the p th percentile. Otherwise, average the two sample values on either side.

Note on Percentiles

- The **first quartile** is the 25th percentile.
- The **median** is the 50th percentile.
- The **third quartile** is the 75th percentile.

Example 4

- Suppose we have the following data:

2, 3, 5, 6, 7, 9, 9, 11, 12, 15

- What is the mean of these data?
- What is the median?
- What is the first quartile?
- What is the third quartile?

Summary Statistics for Categorical Data

- The two most commonly used numerical summaries for categorical data are the **frequencies** and the **sample proportion** (sometimes called **relative frequencies**).
- Example: 100 rivets are checked for their breaking strength. If 4 of the rivets fail (i.e., do not hold up to the standard), find the sample proportion of rivets that fail.

Sample Statistics and Population Parameters

- A numerical summary of a sample is called a **statistic**.
- A numerical summary of a population is called a **parameter**.
- Statistics are often used to estimate parameters.

Section 1.3: Graphical Summaries

- Stem-and-leaf plot
- Dotplot
- Histogram
- Boxplot
- Scatterplot

Stem-and-Leaf Plot

- A simple way to summarize a data set.
- Each item in the sample is divided into two parts: a **stem**, consisting of the leftmost one or two digits, and the **leaf**, which consists of the next digit.
- It is a compact way to represent the data.
- It also gives us some indication of the shape of our data.

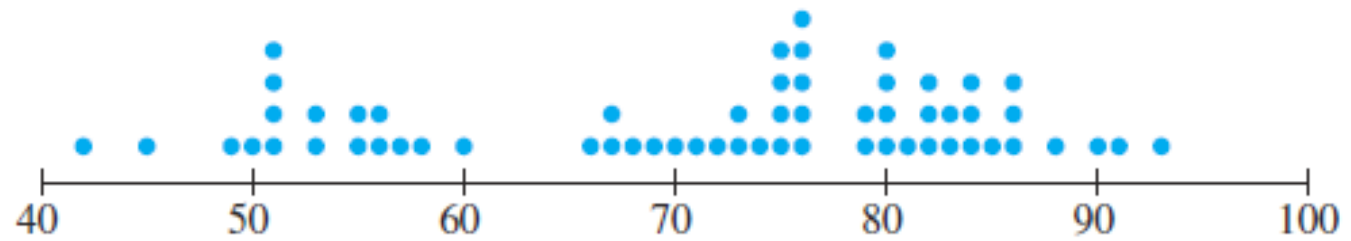
Example 5

- Example: Duration of dormant periods of the geyser Old Faithful in Minutes
- Stem-and-leaf plot:

```
4  259
5  0111133556678
6  067789
7  01233455556666699
8  000012223344456668
9  013
```

- Let's look at the first line of the stem-and-leaf plot. This represents measurements of 42, 45, and 49 minutes.
- A good feature of these plots is that they display all the sample values. One can reconstruct the data in its entirety from a stem-and-leaf plot.

- A **dotplot** is a graph that can be used to give a rough impression of the shape of a sample.
- It is useful when the sample size is not too large and when the sample contains some repeated values.
- It is a good method, along with the stem-and-leaf plot, to informally examine a sample.
- Not generally used in formal presentations.
- Example (geyser data):



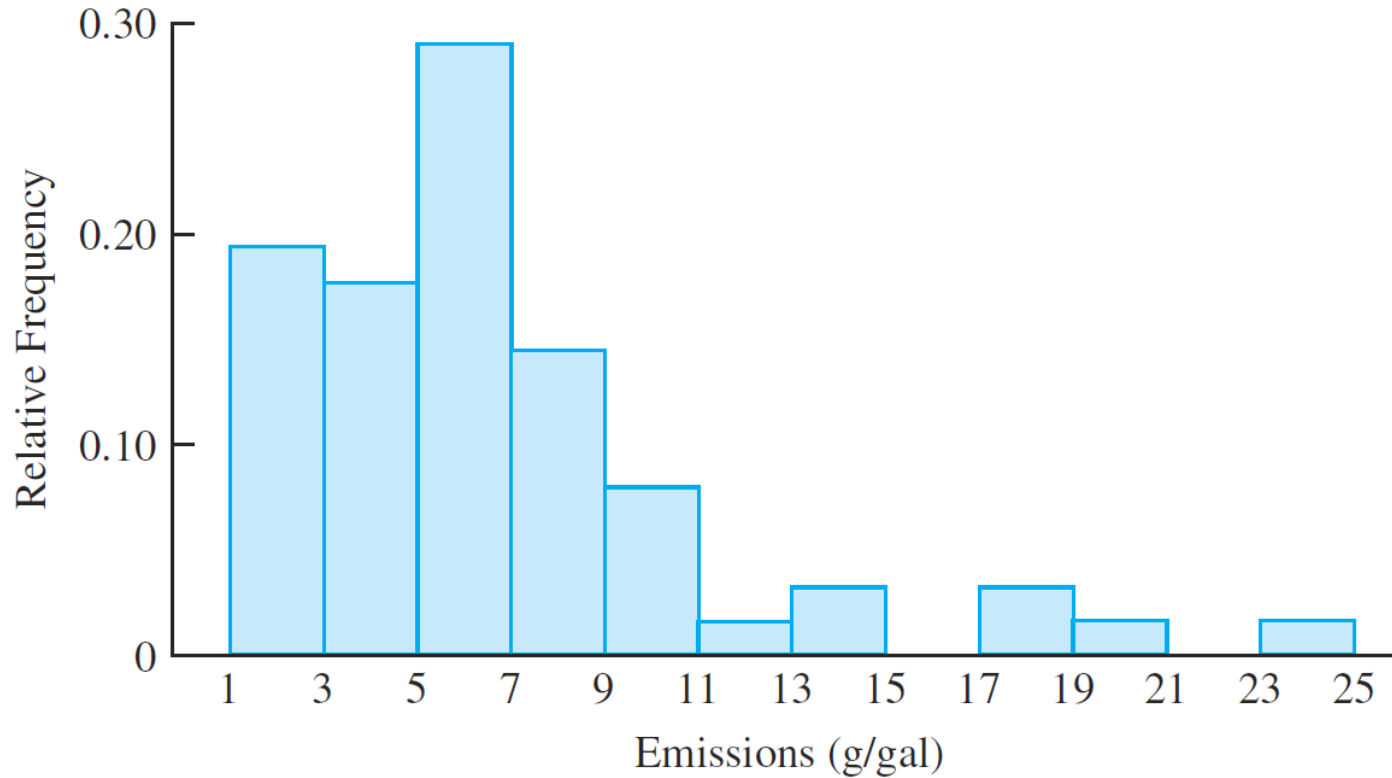
Histogram

- Graphical display that gives an idea of the “shape” of the sample.
- We want a reasonable number of observations in each interval.
- The bars of the histogram touch each other. A space indicates that there are no observations in that interval.

Creating a Histogram

- Choose boundary points for the class intervals. Usually these intervals are the same width.
- Compute the frequencies: this is the number of observations that occur in each interval.
- Compute the relative frequencies for each class: this is the number of observations in each interval divided by the total number of observations.
- If the class intervals are the same width, then draw a rectangle for each class, whose height is equal to the frequencies or relative frequencies.
 - ▣ If the class intervals are of unequal widths, the heights of the rectangles must be set equal to the densities, where **density** is the relative frequency divided by the class width.

Example of Histogram



Symmetry and Skewness

- A histogram is perfectly **symmetric** if its right half is a mirror image of its left half.
 - ▣ For example, heights of randomly selected men are roughly symmetric
- Histograms that are not symmetric are referred to as **skewed**.
- A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**.
 - ▣ For example, incomes are right skewed.
- A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**.
 - ▣ For example, grades on an easy test are left skewed.

Symmetry and Skewness

- When a histogram is roughly **symmetric**, the mean and the median are approximately equal.
- When a histogram is **right-skewed**, the mean is greater than the median.
- When a histogram is **left-skewed**, the mean is less than the median.

Unimodal and Bimodal

- A histogram with only one peak is what we call **unimodal**.
- If a histogram has two peaks then we say that it is **bimodal**.
- If there are more than two peaks in a histogram, then it is said to be **multimodal**.

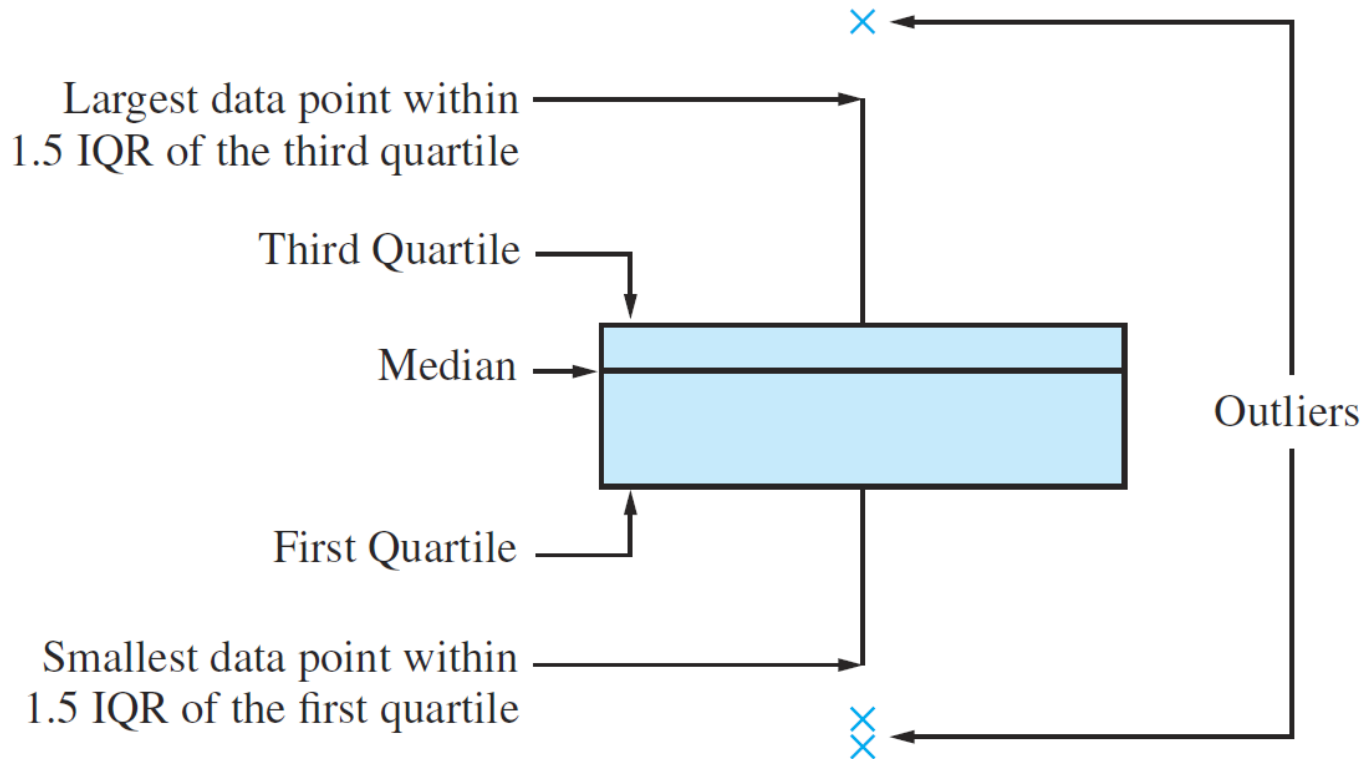
Boxplots

- A **boxplot** is a graphic that presents the median, the first and third quartiles, and any outliers present in the sample.
- The **interquartile range (IQR)** is the difference between the third quartile and the first quartile. This is the distance needed to span the middle half of the data.

Creating a Boxplot

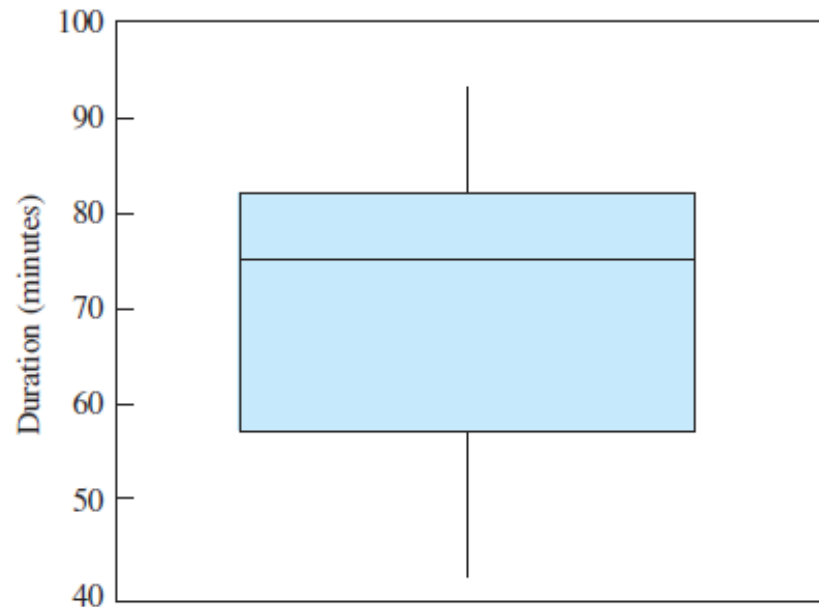
- Compute the median and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box.
- Find the largest sample value that is no more than 1.5 IQR above the third quartile, and the smallest sample value that is no more than 1.5 IQR below the first quartile. Extend vertical lines (whiskers) from the quartile lines to these points.
- Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually.

Anatomy of a Boxplot



Example 5 cont.

- ❖ Notice there are no outliers in these data.
- ❖ Looking at the four pieces of the boxplot, we can tell that the sample values are comparatively densely packed between the median and the third quartile.
- ❖ The lower whisker is a bit longer than the upper one, indicating that the data has a slightly longer lower tail than an upper tail.
- ❖ The distance between the first quartile and the median is greater than the distance between the median and the third quartile.
- ❖ This boxplot suggests that the data are skewed to the left.

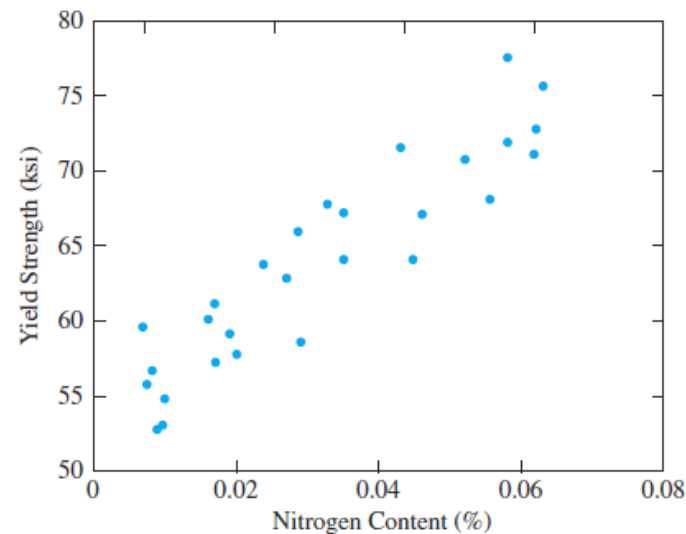


Comparative Boxplots

- Sometimes we want to compare more than one sample.
- We can place the boxplots of the two (or more) samples side-by-side.
- This will allow us to compare how the medians differ between samples, as well as the first and third quartile.
- It also tells us about the difference in spread between the two samples.

Scatterplot

- Data for which items consists of a pair of values is called **bivariate**.
- The graphical summary for bivariate data is a **scatterplot**.
- Display of a scatterplot:



Looking at Scatterplots

- If the dots on the scatterplot are spread out in “random scatter,” then the two variables are not well related to each other.
- If the dots on the scatterplot are spread around a straight line, then one variable may be used to help predict the value of the other variable.

Summary of Chapter 1

- We discussed types of data.
- We looked at sampling, mostly SRS.
- We learned about sample statistics.
- We examined graphical displays of data.