EMSE 4765: DATA ANALYSIS

For Engineers and Scientists

Session 4: Estimator Distributions, Confidence Intervals, Hypothesis Testing

Version: 2/1/2021

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

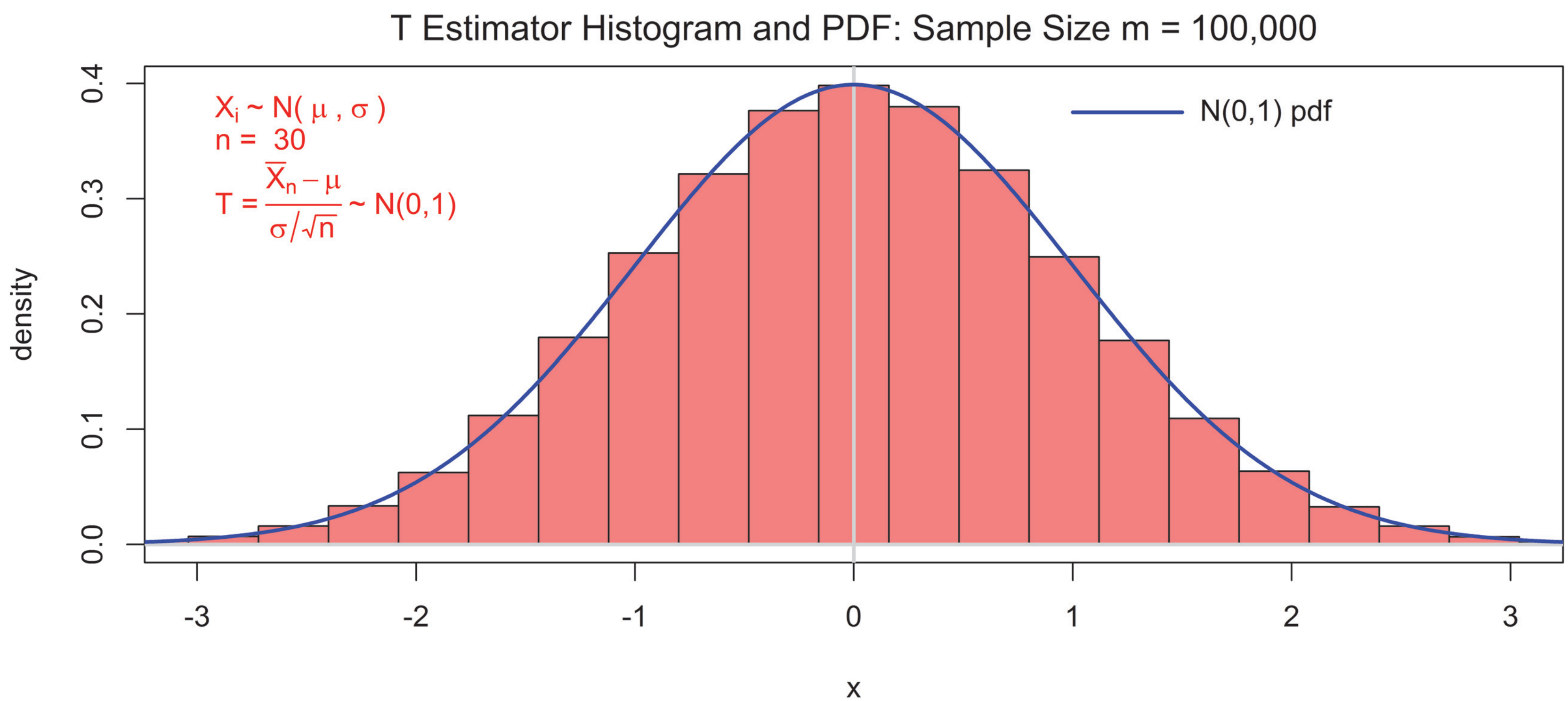
Lecture Notes by: J. René van Dorp[1]

www.seas.gwu.edu/~dorpjr

[1] Department of Engineering Management and Systems Egineering, School of Engineering and Applied Science, The George Washington University, 800 22nd Street, N.W., Suite 2800, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.

- Let $(X_1, \ldots, X_n)$ be a random $i.i.d.$ sample, where $X_i \sim N(\mu, \sigma) \Rightarrow$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sim Normal(0,1)$$ - **standard normal distribution**

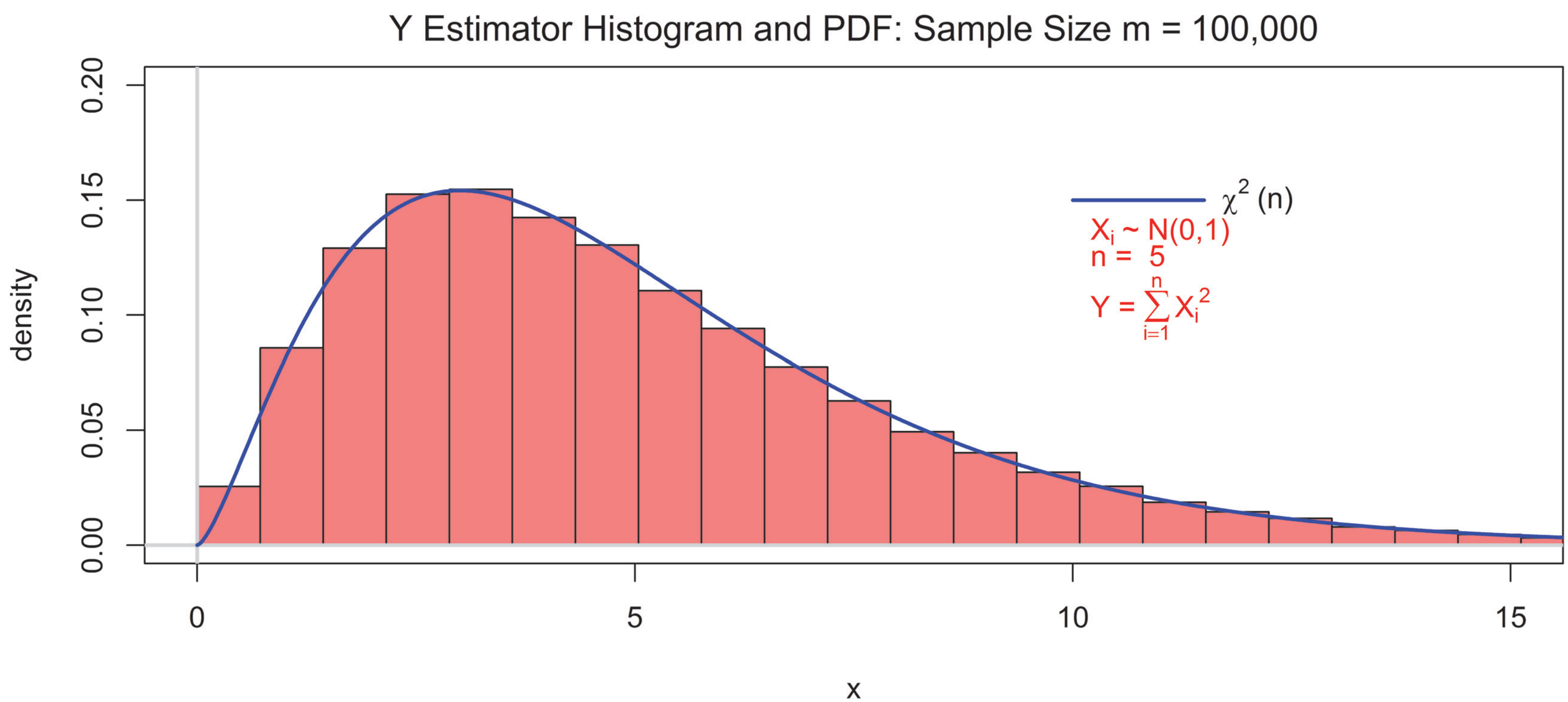


Analysis in "T_Variance_Known.R"

```
# n is the sample size of the dataset, mu and sigma are arbitrary chosen values
n<-30
the_mu<-5
the_sigma<-2

# m is the the number of datasets randomly generated.
m<-100000
t_stat<-replicate(m,0)
for (i in c(1:m))
{
  sample<-rnorm(n,the_mu,the_sigma)
  x_bar<-mean(sample)
  st_dev_x_bar<-the_sigma/(n^0.5)
  t_stat[i]<-(x_bar-the_mu)/st_dev_x_bar
}

# N_bins is the number of intervals for the histogram of the Estimator.
N_bins<-25
LB<--4
UB<-4
bins<-c(1:N_bins)/N_bins
bins<-(UB-LB)*bins+LB
t_probs<-Estimate_empirical_histogram(bins,t_stat)
bins<-c(LB,bins)
bin_widths<-bins[2:(N_bins+1)]-bins[1:N_bins]

# Evaluating the theoretical values for the pdf of the Estimator
x_points<-c(0:1000)/1000
LB<--4
UB<-4
x_points<-(UB-LB)*x_points+LB
n_points<-dnorm(x_points,0,1)
```

- Let $(X_1, \ldots, X_n)$ be a random $i.i.d.$ sample, where $X_i \sim N(0,1) \Rightarrow$:

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$ - **Chi-squared distribution with n degrees of freedom**

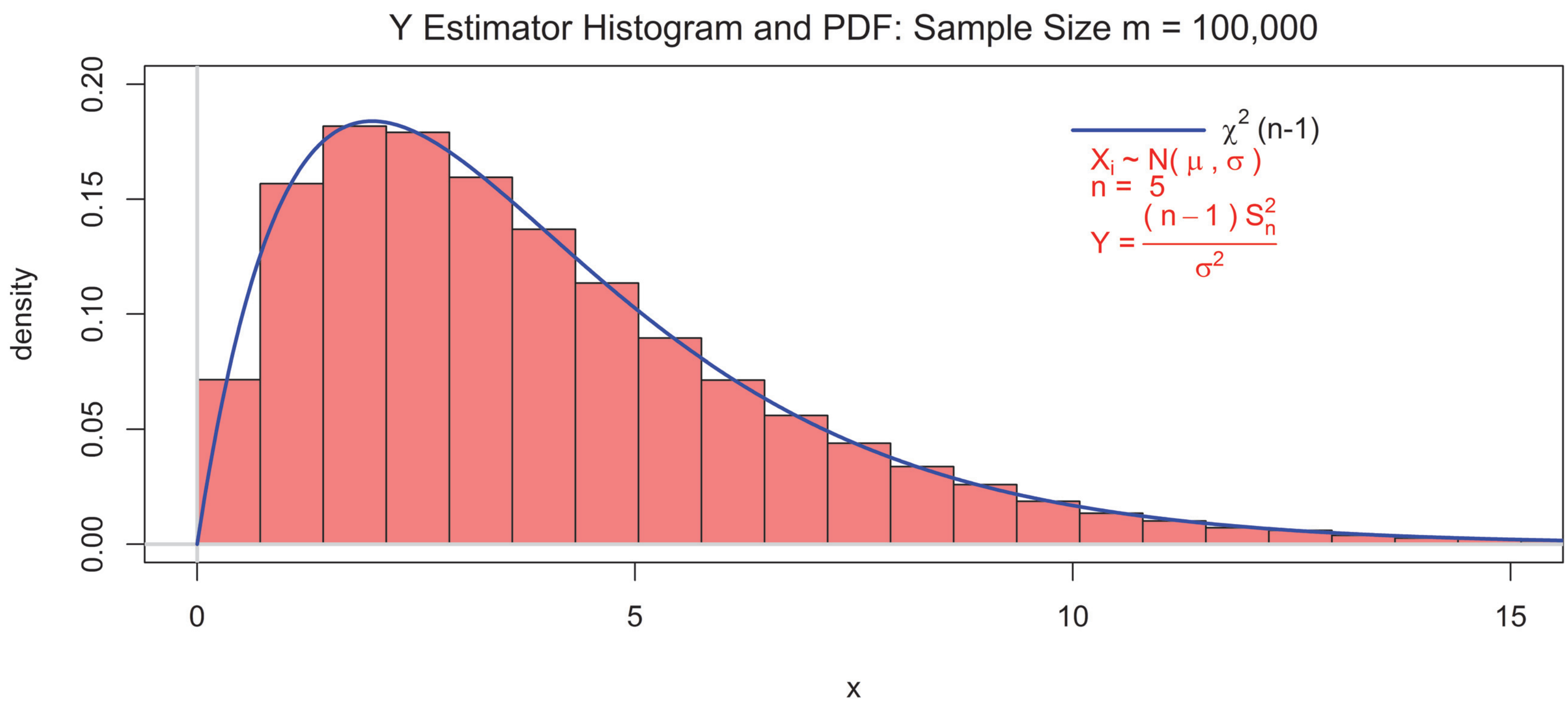


Analysis in "Chi_Squared_n.R"

- Let $(X_1, \ldots, X_n)$ be a random $i.i.d.$ sample, where $X_i \sim N(\mu, \sigma) \Rightarrow$

$$Y = \sum_{i=1}^{n} \left[\frac{X_i - \overline{X}}{\sigma} \right]^2 = \frac{\mathbf{n-1}}{\sigma^2} \frac{1}{\mathbf{n-1}} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Chi-squared distribution with $(n-1)$ degrees of freedom

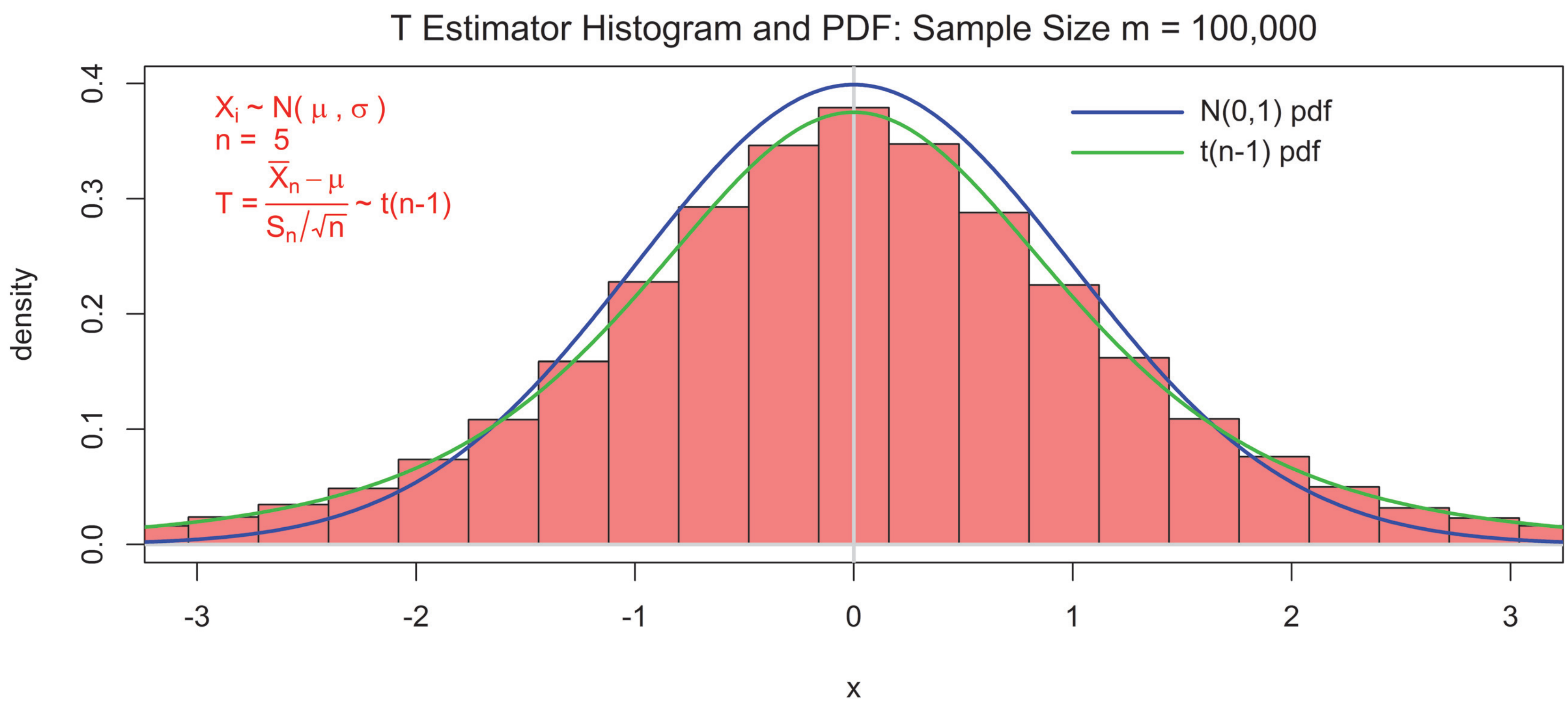


Analysis in "Chi_Squared_Minus_One.R"

- Let $(X_1, \ldots, X_n)$ be a random $i.i.d.$ sample, where $X_i \sim N(\mu, \sigma) \Rightarrow$:

$$\frac{\overline{X} - \mu}{S_n/\sqrt{n}} = \left[\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right] \Big/ \left[\frac{S_n}{\sigma}\right] \sim \frac{Normal(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}$$

Student-t distribution with $(n-1)$ degrees of freedom

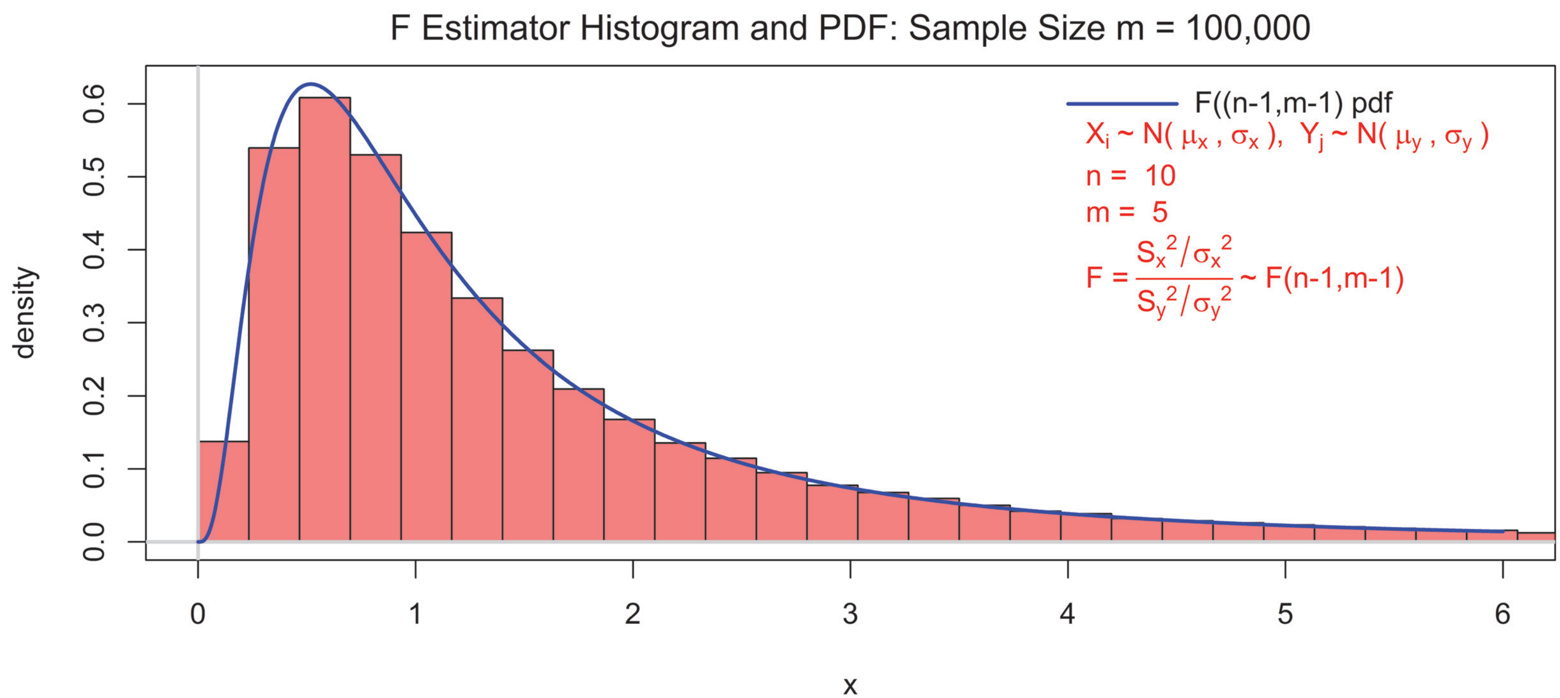


Analysis in "T_Variance_Unknown.R"

- Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be a random $i.i.d.$ samples $X_i \sim N(\mu_x, \sigma_x)$, $Y_i \sim N(\mu_y, \sigma_y)$ (Y_j's independent of the X_i's) $\Rightarrow$

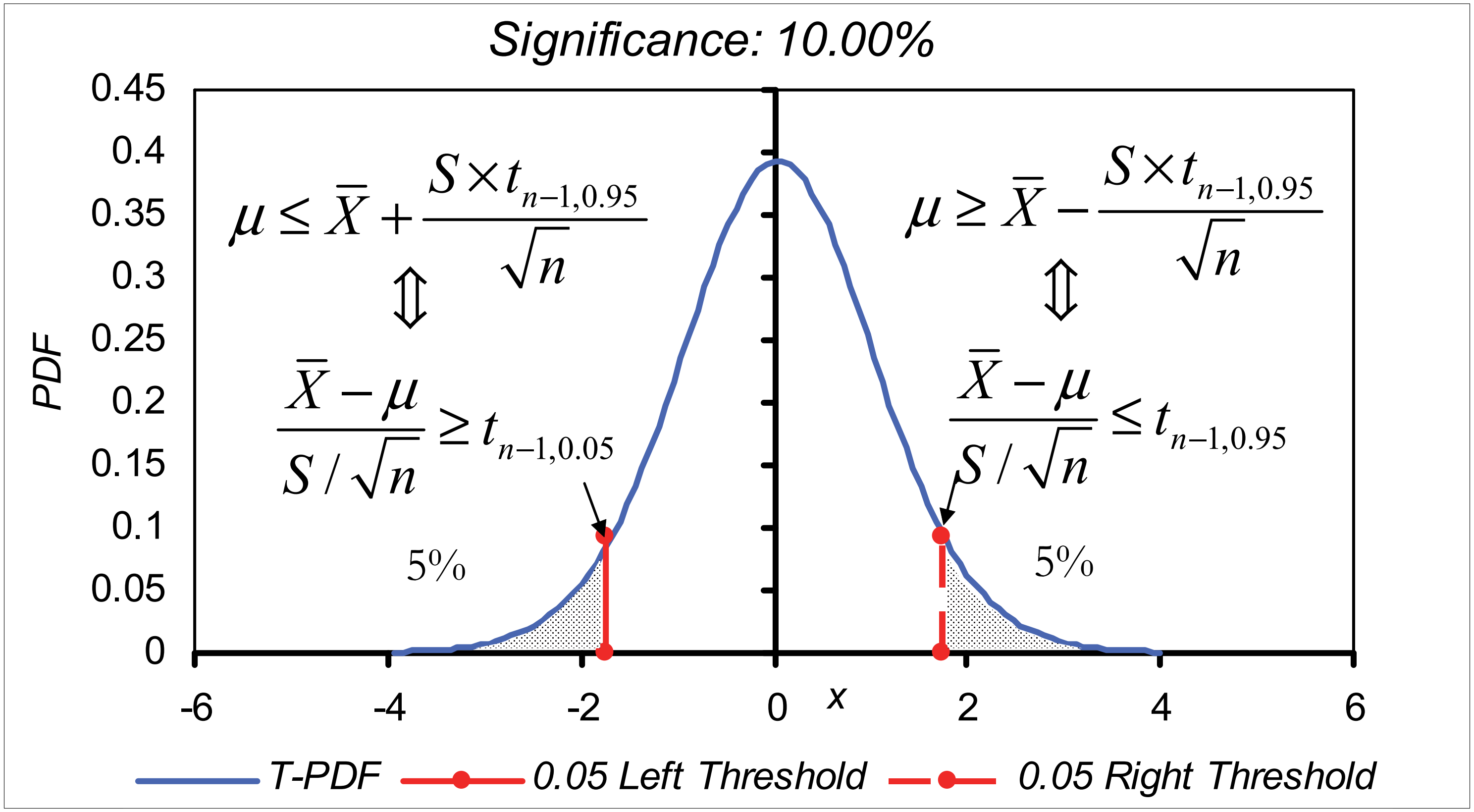
$$\frac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma_x}\right)^2\right] \Big/ \left[\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{Y_i-\overline{Y}}{\sigma_y}\right)^2\right] \sim F_{n-1,m-1}$$

F distribution with $n-1$ and $m-1$ degrees of freedom



Analysis in "F_Estimator.R"

Estimator distributions are important to determine **confidence intervals**.



90% Two-Sided Confidence Interval:

$$[\overline{x} - t_{n-1,0.95} \times \tfrac{s}{\sqrt{n}}, \overline{x} + t_{n-1,0.95} \times \tfrac{s}{\sqrt{n}}]$$

- 90% **two-sided confidence interval** for mean μ is **a realization** of a **random interval** with **two random bounds** because $\overline{X}$ and S are random variables.

- One obtains this 90% **two-sided confidence interval** with two **fixed bounds** by substituting **estimate** $\overline{x}$ for **estimator** $\overline{X}$ and **estimate** s for **estimator** S.

Example 15 (Continued): Dielectric breakdown voltage data

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

$$\overline{x} \approx 27.793,\ s^2 \approx 2.14,\ t_{19,0.95} = 1.73$$

90% Two-Sided Confidence Interval for μ:

$$[27.793 - \frac{1.73 \times \sqrt{2.14}}{\sqrt{20}}, 27.793 + \frac{1.73 \times \sqrt{2.14}}{\sqrt{20}}] = [27.23, 28.36]$$

Analysis in "Voltage_Mu_Sigma_X_Intervals.R"

Excel - Analysis

Mean Confidence Interval	X-Bar	27.793
	Var X	2.137
	St. Dev. X	1.462
	n	20
	St. Dev. X-Bar	0.32688
	α	10%
	$t_{n-1,1-\alpha/2}$	1.729133
	LB	27.22778
	UB	28.35822
	$Pr(T>t_{n-1,1-\alpha/2})$	0.05

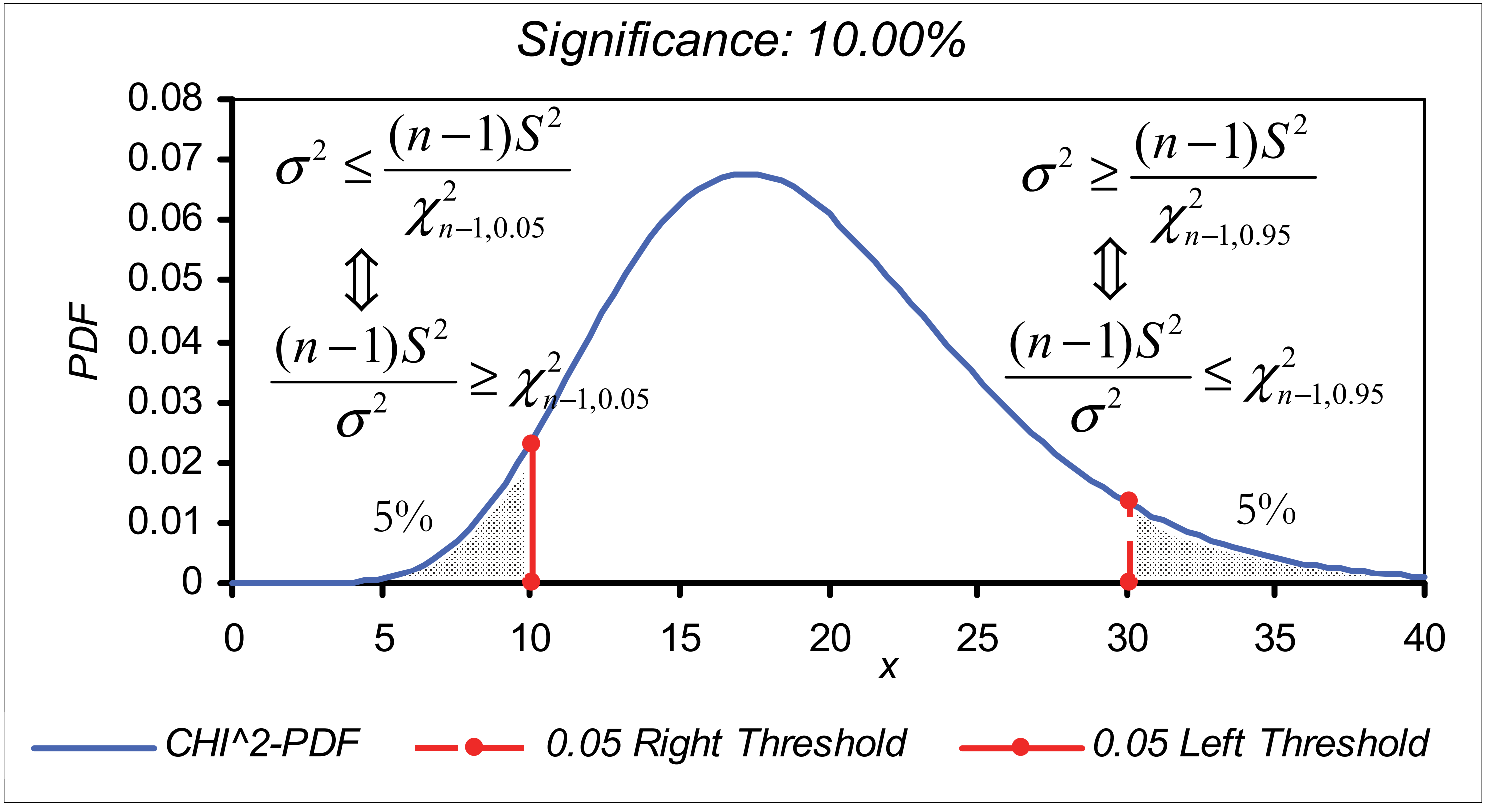
R - Code

```
# loading the readr package
library(readr)
Voltage <- read_csv("Voltage.csv")

# Assigning First Column to Volt
Volt=Voltage[[1]]
mu_0=27

# Evaluating Confidence interval for the mean
alpha=0.10
t.test(Volt, conf.level = 1-alpha, mu = mu_0)
```

R - Output

```
        One Sample t-test

data:  Volt
t = 2.426, df = 19, p-value = 0.02539
alternative hypothesis: true mean is not equal to 27
90 percent confidence interval:
 27.22778 28.35822
sample estimates:
mean of x
   27.793
```

Estimator distributions are important to determine **confidence intervals**.



90% Two-Sided Confidence Interval: $\sigma^2 \in \left[\frac{(n-1)s^2}{\chi^2_{n-1,0.95}}, \frac{(n-1)s^2}{\chi^2_{n-1,0.05}}\right]$

- 90% **two-sided confidence interval** for variance σ^2 is **a realization** of a **random interval** with two random bounds because S^2 is a random variable.

- One obtains this 90% **two-sided confidence interval** with two **fixed bounds** by substituting **the estimate** s^2 for **the estimator** S^2.

Example 15 (Continued): Dielectric breakdown voltage data

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

$$s^2 \approx 2.14,\ \chi^2_{19,0.05} = 10.12,\ \chi^2_{19,0.95} = 30.14$$

90% Two-Sided Confidence Interval for σ^2:

$$[\frac{19 \times 2.14}{30.14}, \frac{19 \times 2.14}{10.12}] = [1.347, 4.014]$$

Analysis in "Voltage_Mu_Sigma_X_Intervals.R"

R - Code

```
# Evaluating Confidence interval for the variance
alpha=0.10
n<-length(Volt)
Variance_Estimate<-var(Volt)
Chi_05=qchisq(alpha/2, n-1)
Chi_95=qchisq(1-alpha/2, n-1)
LL_Var= (n-1)*Variance_Estimate/Chi_95
UPP_Var=(n-1)*Variance_Estimate/Chi_05
LL_Var
UPP_Var
```

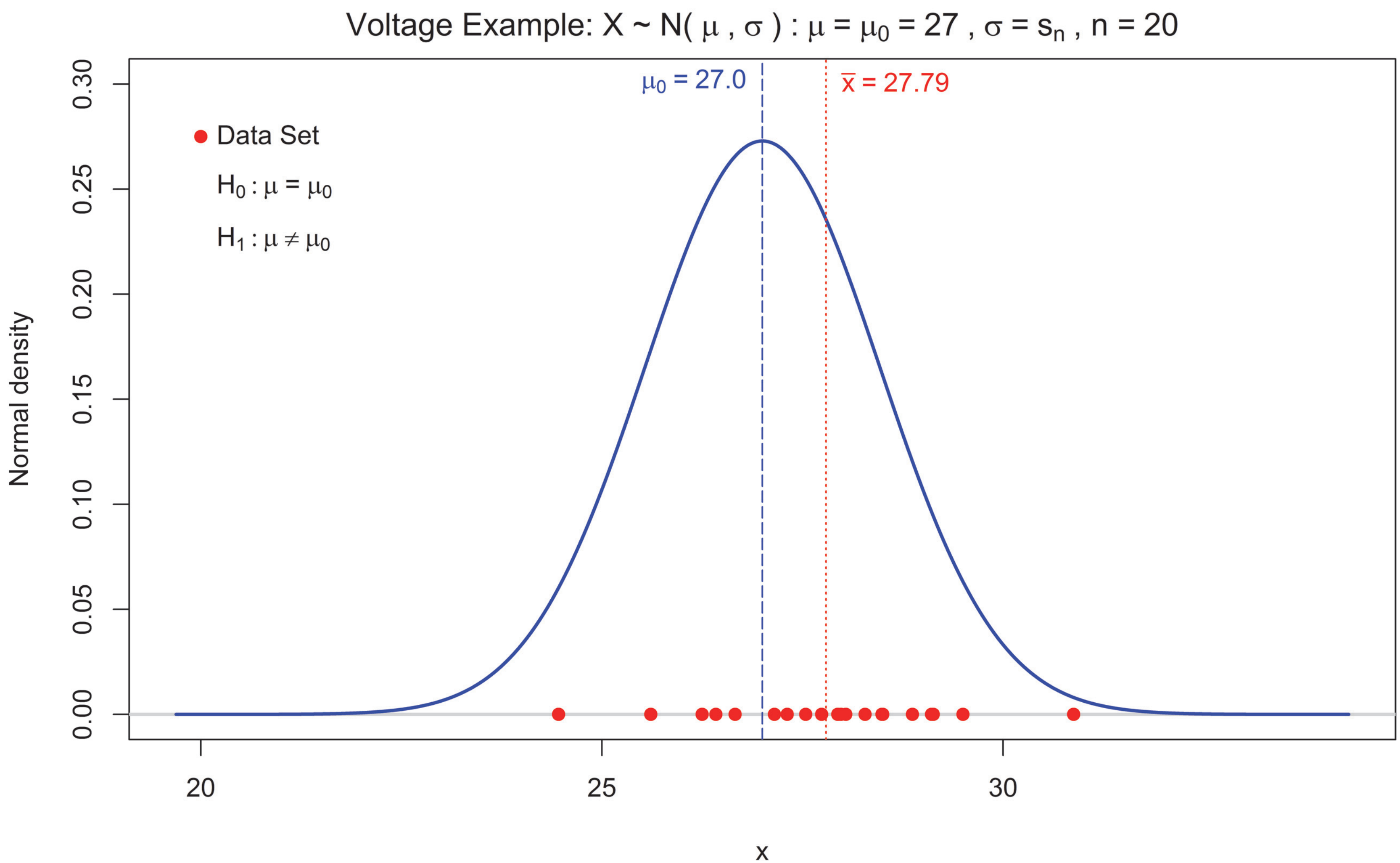
Excel - Analysis

Variance Confidence Interval		
	$\chi^2_{n-1,\alpha/2}$	10.117013
	$\chi^2_{n-1,1-\alpha/2}$	30.143527
	LB	1.347003
	UB	4.013380

R - Output

```
> # Evaluating Confidence interval for the variance
> alpha=0.10
> n<-length(Volt)
> Variance_Estimate<-var(Volt)
> Chi_05=qchisq(alpha/2, length(Volt)-1)
> Chi_95=qchisq(1-alpha/2, length(Volt)-1)
> LL_Var= (n-1)*Variance_Estimate/Chi_95
> UPP_Var=(n-1)*Variance_Estimate/Chi_05
> LL_Var
[1] 1.347003
> UPP_Var
[1] 4.01338
```

Analysis in file "Voltage_Hypothesis_1.R". Should we Reject H_0 in favor of H_1?



- There is a connection between **confidence intervals** and **hypothesis testing.** Let $(x_1, \ldots, x_n)$ be an $i.i.d.$ sample from a **normal distribution** with unknown mean μ and unknown variance σ^2. Consider the hypothesis test.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

High values and low values of $\overline{x}$ are an **indication of support for the alternative hypothesis.** High values and low values of $\overline{x}$ go together with high and low values of the following **t_0 estimate**

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$, Note that: **Estimator $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ if H_0 is true**

- **How high or low do we let $\overline{x}$ (or t_0) get before we reject the null hypothesis?** This is determined by **the distribution of T** and **the significance level α that you specify.** For a two-sided test we divide the significance level α, say 10%, by two **for two tails with equal probability** and by convention:

Too high a value of $\overline{x} \Leftrightarrow t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} > t_{n-1,0.95}$

Too low a value of $\overline{x} \Leftrightarrow t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} < t_{n-1,0.05} = -t_{n-1,0.95}$

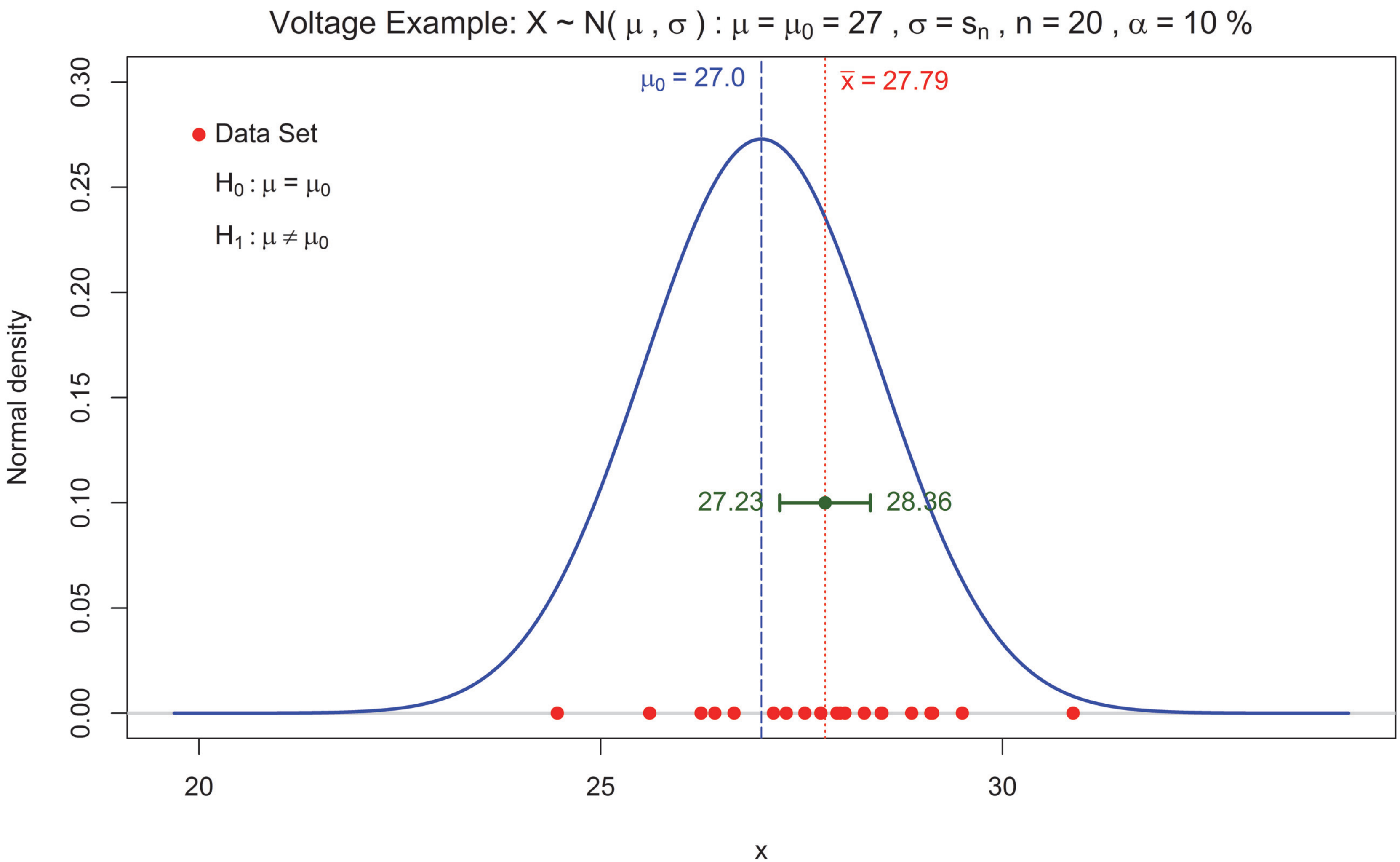
- **Conclusion:**

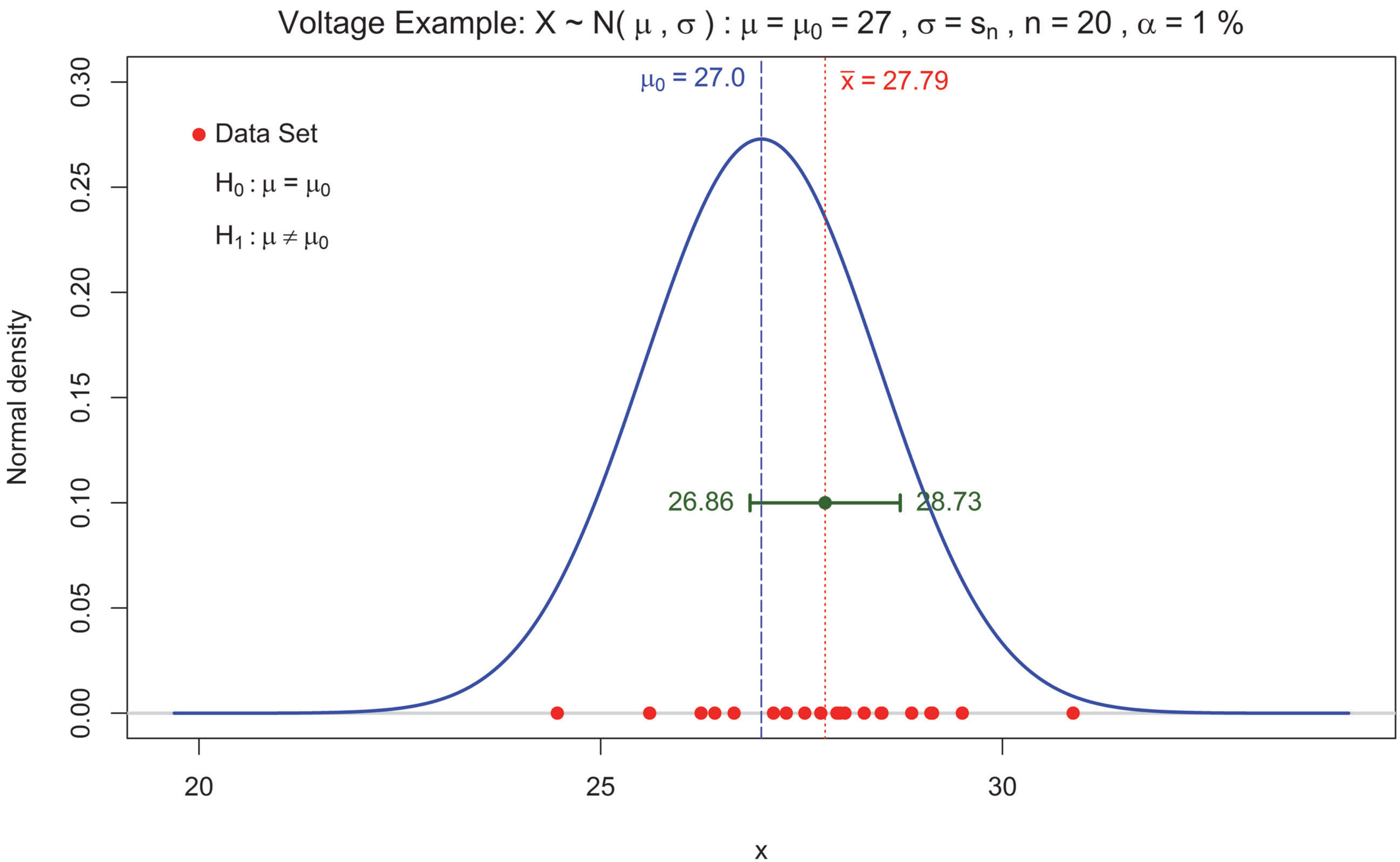
$$\begin{cases} \text{we reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \notin [t_{n-1,0.05}, t_{n-1,0.95}] \\ \text{we fail to reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \in [t_{n-1,0.05}, t_{n-1,0.95}] \end{cases}$$

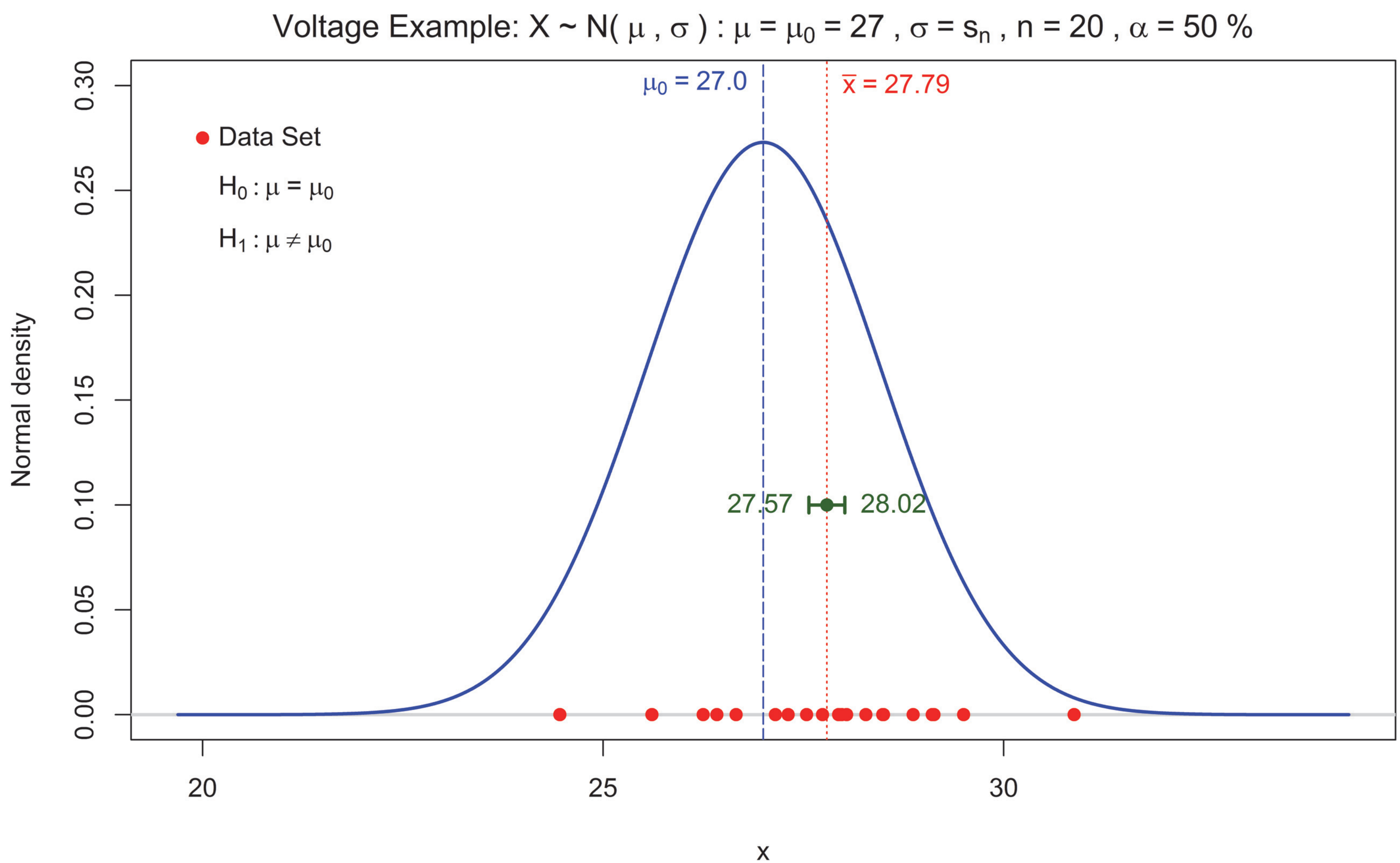
which is equivalent to:

$$\begin{cases} \text{we reject } H_0 \text{ if } \mu_0 \notin [\overline{x} - t_{n-1,0.95} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,0.95} \times \frac{s}{\sqrt{n}}] \\ \text{we fail to reject } H_0 \text{ if } \mu_0 \in [\overline{x} - t_{n-1,0.95} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,0.95} \times \frac{s}{\sqrt{n}}] \end{cases}$$

- **But:** $[\overline{x} - t_{n-1,0.95} \times \frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,0.95} \times \frac{s}{\sqrt{n}}]$ **is the 90% confidence interval for μ.**

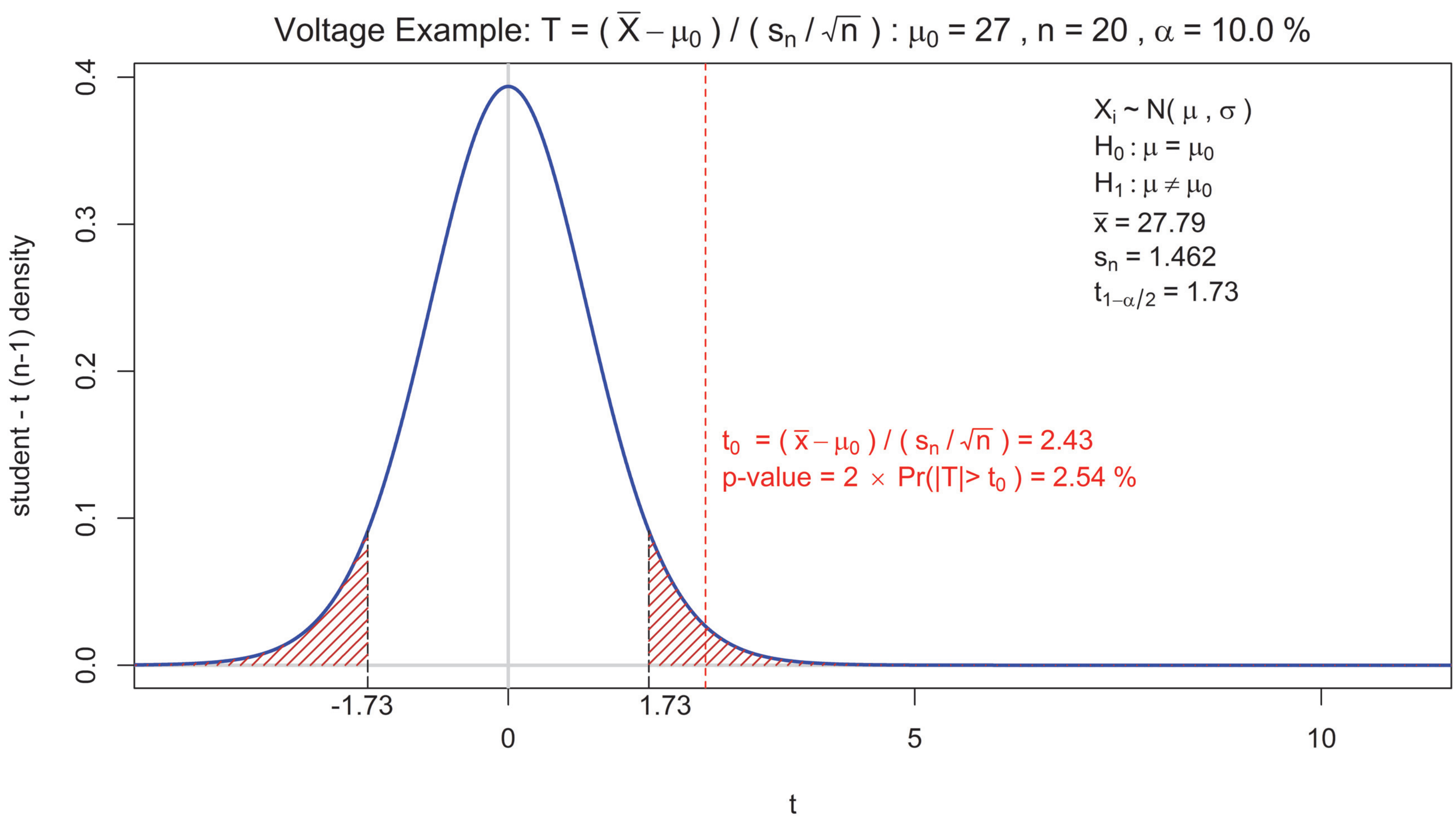

Voltage Example: X ~ N(μ , σ) : μ = μ0 = 27 , σ = sn , n = 20 , α = 10 %
μ0 = 27.0
x̄ = 27.79
Data Set
H0 : μ = μ0
H1 : μ ≠ μ0
27.23
28.36
Normal density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
20
25
30
x


Voltage Example: X ~ N(μ , σ) : μ = μ0 = 27 , σ = sn , n = 20 , α = 1 %
μ0 = 27.0
x̄ = 27.79
Data Set
H0 : μ = μ0
H1 : μ ≠ μ0
26.86
28.73
Normal density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
20
25
30
x


Voltage Example: X ~ N(μ , σ) : μ = μ0 = 27 , σ = sn , n = 20 , α = 50 %
μ0 = 27.0
x̄ = 27.79
Data Set
H0 : μ = μ0
H1 : μ ≠ μ0
27.57
28.02
Normal density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
20
25
30
x

Scenario where we reject H_0 based on the value of t_0 - estimate

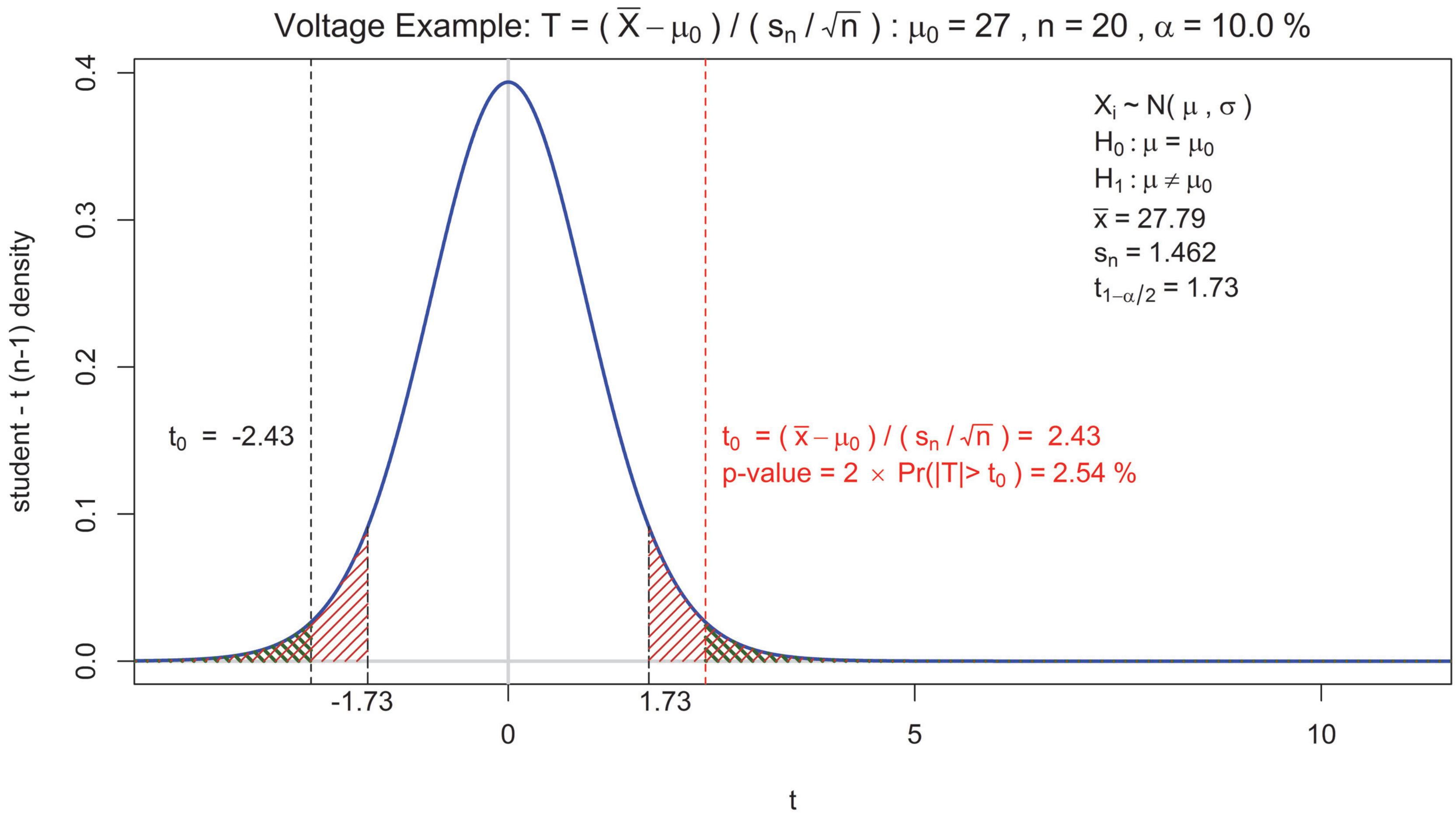
$$\text{Type 1 error} = Pr(\text{Reject } H_0 | H_0 \text{ is true}) = Pr(\text{Reject } H_0 | \mu = \mu_0) = \alpha$$



Analysis in file "Voltage_Hypothesis_2.R"

Scenario where we reject H_0 based on the value of t_0 - estimate

$$p\text{-value} = Pr(T_{n-1} \notin [-t_0, t_0]) = 2Pr(T_{n-1} > t_0) < \alpha$$



Analysis in file "Voltage_Hypothesis_3.R"

Definition: The **p-value** of an hypothesis test is **the largest significance level** at which we would just fail to reject the null-hypothesis. It is also **the probability of observing something more extreme than you have observed.**

- p-value $= Pr(T_{n-1} \notin [-t_0, t_0]) = 2Pr(T_{n-1} > t_0)$. Small p-values indicate level of evidence against the null-hypothesis.

Example 15 (Continued): Dielectric breakdown voltage data

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

$$\overline{x} \approx 27.793,\ s^2 \approx 2.14,\ \alpha = 0.10,\ \ t_{19,0.05} = -1.73,\ t_{19,0.95} = 1.73$$

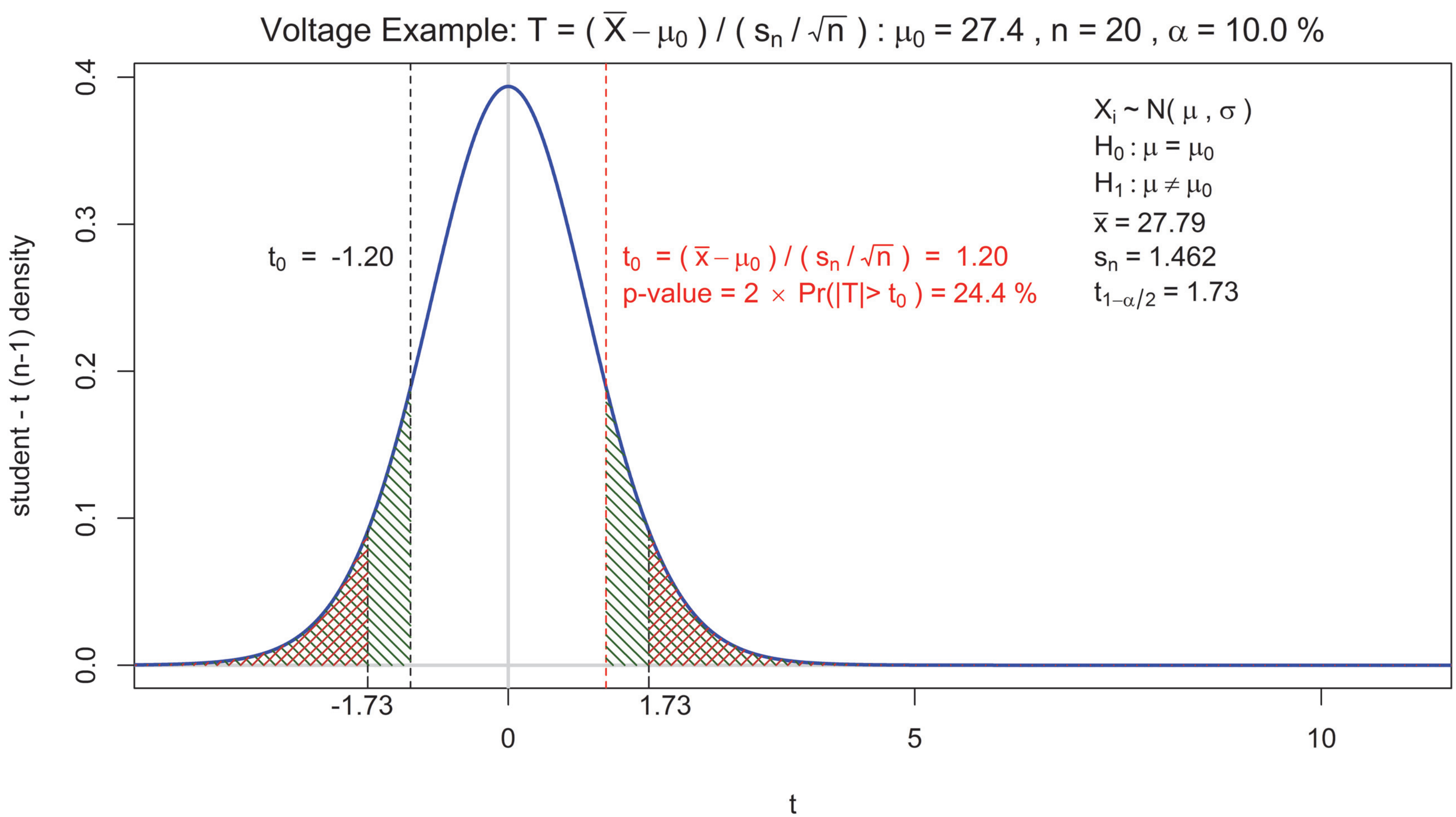
$$H_0 : \mu = 27,\ H_1 : \mu \neq 27 \ \Rightarrow t_0 = \frac{\overline{x} - 27}{\sqrt{2.14}/\sqrt{20}} \approx 2.43$$

Conclusion:

$$t_0 \notin [-1.73, 1.73] \Rightarrow \text{Reject } H_0,\ p\text{ -value} = 2Pr(T_{19} > 2.43) \approx 2.54\%$$

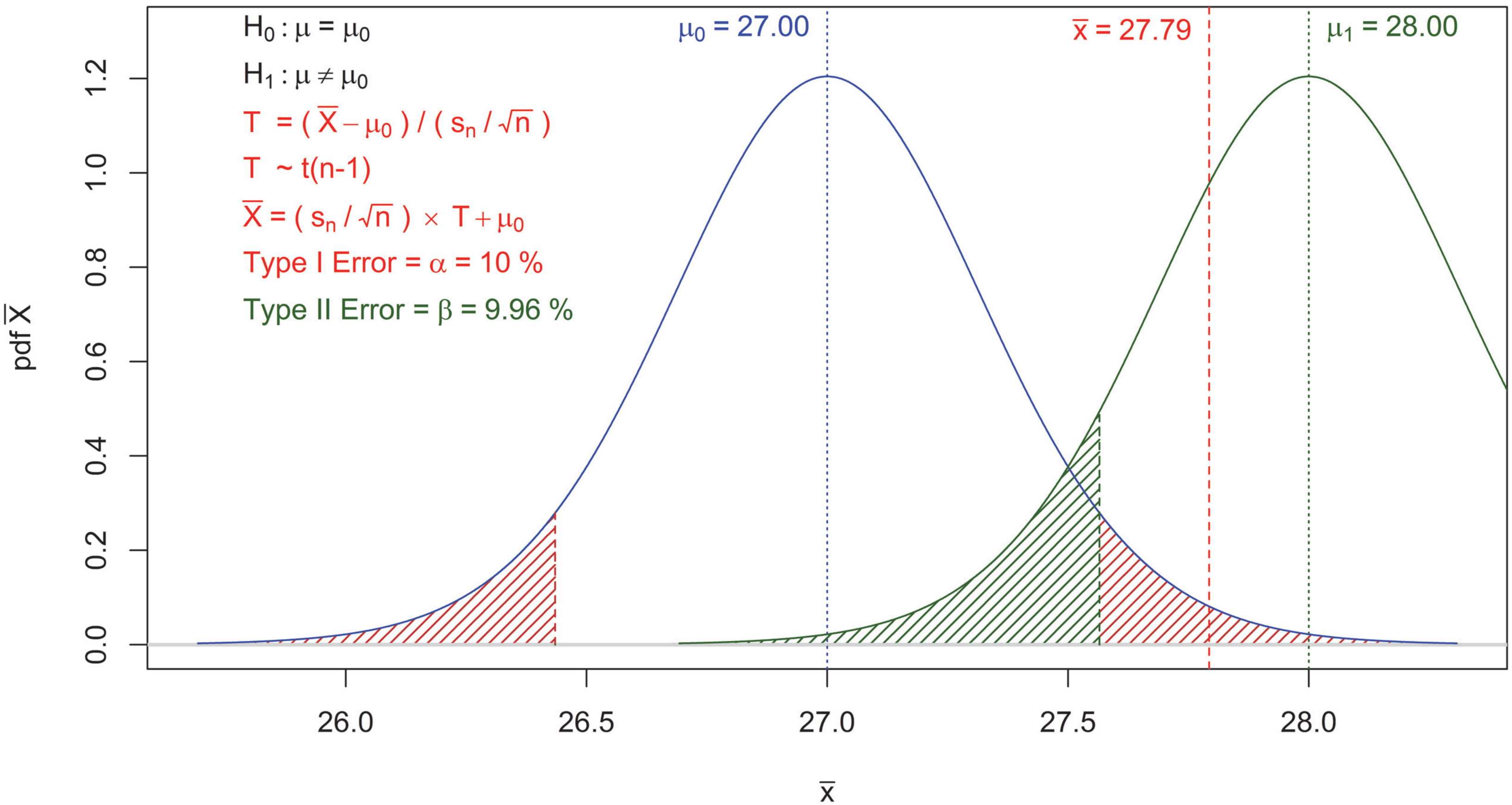
Scenario where we fail to reject H_0 based on the value of t_0 - estimate

$$p\text{-value} = Pr(T_{n-1} \notin [-t_0, t_0]) = 2Pr(T_{n-1} > t_0) > \alpha$$



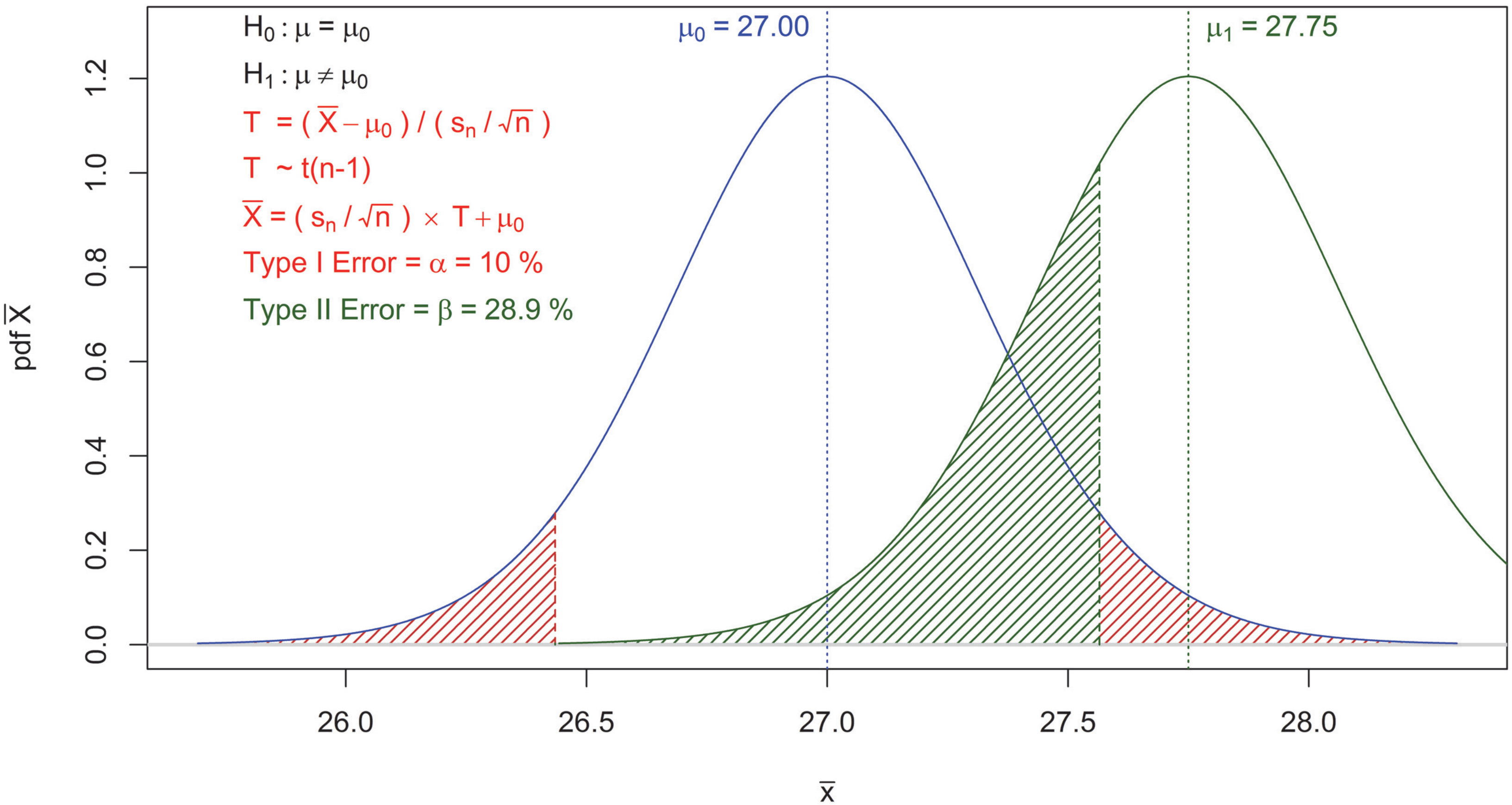
Analysis in file "Voltage_Hypothesis_4.R"

$$\alpha = \text{Type I Error} = Pr(Reject\, H_0 | H_0\, is\, True)$$
$$\beta = \text{Type II Error} = Pr(Accept\, H_0 | H_0\, is\, False)$$



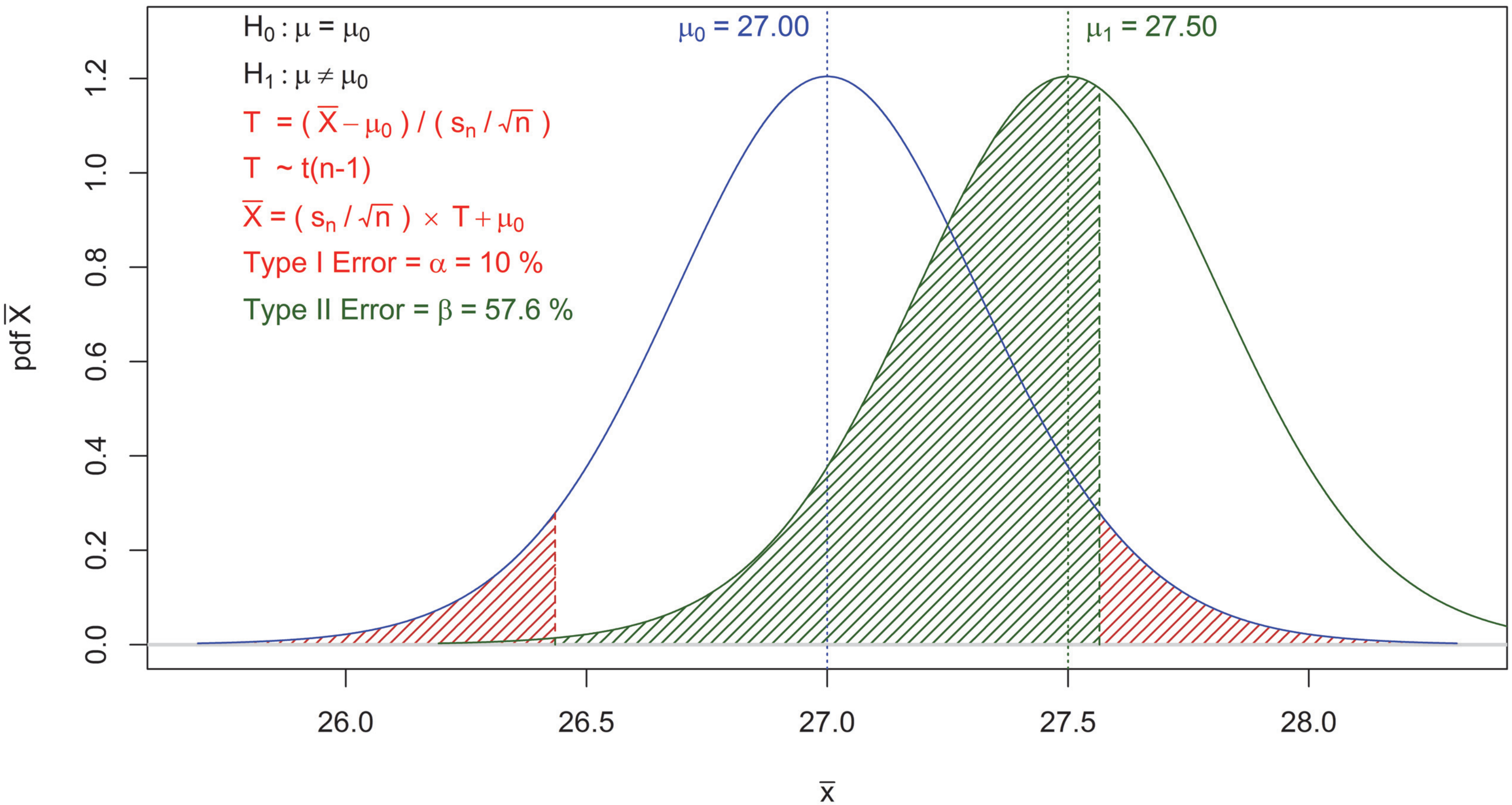
Analysis in file "Voltage_Type_II_Error.R"

$$\alpha = \text{Type I Error} = Pr(Reject\ H_0 | H_0\ is\ True)$$
$$\beta = \text{Type II Error} = Pr(Accept\ H_0 | H_0\ is\ False)$$



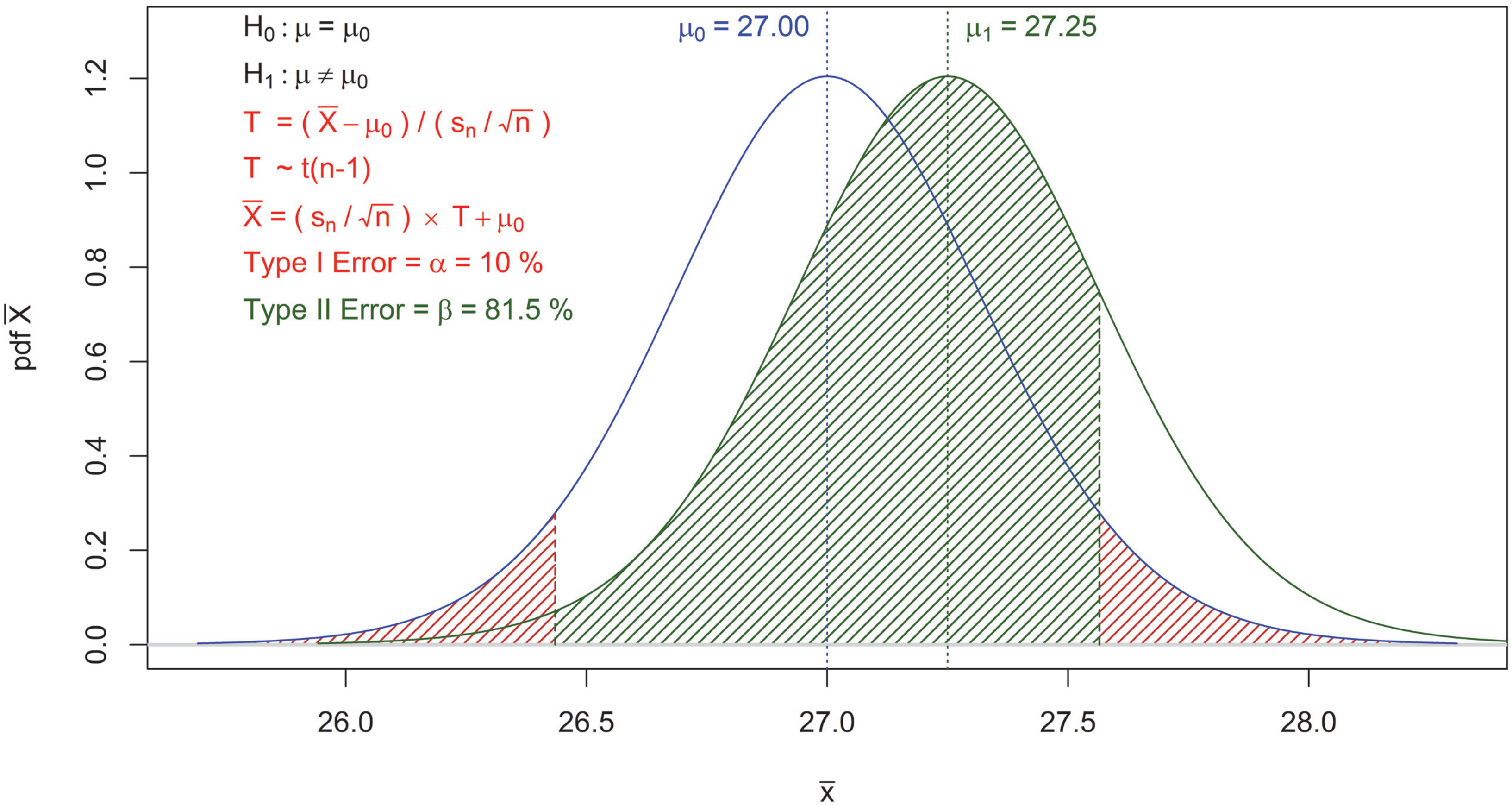
Analysis in file "Voltage_Type_II_Error.R"

$$\alpha = \text{Type I Error} = Pr(Reject\, H_0 | H_0\, is\, True)$$
$$\beta = \text{Type II Error} = Pr(Accept\, H_0 | H_0\, is\, False)$$



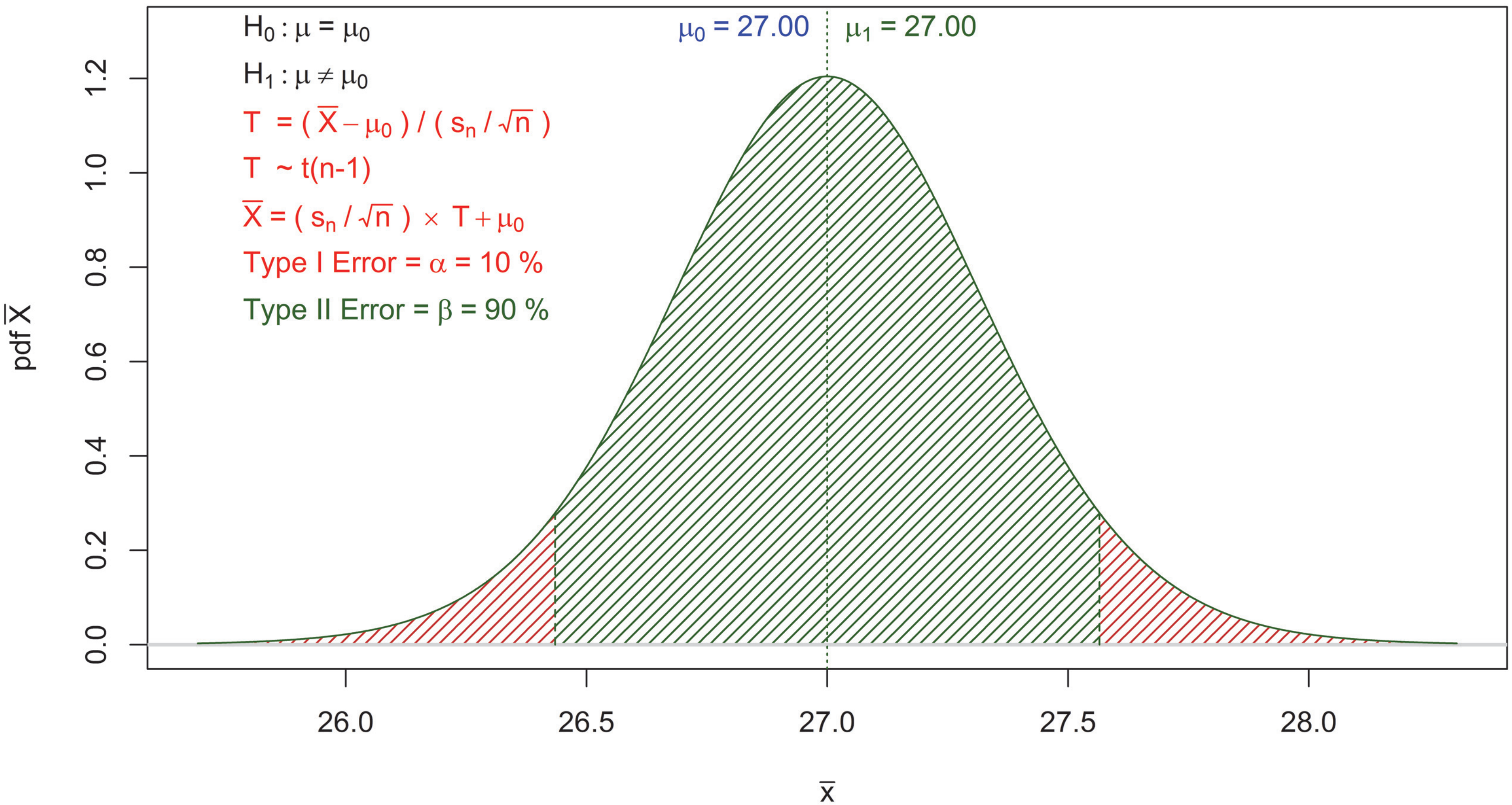
Analysis in file "Voltage_Type_II_Error.R"

$$\alpha = \text{Type I Error} = Pr(Reject\ H_0 | H_0\ is\ True)$$
$$\beta = \text{Type II Error} = Pr(Accept\ H_0 | H_0\ is\ False)$$



Analysis in file "Voltage_Type_II_Error.R"

$$\alpha = \text{Type I Error } = Pr(Reject\, H_0 | H_0\, is\, True)$$
$$\beta = \text{Type II Error } = Pr(Accept\, H_0 | H_0\, is\, False)$$



Analysis in file "Voltage_Type_II_Error.R"

Definition: Type 2 error $\beta = Pr(\text{fail to reject } H_0 | H_0 \text{ is false})$

- The value of β depends on how one defines : **"H_0 is false"** $\equiv \mu = \boldsymbol{\mu_1} \neq \mu_0$

$$\beta(\boldsymbol{\mu_1}) = Pr(\text{fail to reject } H_0 | \boldsymbol{H_0} \textbf{ is false}) = Pr(\text{fail to reject } H_0 | \mu = \boldsymbol{\mu_1})$$

$$= Pr(t_{n-1,0.05} \leq \frac{\overline{X} - \boldsymbol{\mu_0}}{\boldsymbol{S/\sqrt{n}}} \leq t_{n-1,0.95} | \mu = \boldsymbol{\mu_1})$$

$$= Pr(t_{n-1,0.05} + \frac{\boldsymbol{\mu_0}}{\boldsymbol{S/\sqrt{n}}} \leq \frac{\overline{X}}{\boldsymbol{S/\sqrt{n}}} \leq t_{n-1,0.95} + \frac{\boldsymbol{\mu_0}}{\boldsymbol{S/\sqrt{n}}} | \mu = \boldsymbol{\mu_1})$$

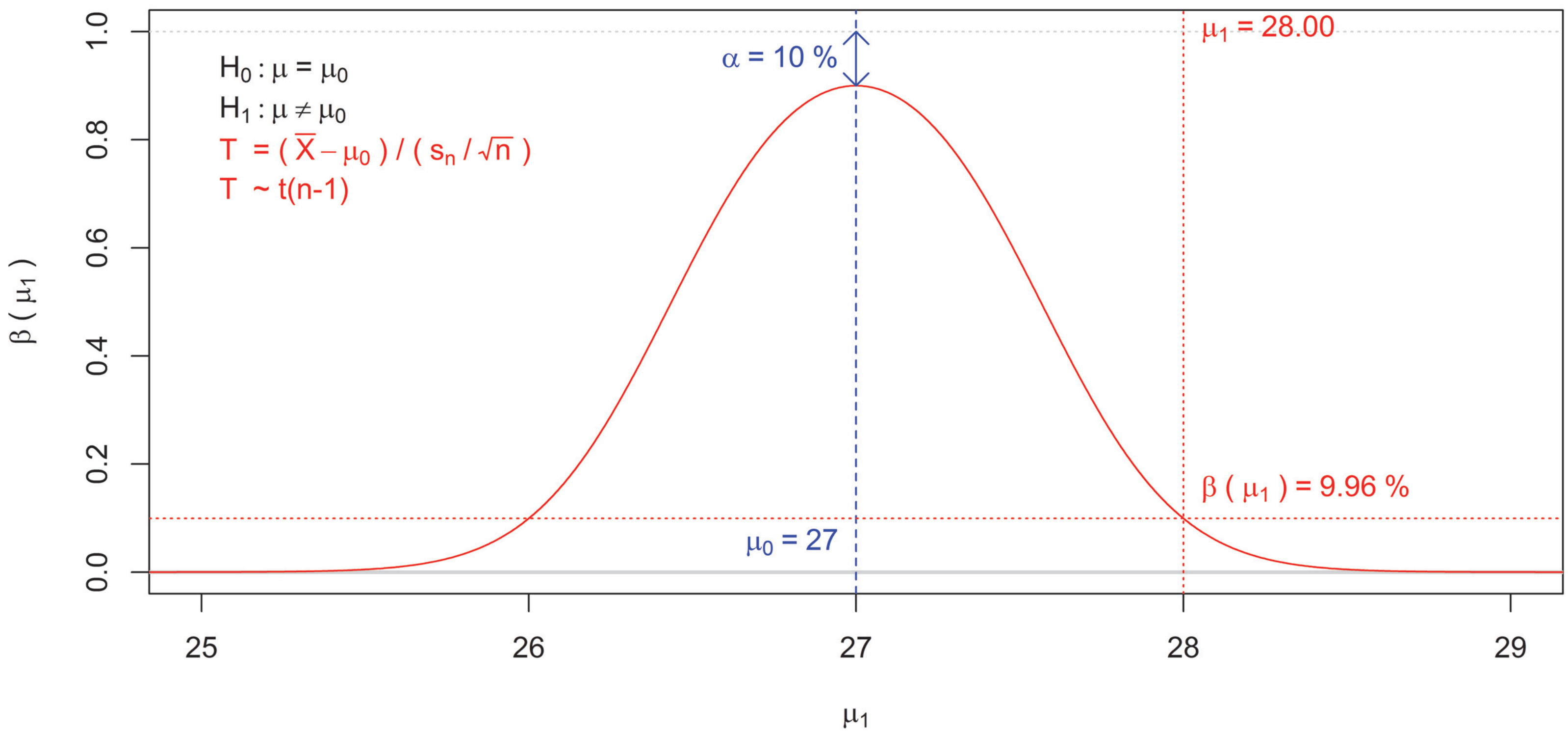
$$= Pr(-t_{n-1,0.95} + \frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{\boldsymbol{S/\sqrt{n}}} \leq \frac{\overline{X} - \boldsymbol{\mu_1}}{\boldsymbol{S/\sqrt{n}}} \leq t_{n-1,0.95} + \frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{\boldsymbol{S/\sqrt{n}}} | \mu = \boldsymbol{\mu_1})$$

Example 15 (Continued): Dielectric breakdown voltage data

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

- Use now $s \approx \sqrt{2.14}, n = 10,\ \alpha = 0.10,$ and **quantiles** $t_{19,0.05} = -1.73, t_{19,0.95} = 1.73$ to estimate Type II Error $\widehat{\beta}(\mu_1)$:

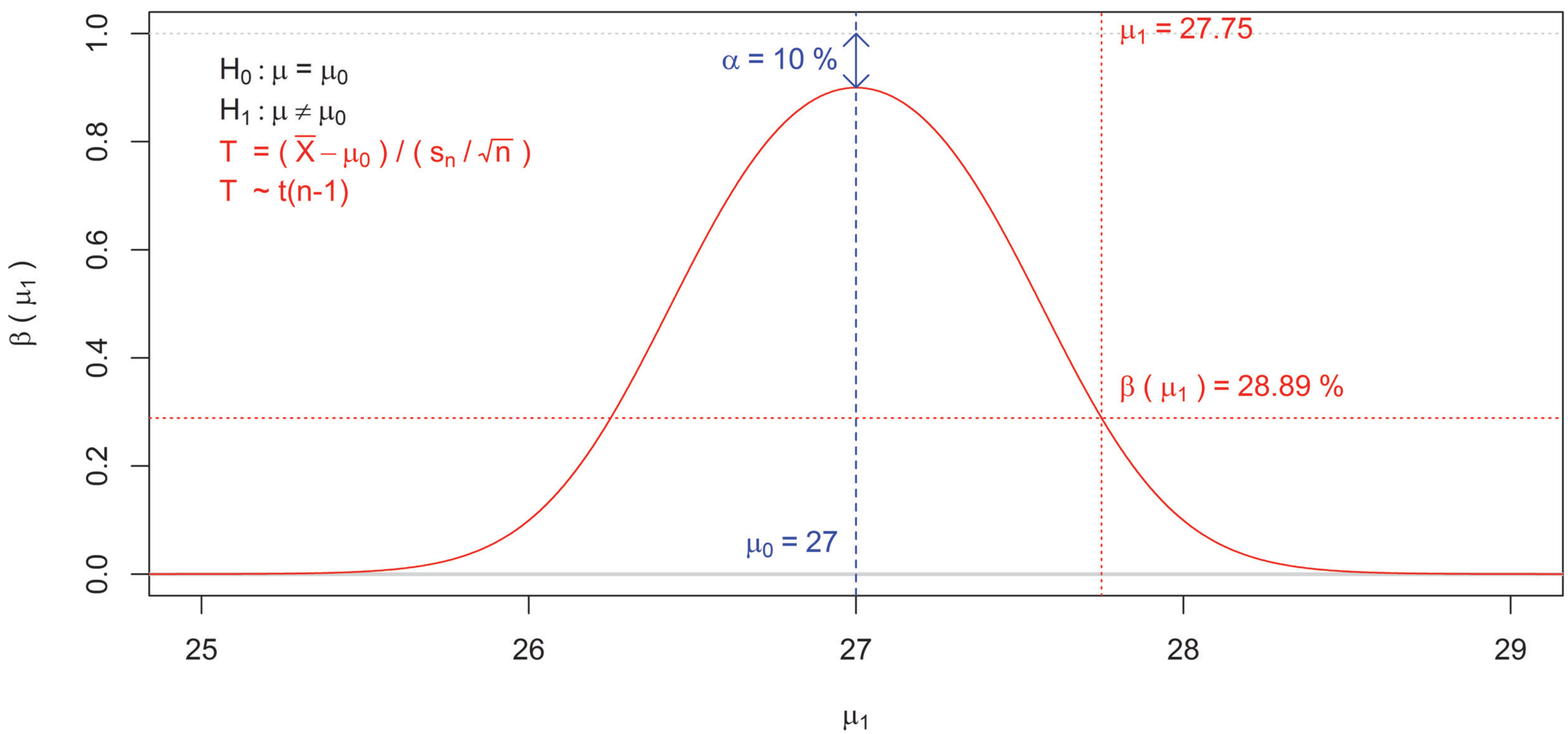
$$\widehat{\beta}(\mu_1) = Pr(t_{n-1,0.05} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} \leq T \leq t_{n-1,0.95} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}), \text{ where } T \sim T(19)$$



Analysis in file "Voltage_OC_Curve.R"

- Use now $s \approx \sqrt{2.14}, n = 10, \alpha = 0.10,$ and **quantiles** $t_{19,0.05} = -1.73, t_{19,0.95} = 1.73$ to estimate Type II Error $\widehat{\beta}(\mu_1)$:

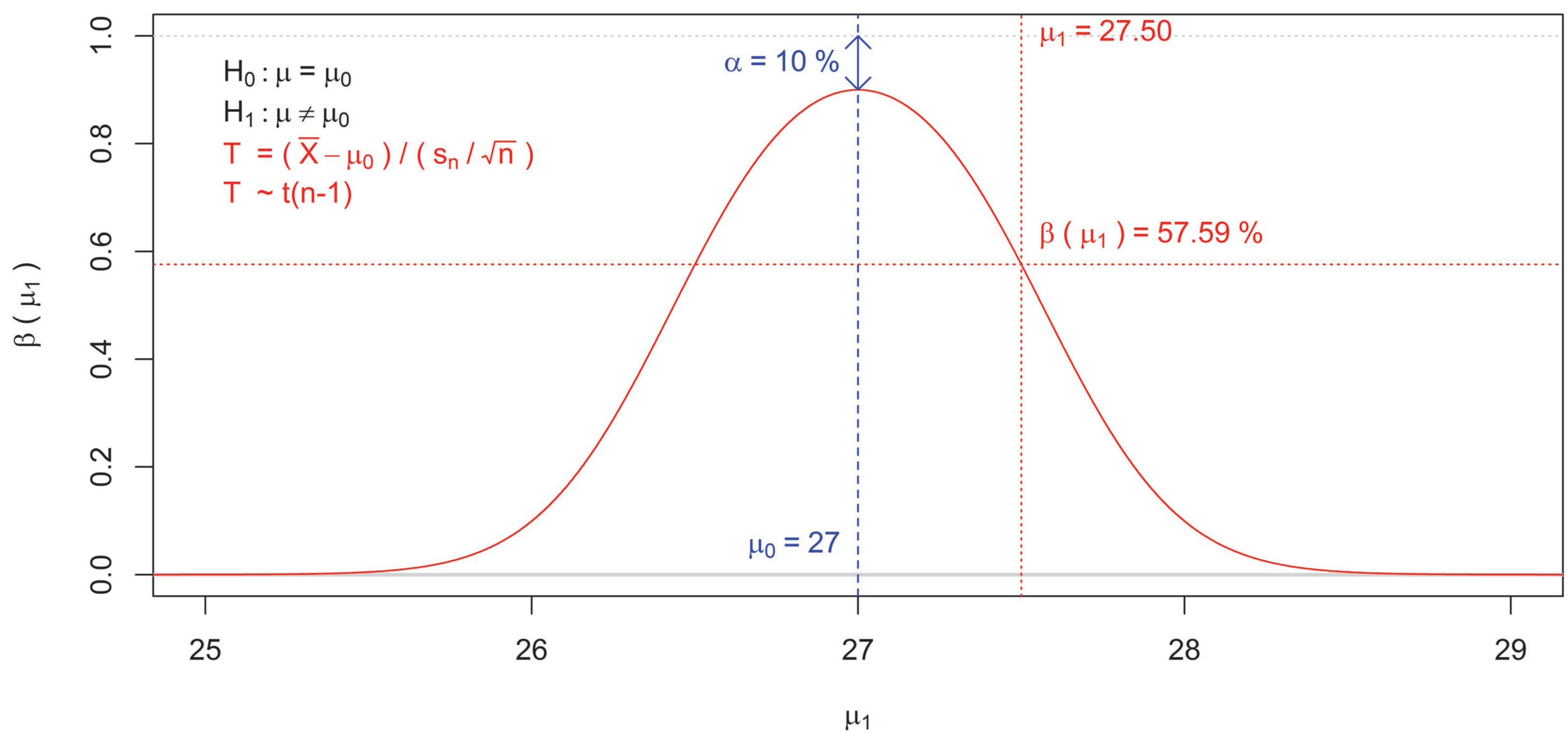
$$\widehat{\beta}(\mu_1) = Pr(t_{n-1,0.05} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} \le T \le t_{n-1,0.95} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}), \text{ where } T \sim T(19)$$



Analysis in file "Voltage_OC_Curve.R"

- Use now $s \approx \sqrt{2.14}$, $n = 10$, $\alpha = 0.10$, and **quantiles** $t_{19,0.05} = -1.73$, $t_{19,0.95} = 1.73$ to estimate Type II Error $\widehat{\beta}(\mu_1)$:

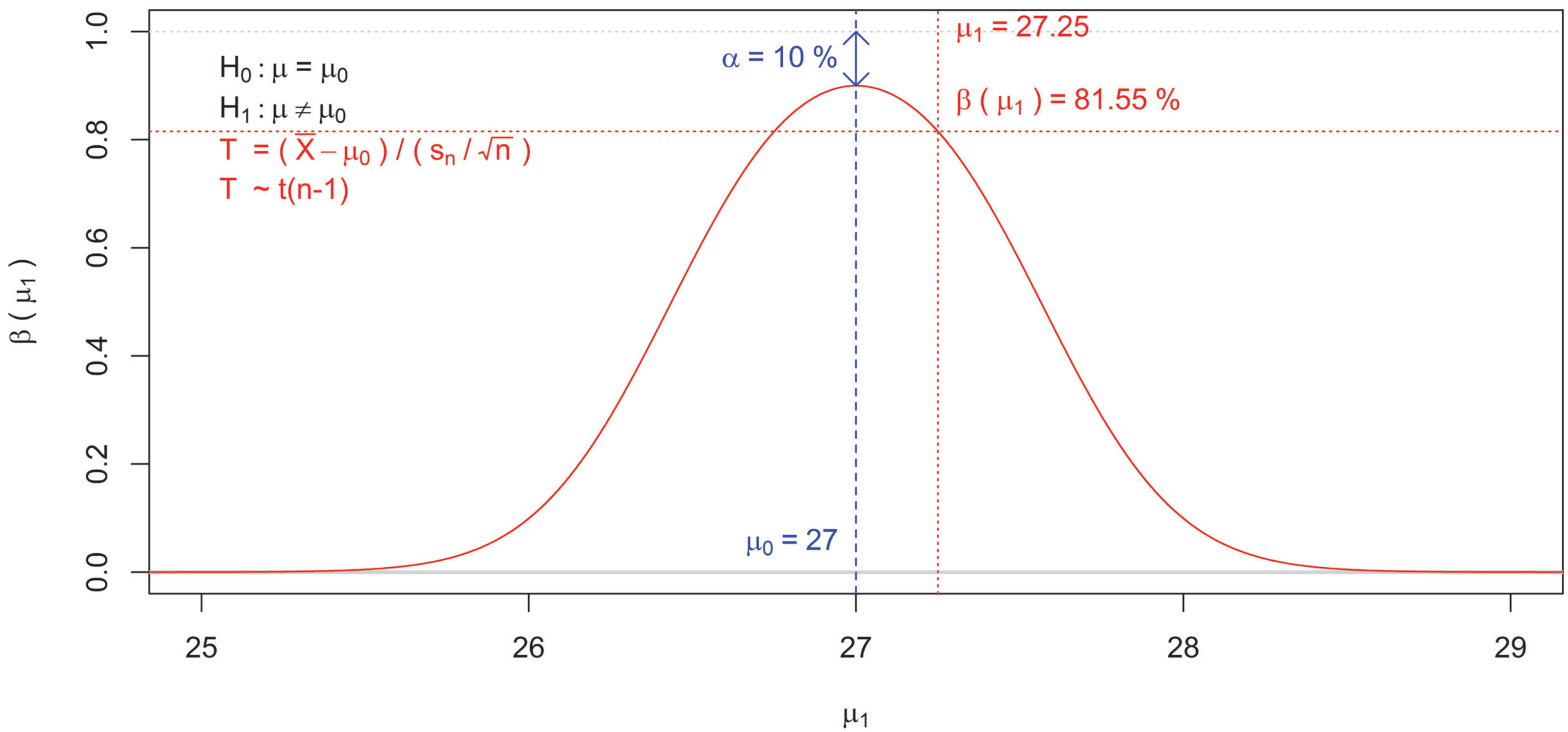
$$\widehat{\beta}(\mu_1) = Pr(t_{n-1,0.05} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} \leq T \leq t_{n-1,0.95} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}), \text{ where } T \sim T(19)$$



Analysis in file "Voltage_OC_Curve.R"

- Use now $s \approx \sqrt{2.14}$, $n = 10,\ \alpha = 0.10,$ and **quantiles** $t_{19,0.05} = -1.73, t_{19,0.95} = 1.73$ to estimate Type II Error $\widehat{\beta}\,(\mu_1)$:

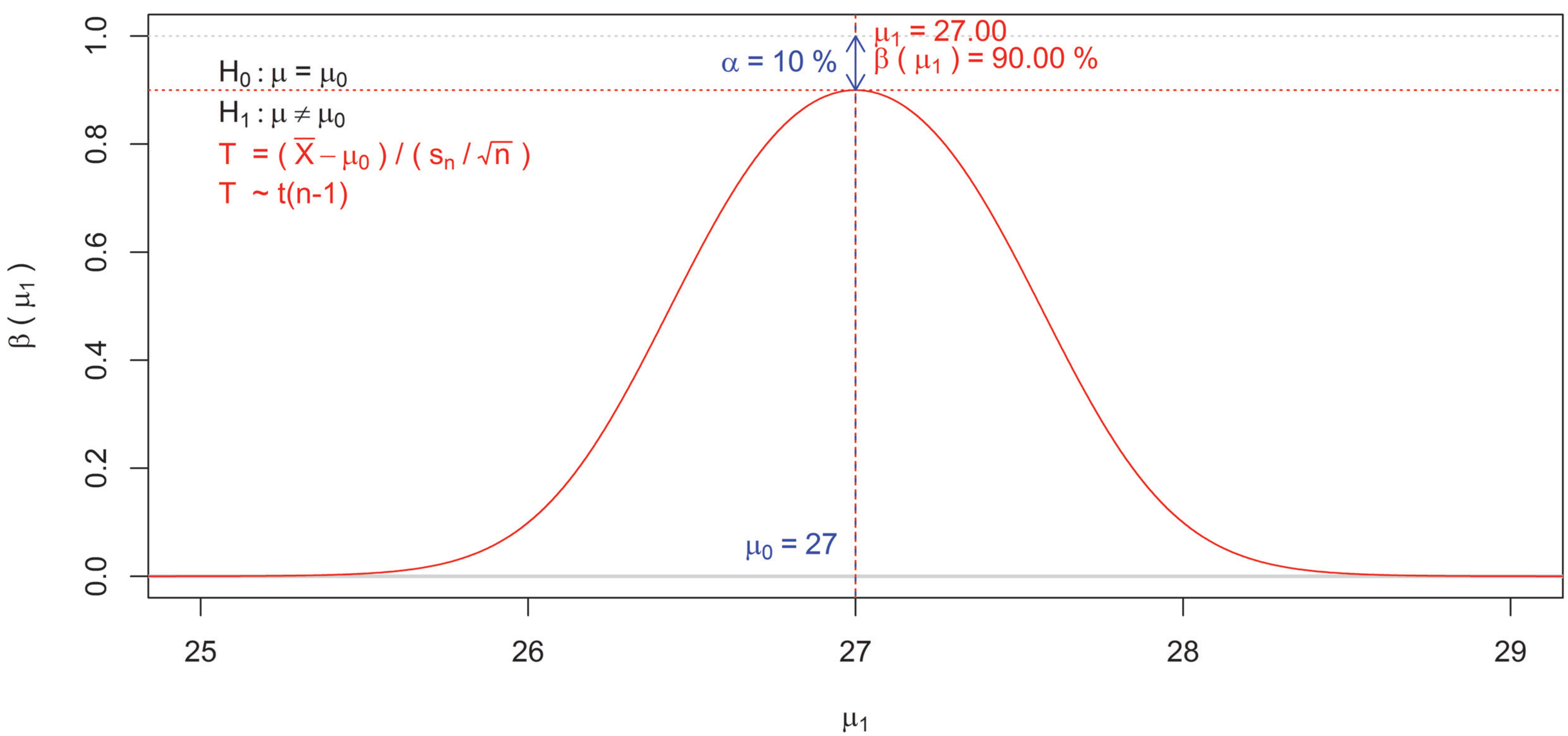
$$\widehat{\beta}\,(\mu_1) = Pr(t_{n-1,0.05} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} \leq T \leq t_{n-1,0.95} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}), \text{ where } T \sim T(19)$$



Analysis in file "Voltage_OC_Curve.R"

- Use now $s \approx \sqrt{2.14}, n = 10, \alpha = 0.10,$ and **quantiles** $t_{19,0.05} = -1.73, t_{19,0.95} = 1.73$ to estimate Type II Error $\widehat{\beta}(\mu_1)$:

$$\widehat{\beta}(\mu_1) = Pr(t_{n-1,0.05} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}} \leq T \leq t_{n-1,0.95} + \frac{\mu_0 - \mu_1}{s/\sqrt{n}}), \text{ where } T \sim T(19)$$



Analysis in file "Voltage_OC_Curve.R"

- **Estimator distributions** are **used for hypothesis testing.** Let $(x_1, \ldots, x_n)$ be a realization of an $i.i.d.$ random sample from a **normal distribution** with unknown mean μ and unknown variance σ^2. Consider **the one-sided hypothesis test.**

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0 \text{ (or } \boldsymbol{\mu_0 < \mu}\text{)}$$

Now, only high values of $\overline{x}$ are an indication of **support for the alternative hypothesis H_1**. **High values of $\overline{x}$ go together with high values of**

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

- **How high should $\overline{x}$ (or t_0) be before we reject the null hypothesis?** This is determined by the significance level α that you specify:

$$\text{Too high a value of } \overline{x} \Leftrightarrow t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \geq t_{n-1,0.95} \ \textbf{(Here } \boldsymbol{\alpha = 5\%}\textbf{!)}$$

- **Conclusion:**

$$\begin{cases} \text{we reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \geq t_{n-1,0.95} \\ \text{we fail to reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} < t_{n-1,0.95} \end{cases}$$

which is equivalent to

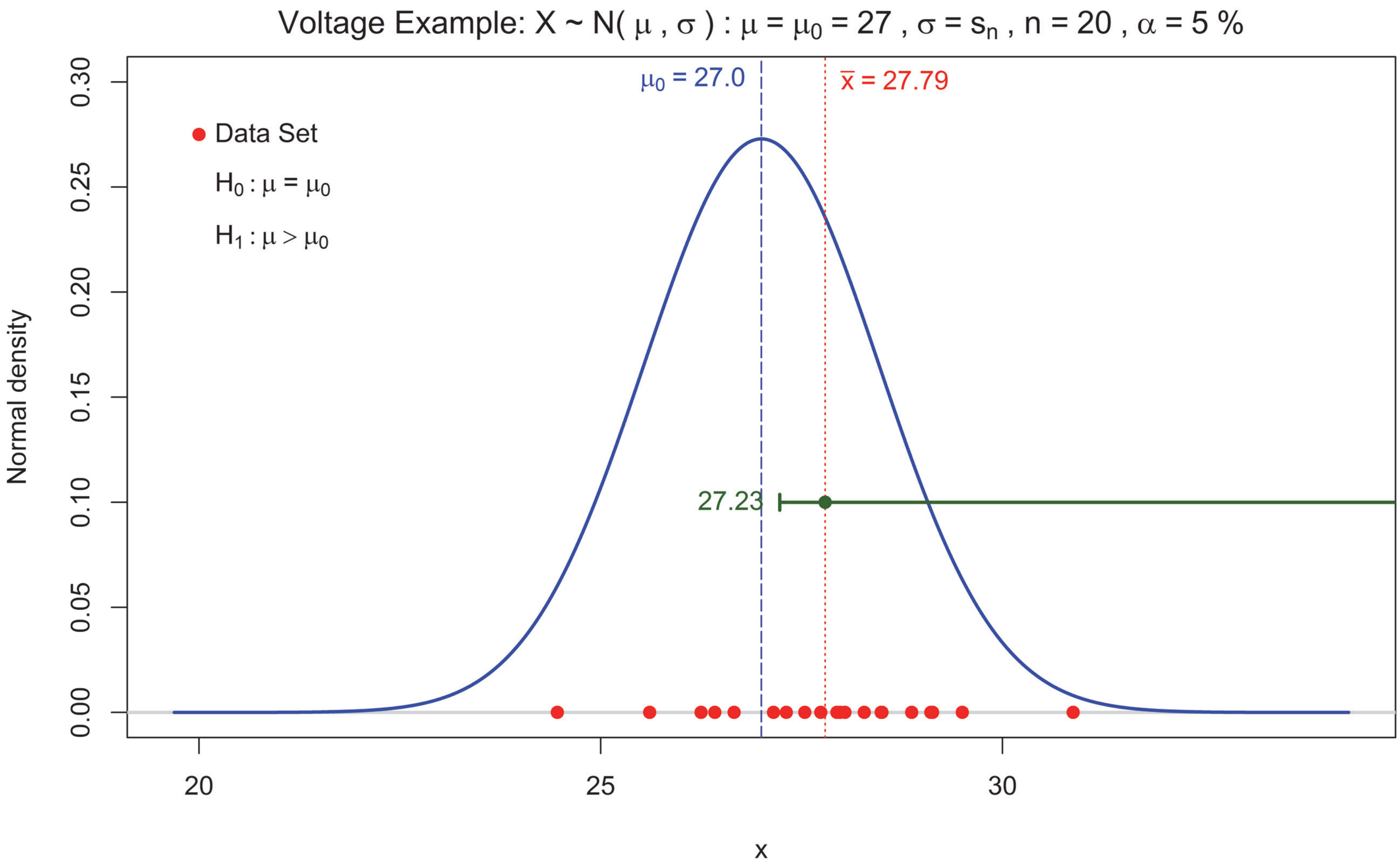
$$\begin{cases} \text{we reject } H_0 \text{ if } \mu_0 \notin (\overline{x} - \frac{s \times t_{n-1,0.95}}{\sqrt{n}}, \infty) \; (\Rightarrow \text{Conclusion: } \boldsymbol{\mu_0 < \mu}) \\ \text{we fail to reject } H_0 \text{ if } \mu_0 \in (\overline{x} - \frac{s \times t_{n-1,0.95}}{\sqrt{n}}, \infty) \end{cases}$$

- **Why?:**

$$\frac{\overline{x}-\mu_0}{s/\sqrt{n}} \geq t_{n-1,0.95} \Leftrightarrow \overline{x} - \mu_0 \geq t_{n-1,0.95} \times s/\sqrt{n} \Leftrightarrow$$

$$-\mu_0 \geq -\overline{x} + t_{n-1,0.95} \times s/\sqrt{n} \Leftrightarrow \mu_0 \leq \overline{x} - t_{n-1,0.95} \times s/\sqrt{n}$$

- $(\overline{x} - \frac{s \times t_{n-1,0.95}}{\sqrt{n}}, \infty)$ is **upper 95% confidence interval for true mean $\boldsymbol{\mu}$.**


Voltage Example: X ~ N(μ , σ) : μ = μ0 = 27 , σ = sn , n = 20 , α = 5 %
μ0 = 27.0
x̄ = 27.79
Data Set
H0 : μ = μ0
H1 : μ > μ0
27.23
Normal density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
20
25
30
x

- **Estimator distributions** are **used for hypothesis testing.** Let $(x_1, \ldots, x_n)$ be a realization of an $i.i.d.$ random sample from a **normal distribution** with unknown mean μ and unknown variance σ^2. Consider **the one-sided hypothesis test.**

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0 \text{ (or } \boldsymbol{\mu_0 > \mu})$$

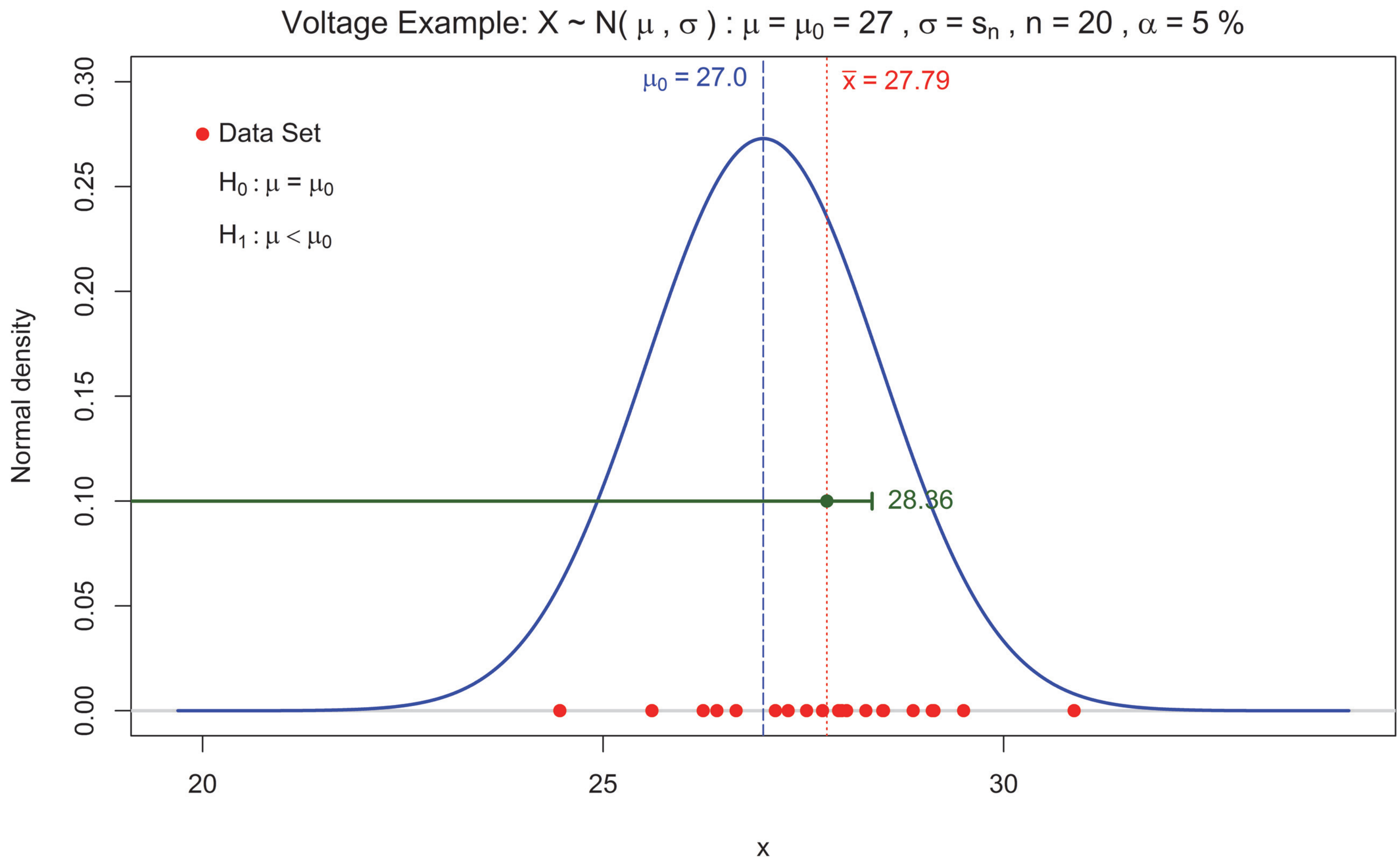
Then at a significance level of $\alpha = 5\%$,

$$\begin{cases} \text{we reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \leq t_{n-1,0.05} = -t_{n-1,0.95} \\ \text{we fail to reject } H_0 \text{ if } t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} > t_{n-1,0.05} = -t_{n-1,0.95} \end{cases}$$

which is equivalent to:

$$\begin{cases} \text{we reject } H_0 \text{ if } \mu_0 \notin (-\infty, \overline{x} + \frac{s \times t_{n-1,0.95}}{\sqrt{n}})(\Rightarrow \text{Conclusion: } \boldsymbol{\mu_0 > \mu}) \\ \text{we fail to reject } H_0 \text{ if } \mu_0 \in (-\infty, \overline{x} + \frac{s \times t_{n-1,0.95}}{\sqrt{n}}) \end{cases}$$

- $(-\infty, \overline{x} + \frac{s \times t_{n-1,0.95}}{\sqrt{n}})$ is **lower 95% confidence interval for true mean μ.**


Voltage Example: X ~ N(μ , σ) : μ = μ0 = 27 , σ = sn , n = 20 , α = 5 %
μ0 = 27.0
x̄ = 27.79
Data Set
H0 : μ = μ0
H1 : μ < μ0
28.36
Normal density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
20
25
30
x

Given a dataset $x_1, \ldots, x_n$ and n **numerical coefficients** $\mu_1, \ldots, \mu_n$, the value y

$$y = \mu_1 x_1 + \ldots + \mu_n x_n = \sum_{i=1}^{n} \mu_i x_i$$

is called **a linear combination** of the x_i datapoints. **The objective of linear regression analysis** deals with **estimating the coefficients $\mu_1, \ldots, \mu_n$** given datasets $\underline{y} = (y_1, \ldots, y_m)$ and $(x_1, \ldots, x_n)_j, j = 1, \ldots m$. The variables $x_1, \ldots, x_n$ are called the **explanatory variables** and the variable y is called the **dependent variable.** The dependent variable may typically be difficult to observe and the independent variables may not be. By observing a new datasets $(x_1, \ldots, x_n)_{new}$ and having established the relationship above, **we may infer/predict** the value of y_{new} associated with this dataset $(x_1, \ldots, x_n)_{new}$.

- Conducting hypothesis tests on the coefficients here are of the form:

$$H_0 : \mu_i = 0, H_1 : \mu_i \neq 0$$

and are standard in regression analysis. **Thus an estimator for each coefficient μ_i needs to formulated** and **their estimates follow from the datasets** $\underline{y} = (y_1, \ldots, y_m)$ and $(x_1, \ldots, x_n)_j, j = 1, \ldots m$.

Let $(X_1, \ldots, X_n)$ be a random $i.i.d.$ sample from a **normal distribution** with mean μ and variance σ^2, then:

$$Y = \sum_{i=1}^{n} \left[\frac{X_i - \overline{X}}{\sigma} \right]^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Chi-squared $(n-1)$ degrees of freedom

- We developed **the $100(1-\alpha)\%$ two-sided confidence interval** for σ^2

$$[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}]$$

Utilizing the above χ^2_{n-1} **Estimator Distribution** one may analogously (as we did for the mean μ) **develop hypothesis tests** of the form:

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_0 : \sigma^2 = \sigma_0^2 \qquad H_0 : \sigma^2 = \sigma_0^2$$
$$H_1 : \sigma^2 \neq \sigma_0^2 \qquad H_1 : \sigma^2 > \sigma_0^2 \qquad H_1 : \sigma^2 < \sigma_0^2$$

Example 15 (Continued): Dielectric breakdown voltage data

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

Hypothesis tests and confidence intervals **involving the F, χ^2 and t distributions** **all utilize an assumption of normality in the data**. Although **minor deviations from normality** are allowable, the procedures above **are not distribution-free**. Alternatives exist to the above tests that are **distribution-free** and **should be used in case of large departures from normality.**

- How can we test for normality of the data?
- How can we test in general whether data fits a particular theoretical distribution?
- To answer to these questions is to execute **a goodness-of-fit test.**