Statistics

A statistical problem:

We wish to study the performance of the lithium batteries used in a particular model of pocket calculator. The purpose of the study is to determine the mean effective life span of these batteries so that we can place a limited warranty on them in the future. Since this type of battery has not been used in this model before, no one can tell us the distribution of the random variable, X, the life span of a battery. We have to discover this distribution. This is a statistical problem which has the following characteristics:

- Associated with the problem is a large group of objects (called ***population*****)** about which inferences are to be made.
- There is at least one random variable whose behavior is to be studied relative to the population
- The population is too large to study in its entirety, or techniques used in the study are destructive in nature. In either case, we must draw conclusions about the population based on observing only a portion or "***sample***" of objects drawn from the population.

How can it be guaranteed that conclusion can be drawn on population based on sampled data?

Sampling

Random sampling:
We want to select a subset of n samples from the population “at random”. The following conditions must be satisfied:

- the selection of one object is independent of the selection of any other.
- Each object in the population is equally likely to make up the sample. Just as in a lottery!

Example: A physical education professor want to study the physical fitness level of students at her university. There are 20,000 students enrolled at the university, and she want to draw a sample of size 100 to take a physical fitness test. She obtains a list of all 20,000 students, numbered from 1 to 20,000. She uses a computer random number generator to generate 100 random integers between 1 and 20,000 and then invites the 100 students corresponding to those numbers to participate in the study. Is this a random sample?

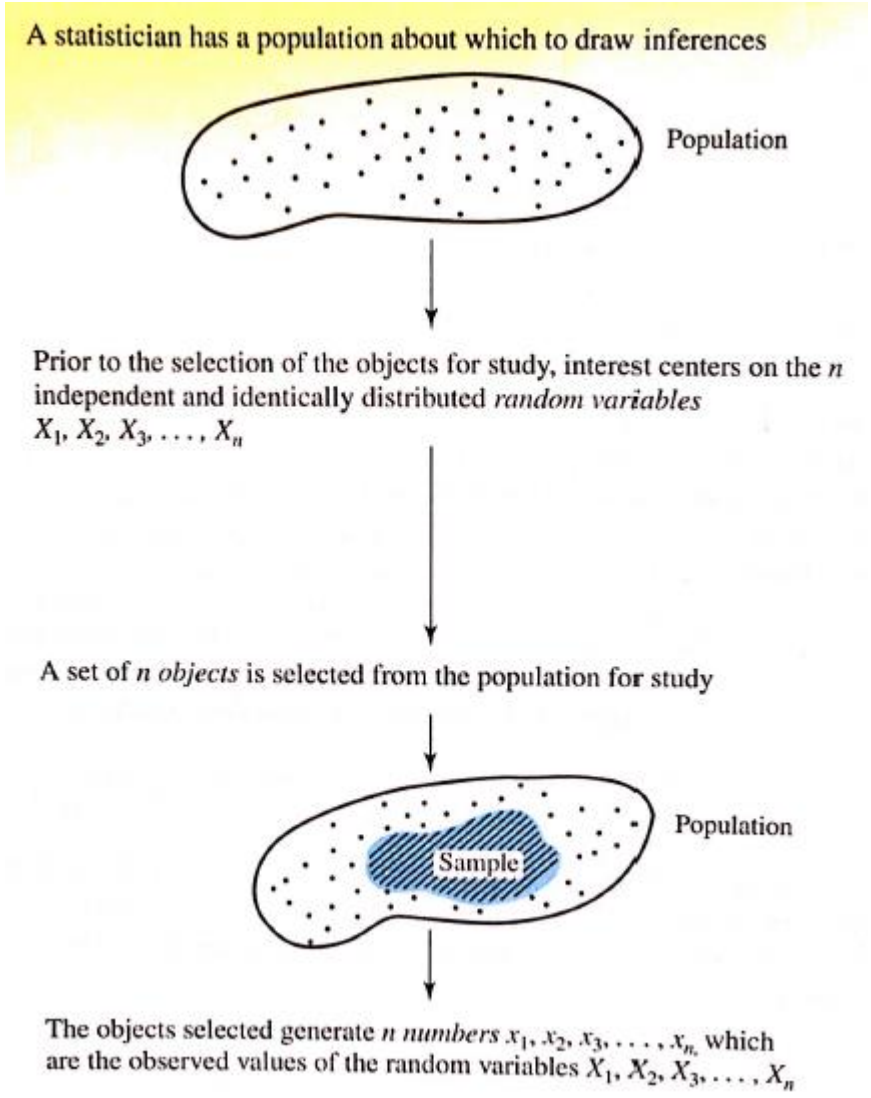
Example: A quality engineer wants to inspect rolls of wallpaper in order to obtain information on the rate at which flaws in the printing are occurring. He decides to draw a sample of 50 rolls of wallpaper from a day’s production. Each hour for hours, he takes the 10 most recently produced rolls and counts the number of flaws on each. Is this a random sample?

Sampling

The meaning of A random sample of size n:

Prior to the actual selection of the batteries to be studied, X_i $(i = 1,2, \dots, n)$, the life span of the i^{th} battery selected is a random variable. It has the same distribution as X, the life span of batteries in the population. These random variables are ***independent*** in the sense that the value assumed by one has no effect on the value assumed by any of the others. **The random variables $X_1, X_2, \dots, X_n$ can be thought of as a "random sample".**

Once we have actually selected n batteries and have observed the life span of each selected batteries, we shall have available n numbers, $x_1, x_2, \dots, x_n$. These numbers are the observed values of the random variables $\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}$ **and can be thought of as a "random sample".**



Sampling

Definition: Random sample:
A random sample of size *n* from the distribution of X is a collection of *n* **independent** random variables, each with the same distribution as X.

Let $X_i, i = 1,2, \dots n$ be a random sample of size n from a distribution with mean μ (*population mean*) and variance σ^2 (*population variance*). We say that $X_i, i = 1,2, \dots n$ are ***independent and identically distributed*** (iid) as the population and we have $E[X_i] = \mu;\ \ VarX_i = \sigma^2$.

Example:
Let $X_i, i = 1,2, \dots 10$ be a random sample of size 10 from a normal random variable with $\mu = 2.0\ and\ \sigma^2 = 15$.
What is the distribution of X_3?
Find $E[X_{10}]\ and\ VarX_{10}$

Sampling

Sampling with replacement

- When objects are selected from a finite population, random sample results only when sampling is done with replacement to ensure that $\boldsymbol{X_1, X_2, \ldots, X_n}$ are indeed independent and identically distributed.

- Usually, sampling from a finite population is done without replacement. It is OK if the sample is small relative to the population itself. In practice, if the sample constitute at most 5% of the population, independence may be assumed.

Sample statistics

Sample mean:

Let $X_1, X_2, X_3, \dots, X_n$ be a random sample from the distribution of X. The statistic $\frac{1}{n}\sum_{i=1}^{n} X_i$ is called the sample mean and is denoted by $\bar{X}$.

Notice that $\bar{X}$ itself is a random variable!!!

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, where\ x_i\ is\ an\ observed\ value\ of\ X_i$$

is an observed value of $\bar{X}$

Q: What is the difference between $\mu_X\ (population\ mean)\ and\ \bar{X}\ (sample\ mean)$?

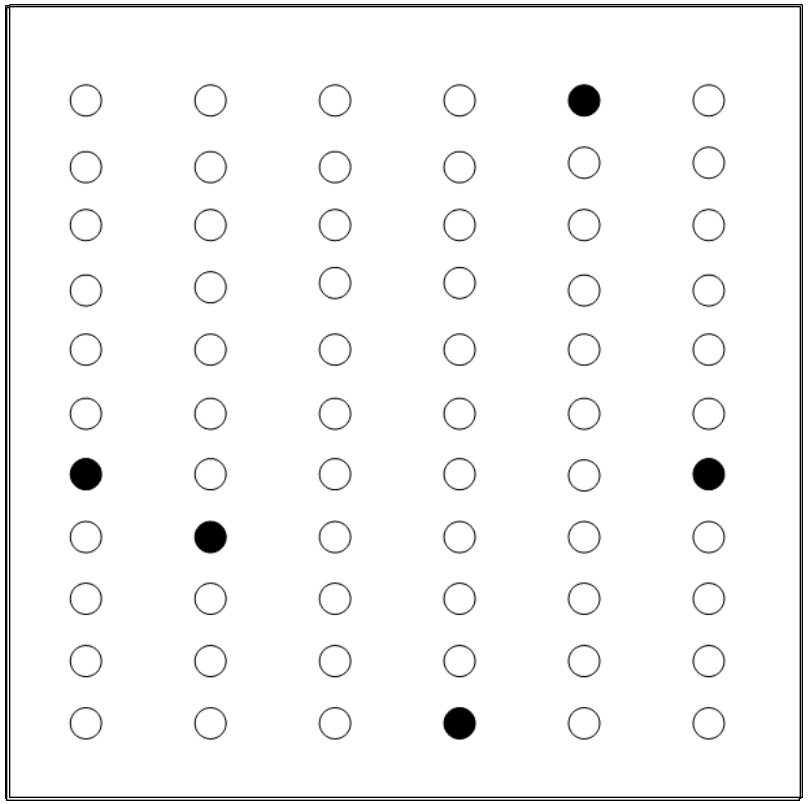
Example: A random sample of size 9 yields the following observations on the random variable X, the coal consumption in millions of tons by electric utilities for a given year:
406, 395, 400, 450, 390, 410, 415, 401, 408
Then, the observed value of sample mean for these data is:

$\bar{x} = \sum_{i=1}^{9} \frac{x_i}{9} = (406 + 395 + 400 + \cdots + 408)/9 \approx 408.3$ million tons

Meanings of random sampling

An example to show $\bar{X}$ as a random variable:

Let X denote the manufacture error (in mm) of a batch of engine parts. Let $X_1, X_2, \ldots, X_5$ be a random sample of size 5 from the population. During a random sampling process, five parts are randomly selected and manufacturing error of each selected part is measured. This process was repeated 5 times and obtain the data in the following table. Calculate the values of $\bar{X}$.



	X1	X2	X3	X4	X5	$\bar{X}$
1	-0.1	0	0.2	0.2	0.1	0.08
2	0.2	0.1	0.1	0	0.1	0.1
3	0.1	0.1	-0.1	-0.1	0.2	0.04
4	-0.1	-0.2	0.1	0	-0.1	-0.06
5	0	0.2	0.1	0.1	-0.1	0.06

Meanings of random sampling

Let $X_1, X_2, \cdots . X_n$ be a random sample of size n from a random variable X with mean μ and variance σ^2.

Then,

- $X_1, X_2, \cdots . X_n$ are random variables with the same distribution of X. i.e., $E[X_i] = \mu, VarX_i = \sigma^2, i = 1,2, \ldots, n$
- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a random variable
- When $X_1, X_2, \cdots . X_n$ assumes the values $x_1, x_2, \cdots . x_n$, then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an observed value of $\bar{X}$
- $x_1, x_2, \cdots . x_n$ is also called a random sample of size n from X

Sample statistics

Sample variance and sample standard deviation:

Let $X_1, X_2, X_3, \dots, X_n$ be a random sample from the distribution of X. Then the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is called the sample variance. Furthermore, the statistic $S = \sqrt{S^2}$ is called the sample standard deviation.

A computational formula for S^2:

$$S^2 = \frac{n\sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} X_i)^2}{n(n-1)}$$

Example: A random sample of size 9 yields the following observations on the random variable X, the coal consumption in millions of tons by electric utilities for a given year:

390, 400, 406, 410, 450, 395, 401, 415, 408

To compute the sample variance, we must evaluate the statistics $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} {X_i}^2$ for this sample. The observed values are: $\sum_{i=1}^{9} x_i = 3675, \quad \sum_{i=1}^{9} {x_i}^2 = 1503051$

Then, the observed value of S^2 is:

$$S^2 = \frac{9 \times (1503051) - (3675)^2}{9 \times 8} \approx 303.25$$

The observed value of S is: $S = \sqrt{303.25} = 17.4$ million tons

Sample statistics

Outliers:
Sometimes a sample may contain a few points that are much larger or smaller than the rest. Such points are called *outliers*.

Sometimes outliers results from data entry errors;

Outliers should always be scrutinized and any outlier that is found to result from an error should be corrected or deleted.

Sample statistics

Sample Median:
Like the mean, the median is a measure of center.
To compute the median of a sample, order the values from smallest to largest. The sample median is the middle number. If the sample size is an even number, it is customary to take the sample median to be the average of the two middle numbers.

Example: Find the mean and median of the following sample: 1,2,3,4,20

The mean is 6 and the median is 3.
The number 20 is likely an outlier and in this case, median is a better representative of the sample than mean.

Sample statistics

Quartiles and Percentiles:

The pth 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 pth percentile, and $(100-p)\%$ are greater.

How to find the pth percentile?

Order the sample values from smallest to the largest, and then compute the quantity $(\frac{p}{100})(n+1)$, where n is the sample size. If this quantity is an integer, the sample value in this position is the pth percentile. Otherwise, average the two sample values on either side.

Note the first quartile is the 25th percentile, the median is the 50th percentile.

Example: Find the 24th percentile and the median of the following sample:

30	75	79	80	80	105	126	138	149	179	179	191
223	232	232	236	240	242	245	247	254	274	384	470

Solution: $n = 24$.

To find 24th percentile, we calculate $\left(\frac{24}{100}\right) \times (24+1) = 6$. Then the 24th percentile is the 6th value from the sample: 105

To find the median, we calculate $\left(\frac{50}{100}\right) \times (24+1) = 12.5$, We then take the average of the 12th and the 13th values from the sample: (191+223)/2=207. The median of this sample is 207

Graphical Description

Histograms:

A histogram is a graphic that gives an idea of the "shape" of the distribution of a sample, indicating regions where sample points are concentrated and regions where they are sparse.

Example: Let's construct a histogram from the PM emission of 62 vehicles driven at high altitude presented in the following table. The sample values range from 1.11 to 23.38

7.59	3.01	4.04	6.95	3.57	2.89	10.14	2.01	5.64
6.28	13.63	17.11	18.64	4.35	4.68	9.2	5.91	2.07
6.07	13.02	12.26	7.1	3.84	1.85	7.31	5.6	1.11
5.23	23.38	19.91	6.04	2.37	9.14	2.09	5.61	3.32
5.54	9.24	8.5	5.66	3.81	8.67	6.32	1.5	1.83
3.46	3.22	7.81	8.86	5.32	9.52	6.53	6.46	7.56
2.44	2.06	7.18	4.4	5.84	2.68	6.32	5.29	

Histograms

First, let's construct a frequency table based on the original data (as shown)
In this table, the class intervals divide the sample into groups. For most histograms, the class intervals all have the same width (in the table, all classes have width 2).

How to decide the number of classes for a sample?

In general, it is good to have more classes rather than fewer, but it is also good to have large number of sample points in the classes. A commonly used practice is, when the sample size n is large (several hundred or more), the number of classes may be $\log_2 n \ \ or \ 2n^{1/3}$. When the sample size is smaller, more classes than these are often needed.

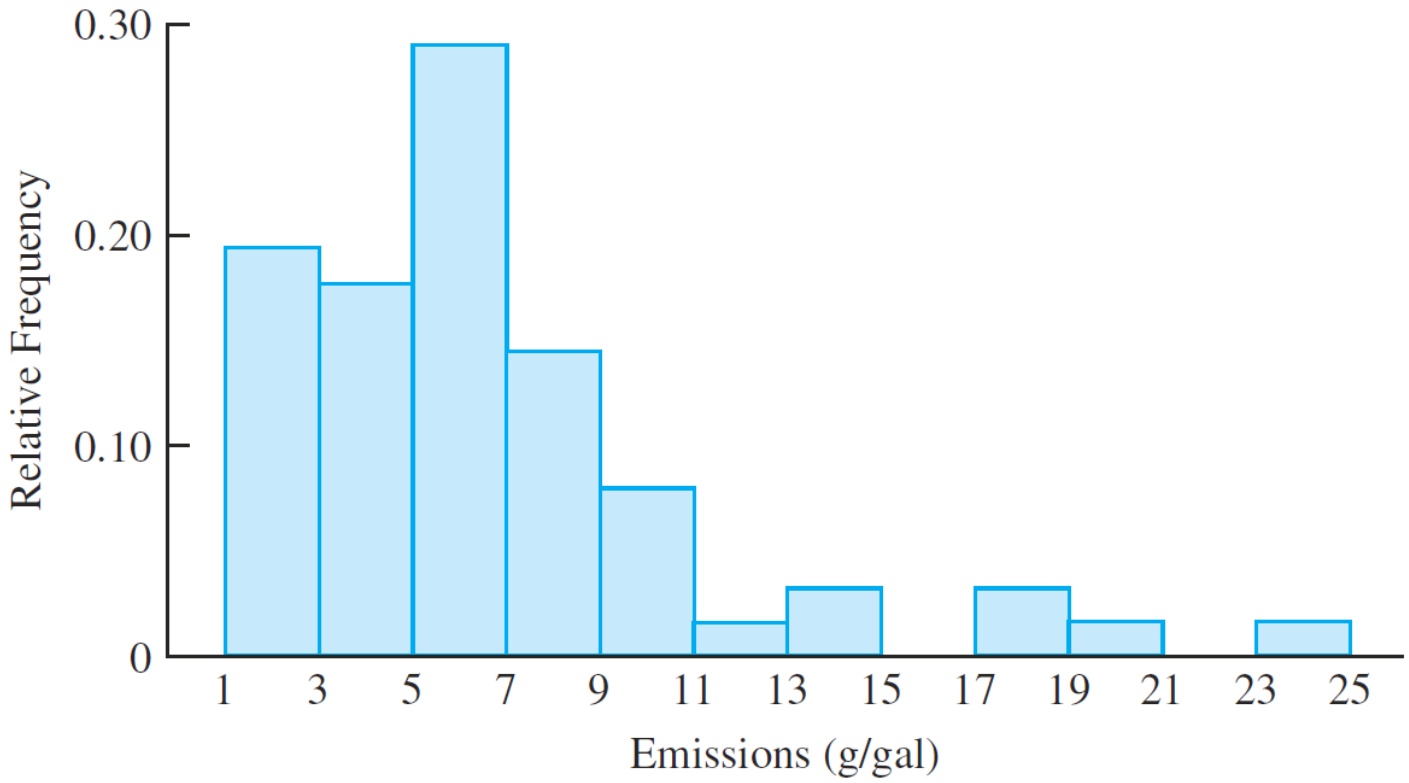
The column labeled "Frequency" in Table 1.4 presents the number of sample points that fall into each of the class intervals.
The column labeled "relative frequency" presents the frequencies divided by the total sample points (62 in the example), i.e., proportion of the sample points fall into the class interval.
The column labeled "Density" presents the relative frequency divided by the class width (2 in the example)

TABLE 1.4 Frequency table for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 13	1	0.0161	0.0081
13–< 15	2	0.0323	0.0161
15–< 17	0	0.0000	0.0000
17–< 19	2	0.0323	0.0161
19–< 21	1	0.0161	0.0081
21–< 23	0	0.0000	0.0000
23–< 25	1	0.0161	0.0081

Histograms

Here is the histogram for Table 1.4.



Boxplots

A boxplot is a graphic that presents the median, the first and third quartiles, and any outliers that are presented in a sample.

The interquartile range (IQR) is the difference between the third quartile and the first quartile. i.e., the IQR is therefore the distance needed to span the middle half of the data.

Outliers: any points that is more than 1.5IQR above the third quartile, or more than 1.5IQR below the first quartile, is considered an ***outlier***. A point that is more than 3 IQR from the first or third quartile is considered an ***extreme outlier***

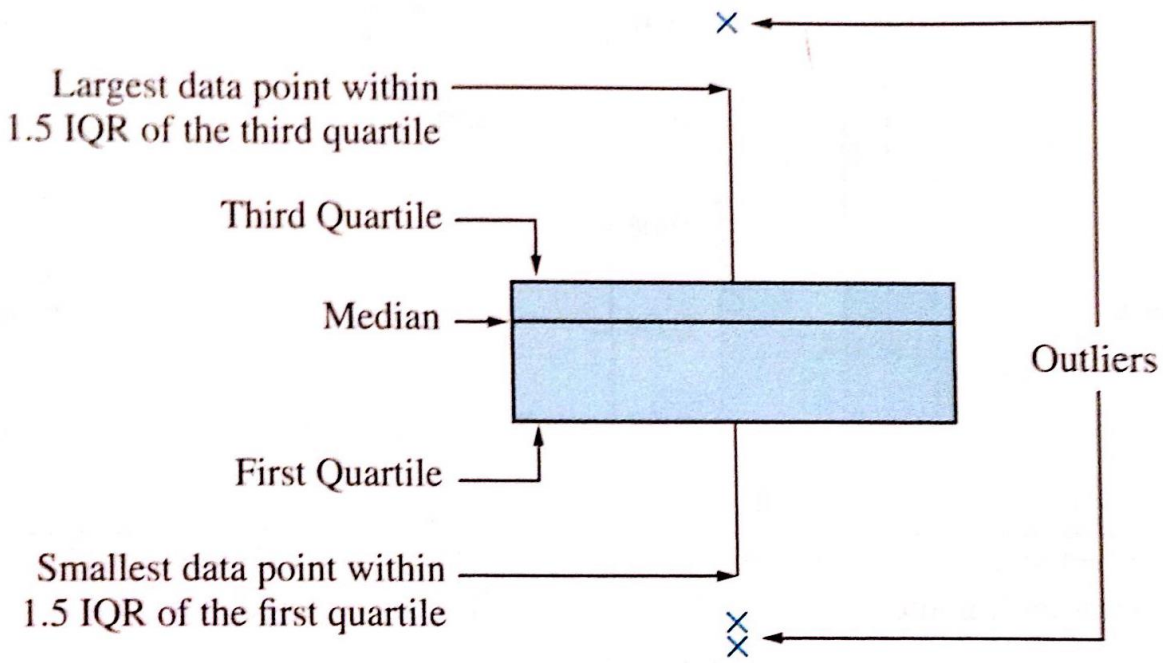


FIGURE 1.13 Anatomy of a boxplot.

Boxplot Example