Lecture 02. Statistics and probability theory.

Numerical characteristics of data sets

Let us consider a sequence $x_1, x_2, x_3, \ldots, x_n$ of the values of some quantitative variable X.
We define the basic characteristics of the distribution of the dataset X:

- the **smallest value** $a_{min} = \min\limits_n x_n$ and the **largest value** $b_{max} = \max\limits_n x_n$ values of X.

- the center of a distribution. It can be taken as

  a) the **mean value** $\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n)$

  b) the **median** M : it is the midpoint of a distribution. Half the observations are smaller than the median and the other half are larger than the median. Here is a rule for finding the median:

  1. Arrange all observations in order of size, from the smallest to the largest:

  $$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

  2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n+1)/2$ observations up from the bottom of the list.

  3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n+1)/2$ from the bottom of the list, i.e. it is the mean value of $(n/2)$-th and $(n/2+1)$-th observations.

  Example



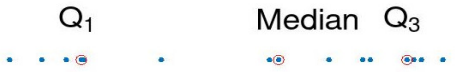


When you use the median to describe the center of the distribution, describe its spread by giving the quartiles.

- the **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.

The quartiles Q_1 and Q_3

To calculate the quartiles:
1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.



The **five-number summary** consisting of

the **median**, the **quartiles**, and the **smallest** and **largest individual observations**

provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

Boxplots based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full

spread of the data. In a modified boxplot, points identified by the 1.5× IQR rule are plotted individually.

A boxplot is a graph of the five-number summary.

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

Example. Fuel consumption in city driving and on the highway

Fuel economy (miles per gallon) for 2004 model vehicles					
Two-Seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

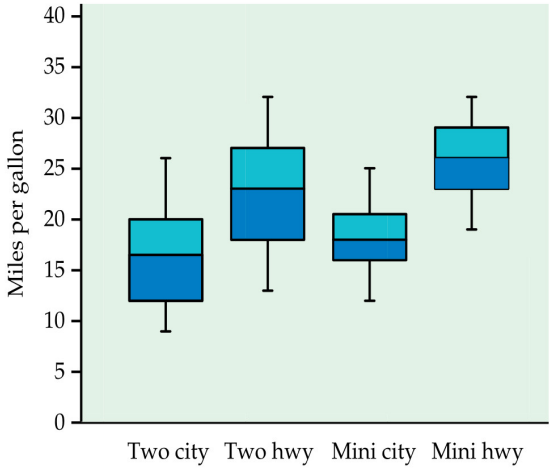
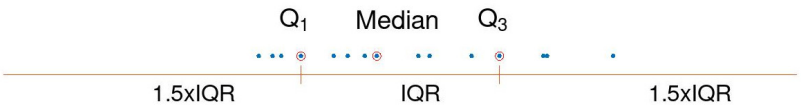


FIGURE 1.19 Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts by the Environmental Protection Agency.

The interquartile range is the difference $IQR = Q_3 - Q_1$ between the quartiles. It is the spread of the center half of the data.



The 1.5×IQR rule for outliers. Call an observation a suspected outlier if it falls more than 1.5×IQR above the third quartile or below the first quartile.

making boxplots:

The five-number summary The five-number summary of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is $$\text{Minimum}, \quad Q_1, \quad \text{M}(= Q_2), \quad Q_3, \quad \text{Maximum}$$

4. Let $x_1, x_2, \ldots, x_{n-1}, x_n$ be a data sequence. We define the **variance** as follows

$$s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_{n-1} - \bar{x})^2 + (x_n - \bar{x})^2\right)$$

and especially its square root, the **standard deviation** s, are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.

Caution.

We average by dividing by $n-1$ rather than n in calculating the variance because the sum of the deviations is always zero $\left(\sum_i^n (x_i - \bar{x}) = 0\right)$, the last deviation $x_n - \bar{x}$ can be found once we know the other $n-1$. So we are not averaging n unrelated numbers. Only $n-1$ of the squared deviations can vary freely, and we average by dividing the total by $n-1$.
The number $n-1$ is called the **degrees of freedom** of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n-1$, so be sure to use $n-1$.

A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.

The **mean** and **standard deviation** are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced below. The **five-number summary** is a better exploratory summary for skewed distributions.

Linear transformations have the form $x_{new} = a + bx$. A linear transformation changes the origin if $a \neq 0$ and changes the size of the unit of measurement if $b > 0$. Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by b and changes a percentile or measure of center m into $a + bm$. Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.

- skewness $= \dfrac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^3}{\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right)^{3/2}}$ is the measure of the asymetry.

Negative, or left-skewed, refers to a longer or fatter tail on the left side of the distribution, while positive, or right-skewed, refers to a longer or fatter tail on the right. These two skews show the direction or weight of the distribution. A distribution is skewed to the right if the right tail (larger values) is much longer than the left tail (smaller values). Money amounts, usually have right-skewed distributions. For example, there are many moderately priced houses, but the few very expensive mansions give the distribution of house prices a strong right-skew.

- kurtosis $= \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^4}{\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2\right]^2} - 3$ - higher kurtosis corresponds to greater extremity of deviations (or outliers), and not the configuration of data near the mean.

Comments

- The functions of data element values $x_1, x_2, x_3, \ldots, x_{n-1}, x_n$ (like mean, variance, median, ..) are called **statistics.**

- Always plot your data: make a graph, usually a stemplot or a histogram. Look for the overall pattern and for striking deviations such as outliers. Calculate an appropriate numerical summary to briefly describe center and spread.

Density Curves and Normal Distributions

Density curve

1. Normalized histogram:
The histogram shows the numbers n_i of data elements contained in the bin with index i. Thus $n_1 + n_2 + \ldots n_k = n$, where k is a number of bins and n is the total number of data elements. A histogram in which the bars have a height of n_i/n instead of n_i is a normalized histogram. It represents the proportion of data elements contained in the interval with index i.

2. As the number n of data points and the number of bins approach infinity, the contour of the normalized histogram approaches a curve known as the probability density of the dataset.

3. The density curve is characterized by the following property: it describes the general shape of the distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range.

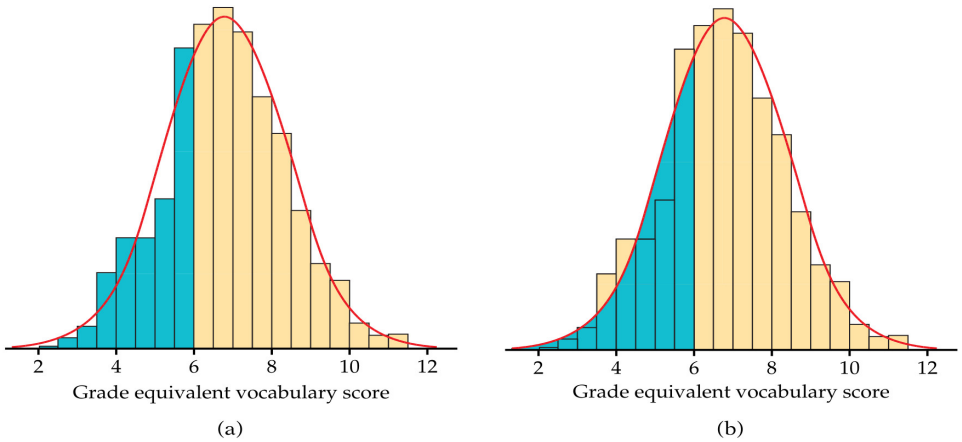


FIGURE 1.25 (a) The distribution of Iowa Test vocabulary scores for Gary, Indiana, seventh-graders. The shaded bars in the histogram represent scores less than or equal to 6.0. The proportion of such scores in the data is 0.303. (b) The shaded area under the Normal density curve also represents scores less than or equal to 6.0. This area is 0.293, close to the true 0.303 for the actual data.

A density curve has the following properties:

- is always on or above the horizontal axis and
- has area exactly 1 underneath it.

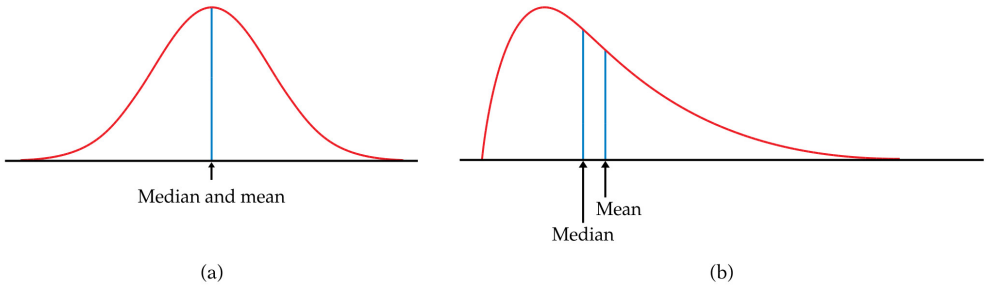


FIGURE 1.26 (a) A symmetric density curve with its mean and median marked. (b) A right-skewed density curve with its mean and median marked.

Normal distributions

One particularly important class of density curves consists of curves which

are symmetric, unimodal and bellshaped. They are called Normal curves given by the mathematical formula

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

and they describe Normal distributions $N(\mu, \sigma)$.

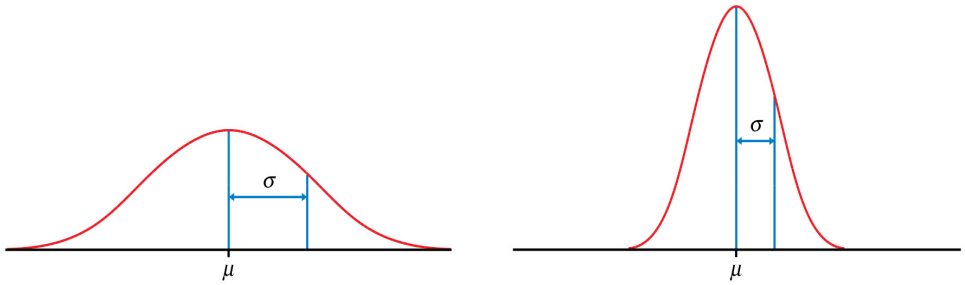


FIGURE 1.28 Two Normal curves, showing the mean μ and standard deviation σ.

Although there are many Normal curves, they all have common properties. Here is one of the most important.

The 68-95-99.7 rule

In the Normal distribution with mean μ and standard deviation σ :

- Approximately 68% of the observations fall within σ of the mean μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

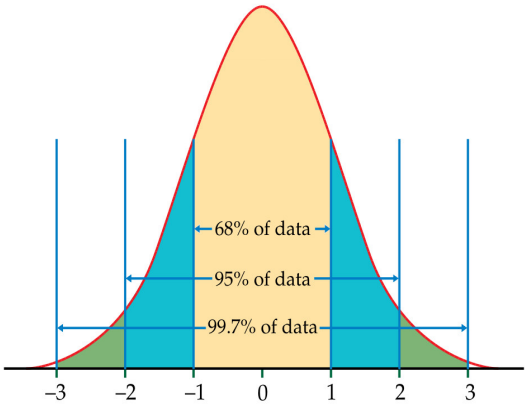


FIGURE 1.29 The 68–95–99.7 rule for Normal distributions.

Example

1.25 **Heights of young women.** The distribution of heights of young women aged 18 to 24 is approximately Normal with mean $\mu = 64.5$ inches and standard deviation $\sigma = 2.5$ inches. Figure 1.30 shows what the 68-95-99.7 rule says about this distribution. Two standard deviations is 5 inches for this distribution. The 95 part of the 68-95-99.7 rule says that the middle 95% of young women are between 64.5 - 5 and 64.5 + 5 inches tall, that is, between 59.5 inches and 69.5 inches. This fact is exactly true for an exactly Normal distribution. It is approximately

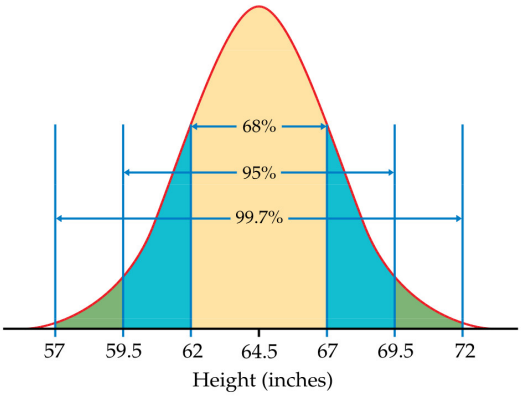


FIGURE 1.30 The 68–95–99.7 rule applied to the heights of young women, for Example 1.25.

true for the heights of young women because the distribution of heights is approximately Normal. The other 5% of young women have heights outside the range from 59.5 to 69.5 inches. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches.

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with parameters μ and σ as $N(\mu, \sigma)$. We will see that μ is the mean value and σ is the standard deviation of the distribution. For example, the distribution of young women's heights is N(64.5, 2.5).

The question arises, why are the Normal distributions important in statistics? Here are three reasons.

- First, Normal distributions are good descriptions for some distributions of real data. Distributions that are often close to Normal include scores on tests taken by many people (such as the Iowa Test of Figure 1.25), repeated careful measurements of the same quantity, and characteristics

of biological populations (such as lengths of baby pythons and yields of corn) and many other distributions.

- Second, Normal distributions are good approximations to the results of many kinds of chance outcomes, such as tossing a coin many times.
- Third, and most important, we will see that many statistical inference procedures based on Normal distributions work well for other roughly symmetric distributions.

Caution

. . . even though many sets of data follow a Normal distribution, many do not. Most income distributions, for example, are skewed to the right and so are not Normal. Non-Normal data, like non-Normal people, not only are common but are sometimes more interesting than their Normal counterparts.

Summary

The overall pattern of a distribution can often be described compactly by a density curve. A density curve has total area 1 underneath it. Areas under a density curve give proportions of observations for the distribution.
The mean μ (balance point), the median (equal-areas point), and the quartiles can be approximately located by eye on a density curve. The standard deviation σ cannot be located by eye on most density curves. The mean and median are equal for symmetric density curves, but the mean of a skewed curve is located farther toward the long tail than is the median.
The Normal distributions are described by bell-shaped, symmetric, unimodal density curves. The mean μ and standard deviation σ completely specify the Normal distribution $N(\mu, \sigma)$. The mean is the center of symmetry, and σ is the distance from μ to the change-of-curvature points on either side.
To standardize any observation x, subtract the mean of the distribution and then divide by the standard deviation. The resulting z-score $z = \frac{x-\mu}{\sigma}$ says how many standard deviations x lies from the distribution mean. All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the 68-95-99.7 rule.
If X has the $N(\mu, \sigma)$ distribution, then the standardized variable

$$Z = (X - \mu)/\sigma$$

has the standard Normal distribution $N(0, 1)$. Proportions for any Normal distribution can be calculated by software or from the standard Normal table

(Table A), which gives the cumulative proportions of $Z < z$ for many values of z.