Graphing: One variable at a Time

Alan Arnholt

7/29/2015

Chapter 8

Graphing: One Variable at a Time

One Categorical Variable

Please watch the GRAPHING 1 video.

Consider the data frame `EPIDURALF` from the `PASWR2` package which records intermediate results from a study to determine whether the traditional sitting position or the hamstring stretch position is superior for administering epidural anesthesia to pregnant women in labor as measured by the number of obstructive (needle to bone) contacts. In this study, there were four physicians. To summarize the number of patients treated by each physician we can use the function `xtabs`.

```
library(PASWR2)
```

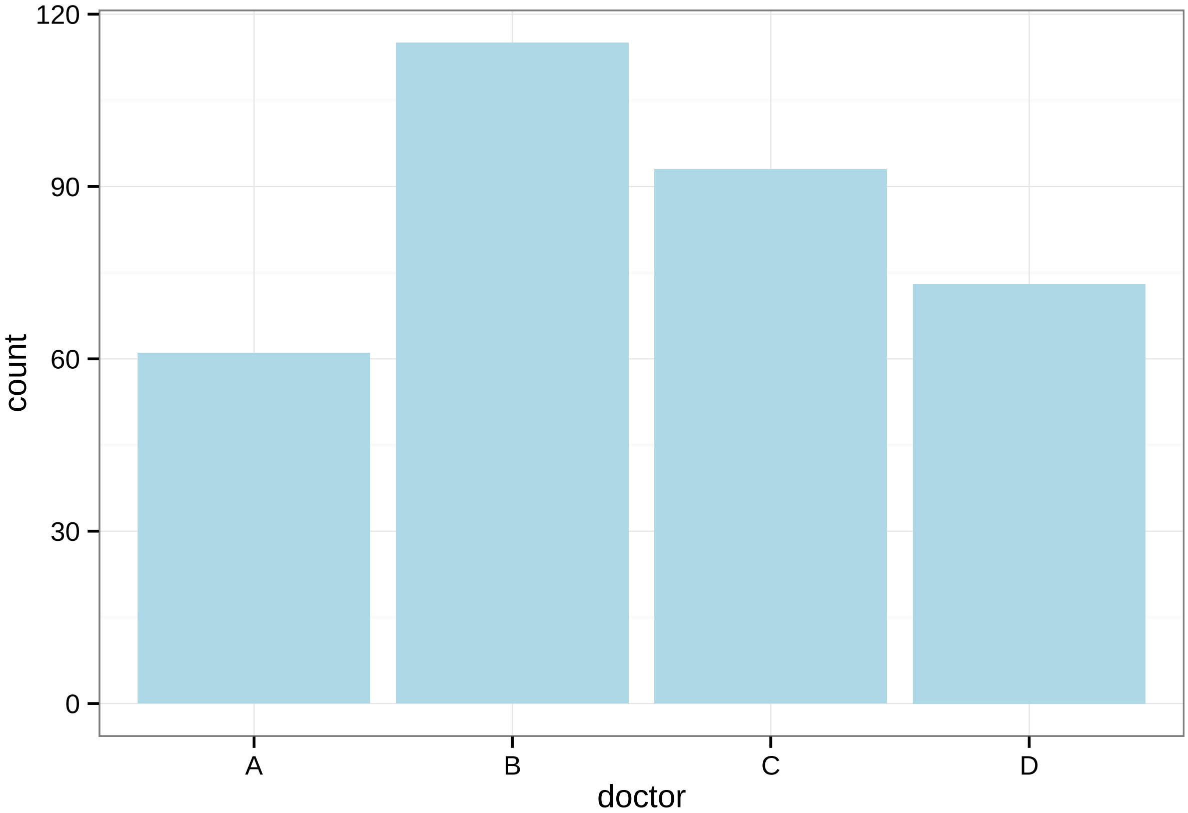
```
## Loading required package: ggplot2
## Loading required package: lattice
```

```
xtabs(~doctor, data = EPIDURALF)
```

```
## doctor
##   A   B   C   D
##  61 115  93  73
```

A barplot of the number of patients treated by each physician (`doctor`) using `ggplot2` is constructed below.

```
library(ggplot2)
ggplot(data = EPIDURALF, aes(x = doctor)) +
  geom_bar(fill = "lightblue") +
  theme_bw()
```



Here is some information that would be interesting to get from these data:

What percentage of the patients were treated by each physician?

```
prop.table(xtabs(~doctor, data = EPIDURALF))
```

```
## doctor
##         A         B         C         D
## 0.1783626 0.3362573 0.2719298 0.2134503
```

How are patients divided across physicians? Are they equally divided? If not, do the percentages follow some other kind of pattern?

One Quantitative Variable

We have explored the distribution of a categorical variable using a bar chart supplemented by numerical measures (percent of observations in each category). In this section, we will learn how to display the distribution of a quantitative variable.

To display data from one quantitative variable graphically, we typically use the histogram.

Example Break the following range of values into intervals and count how many observations fall into each interval.

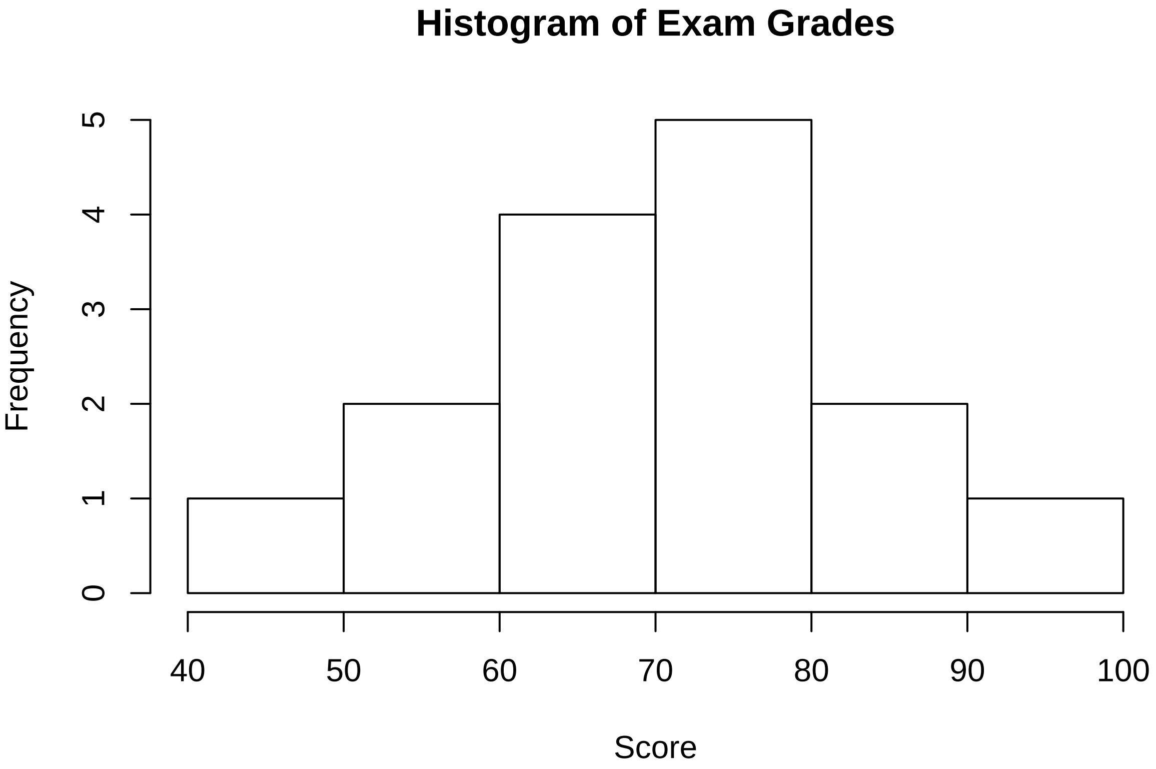
Exam Grades

Here are the exam grades of 15 students: 88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73

We first need to break the range of values into intervals (also called “bins” or “classes”). In this case, since our dataset consists of exam scores, it will make sense to choose intervals that typically correspond to the range of a letter grade, 10 points wide: 40-50, 50-60, ... 90-100. By counting how many of the 15 observations fall in each of the intervals, we get the following table:

SCORE	COUNT
[40,50)	1
[50,60)	2
[60,70)	4
[70,80)	5
[80,90)	2
[90,100)	1

To construct the histogram from this table we plot the intervals on the X-axis, and show the number of observations in each interval (frequency of the interval) on the Y-axis, which is represented by the height of a rectangle located above the interval:



Interpreting the Histogram

Once the distribution has been displayed graphically, we can describe the overall pattern of the distribution and mention any striking deviations from that pattern. More specifically, we should consider the following features of the distribution:

- Shape
- Center
- Spread
- Outliers

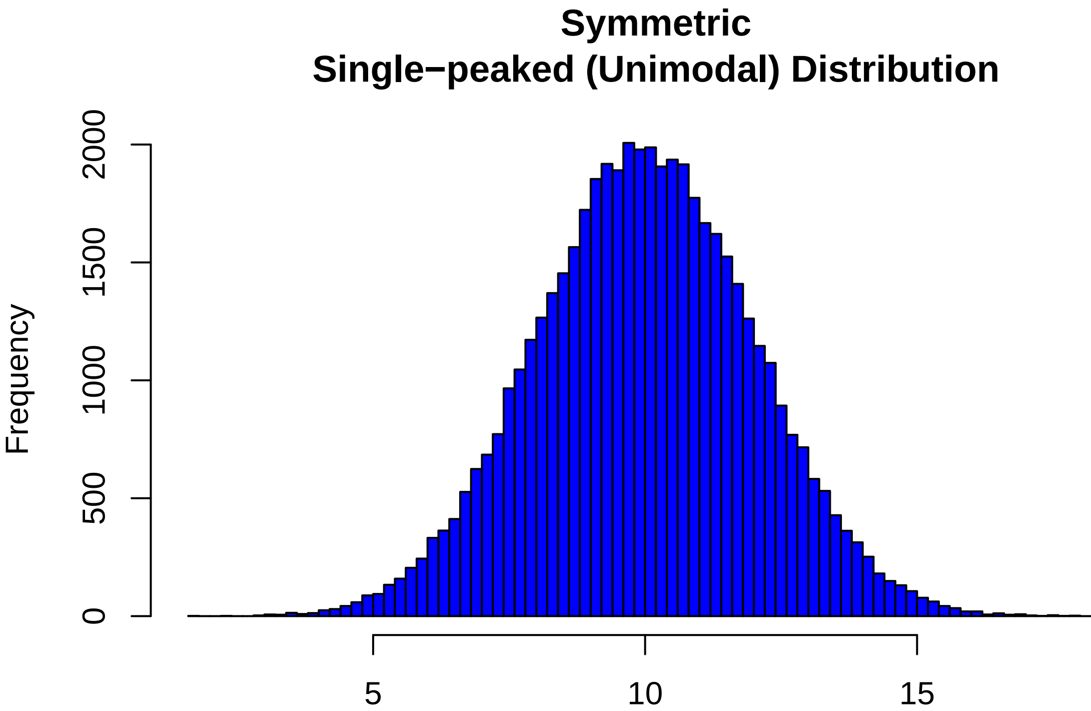
We will get a sense of the overall pattern of the data from the histogram's center, spread, and shape, while outliers will highlight deviations from that pattern.

Shape

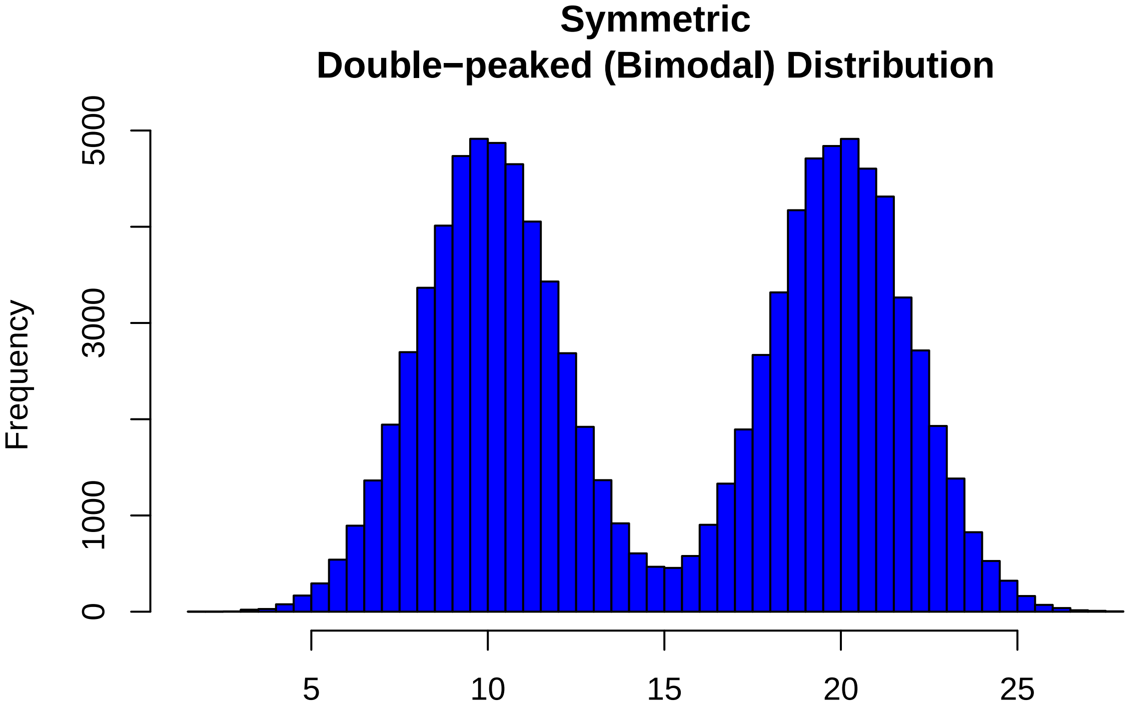
When describing the shape of a distribution, we should consider:

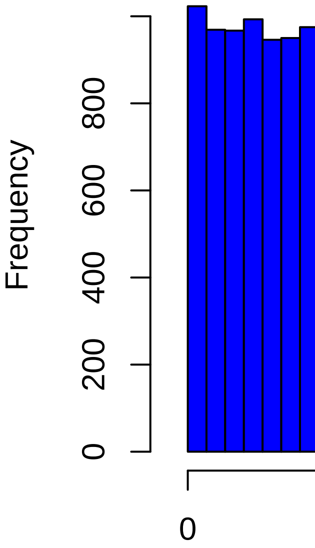
- Symmetry/skewness of the distribution.
- Peakedness (modality)—the number of peaks (modes) the distribution has.

We distinguish between:

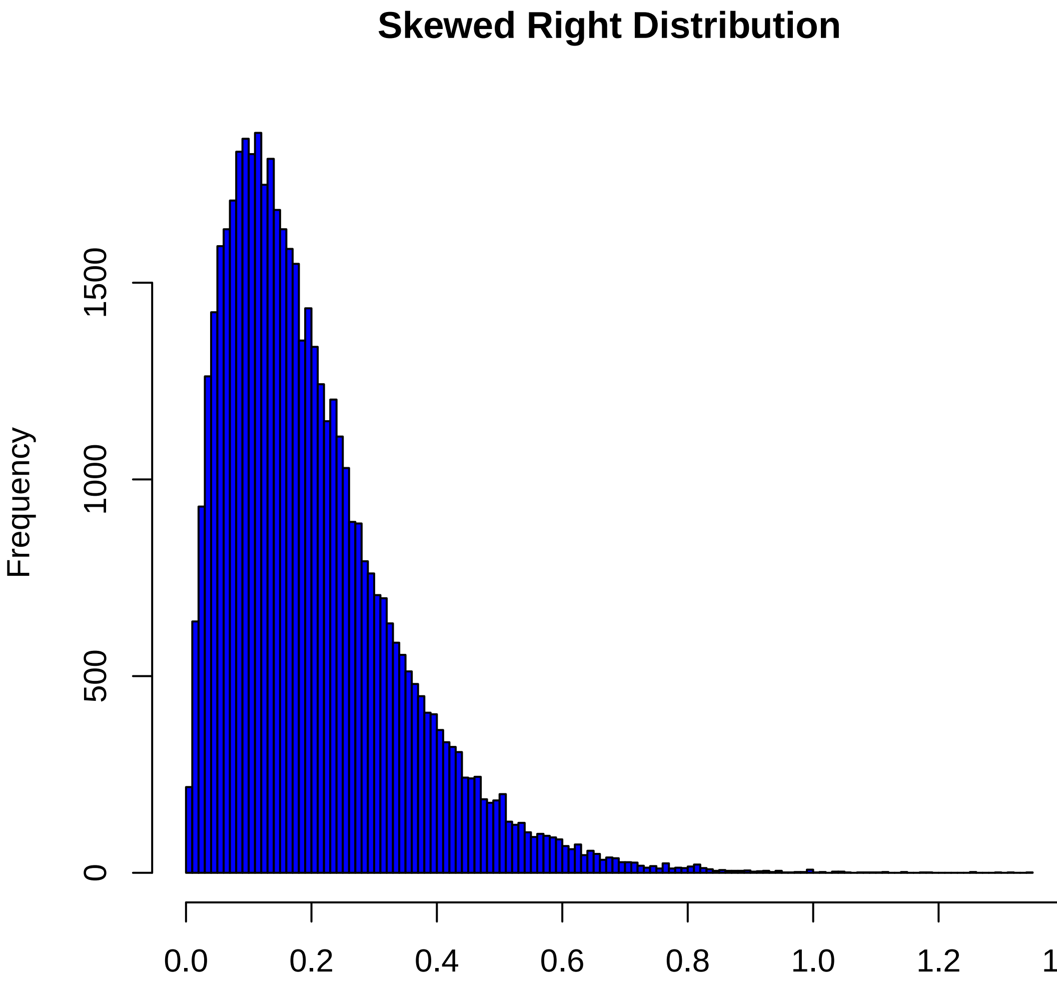


Symmetric Distributions



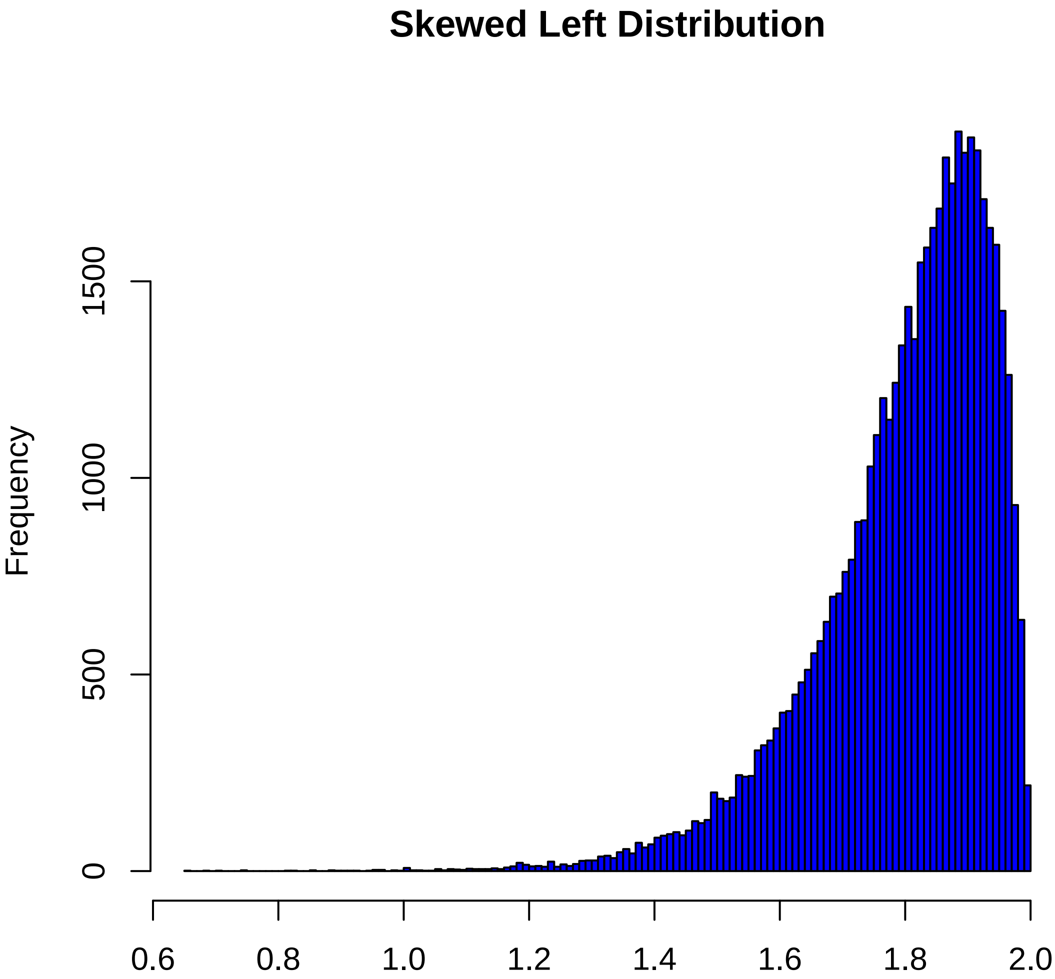


Note that all three distributions are symmetric, but are different in their modality (peakedness). The first distribution is unimodal—it has one mode (roughly at 10) around which the observations are concentrated. The second distribution is bimodal—it has two modes (roughly at 10 and 20) around which the observations are concentrated. The third distribution is kind of flat, or uniform. The distribution has no modes, or no value around which the observations are concentrated. Rather, we see that the observations are roughly uniformly distributed among the different values.



Skewed Right Distributions

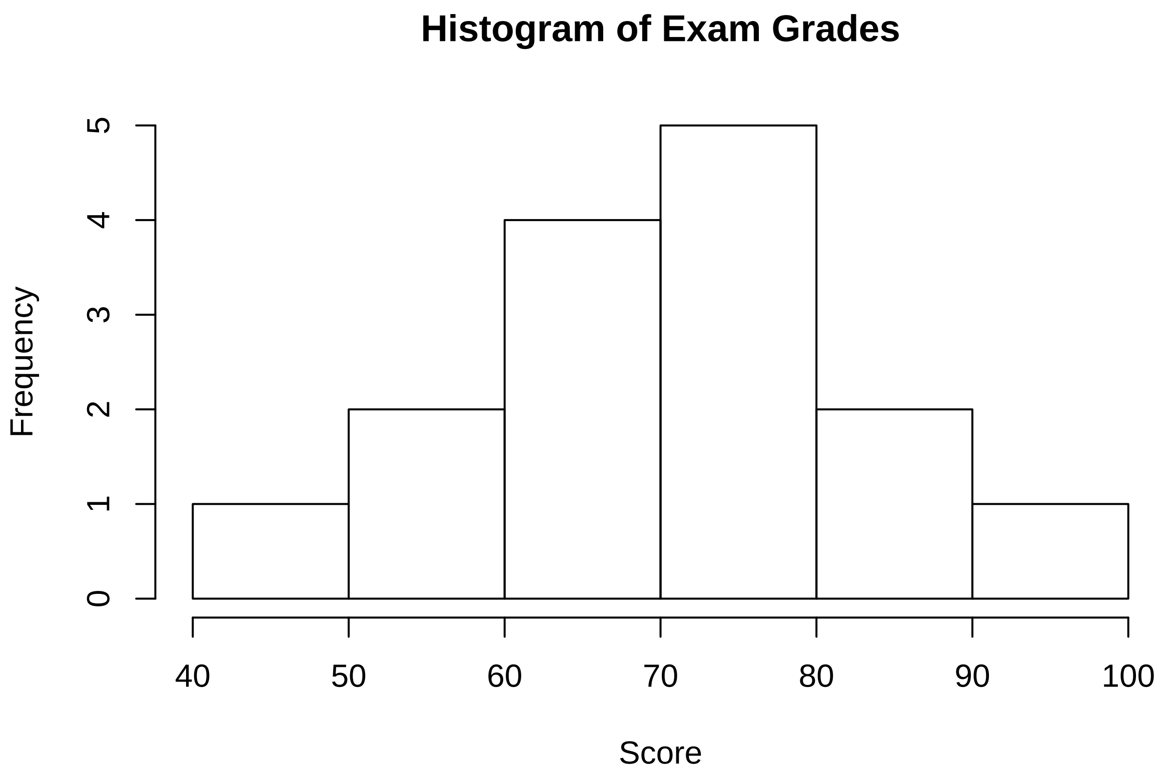
A distribution is called skewed right if, as in the histogram above, the right tail (larger values) is much longer than the left tail (small values). Note that in a skewed right distribution, the bulk of the observations are small/medium, with a few observations that are much larger than the rest. An example of a real-life variable that has a skewed right distribution is salary. Most people earn in the low/medium range of salaries, with a few exceptions (CEOs, professional athletes etc.) that are distributed along a large range (long “tail”) of higher values.



Skewed Left Distributions

A distribution is called skewed left if, as in the histogram above, the left tail (smaller values) is much longer than the right tail (larger values). Note that in a skewed left distribution, the bulk of the observations are medium/large, with a few observations that are much smaller than the rest. An example of a real life variable that has a skewed left distribution is age of death from natural causes (heart disease, cancer, etc.). Most such deaths happen at older ages, with fewer cases happening at younger ages.

Recall our grades example:



As you can see from the histogram, the grades distribution is roughly symmetric.

Center

The center of the distribution is its midpoint—the value that divides the distribution so that approximately half the observations take smaller values, and approximately half the observations take larger values. Note that from looking at the histogram we can get only a rough estimate for the center of the distribution. (More exact ways of finding measures of center will be discussed in the next section.)

Recall our grades example (image above). As you can see from the histogram, the center of the grades distribution is roughly 70 (7 students scored below 70, and 8 students scored above 70).

Spread

The spread (also called variability) of the distribution can be described by the approximate range covered by the data. From looking at the histogram, we can approximate the smallest observation (minimum), and the largest observation (maximum), and thus approximate the range.

In our example:

```
exam <- c(88, 48, 60, 51, 57, 85, 69, 75, 97, 72, 71, 79, 65, 63, 73)
min(exam)
```

```
## [1] 48
```

```
max(exam)
```

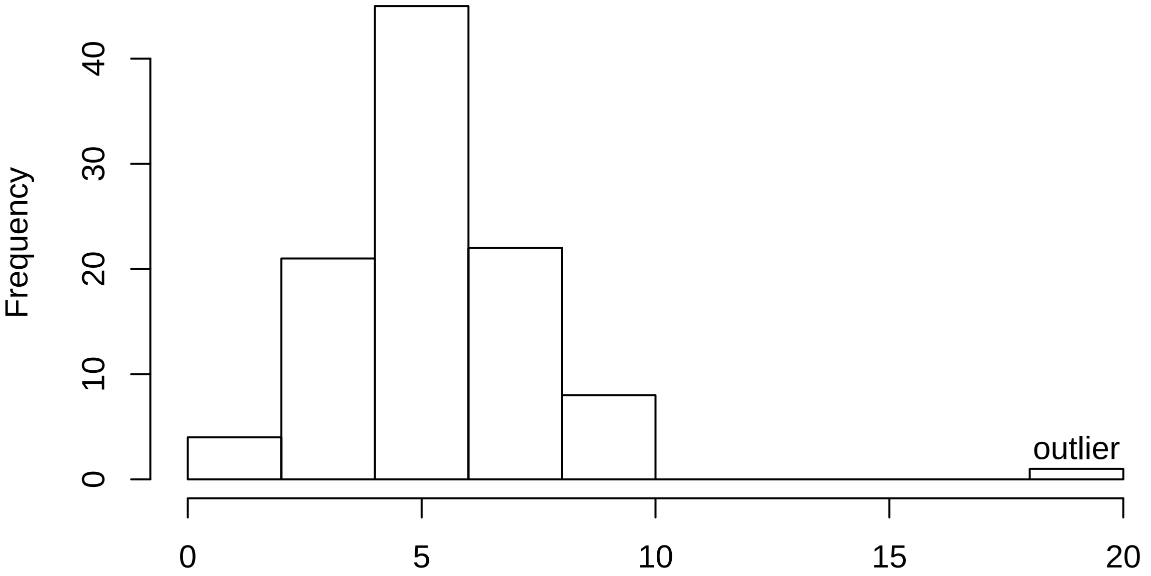
```
## [1] 97
```

```
range(exam)
```

```
## [1] 48 97
```

Outliers

Outliers are observations that fall outside the overall pattern. For example, the following histogram represents a distribution that has a high probable outlier:



The overall pattern of the distribution of a quantitative variable is described by its shape, center, and spread. By inspecting the histogram, we can describe the shape of the distribution, but, as we saw, we can only get a rough estimate for the center and spread. A description of the distribution of a quantitative variable must include, in addition to the graphical display, a more precise numerical description of the center and spread of the distribution.

The two main numerical measures for the center of a distribution are the mean and the median. Each one of these measures is based on a completely different idea of describing the center of a distribution.

Mean

The mean is the average of a set of observations (i.e., the sum of the observations divided by the number of observations). If the n observations are $x_1, x_2, \ldots, x_n$, their mean, which we denote $\bar{x}$ (and read x-bar), is therefore:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

.

World Cup Soccer

The data frame `SOCCER` from the `PASWR2` package contains how many goals were scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002.

Total # of Goals	Game Frequency
0	19
1	49
2	60
3	47
4	32
5	18
6	3
7	3
8	1

To find the mean number of goals scored per game, we would need to find the sum of all 232 numbers, then divide that sum by 232. Rather than add 232 numbers, we use the fact that the same numbers appear many times. For example, the number 0 appears 19 times, the number 1 appears 49 times, the number 2 appears 60 times, etc.

If we add up 19 zeros, we get 0. If we add up 49 ones, we get 49. If we add up 60 twos, we get 120. Repeated addition is multiplication.

Thus, the sum of the 232 numbers = 0(19) + 1(49) + 2(60) + 3(47) + 4(32) + 5(18) + 6(3) + 7(3) + 8(1) = 575. The mean is 575 / 232 = 2.478448.

This way of calculating a mean is sometimes referred to as a weighted average, since each value is "weighted" by its frequency.

```r
FT <- xtabs(~goals, data = SOCCER)
FT
```

```
## goals
##  0  1  2  3  4  5  6  7  8 
## 19 49 60 47 32 18  3  3  1
```

```r
pgoal <- FT/575
pgoal
```

```
## goals
##           0           1           2           3           4           5 
## 0.033043478 0.085217391 0.104347826 0.081739130 0.055652174 0.031304348 
##           6           7           8 
## 0.005217391 0.005217391 0.001739130
```

```r
ngoals <- as.numeric(names(FT))
ngoals
```

```
## [1] 0 1 2 3 4 5 6 7 8
```

```r
weighted.mean(x = ngoals, w = pgoal)
```

```
## [1] 2.478448
```

```r
mean(SOCCER$goals, na.rm = TRUE)
```

```
## [1] 2.478448
```

Meidan

The median (M) is the midpoint of the distribution. It is the number such that half of the observations fall above and half fall below. To find the median:

- Order the data from smallest to largest
- Consider whether n, the number of observations, is even or odd.
- If n is odd, the median M is the center observation in the ordered list. This observation is the one "sitting" in the $(n + 1)/2$ spot in the ordered list
- If n is even,the median M is the mean of the two center observations in the ordered list. These two observations are the ones "sitting" in the $n/2$ and $(n/2) + 1$ spots in the ordered list.