Statistics: Interval Estimates

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2017-2018

Outline

Introduction

- Immediately after the closure of an election, one can watch on TV the first exit polls and, later, the projections obtained by scrutinized votes.

  We can hear statements as: "the party A has 32.4% of the votes with a margin of error of $\pm 2.2\%$".

  What does it mean this statement?

  Moreover, how is it possible do such a prevision only by a small number of interviews (exit polls) with respect to the number of voters or by analyzing only a small part of the votes (projections) with respect to the total number of votes?

  Here, we answer questions like these.

- One of the tasks of the Inferential Statistics is to estimate unknown parameters of a distribution belonging to a given family, by using sampling statistics on a sample of observed values from that distribution.

We recall the definition of an estimator.

Definition

Let θ be one of the unknown parameter of the distribution. An **estimator** of θ is a sampling statistic $\tau : \mathbb{R}^n \to \mathbb{R}$ providing an estimate of θ: if $\boldsymbol{x}^{\text{obs}} = \left(x_1^{\text{obs}}, x_2^{\text{obs}}, \ldots, x_n^{\text{obs}}\right)$ is the sample, i.e. the $n-$tuple of the observed values of the Sample $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, then $\tau\left(\boldsymbol{x}^{\text{obs}}\right)$ is the estimate of θ.

We have already seen examples of estimators: the sample mean is an estimator of the mean, if the mean is one of the unknown parameters, and the sample variance is an estimator of the variance, if the variance is one of the unknown parameters.

Example. Consider a population of notebook batteries and assume that the lifespan of such a battery is a random variable normally distributed of unknown mean.

Suppose that we are interested in estimating such an unknown mean.

We take a sample of the population and compute the sample mean of the observed lifespans of the individuals in the sample (recall that we call "sample" both these observed lifespans and the selected individuals).

This sample mean (it could be 3.5 *years) is the estimate of the unknown mean of the normal distribution.*

- Now, we introduce the concept of an unbaised estimator.

Definition

An estimator τ of the parameter θ is said **unbiased** if $\mathbb{E}(\tau(\boldsymbol{X})) = \theta$, where $\tau(\boldsymbol{X})$, with $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ the Sample, is the Estimator (Sampling Statistic) relevant to the estimator (sampling statistic) τ.

We have seen that the sample mean is an unbiased estimator of the mean and the sample variance is an unbiased estimator of the variance.

Also

$$\tau(\boldsymbol{x}) = \frac{\sum_{k=1}^{n} (x_k - \mu)^2}{n}, \ \boldsymbol{x} \in \mathbb{R}^n,$$

is an unbiased estimator of the variance, to be used when the mean μ is known.

On the other hand,

$$\tau(\boldsymbol{x}) = \frac{\sum_{k=1}^{n} (x_k - \overline{x})^2}{n}, \ \boldsymbol{x} \in \mathbb{R}^n,$$

is an estimator of the variance, but it is not unbiased: its mean is $\frac{n-1}{n}\sigma^2$, not σ^2, where σ^2 is the variance,

Interval Estimates

- An estimate given by the value of an estimator is called a **point estimate**.

  On the other hand, in the above situation of the election, the statement: "the party A has 32.4% of the votes with a margin of error of $\pm 2.2\%$" gives an example of an **interval estimate**.

  Here, we are saying that the percentage of the votes for the party A is in the interval between $32.4\% - 2.2\% = 30.2\%$ and $32.4\% + 2.2\% = 34.6\%$.

  At each interval estimate, we attach a **level of confidence** for that interval, i.e. the probability that the interval contains the parameter.

Definition

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a Sample from a distribution with an unknown parameter $\theta \in \mathbb{R}$.

If τ is an estimator of θ and $\boldsymbol{x}^{\text{obs}}$ is the sample, then $\tau\left(\boldsymbol{x}^{\text{obs}}\right)$ is the **point estimate** of θ.

An **interval estimate** of θ with **level of confidence** $C\% \in [0, 1]$, called a $C\%$ **confidence interval** of θ, is an interval $I\left(\boldsymbol{x}^{\text{obs}}\right)$, which depends on the sample $\boldsymbol{x}^{\text{obs}}$, such that

$$\mathbb{P}\left(\theta \in I\left(\boldsymbol{X}\right)\right) = C\%.$$

Observe that the level of confidence is the probability that the random Interval $I\left(\boldsymbol{X}\right)$ contains the nonrandom (but unknown) parameter θ.

- How to obtain an interval estimate of θ?

  Consider an unbiased estimator τ of θ and let σ_n be the standard deviation of the Estimator $\tau(\boldsymbol{X})$. σ_n is called the **standard error** of the estimator τ.

  We consider interval estimates of θ of the form

$$I(\boldsymbol{x}^{\text{obs}}) = \tau\left(\boldsymbol{x}^{\text{obs}}\right) \pm k\sigma_n := \left(\tau\left(\boldsymbol{x}^{\text{obs}}\right) - k\sigma_n, \tau\left(\boldsymbol{x}^{\text{obs}}\right) + k\sigma_n\right),$$

  where $k > 0$.

  Such an interval estimate is an interval centered at the point estimate $\tau\left(\boldsymbol{x}^{\text{obs}}\right)$ of radius $k\sigma_n$. This radius is called the **margin of error** of the interval estimate.

- The level of confidence of this interval estimate is

$$\mathbb{P}\left(\theta \in I(\boldsymbol{X}) = \tau\left(\overline{X}\right) \pm k\sigma_n\right) = \mathbb{P}\left(\left|\tau\left(\boldsymbol{X}\right) - \mathbb{E}\left(\tau\left(\boldsymbol{X}\right)\right)\right| < k\sigma_n\right).$$

In fact

$$\begin{aligned}
\mathbb{P}\left(\theta \in \tau\left(\boldsymbol{X}\right) \pm k\sigma_n\right) &= \mathbb{P}\left(\theta \in \left(\tau\left(\boldsymbol{X}\right) - k\sigma_n, \tau\left(\boldsymbol{X}\right) + k\sigma_n\right)\right) \\
&= \mathbb{P}\left(\tau\left(\boldsymbol{X}\right) - k\sigma_n < \theta < \tau\left(\boldsymbol{X}\right) + k\sigma_n\right) \\
&= \mathbb{P}\left(-k\sigma_n < \tau\left(\boldsymbol{X}\right) - \theta < k\sigma_n\right) \\
&= \mathbb{P}\left(\left|\tau\left(\boldsymbol{X}\right) - \theta\right| < k\sigma_n\right) \\
&= \mathbb{P}\left(\left|\tau\left(\boldsymbol{X}\right) - \mathbb{E}\left(\tau\left(\boldsymbol{X}\right)\right)\right| < k\sigma_n\right),
\end{aligned}$$

where the last equality follows from the fact that τ is an unbaised estimator for θ.

The Chebyshev's inequality says that

$$\mathbb{P}\left(\left|\tau\left(\boldsymbol{X}\right) - \mathbb{E}\left(\tau\left(\boldsymbol{X}\right)\right)\right| < k\sigma_n\right) \geq 1 - \frac{1}{k^2}$$

and so when $k = 2$ the level of confidence is at least $1 - \frac{1}{4} = 75\%$ and when $k = 3$ it is at least $1 - \frac{1}{9} = 88.9\%$.

Estimating means

- Consider the situation where the distribution from which the Sample is taken depends on a unique unknown parameter, which is the mean of the distribution.

  Examples are the normal distribution $\mathrm{N}(\mu, \sigma^2)$, with μ unknown and σ known, and the Bernoulli distribution $\mathrm{Bernoulli}\,(p)$, where $p \in [0, 1]$ is unknown.

  As an unbiased estimator of the mean we consider the sample mean $\overline{x}$. The standard error of the sample mean is

$$\sigma_n = \mathrm{SD}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}},$$

  where σ is the standard deviation of the distribution.

  Hence, the interval estimates of the mean have the form

$$\overline{x}^{\mathrm{obs}} \pm k\frac{\sigma}{\sqrt{n}},$$

  where $k > 0$.

- A classic situation, where we have a normal distribution with unknown mean and known standard deviation, is the process of measurement.

  Here, the measured value can be considered as a normal random variable with mean the unknown actual value to be measured and known standard deviation σ, related to the precision of the measure instrument: more precise the instrument, smaller the standard deviation.

Example. Consider the previous example of the astronomer measuring the distance of a star from the Earth.

Suppose that the following ten measures of the distance were obtained:

$$10.2,\ \ 9.6,\ \ 9.7,\ \ 10.1,\ \ 10.0,\ \ 9.8,\ \ 10.2,\ \ 9.8,\ \ 9.6,\ \ 9.8$$

all expressed in light years.

The standard deviation is known to be $\sigma = 0.3$ *light years.*

We can compute the point estimate $\overline{x}^{\text{obs}}$ *by hand. Since*

$$\boldsymbol{x}^{\text{obs}} = 10 + \frac{1}{10} \cdot (2, -4, -3, 1, 0, -2, 2, -2, -4, -2)\,,$$

we have

$$\begin{aligned} \overline{x}^{\text{obs}} &= 10 + \frac{1}{10} \cdot \frac{2 - 4 - 3 + 1 + 0 - 2 + 2 - 2 - 4 - 2}{10} \\ &= 10 - 0.12 = 9.88 \text{ light year.} \end{aligned}$$

Hence, the interval estimates of the actual distance have the form

$$9.88 \pm k\frac{0.3}{\sqrt{10}} = 9.88 \pm 0.09k \text{ light year}, \ k > 0.$$

Of course, each interval estimate has its own level of confidence.

The case of the normal distribution

- In case, as in the previous example, of a normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$, where μ is unknown and σ is known, we can easily determine the level of confidence of an interval estimate.

  In fact, in this case the Sample Mean has the normal distribution

$$\mathrm{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

  Thus, the level of confidence of the interval estimate

$$\overline{x}^{\mathrm{obs}} \pm k\frac{\sigma}{\sqrt{n}},$$

  where $k > 0$, is

$$\mathbb{P}\left(\mu \in I(\boldsymbol{X}) = \overline{X} \pm k\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\left|\overline{X} - \mu\right| < k\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\left|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| < k\right)$$

$$= 2\Phi\left(k\right) - 1.$$

  since $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ has the standard normal distribution.

In particular:

- for $k = 1$, the level of confidence is 68.26%;
- for $k = 2$, the level of confidence is 95.44%;
- for $k = 3$, the level of confidence is 97.74%.

In the example of the astronomer measuring the distance of the star from the Earth, we have the following interval estimates:

- 9.88 ± 0.09 light year with level of confidence 68.26%;
- 9.88 ± 0.18 light year with level of confidence 95.44%;
- 9.88 ± 0.27 light year with level of confidence 97.74%.

- Most of times, we are in a situation where a given level of confidence $C\% \in (0,1)$ is assigned. Then, we can determine $k > 0$ in order to have an interval estimate

$$\overline{x}^{\text{obs}} \pm k\frac{\sigma}{\sqrt{n}}$$

with level of confidence $C\%$.

We have

$$C\% = \mathbb{P}\left(\mu \in I(\boldsymbol{X}) = \overline{X} \pm \frac{\sigma}{\sqrt{n}}\right) = 2\Phi(k) - 1$$

for

$$k = \Phi^{-1}\left(\frac{1 + C\%}{2}\right).$$

So k is the

$$100\frac{1 + C\%}{2} = \frac{100 + C}{2}\text{th percentile}$$

of a standard normal random variable.

Exercise. Show that for $C\% = 90\%, 95\%, 99\%$, we have $k = 1.65, 1.96, 2.58$, respectively.

- Sometimes, we are interested in obtaining an interval estimate

$$\overline{x}^{\text{obs}} \pm k\frac{\sigma}{\sqrt{n}},$$

with a given level of confidence $C\%$ (and so with a given k), such that

$$k\frac{\sigma}{\sqrt{n}} \leq \varepsilon,$$

i.e. the margin of error is not larger than a fixed tolerance $\varepsilon > 0$.

In order to obtain this, it is sufficient to take a Sample of size n such that

$$\sqrt{n} \geq \frac{k\sigma}{\varepsilon},$$

i.e.

$$n \geq \left(\frac{k\sigma}{\varepsilon}\right)^2.$$

Therefore, one can take

$$n = \left\lceil \left(\frac{k\sigma}{\varepsilon} \right)^2 \right\rceil .$$

Exercise. Prove that with this choice, the margin of error $k\frac{\sigma}{\sqrt{n}}$ satisfies

$$\sqrt{1 - \frac{1}{n}} \cdot \varepsilon < k\frac{\sigma}{\sqrt{n}} \leq \varepsilon.$$

Note that, we can obtain interval estimates with an arbitrarily small margin of error ε and with an arbitrarily high level of confidence $C\%$, i.e. with an arbitrarily large k, by taking a Sample of a sufficiently large size n.

But, of course, by taking a Sample of large size has a cost.

Example. Once again, consider the case of the astronomer measuring the distance of the star from the Earth.

Suppose that the astronomer wants to be 99% *confident that the margin of error is not larger than* 0.1 *light years.*

To obtain this. it is sufficient to have

$$n = \left\lceil \left(\frac{k\sigma}{\varepsilon} \right)^2 \right\rceil = \left\lceil \left(\frac{2.58 \cdot 0.3}{0.1} \right)^2 \right\rceil = 60.$$

measurements. Since the astronomical measurement requires a careful preparation, this number is quite large.

Exercise. Find the size n of the Sample in order to have a 99% confidence interval with margin of error not larger than m times the standard deviation σ. Compute n for $m = \frac{1}{10}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2$.

- In the case of a Sample from a general distribution with unknown mean μ, what we have established up to now for a normal distribution continues to be valid, although only approximately and for *n* large.

  In fact, for *n* large, the Sample Mean has approximately the normal distribution

$$\mathbf{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right),$$

  where σ is the standard deviation of this general distribution.

  In particular, this holds for the distribution $\mathrm{Bernoulli}\,(p)$, the subject of the next section.

Estimating proportions

- Consider the problem of estimating proportions, where the proportion p of a characteristic of the individuals in a population has to be estimated.

  In order to do this, a sample of n individuals in the population is selected and then the unknown parameter p is estimated by using the Sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from the distribution $\mathrm{Bernoulli}\,(p)$ given by

$$X_k = \begin{cases} 1 \text{ if the } k-\text{th selected individual has the characteristic} \\ 0 \text{ otherwise} \end{cases}$$

  for $k \in \{1, \ldots, n\}$.

  Since p is the mean of $\mathrm{Bernoulli}\,(p)$, p can be estimated by using the sample mean and the point estimate is

$$\widehat{p} := \overline{x}^{\mathrm{obs}},$$

  i.e. the proportion of the individuals in the sample with the characteristic.

Since $S_n = X_1 + X_2 + \cdots + X_n$ has the distribution $\mathrm{Binomial}(n, p)$, we know that, for np and $n(1-p)$ larger than 5, the normal distribution

$$\mathrm{N}\left(np, \left(\sqrt{n}\sigma\right)^2\right)$$

is a quite good approximation of the distribution of S_n, where σ is the standard deviation of the distribution $\mathrm{Bernoulli}\,(p)$.

So, for such n, the Sample Mean $\overline{X} = \frac{S_n}{n}$ has with a quite good approximation the normal distribution

$$\mathrm{N}\left(p, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

Therefore, for such n, the level of confidence of the interval estimate

$$\widehat{p} \pm k\frac{\sigma}{\sqrt{n}},$$

where $k > 0$, is

$$\mathbb{P}\left(|\overline{X} - p| < k\frac{\sigma}{\sqrt{n}}\right) \approx 2\Phi(k) - 1.$$

and we see that the relation between the level of confidence $C\%$ and k is the same (but only approximately) as in case of the normal distribution:

$$k \approx \Phi^{-1}\left(\frac{1 + C\%}{2}\right).$$

- Observe that the standard deviation of $\text{Bernoulli}\,(p)$ is

$$\sigma = \sqrt{p\,(1-p)},$$

and so it is given in terms the unknown parameter p.

We estimate the standard deviation by

$$\widehat{\sigma} = \sqrt{\widehat{p}\,(1-\widehat{p})}.$$

obtained by replacing p with its point estimate $\widehat{p}$.

Then, the interval estimates

$$\widehat{p} \pm k\frac{\sigma}{\sqrt{n}},\ k > 0,$$

are approximated by the **estimated interval estimates**

$$\widehat{p} \pm k\frac{\widehat{\sigma}}{\sqrt{n}},\ k > 0.$$

- Observe that, when the normal approximation is valid, we have with large probability (99.7%)

$$|\widehat{p} - p| < \frac{3\sqrt{p(1-p)}}{\sqrt{n}}$$

and then

$$\left|\frac{\widehat{p} - p}{p}\right| = \frac{|\widehat{p} - p|}{p} < \frac{3\sqrt{\frac{1-p}{p}}}{\sqrt{n}}.$$

Given a function $f : D \subseteq \mathbb{R} \to \mathbb{R}$, D open, and $x \in D$, it is known that if x is perturbed to $\widetilde{x}$ and so the result $y = f(x)$ is perturbed to $\widetilde{y} = f(\widetilde{x})$, then the relation

$$\varepsilon_y \approx K(f, x)\, \varepsilon_x,$$

holds, where

$$\varepsilon_x = \frac{\widetilde{x} - x}{x} \text{ and } \varepsilon_y = \frac{\widetilde{y} - y}{y}$$

are the relative errors of the perturbations $\widetilde{x}$ and $\widetilde{y}$ and

$$K(f, x) = \frac{xf'(x)}{f(x)}$$

is the condition number of f at x.

Exercise. By considering the function

$$f(x) = \sqrt{x(1-x)},\ x \in (0,1),$$

shows that, when the normal approximation is valid, the estimated margin of error $k\frac{\widehat{\sigma}}{\sqrt{n}}$ has with large probability (99.7%) a relative error

$$\left| \frac{k\frac{\widehat{\sigma}}{\sqrt{n}} - k\frac{\sigma}{\sqrt{n}}}{k\frac{\sigma}{\sqrt{n}}} \right| = \left| \frac{\widehat{\sigma} - \sigma}{\sigma} \right|$$

which is approximately smaller than

$$\frac{3\left(\frac{1}{2} - p\right)}{\sqrt{p(1-p)}} \cdot \frac{1}{\sqrt{n}}.$$

This shows that the estimated interval estimates are good approximations of the interval estimates.

- Now, we present two examples involving interval estimates for proportions.

  Example. In a sample of $n = 100$ *students at an university,* 82 *of them are nonsmokers. Based on this, construct* 90%, 95% *and* 99% *confidence intervals of p, p being the proportion of all the students at the university that are nonsmokers.*

  The point estimate of p is $\widehat{p} = \frac{82}{100} = 0.82$.

  Observe that $n\widehat{p} = 82$ *and* $n(1 - \widehat{p}) = 18$ *are both larger than* 5 *and so the normal approximation can be considered valid.*

The estimated interval estimates are

$$\widehat{p} \pm k\frac{\widehat{\sigma}}{\sqrt{n}} = 0.82 \pm k\frac{\sqrt{0.82\,(1-0.82)}}{\sqrt{100}} = 0.82 \pm k0.0384 = 82\% \pm k3.84\%.$$

For $C\% = 90\%, 95\%, 99\%$*, we have* $k = 1.65, 1.96, 2.58$*, respectively, and so*

$$\begin{aligned}
90\%\ \textit{confidence interval:} &\quad 82\% \pm 1.65 \cdot 3.84\% = 82\% \pm 6.3\% \\
95\%\ \textit{confidence interval:} &\quad 82\% \pm 1.96 \cdot 3.84\% = 82\% \pm 7.5\% \\
99\%\ \textit{confidence interval:} &\quad 82\% \pm 2.58 \cdot 3.84\% = 82\% \pm 9.9\%.
\end{aligned}$$

Example. On December 24, 1991, The New York Times reported a poll where it was said that 46% *percent of the US population were in favor of the economic politics of the President George Bush with a margin of error of* $\pm 3\%$.

In this case, the level of confidence is not indicated. It is common practice for media to present 95% *confidence intervals. Indeed, there is the following rule: unless it is mentioned otherwise, the interval estimate has* 95% *level of confidence.*

How many people were contacted for the poll?

The point estimate of the proportion p of the US population were in favor of the economic politic of the president is $\widehat{p} = 46\%$.

With a level of confidence $C\% = 95\%$, *we have* $k = 1.96$.

The margin of error is

$$k\frac{\widehat{\sigma}}{\sqrt{n}} = k\frac{\sqrt{\widehat{p}\left(1-\widehat{p}\right)}}{\sqrt{n}} = 3\%.$$

and then

$$\begin{aligned} n &= \left(k\frac{\sqrt{\widehat{p}\left(1-\widehat{p}\right)}}{3\%}\right)^2 = k^2 \cdot \frac{\widehat{p}\left(1-\widehat{p}\right)}{(3\%)^2} \\ &= 1.96^2 \cdot \frac{0.46 \cdot (1-0.46)}{(0.03)^2} \\ &= 1060.3. \end{aligned}$$

A typical poll contacts around one thousand persons.

Exercise. In the example of the election, where the result of the poll is that "the party A has the 32.4% of the votes with margin of error $\pm 2.2\%$", find the size of the sample in case of the following levels of confidence $C\% = 90\%, 95\%, 99\%$.

- We have presented the formula

$$n = \left\lceil \left(\frac{k\sigma}{\varepsilon} \right)^2 \right\rceil$$

for determining a priori the size of the sample in order to have the margin of error not larger than a fixed tolerance ε.

In case of the distribution $\mathrm{Bernoulli}\,(p)$, we have the problem that the estimate $\widehat{\sigma} = \widehat{p}\,(1-\widehat{p})$ of $\sigma = \sqrt{p\,(1-p)}$ is not known in advance.

We observe that

$$\text{the margin of error } k\frac{\widehat{\sigma}}{\sqrt{n}} \text{ is } \leq \varepsilon \;\Leftrightarrow n \geq \left(\frac{k\,\widehat{\sigma}}{\varepsilon} \right)^2 = \left(\frac{k\sqrt{\widehat{p}\,(1-\widehat{p})}}{\varepsilon} \right)^2$$

We also observe that $\widehat{p}\,(1-\widehat{p})$, $\widehat{p} \in [0,1]$, takes at $\widehat{p} = \frac{1}{2}$ the maximum value $\frac{1}{4}$.

Since

$$\left(\frac{k\sqrt{\frac{1}{4}}}{\varepsilon}\right)^2 \geq \left(\frac{k\sqrt{\widehat{p}\left(1-\widehat{p}\right)}}{\varepsilon}\right)^2,$$

we conclude that if

$$n \geq \left(\frac{k\sqrt{\frac{1}{4}}}{\varepsilon}\right)^2 = \left(\frac{k}{2\varepsilon}\right)^2$$

then the margin of error is $\leq \varepsilon$.

We take

$$n = \left\lceil \left(\frac{k}{2\varepsilon}\right)^2 \right\rceil.$$

Of course, if $\widehat{p}$ is far from $\frac{1}{2}$, the worst case, then the margin of error is remarkably smaller than ε.

Exercise. Let r be the ratio between the margin of error and ε. Prove that

$$\sqrt{4\widehat{p}\left(1-\widehat{p}\right)\left(1-\frac{1}{n}\right)} < r \leq \sqrt{4\widehat{p}\left(1-\widehat{p}\right)}$$

Exercise. Often, politicians require polls able to produce a 95% confidence interval with margin of error not larger than 1%. Find how many people have to be contacted in such a poll.

One-sided interval estimates

- Consider the situation of a Sample from a general distribution with a unique unknown parameter, which is the mean μ and so the sample mean is used as an estimator of μ.

  Suppose that, instead of the previous **two-sided interval estimates**

$$\left(\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}, \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}\right),$$

  where $k > 0$, we are interested in **one-sided interval estimates** of type

$$\left(\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}, +\infty\right) \text{ or } \left(-\infty, \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}\right).$$

  The ends

$$\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}} \text{ and } \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}$$

  in one-sided interval estimates are called a **lower confidence bound** and an **upper confidence bound**, respectively.

- Two-sided interval estimates are used when an accurate estimation of μ is needed.

  On the other hand, one-sided interval estimates are used when we need to compare μ with μ_0, where $\mu_0 \in \mathbb{R}$ is given, i.e. to decide which relation $\mu < \mu_0$ or $\mu > \mu_0$ holds.

  We cannot decide this by comparing the point estimate $\overline{x}^{\text{obs}}$ of μ with μ_0. In fact, we could have $\overline{x}^{\text{obs}} > \mu_0 > \mu$ and so decide the wrong relation $\mu > \mu_0$ because $\overline{x}^{\text{obs}} > \mu_0$, or have $\overline{x}^{\text{obs}} < \mu_0 < \mu$ and so decide the wrong relation $\mu < \mu_0$ because $\overline{x}^{\text{obs}} < \mu_0$.

In order to decide which relation holds we proceed as follows.

When $\overline{x}^{\text{obs}} > \mu_0$, we consider a one-sided confidence interval

$$I(\boldsymbol{x}^{\text{obs}}) = \left(\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}, +\infty \right),$$

where $k > 0$, with level of confidence $C\%$. This level of confidence is also said the level of confidence of the lower confidence bound $\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}$. If

$$\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}} \geq \mu_0, \tag{1}$$

we say that the relation $\mu > \mu_0$ is satisfied with level of confidence $C\%$.

In fact, if $\mu \leq \mu_0$, then $\overline{X} - k\frac{\sigma}{\sqrt{n}} \geq \mu_0$ implies μ not in the random One-Sided Interval Estimate $I(\boldsymbol{X})$, an event with probability $1 - C\%$. So, if $\mu \leq \mu_0$, the probability of $\overline{X} - k\frac{\sigma}{\sqrt{n}} \geq \mu_0$, i.e. the probability of observing (1), is not larger than $1 - C\%$ and then it is a small probability. Thus, we can be quite confident that $\mu > \mu_0$ holds.

On the other hand, when $\overline{x}^{\text{obs}} < \mu_0$ we consider a one-sided confidence interval

$$\left(-\infty, \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}\right),$$

where $k > 0$, with level of confidence $C\%$. This level of confidence is also said the level of confidence of the upper confidence bound $\overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}$. If

$$\overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}} \leq \mu_0,$$

we say that the relation $\mu < \mu_0$ is satisfied with level of confidence $C\%$.

The case of the normal distribution

- In case of a normal distribution, the level of confidence of a lower confidence bound $\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}$ is

$$\mathbb{P}\left(\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu\right) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < k\right) = \Phi(k)$$

and the level of confidence of an upper confidence bound $\overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}$ is

$$\mathbb{P}\left(\mu < \overline{X} + k\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > -k\right) = 1 - \Phi(-k) = \Phi(k).$$

This means that, fixed a given level of confidence $C\%$, we have

$$k = \Phi^{-1}(C\%),$$

i.e. k is the Cth percentile of the standard normal distribution.

- Observe that, given a level of confidence $C\% \in (0,1)$, we have, for the same sample $\boldsymbol{x}^{\text{obs}}$, the two-sided $C\%$ confidence interval

$$\left(\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}, \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}\right), \; k = \Phi^{-1}\left(\frac{1+C\%}{2}\right),$$

as well as the one-sided $C\%$ confidence intervals

$$\left(\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}}, +\infty\right) \text{ and } \left(-\infty, \overline{x}^{\text{obs}} + k\frac{\sigma}{\sqrt{n}}\right), \; k = \Phi^{-1}\left(C\%\right).$$

Since

$$C\% < \frac{1+C\%}{2}$$

and then

$$\Phi^{-1}\left(C\%\right) < \Phi^{-1}\left(\frac{1+C\%}{2}\right),$$

the margin of error $k\frac{\sigma}{\sqrt{n}}$ is smaller for one-sided confidence intervals.

Exercise. For a fixed level of confidence $C\%$ and a fixed tolerance $\varepsilon > 0$, find the size n of the Sample guaranteeing a margin of error not larger than ε in the one-sided confidence intervals and the size of the Sample guaranteeing the same in the two-sided confidence interval. What is the larger size?

- *Example. The life of tires of a certain brand is a random variable normally distributed with unknown mean μ and known standard deviation $\sigma = 5.30 \cdot 10^3$ km. The mean of the lifes of a sample with size $n = 10$ of such tires is $\overline{x}^{\text{obs}} = 45.1 \cdot 10^3$ km.*

  An advertisement says: "With 95 percent certainty, the expected life of the tires is over 42000 km "(consider $42000 = 42.0 \cdot 10^3$). Is it true this advertisement?

  The 95% lower confidence bound, which has

$$k = \Phi^{-1}(C\%) = \Phi^{-1}(95\%) = 1.65,$$

  is

$$\overline{x}^{\text{obs}} - k\frac{\sigma}{\sqrt{n}} = 45.1 \cdot 10^3 - 1.65 \cdot \frac{5.30 \cdot 10^3}{\sqrt{10}} = 42.3 \cdot 10^3 \ \text{km}$$

  and it is larger than $42.0 \cdot 10^3$ km.

  Thus the advertisement can be considered true.

  Exercise. Find the 95% two-sided interval estimate. Can the advertisement be considered true by using this estimate?

- Exercise. In the example of the astronomer measuring the distance of the star from the Earth, find the level of confidence of the upper confidence bound 10 $\mathrm{light\ year}$.

The case of estimating proportions

- Of course, in case of a general distribution the previous considerations are still valid, but only approximately and for n large, i.e. when the normal approximation of the distribution of the Sample Mean is valid.

  In particular, this is true for a Bernoulli distribution $\mathrm{Bernoulli}(p)$, $p \in (0, 1)$, when np and $n(1 - p)$ are larger than 5.

- Now, we present two examples with the Bernoulli distribution where estimated one-sided interval estimates are involved.

  Example. Suppose that one wants to known whether or not the percentage of all workers in a large city that are unsatisfied with their working conditions are over 25%.

  A sample of 125 *workers indicates that* 42 *are unsatisfied.*

  So, the point estimate of the percentage of the unsatisfied workers is

$$\widehat{p} = \frac{42}{125} = 33.6\%.$$

  and we have $\widehat{p} > 25\%$.

  Are we sure with level of confidence 95% *that the true percentage is over* 25%*?*

We are interested in the estimated lower confidence bound

$$\widehat{p} - k\frac{\widehat{\sigma}}{\sqrt{n}}.$$

with level of confidence $C\% = 95\%$ *and so* $k = \Phi^{-1}(95\%) = 1.65$.

Such a bound is

$$\widehat{p} - k\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}} = 0.336 - 1.65 \cdot \frac{\sqrt{0.336(1-0.336)}}{\sqrt{125}} = 26.6\%.$$

Therefore, we are sure with level of confidence 95% *that the true percentage is over* 25%.

Exercise. Are we sure with level of confidence 99% *that the true percentage is over* 25%*?*

Exercise. If the unsatisfied workers were 20*, are we sure with level of confidence* 95% *that the true percentage is under* 25%*?*

Example. The Aid to Families with Dependent Children (AFDC) program was an assistence program in US from 1935 to 1996. It provided financial assistance to children whose familiy has low or no income.

However, since errors cannot be avoided, not every family funded by AFDC met the eligibility requirements.

The California state considered its counties responsible for overseeing the eligibility requirements and it had set a maximum error rate of 4%. The error rate is equal to number of ineligible funded cases divided by number of funded cases.

If more than 4% of the funded cases in a county were found ineligible, then a financial penalty was placed upon the county, with the amount of the penalty determined by the error rate.

Since the California state did not have the resources to check every case for eligibility, it used a random sample to estimate the error rate in each county.

In 1981, a random sample of 152 cases was chosen in Alameda County, California, and 9 were found to be ineligible. So, the estimated rate error was

$$\widehat{p} = \frac{9}{152} = 5.9\%.$$

Since this value is larger than the maximum error rate 4%*, a penalty of* 949597$ *was imposed to the county by the state.*

But, the county appealed to the courts, arguing that 9 errors in 152 trials were not sufficient evidence to prove that the rate error exceeded 4%*.*

With help from statistical experts, the court decided that it was unfair to take only the point estimate $\widehat{p} = 5.9\%$ *of the true rate error p of the county. The court decided it would be fairer to use a* 95% *lower confidence bound of p.*

The 95% *percent lower confidence bound is*

$$\begin{aligned} \widehat{p} - k\frac{\sqrt{\widehat{p}\left(1-\widehat{p}\right)}}{\sqrt{n}} &= 0.059 - 1.65\frac{\sqrt{0.059\left(1-0.059\right)}}{\sqrt{152}} \\ &= 5.9\% - 3.15\% = 2.75\%. \end{aligned}$$

Since this value is less than the maximum error rate of 4%*, the court overturned the state's decision and ruled that no penalty was due.*

Estimating the mean when the standard deviation is unknown

- Consider the situation of a Sample from the normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$, where now both the parameters μ and σ are unknown.

  We want give interval estimates for the mean μ in this situation.

  Above, we considered the situation where σ was known. In that case, we had two-sided interval estimates

$$\overline{x}^{\mathrm{obs}} \pm k\frac{\sigma}{\sqrt{n}},\ k > 0,$$

  with levels of confidence

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < k\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(|Z| < k\right)$$

  determined as a probability for the standard normal variable

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

When σ is unknown, we have the point estimate

$$s_x^{\text{obs}} = \sqrt{\frac{\sum_{k=1}^{n} (x_k^{\text{obs}} - \overline{x}^{\text{obs}})^2}{n-1}}.$$

of σ given by the estimator sample standard deviation.

As in the case of the Bernoulli distribution, we could use estimated interval estimates, i.e. to consider σ as known since it is approximated by s_x^{obs}.

Exercise. Recall that the Sample Variance S_X^2 has standard deviation $\frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}$. By using the Chebyshev's inequality, find an upper bound for the relative error

$$\left| \frac{k\frac{s_x^{\text{obs}}}{\sqrt{n}} - k\frac{\sigma}{\sqrt{n}}}{k\frac{\sigma}{\sqrt{n}}} \right| = \left| \frac{s_x^{\text{obs}} - \sigma}{\sigma} \right|$$

of the margin of error in estimated interval estimates that holds with probability $\geq 95\%$.

Exercise. Now, recall that S_X^2 has, for n large, distribution close to the normal distribution $\mathrm{N}\left(\sigma^2, \left(\frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}\right)^2\right)$ (we saw this in a previous exercise). Find, for n large, a bound for the relative error of the margin of error in estimated interval estimates that holds with probability $\geq 99.7\%$.

The previous two exercises show that estimated interval estimates should be used when n is large.

- When n is not large, we consider a two-sided interval estimate

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}},\ k > 0,$$

not as an estimated interval estimate, i.e. not as an approximation of

$$\overline{x}^{\text{obs}} \pm k\frac{\sigma}{\sqrt{n}},$$

but as a new type of two-sided interval estimate.

The level of confidence of this new two-sided interval estimate is

$$\mathbb{P}\left(|\overline{X}-\mu|<k\frac{S_X}{\sqrt{n}}\right)=\mathbb{P}\left(|T_{n-1}|<k\right),$$

where S_X is the Sample Standard Deviation and

$$T_{n-1}:=\frac{\overline{X}-\mu}{\frac{S_X}{\sqrt{n}}}.$$

Here, the level of confidence is determined by the random variable T_{n-1} not by the standard normal random variable $Z=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$.

By recalling the random variable

$$Y = \frac{(n-1)\, S_X^2}{\sigma^2},$$

whose distribution is χ^2_{n-1}, we obtain

$$T_{n-1} = \frac{\overline{X} - \mu}{\frac{S_X}{\sqrt{n}}} = \frac{\overline{X} - \mu}{\frac{\sigma\sqrt{\frac{Y}{n-1}}}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{Y}{n-1}}} = \frac{Z}{\sqrt{\frac{Y}{n-1}}}$$

Moreover, by recalling that $\overline{X}$ and Y are independent, we have that $Z = \frac{\overline{X}-\mu}{\sigma}$ and Y are independent.

The Student t distributions

- The distribution of the random variable T_{n-1} is a Student t distribution.

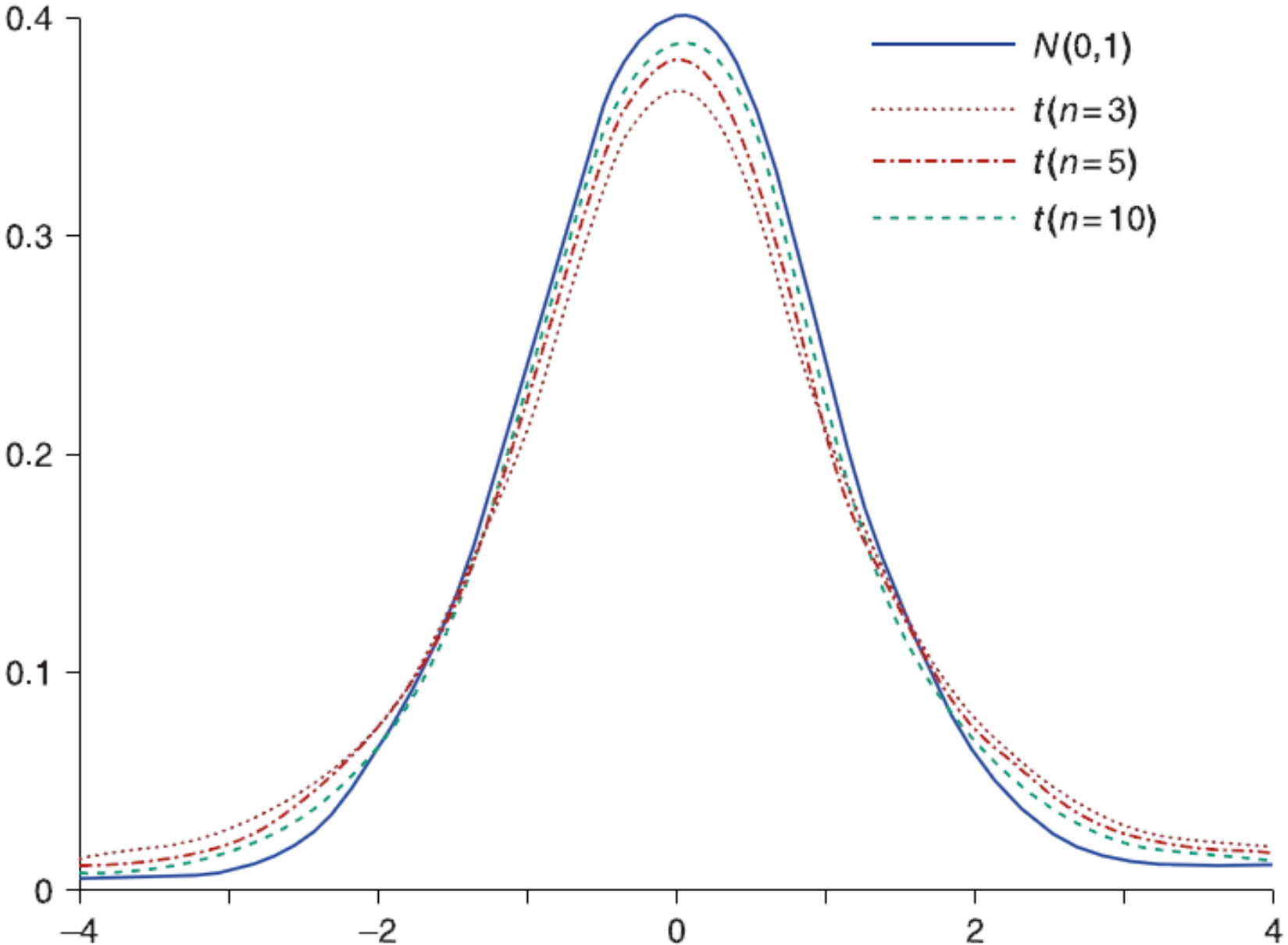
Definition

A continuous random variable T is said to have the **Student t distribution** with ν degrees of freedom, denoted by t_ν, where ν is a positive integer, if

$$T = \frac{Z}{\sqrt{\frac{Y}{\nu}}},$$

where Z is a random variable with the standard normal distribution, Y is a random variable with the chi-squared distribution with ν degrees of freedom and Z and Y are independent.

Some Student t distributions, along with the standard normal distribution, are given in the next figure.



A Student t distribution, like a standard normal distribution, is symmetric around the mean zero.

It looks similar to a standard normal distribution, although it is more spread out: it has "larger tails."

As the number ν of degrees of freedom increases, the distribution becomes more and more similar to the standard normal distribution.

- Now, we prove all these properties.

  We begin by giving a formula for the pdf of a random variable

$$T_\nu = \frac{Z}{\sqrt{\frac{Y}{\nu}}}$$

  with distribution t_ν.

We have, for $x \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(x < T_\nu \leq x + dx) &= \int_{w>0} \mathbb{P}\left(x < \frac{Z}{\sqrt{\frac{Y}{\nu}}} \leq x + dx \cap w < \sqrt{\frac{Y}{\nu}} \leq w + dw\right) \\
&= \int_{w>0} \mathbb{P}\left(x < \frac{Z}{w} \leq x + dx \cap w < \sqrt{\frac{Y}{\nu}} \leq w + dw\right) \\
&= \int_{w>0} \mathbb{P}\left(wx < Z \leq wx + wdx \cap w < \sqrt{\frac{Y}{\nu}} \leq w + dw\right) \\
&= \int_{w>0} \mathbb{P}(wx < Z \leq wx + wdx) \cdot \mathbb{P}\left(w < \sqrt{\frac{Y}{\nu}} \leq w + dw\right)
\end{aligned}$$

where the last equality follows by the independence of Z and $\sqrt{\frac{Y}{\nu}}$.

Thus

$$\begin{aligned} f_{T_\nu}(x)\,dx &= \mathbb{P}(x < T_\nu \leq x + dx) \\ &= \int_{w>0} \mathbb{P}(wx < Z \leq wx + wdx) \cdot \mathbb{P}\left(w < \sqrt{\frac{Y}{\nu}} \leq w + dw\right) \\ &= \int_{w>0} f_Z(wx)\, wdx \cdot f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw \end{aligned}$$

and so

$$f_{T_\nu}(x) = \int_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw.$$

Now, we can prove the properties of the Student t distribution.

f_{T_ν} is symmetric around zero: for $x \geq 0$, we have, since f_Z is symmetric around zero,

$$f_{T_\nu}(-x) = \int_{w>0} f_Z(-wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw = \int_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw = f_{T_\nu}(x).$$

As a consequence of the symmetricity around zero of f_{T_ν}, we have that the mean of the distribution is zero: this is proved exactly as in case of the normal distribution.

f_{T_ν} has maximum value at 0: for $x \in \mathbb{R} \setminus \{0\}$, we have

$$\begin{aligned} f'_{T_\nu}(x) &= \frac{d}{dx} \int_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw = \int_{w>0} f'_Z(wx) \underbrace{w^2 f_{\sqrt{\frac{Y}{\nu}}}(w)}_{\geq 0} dw \\ &= \begin{cases} > 0 \text{ if } x < 0 \text{ (we have } f'_Z(wx) > 0 \text{ for all } w > 0) \\ < 0 \text{ if } x > 0 \text{ (we have } f'_Z(wx) < 0 \text{ for all } w > 0) \end{cases} \end{aligned}$$

$f_{T_\nu}(x) \to 0$ as $x \to \pm\infty$:

$$\begin{aligned} \lim_{x\to\pm\infty} f_{T_\nu}(x) &= \lim_{x\to\pm\infty} \int_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw \\ &= \int_{w>0} \underbrace{\lim_{x\to\pm\infty} f_Z(wx)}_{=0} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw \\ &= 0. \end{aligned}$$

f_{T_ν} decays to zero, as $x \to \pm\infty$, much more slowly than f_Z (it has longer tails): we have

$$\lim_{x\to\pm\infty} \frac{f_{T_\nu}(x)}{f_Z(x)} = \lim_{x\to\pm\infty} \frac{\int\limits_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw}{f_Z(x)}$$

$$= \lim_{x\to\pm\infty} \int\limits_{w>0} \frac{f_Z(wx)}{f_Z(x)} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw = \lim_{x\to\pm\infty} \int\limits_{w>0} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{w^2x^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw$$

$$= \lim_{x\to\pm\infty} \int\limits_{w>0} e^{\frac{\left(1-w^2\right)x^2}{2}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw$$

$$= \lim_{x\to\pm\infty} \left(\int\limits_0^1 e^{\frac{\left(1-w^2\right)x^2}{2}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw + \int\limits_1^{+\infty} e^{\frac{\left(1-w^2\right)x^2}{2}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw \right)$$

$$= \lim_{x\to\pm\infty} \int\limits_0^1 e^{\frac{\left(1-w^2\right)x^2}{2}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw + \lim_{x\to\pm\infty} \int\limits_1^{+\infty} e^{\frac{\left(1-w^2\right)x^2}{2}} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw$$

$$= \int\limits_0^1 \underbrace{\lim_{x\to\pm\infty} e^{\frac{\left(1-w^2\right)x^2}{2}}}_{+\infty} \cdot w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw + \int\limits_1^{+\infty} \underbrace{\lim_{x\to\pm\infty} e^{\frac{\left(1-w^2\right)x^2}{2}}}_{=0} w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw = +\infty$$

$\lim_{\nu\to\infty} f_{T_\nu}(x) = f_Z(x),\ x \in \mathbb{R}$: in fact

$$\sqrt{\frac{Y}{\nu}} = \sqrt{\frac{\sum_{i=1}^{\nu} Z_i^2}{\nu}}$$

is the square root of the Sample Mean for the Sample $Z_1^2, Z_2^2, \ldots, Z_\nu^2$ from the squared standard normal distribution;

so, for $\nu \to \infty$, $\sqrt{\frac{Y}{\nu}}$ concentrates around the square root 1 of the mean 1 of the squared standard normal distribution and then $f_{\sqrt{\frac{Y}{\nu}}}(w)$ becomes a Dirac delta function at $w = 1$ and thus

$$\begin{aligned}
\lim_{\nu\to\infty} f_{T_\nu}(x) &= \lim_{\nu\to\infty} \int_{w>0} f_Z(wx)\, w f_{\sqrt{\frac{Y}{\nu}}}(w)\, dw \\
&= \int_{w>0} f_Z(wx)\, w \underbrace{\lim_{\nu\to\infty} f_{\sqrt{\frac{Y}{\nu}}}(w)}_{\text{Delta Dirac function at } w=1} dw \\
&= f_Z(1x)\,1 = f_Z(x),\ x \in \mathbb{R}.
\end{aligned}$$

Exercise. By using

$$\mathbb{E}\left(\frac{1}{Y}\right) = \frac{1}{\nu - 2}$$

shows that

$$\mathrm{Var}\left(T_\nu\right) = \frac{\nu}{\nu - 2}.$$

Two-sided interval estimates

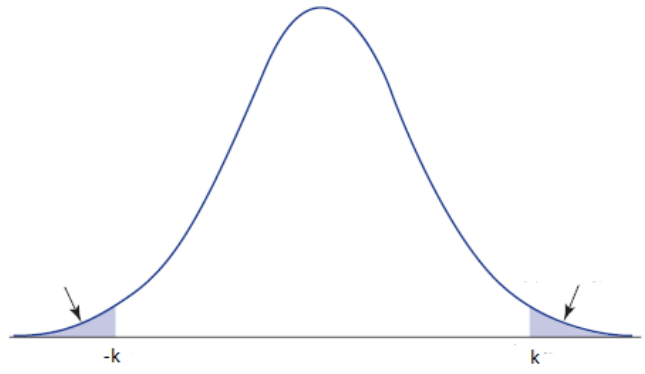
- Let T_ν be a random variable with the distribution t_ν, where ν is a positive integer.

  We denote by Φ_ν the distribution function F_{T_ν} of T_ν.

  Exercise. Prove that $\lim_{\nu\to\infty} \Phi_\nu(x) = \Phi(x),\ x \in \mathbb{R}$, where Φ is the distribution function of a standard normal random variable.

- For the simmetricity of the distribution t_ν around zero, we have

$$\Phi_\nu(-k) = \mathbb{P}(T_\nu \leq -k) = \mathbb{P}(T_\nu \geq k) = 1 - \mathbb{P}(T_\nu < k) = 1 - \Phi_\nu(k),\ k > 0.$$



as in case of the distribution function Φ.

So, similarly to

$$\mathbb{P}(|Z| < k) = 2\Phi(k) - 1,\ k > 0,$$

for a standard normal variable Z, we have

$$\mathbb{P}(|T_\nu| < k) = 2\Phi_\nu(k) - 1,\ k > 0.$$

- We have previously seen that the two-sided interval estimate

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}}$$

of the mean μ has level of confidence $\mathbb{P}(|T_{n-1}| < k)$, where the random variable T_{n-1} has distribution t_{n-1}.

So this level of confidence is

$$\mathbb{P}(|T_{n-1}| < k) = 2\Phi_{n-1}(k) - 1.$$

For a fixed $C\% \in (0,1)$, a $C\%$ confidence interval of μ is given by

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}} \quad \text{with} \quad k = \Phi_{n-1}^{-1}\left(\frac{1+C\%}{2}\right).$$

Observe that k is the

$$100\frac{1+C\%}{2} = \frac{100+C}{2}\text{th percentile}$$

of the random variable T_{n-1}.

- To compute the values of inverse Φ_ν^{-1}, ν positive integer, we can use the quantities

$$t_{\nu,\alpha} := \Phi_\nu^{-1}(1-\alpha) = 100(1-\alpha)\text{th percentile of } T_\nu,$$

where T_ν has distribution t_ν, given in the next table.

Exercise. Prove that, for a given level of confidence $C\% \in (0,1)$, we have

$$k = t_{n-1,\frac{1-C\%}{2}}.$$

Table D.2 Percentiles $t_{n,\alpha}$ of t Distributions

	α									
n	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792

Table D.2 *(Continued)*

23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

n = degrees of freedom.

- *Example. There are places in US, specially close to industrial sites, where the milk of nursing mothers contains PCB (PolyChlorinate Biphenil), a toxic chemical.*

  We assume that the amount of PCB, in the milk of a woman randomly selected in the population of the nursing mothers at one of these places, is a random variable with normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$*, where* μ *and* σ *are unknown.*

  In a sample of $n = 20$ *nursing mothers, the amounts (in parts per million) of PCB were as follows:*

  $$16,\ 0,\ 0,\ 2,\ 3,\ 6,\ 8,\ 2,\ 5,\ 0,\ 12,\ 10,\ 5,\ 7,\ 2,\ 3,\ 8,\ 17,\ 9,\ 1.$$

  We want to construct 95% *and* 99% *confidence intervals of* μ*.*

We have the point estimates

$$\overline{x}^{\text{obs}} = 5.80 \text{ and } s_x^{\text{obs}} = 5.08$$

and

$$\begin{aligned} k &= \Phi_{n-1}^{-1}\left(\frac{1+C\%}{2}\right) = t_{n-1,\frac{1-C\%}{2}} \\ &= \begin{cases} t_{19,2.5\%} \text{ if } C\% = 95\% \\ t_{19,0.5\%} \text{ if } C\% = 99\% \end{cases} = \begin{cases} 2.093 \text{ if } C\% = 95\% \\ 2.861 \text{ if } C\% = 99\%. \end{cases} \end{aligned}$$

A 95% *confidence interval of* μ *is*

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}} = 5.80 \pm 2.093\frac{5.08}{\sqrt{20}} = 5.80 \pm 2.38$$

and a 99% *confidence interval of* μ *is*

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}} = 5.80 \pm 2.861\frac{5.08}{\sqrt{20}} = 5.80 \pm 3.25.$$

One-sided interval estimates

- Observe that the levels of confidence of the one-sided interval estimates

$$\left(\overline{x}^{\text{obs}} - k\frac{s_x^{\text{obs}}}{\sqrt{n}}, +\infty\right) \quad \text{and} \quad \left(-\infty, \overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}}\right),$$

where $k > 0$, are

$$\mathbb{P}\left(\overline{X} - k\frac{S_X}{\sqrt{n}} < \mu\right) = \mathbb{P}\left(T_{n-1} = \frac{\overline{X} - \mu}{\frac{S_X}{\sqrt{n}}} < k\right) = \Phi_{n-1}(k)$$

for the first and

$$\begin{aligned}\mathbb{P}\left(\mu < \overline{X} + k\frac{S_X}{\sqrt{n}}\right) &= \mathbb{P}\left(T_{n-1} = \frac{\overline{X} - \mu}{\frac{S_X}{\sqrt{n}}} > -k\right) = \mathbb{P}\left(T_{n-1} < k\right) \\ &= \Phi_{n-1}(k),\end{aligned}$$

for the second.

So, fixed a level of confidence $C\% \in (0,1)$, $C\%$ lower and upper confidence bounds of μ are given by

$$\overline{x}^{\text{obs}} - k\frac{s_x^{\text{obs}}}{\sqrt{n}} \quad \text{and} \quad \overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}} \text{ with } k = \Phi_{n-1}^{-1}(C\%).$$

Exercise. Prove that

$$k = t_{n-1,1-C\%}.$$

- *Example. Assume that the maximum allowed amount of PCB in the milk of nursing mother is* 8 *(parts for million). Is the mean* μ *of the PCB amount in the milk of a woman randomly selected in the above population of nursing mothers larger than this value?*

  Since the point estimate $\overline{x}^{\text{obs}} = 5.80$ *of* μ *is smaller than the maximum allowed amount* 8*, we are interested in an upper confidence bound*

$$\overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}},\ k > 0.$$

  With level of confidence $C\% = 95\%$*, we have*

$$k = \Phi_{n-1}^{-1}(C\%) = t_{n-1,1-C\%} = t_{19,5\%} = 1.729$$

  and the upper confidence bound is

$$\overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}} = 5.80 + 1.729\frac{5.08}{\sqrt{20}} = 7.76.$$

We conclude that, with level of confidence 95%, *the mean of PCB amount is not larger than then the maximum allowed amount.*

On the other hand, with level of confidence 99%, *and so*

$$k = t_{n-1,1-C\%} = t_{19,1\%} = 2.539,$$

an upper confidence bound is

$$\overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}} = 5.80 + 2.539\frac{5.08}{\sqrt{20}} = 8.68.$$

We cannot conclude, with level of confidence 99%, *the same as above.*

Student t distributions in MATLAB

- In MATLAB, the values of the function Φ_ν, ν positive integer, are computed by the function tcdf:

$$\mathrm{tcdf}(x, \nu),$$

where $x \in \mathbb{R}$, computes $\Phi_\nu(x)$.

On the other hand, the values of the inverse Φ_ν^{-1} are computed by the function tinv:

$$\mathrm{tinv}(\alpha, \nu),$$

where $\alpha \in [0, 1]$, computes $\Phi_\nu^{-1}(\alpha)$.

Exercise. In the previous example of PCB, find the level of confidence of the interval estimate 5.80 ± 1.

Estimating the standard deviation

- We observe that the estimation of the unknown standard deviation σ is not as important as that of the mean μ.

  The reason is that the underlying random variable X, whose distribution is $\mathrm{N}(\mu, \sigma^2)$, is often seen as

  $$X = \mu + E,$$

  where the important term μ is altered by a random error (random noise) E normally distributed with mean 0 and standard deviation σ.

  Therefore, it is clear that we are much more interested in well estimating μ than in well estimating the random error.

  For the estimation of the standard deviation σ of the random error, the point estimation s_x^{obs} is often sufficient.

Historical remark

- The Student t distributions were introduced in 1909 by the british statistician William Sealy Gosset (1876-1937)

William S. Gosset

employed at the Guinness brewery in Dublin. He wrote his statistics papers under the nickname "Student", because the Guinnes did not permit him to use his name in scientific publications.

The introduction of the Student distributions was an important result, because they made possible to deal with situations where only samples of small size *n* were available, as it was often the case at the Guinness brewery, where few data on the malt were observed under the same conditions.

Its importance, however, was not noted at the beginning, and it was mainly ignored by the statistical community.

This was primarily because the idea of learning from samples of small size n was against the prevailing scientific belief at the time: namely

- if n is large, use the estimated confidence intervals

$$\overline{x}^{\text{obs}} \pm k\frac{s_x^{\text{obs}}}{\sqrt{n}},\ k = \Phi^{-1}\left(\frac{1+C\%}{2}\right), \tag{2}$$

  and the lower and upper confidence bounds

$$\overline{x}^{\text{obs}} - k\frac{s_x^{\text{obs}}}{\sqrt{n}} \text{ and } \overline{x}^{\text{obs}} + k\frac{s_x^{\text{obs}}}{\sqrt{n}},\ k = \Phi^{-1}\left(C\%\right), \tag{3}$$

  (this agrees with "Student": when n is large, Φ_{n-1} is close to Φ and so the "Student" confidence intervals and the "Student" confidence bounds become (2) and (3), respectively);

- if n is not large, you cannot use Statistics (this is not true, as "Student" has shown).