**Chapter 5 *Exploring data*: Distributions-part II**

**Topics:**

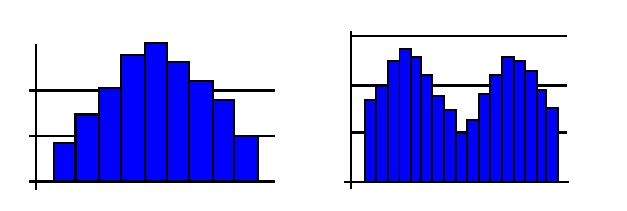
• The numerical descriptions of data spreading: quartiles

• The numerical summary of a data set: five-number summary

• A graphical presentation of data: boxplot

• The numerical descriptions of data spreading: range, variance and standard deviation

**1. Understand the spreading of a data distribution**

A measure of center alone is not sufficient to summarize data. The following two histograms show the annual household incomes in two areas. They have the same mean and the same median. Obviously, two distributions are quite different.****

**2. Quartiles *Q*1 and *Q*3**

The **first quartile** *Q*1 of a set of observations is defined as a point with 75% of the observations above it and 25% of the observations below it.

The **third quartile** *Q*3 of a set of observations is defined as a point with 25% of the observations above it and 75% of the observations below it.

The median is the second quartile, *M* = *Q*2.

**Question**: What percent of observations are between *Q*1 and *Q*3?

**To compute *Q*1 and *Q*3:**

**Step 1**. Arrange observations in an ascending order.

**Step 2**. Divide the ordered data set into two halves, lower half and upper half.

**Step 3**. Q1 is the median of the lower half of the observations in the ordered list.

**Step 4**. Q3 is the median of the upper half of the observations in the ordered list.

**3. Five-number Summary**

The following five statistics together are called **five-number summary:**

min, *Q*1,  M, *Q*3, max

Example 1. Data on the number of surgeries performed in a year by male doctors using the sample size 15.

27 50 33 25 86 25 85 31 37 44 20 36 59 34 28

Find the median and quartiles.

**Step 1**. Arrange the data in an ascending order:

20 25 25 27 28 31 33 34 36 37 44 50 59 85 86

**Step 2**. Find the median. M =

**Step 3**. Find the median of the data to the left of M. *Q*1 =

**Step 4**. Find the median of the data to the right of M. *Q*3 =

Write the Five-number summary:

Example 2. The GPA of 12 students who have applied for financial aid is shown below:

3.15 3.62 2.54 2.81 3.97 2.97 1.85 1.93 2.63 2.50 2.80 2.28

Find the quartiles.

**Step 1**. Arrange the data in an ascending order:

1.85 1.93 2.28 2.50 2.54 2.63 2.80 2.81 2.97 3.15 3.62 3.97

**Step 2**. Find the median. M =

**Step 3**. Find the median of the data to the left of M. *Q*1 =

**Step 4**. Find the median of the data to the right of M. *Q*3 =

Write the Five-number summary:

Practice: Find the Five-number summary for the numbers of surgeries performed by 10 female doctors in a hospital below:

25 7 10 14 19 18 5 29 33 31

**Questions:**

(a) What percent of observations are between Min and *Q*1?

(b) What percent of observations are between *Q*1 and M?

(c) What percent of observations are between M and *Q*3?

(d) What percent of observations are between *Q*3 and max?

**4. Boxplot** A **boxplot** is a graph of the five-number summary. A central box spans the quartiles, with a line marking the median. Whiskers extend out from the box to two extremes, the smallest and the largest observations.

Example 3. Given the data set for the surgeries performed in a year (from Example 1). Draw a box plot. We found the five-number summary in Example 1: 20 27 34 50 86



*Note: The boxplot can be either vertical, or horizontal as the boxplot above. (Error in boxplot on the right side, the whisker should reach 86.)*

**5. Reading and Comparing Box-plots** How to obtain information from a box-plot:

(1) If the median is near the center of the box, and two whiskers are about the same length, the distribution is approximately symmetric.

(2) If the median falls to the left of the center of the box, and the right whisker is longer than the left whiskers, the distribution is right skewed.

(3) If the median falls to the right of the center of the box, and the right whisker is short than the left whiskers, the distribution is left skewed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
| |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | |  | | | | | | | | | |  | |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | |  | | | | |  | |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | |  | | | | |

Example 4: The following side-by-side box-plots compare the number of surgeries performed by male and female doctors in a hospital during a year. Comment on the display.

The five-number summary for the number of surgeries performed by female doctors:

5 10 18.5 29 33

The five-number summary for the number of surgeries performed by male doctors:

20 27 34 50 86



*Note: error in box-plot on the right side, the whisker should reach 86.*

Questions:

(a) Is the distribution for the number of surgeries performed by male doctors symmetric? If not, which direction is it skewed?

(b) Is the distribution for the number of surgeries performed by female doctors symmetric? If not, which direction it is skewed?

(c) Which distribution has less spreading?

(d) Which distribution has wider center (middle 50%)?

(e) Find the approximate proportion of male doctors who perform less than 50 surgeries in the year.

(f) Find the approximate proportion of female doctors who perform 10 to 29 surgeries in the year.

Comparing Box-plots; How to compare two box-plots vertically:

(1) The distribution with higher median has higher center.

(2) The distribution with longer box and longer whiskers has more spreading.

**6. The Range, Standard Deviation and Variance**

**Range**- difference between the highest and lowest observation. It is much affected by outliers.

See previous examples.

The **variance** of a set of observations is an average of squares of the deviations of the observations from their mean, denoted by s2.

If are n observations, then

The **standard deviation** s is the square root of variance *s*2.s =

Example 6. Two experimental brands of outdoor paint were tested to see how they would last before fading. The results (in months) follow.

|  |  |  |
| --- | --- | --- |
| **Brand A** 10 60 50 30 40 20 |  |  |
| **Brand B** 35 45 30 40 35 25 |  |  |

(a) Draw a stemplot for each brand. Comment on plots.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Brand A | | |  |  | Brand B | | |
|  | 1 |  |  |  |  |  |  |  |
|  | 2 |  |  |  |  | 2 |  |  |
|  | 3 |  |  |  |  | 3 |  | |
|  | 4 |  |  |  |  | 4 |  | |
|  | 5 |  |  |  |  |  |  |  |
|  | 6 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

(b) Find the means.

(c) Find both standard deviations, and compare them.

|  |  |  |  |
| --- | --- | --- | --- |
| Brand A: |  |  | |
| Observations | Deviations | Squared Deviations |
|  |  |  |
| 10 | 10 - 35 = - 25 | (-25)2 = 625 |
| 60 | 60 - 35 = 25 | 252 = 625 |
| 50 | 50 - 35 = 15 | 152 = 225 |
| 30 | 30 - 35 = -5 | (-5)2 = 25 |
| 40 | 40 - 35 = 5 | 52 = 25 |
| 20 | 20 - 35 = -15 | (-15)2 = 225 |

Variance is *s*2 = 625+625+225+25+25+225/ 6-1 = Standard deviation is *s* = =

Brand B: Observations(*xi*) Deviations(*xi-*) Squared Deviations (*xi-*)2

Variance is *s*2 = Standard deviation is *s* = =

**When we summarize a data set:**

(1) *s* and come in pairs. Use **s** as a measure of spreading only when you use mean   as a measure of center.

(2) *s* is sensitive to a few extreme observations, while *Q*1 and *Q*3 are not.

(3) *s* has the same units of observation units, while units of variance are the squared units.

(4) *s* > 0 unless all observations have the same value.

(5) The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution.

(6) Use and **s** only for approximately symmetric distribution that is free of outliers.

**Chapter 5 *Normal Distribution-part III***

Many histograms are bell-shaped. For instance, the following histogram is generated from 200 randomly selected math quiz scores. It is approximately bell-shaped. 15 is the max score.



The vertical axis is labeled as “Frequency”.

1) What is the most common quiz score?

2) What is the statistical term for the most common score?

3) What do you expect the mean(average) of the quiz scores to be?

4) What do you expect the median of the quiz scores to be?

5) Describe the shape of the distribution. 6) What is the range of the quiz scores?

The same 200 scores are then converted to percent by the formula:

A new histogram is generated below that has the height of each bin as percent rather than

frequency.



The vertical axis is labeled as “Percent”.

What is the total area of all blocks?

Why does the shape of the histogram remain the same when the vertical scale is changed from frequency to percentage?

***A bell shaped curve is called a normal curve***, which has the following properties:

• The normal curve is symmetric.

• The center is mean *μ* (median = mean), read as “mu”

• The standard deviation is *σ*, read as “sigma”

– The standard deviation of a normal curve is the distance between the center and the change-of-curvature points (from open-upwards to open-downwards)



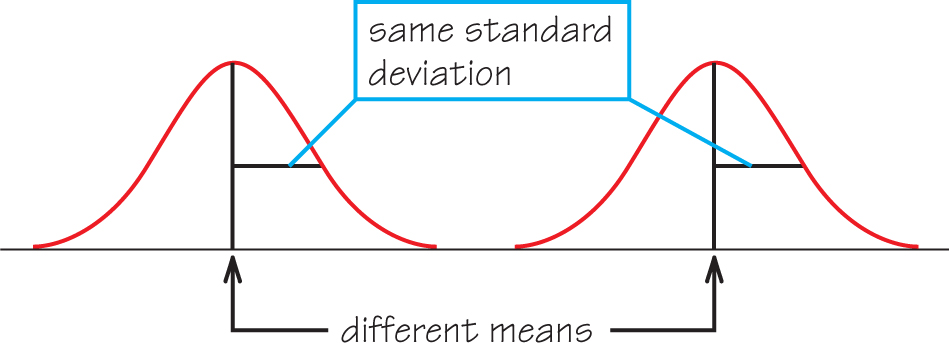
• The whole area under the curve is 1 (100%)

• The area under a normal curve above any interval of values gives the proportion of all values of the variable lie in the interval.

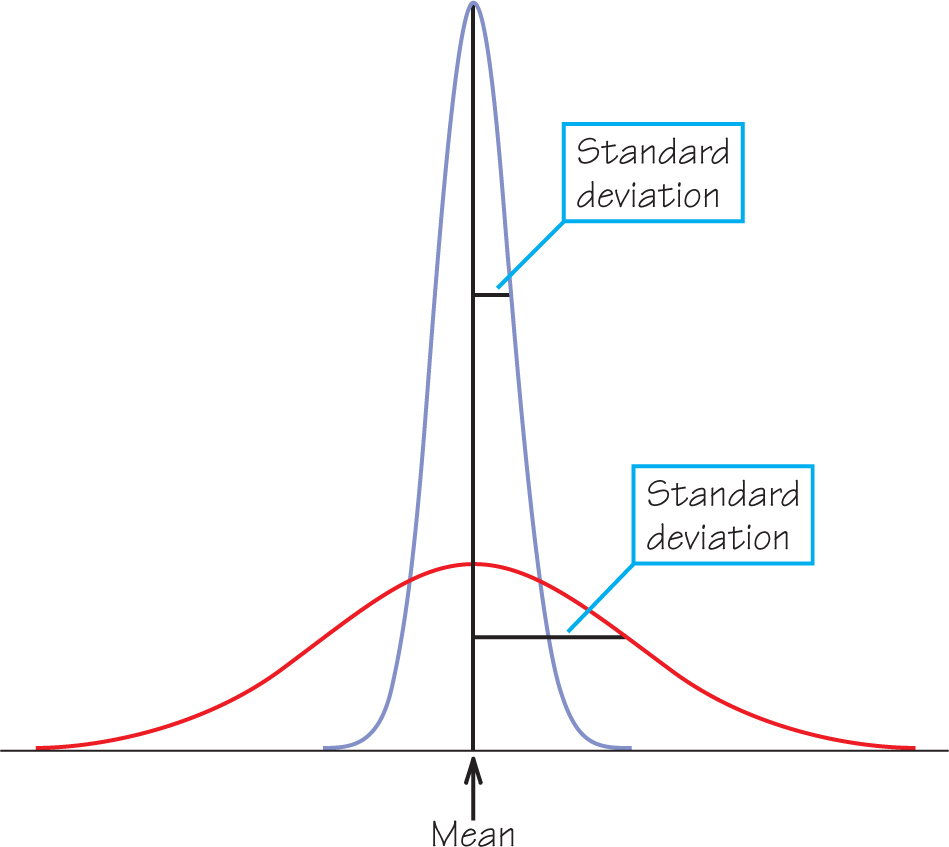
***A normal distribution is described by a normal curve***.

• If the histogram of a data set is bell-shaped, we say that the data follows a normal distribution

• The only two possible differences between two normal curves are

– different means (location of the center, *μ*) 

– different standard deviations or variance (spread, *σ* )



**Quartiles of any normal distribution**

• Given a normal distribution with mean μ and standard deviation *σ*, we have the following formulas to compute the first and the third quartiles:

*Q*1 = μ − 0.67 *σ*.

*Q*3 = μ + 0.67 *σ*.

Example 1. The distribution of certain type of IQ scores is approximately normal with mean *μ* = 100 points and SD *σ* = 20 points. Find the quartiles. Draw and label the normal curve.

*Q*1 =

*Q*3 =

Example 2. The distribution of heights of young women is approximately normal with mean *μ* = 64.5 inches and SD *σ* = 2.5 inches. Draw and label the normal curve.

(a)Find the quartiles.

(b) What is the median height for young women?

(c) How tall must a young woman be to fall in the top 25%?

(d) Find the range of the middle 50% heights of young women.

**68-95-99.7 Rule** For any normal distribution,

• 68% of the observations fall within 1 standard deviation of the mean.

For the graphs below, the mean is 0 and standard deviation is 1.

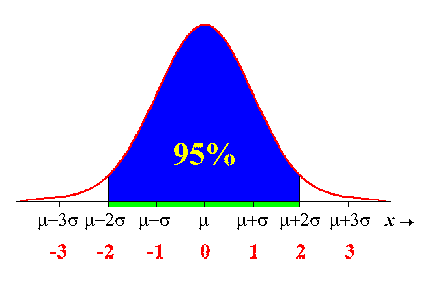


Example 3. The distribution of math quiz scores of a large class is approximately normal with mean 10 and standard deviation 2. Draw and label the normal curve.

(a) What percent of scores falls between 8 and 12 points?

(b) If 500 students are sampled, how many would score between 8 and 12 points?

(c) What percent of scores is higher than 12 points?

• 95% of the observations fall within 2 standard deviations of the mean. 

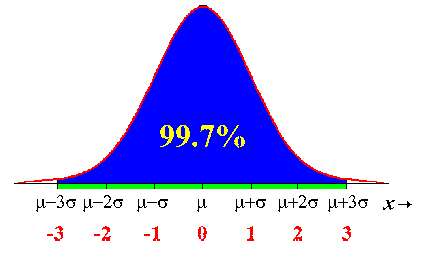
Example 3. (Continuing)

(d) What percent of scores falls between 6 and 14 points?

(e) What percent of scores is lower than 6 points?

(f) If 1000 students took the quiz, how many students scored lower than 14 points?

• 99.7% of the observations fall within 3 standard deviations of the mean.



(g) What percent of scores falls between 4 and 16 points?

(h) What percent of scores is greater than 16?

Workout 1. The distribution of heights of young women is approximately normal with mean *μ* = 64.5 inches and SD *σ* = 2.5 inches. Draw and label a normal curve for each question.

(a) What percent of young women are between 57 inches and 72 inches tall?

(b) What percent of women are taller than 67 inches?

(c) What percent of women are shorter than 64.5 inches?

(d) What is the range of heights for the middle 68% of young women?

(e) In order for a women to be women to be in the top 2.5% she must be taller than \_\_\_\_\_\_.

Workout 2. The distribution of SAT scores in the 1960 and 70’s were approximately normal, with a mean *μ* of 500 and a standard deviation *σ* of 100. Draw and label a normal curve for question b-e.

(a) What is the median SAT score?

(b) How high must a person score to fall in the top 25%?

(c) What percent of scores fall between 300 and 700?

(d) What percent of scores are above 600?

(e) Almost all (99.7%) SAT scores fall in what range?