Chapter 3. Visual and Numerical Summaries

When confronted with a new sample of data the first task is almost always to explore it, i.e. to visualize and summarize the dataset in different ways.

Calculating numerical summaries serves two purposes:

- The first is exploratory. Calculating statistics which characterise general properties of the sample (such as location, dispersion, or symmetry) helps us to understand the data we have gathered. This aim can be greatly aided by the use of graphical displays representing the data.
- The second, as we shall see later in Chapters 8 and 9, is that these summaries will commonly provide the means for relating the sample we have to the wider population in which we are truly interested.

This chapter focuses on some of the most essential visual and numerical summaries of datasets.

3.1 Visualization

3.1.1 The histogram

The **histogram** is one of the most fundamental tools in all of experimental science. It allows us visualize how a sample of data is distributed. Begin with a 1-dimensional data set, a collection of observed values $(x_1, \ldots, x_n)$. The first step is deciding on a set of **bins** that divide the range of x into a series of intervals. I.e. the first bin might be the interval $-3.0 \leq x < -2.5$, the second bin $-2.5 \leq x < -2.0$, and so on (see Fig. 3.1).

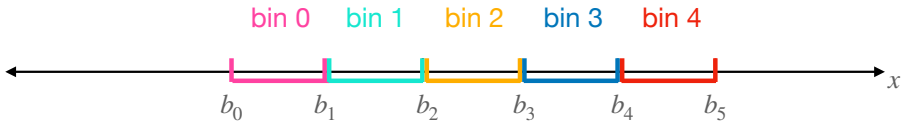


Figure 3.1: Example of binning. The range $b_0 \leq x < b_5$ is divided into 5 bins and bin i covers the range $b_i \leq x < b_{i+1}$.

A histogram then shows the **frequency** for each bin, i.e. how many of the data points lie in each bin. For example, in Fig. 3.2 the height of the first bar is 1 since 1 of the data points lies in the first bin (the data set is given in the figure caption, and shown as black dots along the bottom). The bin running from $x = -3$ to $x = 0$ has a frequency of 4, since 4 of the elements in the data set are between -3 and 0. The heights of all the bars add up to the size of the data set.

The word "bin" is both a noun and a verb. You bin the data into bins to make a histogram. You can also ask, "What binning did you use in that plot?"

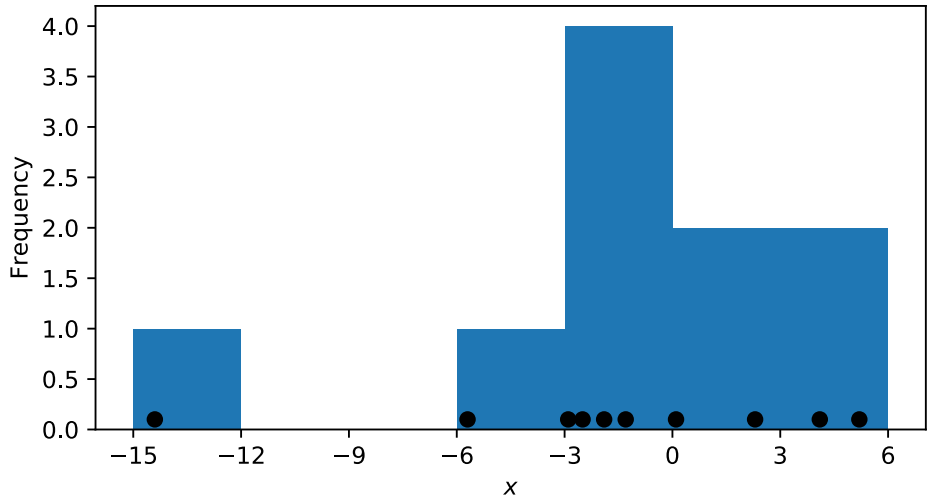


Figure 3.2: Histogram of the data set (-1.3, 5.2, -2.5, 4.1, -1.9, -2.9, 2.3, 0.1, -5.7, -14.4). The seven bins each have a width of $\Delta x = 3$, with the first bin starting at $x = -15$ and last bin ending at $x = 6$. The points in the data set are shown as black circles along the bottom.

Often the histogram's y-axis is normalized in some way. Instead of showing frequency, the height of the histogram can show **relative frequency**: the *fraction* of the data set contained within the bin. In other words, frequency divided by the size of the data set. In this case, the heights of all the bars of the histogram add up to 1 (assuming every data point ends up in some bin). Or the histogram could show the **density**, which is the relative frequency divided by the bin width. This is a way of removing the influence of the bin width (which is fundamentally arbitrary). The normalization condition in this case reminds us of an integral: $1 = \sum_{\text{bins } i} \rho_i \Delta x_i$, where ρ_i is the density for bin i and Δx_i is the width of bin i. This is not a coincidence.

Example

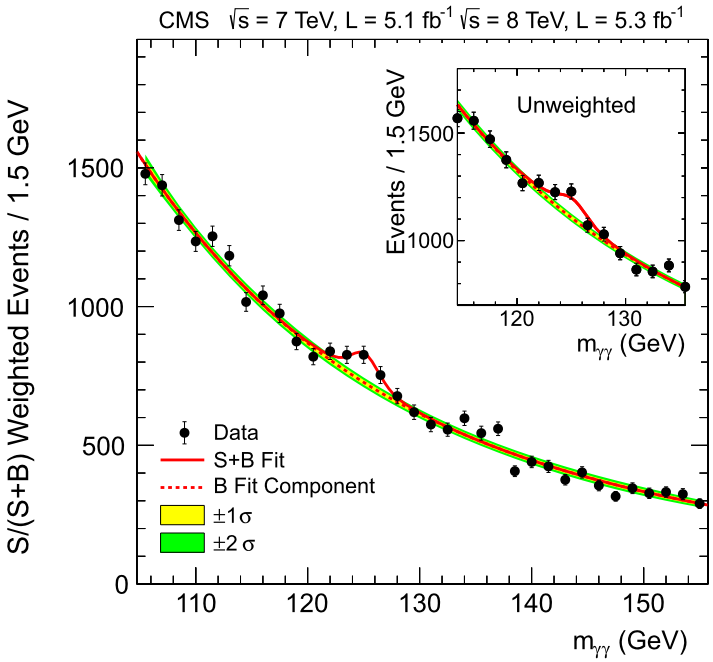


Figure 3.3: A glorified histogram from the 2012 paper discovering the Higgs boson in the CMS detector at the Large Hadron Collider. The small black circles are a histogram: they show the frequency of proton collision events that fall into each bin of invariant mass $m_{\gamma\gamma}$ (bins each have a width of 1.5 GeV). The dashed red line is what was expected from known physics at the time. The "bump" in the histogram at $m_{\gamma\gamma} = 125$ GeV is caused by the collider creating Higgs bosons. This histogram cost at least £6 billion to produce and resulted in a Nobel prize. Figure from S. Chatrchyan *et al.* [CMS], Phys. Lett. B **716**, 30-61 (2012).

■

3.1.2 Empirical CDF

The **empirical cumulative distribution function** (**CDF**) of a sample of real values $(x_1, \ldots, x_n)$ is written $F_n(x)$. At each value x, $F_n(x)$ tells you the proportion of the sample that have values less than or equal to x.

The formal definition is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x).$$

The function $I(x_i \leq x)$ (called an indicator function) is equal to 1 when $x_i \leq x$ and 0 when $x_i > x$. Therefore, when summing over all i, $\sum_{i=1}^{n} I(x_i \leq x)$ counts the number of x_i that are less than or equal to x.

Note that $F_n(x)$ is a step function, which jumps up by $1/n$ at each sample value. While the empirical CDF is somewhat less transparent than the histogram it does not require a choice of binning as the histogram does.

Example In Fig. 3.4 we compare the histogram with the empirical CDF using the same dataset. The histogram shows the concentration of data points around $x = 2$ more clearly. But the CDF makes it easy to make statements like "about 80% of the data has values less than $x = 3$."

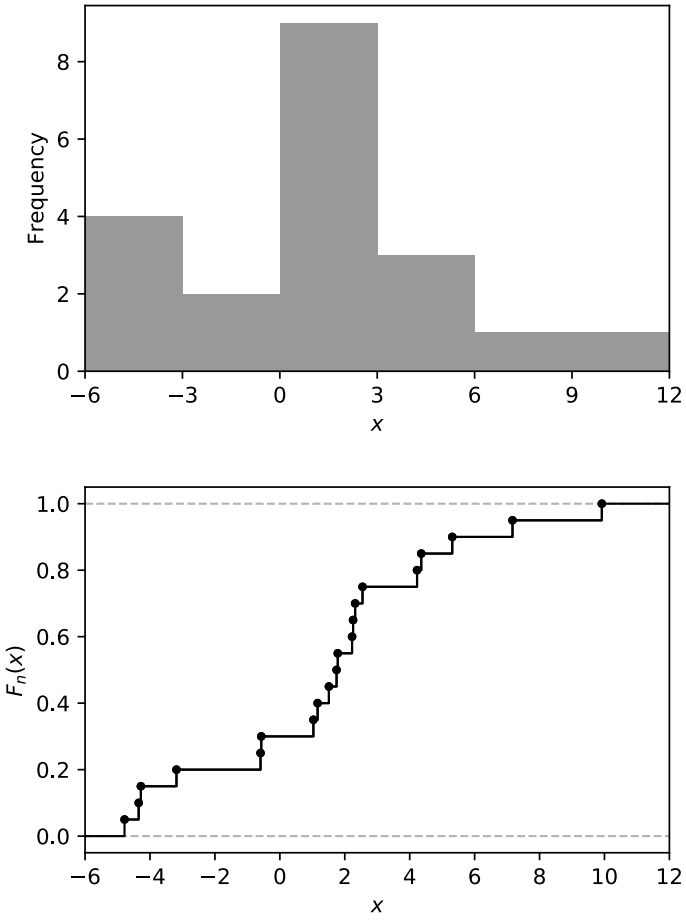


Figure 3.4

■

3.2 Summary Statistics

3.2.1 Measures of Location

The **arithmetic mean** (or just mean for short) of a sample of real values $(x_1, \ldots, x_n)$ is the sum of the values divided by their number. That is,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

This is often colloquially referred to as the average.

Example The mean of $(8, 3, 2, 12, 5)$ is

$$\frac{8+3+2+12+5}{5} = \frac{30}{5} = 6.$$

■

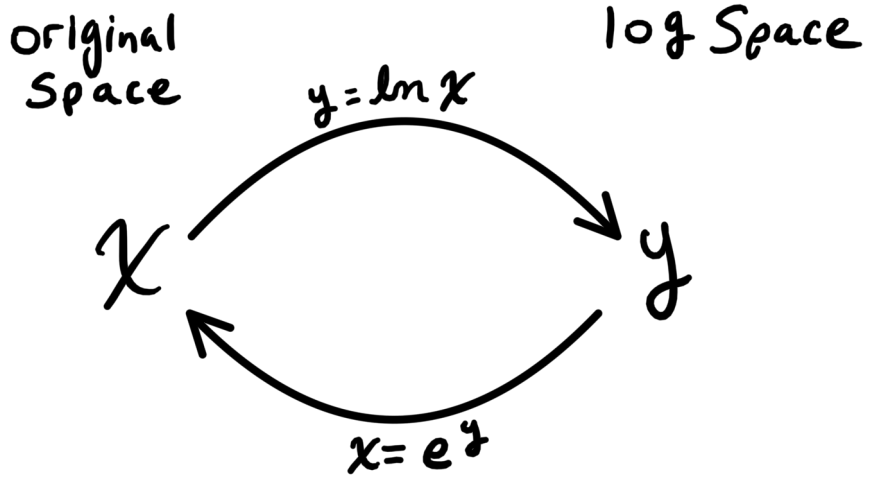
Two other useful measures of location related to the average are the geometric and harmonic means.

For positive data, the **geometric mean** is given by

$$x_G = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

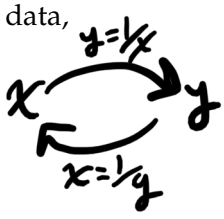
Note The geometric mean is equivalent to the regular arithmetic mean in "log space". I.e. we imagine *transforming* each data point x_i by taking its log. Then the arithmetic mean of these logs is just the log of the geometric mean.

Specifically, define $y_i = \log x_i$ for each i. Now calculate $\overline{y}$ and show this is equal to $\log x_G$.



The **harmonic mean** x_H is defined using the average of the reciprocals of the data,

$$\frac{1}{x_H} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \dots \frac{1}{x_n}\right),$$



which comes up naturally when the x_i are rates.

Note For positive data $(x_1, \dots, x_n)$,

$$\text{Arithmetic mean} \geq \text{geometric mean} \geq \text{harmonic mean}.$$

For the **median** we start with an intuitive definition. The median of a sample $(x_1, \ldots, x_n)$ is the "middle value" when the data are sorted into increasing order. If the number of data points is even we conventionally take the average of the two middle values.

Example

- The median of $(7, 2, 4, 12, 5)$ is 5.

- The median of $(7, 2, 4, 12, 5, 15)$ is $\dfrac{5+7}{2} = 6$.

■

More formally, for a sample of real values $(x_1, \ldots, x_n)$, define the i^{th} **order statistic** $x_{(i)}$ to be the i^{th} smallest value of the sample.

So

- $x_{(1)} \equiv \min(x_1, \ldots, x_n)$ is the smallest value;
- $x_{(2)}$ is the next smallest, and so on, up to

$\vdots$

- $x_{(n)} \equiv \max(x_1, \ldots, x_n)$ being the largest value.

In other words $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is just the dataset sorted into increasing order.

Example For the sample $(8, 3, 2, 12, 5)$ we have

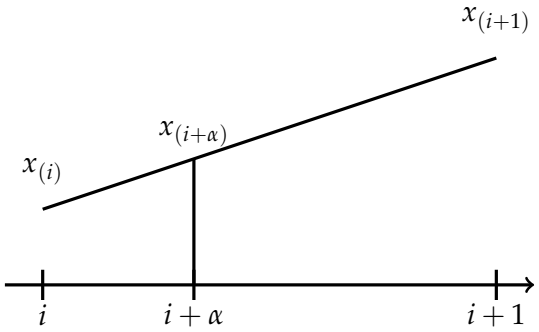
$$x_{(1)} = 2, \qquad x_{(2)} = 3, \qquad x_{(3)} = 5, \qquad x_{(4)} = 8, \qquad x_{(5)} = 12.$$

■

Furthermore, it will be useful to define $x_{(i)}$ when the number in parenthesis is not an integer. For integer $1 \leq i < n$ and non-integer $\alpha \in (0, 1)$ we define $x_{(i+\alpha)}$ as the linear interpolant

$$x_{(i+\alpha)} = (1 - \alpha)\ x_{(i)} + \alpha\ x_{(i+1)},$$

where the order statistics $x_{(i)}$ are defined as before.



Example

$$x_{(4.2)} = 0.8x_{(4)} + 0.2x_{(5)}.$$

■

The **median** of a sample of real values $(x_1, \ldots, x_n)$ is the middle value of the order statistics. That is, using our extended notation,

$$\text{median} = x_{\left(\frac{n+1}{2}\right)} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

The mean is sensitive to outlying points, whilst the median is not.

Example

$(1, 2, 3, 4, 5)$ has median = mean = 3

$(1, 2, 3, 4, 40)$ has median = 3, but now mean = 10

■

The arithmetic mean is the most commonly used *location* statistic, followed by the median.

The **mode** of a sample of real values $(x_1, \ldots, x_n)$ is the value of the x_i which occurs most frequently in the sample.

Example The mode of $(3, 5, 7, 2, 10, 14, 12, 2, 5, 2)$ is 2. ■

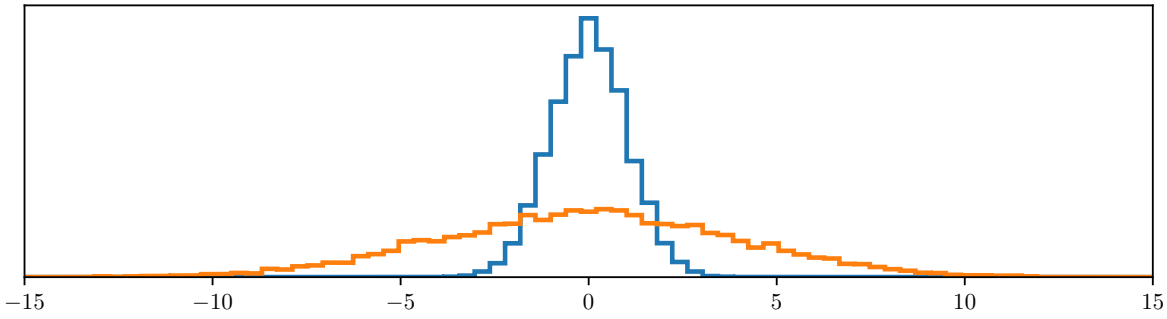
Note Some data sets are *multimodal*.

The mode of a sample is only useful when there can be repeated values (e.g. the x_i are integers or can only take on discrete values). If the x_i can take continuous values, e.g. if they are real numbers, then the idea of the "most common" value is pretty useless. Nonetheless, once we come to *probability distributions* we will see that the mode makes sense even for continuously distributed values.

3.2.2 Measures of Dispersion

The concept of dispersion of a sample relates to how spread out the values are. Are all the data points close together or far apart? Common words you'll come across that get at this notion are "spread" and "concentration". We'll talk about a few numerical measures that quantify this concept.

Example Use histograms to visualize



■

The most widely used measure of dispersion is based on the squared differences between the data points and their mean, $(x_i - \overline{x})^2$. The average (the mean) of these squared differences is the **mean square** or **sample variance**

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Equivalently, it is often more convenient to rewrite this formula as

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

That is, the mean of the squares minus the square of the mean.

The square root of the variance is the **root mean square** or **sample standard deviation**

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Unlike the variance, the standard deviation is in the same units as the x_i.

The **range** of a sample of real values $(x_1, \ldots, x_n)$ is the difference between the largest and the smallest values. That is

$$\text{range} = x_{(n)} - x_{(1)}$$

Example The range of $(7, 1, 4, 15, 5)$ is $15 - 1 = 14$. ■

Consider again the order statistics of a sample, $\left(x_{(1)}, \ldots, x_{(n)}\right)$.

We defined the *median* so that it lay approximately $\frac{1}{2}$ of the way through the ordered sample — not necessarily exactly or uniquely since there may be tied values or n even.

Similarly, we can define the **first** and **third quartiles** respectively as being values $\frac{1}{4}$ and $\frac{3}{4}$ of the way through the ordered sample:

$$\text{first quartile} = x_{\left(\frac{1}{4}(n+1)\right)}$$
$$\text{third quartile} = x_{\left(\frac{3}{4}(n+1)\right)}$$

and thus we define the **interquartile range** as the range of the data lying between the first and third quartiles,

$$\begin{aligned}\text{interquartile range} &= \text{third quartile} - \text{first quartile} \\ &= x_{\left(\frac{3}{4}(n+1)\right)} - x_{\left(\frac{1}{4}(n+1)\right)}\end{aligned}$$

The five point summary of a set of data lists, in order:

- The minimum value in the sample
- The lower quartile
- The sample median
- The upper quartile
- The maximum value

Each interval contains a quarter of the data points. E.g. between the minimum and lower quartile, or between the median and the upper quartile.

We can see analogies between the numerical summaries for location and dispersion, and their robustness properties are comparable.

	Least Robust	More Robust	Most Robust
Location	$\frac{x_{(1)} + x_{(n)}}{2}$	$\overline{x}$	$x_{\left(\frac{n+1}{2}\right)}$
Dispersion	$x_{(n)} - x_{(1)}$	s	$x_{\left(\frac{3}{4}(n+1)\right)} - x_{\left(\frac{1}{4}(n+1)\right)}$

(where $\frac{x_{(1)} + x_{(n)}}{2}$ would be the midpoint of our data halfway between the minimum and maximum values in the sample, which provides another alternative descriptor of location.)

3.2.3 Covariance and Correlation

The covariance and correlation apply to multidimensional data sets. For instance, say each sample is associated with an x and a y value, i.e. the ith sample is the pair (x_i, y_i). For example, each person in a sample has a height (x_i) and a weight (y_i). The data set can be represented by a sequence of ordered pairs $((x_1, y_1), \ldots (x_n, y_n))$.

Then the **covariance** between x and y for this sample is given by

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

It gives a measurement of relatedness between the two quantities x and y.

The covariance can be rewritten equivalently as

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{x}\,\overline{y}.$$

Note that the magnitude of s_{xy} varies according to the scale on which the data have been measured. The **correlation** normalizes the covariance by to the spread of the individual variables x and y. Specifically, the correlation r_{xy} is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ respectively.

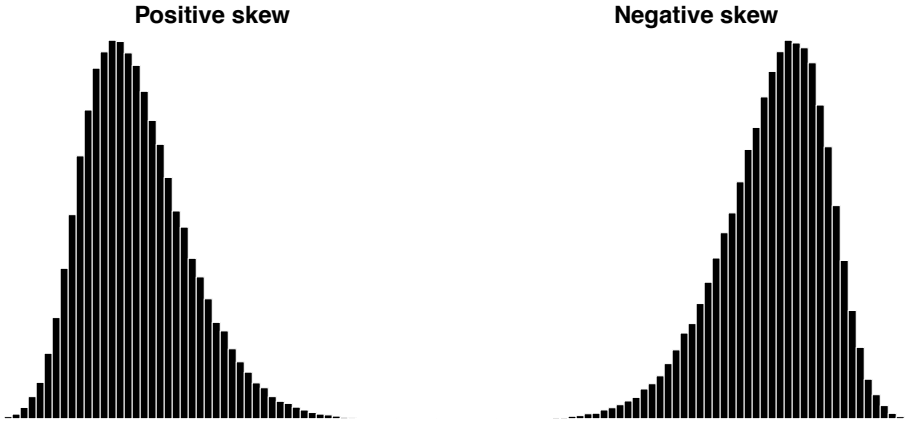
Unlike covariance, correlation gives a measurement of relatedness between the two quantities x and y which is scale-invariant. To see this pick two numbers a and b and create a new dataset where every x_i is multiplied by a and every y_i is multiplied by b. Show that the correlation of this new dataset is the same as in the original dataset. (Hint: $s_{xy} \to abs_{xy}$, $s_x \to as_x$, and $s_y \to bs_y$.)

3.2.4 Skewness

Skewness is a measure of asymmetry. The **skewness** of a sample of real values $(x_1, \ldots, x_n)$ is given by

$$\text{skewness} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3.$$

A sample is **positively** (**negatively**) or **right** (**left**) **skewed** if the upper tail of the histogram of the sample is longer (shorter) than the lower tail.



Since the mean is more sensitive to outlying points than is the median, one might choose the median as a more suitable measure of 'average value' if the sample is skewed.

We expect skewness for example when the data can only take positive (or only negative values) and if the values are not far from zero.

We can remove skewness by transforming the data. In the case above, we need a transformation which has larger effect on the larger values: e.g. square root, log (though beware 0 values).

Note For a positively skewed sample the mean is greater than the median.

3.3 One more visualization: the box-and-whisker plot

Based on the five point summary.

- Median – middle line in the box
- 3rd & 1st Quartiles – top and bottom of the box
- 'Whiskers' – extend out to any points which are within $(\frac{3}{2} \times$ interquartile range$)$ from the box
- Any extreme points out to the maximum and minimum which are beyond the whiskers are plotted individually.

Example Figure 3.5 are box plots of the counts of insects found in agricultural experimental units treated with six different insecticides (A-F).

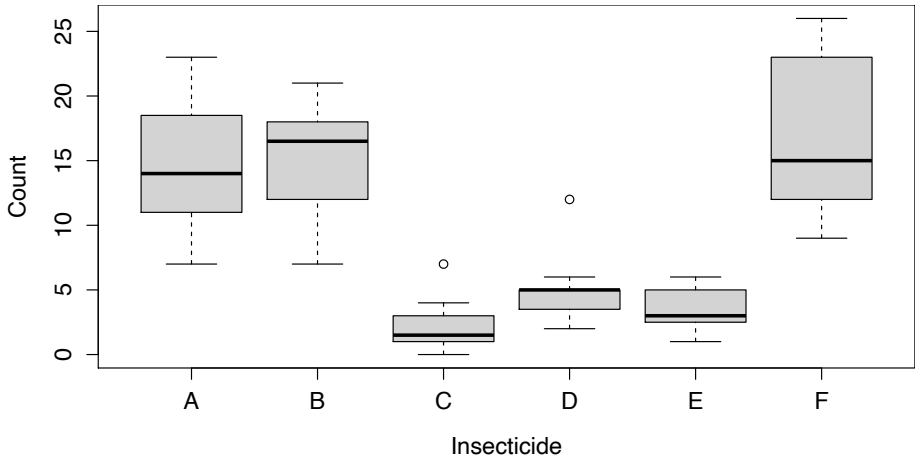


Figure 3.5

■