

Recall that last time we considered the distribution of cholesterol levels for US men aged 20 - 74. We assumed the mean to be 211 mg/100ml and the standard deviation to be 46 mg/100ml. We hypothetically selected a sample of size $n = 25$ from the population, calculated the sample mean \bar{x}_{25} , and wanted to know if it looked like an outlier.

To answer this, we figured out the interval that enclosed 95% of the sample means and then we could see if the sample mean that we were interested in (\bar{x}_{25}) fell inside or outside this range.

For general n , we want to find x_l and x_u such that

$$P[x_l \leq \bar{X}_n \leq x_u] = 0.95. \text{ We know } P[-1.96 \leq Z \leq 1.96] = 0.95$$

For $n = 25$,

$$\begin{aligned} P[x_l \leq \bar{X} \leq x_u] &= P[-1.96 \leq \frac{\bar{X} - 211}{46/\sqrt{25}} \leq 1.96] \\ &= P[-1.96 \cdot \left(\frac{46}{\sqrt{25}}\right) \leq \bar{X} - 211 \leq 1.96 \cdot \left(\frac{46}{\sqrt{25}}\right)] \\ &= P[211 - 1.96 \cdot 9.2 \leq \bar{X} \leq 211 + 1.96 \cdot 9.2] \\ &= P[192.97 \leq \bar{X} \leq 229.03] \end{aligned}$$

Suppose that we collect a sample of 25 Japanese men aged 20-74 and we record the cholesterol level of each. Suppose the sample mean is $\bar{x} = 189$. Is this compatible with a true mean of 211? Recall that 211 was the mean of the cholesterol levels for the population of US men aged 20-74.

n	$\frac{\sigma}{\sqrt{n}}$	Int. Enclosing 95% of Sample means	Length
1	46.0	$120.8 \leq \bar{X} \leq 301.2$	180.4
10	14.5	$182.5 \leq \bar{X} \leq 239.5$	57.0
25	9.2	$193.0 \leq \bar{X} \leq 229.0$	36.0
50	6.5	$198.2 \leq \bar{X} \leq 223.8$	25.6
100	4.6	$202.0 \leq \bar{X} \leq 220.0$	18.0

Notes:

As the size of the sample increases, the amount of variability among the sample means (quantified by $\frac{\sigma}{\sqrt{n}}$) decreases.

The interval enclosing 95 % of the sample means gets smaller.

Estimation

Now that we understand sampling distributions of statistics, we are equipped to study an important area of statistical inference, *parameter estimation*.

A *point estimate* is a single numerical value used to estimate the corresponding population parameter.

For example, \bar{X} is a point estimate of μ . We'd like to know how close the estimator is to the true value.

We saw that statistics such as \bar{X} are random variables with probability distributions. By finding intervals associated with high probability in the sampling distribution, we can find intervals likely to contain the population parameter of interest.

For the calculations that we started last last time, we knew that the population mean was 211 and the population standard deviation was 46. Recall that we don't know how the data values are distributed. Suppose now that we don't know the population mean.

From the Central Limit Theorem, we know that \bar{X} , calculated from a sample of size n , is approximately Normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

How large does n have to be so that 95 % of the sample means are within ± 5 from the true population mean ?

Confidence interval for a population mean

Let's assume that we are sampling from a normal distribution, or that n is large enough that we may assume the distribution of \bar{X} is approximately normal (central limit theorem).

Recall that the mean of \bar{X} is μ and the standard deviation of \bar{X} is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. It is common to refer to $\frac{\sigma}{\sqrt{n}}$ as the *standard error*.

Find

$$P[-1.96 \leq \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \leq 1.96]$$

By now, we know that the distribution of $(\bar{X} - \mu)/\frac{\sigma}{\sqrt{n}}$ is approximately standard normal for large sample sizes n , so $P[-1.96 \leq \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \leq 1.96] = 0.95$.

Knowing this, let's rearrange the inequality a bit.

$$\begin{aligned} 0.95 &= P[-1.96 \leq \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \leq 1.96] \\ &= P[-1.96 \frac{\sigma}{\sqrt{n}} \leq (\bar{X} - \mu) < 1.96 \frac{\sigma}{\sqrt{n}}] \\ &= P[-1.96 \frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq 1.96 \frac{\sigma}{\sqrt{n}} - \bar{X}] \\ &= P[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}] \end{aligned}$$

Thus, the interval $(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$ will include the population mean μ in 95% of repeated random samples.

Accordingly, $(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$ is called *confidence interval* for μ with *confidence level* 0.95.

The width of a confidence level is $2 \times 1.96 \frac{\sigma}{\sqrt{n}}$ and $1.96 \frac{\sigma}{\sqrt{n}}$ is sometimes called the *precision* of the estimate or the *margin of error*.

Notation

Define the values $z_{(\alpha/2)}$ and $-z_{(\alpha/2)}$ as follows:

$$P[Z > z_{(\alpha/2)}] = \alpha/2 \text{ and } P[Z < -z_{(\alpha/2)}] = \alpha/2$$

where Z is a standard normal random variable.

Find

$$P[-z_{(\alpha/2)} < (\bar{X} - \mu)/\sigma_{\bar{x}} < z_{(\alpha/2)}]$$

Confidence Intervals

We know that the distribution of $(\bar{X} - \mu)/\sigma_{\bar{x}}$ is approximately standard normal for large sample sizes n ,

so $P[-z_{(\alpha/2)} < (\bar{X} - \mu)/\sigma_{\bar{x}} < z_{(\alpha/2)}] = 1 - \alpha$.

Knowing this, let's rearrange the inequality a bit.

$$\begin{aligned} 1 - \alpha &= P[-z_{(\alpha/2)} < (\bar{X} - \mu)/\sigma_{\bar{x}} < z_{(\alpha/2)}] \\ &= P[-z_{(\alpha/2)}\sigma_{\bar{x}} < (\bar{X} - \mu) < z_{(\alpha/2)}\sigma_{\bar{x}}] \\ &= P[-z_{(\alpha/2)}\sigma_{\bar{x}} - \bar{X} < -\mu < z_{(\alpha/2)}\sigma_{\bar{x}} - \bar{X}] \\ &= P[\bar{X} - z_{(\alpha/2)}\sigma_{\bar{x}} < \mu < \bar{X} + z_{(\alpha/2)}\sigma_{\bar{x}}] \end{aligned}$$

Thus, the interval

$$(\bar{X} - z_{(\alpha/2)}\sigma_{\bar{x}}, \bar{X} + z_{(\alpha/2)}\sigma_{\bar{x}})$$

will include the population mean μ in $100 \times (1 - \alpha)\%$ of all random samples.

Accordingly,

$$(\bar{X} - z_{(\alpha/2)}\sigma_{\bar{x}}, \bar{X} + z_{(\alpha/2)}\sigma_{\bar{x}})$$

is called *confidence interval* for μ with *confidence level* $1 - \alpha$.

Consider the confidence interval for μ with confidence level $1 - \alpha$:

$$(\bar{X} - z_{(\alpha/2)}\sigma_{\bar{x}}, \bar{X} + z_{(\alpha/2)}\sigma_{\bar{x}})$$

Notes:

1. \bar{X} is a random variable, μ is fixed.
2. Given a 95 % confidence interval for μ , we say that we are 95 % confident that the interval will contain μ . For example, consider 100 random samples of size n . For each sample, we can calculate a sample mean and construct the associated confidence interval. We expect 95 of these intervals to contain μ .
3. The interval shown is random and has a 95 % chance of covering μ *before* a sample is selected. Since μ is fixed (NOT random), once a sample has been drawn and an interval constructed, either μ is within the interval or it is not.

Figure to remember...

Q: How does sample size affect the length of the confidence interval ?

Q: How does confidence level affect the length of the confidence interval ?

Consider again the distribution of cholesterol levels for US men aged 20 - 74. Assume the mean $\mu = 211$ mg/100ml and the standard deviation is $\sigma = 46$ mg/100ml.

n	$\frac{\sigma}{\sqrt{n}}$	95% Confidence Interval for μ	Length
n	$\frac{\sigma}{\sqrt{n}}$	$(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$	$2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{n}}$
1	46.0	$(\bar{X} - 1.96 \cdot 46, \bar{X} + 1.96 \cdot 46)$	180.4
10	14.5	$(\bar{X} - 1.96 \cdot 14.5, \bar{X} + 1.96 \cdot 14.5)$	57.0
25	9.2	$(\bar{X} - 1.96 \cdot 9.2, \bar{X} + 1.96 \cdot 9.2)$	36.0
50	6.5	$(\bar{X} - 1.96 \cdot 6.5, \bar{X} + 1.96 \cdot 6.5)$	25.6
100	4.6	$(\bar{X} - 1.96 \cdot 4.6, \bar{X} + 1.96 \cdot 4.6)$	18.0

Suppose our sample is of size 25. Then, $\frac{\sigma}{\sqrt{n}} = 9.2$

Confidence Level	Confidence Interval for μ	Length
$1 - \alpha$	$(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$	$2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
0.80	$(\bar{X} - 1.28 \cdot 9.2, \bar{X} + 1.28 \cdot 9.2)$	23.6
0.90	$(\bar{X} - 1.645 \cdot 9.2, \bar{X} + 1.645 \cdot 9.2)$	30.3
0.95	$(\bar{X} - 1.96 \cdot 9.2, \bar{X} + 1.96 \cdot 9.2)$	36.0
0.99	$(\bar{X} - 2.56 \cdot 9.2, \bar{X} + 2.56 \cdot 9.2)$	47.1

One-sided Confidence Interval

Perhaps we are interested in finding an upper bound for some true population mean μ . For example, we might want to find some value x_U such that

$$P(\mu < x_U) = 1 - \alpha$$

Since $P(Z < -z_\alpha) = \alpha$, $P(Z \geq -z_\alpha) = 1 - \alpha$. If n is large enough, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is approximately standard normal. So,

$$\begin{aligned} 1 - \alpha &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq -z_\alpha\right) \\ &= P(-\mu \geq -\bar{X} - z_\alpha \sigma / \sqrt{n}) \\ &= P(\mu \leq \bar{X} + z_\alpha \sigma / \sqrt{n}) \end{aligned}$$

Consider again the distribution of cholesterol levels for US men aged 20 - 74 (with standard deviation $\sigma = 46$ mg/100ml). Suppose we don't know the mean and we take a sample of size 25 and want a 95 % one-sided confidence interval for μ . Suppose the sample mean is 217.

$$P(\mu \leq 217 + 1.65 \cdot 9.2) = P(\mu \leq 232.8) \neq 0.95$$

So far, we have derived confidence intervals for μ using the knowledge that

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

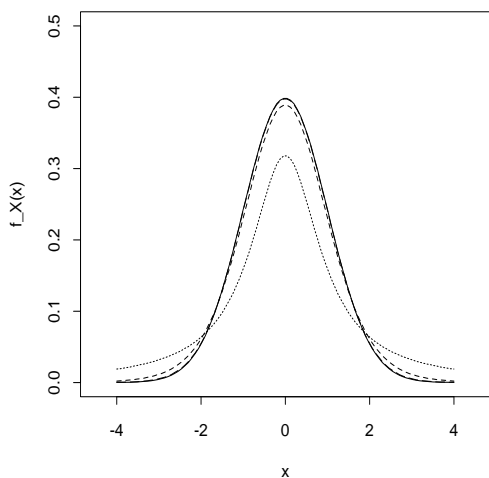
has a standard normal distribution. However, this requires knowing the population standard deviation σ .

In most applications, we need to estimate σ with the sample standard deviation,

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

If we are sampling from a normal population, what is the distribution of

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



Because the constant σ has been replaced with its estimate s , the distribution is no longer standard normal.

Instead,

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

has a t distribution.

In fact, there is a whole family of t distributions, and they depend on a parameter called *degrees of freedom*.

The degrees of freedom associated with the t -distributed random variable

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

is $n - 1$.

When sampling from a normal distribution with an unknown standard deviation, a $100(1 - \alpha)$ percent confidence interval for the population mean μ is given by,

$$\left(\bar{X} - t_{(n-1),(\alpha/2)} \frac{s}{\sqrt{n}}, \bar{X} + t_{(n-1),(\alpha/2)} \frac{s}{\sqrt{n}} \right),$$

where $t_{(n-1),(\alpha/2)}$ refers to the value above which $\alpha/2$ % of the area of the t distribution (with $n - 1$ degrees of freedom) lies.

Consider again the distribution of cholesterol levels for US men aged 20 - 74. Assume the mean $\mu = 211$ mg/100ml and the standard deviation is $\sigma = 46$ mg/100ml.

n	$\frac{\sigma}{\sqrt{n}}$	95% Confidence Interval for μ	Length
n	$\frac{\sigma}{\sqrt{n}}$	$(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}})$	$2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
10	14.5	$(\bar{X} - 1.96 \cdot 14.5, \bar{X} + 1.96 \cdot 14.5)$	57.0
25	9.2	$(\bar{X} - 1.96 \cdot 9.2, \bar{X} + 1.96 \cdot 9.2)$	36.0
50	6.5	$(\bar{X} - 1.96 \cdot 6.5, \bar{X} + 1.96 \cdot 6.5)$	25.6
100	4.6	$(\bar{X} - 1.96 \cdot 4.6, \bar{X} + 1.96 \cdot 4.6)$	18.0
200	3.3	$(\bar{X} - 1.96 \cdot 3.3, \bar{X} + 1.96 \cdot 3.3)$	12.94

Suppose we don't know the standard deviation, but it just so happens that in each sample we take, the sample standard deviation, s , is 46. What would the confidence intervals look like? Recall that since we had to estimate the true standard deviation by the sample standard deviation, we have $\frac{\bar{X} - \mu}{s/\sqrt{n}}$, which is not standard normal.

n	$\frac{s}{\sqrt{n}}$	95% Confidence Interval for μ	Length
n	$\frac{s}{\sqrt{n}}$	$(\bar{X} - t_{(n-1),(\alpha/2)} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{(n-1),(\alpha/2)} \cdot \frac{s}{\sqrt{n}})$	$2 \cdot t_{(n-1),(\alpha/2)} \cdot \frac{s}{\sqrt{n}}$
10	14.5	$(\bar{X} - 2.26 \cdot 14.5, \bar{X} + 2.26 \cdot 14.5)$	65.5
25	9.2	$(\bar{X} - 2.06 \cdot 9.2, \bar{X} + 2.06 \cdot 9.2)$	37.9
50	6.5	$(\bar{X} - 2.01 \cdot 6.5, \bar{X} + 2.01 \cdot 6.5)$	26.13
100	4.6	$(\bar{X} - 1.98 \cdot 4.6, \bar{X} + 1.98 \cdot 4.6)$	18.2
200	3.3	$(\bar{X} - 1.97 \cdot 3.3, \bar{X} + 1.97 \cdot 3.3)$	13.0

For the population of infants subjected to fetal surgery for congenital anomalies, the distribution of gestational ages at birth is approximately normal with unknown mean μ and unknown standard deviation σ . A random sample of 14 such infants has sample mean $\bar{x} = 29.6$ weeks and standard deviation $\sigma = 3.6$ weeks.

1. Construct a 95 % confidence interval for the true population mean μ .
2. What is the length of the interval ?
3. What is the margin of error ?
4. How large a sample would be required for the 95 % CI to have length 3 weeks. For this one, assume that the population standard deviation is known and is $\sigma = 3.9$ weeks.
5. How large a sample would be needed for the 95 % CI to have length 3 weeks if σ is not known ?