1

Basic Statistical Background

1.1 Introduction

Statistics and data analytics play a central role in improving processes and systems and in decision-making for strategic planning and manufacturing (Roberts, H. V., 1987). During experimental research, statistical tools can allow the experimenter to better organize observations, to specify working hypotheses and possible alternative hypotheses, to collect data efficiently, and to analyze the results and come to some conclusions about the hypotheses made.

This is an introductory chapter where readers can review several basic statistical concepts before moving on to the next chapters. Sixteen sections titled Stat Tools will introduce some key terms and procedures that will be further elaborated and referred to throughout the text.

Specifically, this chapter deals with the following:

Topics	Stat tools
Statistical variables and types of data	1.1
Statistical Units, populations, samples	1.2
Introduction to descriptive and inferential analyses	1.3, 1.12, 1.13
Data distributions	1.4, 1.5
Mean values	1.6, 1.7
Measures of variability	1.8, 1.9, 1.10
Boxplots	1.11
Introduction to confidence intervals	1.14
Introduction to hypothesis testing procedures, including the *p*-value approach	1.15, 1.16

COPYRIGHTED MATERIAL

End-to-End Data Analytics for Product Development: A Practical Guide for Fast Consumer Goods Companies, Chemical Industry and Processing Tools Manufacturers, First Edition.
Rosa Arboretti, Mattia De Dominicis, Chris Jones, and Luigi Salmaso.
© 2020 John Wiley & Sons Ltd. Published 2020 by John Wiley & Sons Ltd.
Companion website: www.wiley.com/go/salmaso/data-analytics-for-pd

Learning Objectives and Outcomes

Upon completion of the review of these basic statistical concepts, you should be able to do the following:

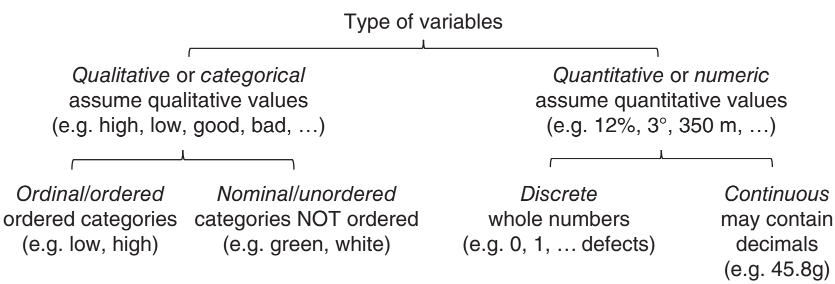
- Recognize and distinguish between different types of variables.
- Distinguish between a population and a sample and know the meaning of random sampling.
- Detect the shape of data distributions.
- Calculate and interpret descriptive measures (means, measures of variability).
- Understand the basic concept and interpretation of a confidence interval.
- Understand the general idea of hypothesis testing.
- Understand the *p*-value approach to hypothesis testing.

Stat Tool 1.1 Statistical Variables and Types of Data

In statistical studies, several characteristics are observed or measured to obtain information on a phenomenon of interest. The observed or measured characteristics are called *statistical variables*. Statistical variables differ according to the type of values they store.

Qualitative or *categorical* variables can assume values that are qualitative categories and can be either ordinal or nominal.

Quantitative or *numeric* variables can assume numeric values and can be discrete or continuous. Discrete data (or count data) are numerical values only measurable as integers. Continuous data are numeric values (typically instrumental measures) that can be meaningfully subdivided into fractions.



➢ *Example 1.1*. For a new shaving oil, it is of interest to compare fragrance A and fragrance B to investigate which is preferred. A sample of female respondents is presented with the two fragrances and asked their age and to answer the following question: How suitable or unsuitable is the fragrance to a shaving aid? Each was asked to assign an integer score from 0 (very unsuitable) to 10

Stat Tool 1.1 (Continued)

(very suitable). Respondents also expressed their purchase intentions for selecting one of the following categories: Probably would not buy it, Neither, Probably would buy it.

The variable "Fragrance" is a nominal categorical variable, assuming two different categories: A and B. The variable "Appropriateness" is a discrete quantitative variable assuming values from 0 to 10. "Age" of the respondents is a continuous quantitative variable, and "purchase intent" is an ordinal categorical variable assuming three different ordered categories.

In some contexts, you may find different terminology used to refer to similar data types. In quality control, categorical and discrete data are referred to as *attributes* and continuous data as *variables*.

When performing a statistical analysis, take into account the type of variable(s) you have, i.e. is it qualitative or quantitative? Different graphs, descriptive statistics, and inferential procedures must be used to study different types of data.

Stat Tool 1.2 Statistical Unit, Population, Sample

A *statistical unit* is the unit of observation (e.g. entity, person, object, product) for which data are collected. For each statistical unit, qualitative or quantitative variables are observed or measured.

The whole set of statistical units is the *population*. It may also be virtually infinite (e.g. all products of a production process).

A *sample* is a subset of statistical units (sampling units) selected from the population in a suitable way. The sample size of a study is the total number of sampling units (see Figure 1.1).

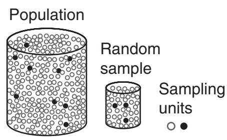


Figure 1.1 Population, samples, sampling units.

When we use a sample to draw conclusions about a population, sample selection must be performed *at random*. Random sampling is carried out in such a way as to ensure that no element in the population is given preference over any other. Random sampling is used to avoid nonrepresentative samples of the population.

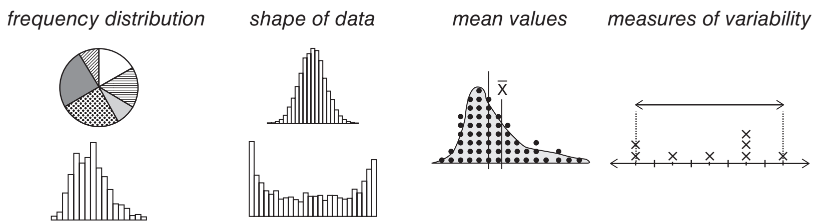
In *Example 1.1* the statistical unit is the single respondent.

Stat Tool 1.3 Descriptive and Inferential Analysis

Usually, the first step of a statistical analysis is *descriptive analysis*, where tables, graphs, and simple measures help to quickly assess and summarize important aspects of sample data.

When performing descriptive analysis, take into account the type of variable(s) present, i.e. is it qualitative (categorical) or quantitative? Different graphs, descriptive statistics, and inferential procedures have to be used to study different types of data.

The descriptive phase evaluates the following aspects:



After outlining important sample data characteristics through descriptive statistics, the second step of a statistical analysis is *inferential analysis*, where sample findings are generalized to the referring population.

We often wish to answer questions about our processes or products to make improvements and predictions, save money and time, and increase customer satisfaction:

- What is the stability of a new formulation?
- What is the performance of a new product compared with the industry standard or products currently on the market?
- What is causing high levels of variation and waste during processing?

These questions are examples of *inferential problems*.

Inferential problems are usually related to:

Estimation of a population parameter (e.g. a mean)	⟶	What is the stability of a new formulation?
Comparison among groups	⟶	What is the performance of a new product compared with the industry standard or products currently on the market?
Assessing *relationships* among variables	⟶	What is causing high levels of variation and waste during processing?

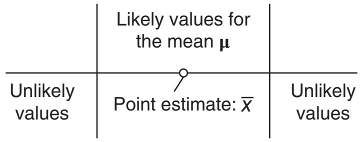
Stat Tool 1.3 (Continued)

We can use several inferential techniques to answer different questions. Later on, we will review the following ones:

Estimation of a population parameter:

- Point estimate
- Confidence intervals

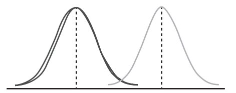
→



Comparison among groups:

- Hypothesis testing (one-sample tests; two-sample tests; ANOVA)

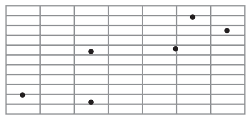
→



Assessing *relationships* among variables:

- Regression models

→



Stat Tool 1.4 Shapes of Data Distributions

Frequency distributions may be shown by tables or graphs. Use *bar charts* for categorical or quantitative discrete variables, *histograms* for continuous variables, and *dot plots* (especially useful for small data sets) for discrete or continuous variables.

By observing the frequency distribution of a categorical or quantitative variable, several *shapes* may be detected:

- When values or classes have similar percentages, the distribution is said to be fairly *uniform*. In a fairly uniform distribution there are no values or classes predominant over the others (a).
- When there is one value or class predominant over the others, the distribution is said to be *nonuniform* and *unimodal* with one peak (b).
- When there is more than one value or class predominant over the others, the distribution is said to be nonuniform and *multimodal* with more than one peak (c).

The value or class with the highest frequency is the *mode* of the distribution (see Figure 1.2).

Stat Tool 1.4 (Continued)

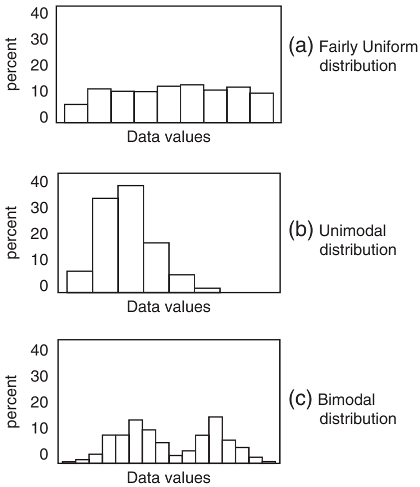


Figure 1.2 Shapes of distributions.

Stat Tool 1.5 Shapes of Data Distributions for Quantitative Variables

By observing the frequency distribution of a *quantitative* discrete or continuous variable, several *shapes* may be detected related also to the presence or absence of *symmetry* (Figures 1.3 and 1.4).

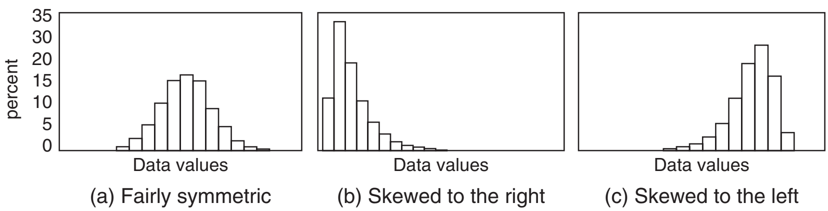


Figure 1.3 Shapes of distributions (symmetric and skewed distributions).

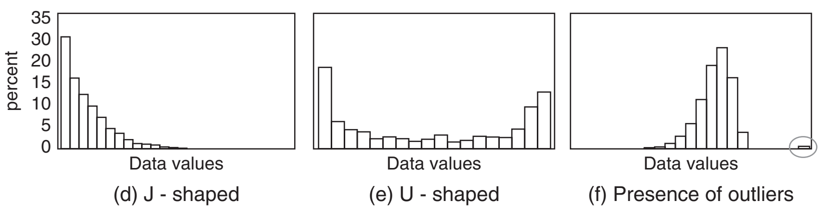


Figure 1.4 Other shapes of distributions.

Stat Tool 1.5 (Continued)

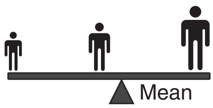
If one side of the histogram (or bar chart for quantitative discrete variables) is close to being a mirror image of the other, then the data are *fairly symmetric* (a). Middle values are more frequent, while low and high values are less frequent. If data are not symmetric, they may be *skewed to the right* (b) or *skewed to the left* (c). In (b) low and middle values are more frequent than high values. In (c) high and middle values are more frequent than low values.

If histograms (or bar charts for quantitative discrete variables) show ever-decreasing or ever-increasing frequencies, the distribution is said to be *J-shaped* (d). If frequencies are decreasing on the left side of the graph and increasing on the right side, the distribution is said to be *U-shaped* (e). Sometimes there are values that do not fall near any others. These extremely high or low values are called *outliers* (f).

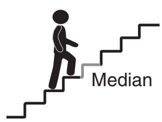
Stat Tool 1.6 Measures of Central Tendency: Mean and Median

When quantitative data distributions tend to concentrate around certain values, we can try to locate these values by calculating the so-called measures of *central tendency*: the *mean* and the *median*. These measures describe the area of the distribution where most values occur.

The *mean* is the sum of all data divided by the number of data. It represents the "balance point" of a set of values.



The *median* is the middle value in a *sorted* list of data. It divides data in half: 50% of data are greater than the median, 50% are less than the median.



For *symmetric* data, mean and median tend to be close in value (Figure 1.5):

In *skewed data* or *data with extreme values*, mean and median can be quite different. Usually for such data, the median tends to be a better indicator of the

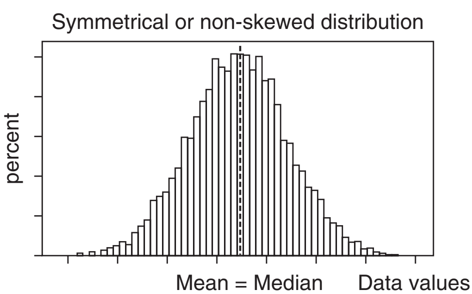


Figure 1.5 Mean and median in symmetric distributions.

Stat Tool 1.6 (Continued)

central tendency rather than the mean, because while the mean tends to be pulled in the direction of the skew, the median remains closer to the majority of the observations (Figure 1.6).

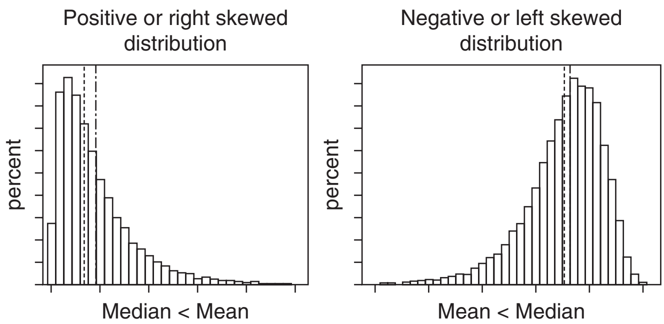
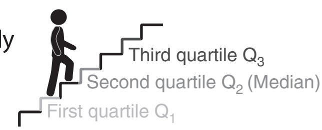


Figure 1.6 Mean and median in skewed distributions.

Stat Tool 1.7 Measures of Non-Central Tendency: Quartiles

Particularly when *numeric* data do not tend to concentrate around a unique central value (e.g. fairly uniform distributions), more than one descriptive measure is needed to summarize the data distribution. These measures are called *quantiles*.



The most common quantiles are *quartiles,* which are three values (first quartile Q_1, second quartile Q_2, and third quartile Q_3) corresponding to specific positions in the *sorted* list of data values (Figure 1.7).

75% of the data are less than Q_3 and 25% are greater than Q_3.
50% of the data are less than Q_2 and 50% are greater than Q_2.
25% of the data are less than Q_1 and 75% are greater than Q_1.

The first quartile is also known as the 25th percentile, the median as the 50th percentile, and the third quartile as the 75th percentile.

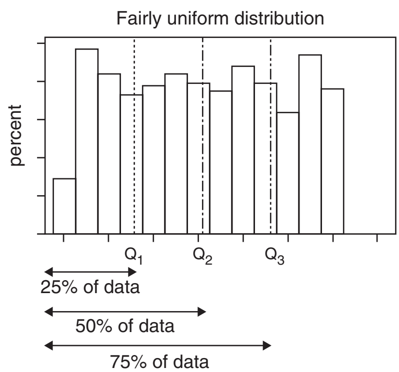
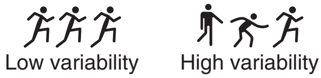


Figure 1.7 Quartiles.

Stat Tool 1.8 Measures of Variability: Range and Interquartile Range

Variability refers to how spread out a set of data values is.



Consider the following graphs (see Figure 1.8):

- The two data distributions are quite different in terms of variability: the graph on the left shows more densely packed values (less variability), while the graph on the right reveals more spread out data (higher variability).
- The terms *variability*, *spread*, *variation*, and *dispersion* are synonyms, and refer to how spread out a distribution is.

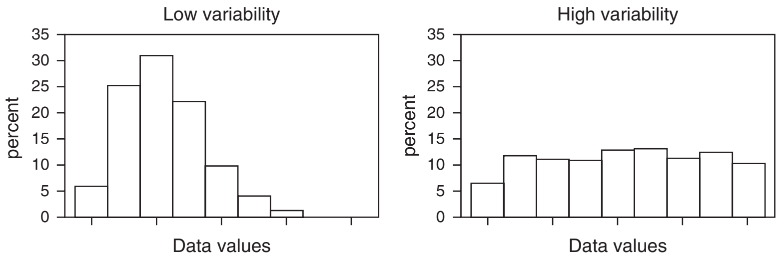


Figure 1.8 Frequency distributions and variability.

How can the spread of a set of *numeric* values be quantified?

The *range*, commonly represented as R, is a simple way to describe the spread of data values. It is the difference between the maximum value and the minimum value in a data set. The range can also be represented as the interval: (minimum value; maximum value).

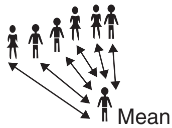
A large range value (or a wide interval) indicates greater dispersion in the data. A small range value (or a narrow interval) indicates that there is less dispersion in the data.

Note that the range only uses two data values. For this reason, it is most useful in representing dispersion when data doesn't include outliers.

A second measure of variation is the *interquartile range*, commonly represented as IQR. It is the difference between the third quartile Q_3 and the first quartile Q_1 in a data set. IQR can also be represented as the interval: (Q_1; Q_3). Fifty percent of the data are within this range: as the spread of these data increases, the IQR becomes larger.

The IQR is not affected by the presence of outliers.

Stat Tool 1.9 Measures of Variability: Variance and Standard Deviation



For *numeric* data, spread can also be measured by the *variance*. It accounts for *all the data* by measuring the *distance or difference between each value and the mean*. These differences are called *deviations*. The variance is the sum of squared deviations, divided by the number of values minus one. Roughly speaking, the variance (usually denoted by S^2) is the average of the squared deviations from the mean.

➢ *Example 1.2*. Suppose you observed the following numeric data with their dotplot (Figure 1.9):

8.1 8.2 7.6 9.0 7.5 6.9 8.1 9.0 8.3 8.1 8.2 7.6

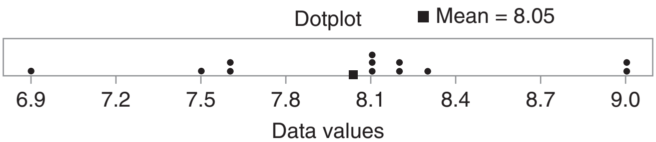


Figure 1.9 Dotplot.

The mean is equal to 8.05. Let's calculate the *deviations from the mean* and their squares:

Deviations	Squared deviations
8.1 − 8.05 = 0.05	$(0.05)^2 = 0.0025$
8.2 − 8.05 = 0.15	$(0.15)^2 = 0.0225$
7.6 − 8.05 = −0.45	$(-0.45)^2 = 0.2025$
9.0 − 8.05 = 0.95	$(0.95)^2 = 0.9025$
7.5 − 8.05 = −0.55	$(-0.55)^2 = 0.3025$
6.9 − 8.05 = −1.15	$(-1.15)^2 = 1.3225$
8.1 − 8.05 = 0.05	$(0.05)^2 = 0.0025$
9.0 − 8.05 = 0.95	$(0.95)^2 = 0.9025$
8.3 − 8.05 = 0.25	$(0.25)^2 = 0.0625$
8.1 − 8.05 = 0.05	$(0.05)^2 = 0.0025$
8.2 − 8.05 = 0.15	$(0.15)^2 = 0.0225$
7.6 − 8.05 = −0.45	$(-0.45)^2 = 0.2025$

Stat Tool 1.9 (Continued)

Now, by adding up the squared deviations to get the sum and dividing it by the number of values minus one, you obtain the variance:

$$S^2 = \frac{0.0025 + 0.0225 + \ldots + 0.2025}{11} = 0.359$$

The variance measures how spread out the data are around their mean. The greater the variance, the greater the spread in the data.

The variance is not in the same units as the data, but in squared units. If the data are in grams, the variance is expressed in squared grams, and so on. Thus, for descriptive purposes, its square root, called *standard deviation*, is used instead.

The *standard deviation* (usually denoted by S) quantifies variability *in the same units of measurement* as we measure our data.

Considering the previous example, the standard deviation is:

$$S = \sqrt{0.359} = 0.599$$

The greater the standard deviation, the greater the spread of data values around the mean.

Considering the mean and the standard deviation together and computing the range: mean ± S, we can say that data values vary on average from (mean − S) to (mean + S).

From the previous example the average range is:

$$\left(8.05 - 0.599;\ 8.05 + 0.599\right) = \left(7.451;\ 8.649\right)$$

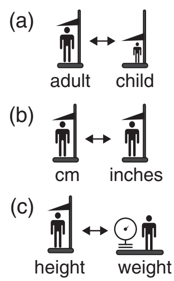
The observed data vary on average from 7.5 to 8.6.

Stat Tool 1.10 Measures of Variability: Coefficient of Variation

Another measure of variability for *numeric data* is the *coefficient of variation*.

It is calculated as follows:

$$\text{CV\%} = \left(\text{standard deviation} / \text{mean}\right) \times 100.$$



Being a dimensionless quantity, the coefficient of variation is a useful statistic for *comparing the spread* among several datasets, even if the means are different from one another (a), or data have different units (b), or data refer to different variables (c).

The higher the coefficient of variation, the higher the variability.

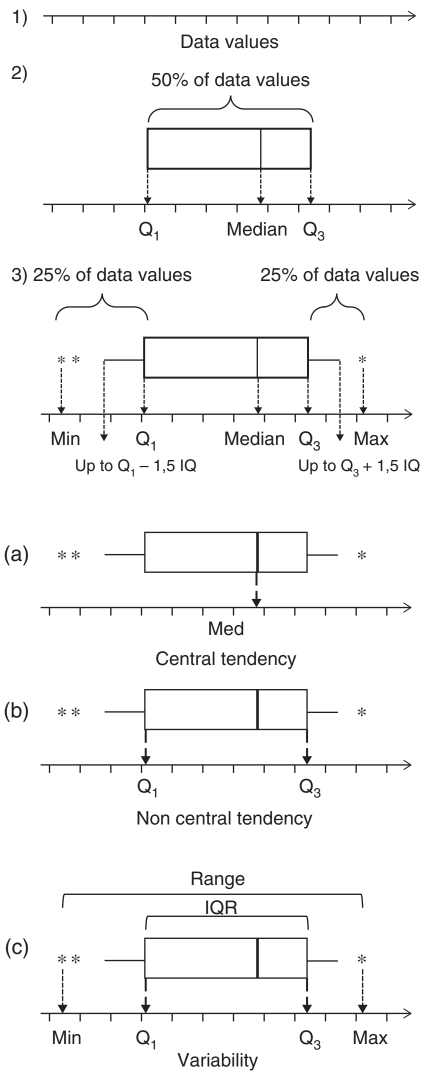
Stat Tool 1.11 Boxplots

So far, we have looked at three different aspects of numerical data analysis: *shape of the data*, *central and non-central tendency*, and *variability*.

Boxplots can be used to assess and compare these three aspects of *quantitative* data distributions, and to look for outliers.

Like histograms, boxplots work best with moderate to large sample sizes (at least 20 values).

Let's look at how a boxplot is constructed. It can be displayed horizontally or vertically:



1) Start by drawing a horizontal or vertical axis in the units of the data values.
2) Draw a box to encompass 50% of middle data values. The left edge of the box is the first quartile Q_1. The right edge of the box is the third quartile Q_3. The width of the box is the interquartile range, IQR. Draw a line inside the box to denote the median.
3) Draw lines, called whiskers, on the left (to the minimum) and on the right (to the maximum) of the box to show the spread of the remaining data (25% of data points are below Q_1 and 25% are above Q_3). Several statistical softwares do not allow the whiskers to extend beyond one and a half times the interquartile range ($1.5 \times$ IQR). Any points outside of this range are outliers and are displayed individually by asterisks.

Boxplots help to summarize:

a) *Central tendency*. Look at the value of the median.
b) *Non-central tendency*. Look at the values of the first quartile Q_1 and the third quartile Q_3.

Stat Tool 1.11 (Continued)

c) *Variability*. Look at the length of the boxplot (range) and the width of the box (IQR).
d) *Shape of data*. Look at the position of the line of the median in the box and the position of the box between the two whiskers. In a symmetric distribution, the median is in the middle of the box and the two whiskers have the same length. In a skewed distribution, the median is closer to Q_1 (skewed to the right) or to Q_3 (skewed to the left) and the two whiskers do not have the same length (Figure 1.10).

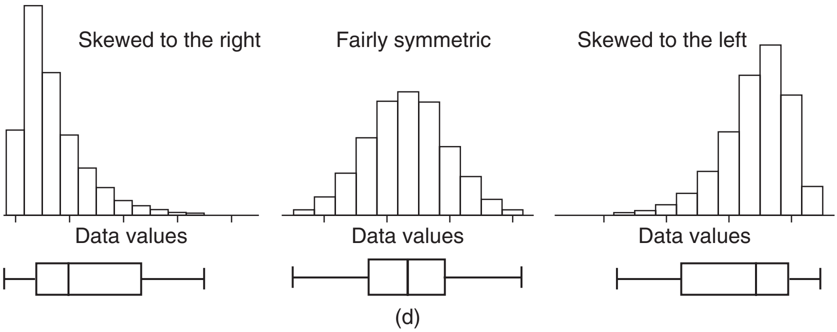


Figure 1.10 Histograms and boxplots.

Stat Tool 1.12 Basic Concepts of Statistical Inference

After describing important characteristics of sample data through descriptive statistics, the second step of a statistical analysis is usually *inferential analysis,* where sample findings are generalized to the referring population.

Inferential techniques use descriptive statistics such as:

sample mean "$\bar{x}$*"   sample proportion "p"   sample standard deviation "S"*

to draw conclusions about the corresponding unknown quantities of the population, called *parameters*:

population mean: "μ"   population proportion "π"   population standard deviation "σ"

Note that it is standard to use Greek letters for certain *parameters*, such as μ to stand for a population mean, σ for a population standard deviation, σ^2 for a population variance, and π for a proportion of statistical units having a characteristic of interest.

Stat Tool 1.12 (Continued)

A *statistic* (mean, proportion, variance) describes a characteristic of the sample (central tendency, variability, shape of data) and is *known*.

A *parameter* (mean, proportion, variance) describes a characteristic of the population (central tendency, variability, shape of data) and is *unknown*.

Statistical inference uses sample data to draw conclusions about a population with a known level of risk. In general, statistical inference proceeds as follows:

1) We are interested in a *population*.
2) We identify *parameters* of that population that will help us understand it better.
3) We take a *random sample* and compute *sample statistics*.
4) Through inferential techniques, we use the sample statistics to *infer* facts about the population parameters of interest.

Stat Tool 1.13 Inferential Problems

As mentioned in Stat Tool 1.3, we often want to answer questions about our processes or products to make improvements and predictions, save money and time, and increase customer satisfaction:

- What is the stability of a new formulation?
- Which attributes of a product do consumers find most appealing?
- What is the performance of a new product compared with products currently on the market?
- What is causing high levels of variation and waste during processing?
- Can a process change reduce production time to get the product in stores more quickly?

Stat Tool 1.13 (Continued)

These questions are examples of *inferential problems*.
How can we use *inferential techniques* to answer these questions?
Inferential problems are usually related to:

Estimation of a population parameter:
- What is the stability of a new formulation?
- Which attributes of a product do consumers find most appealing?

Comparison of a population parameter to a specified value or among groups:
- What is the performance of a new product compared with the industry standard or products currently on the market?

Assessing *relationships* among variables:
- What is causing high levels of variation and waste during processing?
- Can a process change reduce production time to get the product in stores more quickly?

We may use several inferential techniques to answer different questions:

Estimation of a population parameter:
- Point estimate and confidence intervals

Comparison among groups:
- Hypothesis testing (one-sample tests; two-sample tests; analysis of variance, ANOVA)

Assessing *relationships* among variables:
- Regression models

Stat Tool 1.14 Estimation of Population Parameters and Confidence Intervals

Let's introduce the problem of the estimation of a population parameter.

Because it is often impractical or impossible to gather data on the entire population, we must *estimate* the *population parameters* using *sample statistics*.

Statistics, such as the sample mean and standard deviation, are called *point estimators*.

A *point estimate* is a *single sample value* that approximates the true unknown value of a population parameter.

Stat Tool 1.14 (Continued)

- Point estimators:

$$\textit{sample mean } \bar{x} \quad \textit{sample proportion } p \quad \textit{sample standard deviation } S$$

- Population parameters:

$$\textit{population mean } \mu \quad \textit{population proportion } \pi \quad \textit{population standard deviation } \sigma$$

Point estimates, such as the sample mean or standard deviation, provide a lot of information, but they don't give us the full picture.

As it is highly unlikely that, for example, the sample mean and standard deviation we obtain are exactly the same as the population parameters, and to get a better sense of the true population values, we can use *confidence intervals*.

A *confidence interval* is a *range of likely values for a population parameter*, such as the population mean or standard deviation.

Usually, a confidence interval is a range:

$$\left(\textit{point estimate} - \text{something};\ \textit{point estimate} + \text{something}\right)$$

Using confidence intervals, we can say that it is likely that the population parameter is somewhere within this range.

➢ *Example 1.3*. To illustrate this point, suppose that a research team wants to know the *mean* satisfaction score (from 0: completely not satisfied, to 10: completely satisfied) for the population of people who use a new formulation of a product.

Mean satisfaction score (population parameter) = ?

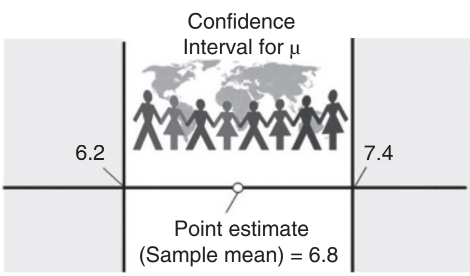
From a random sample of consumers, the *sample mean* is 6.8, and the *confidence interval* is CI = (6.2; 7.4).

Sample mean = 6.8
C.I. = (6.2; 7.4)

So the true unknown population mean satisfaction score is likely to be somewhere between 6.2 and 7.4.

The central point of the confidence interval is the sample mean: $\bar{x} = 6.8$ (point estimate of μ).

There's always a chance that the confidence interval won't contain the true population mean.



Stat Tool 1.14 (Continued)

When we use confidence intervals, we must decide how sure we need to be that the confidence interval contains the actual population parameter value, taking into account that we cannot be 100% sure.

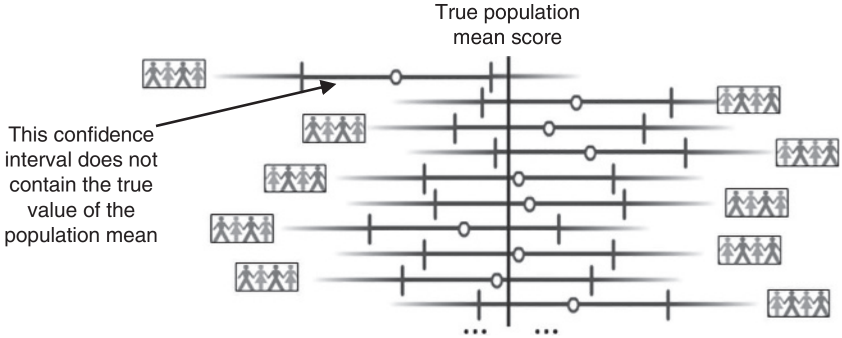
We quantify how sure we need to be with a value called the *confidence level*, usually denoted by $(1-\alpha)$.

The confidence level is set by the researcher before calculation of a confidence interval.

The most common confidence level is 95% (0.95). Other common levels are 90% and 99%.

The confidence level is how sure we are that the confidence interval contains the actual population parameter value.

➢ *Example 1.4*. To illustrate the meaning of the confidence level, let's return to the previous example and suppose we drew 100 samples from the same population and calculated the confidence interval for each sample.
If we used 95% confidence intervals, on average 95 out of 100 of the confidence intervals will contain the population parameter, while 5 out of 100 will not.
In practice when we calculate a 95% confidence interval for our sample, we are confident that our sample is one of the 95% samples for which CI covers the true parameter value.



Stat Tool 1.15 Hypothesis Testing

A common task in statistical studies is the *comparison* of *mean values*, *variances*, *proportions*, and so on, to a hypothesized value of interest or among different groups, for example:

➢ What is the performance of a new product compared with the industry standard or products currently on the market?

To investigate any such questions, we can conduct an inferential procedure called a *hypothesis test*. *Hypothesis tests* allow us to make decisions on business problems based on statistically significant results, not based on intuition alone.

To begin with, the researchers need to determine which question they want to focus on and then define the *hypotheses*. A *statistical hypothesis* is a *claim about a population parameter* (e.g. about the mean or the standard deviation of a variable of interest).

Hypotheses should be based on our knowledge of the process, such as how a process has performed in the past or customers' expectations.

To perform a hypothesis test, we need to define the *null hypothesis* and the *alternative hypothesis*.

- *Null hypothesis* H_0: usually states that a population parameter, such as the population mean, *equals* a specified value or parameters from other populations.
  E.g. H_0: the mean performance of the new product is equal to the industry standard.
- *Alternative hypothesis* H_1: is the opposite of the null hypothesis, so it usually states that the population parameter *does not equal* a specified value or parameters from other populations.

  H_1: the mean performance of the new product is NOT equal to the industry standard.

Sometimes the alternative hypothesis is *directional* or *one-sided*; that is, we suspect the population parameter is greater than or less than a given value.

- *Null hypothesis* H_0:
  E.g. H_0: the mean performance of the new product is equal to the industry standard.
- *One-sided alternative hypothesis* H_1:

  H_1: the mean performance of the new product is GREATER THAN the industry standard.
  Or
  H_1: the mean performance of the new product is LESS THAN the industry standard.

Stat Tool 1.15 (Continued)

After defining null and alternative hypotheses, the hypothesis test uses *sample data* to decide on *two possible conclusions*:

1) We can *reject the null hypothesis* in favor of the alternative hypothesis. If we reject the null hypothesis, we say that our result is *statistically significant*:
   E.g. H_0 (the mean performance of the new product is equal to the industry standard) is REJECTED.
2) We can *fail to reject the null hypothesis* and conclude that we do not have enough evidence to claim that the alternative hypothesis is true. We will say that our results are *NOT statistically significant*:
   E.g. H_0 (the mean performance of the new product is equal to the industry standard) is NOT REJECTED.

Because we are using sample data, decisions based on those hypothesis tests could be wrong.

Recall the *confidence level* $(1-\alpha)$ that we discussed earlier (*Stat Tool 1.14*). The confidence level is how sure we are that the confidence interval contains the true population parameter value. This confidence level is set by the researcher and is usually equal to 95% (0.95) or 99% (0.99).

Let's consider the outcomes of a *hypothesis test*. If the null hypothesis is true and based on our sample data we fail to reject it, we make the *correct decision*, but if we reject it, we make an *error*. In hypothesis testing, the *probability of rejecting a null hypothesis that instead is true* is called *significance level* and is denoted by α. We always select it before performing the hypothesis test and it is usually equal to 5% (0.05) or 1% (0.01).

Confidence level and significance level are tools to quantify the *uncertainty* about our inferential conclusions.

Stat Tool 1.16 The p-Value

After establishing the null and alternative hypotheses and setting the significance level α, how do we decide to reject the null hypothesis?

When we conduct a hypothesis test, the results include a probability called *p-value*.

We use the *p-value* to determine whether we should *reject* or *fail to reject the null hypothesis*, by comparing it to the significance level α.

If the *p-value is less than* α, we *reject the null hypothesis* in favor of the alternative hypothesis (our result is *statistically significant*):

E.g. H_0 (the mean performance of the new product is equal to the industry standard) is REJECTED.

Stat Tool 1.16 (Continued)

If the *p-value is greater than or equal to* α, we *fail to reject the null hypothesis*. There is not enough evidence to claim that the alternative hypothesis is true (our results are *NOT statistically significant*):

E.g. H_0 (the mean performance of the new product is equal to the industry standard) is NOT REJECTED.

➢ *Example 1.5*. A researcher wants to investigate whether the performance of a new product differs with respect to the industry standard.

Suppose they set the significance level α equal to 0.05 (5%). The two hypotheses are:

H_0: The mean performance of the new product is equal to the industry standard.

H_1: The mean performance of the new product is NOT equal to the industry standard.

What if the p-value is 0.032? Would they reject or fail to reject the null hypothesis?

They should *reject the null hypothesis* H_0 because the p-value is less than the significance level α and conclude that the new product has a different mean performance with respect to the industry standard. The result of the test is *statistically significant* at $\alpha = 5\%$.

And, what if the p-value is 0.076? Would they reject or fail to reject the null hypothesis?

They should *fail to reject the null hypothesis* H_0 because the p-value is greater than the significance level α and conclude that there is not enough evidence to claim that the new product has a different mean performance with respect to the industry standard. The result of the test is *NOT statistically significant* at $\alpha = 5\%$.

The p-value indicates whether our results are statistically significant. However, just because our results are *statistically significant* doesn't mean that they are *practically significant*.

➢ *Example 1.6*. A production line manager attempts to reduce production time by modifying the process. They compare the mean production time of the old process with the mean production time of the new process using a hypothesis test.

The p-value is 0.032. Using an alpha of 0.05, would they reject or fail to reject the null hypothesis?

Stat Tool 1.16 (Continued)

Because the p-value is less than alpha, they reject the null hypothesis that the production times are equal. The difference between the mean production times is statistically significant.

Now, consider practical significance. If the difference between the two production times is five seconds, is that really practically significant? Is it worth the cost of implementing the process change?

Always consider the practical significance of your results and your knowledge of the process before reaching a final conclusion.