

Lecture 8.2, MATH-57091 Probability and Statistics for High-School Teachers.

Artem Zvavitch

Department of Mathematical Sciences, Kent State University

October 12 - October 15, 2014.

Our goal now will be to learn how to estimate population mean, population proportion, population variance and understand how to measure an error of our estimation process for those parameters.

Our goal now will be to learn how to estimate population mean, population proportion, population variance and understand how to measure an error of our estimation process for those parameters.

A recent poll of 2000 randomly chosen Americans indicates that 52 percent for entire U.S. population likes to eat banana, with margin of error of ± 3 percent."

So what the above claim mean?

Our goal now will be to learn how to estimate population mean, population proportion, population variance and understand how to measure an error of our estimation process for those parameters.

A recent poll of 2000 randomly chosen Americans indicates that 52 percent for entire U.S. population likes to eat banana, with margin of error of ± 3 percent."

So what the above claim mean? What is "margin ± 3 percent?"

Our goal now will be to learn how to estimate population mean, population proportion, population variance and understand how to measure an error of our estimation process for those parameters.

A recent poll of 2000 randomly chosen Americans indicates that 52 percent for entire U.S. population likes to eat banana, with margin of error of ± 3 percent."

So what the above claim mean? What is "margin ± 3 percent? How can you check love to banana among just a few people (yes 2000 out of 150 million is just a few!) and make a statement about all population?

Our goal now will be to learn how to estimate population mean, population proportion, population variance and understand how to measure an error of our estimation process for those parameters.

A recent poll of 2000 randomly chosen Americans indicates that 52 percent for entire U.S. population likes to eat banana, with margin of error of ± 3 percent."

So what the above claim mean? What is "margin ± 3 percent? How can you check love to banana among just a few people (yes 2000 out of 150 million is just a few!) and make a statement about all population?

An estimator is a static whose value depends on the particular sample drawn. The value of the estimator, called, estimate, is used to predict the value of a population parameter.

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN.

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN. The sample mean $\bar{X} = \sum X_i/n$ can be used as an estimator of μ .

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN. The sample mean $\bar{X} = \sum X_i/n$ can be used as an estimator of μ . There is something quite good about this estimator, at least, in average we do know that it gives a correct value, i.e.

$$\mathbb{E}\bar{X} = \mu.$$

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN. The sample mean $\bar{X} = \sum X_i/n$ can be used as an estimator of μ . There is something quite good about this estimator, at least, in average we do know that it gives a correct value, i.e.

$$\mathbb{E}\bar{X} = \mu.$$

An estimator whose expected value is equal to the parameter it is estimating is said to be an **unbiased** estimator of that parameter.

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN. The sample mean $\bar{X} = \sum X_i/n$ can be used as an estimator of μ . There is something quite good about this estimator, at least, in average we do know that it gives a correct value, i.e.

$$\mathbb{E}\bar{X} = \mu.$$

An estimator whose expected value is equal to the parameter it is estimating is said to be an **unbiased** estimator of that parameter.

Example

To estimate the average of amount of cash a student have with her/him on Kent Campus we asked 12 randomly selected on campus students to provide us with this information, we got:

125, 87, 5, 50, 45, 76, 12, 45, 91, 54, 1, 17.

Point Estimator of A population mean

Let X_1, \dots, X_n denote a sample from a population whose mean is μ , AND μ IS UNKNOWN. The sample mean $\bar{X} = \sum X_i/n$ can be used as an estimator of μ . There is something quite good about this estimator, at least, in average we do know that it gives a correct value, i.e.

$$\mathbb{E}\bar{X} = \mu.$$

An estimator whose expected value is equal to the parameter it is estimating is said to be an **unbiased** estimator of that parameter.

Example

To estimate the average of amount of cash a student have with her/him on Kent Campus we asked 12 randomly selected on campus students to provide us with this information, we got:

125, 87, 5, 50, 45, 76, 12, 45, 91, 54, 1, 17.

To estimate the average for all Kent Student amount we compute the sample mean

$$\bar{X} = \frac{125 + 87 + 5 + 50 + 45 + 76 + 12 + 45 + 91 + 54 + 1 + 17}{12} = 50.7$$

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ .

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} .

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value.

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value. Thus it is essential for us to determine the standard deviation of \bar{X}

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value. Thus it is essential for us to determine the standard deviation of \bar{X} (for example if we put $m = 2$ then the probability that the random variable is 2 standard deviations different from it's expected value is less then $1/4$,

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value. Thus it is essential for us to determine the standard deviation of \bar{X} (for example if we put $m = 2$ then the probability that the random variable is 2 standard deviations different from it's expected value is less then $1/4$, moreover, if we know that \bar{X} is close to the Normal Distribution, which is the case when n is large, this probability becomes almost zero (check the table!) so we can almost certainly assume that \bar{X} within two standard deviations from μ).

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value. Thus it is essential for us to determine the standard deviation of \bar{X} (for example if we put $m = 2$ then the probability that the random variable is 2 standard deviations different from it's expected value is less then $1/4$, moreover, if we know that \bar{X} is close to the Normal Distribution, which is the case when n is large, this probability becomes almost zero (check the table!) so we can almost certainly assume that \bar{X} within two standard deviations from μ).

As we learned before

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

We know that $\mathbb{E}\bar{X} = \mu$ but clearly, we would like to know how different \bar{X} may be from μ . Chebychev inequality tell us that

$$P(|\bar{X} - \mu| \geq k) \leq \frac{\sigma^2}{k^2},$$

where σ^2 is the variance of \bar{X} . If we substitute $k = m\sigma$ in the above inequality, we get

$$P(|\bar{X} - \mu| \geq m\sigma) \leq \frac{1}{m^2}.$$

The above inequality tell us that the random variable (in this case \bar{X}) is not likely to be too many standard deviation away from the expected value. Thus it is essential for us to determine the standard deviation of \bar{X} (for example if we put $m = 2$ then the probability that the random variable is 2 standard deviations different from it's expected value is less then $1/4$, moreover, if we know that \bar{X} is close to the Normal Distribution, which is the case when n is large, this probability becomes almost zero (check the table!) so we can almost certainly assume that \bar{X} within two standard deviations from μ).

As we learned before

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Is population standard deviation (we will "know" it in the next examples, but will learn how to deal with it if it is "like in real life" is not given to us). $SD(\bar{X})$ is sometimes called **the standard error** of \bar{X} .

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

with the standard error of the estimate being equal to

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = .15.$$

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

with the standard error of the estimate being equal to

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = .15.$$

Therefore we can be quite sure that the actual mean will not differ from 3.6 by more than .3.

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

with the standard error of the estimate being equal to

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = .15.$$

Therefore we can be quite sure that the actual mean will not differ from 3.6 by more than .3. Assume we wanted to estimate with the standard error of .05.

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

3.6, 3.9, 3.4, 3.5.

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

with the standard error of the estimate being equal to

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = .15.$$

Therefore we can be quite sure that the actual mean will not differ from 3.6 by more than .3. Assume we wanted to estimate with the standard error of .05. Then, we would need

$$\frac{\sigma}{\sqrt{n}} = .05$$

or

Example:

Successive tests for the level of potassium in an individual's blood vary because of the basic imprecision of the test and because the actual level itself varies depending on such things as the amount of food recently eaten and the amount of exertion recently undergone. Suppose it is known that, for a given individual, the successive readings of potassium level vary around a mean value μ with a standard deviation of .3. If a set of four readings on particular individual yields the data

$$3.6, 3.9, 3.4, 3.5.$$

Then the estimate of the mean potassium level of that person is

$$\bar{X} = \frac{3.6 + 3.9 + 3.4 + 3.5}{4} = 3.6,$$

with the standard error of the estimate being equal to

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{0.3}{2} = .15.$$

Therefore we can be quite sure that the actual mean will not differ from 3.6 by more than .3. Assume we wanted to estimate with the standard error of .05. Then, we would need

$$\frac{\sigma}{\sqrt{n}} = .05$$

or

$$\frac{.3}{\sqrt{n}} = .05$$

which gives $n = 36$.

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study.

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion.

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q .

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p .

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

The above quantity is called the **Standard Error of \hat{p}** .

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

The above quantity is called the **Standard Error of \hat{p}** . What is very cool that we can do a bit of calculus and provide a very useful estimate of $SD(\hat{p})$ which would be independent of p (remember p is unknown so it is not so useful for us). Indeed, if we study the parabola $f(p) = p(1-p)$ on the interval $[0, 1]$ we can see that it achieves its maximum at $p = .5$ and thus

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

The above quantity is called the **Standard Error of \hat{p}** . What is very cool that we can do a bit of calculus and provide a very useful estimate of $SD(\hat{p})$ which would be independent of p (remember p is unknown so it is not so useful for us). Indeed, if we study the parabola $f(p) = p(1-p)$ on the interval $[0, 1]$ we can see that it achieves its maximum at $p = .5$ and thus

$$p(1-p) \leq \frac{1}{4}, \text{ FOR ALL } p \in [0, 1].$$

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

The above quantity is called the **Standard Error of \hat{p}** . What is very cool that we can do a bit of calculus and provide a very useful estimate of $SD(\hat{p})$ which would be independent of p (remember p is unknown so it is not so useful for us). Indeed, if we study the parabola $f(p) = p(1-p)$ on the interval $[0, 1]$ we can see that it achieves its maximum at $p = .5$ and thus

$$p(1-p) \leq \frac{1}{4}, \text{ FOR ALL } p \in [0, 1].$$

Finally, we get that always

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}.$$

Point estimator for population proportion

Suppose that we are trying to estimate the proportion of a large population that satisfy some property Q that we would like to study. Let p denote the unknown proportion. To estimate p we should choose a random sample and then we will estimate p by proportion of the sample that satisfy the property Q :

$$\hat{p} = \frac{X}{n},$$

where X is the number of members of the sample who satisfy the property Q . As we know from the previous lecture

$$\mathbb{E}\hat{p} = p,$$

and thus \hat{p} is unbiased estimator of p . Moreover we know that

$$SD(\hat{p}) = SD\left(\frac{X}{n}\right) = \frac{1}{n}SD(X) = \sqrt{\frac{p(1-p)}{n}}.$$

The above quantity is called the **Standard Error of \hat{p}** . What is very cool that we can do a bit of calculus and provide a very useful estimate of $SD(\hat{p})$ which would be independent of p (remember p is unknown so it is not so useful for us). Indeed, if we study the parabola $f(p) = p(1-p)$ on the interval $[0, 1]$ we can see that it achieves its maximum at $p = .5$ and thus

$$p(1-p) \leq \frac{1}{4}, \text{ FOR ALL } p \in [0, 1].$$

Finally, we get that always

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}.$$

For example, if we have a random sample of 900. Then **no matter what proportion of the population is actually satisfy property Q** it follows that the standard error of the estimator \hat{p} is less then $1/(2\sqrt{900}) = 1/60$.

Example

A school district is trying to determine its students reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 were in favor of the proposal, then

- Estimate the proportion of all students who are in favor.
- Estimate the standard error of the estimate.

A school district is trying to determine its students reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 were in favor of the proposal, then

- Estimate the proportion of all students who are in favor.
- Estimate the standard error of the estimate.

Solution: The estimate of the proportion of all students who are in favor of the dress code is

$$\hat{p} = \frac{20}{50} = .4$$

A school district is trying to determine its students reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 were in favor of the proposal, then

- Estimate the proportion of all students who are in favor.
- Estimate the standard error of the estimate.

Solution: The estimate of the proportion of all students who are in favor of the dress code is

$$\hat{p} = \frac{20}{50} = .4$$

There are two ways to estimate $SD(\hat{p})$, first we notice that $SD(\hat{p}) \leq 1/2\sqrt{20} = 0.1$.

A school district is trying to determine its students reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 were in favor of the proposal, then

- Estimate the proportion of all students who are in favor.
- Estimate the standard error of the estimate.

Solution: The estimate of the proportion of all students who are in favor of the dress code is

$$\hat{p} = \frac{20}{50} = .4$$

There are two ways to estimate $SD(\hat{p})$, first we notice that $SD(\hat{p}) \leq 1/2\sqrt{20} = 0.1$. But we may also do a bit better. Consider

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

A school district is trying to determine its students reaction to proposed dress code. To do so, the school selected a random sample of 50 students and questioned them. If 20 were in favor of the proposal, then

- Estimate the proportion of all students who are in favor.
- Estimate the standard error of the estimate.

Solution: The estimate of the proportion of all students who are in favor of the dress code is

$$\hat{p} = \frac{20}{50} = .4$$

There are two ways to estimate $SD(\hat{p})$, first we notice that $SD(\hat{p}) \leq 1/2\sqrt{20} = 0.1$. But we may also do a bit better. Consider

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

use the estimate we now have for p , i.e. \hat{p} :

$$SD(\hat{p}) \approx \sqrt{\frac{.4(1-.4)}{50}} = 0.07.$$

Estimating A population Variance

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

The sample variance S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

is an estimator of the population σ^2 .

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

The sample variance S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

is an estimator of the population σ^2 .

It is always a big surprise for students to see $(n-1)$ in the above formula!

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

The sample variance S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

is an estimator of the population σ^2 .

It is always a big surprise for students to see $(n-1)$ in the above formula! But the reason is quite "simple"

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

The sample variance S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

is an estimator of the population σ^2 .

It is always a big surprise for students to see $(n-1)$ in the above formula! But the reason is quite "simple" - to make the estimator unbiased, using direct computation (i.e. playing with sums) one can prove that:

$$\mathbb{E}S^2 = \sigma^2.$$

Suppose that we have a sample of size n : X_1, X_2, \dots, X_n , from a population whose variance σ^2 is unknown, and that we are interested in using the sample data to estimate σ^2 .

The sample variance S^2 is defined by

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

is an estimator of the population σ^2 .

It is always a big surprise for students to see $(n-1)$ in the above formula! But the reason is quite "simple" - to make the estimator unbiased, using direct computation (i.e. playing with sums) one can prove that:

$$\mathbb{E}S^2 = \sigma^2.$$

We will prove it on the next slide just for fun....

S^2 is unbiased estimate for σ^2

Note that, using $\bar{X} = \sum_{k=1}^n X_k/n$ and $(\sum_{k=1}^n a_k)^2 = \sum_{k=1}^n a_k^2 + \sum_{k \neq j} a_k a_j$ we get

Note that, using $\bar{X} = \sum_{k=1}^n X_k/n$ and $(\sum_{k=1}^n a_k)^2 = \sum_{k=1}^n a_k^2 + \sum_{k \neq j} a_k a_j$ we get

$$(X_i - \bar{X})^2 = X_i^2 - 2X_i\bar{X} + (\bar{X})^2 = X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n} \sum_{k:k \neq i} X_i X_k + \frac{1}{n^2} \sum_{k=1}^n X_k^2 + \frac{1}{n^2} \sum_{k \neq i} X_i X_k$$

S^2 is unbiased estimate for σ^2

Note that, using $\bar{X} = \sum_{k=1}^n X_k/n$ and $(\sum_{k=1}^n a_k)^2 = \sum_{k=1}^n a_k^2 + \sum_{k \neq j} a_k a_j$ we get

$$(X_i - \bar{X})^2 = X_i^2 - 2X_i\bar{X} + (\bar{X})^2 = X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n} \sum_{k:k \neq i} X_i X_k + \frac{1}{n^2} \sum_{k=1}^n X_k^2 + \frac{1}{n^2} \sum_{k \neq i} X_i X_k$$

Then

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \left(1 - \frac{2}{n}\right) \sum_{i=1}^n X_i^2 - \frac{2}{n} \sum_{k \neq i} X_i X_k + \frac{1}{n} \sum_{k=1}^n X_k^2 + \frac{1}{n} \sum_{k \neq j} X_j X_k$$

Note that, using $\bar{X} = \sum_{k=1}^n X_k/n$ and $(\sum_{k=1}^n a_k)^2 = \sum_{k=1}^n a_k^2 + \sum_{k \neq j} a_k a_j$ we get

$$(X_i - \bar{X})^2 = X_i^2 - 2X_i\bar{X} + (\bar{X})^2 = X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n} \sum_{k:k \neq i} X_i X_k + \frac{1}{n^2} \sum_{k=1}^n X_k^2 + \frac{1}{n^2} \sum_{k \neq i} X_i X_k$$

Then

$$\sum_{i=1}^n (X_i - \bar{X})^2 = (1 - \frac{2}{n}) \sum_{i=1}^n X_i^2 - \frac{2}{n} \sum_{k \neq i} X_i X_k + \frac{1}{n} \sum_{k=1}^n X_k^2 + \frac{1}{n} \sum_{k \neq j} X_j X_k$$

Now we use that all X_i are independent and identically distributed:

$$\mathbb{E} \sum_{i=1}^n (X_i - \bar{X})^2 = (1 - \frac{1}{n}) \sum_{i=1}^n \mathbb{E} X_i^2 - \frac{1}{n} \sum_{k \neq i} \mathbb{E} X_i X_k = (1 - \frac{1}{n}) n \mathbb{E} X_1^2 - \frac{1}{n} (n-1)n (\mathbb{E} X_1)^2$$

S^2 is unbiased estimate for σ^2

Note that, using $\bar{X} = \sum_{k=1}^n X_k/n$ and $(\sum_{k=1}^n a_k)^2 = \sum_{k=1}^n a_k^2 + \sum_{k \neq j} a_k a_j$ we get

$$(X_i - \bar{X})^2 = X_i^2 - 2X_i\bar{X} + (\bar{X})^2 = X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n} \sum_{k:k \neq i} X_i X_k + \frac{1}{n^2} \sum_{k=1}^n X_k^2 + \frac{1}{n^2} \sum_{k \neq i} X_i X_k$$

Then

$$\sum_{i=1}^n (X_i - \bar{X})^2 = (1 - \frac{2}{n}) \sum_{i=1}^n X_i^2 - \frac{2}{n} \sum_{k \neq i} X_i X_k + \frac{1}{n} \sum_{k=1}^n X_k^2 + \frac{1}{n} \sum_{k \neq j} X_j X_k$$

Now we use that all X_i are independent and identically distributed:

$$\begin{aligned} \mathbb{E} \sum_{i=1}^n (X_i - \bar{X})^2 &= (1 - \frac{1}{n}) \sum_{i=1}^n \mathbb{E} X_i^2 - \frac{1}{n} \sum_{k \neq i} \mathbb{E} X_i X_k = (1 - \frac{1}{n}) n \mathbb{E} X_1^2 - \frac{1}{n} (n-1)n (\mathbb{E} X_1)^2 \\ &= (n-1) \mathbb{E} X_1^2 - (n-1) (\mathbb{E} X_1)^2 = (n-1) \sigma^2. \end{aligned}$$

Example

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

Solution: To answer this question we need to find the sample variance S^2 .

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

Solution: To answer this question we need to find the sample variance S^2 . We will do a trick which would save us quite a bit of time and calculation.

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

Solution: To answer this question we need to find the sample variance S^2 . We will do a trick which would save us quite a bit of time and calculation. We note that if we subtract the same constant value from each point of the data, this action will not change the value of variance.

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

Solution: To answer this question we need to find the sample variance S^2 . We will do a trick which would save us quite a bit of time and calculation. We note that if we subtract the same constant value from each point of the data, this action will not change the value of variance. So we subtract 1200:

11, 24, -3, 8, 20, 16, 13, -2, -3.

A random sample of nine electronic components produced by a certain company yields the following sizes:

1211, 1224, 1197, 1208, 1220, 1216, 1213, 1198, 1197.

What are the estimates of the population standard deviation and the population variance?

Solution: To answer this question we need to find the sample variance S^2 . We will do a trick which would save us quite a bit of time and calculation. We note that if we subtract the same constant value from each point of the data, this action will not change the value of variance. So we subtract 1200:

11, 24, -3, 8, 20, 16, 13, -2, -3.

Now do direct calculation (or use calculator) to get estimator for the variance $S^2 = 103$ and for the standard deviation $S = 10.15$.