Course Information

- Schedule and Homework
- Course Requirements
- Instructor's Information
- Links
- Main Page

Topics

1. Introduction
2. Summarizing Data: listing and grouping
3. Summarizing Data: Statistical Descriptions
4. Possibilities and Probabilities
5. Some rules of probability
6. Probability Distributions
7. Normal Distributions
8. Sampling Distributions
9. Problems of Estimation
10. Hypothesis Tests
11. Tests based on count data
12. Regression
13. Non-parametric Tests

Projects

1. Stock return distribution

- Forum

edit sec src prt

Site Manager

Chapter 3. Summaring Data: Statistical descriptions

Besides being the subject matter of the course, the word *statistics* is also the plural of the word **statistic**, which is a quantity computed from sample data, while a **parameter** is a quantity computed from the data from the whole population.
A statistical description is a synonym of statistic.

Statistical descriptions are usually classified by the features of the sample data that they are trying to describe. The most common ones are *measures of location* or center, which are indicative of the 'center' of the data, while the *measures of variation* are indicative of the variability of the data.

Measures of Location: The mean

Sample mean

The sample mean is obtained by adding all the values in your sample and dividing by the sample size (which is usually denoted by small *n*). In mathematical notation, we have

$$\bar{x} = \frac{\sum x}{n} \qquad (1)$$

Notice that the symbol for the sample mean is $\bar{x}$, an *x* with a bar above, and it is read `x bar`. The symbol
Σ is the sum sign in mathematics, which means that you should add up all the values in your sample.

Say, for example that you collected the age of 4 students in the class to estimate

the average age of the whole class. Then the 4 students are the *sample*, and the whole class the *population*.

The population size is $n = 4$.

If the ages you collected are 22, 21, 48, and 21, then the sample values are written as

$$x_1 = 22,\ x_2 = 21,\ x_3 = 48,\ x_4 = 21. \tag{2}$$

and for this particular example, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{22 + 21 + 48 + 21}{4} = \frac{112}{4} = 28 \tag{3}$$

Population mean

If we were to calculate the mean from a population instead of a sample, then we would still proceed in the same way, we would add up all the values in the population (more values to add) and we would divide by the population size (denoted by *N*).

The symbol for the population mean is the Greek letter mu: μ, so we obtain

$$\mu = \frac{\sum x}{N} \tag{4}$$

In the case of the population of students in the classroom with the following ages: 21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28, the population mean is 26.2.

In this case, it is understood that the $\sum$ (Sigma)- sign indicates the sum over all the elements of the population (not only the sample), therefore, in this case, the sum indicates
$x_1 + x_2 + x_3 + \ldots + x_N$, where in this case N= 20, instead of $x_1 + x_2 + \ldots + x_n$ in the case of the sample mean, where n=4 in our example above. When we want to make this difference more evident, then we write $\sum_{i=1}^{N}$ for the population, and $\sum_{i=1}^{n}$ for the sample.

Properties of the mean

The mean as a measure of center is probably the most frequently used measure of center, partly because it has the following properties:

1. it can be calculated for any numerical data set.
2. it's value is unambiguous and unique for a given data set
3. it lends itself to further statistical treatment
4. if each value in the sample would be replaced by the mean, $\sum x$ would remain unchanged
5. it takes into account every value in the data set
6. it is relatively reliable (does not fluctuate widely when selecting different samples)

However, **outliers** have a strong effect on the mean, particularly if the sample size is small. Therefore, at times the **trimmed mean** is used instead, in which case the upper and lower 5% is deleted, and the mean is taken without those values.

In our population above, we would *trim* the data set 58 and the data set 20, and then add the values and divide by 18, to obtain 23.3.

Measures of Location: The weighted mean

When taking averages, it is often important to take into account that data have different *weights*; some data points are more "important" than others. For example, if we have the household income per state, we see that Alaska has the 4th highest median household income (64,333) , while New York has the 19th

highest (53,514).
If we want to calculate the national average, however, New York's value counts much more than Alaska's because New York's population is much larger than Alaska's (19,490,297 *vs.* 686,293).

Therefore, to get the correct mean, we would have to weight each state's household income with the appropriate weight, which is a measure of relative importance (in this case the population size).

The trimmed mean is given by

$$x_w = \frac{\sum w \cdot x}{\sum w} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + x_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \quad (5)$$

For example, suppose that we have a sample with the 2 states named above and Texas ($47,548) and Mississippi ($36,338), with corresponding populations 24,326,974 and 2,938,618 respectively.

Then, the weighted mean is

$$x_w = \frac{19490297 \cdot 53514 + 686,293 \cdot 64333 + 24326974 \cdot 47548 + 2938618 \cdot 36338}{19,490,297 + 686,293 + 24,326,974 + 2,938,618} \quad (6)$$

which gives a weighted average of $49, 547. The national median household income is $50,740.

Grand mean

A special case of the formula for the weighted average is the **grand mean**, which is the overall mean. In that case, it has a special notation and slightly differenct formula:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \cdot \bar{x_i}}{\sum_{i=1}^{k} n_i} = \frac{n_1 \bar{x_1} + n_2 \bar{x_2} + \cdots + n_k \bar{x_n}}{n_1 + n_2 + \cdots + n_k} \quad (7)$$

Measures of Location: The Median and other fractiles

The median is a measure of center, like the mean, which is not affected by outliers like the mean is. The symbol used for the sample median is $\tilde{x}$ and for the population median $\tilde{\mu}$.

To obtain the median, we first need to re-arrange the data in ascending order (sorted), and then find the middle value, namely,

- *when n is odd, the median is the value in the middle (after sorting)*
- *when n is even, it is the mean of the two items nearest to the middle*

For example, suppose that we select a sample of size 4 from the population of students in the class listed above, obtaining the following ages: 20, 26, 22, 46.

Then, since *n=4*, the median is the average of the two values in the middle (after sorting, hence the average of 22 and 26, or $\tilde{x}$=24.

If we added another value to the sample, say 28, then the sorted sample would be 20, 22, 26, 28, and 46, and hence the median would be $\tilde{x}$=26.

When dealing with a small population, or larger sample, it is sometimes convenient to make a stem-and-leaf plot to find the median. The reason for this is that the stem-and-leaf plot **sorts the data**.

For example, in the case of the GDP per capita of the 20 countries listed in chapter 2, we got the following stem and leaf plot:

```
The decimal point is 1 digit(s) to the right of the |

3 | 99999
4 | 02445669
5 | 5567
6 |
7 | 06
8 | 1
```

Since *n=20*, we need to find the average between the values in the 10th and 11th position, namely 45 and 46 thousand USD. Therefore, the median GDP per capita, for the 20 richest countries (per capita) is 45,500 USD.

Fractiles

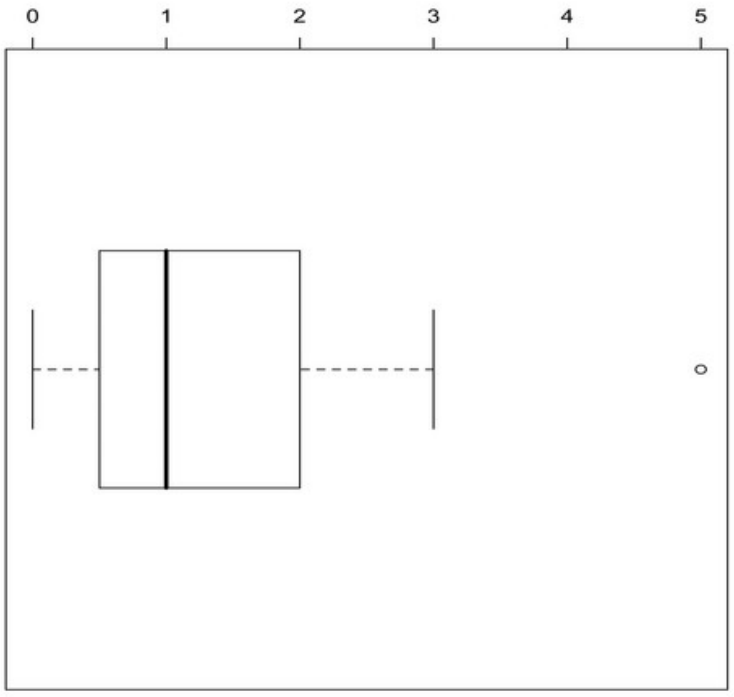
The median is one example of a fractile, a measure (value) that divides the data set into two or more equal parts. Another such example are quartiles, that divide the data into 4 equal parts, obtaining the values *Q1, Q2,* and *Q3*. Q4 is the maximum value, so no need to calculate it. Namely, Q1 is the value such that 25% of the data fall below it, and Q3 is such that 75% of the data fall below it, while Q2 is the median, and hence 50% of the data values fall below.

In the list above, there is 5 countries whose GDP per capita is 39,000 and the next one has a GDP per capita of 40,000, hence, Q1=39,500. Similarly, Q2 = 45,500 = $\tilde{x}$ and Q3 = 55,500.

Boxplot

A box plot is a graph that summarizes the data by representing 5 values, the minimum and maximum value, the Q1, Q2 and Q3.

In the graph below, the 3 central horizontal lines represent Q1, Q2 and Q3, while the point on the extreme represents an outlier value. This is a variation of the boxplot as described in the previous line, since the other two horizontal lines cannot be the minimum and maximum. They are the 5th and 95th percentile.



Measures of Location: The mode

The mode is a measure of center that is usually used for data that is non-numerical. Namely, the mode is the *value that occurs most frequently*, and it is the only value that can be collected for qualitative data.

For example, in example 3.17, they list the size of dresses sold by a store to be 10,7, 14, 9, 9, 14, 18, 9, 11, 12, 16, 14, 9, 14, 14, 11, 9 and 20. In this case, the number 9 and 14 appear most frequently, both showing exactly 5 times. Therefore, you would say in this case that this data is bimodal (has two modes), namely 9 and 14.

Measures of Variation: The range

The range is a measure of variation or variability of the data. *The range of a data set is the largest value minus the smallest value*.

For example, for the age distribution in the class, which we write here again for convenience:
21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28,

the range is 58-20 = 38 years.

One disadvantage of the range is that it is highly influenced by outliers. For that reason, we use the following measure of variation more often.

Measures of Variation: The standard deviation

The standard deviation is the most general measure of variation.

To calculate the standard deviation of a population, one first takes the difference of each data point to the mean (the variation), and squares that difference (to insure it's positive). Then, all those squared differences
are added together and divided by *N*, the size of the population. Finally, the square root is taken. This is summarized in the following formula:

Population standard deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{8}$$

The square of this quantity is called the **population variance**, and even though it is not a measure of variation *per se*, it is a notion that we will widely use during the course.

Population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \tag{9}$$

Population example

Suppose that we had 20 students in the classroom with the following ages:
21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28 23.

The mean μ=26 in that case.

Then, one way to calculate the standard deviation of the population would be to do the following table:

x	**x-μ**	**(x- μ)^2**
21	21 - 26=-5	25
21	21 - 26=-5	25
25	25 - 26=-1	1
20	20 - 26=-6	36

and so on, until the last two rows:

28	28 - 26=2	4

23	23 - 26=-3	9
sum	**=**	**1678**

Therefore, the variance of the population is σ^2=1678/20=83.9, and the standard deviation is the square root of this result, namely σ=9.16.

Sample standard deviation

The sample standard deviation is calculated almost in the same way as the population standard deviation, but substituting the sample mean $\bar{x}$ for the population mean σ and the population size *N* for the **sample size minus 1: *n-1***.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \tag{10}$$

The square of this quantity is called the **sample variance**.

Sample variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \tag{11}$$

Sample example

Suppose that we take a sample of 4 students out of the population listed above, say the ones with ages 26 22 46 25.

Then the sample standard deviation can be calculated by though the table:

x	**x-** $\bar{x}$	**(x-** $\bar{x}$**)^2**
27	27 - 30=-3	9
22	22 - 30=-8	64
46	46 - 30=16	256
25	25 - 30=-5	5
sum	**=**	**354**

This total we would have to divide by *n-1=3* to obtain the sample variance s^2=118.
The sample standard deviation is then *s*=10.86.

"Fast" formula

An alternative but equivalent formula for calculating the sample standard deviation is

$$s = \sqrt{\frac{S_{xx}}{n - 1}} \tag{12}$$

where

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}\,. \tag{13}$$

The advantage of this formula is that you need one less column to calculate the sample standard deviation, which could lead to faster calculations, specially when a lot of data is involved. For the example above, we obtain:

x	**x^2**

27	729
22	484
46	2116
25	625
$\sum x$**=120**	$\sum x$**=3954**

Therefore,

$S_{xx}=3954-(120)^2/4=354$

Dividing this by $n-1=3$, we obtain 118. This is again the value of the sample variance, and therefore we obtain the same value for the sample standard deviation, namely, $s=10.86$.

Application of the standard deviation

Chebyshev's theorem

> For any set of data and any constant k greater than one, at least $1-1/k^2$ of the data must lie within k standard deviations on either side of the mean.

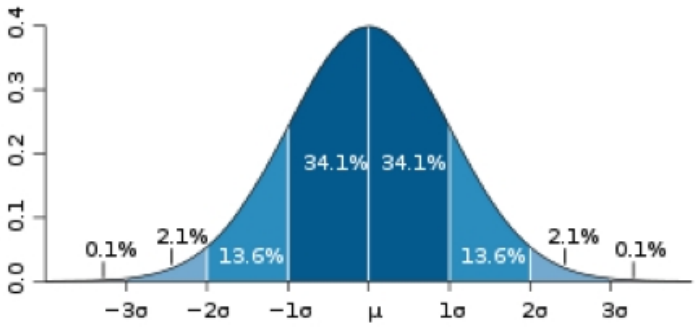
This theorem gives us a broad bound of how much data should be inside the mean plus or minus k standard deviations.

For example, when $k=3$, it is telling us that at least

$1-1/(3^2)=1-1/9=8/9$ or approximately 89%

of the data must lie within 3 standard deviations from the mean.

We will see in this course that many data sets follow a *normal distribution*, which is a bell-shaped distribution like the one in the graph below:



For the normal distribution, Chebyshev's theorem applies, but there is actually a more precise **empirical rule**:

> 1. About 68% of all the data values lie within 1 standard deviation from the mean
> 2. About 95% of all the data values lie within 2 standard deviation from the mean
> 3. About 99.7% of all the data values lie within 3 standard deviation from the mean

For example, if the height of women in the US is normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches, this implies that 95% of all women in the US are between 59 and 69 inches tall.

z-scores

The z-score is a measure of relative value which is useful to compare values from different data sets, it is calculated using the following formula in the case of a

population:

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

and in the case of the sample by

$$z = \frac{x - \bar{x}}{s}. \tag{15}$$

For example, Rebecca Lobo, a female basketball player, is 76 inches tall, therefore here *z*-score is

$$z = \frac{76 - 64}{2.5} = 4.8, \tag{16}$$

using the values for the population mean and standard deviation given above.

In other words, she is 4.8 standard deviations higher than the average US woman. That is extraordinarily tall!

Coefficient of variation

A measure that allows us to compare variation from different data sets is the **coefficient of variation**, given by

$$V = \frac{s}{\bar{x}} \cdot 100\% \text{ or } V = \frac{\sigma}{\mu} \cdot 100\%$$

For example, if one statistics class averaged 75 with a standard deviation of 10 points and another one averaged 65 with a standard deviation of 8 points, the coefficient of variation would allow us to find which class has less variability (is more homogeneous).

The coefficients of variation are $\frac{10}{75}\ 100\% = 13.3\%$ and $\frac{8}{65}\ 100\% = 12.3\%$, so that the second class is more homogeneous (or consistent).

Grouped data

When we create or receive a frequency distribution, the data has been grouped already, and therefore we have lost some of the original information.

For example, in the last section we saw the frequency distribution of GDP per capita of the 20 richest countries:

GDP range	Number of countries
80,000 -89,999	1
70,000-79,999	2
60,000-69,999	4
50,000-59,999	0
40,000-49,999	8
30,000-39,999	5
Total	**20**

Even though we have lost some information, we can still calculate and approximate mean and standard deviation. Namely, let $f_1, f_2, \ldots f_k$ be the class frequencies, and let $x_1, x_2, \ldots x_k$ be the midpoints of every class, then we can approximate the mean by

$$\bar{x} = \frac{\sum x_k f_k}{n}$$

and the standard deviation using the "fast formula" by

$$s = \sqrt{\frac{S_{xx}}{n-1}} \text{ where } S_{xx} = \sum x^2 f - \frac{(\sum xf)^2}{n}$$

Extending the table above, we obtain

GDP range	f	x	xf	x^2f
80,000 -89,999	1	85K	85	7225
70,000-79,999	2	75K	150	11250
60,000-69,999	4	65K	260	16900
50,000-59,999	0	55K	0	0
40,000-49,999	8	45K	350	16200
30,000-39,999	5	35K	175	6125
Total	**20**		**1030**	**57700**

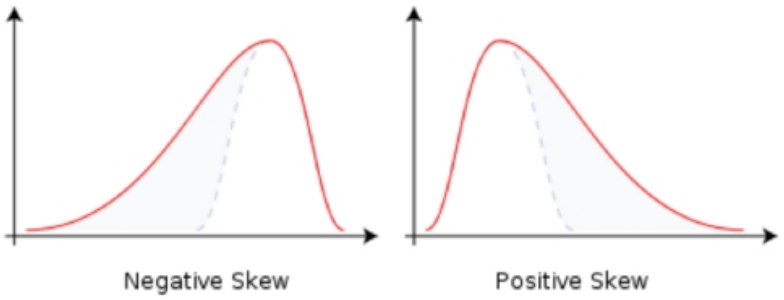
Therefore,

$$\bar{x} = \frac{1030}{20} = \$51.5K$$
$$S_{xx} = 57,700 - (1030)^2/20 = 4655$$

Therefore the variance from grouped data is 4655/19=245.52 and the standard deviation is 15.67.

Skewness

When the data in a distribution is tilted to the left of center or to the right of center, then you say that it is *skewed*, namely *skewed to the left (positive skew)* or *skewed to the right (negative skew)*, respectively, as can be seen in the image below.



If the data is not skewed either way, then we call it symmetric, like in the case of the normal distribution depicted above.

http://statistics.wikidot.com/local--files/ch3/LifeExp.xls

http://statistics.wikidot.com/local--files/ch3/LifeExp2.xls