Lecture 7, MATH-57091 Probability and Statistics for High-School Teachers.

Artem Zvavitch

Department of Mathematical Sciences, Kent State University

October 5 - October 10, 2014.

Welcome to Statistics.

Our goal will be to try to make a conclusion from a set of observed data.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.

Welcome to Statistics.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.
To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.
To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.
Clearly, it is almost impossible to deal with a super general case, we will need to make some assumptions.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.
To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.
Clearly, it is almost impossible to deal with a super general case, we will need to make some assumptions. We assume that there is a probability distribution of the above values, and the values assigned to a different members of the population are independent random variables, having the same distribution.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.

To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.

Clearly, it is almost impossible to deal with a super general case, we will need to make some assumptions. We assume that there is a probability distribution of the above values, and the values assigned to a different members of the population are independent random variables, having the same distribution. In most applications the distribution is not known (we may know some parameters).

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.
To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.
Clearly, it is almost impossible to deal with a super general case, we will need to make some assumptions. We assume that there is a probability distribution of the above values, and the values assigned to a different members of the population are independent random variables, having the same distribution. In most applications the distribution is not known (we may know some parameters). For example a factory produce cars, each car have some (unknown at start!) time for which it will work without any problems (this is our distribution and we have no idea about it). We would like to study how long those cars will survive, i.e. our goal is to get some information about the distribution. We can not check all cars - we will check a few (a sample) and will try to make a conclusion.

Our goal will be to try to make a conclusion from a set of observed data. The data will usually consists of sample of certain elements of a population, and we will try draw a conclusion about the entire population.
To make it a bit more "mathematical" from here let us assume that each member of population has a numerical value associated with it. The value, depends of what we are interested, or often "told", to study.
Clearly, it is almost impossible to deal with a super general case, we will need to make some assumptions. We assume that there is a probability distribution of the above values, and the values assigned to a different members of the population are independent random variables, having the same distribution. In most applications the distribution is not known (we may know some parameters). For example a factory produce cars, each car have some (unknown at start!) time for which it will work without any problems (this is our distribution and we have no idea about it). We would like to study how long those cars will survive, i.e. our goal is to get some information about the distribution. We can not check all cars - we will check a few (a sample) and will try to make a conclusion.

If $X_1, \dots, X_n$ are independent random variables having a common probability distribution, we say they constitute a sample from that distribution.

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*)

Sample Mean

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*) and variance σ^2 (*population variance*).

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*) and variance σ^2 (*population variance*). Let $X_1, X_2, \dots, X_n$ be a sample of values from the population. Then the sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*) and variance σ^2 (*population variance*). Let $X_1, X_2, \ldots, X_n$ be a sample of values from the population. Then the sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

NOTE: $\bar{X}$ is a random variable!.

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*) and variance σ^2 (*population variance*). Let $X_1, X_2, \ldots, X_n$ be a sample of values from the population. Then the sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

NOTE: $\bar{X}$ is a random variable!. It is good point to remark that

$$\mathbb{E}\bar{X} = \mu \text{ and } Var\bar{X} = \frac{\sigma^2}{n}$$

Thus we see that the sample mean $\bar{X}$ has the same expected value as the individual data value, but its variance is much smaller, i.e.

We assume that value of the distribution we are looking for associated to any member of the population has mean μ (*population mean*) and variance σ^2 (*population variance*). Let $X_1, X_2, \dots, X_n$ be a sample of values from the population. Then the sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

NOTE: $\bar{X}$ is a random variable!. It is good point to remark that

$$\mathbb{E}\bar{X} = \mu \text{ and } Var\bar{X} = \frac{\sigma^2}{n}$$

Thus we see that the sample mean $\bar{X}$ has the same expected value as the individual data value, but its variance is much smaller, i.e. $\bar{X}$ for large n becomes more and more concentrated around μ.

Central Limit Theorem (and restatement for sample mean).

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having common distribution and let $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2$. Then, the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$. That is

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < a\right) \to P(N(0,1) < a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx$$

as $n \to \infty$.

Central Limit Theorem (and restatement for sample mean).

Central Limit Theorem

Let $X_1, X_2, \dots$ be a sequence of independent random variables having common distribution and let $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2$. Then, the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$. That is

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < a\right) \to P(N(0,1) < a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}}\, dx$$

as $n \to \infty$.

We notice that $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ and thus

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having common distribution and let $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2$. Then, the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$. That is

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < a\right) \to P(N(0,1) < a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx$$

as $n \to \infty$.

We notice that $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ and thus

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem for Sample Mean

Let $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$.

Central Limit Theorem (and restatement for sample mean).

Sample mean: $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ thus

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem for Sample Mean

Let $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$.

Sample mean: $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ thus

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem for Sample Mean

Let $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution $N(0, 1)$ as $n \to \infty$.

Thus if $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2 (the *Standard Deviation* σ),

Central Limit Theorem (and restatement for sample mean).

Sample mean: $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$ thus

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Central Limit Theorem for Sample Mean

Let $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends to the standard normal distribution $N(0,1)$ as $n \to \infty$.

Thus if $\bar{X}$ be a sample mean of a sample of size n from a population having mean μ and variance σ^2 (the *Standard Deviation* σ), then

$$P(\bar{X} \leq a) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right) \approx P\left(N(0,1) \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right).$$

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Solution: In this case $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$.

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Solution: In this case $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Thus

$$\frac{\bar{X} - 202}{7/3}$$

can be approximated by $N(0,1)$.

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Solution: In this case $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Thus

$$\frac{\bar{X} - 202}{7/3}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Solution: In this case $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Thus

$$\frac{\bar{X} - 202}{7/3}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198-202}{7/3} \leq \frac{\bar{X}-202}{7/3} \leq \frac{206-202}{7/3}\right)$$

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

Solution: In this case $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{36} = 7/3$. Thus

$$\frac{\bar{X} - 202}{7/3}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198-202}{7/3} \leq \frac{\bar{X}-202}{7/3} \leq \frac{206-202}{7/3}\right)$$

$$= P\left(-1.714 \leq \frac{\bar{X}-202}{7/3} \leq 1.714\right) \approx P\left(-1.714 \leq N(0,1) \leq 1.714\right) = .913$$

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Solution: Now $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{64} = 7/4$.

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Solution: Now $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{64} = 7/4$. Thus

$$\frac{\bar{X} - 202}{7/4}$$

can be approximated by $N(0, 1)$.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Solution: Now $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{64} = 7/4$. Thus

$$\frac{\bar{X} - 202}{7/4}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198 - 202}{7/4} \leq \frac{\bar{X} - 202}{7/4} \leq \frac{206 - 202}{7/4}\right)$$

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Solution: Now $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{64} = 7/4$. Thus

$$\frac{\bar{X} - 202}{7/4}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198 - 202}{7/4} \leq \frac{\bar{X} - 202}{7/4} \leq \frac{206 - 202}{7/4}\right)$$

$$= P\left(-2.286 \leq \frac{\bar{X} - 202}{7/3} \leq 2.286\right) \approx P\left(-2.286 \leq N(0,1) \leq 2.286\right) = .978$$

Example.

The blood cholesterol levels of a population of workers have mean 202 and the standard deviation 14.

- If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.
- How would this question change if the sample size would be 64?

Solution: Now $\mu = 202$ and $\sigma/\sqrt{n} = 14/\sqrt{64} = 7/4$. Thus

$$\frac{\bar{X} - 202}{7/4}$$

can be approximated by $N(0,1)$. So to compute $P(198 \leq \bar{X} \leq 206)$ we must rewrite this inequality in standardized form:

$$P(198 \leq \bar{X} \leq 206) = P\left(\frac{198-202}{7/4} \leq \frac{\bar{X}-202}{7/4} \leq \frac{206-202}{7/4}\right)$$

$$= P\left(-2.286 \leq \frac{\bar{X}-202}{7/3} \leq 2.286\right) \approx P\left(-2.286 \leq N(0,1) \leq 2.286\right) = .978$$

So as we were expected, when we increase the sample size the probability that the sample mean will be close to the population mean increases!

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star.

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it.

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance.

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Solution: Our goal is to estimate

$$P(-.5 < \bar{X} - d < .5),$$

where $\bar{X}$ is the sample mean of the 10 measurements.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Solution: Our goal is to estimate

$$P(-.5 < \bar{X} - d < .5),$$

where $\bar{X}$ is the sample mean of the 10 measurements. We need to rewrite it in form of

$$\frac{\bar{X} - d}{3/\sqrt{10}}.$$

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Solution: Our goal is to estimate

$$P(-.5 < \bar{X} - d < .5),$$

where $\bar{X}$ is the sample mean of the 10 measurements. We need to rewrite it in form of

$$\frac{\bar{X} - d}{3/\sqrt{10}}.$$

Thus

$$P(-.5 < \bar{X} - d < .5) = P\left(\frac{-.5}{3/\sqrt{10}} < \frac{\bar{X} - d}{3/\sqrt{10}} < \frac{.5}{3/\sqrt{10}}\right)$$

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Solution: Our goal is to estimate

$$P(-.5 < \bar{X} - d < .5),$$

where $\bar{X}$ is the sample mean of the 10 measurements. We need to rewrite it in form of

$$\frac{\bar{X} - d}{3/\sqrt{10}}.$$

Thus

$$P(-.5 < \bar{X} - d < .5) = P\left(\frac{-.5}{3/\sqrt{10}} < \frac{\bar{X} - d}{3/\sqrt{10}} < \frac{.5}{3/\sqrt{10}}\right)$$

$$\approx P(-.527 < N(0,1) < .527) = .402$$

Example.

An astronomer is interested in measuring, in units of light-years, the distance from her observatory to a distant star. However, the astronomer knows that due to differing atmospheric conditions and normal errors, each time a measurement is made, it will yield not the exact distance, but an estimate of it. As a result, she is planning on making a series of 10 measurements and using the average of these measurements as her estimated value for actual distance. If the value of the measurements constitute a sample from population having mean d (the actual **unknown !** distance) and a standard deviation of 3 light-years, approximate the probability that the astronomer's estimated value of the distance will be within .5 light-years of the actual distance?

Solution: Our goal is to estimate

$$P(-.5 < \bar{X} - d < .5),$$

where $\bar{X}$ is the sample mean of the 10 measurements. We need to rewrite it in form of

$$\frac{\bar{X} - d}{3/\sqrt{10}}.$$

Thus

$$P(-.5 < \bar{X} - d < .5) = P\left(\frac{-.5}{3/\sqrt{10}} < \frac{\bar{X} - d}{3/\sqrt{10}} < \frac{.5}{3/\sqrt{10}}\right)$$

$$\approx P(-.527 < N(0,1) < .527) = .402$$

So with 10 measurements there is about 40 percent chance that the estimates distance will be within .5 light -years of the actual distance.