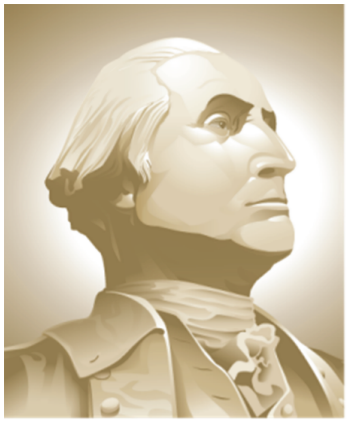
Lecture Notes EMSE 4765: Statistical Analysis Review

Chapter 23: Confidence Intervals for the Mean

Version: 1/25/2021



THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Text Book: A Modern Introduction to Probability and Statistics, Understanding Why and How

By: F.M. Dekking. C. Kraaikamp, H.P.Lopuhaä and L.E. Meester

23 Confidence Intervals

23.1 General Principle . . .

- We discussed **Estimators $T(X_1, X_2, \ldots, X_n)$** to **estimate probability distribution features**, **where $(X_1, X_2, \ldots, X_n)$ is a random sample and $X_i \sim X$.**

- If we have at our disposal **an Estimator $T(X_1, \ldots, X_n)$ for an unknown parameter θ**, and **$(x_1, \ldots, x_n)$ is a collected data set** we use the estimators **realization $t = T(x_1, \ldots, x_n)$ as our estimate for θ**.

- Example estimators are **the (random) sample mean and the (random) sample variance:**

$$\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n), \; S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

For an observed dataset $(x_1, \ldots, x_n)$ we evaluate $\overline{x}_n$ and s_n^2 as estimates for μ and σ^2. These estimates change with each dataset $(x_1, \ldots, x_n)$, but are **usually close** to $E[X] = \mu$ and $Var(X) = \sigma^2$.

23 Confidence Intervals

23.1 General Principle . . .

- **Investigate methods** to make **confidence statements about unknown parameter θ**, by **taking advantage of our knowledge of the estimator distribution of $T(X_1, \ldots, X_n)$, where $(X_1, X_2, \ldots, X_n)$ is a random sample and $X_i \sim X$.**

Speed of Light Example: Estimator $T(X_1, \ldots, X_n)$, such that $E[T] = \theta$, **θ is the unknown speed of light.** Suppose we know for standard deviation of T that $\sigma_T = 100\text{km/sec}$. **The Chebyshev 's inequality (see Section 13.2)**, implies

$$Pr(|T(X_1, \ldots, X_n) - \theta| < 2\sigma_T) \geq \frac{3}{4} = 0.75.$$

In other words (recall θ is a constant that we do not know the value of):

$T \in (\theta - 200, \theta + 200)$ with probability of at least 75%

This is a probability statement about what?

Answer:
It is a statement about **a random variable T** being **in a fixed unknown interval.**

- But, **if I am close the city of Paris, then the city of Paris is close to me.**

"T is within 200 of unknown θ" $\Leftrightarrow$ "unknown θ is within 200 of T"

- Recall **using Chebyshev 's inequality (see Section 13.2)**, we had

$$Pr(|T(X_1, \ldots, X_n) - \theta| < 2\sigma_T) \geq \frac{3}{4} = 0.75$$

In other words (recall θ is a constant that we do not know the value of):

$\theta \in (T - 200, T + 200)$ with probability of at least 75%

This is a probability statement about what?

Answer: It is a statement about **the probability that the interval** $(T - 200, T + 200)$ **with random bounds contains the unknown but constant value** θ.

- The interval $(T - 200, T + 200)$ is called **an Interval Estimator**, and **its realization is an Interval Estimate**.

Speed of Light Example: Suppose we now have as an estimate $t = 299,852.4$. Thus, t is a **realization for the estimator** T. This yields **the interval estimate:**

$$\theta \in (299852.4 - 200,\ 299852.4 + 200) = (299652.4,\ 300052.4)$$

However, because we substituted the realization t **for the random variable** T, **we cannot claim (and I repeat cannot claim!):**

$$Pr(\theta \in (299652.4,\ 300052.4)) \geq 75\%.$$

23 Confidence Intervals

23.1 General Principle . . .

- The statement **"$\theta \in (299652.4, 300052.4)$ holds with probability at least 75%"**, **does not make sense** since **nothing in that statement is random, because θ is a constant.** The model parameter θ is a constant (as per the frequentist paradigm) of which we simply do not know its value.

- **The statement "$\theta \in (299652.4, 300052.4)$" can only be true of false**, (as per the frequentist) and **we do not know which one is the case.**

- However, **if one plans to conduct the experiment again, one knows beforehand that** $Pr(\theta \in (\boldsymbol{T} - 200, \boldsymbol{T} + 200)) \geq 75\%$, because $\boldsymbol{T}$ is a random variable, not θ.

- Thus, the interval $(299652.4, 300052.4)$ is **a realization of a random interval. Only that random interval has the probability of at least 75% of capturing the true value of θ. The interval estimate** $(299652.4, 300052.4)$ **only provides us a sense of location and accuracy.**

23 Confidence Intervals

23.1 General Principle . . .

Confidence Interval: Suppose a data set $(x_1, \ldots, x_n)$ is given and thought of as **a realization of random sample** $(X_1, \ldots, X_n)$, $X_i \sim X$.
Let θ be the parameter of interest, and $\gamma \in (0, 1)$. Given **Estimators** L_n, U_n

$$L_n = g(X_1, \ldots, X_n) \text{ and } U_n = h(X_1, \ldots, X_n)$$

such that $Pr(L_n < \theta < U_n) = \gamma$, and given their **Estimates** l_n, u_n, **the interval estimate** (l_n, u_n), **where**

$$l_n = g(x_1, \ldots, x_n) \text{ and } u_n = h(x_1, \ldots, x_n)$$

is called **a** $100 \times \gamma\%$ **confidence interval for** θ. **The number** γ **is called the confidence level.**

- **No probability interpretation** can be assigned to that **the interval estimate** (l_n, u_n). Thus, no probability interpretation can be assigned to **the** $100 \times \gamma\%$ **confidence interval.** It may be considered **a reasonable range.**

23 Confidence Intervals for the Mean

23.2 Normal Data . . .

- Given $i.i.d.$ sample $(X_1, \ldots, X_n)$, $X_i \sim X$, $X \sim N(\mu, \sigma^2)$. **Assume μ is the unknown, but constant, parameter of interest and σ^2 is known (for now).**

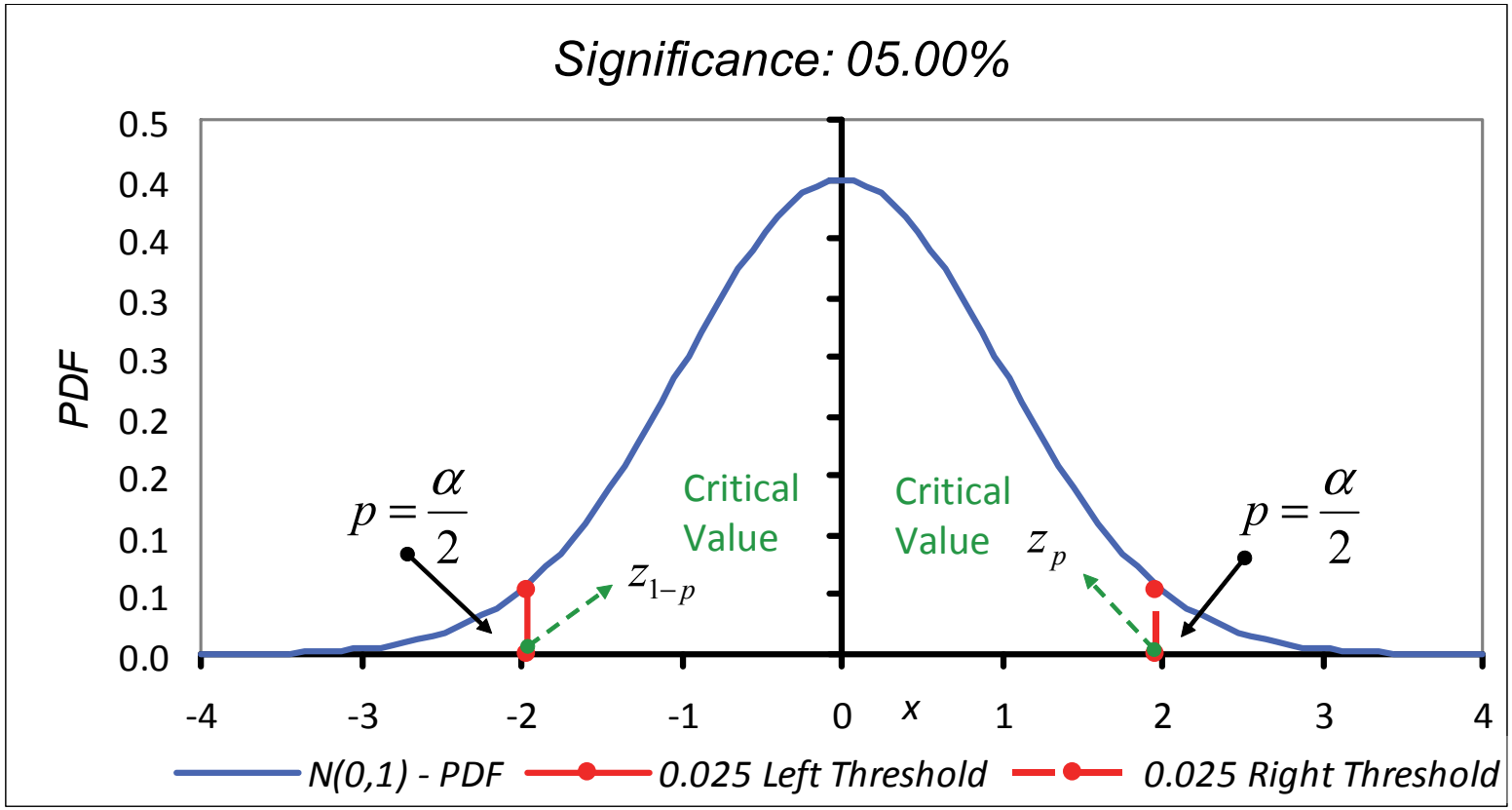
> **Definition: Let $Z \sim N(0,1)$. The critical value z_p** of an $N(0,1)$ distribution is **the number z_p that has right tail probability p,** that is:
>
> $$Pr(Z \geq z_p) = p.$$

- Thus the critical value **z_p of an $N(0,1)$ distribution is exactly the same as its $(1-p)$-th quantile z_{1-p}** , because:

$$Pr(Z \geq z_p) = p \Leftrightarrow 1 - Pr(Z \geq z_p) = 1 - p \Leftrightarrow Pr(Z < z_p) = 1 - p$$

- **Conclusion:** **Critical Value $z_p \equiv$ Quantile z_{1-p}**

Set, say, $\alpha = 5\%$. Because of **symmetry of the normal distribution**, we have:



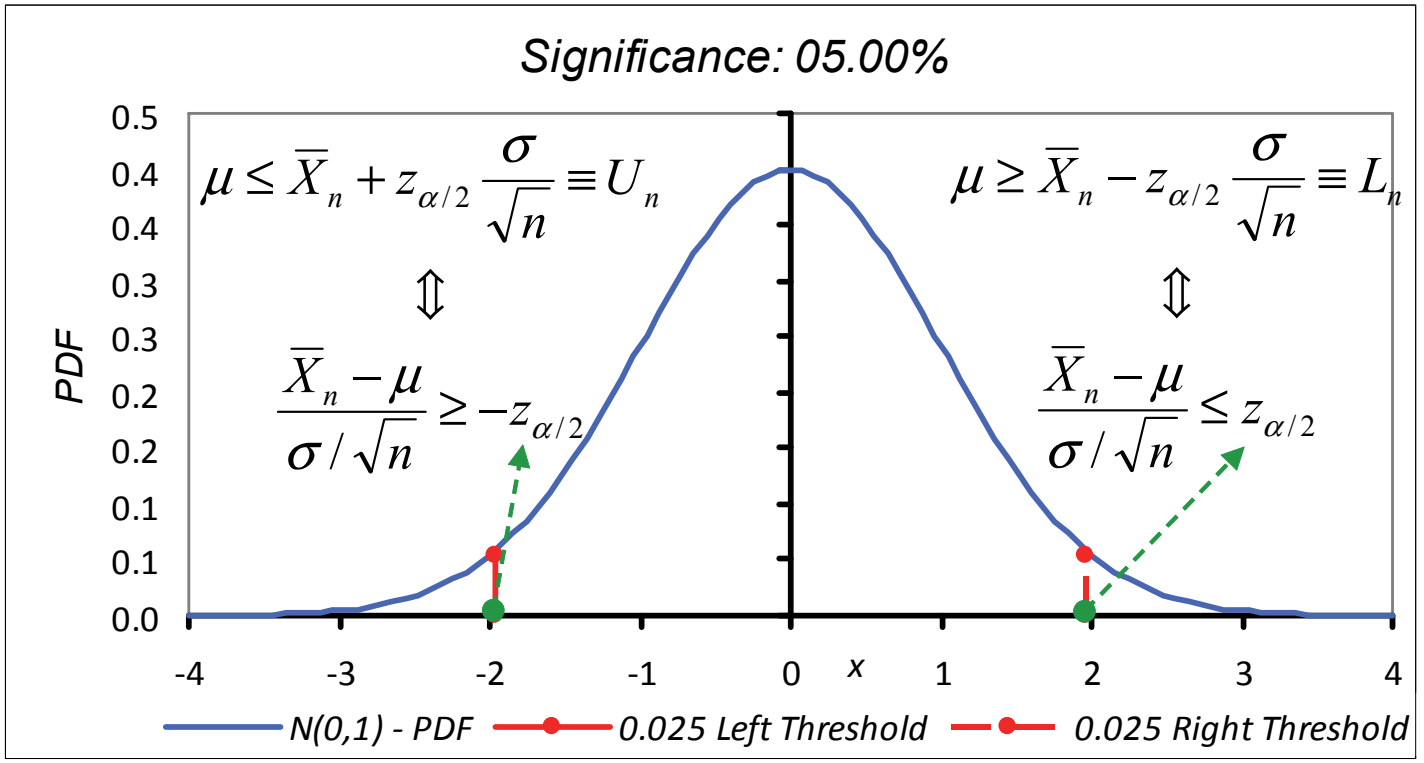
$$Pr(Z \leq -z_p) = p = \frac{\alpha}{2} \Leftrightarrow Pr(Z > z_p) = p = \frac{\alpha}{2} \Leftrightarrow z_{1-p} = -z_p$$

23 Confidence Intervals for the Mean

23.2 Normal Data, Variance Known . . .

Variance Known: Setting $\alpha = 5\%$ and given random sample $(X_1, \ldots, X_n)$, $X_i \sim X$, where $X \sim N(\mu, \sigma^2)$, we have **for the estimator distribution :**

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n) \Leftrightarrow Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$



- Hence, Estimators $L_n = \overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $U_n = \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ satisfy **interval definition with random bounds**, $Pr(\mu \in (L_n, U_n)) = 1 - \alpha$.

Normal Confidence Interval, Variance Known: Let $(X_1, \ldots, X_n)$ be a random sample such that $X_i \sim X$, $X \sim N(\mu, \sigma^2)$ with σ^2 known. Then

$$\left(\overline{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

is a $(100 - \alpha)\%$ confidence interval for the unknown mean μ, where $z_{\alpha/2}$ is the $\frac{\alpha}{2}$ critical value of an $N(0, 1)$ distribution. This is **a realization of a randomly changing interval** that has a $(1 - \alpha) \times 100\%$ probability of capturing μ.

23 Confidence Intervals for the Mean

23.1 General Principle . . .

Exercise: Randomly sample $(X_1, \ldots, X_n)$, $n = 20$ from an $N(0,1)$ distribution. Next, **pretend that it is known that the data are from a normal distribution with variance 1** but that μ is unknown. Construct the 90% confidence interval (l_n, l_n) for the expectation μ, such that

$$l_n = \bar{l}_n + \Phi^{-1}(0.05) \cdot \frac{\sigma}{\sqrt{n}}, u_n = \overline{x}_n + \Phi^{-1}(0.95) \cdot \frac{\sigma}{\sqrt{n}}$$

where $\overline{X}_n$ is the sample mean and $\Phi^{-1}(\,\cdot\,)$ is the quantile function of $N(0,1)$ distribution $\Rightarrow \alpha = 10\%$ and

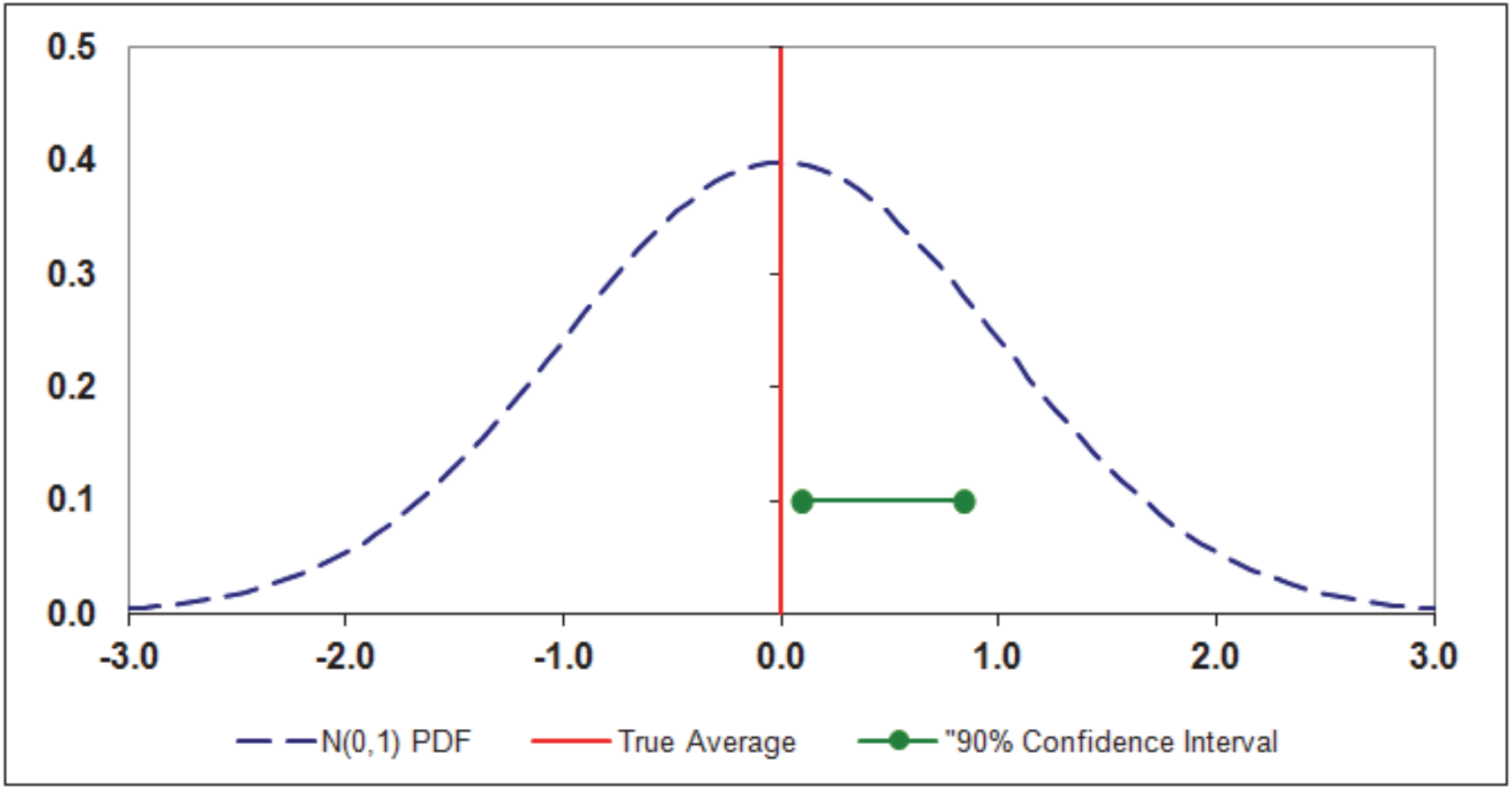
$$z_{1-\alpha/2} = z_{0.95} = N^{-1}(0.05) = -1.645,\ z_{\alpha/2} = z_{0.05} = \Phi^{-1}(0.95) = 1.645.$$

Finally, check whether the "true μ," in this case 0, is in the confidence interval (l_n, u_n).

Solution: See Excel spreadsheet Normconf.xls

23 Confidence Intervals for the Mean

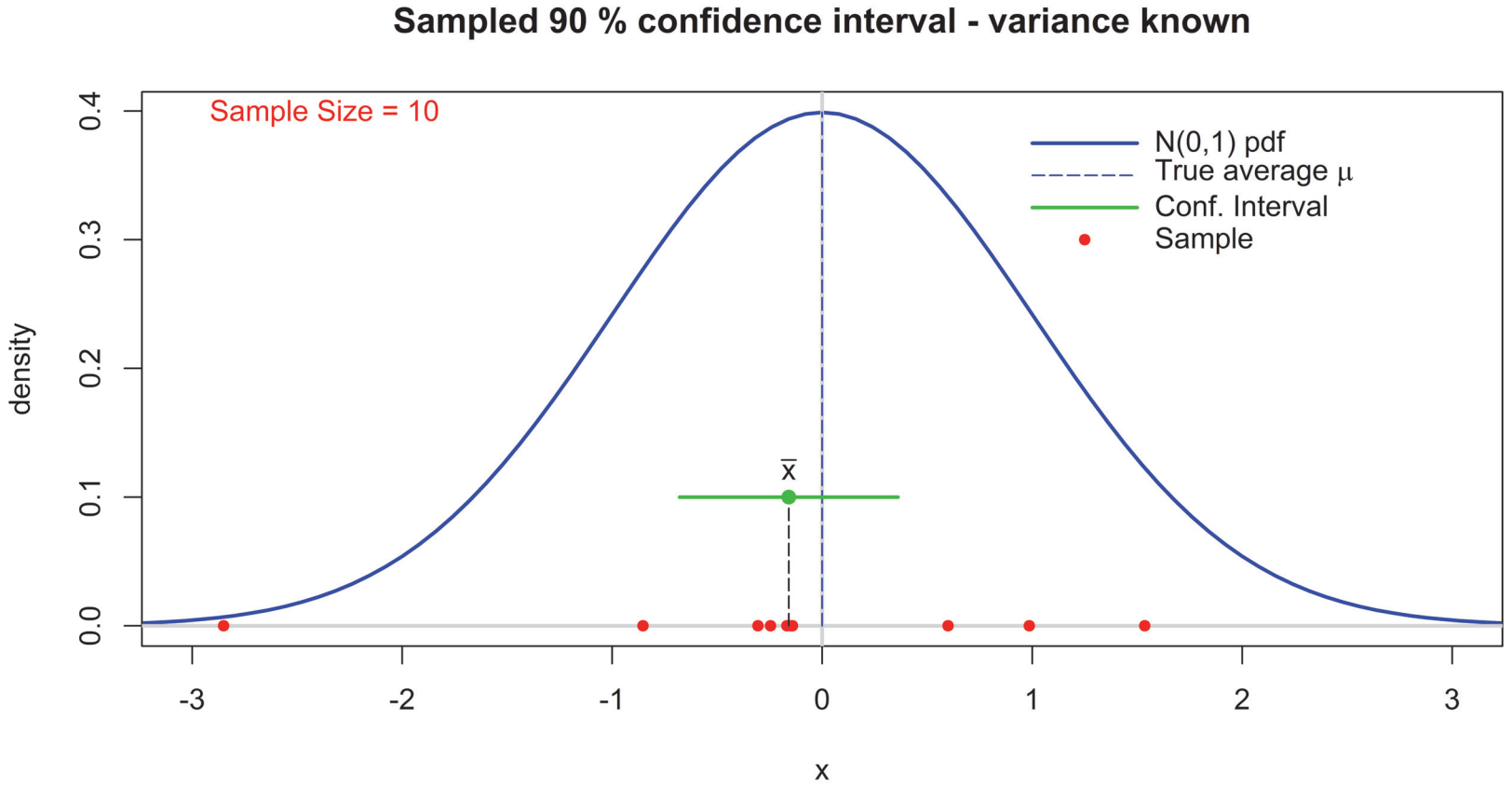
23.1 General Principle . . .



23 Confidence Intervals for the Mean

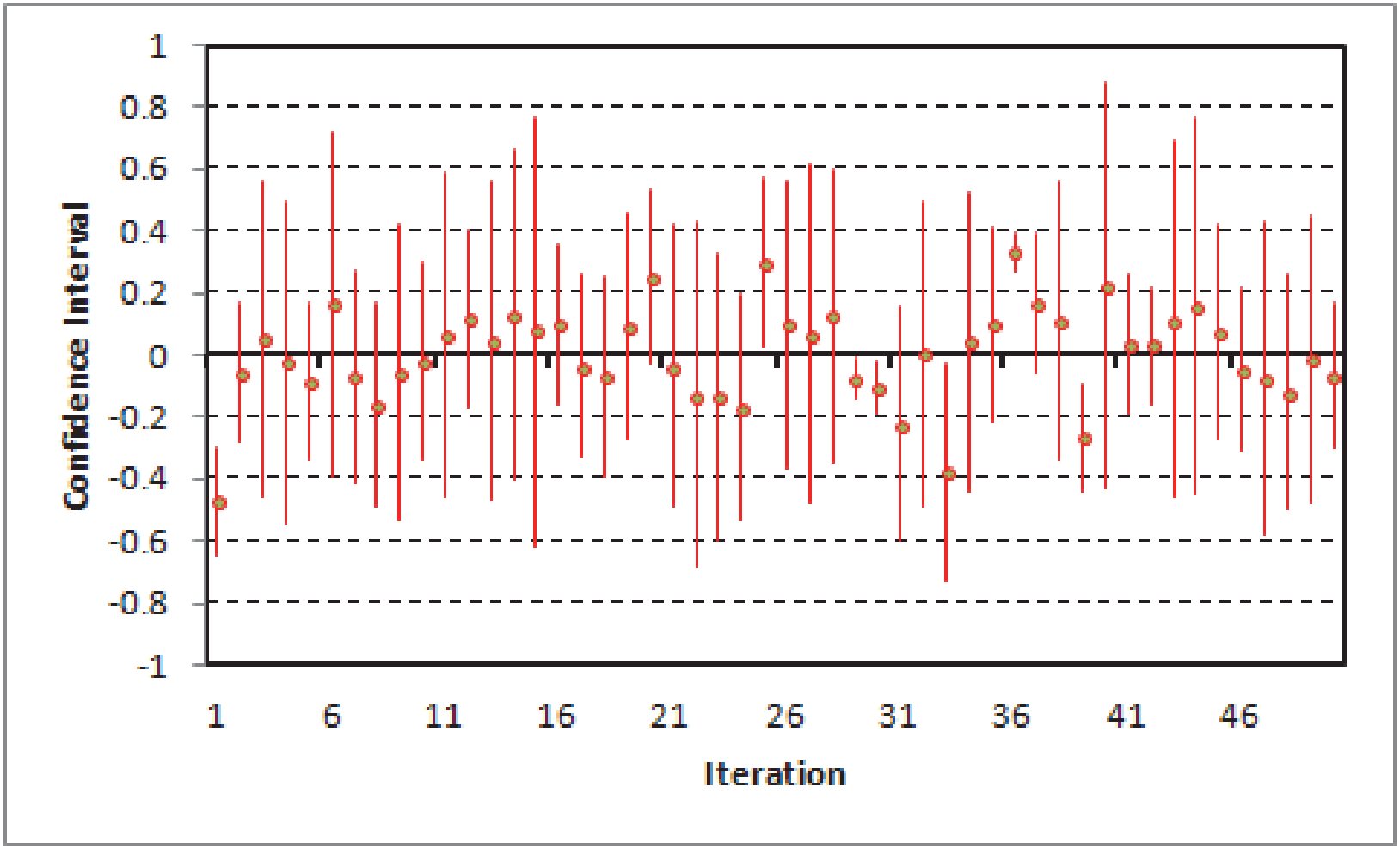
23.1 General Principle . . .

Same analysis in file "Norm_Conf_Var_Known.R"



23 Confidence Intervals for the Mean

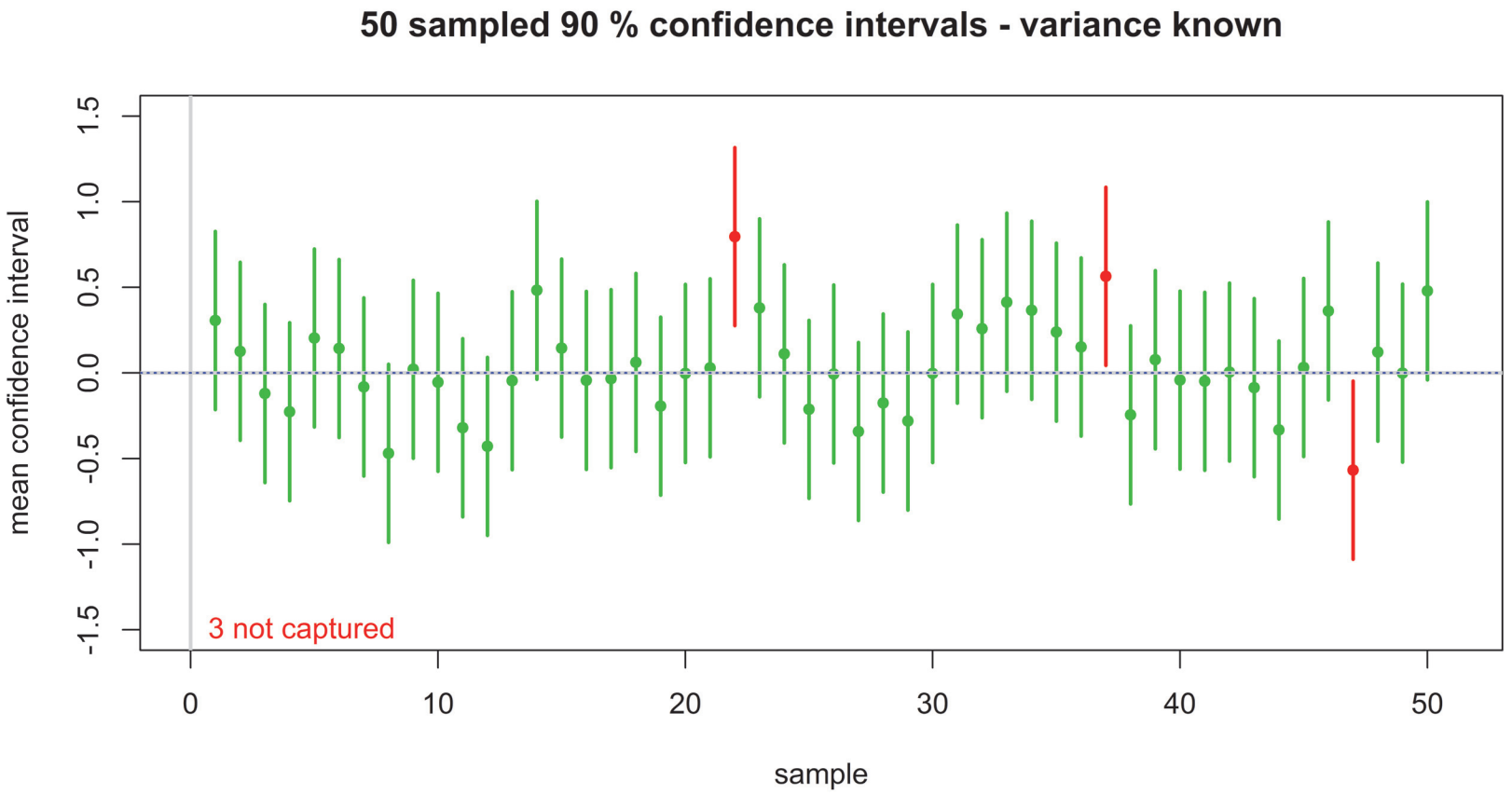
23.1 General Principle . . .



23 Confidence Intervals for the Mean

23.1 General Principle . . .

Same analysis in file "Norm_Conf_Var_Known_Sample.R"



Example: gross calorific content of coal: When a shipment of coal is traded, a number of its properties should be known accurately, because the value of the shipment is determined by them.

- **An important example is the so-called gross calorific value**, which characterizes the heat content and is a numerical value in megajoules per kilogram (MJ/kg).

- **The International Organization of Standardization (ISO)** issues standard procedures for the determination of these properties. For the gross calorific value, there is a method known as ISO 1928.

- When ISO 1928 procedure is followed, measurement errors are known to be approximately normal, **with a standard deviation of about 0.1 MJ/kg**.

Table 23.1. Gross calorific value measurements for Osterfeld 262DE27.

23.870	23.730	23.712	23.760	23.640	23.850	23.840	23.860
23.940	23.830	23.877	23.700	23.796	23.727	23.778	23.740
23.890	23.780	23.678	23.771	23.860	23.690	23.800	

Source: A.M.H. van der Veen and A.J.M. Broos. Interlaboratory study programme "ILS coal characterization"—reported data. Technical report, NMi Van Swinden Laboratorium B.V., The Netherlands, 1996.

From data we have $\overline{x}_n = 23.788, n = 23$. We know $\sigma = 0.1$. Setting $\alpha = 5\%$, we have from Table B.1, $z_{0.025} = 1.96$. This yields for the 95% confidence interval for the gross calorific content of Osterfeld 262DE27.

$$\left(23.788 - 1.96\frac{0.1}{\sqrt{23}}, 23.788+1.96\frac{0.1}{\sqrt{23}}\right) = (23.747, 23.839\,)\,MJ/kg$$

- **Variance Unknown:** Given $i.i.d.$ sample $(X_1, \ldots, X_n)$, $X_i \sim X$, $X \sim N(\mu, \sigma^2)$, we have :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n) \Leftrightarrow Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

However, because σ is unknown this information is now useless!

Definition: A continuous random variable X has a Student t-distribution with parameter $m > 1$ and m is integer, if its pdf is given by:

$$f(x) = k_m(1 + \frac{x^2}{m})^{-\frac{m+2}{2}} \text{ for } -\infty < x < \infty, \text{ where } k_m = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{m\pi}}.$$

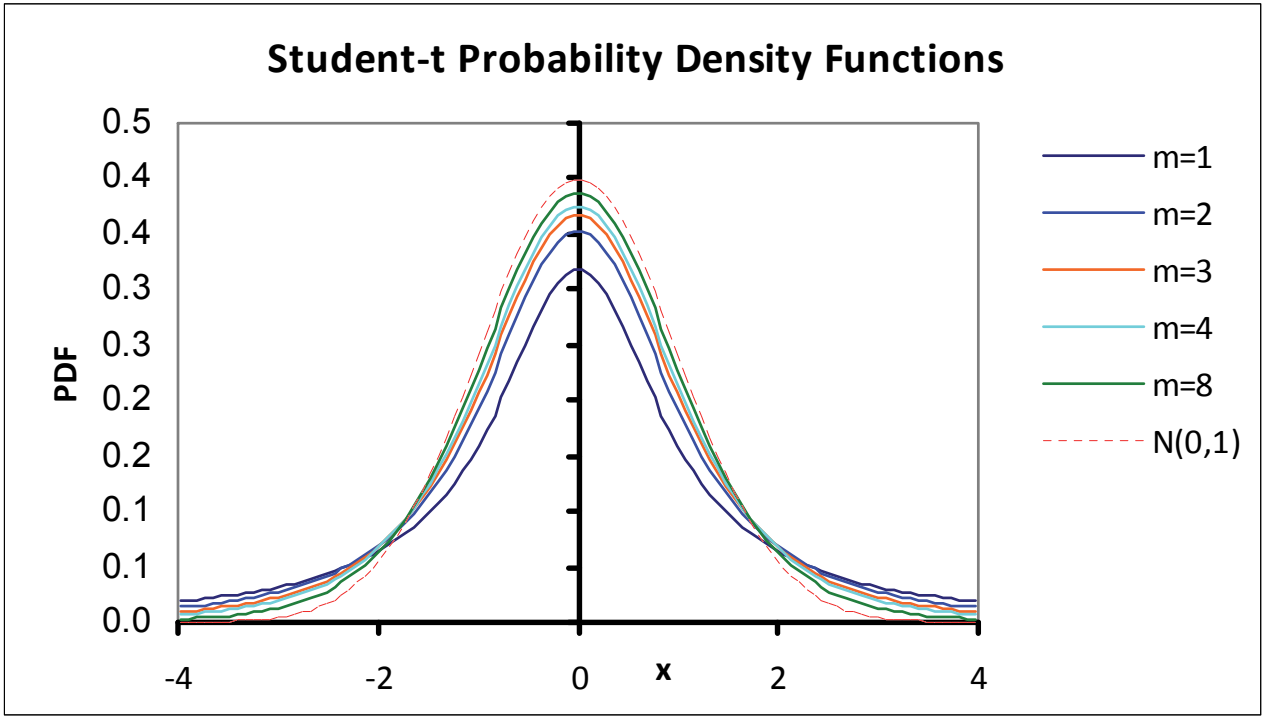
This distribution is denoted $X \sim t(m)$. Its parameter m has been given the name: **"Degrees of Freedom".**

23 Confidence Intervals for the Mean

23.2 Normal Data, Variance Unknown . . .

- If we substitute the Estimator S_n for σ in the formula for the Estimator Z, one obtains :

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

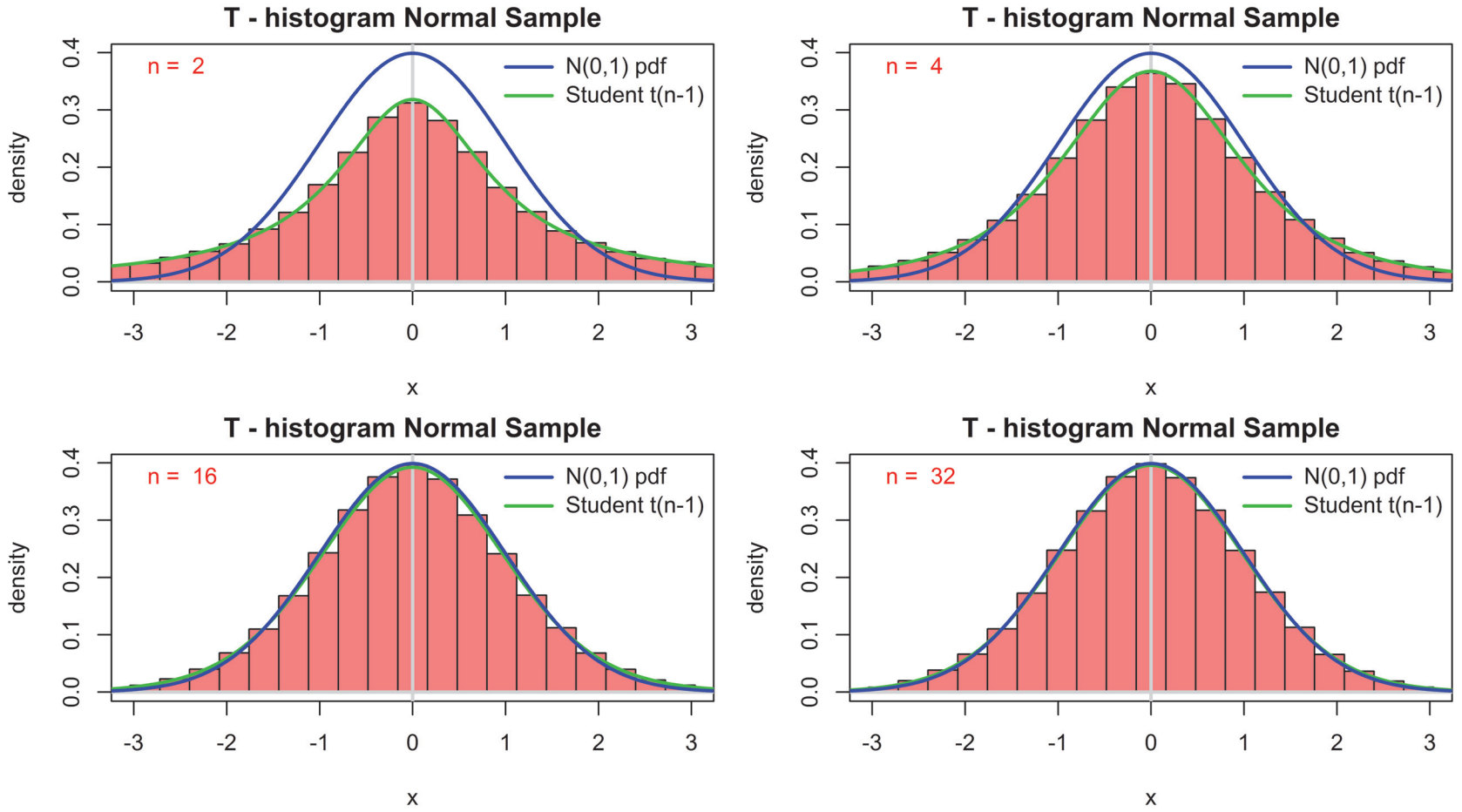


When $m \to \infty$, a Student-$t(m)$ distribution converges to a $N(0,1)$ one.

23 Confidence Intervals for the Mean

23.2 Normal Data, Variance Unknown . . .

Analysis in file "Student_T_Normal_Sample.R"



Definition: The critical value $t_{m,p}$ of $T \sim t(m)$, where $t(m)$ is a student t distribution with parameter m is **the number $t_{m,p}$ that has right tail probability p**, that is:

$$Pr(T \geq t_{m,p}) = p.$$

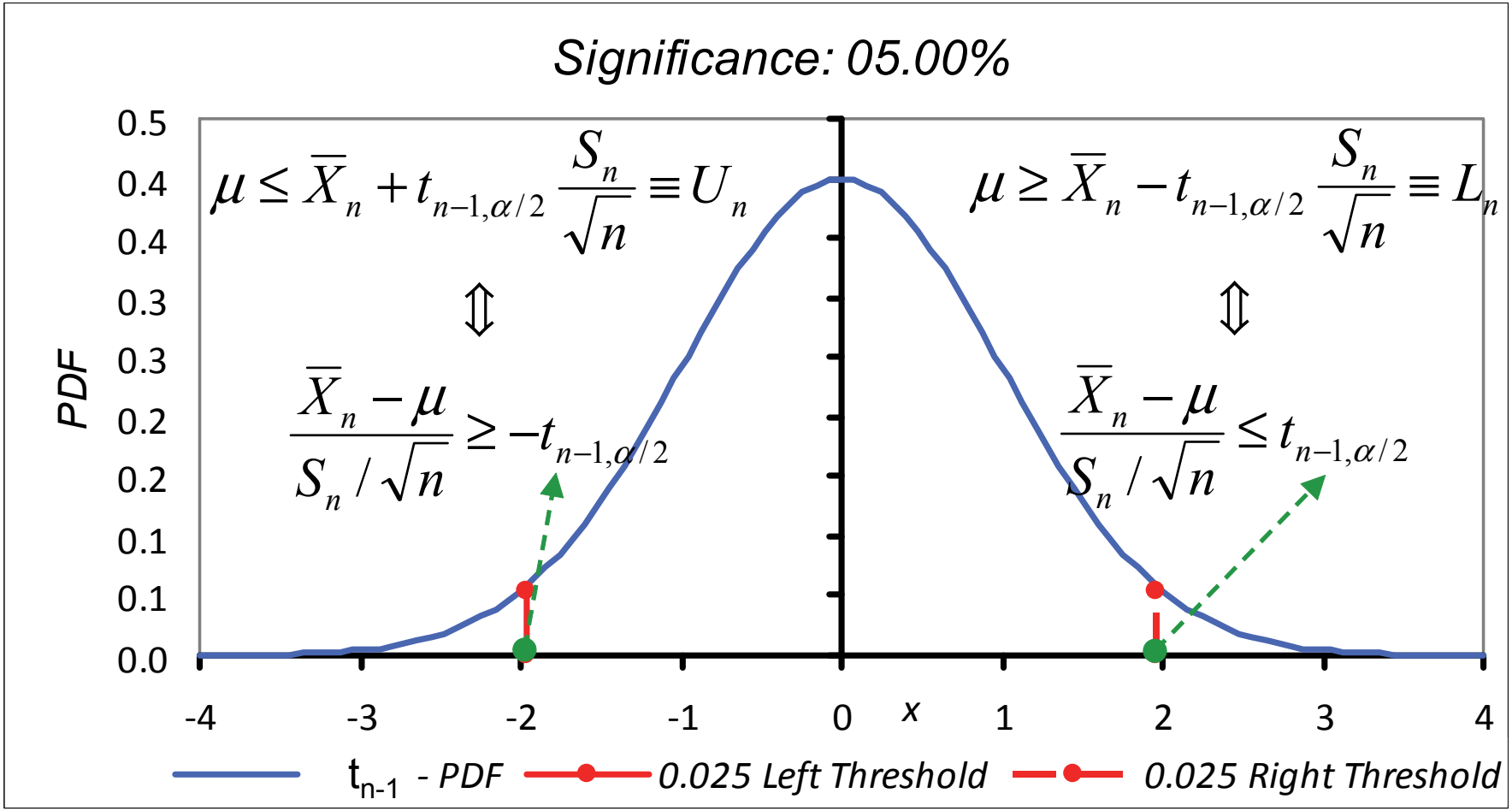
- Thus the critical value $t_{m,p}$ **of an $T(m)$ distribution is exactly the same as its $(1-p)$-th quantile $t_{m,1-p}$,** because:

$$Pr(T \geq t_{m,p}) = p \Leftrightarrow 1 - Pr(T \geq t_{m,p}) = 1 - p \Leftrightarrow Pr(T < t_{m,p}) = 1 - p$$

- **Conclusion:** Critical Value $t_{m,p} \equiv$ Quantile $t_{m,1-p}$

- Because of **symmetry of the Student-$t(m)$ distribution**, we also have:

$$\text{Critical Value } t_{m,1-p} \equiv -\text{ Critical Value } t_{m,p}$$


Significance: 05.00%
$\mu \le \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \equiv U_n$
$\mu \ge \bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \equiv L_n$
$\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \ge -t_{n-1,\alpha/2}$
$\frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \le t_{n-1,\alpha/2}$
PDF
0.5
0.4
0.4
0.3
0.3
0.2
0.2
0.1
0.1
0.0
-4
-3
-2
-1
0
x
1
2
3
4
t_{n-1} - PDF
0.025 Left Threshold
0.025 Right Threshold

- Hence, Estimators $L_n = \overline{X}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, U_n = \overline{X}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}$ satisfy **the interval definition with random bounds such that**,

$$Pr(\mu \in (L_n, U_n)) = 1 - \alpha.$$

Normal Confidence Interval, Variance unknown: Let $(X_1, \ldots, X_n)$ be a random sample such that $X_i \sim X$, $X \sim N(\mu, \sigma^2)$ with σ^2 unknown. Then:

$$\left(\overline{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \overline{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right)$$

is a $(100 - \alpha)\%$ confidence interval for the unknown mean μ, where $t_{n-1,\alpha/2}$ is the $\frac{\alpha}{2}$ critical value of an Student - t distribution with $n - 1$ degrees of freedom. Recall this is **a realization of a randomly changing interval** that has a $(1 - \alpha) \times 100\%$ probability of capturing μ.

Example: Gross calorific content of coal: Suppose σ is unknown.

Table 23.1. Gross calorific value measurements for Osterfeld 262DE27.

23.870	23.730	23.712	23.760	23.640	23.850	23.840	23.860
23.940	23.830	23.877	23.700	23.796	23.727	23.778	23.740
23.890	23.780	23.678	23.771	23.860	23.690	23.800	

Source: A.M.H. van der Veen and A.J.M. Broos. Interlaboratory study programme "ILS coal characterization"—reported data. Technical report, NMi Van Swinden Laboratorium B.V., The Netherlands, 1996.

- From data we have $\overline{x}_n = 23.788$, $s_n = 0.078$, $n = 23$, Setting $\alpha = 5\%$, we have from Table B.2, $t_{22,0.025} = 2.074$. This yields for the 95% confidence interval for the gross calorific content of Osterfeld 262DE27.

$$\left(23.788 - 2.074\frac{0.078}{\sqrt{23}}, 23.788+2.074\frac{0.078}{\sqrt{23}}\right) = (23.754, 23.822\,)\,MJ/kg$$

- **A variant of the central limit theorem** states that as $n \to \infty$ goes , **the estimator distribution** of

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \to N(0,1)$$

This fact is the basis for so called **large sample confidence intervals.**

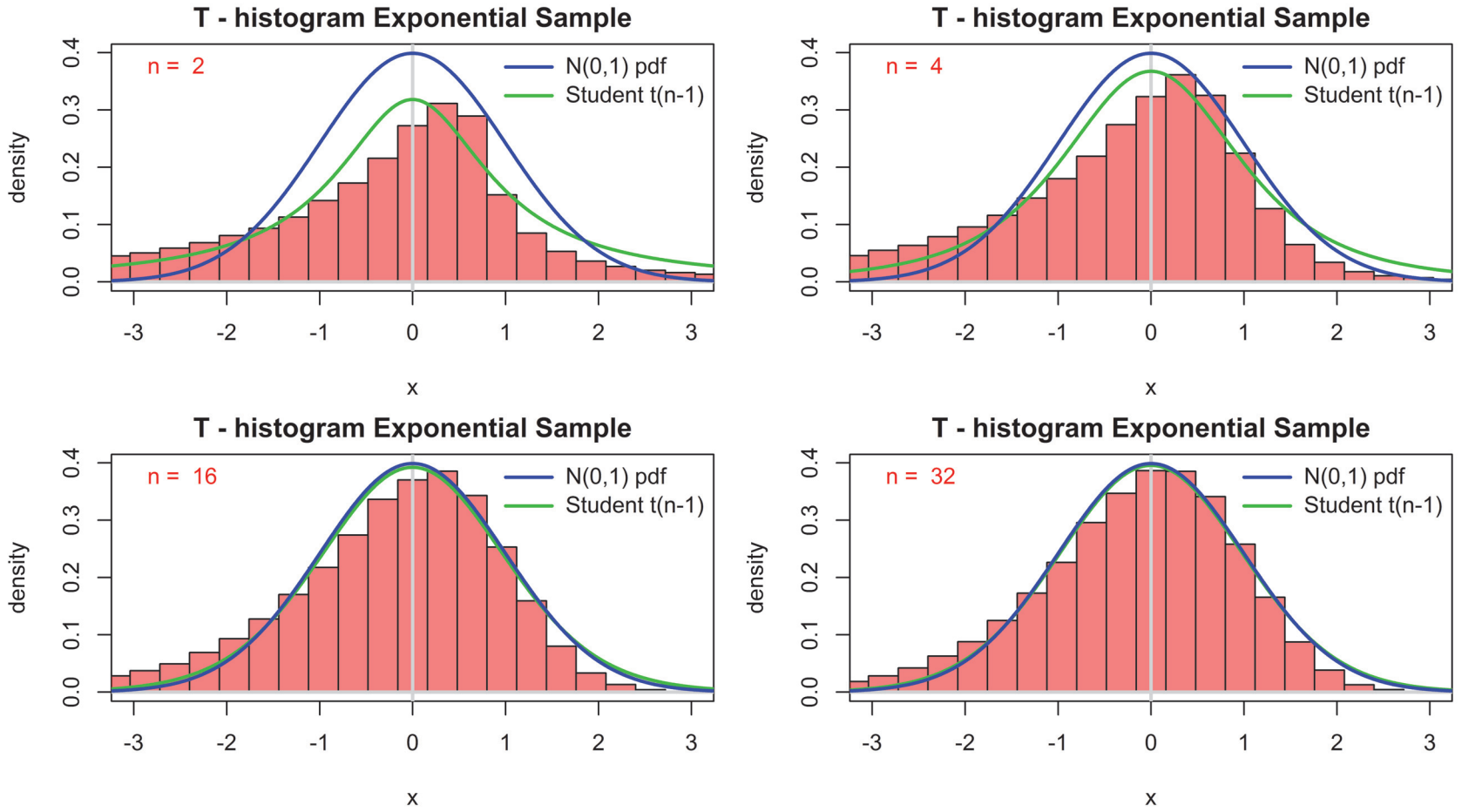
- **In words:** Procedures in this chapter for confidence intervals can be applied for $\mu = E[X]$, using a random sample $(X_1, \ldots, X_n)$, where $X_i \sim X$ and $X \sim F(\,\cdot\,)$. Hence, **it is not necessary $X \sim N(\mu, \sigma)$ as long as n is large enough. Typically, the rule of thumb is $n > 30$.**

- The next page contains a graph when $X \sim Exp(2)$ for different values of the sample size n.

23 Confidence Intervals for the Mean

23.2 Non-Normal Data, Variance Unknown . . .

Analysis in file "Student_T_Expon_Sample.R"



23 Confidence Intervals for the Mean

23.2 Normal Data, Variance Unknown . . .

Analysis in file "Student_T_Normal_Sample.R"