Quantifying Distributions

Quantifying Distributions

Measuring the "Center"

Measuring Variability

Central Tendency

A measure of *central tendency* is a number that represents the *middle* of a distribution

Typical Measures of Central Tendency

- Mean
- Median
- Mode

Typical Measures of Central Tendency

Mean | Median | Mode

The Mean

The mean is the sum of a set of scores divided by the number of scores in the set

- Commonly known as the "average"
- Appropriate for interval, or ratio scale data
- Not valid for nominal or ordinal scale data

Population

μ

mu
"mew"

Sample

M "M"

$\overline{X}$ "X bar"

Calculating the Mean - Population Formula -

$$\mu = \frac{\text{Sum of Scores}}{\text{Number of Scores}}$$

Calculating the Mean
- Population Formula -

$$\mu = \frac{\Sigma X}{N}$$

9

Calculating the Mean
- Population Formula -

$$\mu = \frac{\Sigma X}{N}$$

10

$$\Sigma X$$

11

Summation Notation

$$\Sigma$$

Take the sum of

12

X
8
6
7
9

13

i	X
1	8
2	6
3	7
4	9

14

Index Number

i	X
1	8
2	6
3	7
4	9

15

i	X
1	8
2	6
3	7
4	9

ΣX ?

16

i	X
1	8
2	6
3	7
4	9

$$\Sigma X = 8 + 6 + 7 + 9$$

17

i	X
1	8
2	6
3	7
4	9

$$\Sigma X = 30$$

18

i	X
1	8
2	6
3	7
4	9

19

i	X
1	8
2	6
3	7
4	9

$$\Sigma X_i \, ?$$

20

i	X
1	8
2	6
3	7
4	9

$$\Sigma X_i = X_1 + X_2 + X_3 + X_4$$

21

i	X
1	8
2	6
3	7
4	9

$$\Sigma X_i = 8 + 6 + 7 + 9$$

22

i	X
1	8
2	6
3	7
4	9

$$\Sigma X_i = 30$$

23

i	X
1	8
2	6
3	7
4	9

24

i	X
1	8
2	6
3	7
4	9

$$\Sigma(X_i - 5)^2 \, ?$$

25

i	X
1	8
2	6
3	7
4	9

$$\Sigma(X_i - 5)^2 = (X_1 - 5)^2 + (X_2 - 5)^2 + (X_3 - 5)^2 + (X_4 - 5)^2$$

26

i	X
1	8
2	6
3	7
4	9

$$\Sigma(X_i - 5)^2 = 3^2 + 1^2 + 2^2 + 4^2$$

27

i	X
1	8
2	6
3	7
4	9

$$\Sigma(X_i - 5)^2 = 9 + 1 + 4 + 16$$

28

i	X
1	8
2	6
3	7
4	9

$$\Sigma(X_i - 5)^2 = 30$$

29

Summation Notation

Σ

Take the sum of

30

ΣX

31

Calculating the Mean - Population Formula -

$$\mu = \frac{\Sigma X}{N}$$

32

Calculating the Mean
- Population Formula -

$$\mu = \frac{\Sigma X}{N}$$

33

Calculating the Mean
- Population Formula -

$$\mu = \frac{\Sigma X_i}{N}$$

34

Population

$$\mu = \frac{\Sigma X_i}{N}$$

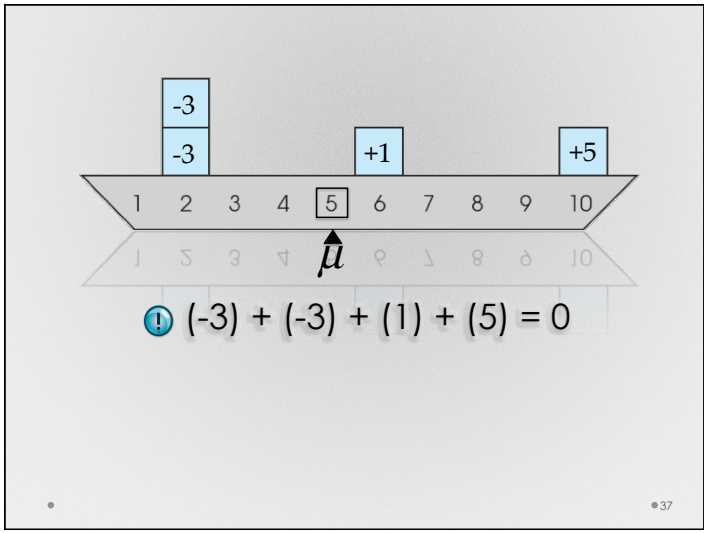
Sample

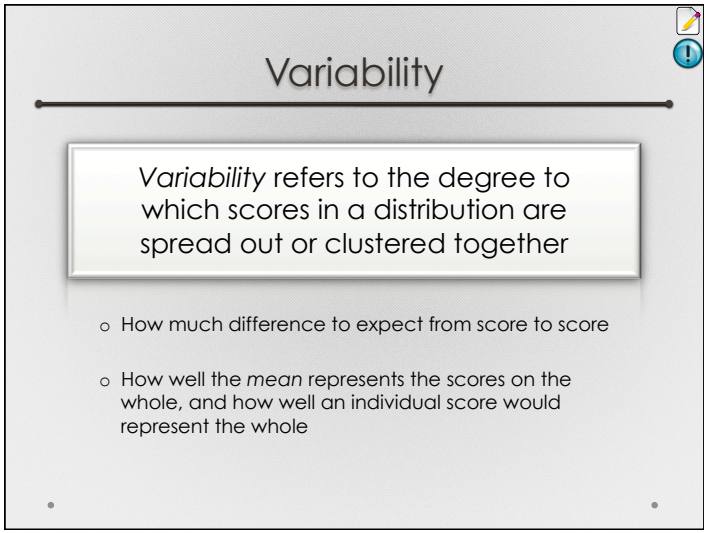
$$\bar{X} = \frac{\Sigma X_i}{n}$$

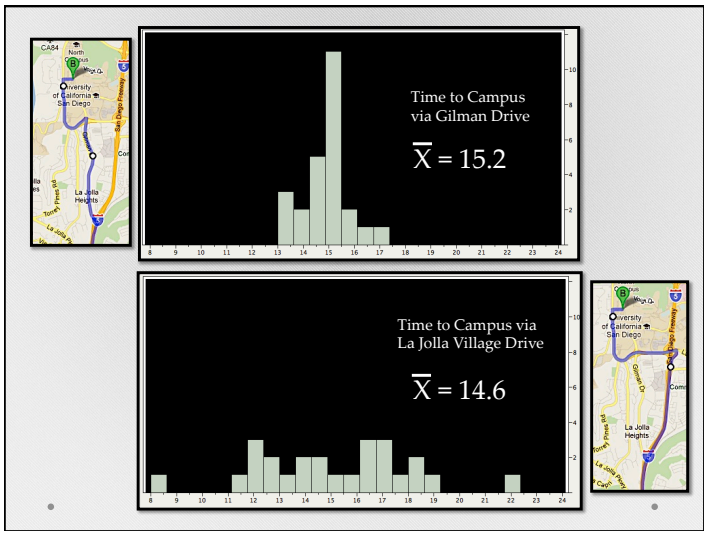
35

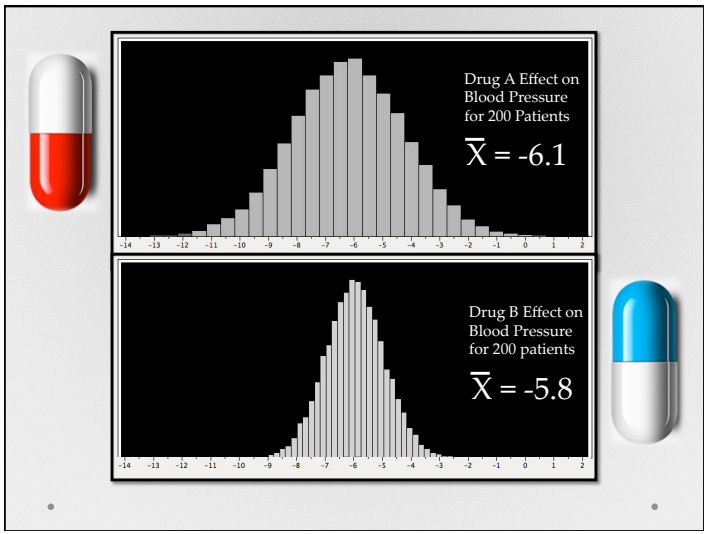
Visualizing the Mean

36


-3
-3
+1
+5
1 2 3 4 5 6 7 8 9 10
μ
(-3) + (-3) + (1) + (5) = 0
37


Variability
Variability refers to the degree to which scores in a distribution are spread out or clustered together
How much difference to expect from score to score
How well the *mean* represents the scores on the whole, and how well an individual score would represent the whole


Time to Campus via Gilman Drive
$\overline{X}$ = 15.2
Time to Campus via La Jolla Village Drive
$\overline{X}$ = 14.6


Drug A Effect on Blood Pressure for 200 Patients
$\overline{X}$ = -6.1
Drug B Effect on Blood Pressure for 200 patients
$\overline{X}$ = -5.8

Standard Deviation

The *standard* (typical) amount scores *deviate* from the mean

Population

σ

"sigma"

Sample

s

"s"

Population

σ

"sigma"

Sample

s

"s"

Population Standard Deviation

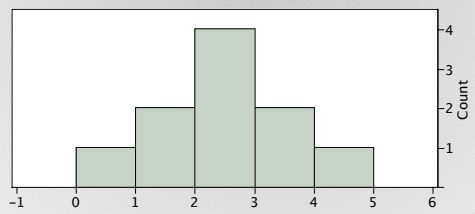
Cups of Coffee Consumed Each Day
N = 10 Individuals

2, 4, 3, 2, 2, 3, 1, 0, 2, 1

Cups of Coffee Consumed Each Day
N = 10 Individuals

2, 4, 3, 2, 2, 3, 1, 0, 2, 1

Cups of Coffee Consumed Each Day
N = 10 Individuals



i	X
1	2
2	4
3	3
4	2
5	2
6	3
7	1
8	0
9	2
10	1

Standard Deviation

The *standard* (typical) amount scores *deviate* from the mean

$$deviation_i = X_i - \mu$$

i	X
1	2
2	4
3	3
4	2
5	2
6	3
7	1
8	0
9	2
10	1

i	X	$X_i - \mu$
1	2	
2	4	
3	3	
4	2	
5	2	
6	3	
7	1	
8	0	
9	2	
10	1	

$\mu = 20 / 10 = 2$

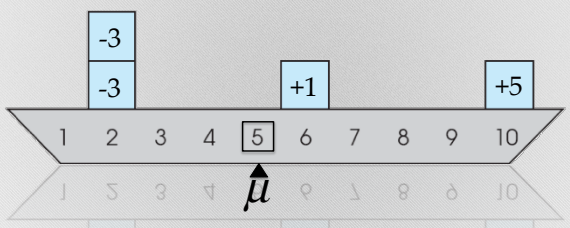
i	X	$X_i - \mu$
1	2	
2	4	
3	3	
4	2	
5	2	
6	3	
7	1	
8	0	
9	2	
10	1	
	20	

$\mu = 20 / 10 = 2$

i	X	$X_i - \mu$
1	2	0
2	4	2
3	3	1
4	2	0
5	2	0
6	3	1
7	1	-1
8	0	-2
9	2	0
10	1	-1
	20	***0***

$\mu = 20 / 10 = 2$

i	X	$X_i - \mu$
1	2	0
2	4	2
3	3	1
4	2	0
5	2	0
6	3	1
7	1	-1
8	0	-2
9	2	0
10	1	-1
	20	***0*** ← Will *always*, by definition, be zero



(-3) + (-3) + (1) + (5) = 0

$\mu = 20 / 10 = 2$

i	X	$X_i - \mu$
1	2	0
2	4	2
3	3	1
4	2	0
5	2	0
6	3	1
7	1	-1
8	0	-2
9	2	0
10	1	-1
	20	***0*** ← Will *always*, by definition, be zero

$|X_i - \mu|$

Absolute Deviation

$(X_i - \mu)^2$

Squared Deviation

Sir Arthur Eddington

Sir Ronald Fisher

$|X_i - \mu|$ $\quad$ $(X_i - \mu)^2$

$\uparrow$ $\quad$ $\uparrow$

$|X_i - \bar{X}|$ $\quad$ $(X_i - \bar{X})^2$

Sampling Error

The discrepancy (or amount of error) that exists between a sample statistic and the corresponding population parameter

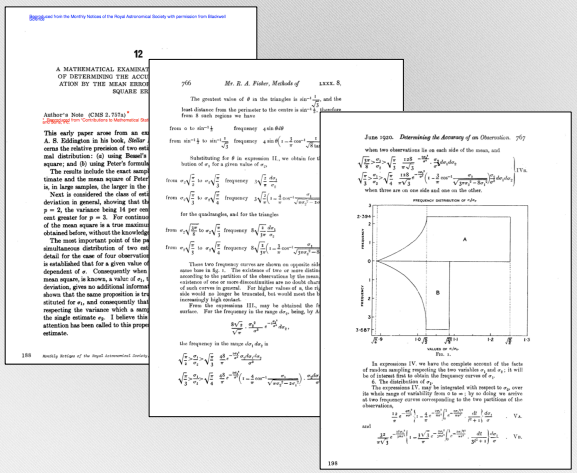
Population Parameters	*Sample 1 Statistics*	*Sample 2 Statistics*
Average Age: 19.8	*Average Age: 21.8*	*Average Age: 19.2*
Average IQ: 110.5	*Average IQ: 106.2*	*Average IQ: 112.2*
65.1% Female, 34.9% Male	*40% Female, 60% Male*	*80% Female, 20% Male*

$|X_i - \mu|$ $\quad$ $(X_i - \mu)^2$

$\uparrow$ $\quad$ $\uparrow$

$|X_i - \bar{X}|$ $\quad$ $(X_i - \bar{X})^2$

Wiki ➔ Readings ➔ Fisher (1920)



Properties of Estimators

- Consistency:
  "Does it get better with more data?"
- (Relative) Efficiency:
  "Does it err less than other estimators?"
- Sufficiency:
  "Does it use all the data?"

Properties of Estimators

- Consistency:
  "Does it get better with more data?"
- **(Relative) Efficiency:**
  "Does it err less than other estimators?"
- Sufficiency:
  "Does it use all the data?"

$$\begin{array}{cc} |X_i - \mu| & (X_i - \mu)^2 \\ \uparrow & \uparrow \\ |X_i - \bar{X}| & (X_i - \bar{X})^2 \end{array}$$

$|X_i - \mu|$

Absolute Deviation

$(X_i - \mu)^2$

Squared Deviation

Sir Arthur Eddington

Sir Ronald Fisher

$(X_i - \mu)^2$

Squared Deviation

The Squared Deviation

the base unit of variability in statistics

$(X_i - \mu)^2$

Squared Deviation

The Squared Deviation

the base unit of variability in statistics

$(X_i - \mu)^2$

Squared Deviation
"Squares"

i	X	$X_i - \mu$	$(X_i - \mu)^2$
1	2	0	
2	4	2	
3	3	1	
4	2	0	
5	2	0	
6	3	1	
7	1	-1	
8	0	-2	
9	2	0	
10	1	-1	
	20	***0***	

i	X	$X_i - \mu$	$(X_i - \mu)^2$
1	2	0	0
2	4	2	4
3	3	1	1
4	2	0	0
5	2	0	0
6	3	1	1
7	1	-1	1
8	0	-2	4
9	2	0	0
10	1	-1	1
	20	***0***	

i	X	$X_i - \mu$	$(X_i - \mu)^2$
1	2	0	0
2	4	2	4
3	3	1	1
4	2	0	0
5	2	0	0
6	3	1	1
7	1	-1	1
8	0	-2	4
9	2	0	0
10	1	-1	1
	20	***0***	***12***

Sum of the Squared Deviations
aka *sum of squares*

SS

"S - S"

i	X	$X_i - \mu$	$(X_i - \mu)^2$
1	2	0	0
2	4	2	4
3	3	1	1
4	2	0	0
5	2	0	0
6	3	1	1
7	1	-1	1
8	0	-2	4
9	2	0	0
10	1	-1	1
	20	***0***	***12*** ← ***SS***

$$\frac{SS}{N} = \frac{12}{10} = 1.2 = \text{Mean Squared Deviation}$$

Population Variance

Mean Squared Deviation

$$\frac{SS}{N} = \frac{12}{10} = 1.2 = \text{Mean Squared Deviation}$$

Population Variance

Mean Squared Deviation

Mean *squared* distance of scores to the mean

$$\text{Variance} = \frac{\text{Sum of Squared Deviations}}{\text{Number of Scores}}$$

Population Variance

Mean Squared Deviation

Mean *squared* distance of scores to the mean

$$\sigma^2 = \frac{SS}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

$$Standard\ Deviation = \sqrt{Variance}$$

$$(Standard\ Deviation)^2 = Variance$$

Population Variance

Mean Squared Deviation

Mean *squared* distance of scores to the mean

$$\sigma^2 = \frac{SS}{N}$$

Population Standard Deviation

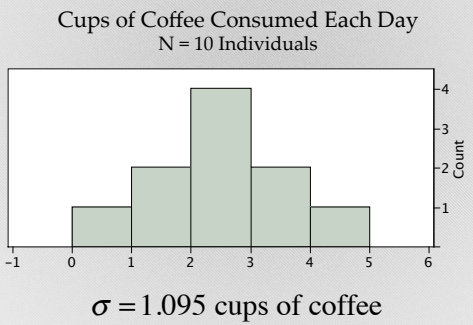
*Average** distance of scores to the mean

$$\sigma = \sqrt{\frac{SS}{N}}$$

** Not actually the average. The average distance of scores to the mean is the: "mean absolute deviation"*

$$\frac{SS}{N} = \frac{12}{10} = 1.2 = \text{Mean Squared Deviation}$$

$$\sigma = \sqrt{1.2} = 1.095 \text{ cups of coffee}$$



$$\sigma = 1.095 \text{ cups of coffee}$$

Complete Steps for Calculating the Population Standard Deviation

$X_i - \mu$	Find each deviation score
$(X_i - \mu)^2$	Square each deviation score
$SS = \Sigma(X_i - \mu)^2$	Sum the squared deviations
$\sigma^2 = SS/N$	Divide SS by the number of scores
$\sigma = \sqrt{\sigma^2}$	Take the square root of the result

Population Standard Deviation

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

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Population

σ

"sigma"

Sample

s

"s"

Population

σ

"sigma"

Sample

s

"s"

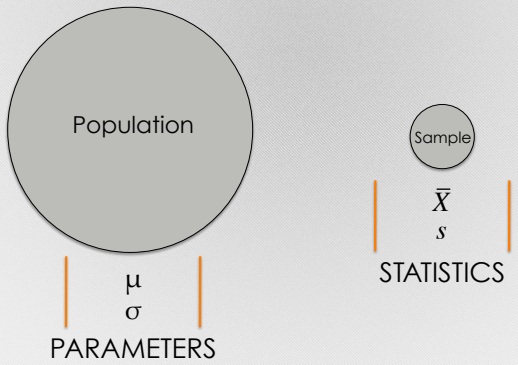
Sample Standard Deviation

Standard Deviation

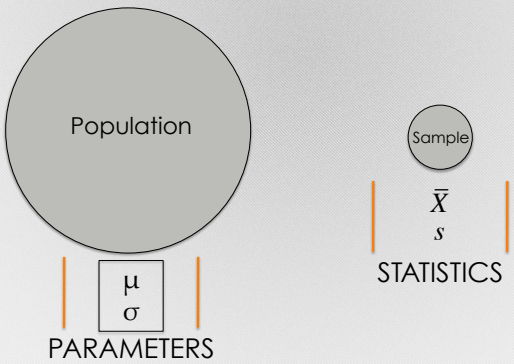
FROM BEFORE

The *standard* (typical) amount scores *deviate* from the mean

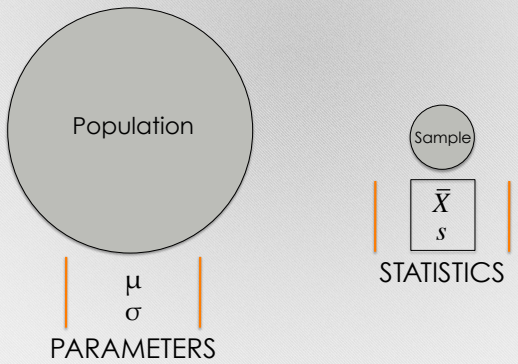
Parameters and Statistics



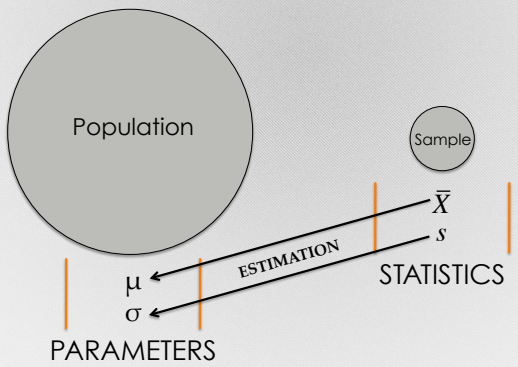
Parameters and Statistics



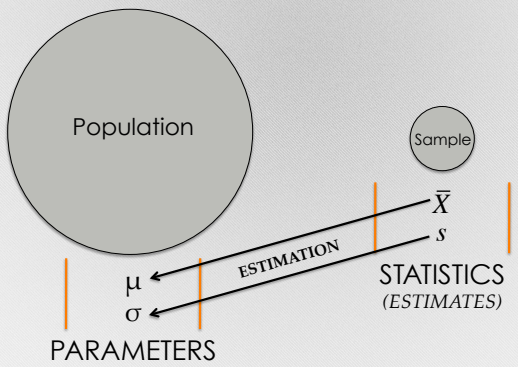
Parameters and Statistics



Parameters and Statistics



Parameters and Statistics



Properties of Estimators

- Consistency:
  "Does it get better with more data?"
- (Relative) Efficiency:
  "Does it err less than other estimators?"
- Sufficiency:
  "Does it use all the data?"

Properties of Estimators

- Consistency:
  "Does it get better with more data?"
- (Relative) Efficiency:
  "Does it err less than other estimators?"
- Sufficiency:
  "Does it use all the data?"
- Bias:
  "Does it over- or under-estimate the true value on average?"

Standard Deviation

The *standard* (typical) amount scores *deviate* from the mean

Sample Standard Deviation

An estimate based on sample data of the *standard deviation of the population* from which the *sample* was drawn

Sample Standard Deviation

corrected for the purpose of estimation

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2$$

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2$$

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s = \sqrt{\frac{SS}{n-1}}$$

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s = \sqrt{\frac{SS}{n-1}}$$

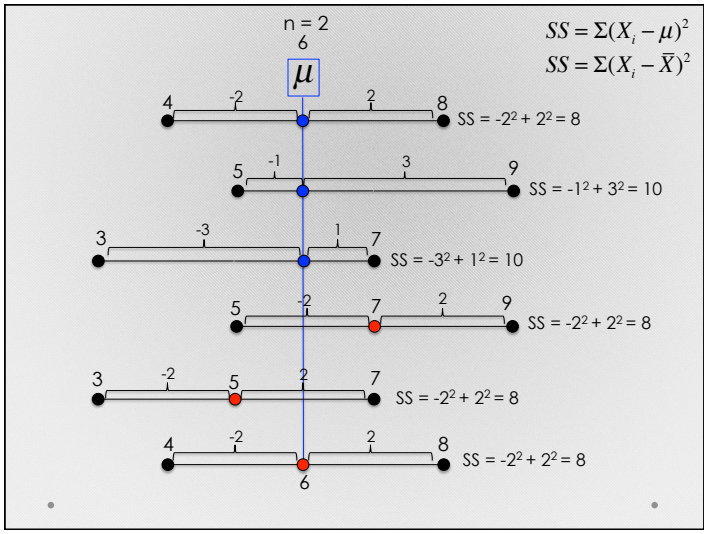
Calculating the Standard Deviation

Population

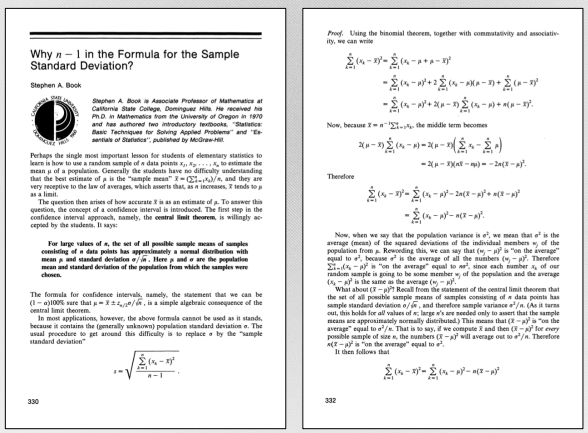
$$SS = \Sigma(X_i - \mu)^2 \quad \Rightarrow \quad \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \quad \Rightarrow \quad s = \sqrt{\frac{SS}{n-1}}$$



Wiki ➔ Readings ➔ Why n-1? (1979)



Why $n-1$ in the Formula for the Sample Standard Deviation?

Stephen A. Book

Stephen A. Book is Associate Professor of Mathematics at California State College, Dominguez Hills. He received his Ph.D. in Mathematics from the University of Oregon in 1970 and has authored two introductory textbooks, "Statistics: Basic Techniques for Solving Applied Problems" and "Essentials of Statistics", published by McGraw-Hill.

Perhaps the single most important lesson for students of elementary statistics to learn is how to use a random sample of n data points $x_1, x_2, \ldots, x_n$ to estimate the mean μ of a population. Generally the students have no difficulty understanding that the best estimate of μ is the "sample mean" $\bar{x} = (\sum_{k=1}^{n} x_k)/n$, and they are very receptive to the law of averages, which asserts that, as n increases, $\bar{x}$ tends to μ as a limit.

The question then arises of how accurate $\bar{x}$ is as an estimate of μ. To answer this question, the concept of a confidence interval is introduced. The first step in the confidence interval approach, namely, the **central limit theorem**, is willingly accepted by the students. It says:

For large values of n, the set of all possible sample means of samples consisting of n data points has approximately a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. Here μ and σ are the population mean and standard deviation of the population from which the samples were chosen.

The formula for confidence intervals, namely, the statement that we can be $(1-\alpha)100\%$ sure that $\mu = \bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$, is a simple algebraic consequence of the central limit theorem.

In most applications, however, the above formula cannot be used as it stands, because it contains the (generally unknown) population standard deviation σ. The usual procedure to get around this difficulty is to replace σ by the "sample standard deviation"

$$s = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1}}.$$

330

Proof. Using the binomial theorem, together with commutativity and associativity, we can write

$$\sum_{k=1}^{n}(x_k - \bar{x})^2 = \sum_{k=1}^{n}(x_k - \mu + \mu - \bar{x})^2$$
$$= \sum_{k=1}^{n}(x_k - \mu)^2 + 2\sum_{k=1}^{n}(x_k - \mu)(\mu - \bar{x}) + \sum_{k=1}^{n}(\mu - \bar{x})^2$$
$$= \sum_{k=1}^{n}(x_k - \mu)^2 + 2(\mu - \bar{x})\sum_{k=1}^{n}(x_k - \mu) + n(\mu - \bar{x})^2.$$

Now, because $\bar{x} = n^{-1}\sum_{k=1}^{n} x_k$, the middle term becomes

$$2(\mu - \bar{x})\sum_{k=1}^{n}(x_k - \mu) = 2(\mu - \bar{x})\left(\sum_{k=1}^{n} x_k - \sum_{k=1}^{n}\mu\right)$$
$$= 2(\mu - \bar{x})(n\bar{x} - n\mu) = -2n(\bar{x} - \mu)^2.$$

Therefore

$$\sum_{k=1}^{n}(x_k - \bar{x})^2 = \sum_{k=1}^{n}(x_k - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2$$
$$= \sum_{k=1}^{n}(x_k - \mu)^2 - n(\bar{x} - \mu)^2.$$

Now, when we say that the population variance is σ^2, we mean that σ^2 is the average (mean) of the squared deviations of the individual members x_i of the population from μ. Rewording this, we can say that $(x_i - \mu)^2$ is "on the average" equal to σ^2, because σ^2 is the average of all the numbers $(x_i - \mu)^2$. Therefore $\sum_{k=1}^{n}(x_k - \mu)^2$ is "on the average" equal to $n\sigma^2$, since each number x_k of our random sample is going to be some member x_j of the population and the average $(x_k - \mu)^2$ is the same as the average $(x_j - \mu)^2$.

What about $(\bar{x} - \mu)^2$? Recall from the statement of the central limit theorem that the set of all possible sample means of samples consisting of n data points has sample standard deviation $\sigma/\sqrt{n}$, and therefore sample variance σ^2/n. (As it turns out, this holds for *all* values of n; large n's are needed only to assert that the sample means are approximately normally distributed.) This means that $(\bar{x} - \mu)^2$ is "on the average" equal to σ^2/n. That is to say, if we compute $\bar{x}$ and then $(\bar{x} - \mu)^2$ for *every* possible sample of size n, the numbers $(\bar{x} - \mu)^2$ will average out to σ^2/n. Therefore $n(\bar{x} - \mu)^2$ is "on the average" equal to σ^2.

It then follows that

$$\sum_{k=1}^{n}(x_k - \bar{x})^2 = \sum_{k=1}^{n}(x_k - \mu)^2 - n(\bar{x} - \mu)^2$$

332

Sampling Distn Demo of Variance

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s = \sqrt{\frac{SS}{n-1}}$$

Calculating the Standard Deviation

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma = \sqrt{\frac{SS}{N}}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s = \sqrt{\frac{SS}{n-1}}$$

Calculating the Variance

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma^2 = \frac{SS}{N}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s^2 = \frac{SS}{n-1}$$

Calculating the Variance

Population

$$SS = \Sigma(X_i - \mu)^2 \Rightarrow \sigma^2 = \frac{SS}{N}$$

Sample

$$SS = \Sigma(X_i - \bar{X})^2 \Rightarrow s^2 = \frac{SS}{n-1}$$

$$s^2 \hat{=} \sigma^2$$

$$s^2 \hat{=} \sigma^2$$

is an unbiased estimator of

$$s^2 \hat{=} \sigma^2$$

$$\bar{X} \hat{=} \mu$$