Introduction

We usually care about two things of a given data set. Where is the center of data, and the spread of data. These two would give us a rough idea of a given data set.

Furthermore, the ultimate goal of statistics is make inference about the population. For example, if I want to know the average height of all UW-Madison students, but with limited time and money, I am unable to ask all the students, then we ask — can I use the average height of randomly-chosen 2,000 UW-Madison students to guess? How precise can this guess be?

In this chapter, important concepts:

- sample mean/ average
- standard deviation
- population, parameter, sample, statistic
- Expected value
- standard error
- law of large number

Prompt:
I want to know the average height of all UW-Madison students, but I can only ask random 2,000 students' heights

The first thing I consider is the numerical summaries of my data — the heights of the random 2,000 students.

- The sample average / sample mean

  The average of the data is the sum of all the numbers divided by how many numbers there are.

- The standard deviation (SD)

  tells us how far a typical data point is away from the mean of data. A larger SD means the points are usually further away from the average.

  Steps to compute SD:

  Calculate the average → subtract the average from all the data points to find the deviation → square all the deviations → add all squared deviation → ask if you have all the data from the population, or you just have the sample, which is a subset of population. If you have the population, divide the number of data points, if not divide by number of data points minus 1. → take the square root of previous result.

※: If you like math notation, the above steps are the same as $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ $\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$,

for x_i, $i=1.2.3\_\_\_n$ are the data points.

Exercises with solutions for average and SD.

- For each of the list below, work out the average and SD.

(a) 1, 3, 4, 5, 7

Solution:

average is $\frac{1+3+4+5+7}{5} = 4$

SD is, we make a table

	1	3	4	5	7
subtract	-3	-1	0	1	3
square	9	1	0	1	9

add $9+1+0+1+9 = 20$

assuming this is population, then we divide by 5, (if this is sample, divide by 4.).

$\frac{20}{5} = 4$. and $\sqrt{4} = 2$.

Therefore, SD is 2

(b) 6, 8, 9, 10, 12

Solution: the average is $\frac{6+8+9+10+12}{5} = 9$

	6	8	9	10	12
subtract mean	-3	-1	0	1	3
square	9	1	0	1	9, $1+9+0+1+9=20$

$20/5 = 4$ $\sqrt{4} = 2$, SD is 2

(c) How is list (a) related to list (b)? Why do they share the same SD?

Solution:

list (ii) is created by adding 5 to each data point in list (i). This shows, since adding a number to each element would not influence their spread, SD would be the same.

(d) calculate the mean and SD for the list: 3, 9, 12, 15, 21, how does this list related to list (a)?

Solution:

the average is $\frac{3+9+12+15+21}{5} = 12$

SD table	3	9	12	15	21
subtract the mean	-9	-3	0	3	9
square	81	9	0	9	81

$81+9+0+9+81 = 180$

$\frac{180}{5} = 36$, $\sqrt{36} = 6$, therefore, the SD of list (d) is 6. Notice that the elements in list (d) is created by multiplying elements in list (a) by 3, thus we notice, the SD is also multiplied by 3.

Hint: If you add the same number to a list, SD would not change, If you multiply a number to the list, the SD would expand/ shrink by the absolute sign of the number.

■ Find the mean and standard deviation for the following scores (sample formula)

92, 95, 85, 80, 75, 50

Solution:

the mean is $\frac{92+95+85+80+75+50}{6} = 79.5$

SD:

	92	95	85	80	75	50
subtract the mean	12.5	15.5	5.5	0.5	-4.5	-29.5
square	156.25	240.25	30.25	0.25	20.25	870.25

add together : $156.25 + 240.25 + 30.25 + 0.25 + 20.25 + 870.25 = 1317.5$

By sample formula, we need to divide by $(6-1)=5$

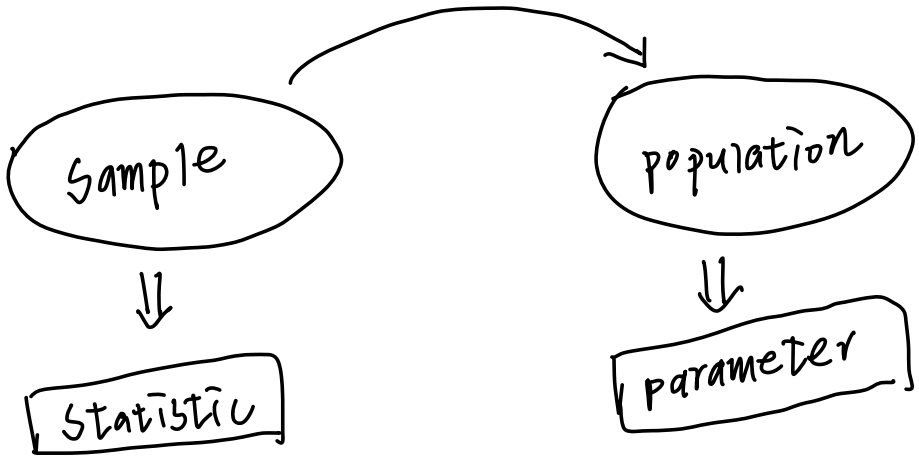
$\frac{1317.5}{5} = 263.5$

and take square root, $\sqrt{263.5} = 16.2327$

thefore, the SD, under sample formula, is 16.2327

■ Population, parameter, sample, statistic

the ultimate goal of statistics: Inference of population



Definitions

- **Population**

A population is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired.

A population is the collection of all subjects of interest.

- **Parameter**

A parameter is any summary number, like an average or percentage, that describes the entire population.

A parameter is a descriptive measure for an entire population.

- Sample

A sample is a representative group drawn from the population.

A sample is a subset of the population of interest.

- Statistic

A statistic is any summary number, like an average or percentage, that describes the sample.

A descriptive measure for a sample is a statistic.

In our prompt, the population is all UW-Madison students; the parameter is the average of their heights; the sample is the randomly-chosen 2,000 students; statistic is the average height of randomly-chosen 2,000 students. The reason why we distinguish these concepts is, the population is fixed — if no one drops out overnight, all UW-Madison students are the same group of people, and their average height is fixed. However, our sample might be different, for example, the randomly-chosen 2,000 students would be different depending on how I draw, as a result, the sample average height of randomly-chosen students might be different than the average height of UW-Madison students.

Examples of Population, parameter, sample, and statistic.

- Of all the U.S. adults, 36% has an allergy. A sample of 1,200 randomly-selected adults resulted in 33.2% reporting an allergy.

  population: U.S. adults

  sample: 1,200 randomly-chosen adults

  parameter: allergy ratio, 36%

  statistic: allergy ratio of the sample, 33.2%

- Why is the parameter fixed and the statistic varies?

  The parameter comes from the population, and the population is set. The statistic comes from the sample, and different samples can be taken from the population. Therefore, the observed / sample statistic might be different from parameter. We call the difference between parameter and statistic chance error.

- An experiment was conducted to learn more about the percentage of potato plants that exhibit signs of a disease. A 10-acre field of potatoes was planted. Researchers enumerated all the plants, then sampled 300 of them at random. 17% of them showed signs of disease.

  The population is all the plants in the field. The sample is the 300 potatoes that were randomly chosen. The parameter is the percentage of potato plants showing disease. The statistic, which is the percentage of potatoes got the disease in the sample, is 17%.

Expeuted Value

Recall our prompt is we want to understand the average height of all UW-Madison students. Which is a "parameter" we want to infer about the population.

Therefore, we define Expeuted value, which can be thought as population average, is defined as follows.

- The expeutation, or expeuted value of a random variable, is a weighted average of the possible values that the random variable can take on, being weighted by the probability, usually we denote expeuted value of a random variable X as $E(X)$.

For example, assume a very stupid and unreasonable case that 60% of the UW-Madison students is 5'2" tall (157 cm), and 40% of the UW-Madison students is 5'10" tall (177 cm)

assume that we have 10,000 students. Then the population average height is

$$\frac{157 \times 60\% \times 10{,}000 + 177 \times 40\% \times 10{,}000}{10{,}000}$$

$$= 157 \times 60\% + 177 \times 40\% = 165 \text{ cm}.$$

Therefore, the expeuted value of the height of UW-Madison students is 165 cm.

In other words, we want to use the average height of 2,000 students to guess the average height of all UW-Madison students.

Standard Error

The standard error of a statistic is the standard deviation of the sampling population. In other words, we're looking for a measure to tell us how big our chance error would typically be, when we use a statistic to guess a parameter.
For example, we use sample mean to guess the expected value/population mean, thus, standard error of the sample mean is an estimate of how far

the sample mean is likely to be from the population mean.

If the variable has higher spread, then it is more possible for us to have higher chance error/standard error. Thus, to be able to estimate standard error of the sample sum or sample mean, we should first define the following.

Variance

the variance of a random variable X, by the previous notation, is $E[X-E(X)]^2$

that is, you first compute the expectation of X, and compute the expectation of $[X-E(X)]^2$.

population standard deviation.

The standard deviation of a random variable X is the positive square root of variance of X.

that is, $\sqrt{E[X-E(X)]^2}$, usually denoted as σ.

Important relationships among the population standard deviation, standard error of the average, and standard error of the sum.

- The standard error of the average of n draws / sample size n is

$$se = \frac{\sigma}{\sqrt{n}}.$$

Recall that we usually use sample average to guess the expected value.

The standard error of the average is the standard deviation among the averages obtained in infinitely many samples of size n. The standard error of the average is the likely size of the difference between the sample average and the expected value.

- The standard error of the sum of n draws / sample size n. is

$$\text{s.e.} = \sqrt{n}\,\sigma$$

Analogously to standard error of the average, the standard error of the sum is the likely size of the difference between the sample sum and expected sum, which is n times the expected value of an experiment.

→ Is it possible for me to reduce chance error?
yes, by increasing the sample size.

- The law of large number

If you repeat an experiment independently a large number of times and average the results, the sample mean would be close to the expected value.

Notice that the standard error of the average is $\sigma/\sqrt{n}$, σ is fixed for a given population, $\sigma/\sqrt{n}$ is a decreasing function in n.

Notice that the statement of the law of large number doesn't take about the shape / distribution of sample mean, students should pay attention to the difference between law of large number and central limit theorem.

Exercises with Solutions

- Find the expected value for the sum of 100 draws at random with replacement from the box —

(a) | 0 | 1 | 1 | 6 |

Solution:

Recall that expected value is the sum of possible outcomes weighted by the probability.

For a single draw, the possible outcomes and probabilities are:

outcome	0	1	6
probability	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

thus, the expected value for a single draw is $0 \times \frac{1}{4} + 1 \times \frac{2}{4} + 6 \times \frac{1}{4} = 2$

and then expected value of 100 draws is

$2 \times 100 = 200$.

(b) | -2 | -1 | 0 | 2 |

Solution:

the expected value for a single draw is:

$(-2) \times \frac{1}{4} + (-1) \times \frac{1}{4} + 0 \times \frac{1}{4} + 2 \times \frac{1}{4} = \frac{-1}{4}$

Then the expected value for 100 draws is

$\frac{-1}{4} \times 100 = -25$

- Find the expected value of the outcome of rolling a die

Solution:

possible outcome	1	2	3	4	5	6
probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

then the expectation is

$$1\times\tfrac{1}{6}+2\times\tfrac{1}{6}+3\times\tfrac{1}{6}+4\times\tfrac{1}{6}+5\times\tfrac{1}{6}+6\times\tfrac{1}{6}=\frac{7}{2}$$

- A school class of 120 students is driven in 3 buses to a symphonic performance. There are 36 students in one of the buses, 40 in another, and 44 in the third bus. When the buses arrive, one of the 120 students is randomly chosen. Find the expected value of the number of students on the bus of that randomly-chosen student.

Solution:

Since the randomly-chosen student is equally likely to be any of the 120 students, then we have $\frac{36}{120}$ probability to choose student for the bus with 36 students.

The expected value is then

$$36\times\frac{36}{120}+40\times\frac{40}{120}+44\times\frac{44}{120}=\frac{1208}{30}=40.2667.$$

- Let X denote a random variable that takes on any of the values -1, 0, and 1, with probability 0.2, 0.5, and 0.3, compute the expected value of its square.

Solution:

$P(X=-1)=0.2$, $P(X=0)=0.5$, $P(X=1)=0.3$

The question asks for $E(X^2)$

Let $Y=X^2$, then

$P(Y=1)=P(X=-1)+P(X=1)=0.5$

$P(Y=0)=P(X=0)=0.5$

Hence, $E(X^2)=E(Y)=1\times 0.5+0\times 0.5=0.5$

■ Calculate the population standard deviation of the outcome of rolling a die.

Solution:

Recall that $SD(X) = \sqrt{E[X - E(X)]^2}$

In this case, the X represents the outcomes of rolling a die.

Make the table.

possible outcome $X = ?$	1	2	3	4	5	6
probability $P(X = ?)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
subtract the $E(X)$: $X - E(X)$	-2.5	-1.5	-0.5	0.5	1.5	2.5
square $[X - E(X)]^2$	6.25	2.25	0.25	0.25	2.25	6.25
probability of $[X - E(X)]^2 = ??$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$E[X - E(X)]^2$ $6.25 \times \frac{1}{6} + 2.25 \times \frac{1}{6} + 0.25 \times \frac{1}{6} + 0.25 \times \frac{1}{6}$

$+ 2.25 \times \frac{1}{6} + 6.25 \times \frac{1}{6} = 2.9166$

take square root $\sqrt{2.9166} = 1.7078$

Thus, the standard deviation is 1.7078

■ One hundred draws are going to be made at random with replacement from the box [1 2 3 4 5 6 7]

Find the expected value and standard error for the sum.

Solution:-

the expected value of a single draw is

$1\times\frac{1}{7}+2\times\frac{1}{7}+3\times\frac{1}{7}+4\times\frac{1}{7}+5\times\frac{1}{7}+6\times\frac{1}{7}+7\times\frac{1}{7}=4$

thus,

the expected value of 100 draws is

$4\times100=400.$

the population standard deviation is, let X be the value of a draw

possible X=?	1	2	3	4	5	6	7
probability X=?	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$
subtract E(X)	-3	-2	-1	0	1	2	3
square $[X-E(X)]^2$	9	4	1	0	1	4	9

$E[X-E(X)]^2$ $9\times\frac{1}{7}+4\times\frac{1}{7}+1\times\frac{1}{7}+0\times\frac{1}{7}+1\times\frac{1}{7}$

$+4\times\frac{1}{7}+9\times\frac{1}{7}=4$

square root, $\sqrt{4}=2$

SD for a single draw is 2, then

i.e. for the sum is $\sqrt{100}\times2=20$

standard error for the sum is 20

- You gamble 100 times on the toss of a coin. If it lands heads, you win $1. If it lands tails, you lose $1. Your net gain will be around ?, give or take ??

Solution:

First compute the Expected value.

For a single toss, your expected gain is $1 \times 0.5 + (-1) \times 0.5 = 0$.

thus, for tossing 100 times, you expected gain is $0 \times 100 = 0$

the standard error of the sum ??

First compute population standard deviation

* for a single draw

possible outcome X	-1	1
subtract $E(X)$	-1	1
square $[X-E(X)]^2$	1	1

$E(X-E(X))^2 = 1 \times 0.5 + 1 \times 0.5 = 1$

thus, the SD for a single draw is 1,

then s.e. for the sum is $\sqrt{100} \times 1 = 10$.

Your net gain will be around $0, give or take $10.

■ A box model contains eight tickets marked 1, twenty marked 6, and twelve marked 8.

(a) What is the expected value of the sum of the 150 tickets drawn? What is the expected value of the average of the 150 tickets drawn?

Solution:

You have 40 tickets in total. Each ticket is equally likely to be drawn, Thus, the chance to get a ticket marked 1 is $\frac{8}{40}$. Therefore, the expected value for a single ticket is

$1 \times \frac{8}{40} + 6 \times \frac{20}{40} + 8 \times 12/40 = 224/40 = 5.6$

thus, the expected sum from 150 draws is $150 \times 5.6 = 840$. The expected average is 5.6

(b) What is the standard error of the sum of 150 tickets? What is the standard error of the average of the 150 tickets?

Solution:

First compute the standard deviation of a single draw.

possible outcome	1	6	8
probability	$\frac{8}{40}$	$\frac{20}{40}$	$\frac{12}{40}$
subtract the expected value	(-4.6)	0.4	2.4
square	21.16	0.16	5.76

$E[X-E(X)]^2 = \frac{8}{40} \times 21.16 + \frac{20}{40} \times 0.16 + \frac{12}{40} \times 5.76 = 6.04$

square root $= \sqrt{6.04} = 2.4576$

then s.e. for the sum is $\sqrt{150} \times 2.4576 = 30.1$

the s.e. for the average is $\frac{\sqrt{6.04}}{\sqrt{150}} = 0.2$