STA286 Lecture 25

Neil Montgomery

Last edited: 2017-03-23 11:01

watching even more paint dry

The paint company wants to estimate the mean paint drying time to within a margin of error of 10 minutes, with 95% confidence.

watching even more paint dry

The paint company wants to estimate the mean paint drying time to within a margin of error of 10 minutes, with 95% confidence.

Sample size requirement: The formula is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

watching even more paint dry

The paint company wants to estimate the mean paint drying time to within a margin of error of 10 minutes, with 95% confidence.

Sample size requirement: The formula is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

We could use our best guess from the available information $s = 0.971$ (in hours) in place of σ in the calculation:

$$\left(\frac{1.96 \cdot 0.971}{10/60}\right)^2 = 130.369$$

watching even more paint dry

The paint company wants to estimate the mean paint drying time to within a margin of error of 10 minutes, with 95% confidence.

Sample size requirement: The formula is:

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

We could use our best guess from the available information $s = 0.971$ (in hours) in place of σ in the calculation:

$$\left(\frac{1.96 \cdot 0.971}{10/60}\right)^2 = 130.369$$

So just to bother you, I'll use $n = 130$.

gather a sample of size $n = 130$

Here is a relevant summary of the dataset:

x_bar	s	n
4.14	1.04	130

gather a sample of size $n = 130$

Here is a relevant summary of the dataset:

x_bar	s	n
4.14	1.04	130

From the t_{129} distribution we get $t_{129,0.025} = 1.979$. So the 95% confidence interval is:

$$\overline{x} \pm t_{129,0.025} \frac{s}{\sqrt{n}} = 4.143 \pm 1.979 \frac{1.037}{\sqrt{130}}$$

gather a sample of size $n = 130$

Here is a relevant summary of the dataset:

x_bar	s	n
4.14	1.04	130

From the t_{129} distribution we get $t_{129,0.025} = 1.979$. So the 95% confidence interval is:

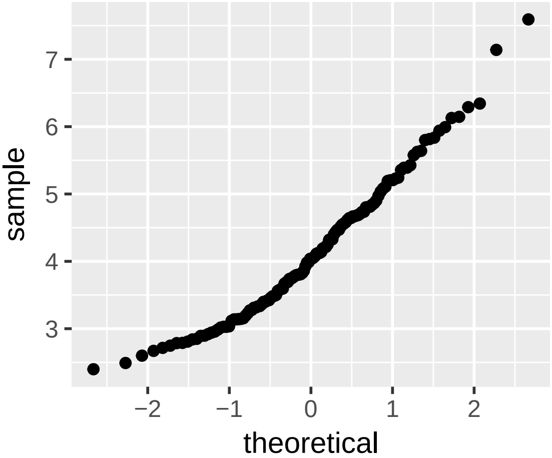
$$\overline{x} \pm t_{129,0.025} \frac{s}{\sqrt{n}} = 4.143 \pm 1.979 \frac{1.037}{\sqrt{130}}$$

or

$$[3.963, 4.323]$$

verifying the model assumption(s)

In this case there is only one assumption (that can be verified)—that the underlying distribution is normal. Here is a normal quantile plot of the data:



model assumption conclusion | robustness of "t procedure"

The normal distribution assumption has been violated. It seems the underlying distribution is skewed right.

However, the sample size $n = 130$ is large, so by the speed of convergence of the CLT and its buddy Mr. Slutsky, we're still OK.

model assumption conclusion | robustness of "t procedure"

The normal distribution assumption has been violated. It seems the underlying distribution is skewed right.

However, the sample size $n = 130$ is large, so by the speed of convergence of the CLT and its buddy Mr. Slutsky, we're still OK.

Dirty secret: I simulated the data to get the example sample of size 130.

model assumption conclusion | robustness of "t procedure"

The normal distribution assumption has been violated. It seems the underlying distribution is skewed right.

However, the sample size $n = 130$ is large, so by the speed of convergence of the CLT and its buddy Mr. Slutsky, we're still OK.

Dirty secret: I simulated the data to get the example sample of size 130.

As an example of what is called the "robustness" of this confidence interval against violations of the normality assumption, I did a quick simulation (code embedded in notes).

model assumption conclusion | robustness of "t procedure"

The normal distribution assumption has been violated. It seems the underlying distribution is skewed right.

However, the sample size $n = 130$ is large, so by the speed of convergence of the CLT and its buddy Mr. Slutsky, we're still OK.

Dirty secret: I simulated the data to get the example sample of size 130.

As an example of what is called the "robustness" of this confidence interval against violations of the normality assumption, I did a quick simulation (code embedded in notes).

The proportion of the 10^4 simulated confidence intervals that captured the true mean is (for this simulation—changes every time I render the lecture notes):

$$0.9444$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\mathrm{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\mathrm{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\mathrm{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\text{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\mathrm{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\mathrm{Var}\left(\overline{X} - X\right)$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\text{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\text{Var}\left(\overline{X} - X\right) = \text{Var}\left(\overline{X}\right) + \text{Var}(X)$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\mathrm{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\mathrm{Var}\left(\overline{X} - X\right) = \mathrm{Var}\left(\overline{X}\right) + \mathrm{Var}(X) = \frac{\sigma^2}{n} + \sigma^2$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\text{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\text{Var}\left(\overline{X} - X\right) = \text{Var}\left(\overline{X}\right) + \text{Var}(X) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

prediction, as opposed to estimation

To get the interval estimate of μ we used the fact that $\overline{X} - \mu$ is normal with variance $\text{Var}\left(\overline{X} - \mu\right) = \sigma^2/n$ to obtain $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, etc.

Suppose instead we want to predict the next actual value of the random process under consideration.

For example, we open a new can of paint. How long will *this* paint take to dry?

We can use the sample $X_1, \ldots, X_n$ to make the prediction. Simply use $\overline{X}$. But the difference between predition and actual is now $\overline{X} - X$.

The variance of this expression is:

$$\text{Var}\left(\overline{X} - X\right) = \text{Var}\left(\overline{X}\right) + \text{Var}(X) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

If the population is normal, so will be $\overline{X} - X$, and its mean will be $E\left(\overline{X} - X\right) = \mu - \mu = 0$

prediction "interval"

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0, 1)$$

prediction "interval"

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0, 1)$$

Deal with unknown σ right now—replace it with S from the sample, to get:

$$\frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} \sim$$

prediction "interval"

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0,1)$$

Deal with unknown σ right now—replace it with S from the sample, to get:

$$\frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} \sim t$$

prediction "interval"

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0, 1)$$

Deal with unknown σ right now—replace it with S from the sample, to get:

$$\frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

prediction "interval"

Put it all together to get:

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}} \sim N(0, 1)$$

Deal with unknown σ right now—replace it with S from the sample, to get:

$$\frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} \sim t_{n-1}$$

A $100 \cdot (1 - \alpha)\%$ *prediction interval* can be obtained by solving for X in:

$$P\left(-t_{n-1,\alpha/2} < \frac{\overline{X} - X}{S\sqrt{1 + \frac{1}{n}}} < t_{n-1,\alpha/2}\right)$$

prediction interval example

The formula is:

$$\overline{X} \pm t_{n-1,\alpha/2} S \sqrt{1 + \frac{1}{n}}$$

prediction interval example

The formula is:

$$\overline{X} \pm t_{n-1,\alpha/2} S \sqrt{1 + \frac{1}{n}}$$

Just that little '1' under the square root—but it makes all the difference. It guarantees that the prediction can never be better than the variance in the population itself, no matter what the sample size.

prediction interval example

The formula is:

$$\overline{X} \pm t_{n-1,\alpha/2} S \sqrt{1 + \frac{1}{n}}$$

Just that little '1' under the square root—but it makes all the difference. It guarantees that the prediction can never be better than the variance in the population itself, no matter what the sample size.

Using the paint example, with $t_{129,0.025} = 1.97852$ and

x_bar	s	n
4.14	1.04	130

in the formula gives:

$$4.143 \pm 1.979 \cdot 1.037 \sqrt{1 + \frac{1}{130}} \qquad \text{or} \qquad [2.084, 6.203]$$

prediction interval model assumptions

Normal population is the only assumption.

Suppose the population is not normal. What might happen to the following as n gets large?

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}}$$

prediction interval model assumptions

Normal population is the only assumption.

Suppose the population is not normal. What might happen to the following as n gets large?

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}}$$

There is no way of knowing. X just sits there in the numerator, with properties that never change no matter what the sample size.

prediction interval model assumptions

Normal population is the only assumption.

Suppose the population is not normal. What might happen to the following as n gets large?

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}}$$

There is no way of knowing. X just sits there in the numerator, with properties that never change no matter what the sample size.

So the population really has to be normal, or the P.I. formula doesn't work.

prediction interval model assumptions

Normal population is the only assumption.

Suppose the population is not normal. What might happen to the following as n gets large?

$$\frac{\overline{X} - X}{\sigma\sqrt{1 + \frac{1}{n}}}$$

There is no way of knowing. X just sits there in the numerator, with properties that never change no matter what the sample size.

So the population really has to be normal, or the P.I. formula doesn't work.

The paint drying P.I. we calculated is therefore not that useful.

the two-sample problem (normal populations)

We've solved the case of one numerical variable in a dataset with a normal population.

Often you'll have a numerical variable in one column, and a "grouping" variable in another column that categorizes the observations into two groups.

Variable	Group
3.85	2
6.06	2
3.28	1
4.85	2
5.34	1
6.03	2
$\vdots$	$\vdots$

the two-sample problem (normal populations)

We've solved the case of one numerical variable in a dataset with a normal population.

Often you'll have a numerical variable in one column, and a "grouping" variable in another column that categorizes the observations into two groups.

Variable	Group
3.85	2
6.06	2
3.28	1
4.85	2
5.34	1
6.03	2
$\vdots$	$\vdots$

Variable	Group
X_{21}	2
X_{22}	2
X_{11}	1
X_{23}	2
X_{12}	1
X_{24}	2
$\vdots$	$\vdots$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$E\left(\overline{X}_1 - \overline{X}_2\right)$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$E\left(\overline{X}_1 - \overline{X}_2\right) = \mu_1 - \mu_2$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$
\begin{aligned}
E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\
\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right)
\end{aligned}
$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$
\begin{aligned}
E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\
\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \mathrm{Var}\left(\overline{X}_1\right) + \mathrm{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}
\end{aligned}
$$

the two-sample problem (normal populations) with equal variances

We have two populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, and the goal is to estimate $\theta = \mu_1 - \mu_2$.

Gather independent samples: $X_{11}, \ldots, X_{1n_1}$ i.i.d. $N(\mu_1, \sigma)$ and $X_{21}, \ldots, X_{2n_2}$ i.i.d. $N(\mu_2, \sigma)$.

The "obvious" estimator is $\overline{X}_1 - \overline{X}_2$, with the following properties:

$$
\begin{aligned}
E\left(\overline{X}_1 - \overline{X}_2\right) &= \mu_1 - \mu_2 \\
\mathrm{Var}\left(\overline{X}_1 - \overline{X}_2\right) &= \mathrm{Var}\left(\overline{X}_1\right) + \mathrm{Var}\left(\overline{X}_2\right) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)
\end{aligned}
$$

We need to figure out what to do about σ^2.