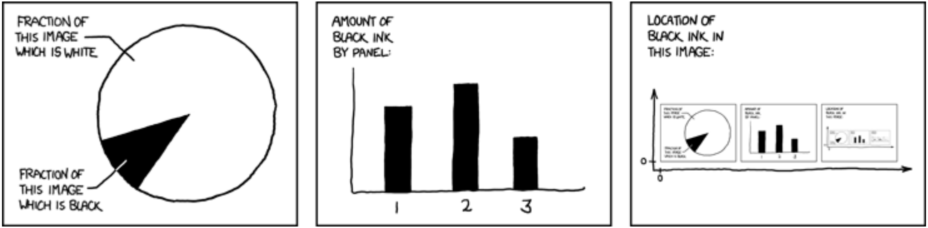
Ch. 1 – Overview and Descriptive Statistics



xkcd.com

Example 1 – Concrete Strength Data

The following measurements of flexural strength (in MPa) were taken from 27 specimens of high-performance concrete obtained by using superplasticizers and certain binders[1]:

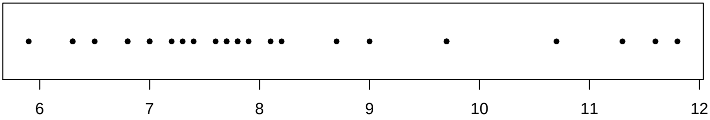
5.9, 7.2, 7.3, 6.3, 8.1, 6.8, 7.0, 7.6, 6.8, 6.5, 7.0, 6.3, 7.9, 9.0, 8.2, 8.7, 7.8, 9.7, 7.4, 7.7, 9.7, 7.8, 7.7, 11.6, 11.3, 11.8, 10.7

- How can we **visualize** this distribution?
- How can we quantify the **center** of the distribution?
- How can we quantify the **spread** of the distribution?

[1] From "Effects of Aggregates and Microfillers on the Flexural Properties of Concrete", **Magazine of Concrete Research**, 1997: 81–98

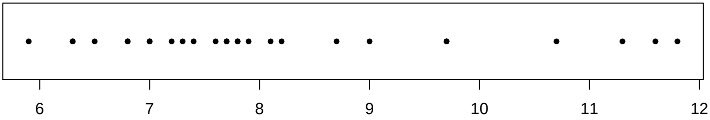
Dotplot

One simple method of visualization is a **dotplot** (or strip chart). The 27 observations are simply plotted on a line:

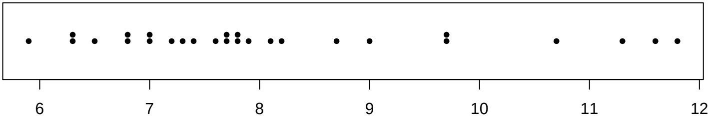


Dotplot

One simple method of visualization is a **dotplot** (or strip chart). The 27 observations are simply plotted on a line:



Identical values may be represented by stacking the dots:



Measures of center: Sample Mean and Median

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

2. The **sample median**: Sort the data from smallest to largest. If n is odd, then the sample median $\tilde{x}$ is the middle observation in the list; if n is even, then $\tilde{x}$ is the average of the two middle observations. Explicitly, if $x_1 \leq x_2 \leq \cdots \leq x_n$,

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

Measures of center: Sample Mean and Median

1. The **sample mean**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

2. The **sample median**: Sort the data from smallest to largest. If n is odd, then the sample median $\tilde{x}$ is the middle observation in the list; if n is even, then $\tilde{x}$ is the average of the two middle observations. Explicitly, if $x_1 \le x_2 \le \cdots \le x_n$,

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

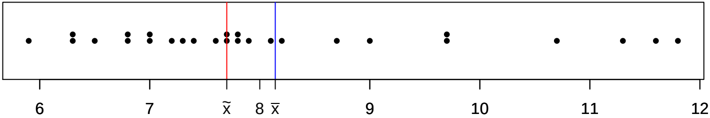
For example, given data 2, 3, 9, 11, 15, 17 we would have

$$\text{Sample mean: } \overline{x} = \frac{1}{6}(2 + 3 + 9 + 11 + 15 + 17) = 9.5$$

$$\text{Sample median: } \tilde{x} = \frac{1}{2}(9 + 11) = 10$$

Mean vs. Median

The sample mean and sample median for the strength data are $\overline{x} = 8.14$, $\tilde{x} = 7.7$.



- The mean may be strongly influenced by a few extreme observations, whereas the median is **robust** against such influence.
- If the largest measurement 11.8 were replaced 118, then the mean would increase to $\overline{x} = 12.07$, while the median would remain $\tilde{x} = 7.7$.

Mean vs. Median

- If the median is robust but the mean isn't, why use the mean?

Mean vs. Median

- If the median is robust but the mean isn't, why use the mean?
- In many cases, the mean is more **efficient** than the median: it achieves greater accuracy with a smaller sample size.

Mean vs. Median

- If the median is robust but the mean isn't, why use the mean?
- In many cases, the mean is more **efficient** than the median: it achieves greater accuracy with a smaller sample size.
- To address the mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.

Mean vs. Median

- If the median is robust but the mean isn't, why use the mean?
- In many cases, the mean is more **efficient** than the median: it achieves greater accuracy with a smaller sample size.
- To address the mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.
- To achieve a balanced tradeoff between efficiency and robustness, there are hybrid approaches such as a **trimmed mean**, e.g., discarding the top 10% and bottom 10% and then taking the mean of the remaining data.

Mean vs. Median

- If the median is robust but the mean isn't, why use the mean?
- In many cases, the mean is more **efficient** than the median: it achieves greater accuracy with a smaller sample size.
- To address the mean's sensitivity to outliers, we may carefully check in advance to screen out any bad data.
- To achieve a balanced tradeoff between efficiency and robustness, there are hybrid approaches such as a **trimmed mean**, e.g., discarding the top 10% and bottom 10% and then taking the mean of the remaining data.
- We'll discuss these issues in Chapter 6 (Point Estimation).

Measure of Spread: Sample variance

The **sample variance** is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

Measure of Spread: Sample variance

The **sample variance** is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

The **sample standard deviation** is the square root of the sample variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

Measure of Spread: Sample variance

The **sample variance** is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

The **sample standard deviation** is the square root of the sample variance:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}}$$

Like the mean, the sample variance and sample standard deviation may be strongly influenced by extreme observations.

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **lower fourth**.

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **lower fourth**.
- The median of the larger half is called the **upper fourth**.

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **lower fourth**.
- The median of the larger half is called the **upper fourth**.
- Their difference is called the **fourth spread** f_s:

$$f_s = \text{upper forth} - \text{lower fourth}$$

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **lower fourth**.
- The median of the larger half is called the **upper fourth**.
- Their difference is called the **fourth spread** f_s:

$$f_s = \text{upper forth} - \text{lower fourth}$$

Example: Given data 2,3,5,6,8,9,12,13,15 the median is $\tilde{x} = 8$, the lower fourth is 5, the upper fourth is 12, and the fourth spread is 7.

Robust Measure of Spread: Fourth spread

- Separate the data into the smaller half and larger half. (Include the median $\tilde{x}$ in both halves if n is odd.)
- The median of the smaller half is called the **lower fourth**.
- The median of the larger half is called the **upper fourth**.
- Their difference is called the **fourth spread** f_s:

$$f_s = \text{upper forth} - \text{lower fourth}$$

Example: Given data 2,3,5,6,8,9,12,13,15 the median is $\tilde{x} = 8$, the lower fourth is 5, the upper fourth is 12, and the fourth spread is 7.

	Efficient	Robust
Measure of center	$\overline{x}$	$\tilde{x}$
Measure of spread	s	f_s

Five-number summary and Boxplot

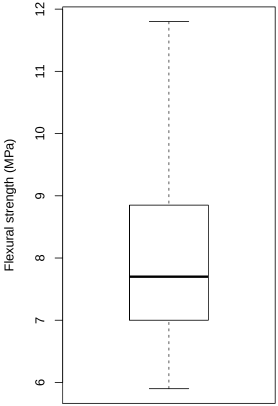
The **five-number summary** consists of

1. The minimum observation
2. The lower fourth
3. The median
4. The upper fourth
5. The maximum observation

For the concrete strength data:

$$5.90, 7.00, 7.70, 8.85, 11.80$$

This is shown graphically in a **boxplot**: The bottom and top of the box show the lower and upper fourths; the horizontal line inside the box shows the median; the whiskers (dotted lines) extend to the minimum and maximum observations.



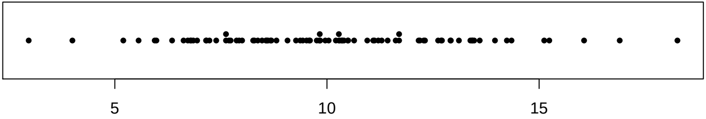
Example 2 – Power Usage

Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from a power company determined energy consumption (in BTU) for a sample of 90 homes over a fixed time interval:

2.97, 4.00, 5.20, 5.56, 5.94, 5.98, 6.35, 6.62, 6.72, 6.78, 6.80, 6.85, 6.94, 7.15, 7.16, 7.23, 7.39, 7.62, 7.62, 7.69, 7.73, 7.87, 7.93, 8.00, 8.26, 8.29, 8.37, 8.47, 8.56, 8.58, 8.61, 8.67, 8.69, 8.81, 9.07, 9.27, 9.37, 9.43, 9.52, 9.58, 9.60, 9.76, 9.82, 9.83, 9.83, 9.84, 9.96, 10.04, 10.21, 10.28, 10.28, 10.30, 10.35, 10.36, 10.40, 10.49, 10.50, 10.64, 10.95, 11.09, 11.12, 11.21, 11.29, 11.43, 11.62, 11.70, 11.70, 12.16, 12.19, 12.28, 12.31, 12.62, 12.69, 12.71, 12.91, 12.92, 13.11, 13.38, 13.42, 13.43, 13.47, 13.60, 13.96, 14.24, 14.35, 15.12, 15.24, 16.06, 16.90, 18.26

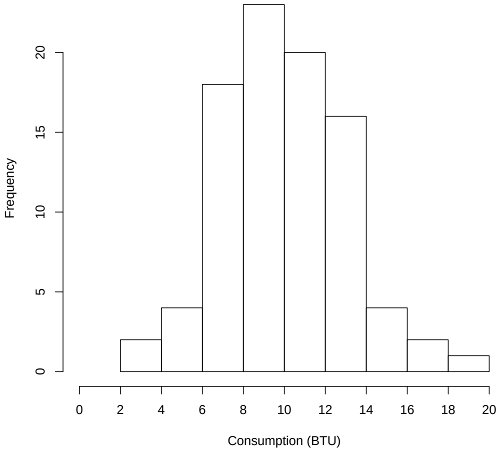
Example 2 – Power Usage

The dotplot is less effective here: many dots are packed close together, making it difficult to see where the dots are most concentrated:



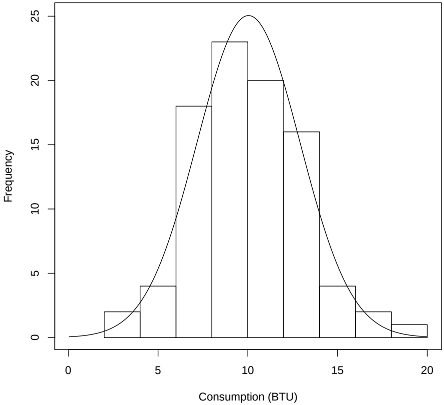
Histogram

A **histogram** is constructed by grouping the data into bins of equal width and counting the number of observations within each bin:



Here the bins are the intervals (0,2], (2,4], (4,6], (6,8], . . . , (16,18], (18,20].

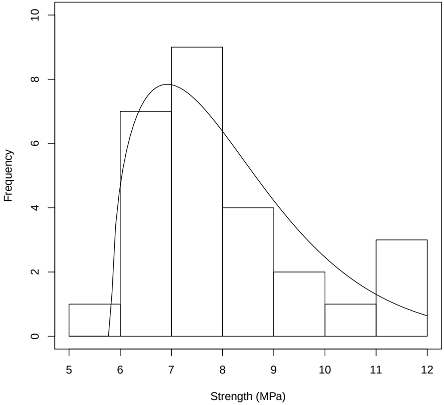
This histogram appears to be a fairly close fit to the symmetric, bell-shaped density curve of a **normal distribution**:



Normal distributions are very common in statistics and will be discussed in Chapter 4.

Weibull Distribution

In contrast, the concrete strength data appears asymmetric and does not seem to fit a normal distribution. A better fit is provided by a **Weibull distribution**.



Stem-and-leaf diagram

3	|	0
4	|	0
5	|	269
6	|	03678889
7	|	2224667799
8	|	03345666778
9	|	134456688888
10	|	0023333445569
11	|	11234677
12	|	223367799
13	|	144456
14	|	023
15	|	12
16	|	19
17	|	
18	|	3

A **stem-and-leaf diagram** is constructed by grouping the data according to their first one or two digits (the **stem**), shown on the left part of the diagram, while the remaining digit (the **leaf**) is displayed on the right.

Here the first few measurements are 3.0, 4.0, 5.2, 5.6, 5.9, 6.0, 6.3, etc.

A stem-and-leaf diagram is similar to a "sideways" histogram. It is visually less appealing but provides more precise information about the data.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.
- A **parameter** is a quantity which describes a population, e.g., the population mean μ.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.
- A **parameter** is a quantity which describes a population, e.g., the population mean μ.
- A **statistic** is a quantity computed from sample data, e.g., the sample mean $\overline{x}$.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.
- A **parameter** is a quantity which describes a population, e.g., the population mean μ.
- A **statistic** is a quantity computed from sample data, e.g., the sample mean $\overline{x}$.
- **Statistical inference** is the method of drawing *probabilistic* conclusions about a population based on sample data.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.
- A **parameter** is a quantity which describes a population, e.g., the population mean μ.
- A **statistic** is a quantity computed from sample data, e.g., the sample mean $\overline{x}$.
- **Statistical inference** is the method of drawing *probabilistic* conclusions about a population based on sample data.
  - A **confidence interval** gives a range of values in which we can expect a parameter to be, with a specific degree of confidence.

Statistical Inference: Parameters vs. Statistics

- A **population** is a group that we wish to study.
- A **sample** is an observed subset of a population.
- A **parameter** is a quantity which describes a population, e.g., the population mean μ.
- A **statistic** is a quantity computed from sample data, e.g., the sample mean $\overline{x}$.
- **Statistical inference** is the method of drawing *probabilistic* conclusions about a population based on sample data.
  - A **confidence interval** gives a range of values in which we can expect a parameter to be, with a specific degree of confidence.
  - A **hypothesis test** provides a way to test whether a particular theory is consistent with the data or not. We cannot be certain our conclusion is correct, but we can quantify the degree of uncertainty using probability.

Statistical Inference: Confidence Intervals

- In the concrete strength data, based on 27 specimens, the sample mean was $\overline{x} = 8.14$.

Statistical Inference: Confidence Intervals

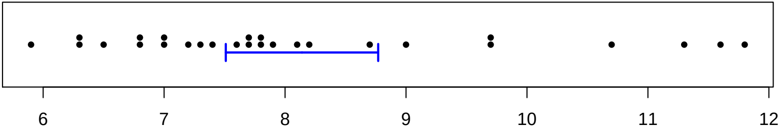
- In the concrete strength data, based on 27 specimens, the sample mean was $\overline{x} = 8.14$.
- If the process were repeated under the same conditions, creating a new sample of 27 specimens, the new sample may have a somewhat different sample mean.

Statistical Inference: Confidence Intervals

- In the concrete strength data, based on 27 specimens, the sample mean was $\overline{x} = 8.14$.
- If the process were repeated under the same conditions, creating a new sample of 27 specimens, the new sample may have a somewhat different sample mean.
- The sample mean $\overline{x}$ is only an estimate of the true **population mean** μ.

Statistical Inference: Confidence Intervals

- In the concrete strength data, based on 27 specimens, the sample mean was $\overline{x} = 8.14$.
- If the process were repeated under the same conditions, creating a new sample of 27 specimens, the new sample may have a somewhat different sample mean.
- The sample mean $\overline{x}$ is only an estimate of the true **population mean** μ.
- The accuracy of this estimate may be quantified using a **confidence interval**: In this case, given the sample of 27 specimens, we can say with 95% confidence that the population mean μ is between 7.51 and 8.77.



Statistical Inference: Hypothesis Tests

- Suppose a second method of producing high-strength concrete is tested: 30 specimens have a sample mean flexural strength of $\overline{y} = 8.90$, compared to $\overline{x} = 8.14$ for the original method.

Statistical Inference: Hypothesis Tests

- Suppose a second method of producing high-strength concrete is tested: 30 specimens have a sample mean flexural strength of $\overline{y} = 8.90$, compared to $\overline{x} = 8.14$ for the original method.
- Does the second method produce stronger concrete? Or could the difference be due to chance?

Statistical Inference: Hypothesis Tests

- Suppose a second method of producing high-strength concrete is tested: 30 specimens have a sample mean flexural strength of $\overline{y} = 8.90$, compared to $\overline{x} = 8.14$ for the original method.
- Does the second method produce stronger concrete? Or could the difference be due to chance?
- To answer this, we need to know the sample standard deviations of the two samples: $s_1 = 1.66$, $s_2 = 1.91$.

Statistical Inference: Hypothesis Tests

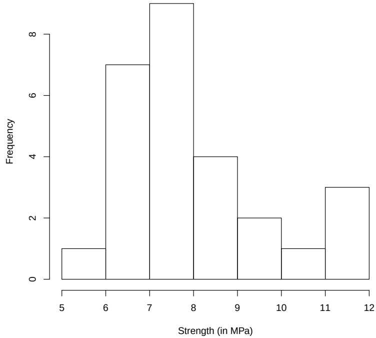
- Suppose a second method of producing high-strength concrete is tested: 30 specimens have a sample mean flexural strength of $\overline{y} = 8.90$, compared to $\overline{x} = 8.14$ for the original method.
- Does the second method produce stronger concrete? Or could the difference be due to chance?
- To answer this, we need to know the sample standard deviations of the two samples: $s_1 = 1.66$, $s_2 = 1.91$.
- The appropriate hypothesis test (two-sample t test) gives a P-**value** of .06, which means that if there is no difference in the population means of the two methods, there is only a 6% chance that the sample means $\overline{x}$ and $\overline{y}$ would differ by as much as they did.

Statistical Inference: Hypothesis Tests

- Suppose a second method of producing high-strength concrete is tested: 30 specimens have a sample mean flexural strength of $\overline{y} = 8.90$, compared to $\overline{x} = 8.14$ for the original method.
- Does the second method produce stronger concrete? Or could the difference be due to chance?
- To answer this, we need to know the sample standard deviations of the two samples: $s_1 = 1.66$, $s_2 = 1.91$.
- The appropriate hypothesis test (two-sample t test) gives a P-**value** of .06, which means that if there is no difference in the population means of the two methods, there is only a 6% chance that the sample means $\overline{x}$ and $\overline{y}$ would differ by as much as they did.
- This provides some evidence (but not strong evidence) that the second method outperforms the first in mean flexural strength.

Statistical Inference: Hypothesis Tests

Based on the histogram, we remarked that the concrete strength data did not appear to have a normal distribution. However, appearances can be misleading, especially in small samples.



A normality test (Shapiro-Wilk test) gives a P-value of .008, providing strong evidence that indeed the data is not normal.

Key concepts

- Ways to display the distribution of a numerical variable:
  - dotplot, boxplot, histogram, stem-and-leaf diagram

Key concepts

- Ways to display the distribution of a numerical variable:
  - dotplot, boxplot, histogram, stem-and-leaf diagram
- Measures of center: mean, median

Key concepts

- Ways to display the distribution of a numerical variable:
  - dotplot, boxplot, histogram, stem-and-leaf diagram
- Measures of center: mean, median
- Measures of spread: variance, standard deviation, fourth spread

Key concepts

- Ways to display the distribution of a numerical variable:
  - dotplot, boxplot, histogram, stem-and-leaf diagram
- Measures of center: mean, median
- Measures of spread: variance, standard deviation, fourth spread
- Statistical inference: using *sample* data to draw probabilistic conclusions about a broader *population*.