EMSE 4765: DATA ANALYSIS

For Engineers and Scientists

Session 5: Method-of-Moments, Maximum Likelihood, Goodness-of-Fit, Credibility Intervals

Version: 2/9/2021

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.

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 **involve the F, χ^2 and t distributions that 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.**

- **To execute a goodness-of-fit test** we first need to **fit a theoretical distribution to the data.** How does one accomplish that?

- **Let the dataset $(x_1, \ldots, x_n)$ be a realization of an $i.i.d.$ random sample** from a theoretical distribution with density function $f(x|\Theta)$, where $\Theta = (\theta_1, \theta_2)$. We can next evaluate **the expressions for population mean $E[X|\theta]$ and variance $V[X|\Theta]$** and solve for $\Theta = (\theta_1, \theta_2)$, by setting:

$$\begin{cases} E[X|\theta] = \overline{x}, \text{ i.e. the sample mean} \\ V[X|\theta] = s^2, \text{ i.e. the sample variance} \end{cases}$$

Example 15 (Continued): Dielectric breakdown voltage data

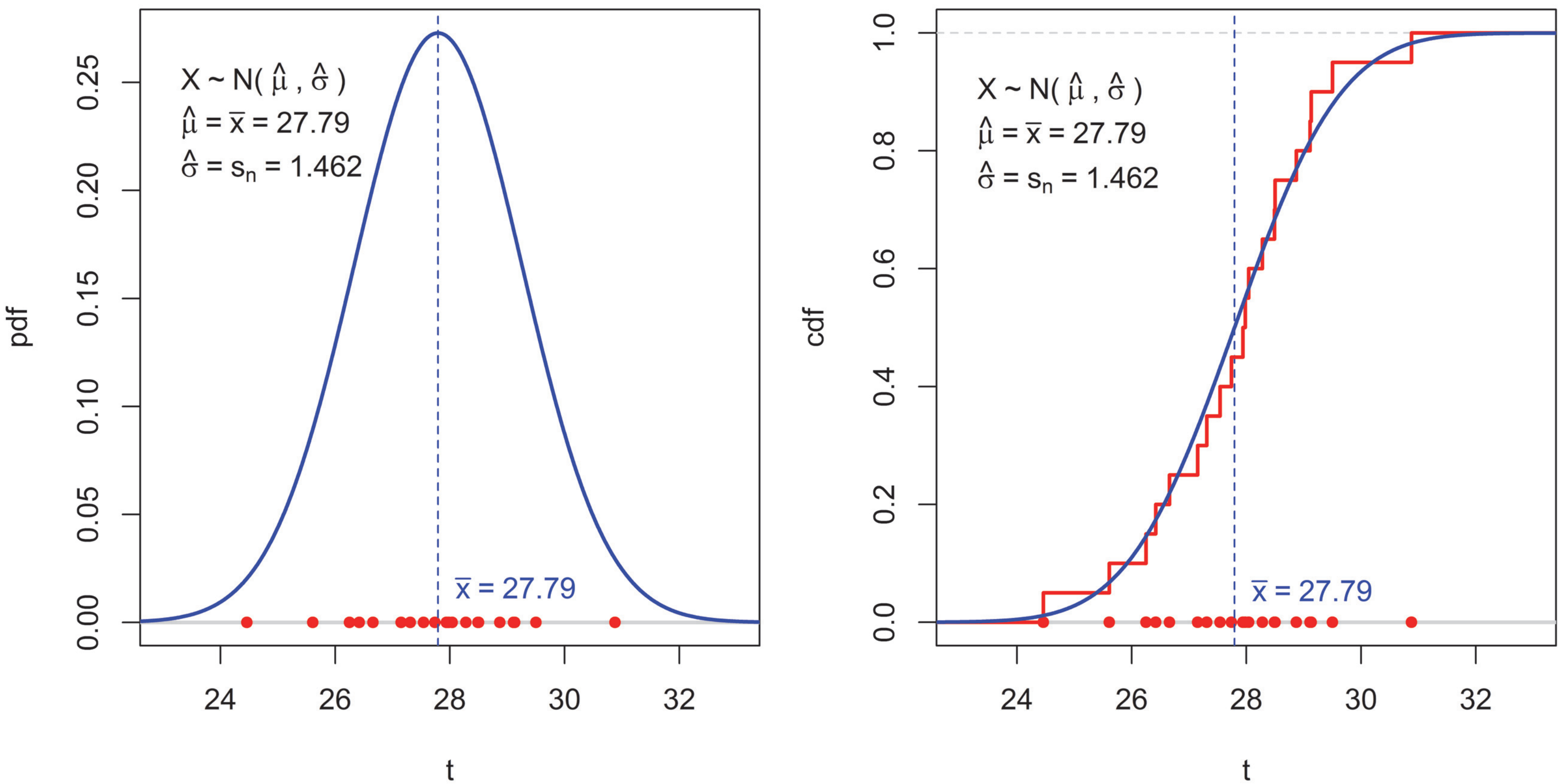
24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

$$\overline{x} \approx 27.793, s^2 \approx 2.137 \text{ or } s \approx 1.462$$

- Assuming a normal density $f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ as **the model distribution** with $E[X \mid \mu, \sigma] = \mu$ and $V[X \mid \mu, \sigma] = \sigma^2$, we have

$$\begin{cases} \mu = \overline{x} = 27.793 \\ \sigma^2 = s^2 = 2.137 \end{cases} \Leftrightarrow \begin{cases} \mu = \overline{x} = 27.793 \\ \sigma = \sqrt{s^2} = s = \sqrt{2.137} = 1.462 \end{cases}$$

Voltage Normal MOM Fit: n = 20 , $\overline{x}$ = 27.79 , s_n = 1.462



Analysis in "Voltage_Normal_MOM_Fit.R"

- Assuming a gamma density $f(x|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ as **the model distribution** with $E[X\,|\,\alpha,\beta]=\alpha/\beta$ and $V[X\,|\,\alpha,\beta]=\alpha/\beta^2$, we have

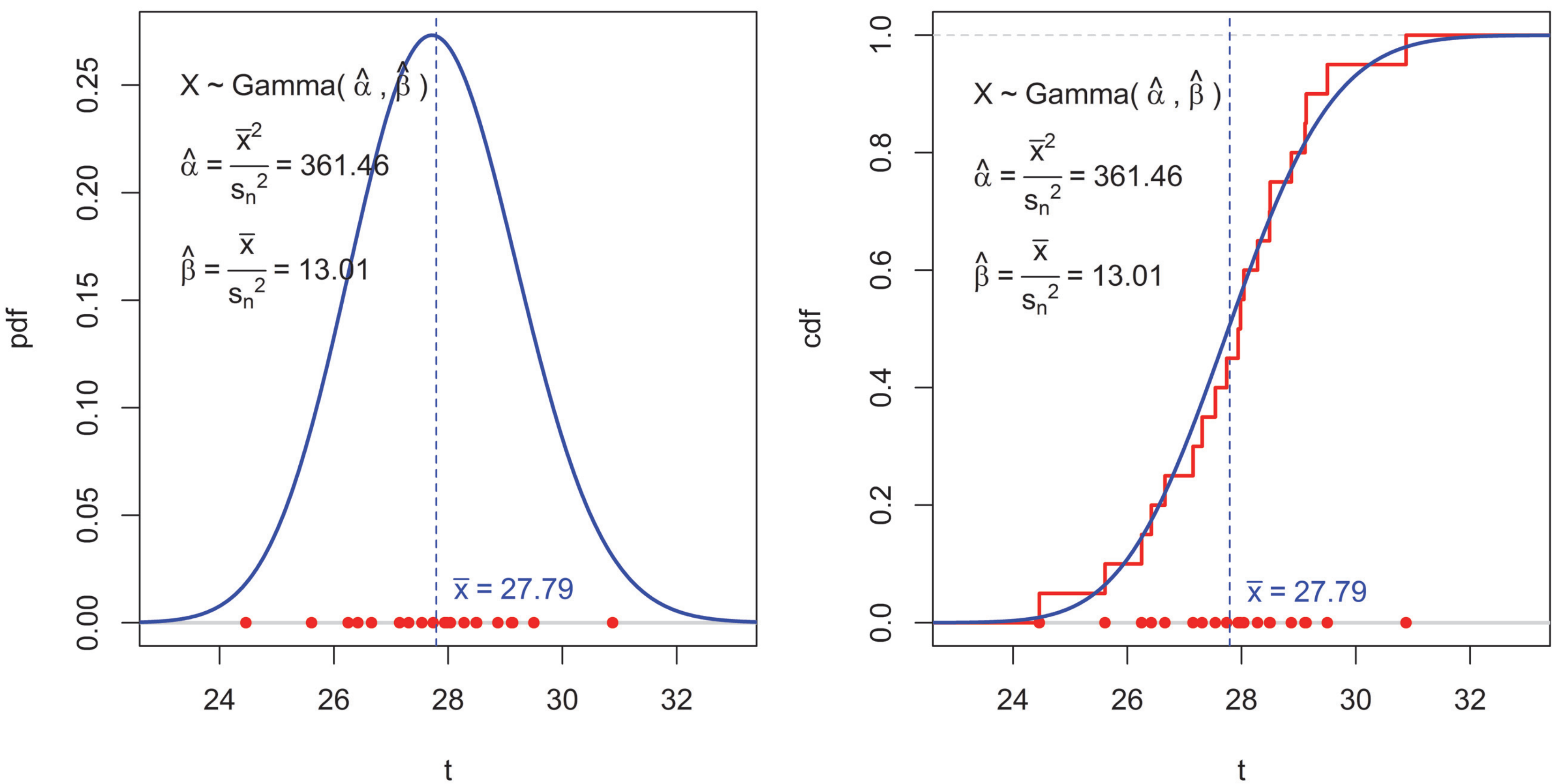
$$\begin{cases}\frac{\alpha}{\beta}=\overline{x}\\ \frac{\alpha}{\beta^2}=\frac{\alpha}{\beta}\times\frac{1}{\beta}=s^2\end{cases}\Leftrightarrow\begin{cases}\frac{\alpha}{\beta}=\overline{x}\\ \overline{x}\times\frac{1}{\beta}=s^2\end{cases}\Leftrightarrow$$

$$\begin{cases}\alpha=\overline{x}\times\beta\\ \beta=\frac{\overline{x}}{s^2}\end{cases}\Leftrightarrow\begin{cases}\alpha=\frac{\overline{x}^2}{s^2}=\frac{(27.793)^2}{2.137}\approx 361.46\\ \beta=\frac{\overline{x}}{s^2}=\frac{27.793}{2.137}\approx 13.01\end{cases}$$

$$\Leftrightarrow\begin{cases}\alpha=\frac{\overline{x}^2}{s^2}=\frac{(27.793)^2}{2.137}\approx 361.46\text{ - Shape Parameter}\\ \frac{1}{\beta}\approx 0.0769\text{ - Scale Parameter}\end{cases}$$

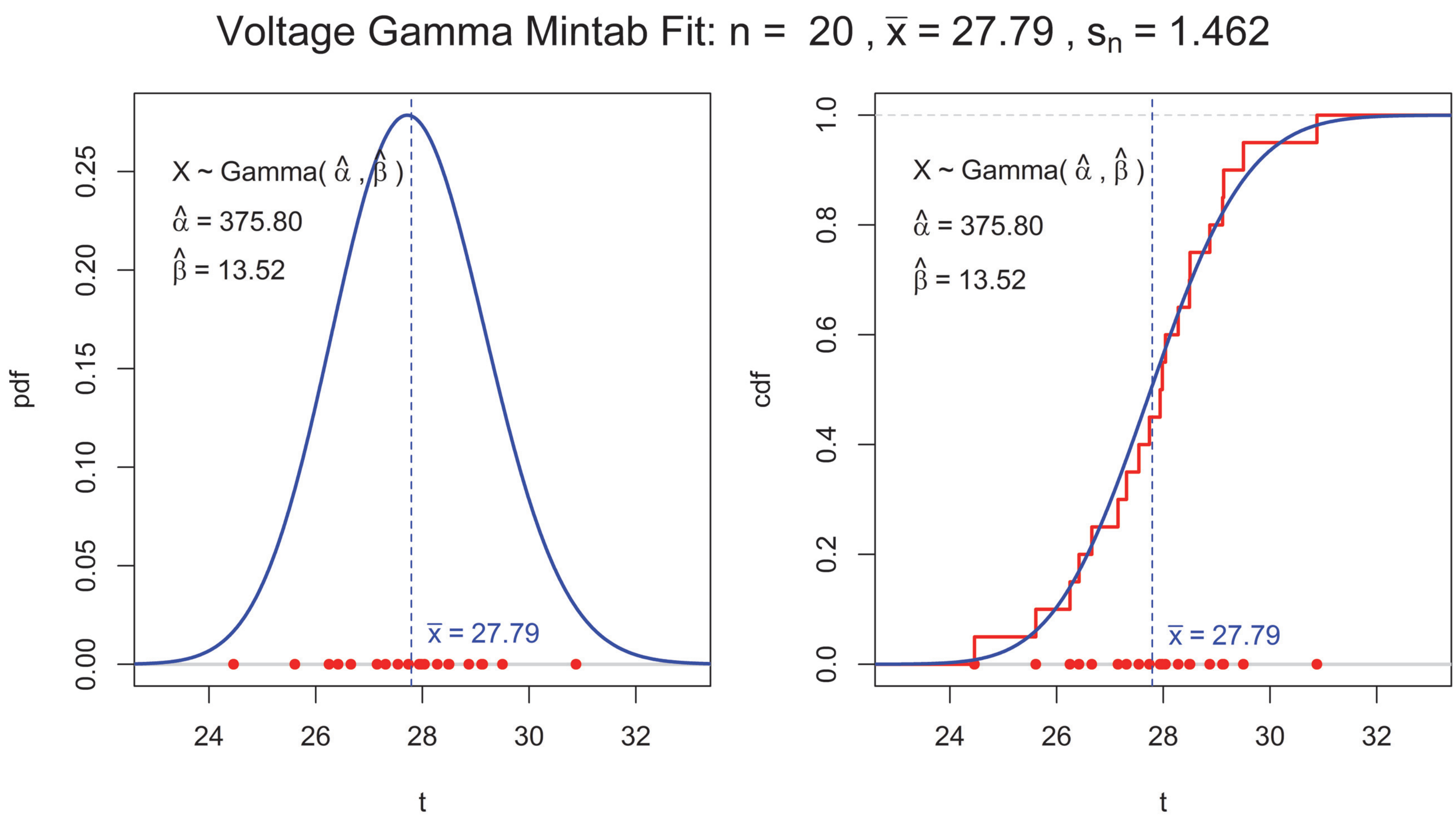
$$\begin{cases} \alpha = \overline{x} \times \beta \\ \beta = \frac{\overline{x}}{s^2} \end{cases} \Leftrightarrow \begin{cases} \alpha = \frac{\overline{x}^2}{s^2} = \frac{(27.793)^2}{2.137} \approx 361.46 \\ \beta = \frac{\overline{x}}{s^2} = \frac{27.793}{2.137} \approx 13.01 \end{cases}$$

Voltage Gamma MOM Fit: $n = 20$, $\overline{x} = 27.79$, $s_n = 1.462$



Analysis in "Voltage_Gamma_MOM_Fit.R"

$$\begin{cases} \alpha \approx 375.8 \\ \beta \approx 1/0.07395 = 13.52 \end{cases} \Rightarrow \text{Minitab does not uses the method-of-moments}$$



Analysis in "Voltage_Gamma_Minitab_Fit.R"

- **Let the dataset $(x_1, \ldots, x_n)$ be a realization of an $i.i.d.$ random sample** from a theoretical distribution with density function $f(x|\Theta)$, where $\Theta = (\theta_1, \theta_2)$. Next, we can formulate **the likelihood function**

$$L(\Theta|(x_1, \ldots, x_n)) = \prod_{i=1}^{n} f(x_i|\Theta)$$

- Next, we select the **Maximum Likelihood Estimates (MLE's)** such that $\widehat{\theta}_1, \widehat{\theta}_2$ **maximize the likelihood function $L(\Theta|(x_1, \ldots, x_n))$** as a function of (θ_1, θ_2) given **the dataset $(x_1, \ldots, x_n)$** .

- For most distributions the $L(\Theta|(x_1, \ldots, x_n))$ is differentiable and in that case **a necessary condition for optimality** is:

$$\frac{\partial}{\partial \theta_i} L(\Theta|(x_1, \ldots, x_n)) = 0, \; i = 1, 2$$

- For many distributions these conditions are also sufficient (**BE CAREFUL**).

- For many distributions maximizing $L(\Theta|(x_1,\ldots,x_n))$ is **equivalent to maximizing** $\boldsymbol{Log\{L(\Theta|(x_1,\ldots,x_n))\}}$. (**BE CAREFUL**).

Example 17: Assuming a normal density $f(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ as **the model distribution**, we have

$$L(\mu,\sigma|(x_1,\ldots,x_n))=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$=\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n\prod_{i=1}^{n}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

- Maximizing the likelihood is here **equivalent to** maximizing the log-likelihood:

$$Ln\{L(\mu,\sigma|(x_1,\ldots,x_n))\} = -\frac{n}{2}Ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}$$

- **Taking partial derivatives with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ yields**:

$$\frac{\partial}{\partial\mu}Ln\{L(\mu,\sigma|(x_1,\ldots,x_n))\} = \frac{1}{2\sigma^2} \times \left[\sum_{i=1}^{n}2(x_i-\mu)\right]$$

$$\frac{\partial}{\partial\sigma}Ln\{L(\mu,\sigma|(x_1,\ldots,x_n))\} = -\frac{n}{2}\frac{2\pi\cdot 2\sigma}{2\pi\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\left[-\frac{2\sigma}{\sigma^4}\right]$$

$$= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = -\frac{n\sigma^2 - \sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3}$$

- **Setting both partial derivatives equal to zero** and **solving for μ and σ yields:**

$$\sum_{i=1}^{n} 2(x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0 \Leftrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$-\frac{n\sigma^2 - \sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = 0 \Leftrightarrow n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

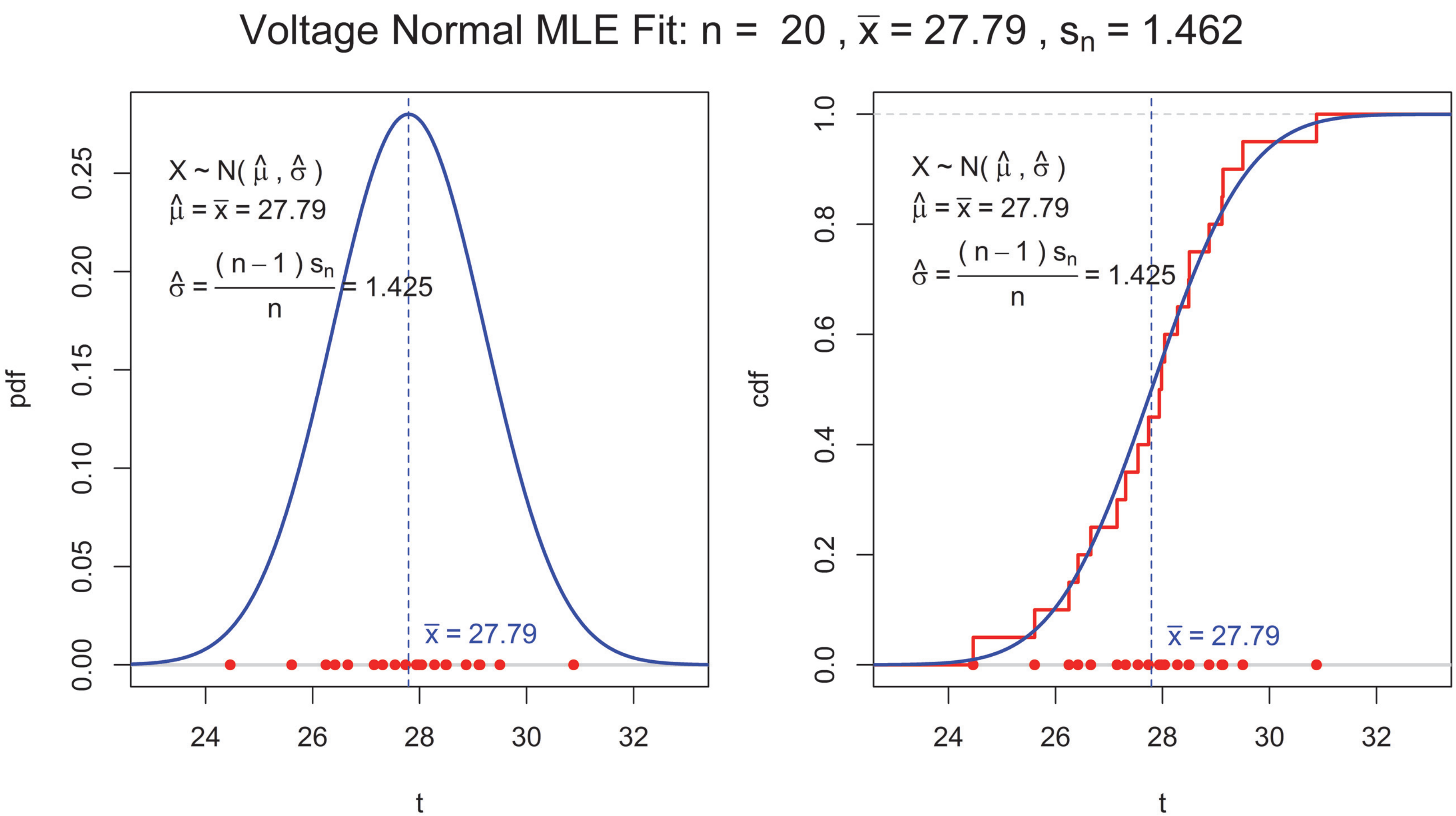
- **Note that the MLE for σ^2** is **not the unbiased estimator S^2**, so these two different principles of estimation may yield different estimates:

Example 15 (Continued): Dielectric breakdown voltage data

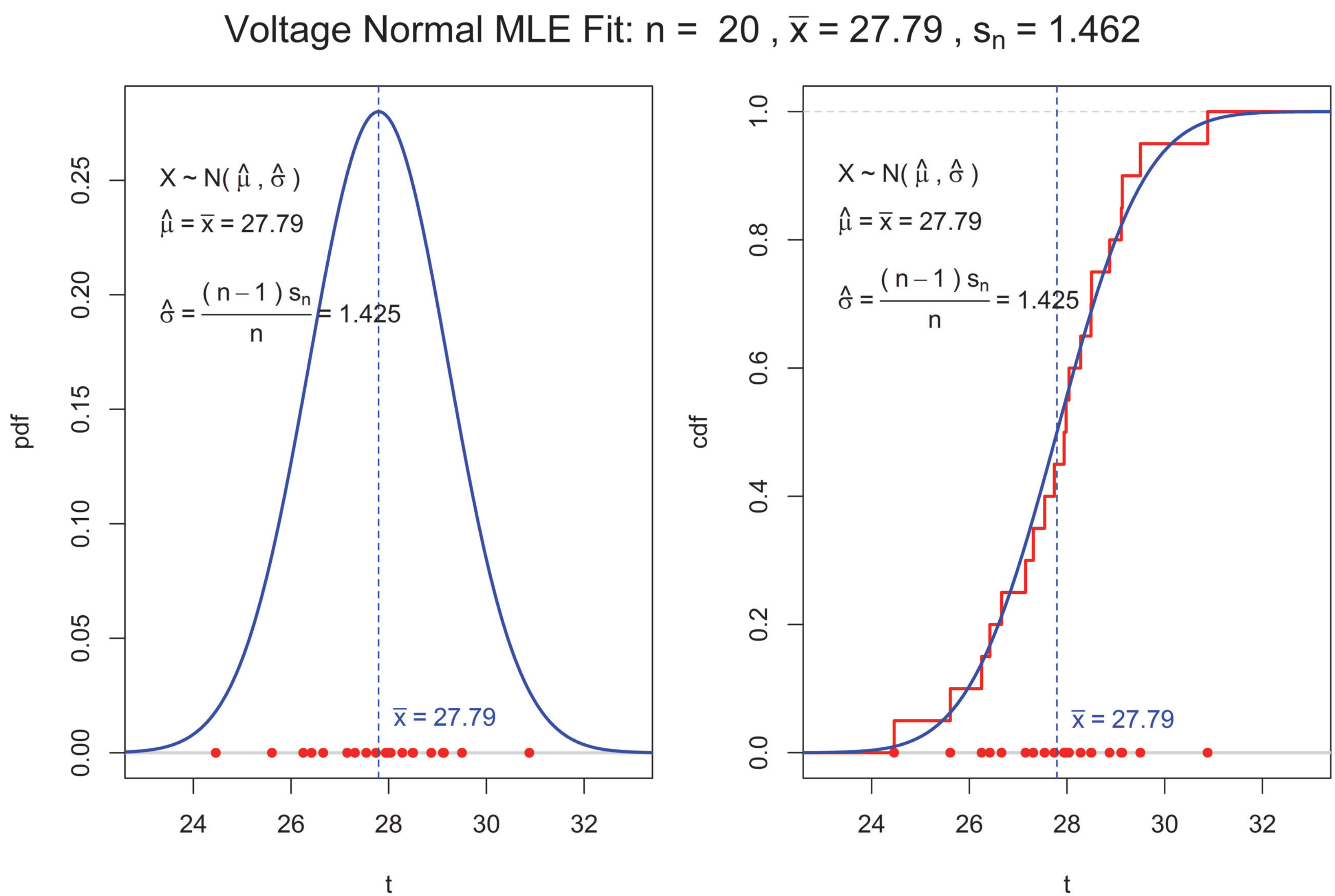
24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

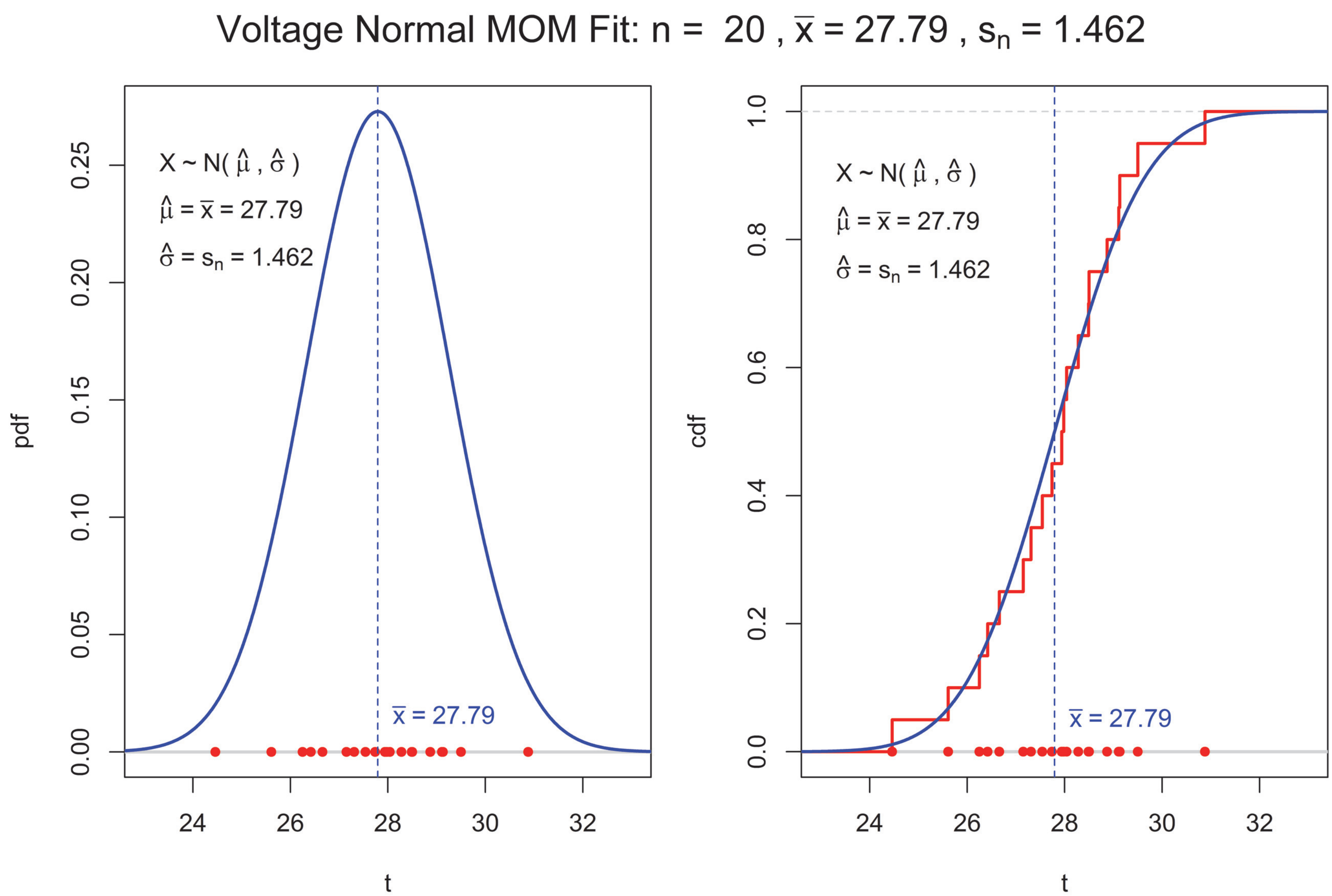
$$\overline{x} \approx 27.793,\ (n-1)s^2/n \approx 2.030 \text{ or } \sqrt{(n-1)s^2/n} \approx 1.425$$

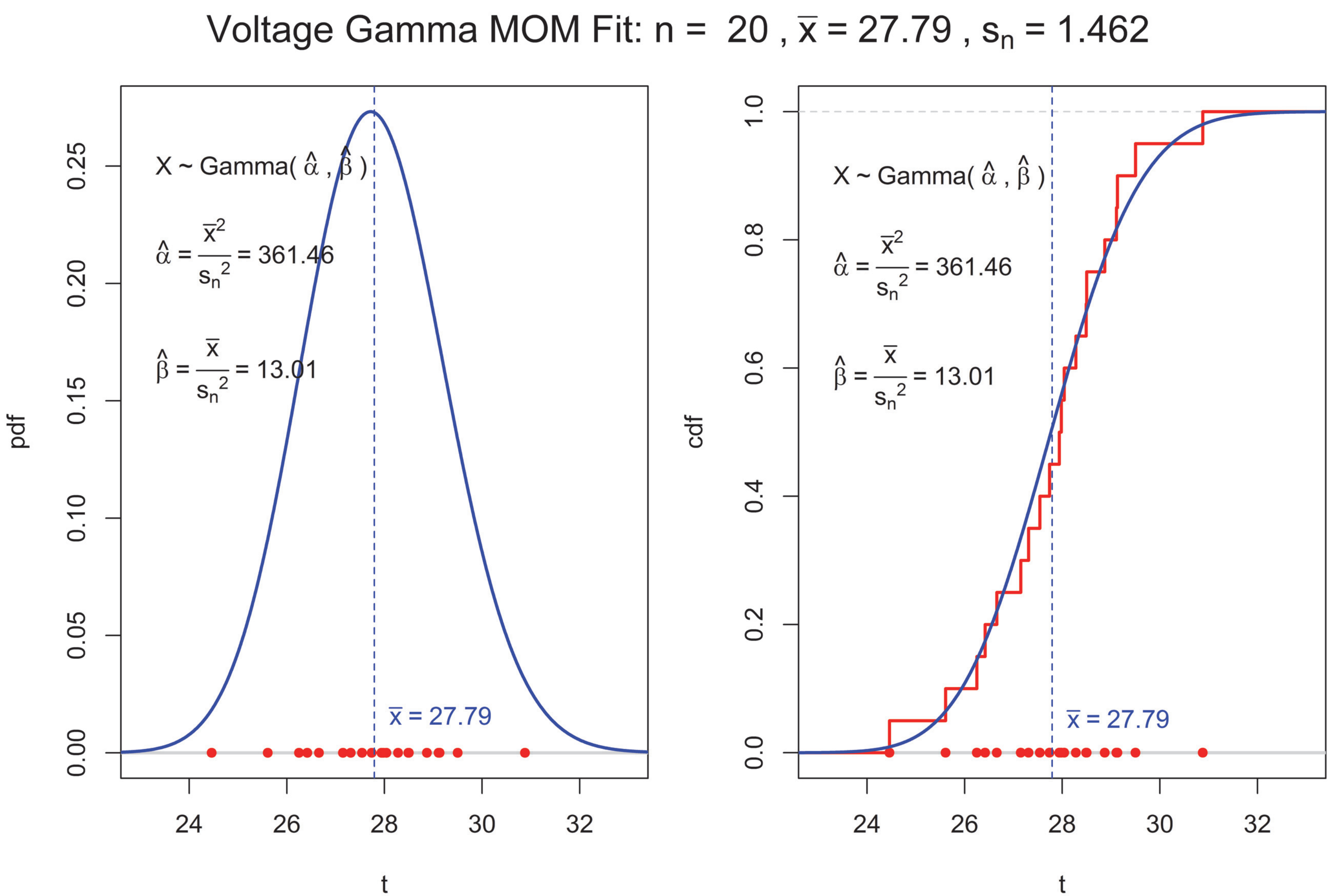
$$\overline{x} \approx 27.793, (n-1)s^2/n \approx 2.030 \text{ or } \sqrt{(n-1)s^2/n} \approx 1.425$$



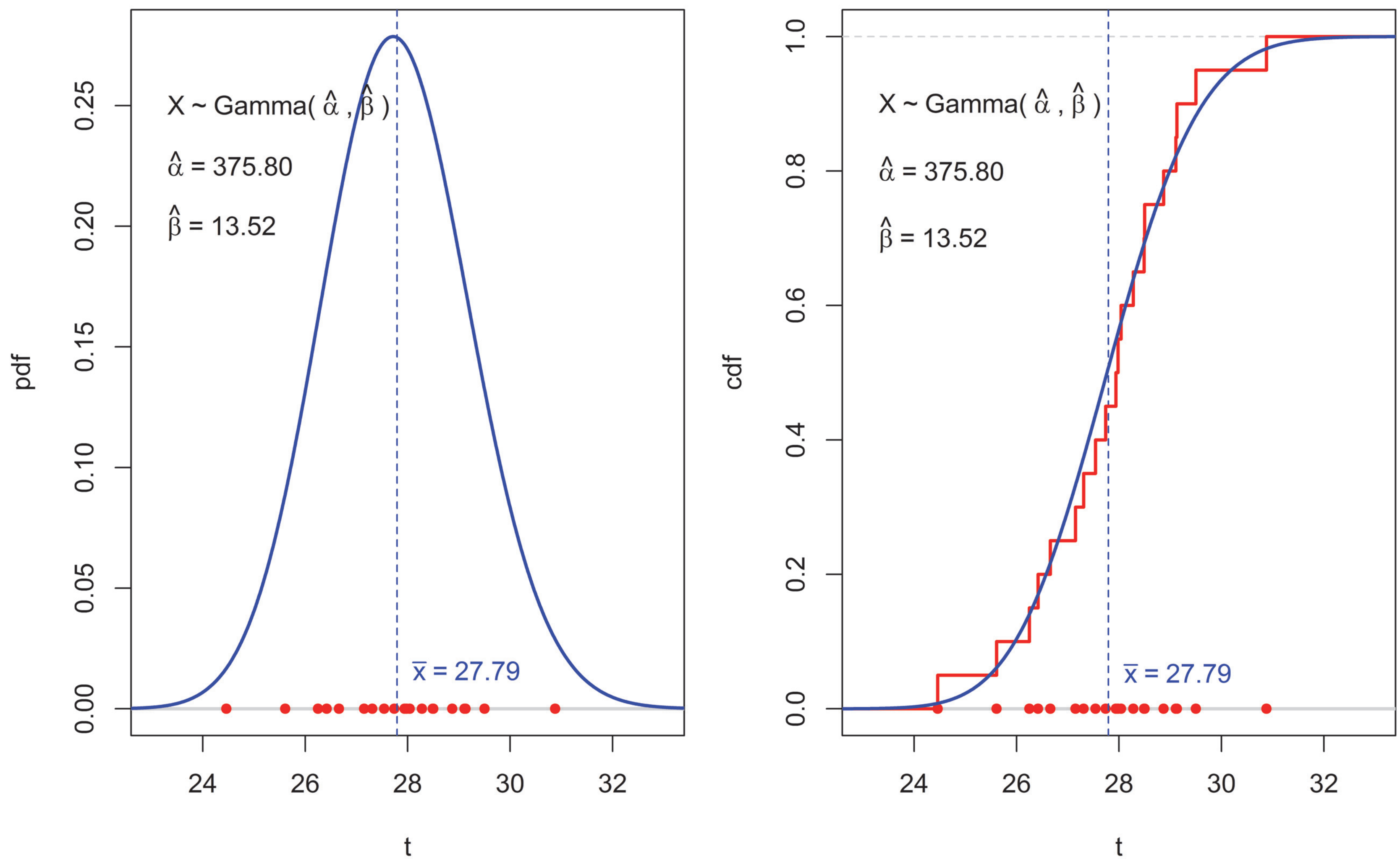
Analysis in "Voltage_Gamma_Minitab_Fit.R"


Voltage Normal MLE Fit: n = 20 , x̄ = 27.79 , s_n = 1.462
X ~ N(μ̂ , σ̂)
μ̂ = x̄ = 27.79
σ̂ = (n − 1) s_n / n = 1.425
x̄ = 27.79
pdf
cdf
t
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
24
26
28
30
32


Voltage Normal MOM Fit: n = 20 , $\bar{x}$ = 27.79 , s_n = 1.462
$X \sim N(\hat{\mu}, \hat{\sigma})$
$\hat{\mu} = \bar{x} = 27.79$
$\hat{\sigma} = s_n = 1.462$
$\bar{x}$ = 27.79
pdf
0.00
0.05
0.10
0.15
0.20
0.25
24
26
28
30
32
t
$X \sim N(\hat{\mu}, \hat{\sigma})$
$\hat{\mu} = \bar{x} = 27.79$
$\hat{\sigma} = s_n = 1.462$
$\bar{x}$ = 27.79
cdf
0.0
0.2
0.4
0.6
0.8
1.0
24
26
28
30
32
t


Voltage Gamma MOM Fit: n = 20 , x̄ = 27.79 , s_n = 1.462
X ~ Gamma(α̂ , β̂)
α̂ = x̄²/s_n² = 361.46
β̂ = x̄/s_n² = 13.01
x̄ = 27.79
pdf
cdf
t
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
24
26
28
30
32

Voltage Gamma Mintab Fit: n = 20 , $\bar{x}$ = 27.79 , s_n = 1.462



- The Chi-Square test compares **the empirical histogram density**, constructed from sample data, to **a candidate theoretical density**.

- Assume that the **dataset $(x_1, \ldots, x_n)$** is **a realization of $i.i.d.$ random sample** from an **underlying random variable $X \sim F(\cdot \,|\underline{\theta})$**.

- This sample is then used to construct an empirical histogram with **m bins** where **Bin j corresponds to the interval $[LB_j, UB_j]$**. The chi-square test allows some flexibility in the choice on bin-boundaries.

- The estimator of the probability $p_j = Pr\{X \in [LB_j, UB_j]\}$ of cell j is:

$$\widehat{p}_j = \frac{O_j}{N}, \; j = 1, \ldots, m,$$

where **O_j is the number of observations in Bin j**. (These can be determined using the FREQUENCY array function in Micro Soft Excel).

- Let $F_X(x|\underline{\theta})$ be some **theoretical candidate model distribution** with **model parameter vector $\underline{\theta}$** of the random variable X whose goodness-of-fit is to be assessed. Then we can **mathematically evaluate after estimating $\widehat{\underline{\theta}}$ from the dataset $(x_1, \ldots, x_n)$:**

$$\begin{aligned} p_j &= Pr\{X \in [LB_j, UB_j]\} \\ &= F_X(UB_j|\widehat{\underline{\theta}}) - F_X(LB_j|\widehat{\underline{\theta}}),\ j = 1, \ldots, m. \end{aligned}$$

- Define next:

$$\begin{aligned} O_j &= \text{Number of Observations in Bin } j = \ \widehat{p}_j \times N \\ E_j &= \text{Expected Number of Observations in Bin } j = \ p_j \times N. \end{aligned}$$

and the **"distance measure" between the O_j's and E_j's (both estimated using the same dataset $(x_1, \ldots, x_n)$):**

$$S^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} > 0.$$

- **Intuition:** If $F_X(\cdot\,|\widehat{\underline{\theta}})$ is a good fit then "the theoretical value" of p_j should be close to the empirical value $\widehat{p}_j$ (and **thus O_j should be close to E_j, $j = 1, \ldots, m$**). Hence a good fit would have a **small "distance" S^2-value.**

- It can be shown that **S^2 is a realization of χ^2_k-random variable (asymptotically)**: i.e. a realization of **a chi-squared random variable** with **k degrees of freedom**, where

$$k = m - |\underline{\theta}| - 1.$$

Here is **$|\underline{\theta}|$ equal to dimension or the number of parameters in the vector $\underline{\theta}$.** Note that, χ^2_k is a random variable with support $[0, \infty)$ (i.e. it only takes on non-negative values).

- Using the CHI.DIST function in Microsoft Excel we can calculate **the probability that χ^2_k is greater than the observed value S^2**. **If this probability is small** (**large**), than clearly **the observed value S^2 may be considered "big"** (**"small"**).

- The ***p-value*** of the Chi-Squared goodness-of-fit test is defined as:

$$p\text{-}value \equiv Pr(\chi_k^2 > S^2)$$

- It is common to reject the candidate theoretical distribution when the ***p-value*** **is smaller than 0.01, 0.05 or even 0.10.**

- **Rule of thumb for the number of Bins:**

$Sample\ Size\ N$	$Number\ of\ Bins$
< 20	$Do\ not\ use\ \chi^2 - Test$
50	$5\ to\ 10$
100	$10\ to\ 20$
> 100	$\sqrt{N}\ to\ \frac{N}{5}$

- **Rule of thumb for the size of E_j's**: (which allows for the Chi-Squared distribution assumption): It has been suggested that $E_j > 3,\ 4$ or 5, there is no real agreement on this issue.

- The Chi-Squared Test allows for **flexibility in the choice of bin boundaries**. **Nowadays it is preferred that boundaries are selected such that the expected number of observations is the same in each bin**. This **weighs each part of the theoretical fit $F_X(\cdot\,|\widehat{\underline{\theta}})$** equally in the chi-squared analysis.

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

Equal Bin Width Method: MOM fit

Bin	LB	UB	O_i	p_i	E_i	$(O_i-E_i)^2/E_i$
1	< 24.46	26.07	2	11.86%	2.37	0.06
2	26.07	27.67	6	34.79%	6.96	0.13
3	27.67	29.28	10	37.82%	7.56	0.78
4	29.28	> 30.88	2	15.53%	3.11	0.39
			20	100.00%		1.37

# Bins	4
# Parameters	2
# Degrees of Freedom	1
P-Value	24.19%

Equal Bin Width Method: MLE fit

Bin	LB	UB	O_i	p_i	E_i	$(O_i-E_i)^2/E_i$
1	< 24.46	26.07	2	11.26%	2.25	0.03
2	26.07	27.67	6	35.30%	7.06	0.16
3	27.67	29.28	10	38.53%	7.71	0.68
4	29.28	> 30.88	2	14.91%	2.98	0.32
			20	100.00%		1.19

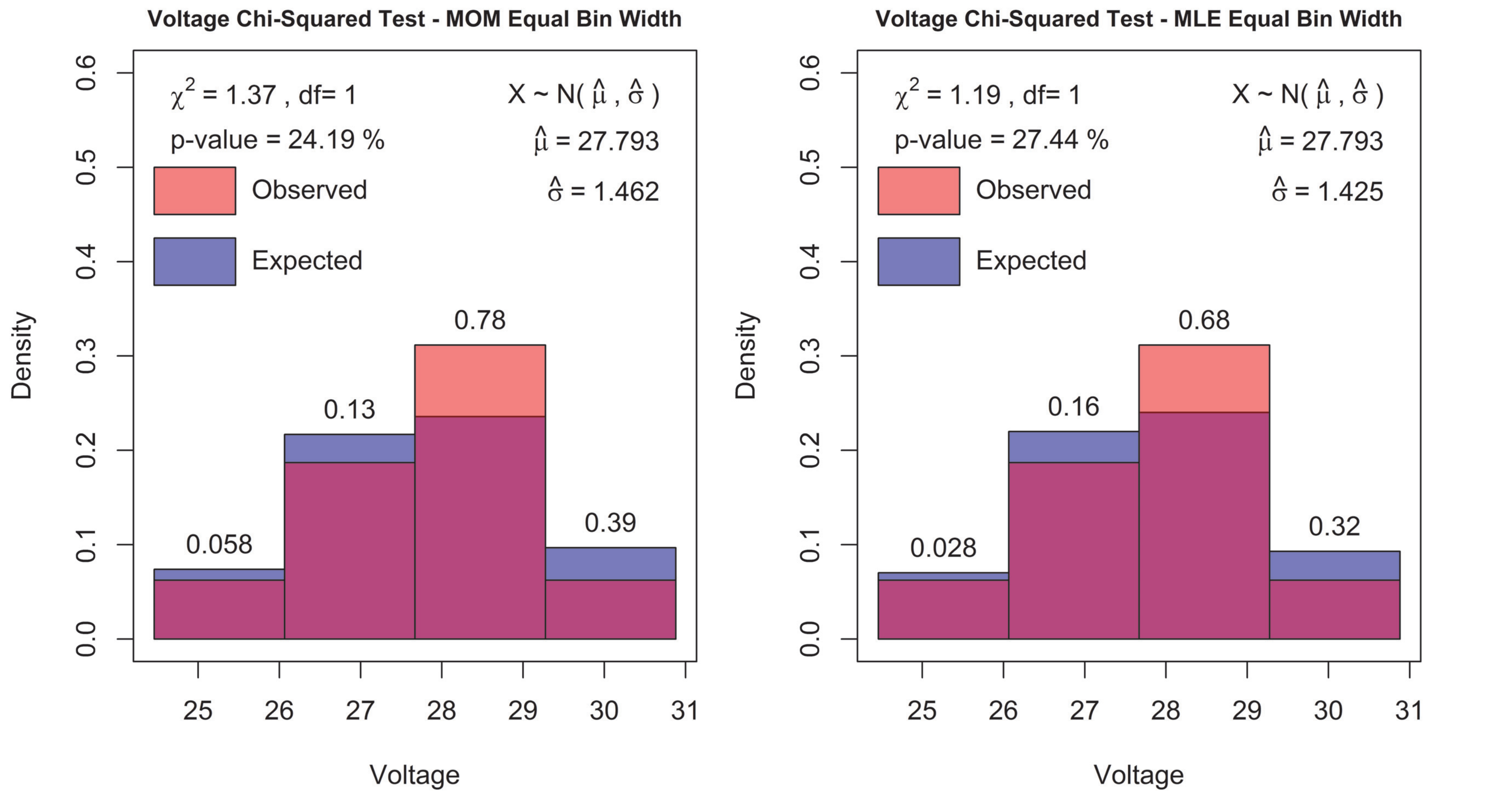
# Bins	4
# Parameters	2
# Degrees of Freedom	1
P-Value	27.44%

Equal Average Observation in Bin Method: MLE and MOM fit

Bin	LB	UB	O_i	$\Sigma\, p_i$	p_i	E_i	$(O_i-E_i)^2/E_i$
1	< 24.46	26.81	5	25.00%	25.00%	5.00	0.00
2	26.81	27.79	4	50.00%	25.00%	5.00	0.20
3	27.79	28.78	6	75.00%	25.00%	5.00	0.20
4	28.78	> 28.78	5	100.00%	25.00%	5.00	0.00
			20				0.40

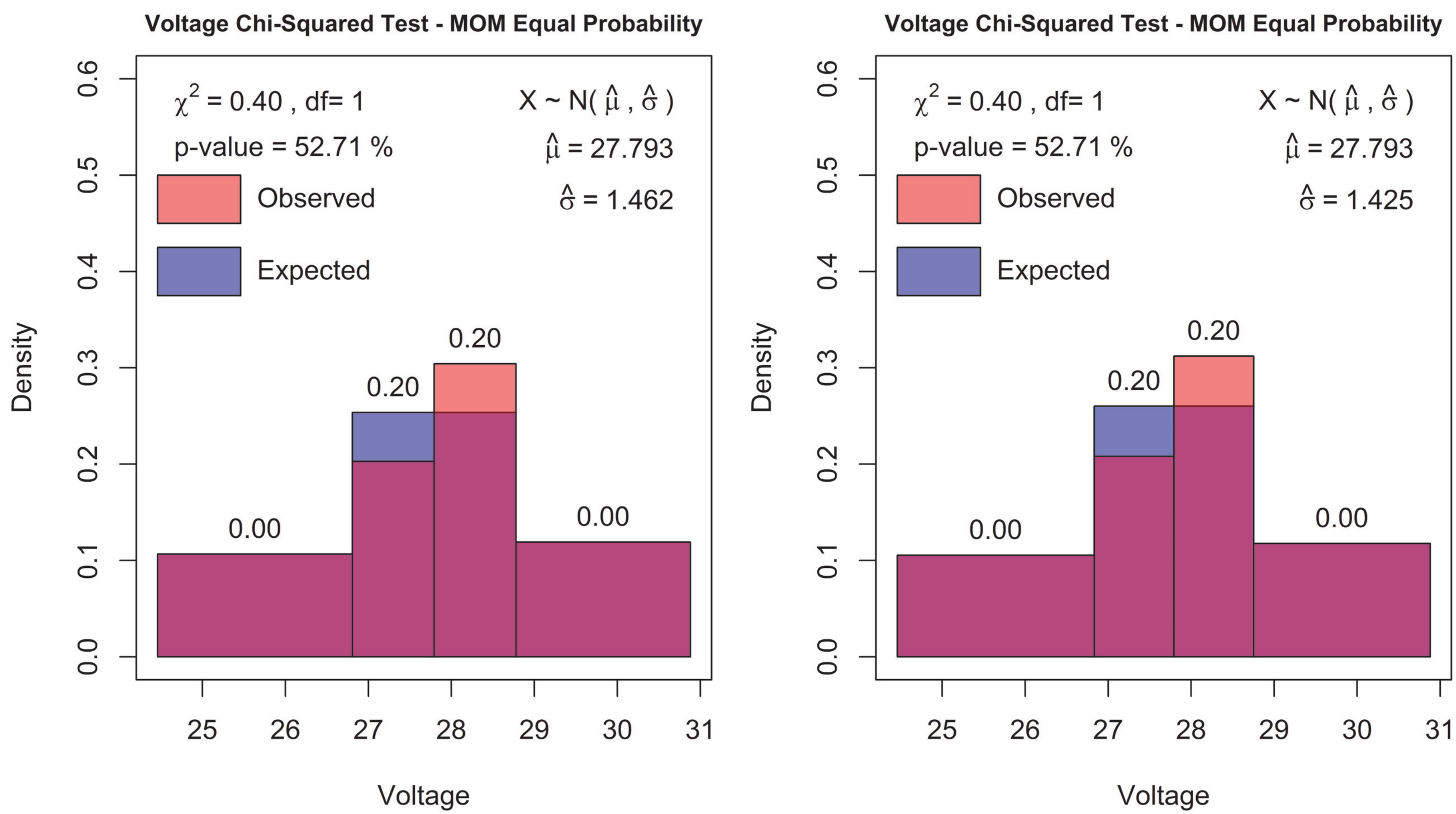
# Bins	4
# Parameters	2
# Degrees of Freedom	1
P-Value	52.71%

Graphical Depiction of Equal-Bin Width χ^2- goodness-of-fit test

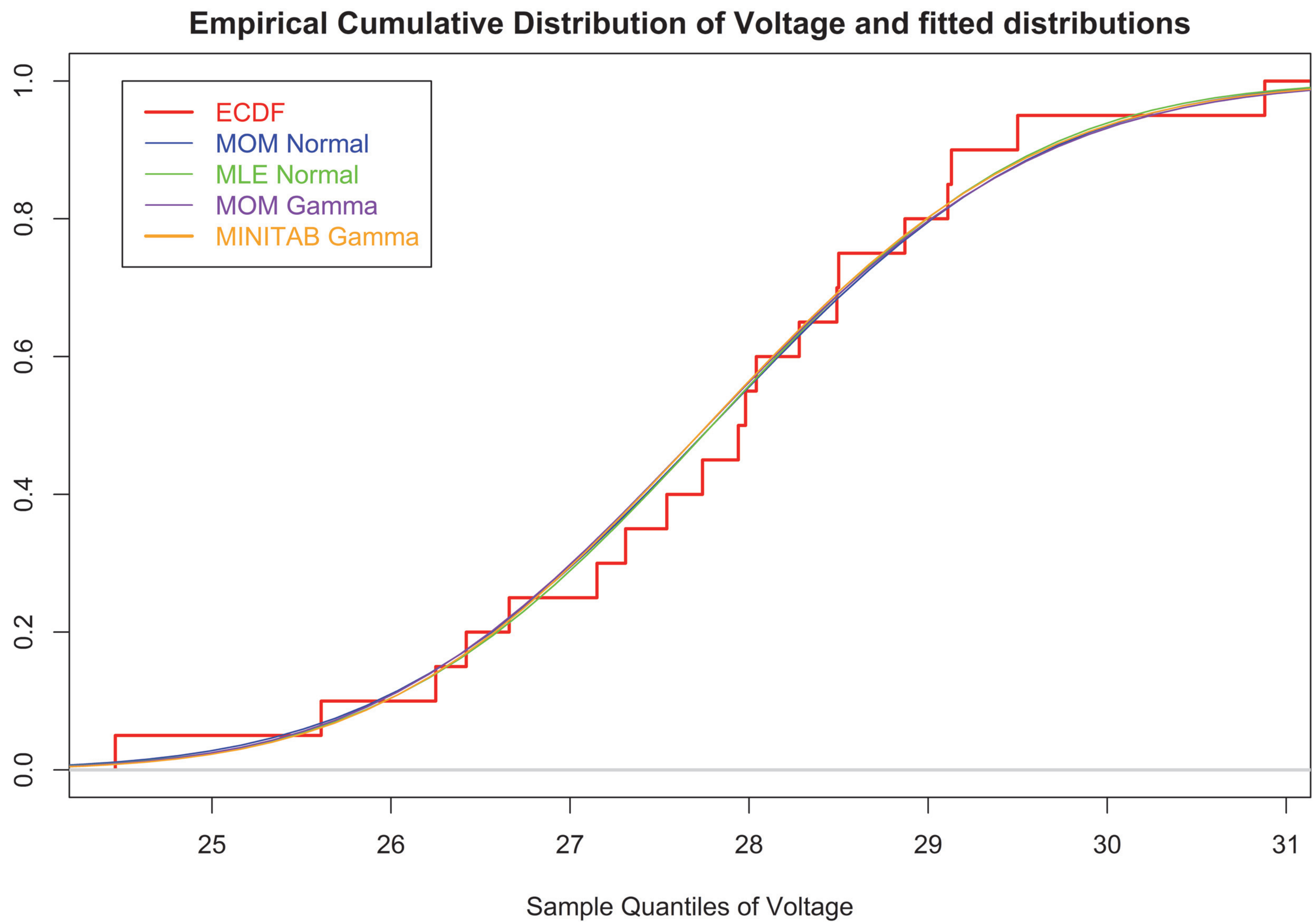


"Voltage_ChiSquared_Test_Normal_Distribution_Equal_Bin_Width.R"

Graphical Depiction of **Equal-Probability χ^2- goodness-of-fit test**



"Voltage_ChiSquared_Test_Normal_Distribution_Equal_Probability.R"


Empirical Cumulative Distribution of Voltage and fitted distributions
ECDF
MOM Normal
MLE Normal
MOM Gamma
MINITAB Gamma
0.0
0.2
0.4
0.6
0.8
1.0
25
26
27
28
29
30
31
Sample Quantiles of Voltage

- **Given an "adequately" fitted theoretical distribution $F_X(\cdot|\underline{\widehat{\theta}})$** to a dataset $(x_1, \ldots, x_n)$ from an $i.i.d.$ random sample, we can establish a $p \times 100\%$ credibility interval estimate $[l, u]$ such that

$$Pr(X \in [l, u]) \approx p$$

by setting:

$$\begin{cases} l = x_{(1-p)/2} = F_X^{-1}(\frac{1-p}{2}|\underline{\widehat{\theta}}), \\ u = x_{1-(1-p)/2} = F_X^{-1}(1 - \frac{1-p}{2}|\underline{\widehat{\theta}}) \end{cases}$$

where $x_{(1-p)/2}$ and $x_{1-(1-p)/2}$ are quantiles of **the cumulative distribution function $F_X(x|\underline{\widehat{\theta}})$.**

- For example, if we set $p = 0.90$, we have

$$\begin{cases} l = x_{0.05} = F_X^{-1}(0.05|\underline{\widehat{\theta}}), & \text{The 5-th percentile,} \\ u = x_{0.95} = F_X^{-1}(0.95|\underline{\widehat{\theta}}), & \text{The 95-th percentile.} \end{cases}$$

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

$$\overline{x} \approx 27.793,\ s^2 \approx 2.137 \text{ or } s \approx 1.462$$

We have for a 90% credibility interval **using the normal distribution fit**:

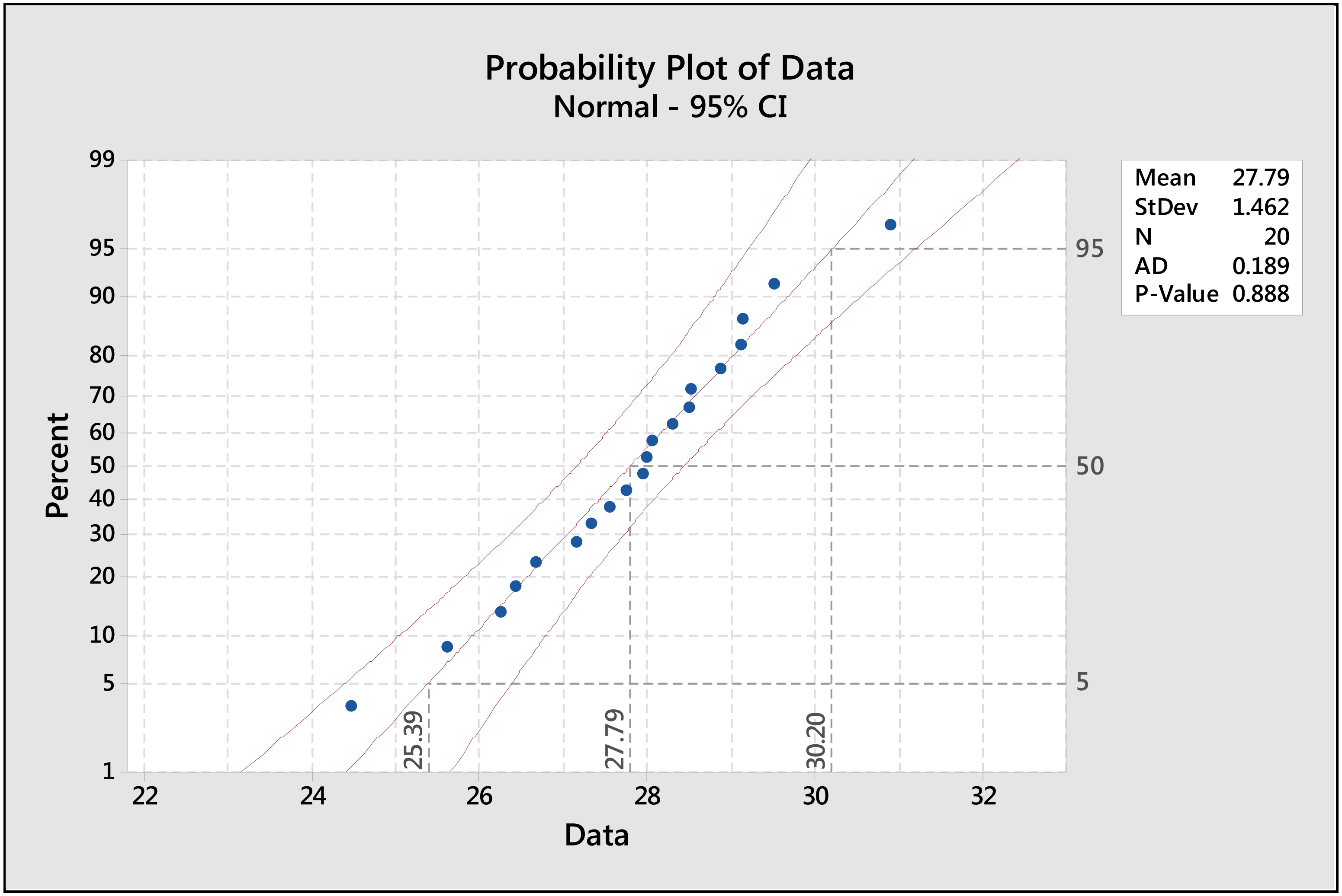
$$\begin{cases} l = x_{0.05} \approx 25.39, \\ u = x_{0.95} \approx 30.20 \end{cases}$$

Compare this with the earlier established **90% two-sided confidence interval:**

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

WHAT IS THE DIFFERENCE?

Probability Plots in MINITAB to calculate Credibility Intervals



- **Probability plots** in MINITAB are a **powerful visual tool** for **testing goodness of fit**.

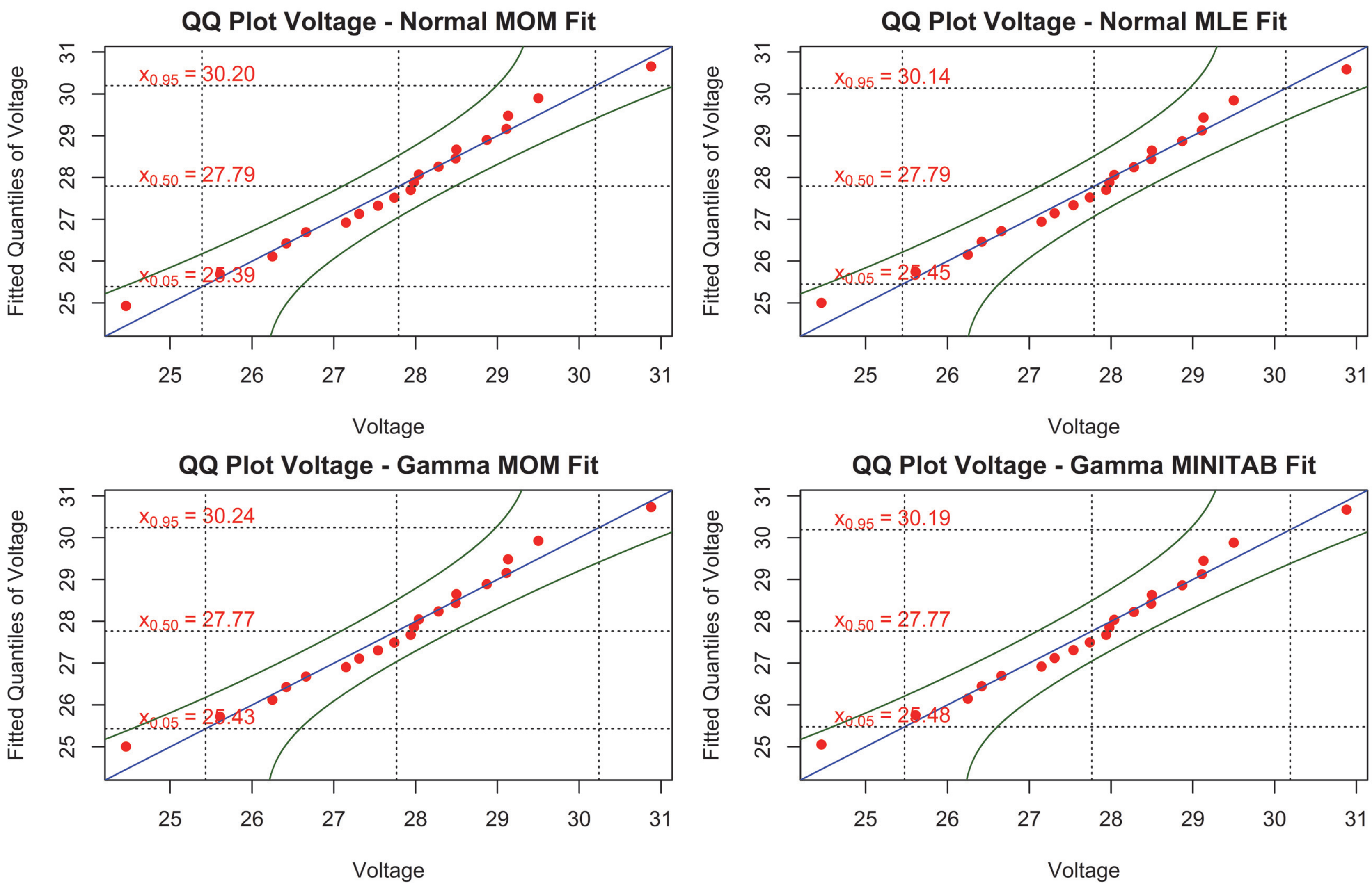
- **The AD value is the statistic value of the Anderson-Darling goodness-of-fit test** (**similar in spirit as the χ^2-test**). Large values of the AD-statistic indicate a larger deviation from the fitted theoretical distribution.

- The **larger the *p*-value the larger the support** for $F_X(\cdot \mid \underline{\widehat{\theta}})$.

- If the theoretical distribution is **a perfect fit** of all data points should form **a straight line.**

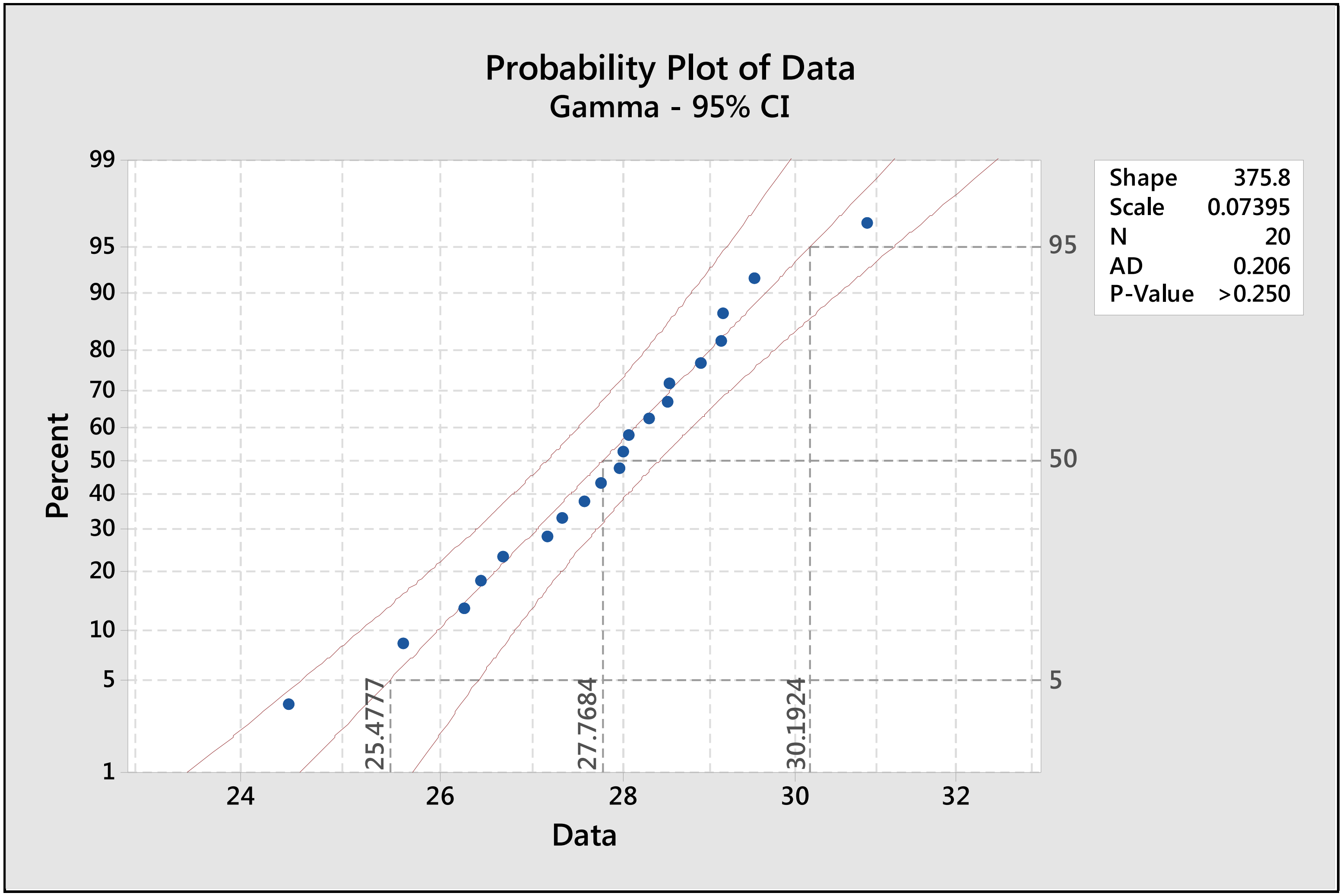
- **Deviations from the straight line** show **deviations from the fitted theoretical distribution.**

- **When can a data point be considered an outlier? Answer:** when a data point is **outside the boundaries that are drawn**. The boundaries in the above figure are **95% confidence intervals** for the cdf-value $F_X(\cdot \mid \underline{\widehat{\theta}})$.



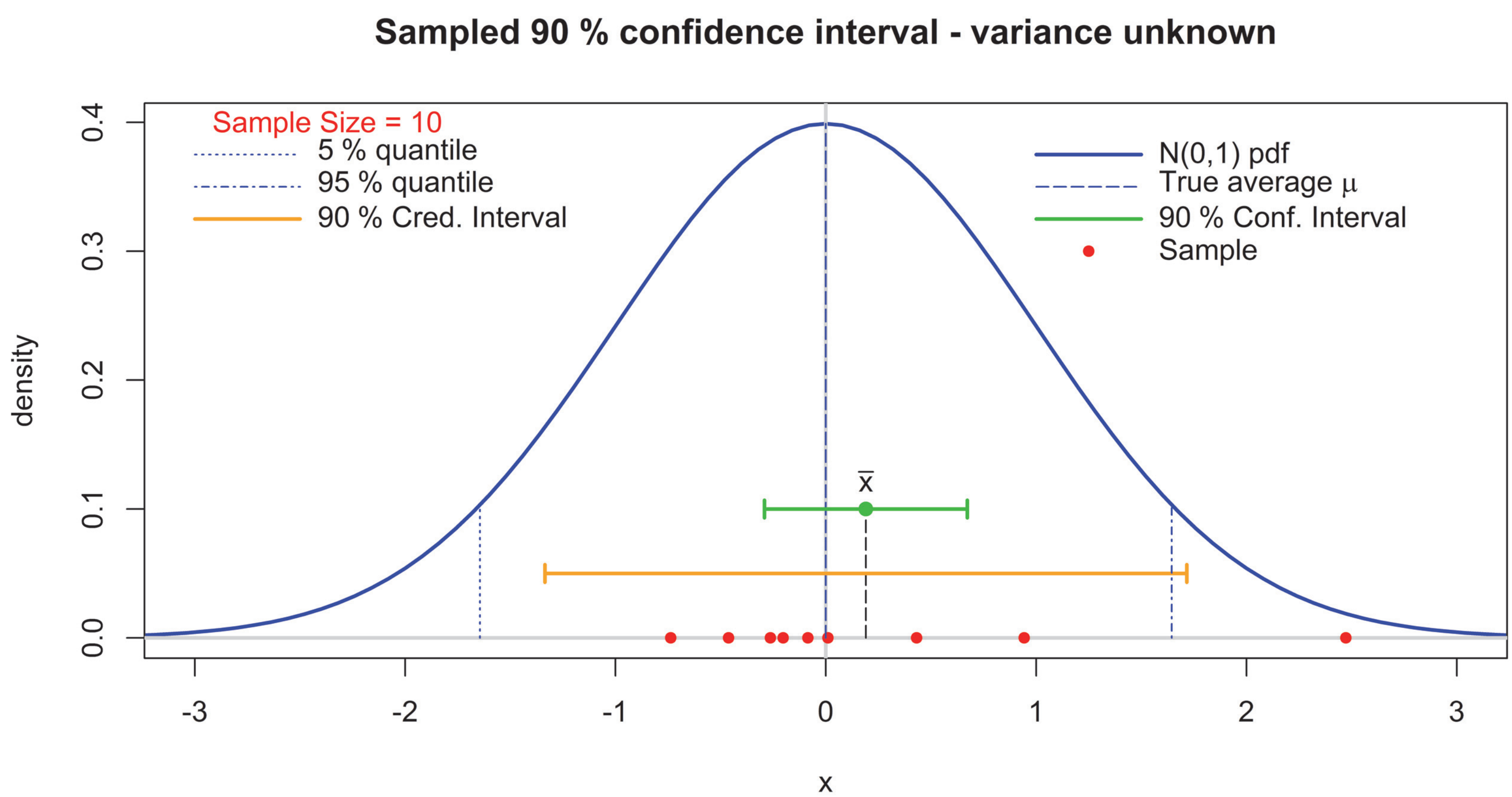
Analysis in file "Voltage_QQ_Plots.R"

Probability Plots in MINITAB to calculate Credibility Intervals



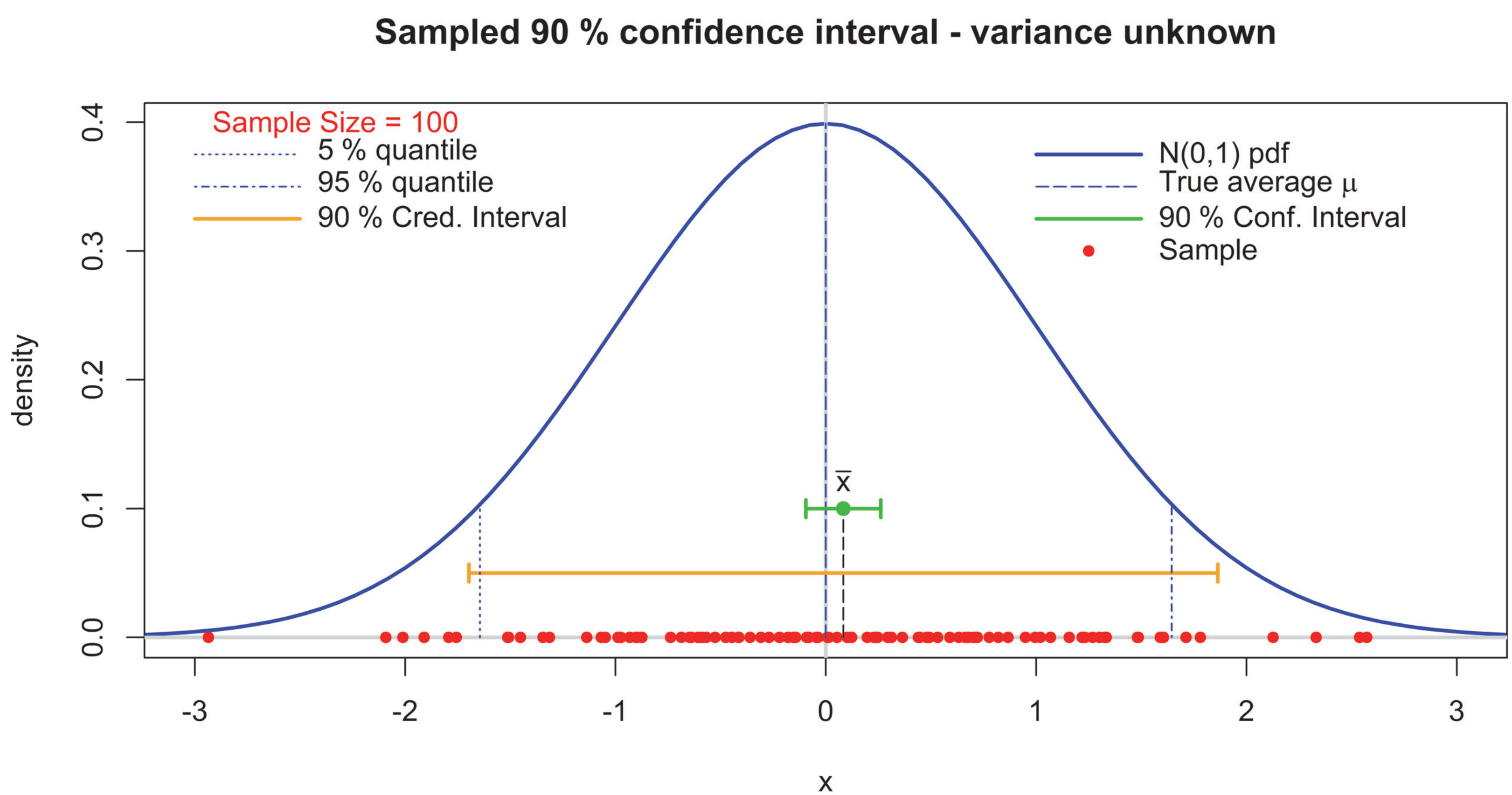
- The 90% confidence interval above was calculated for $E[X]$, not X!
- This is true in general: **confidence intervals are calculated for a characteristic of X**, such as for example $E[X]$, $Var[X]$, etc.
- **For a dataset $(x_1, \ldots, x_n)$** no probability interpretation can be assigned to an evaluated $(1-\alpha)100\%$ confidence interval. Recall, it is **a realization of a randomly changing interval based on an $i.i.d.$ random sample $(X_1, \ldots, X_n)$ where that randomly changing interval has $(1-\alpha)100\%$ probability of capturing $E[X]$.**
- When **the sample size n increases the width of confidence intervals decrease**. They converge to the true value (a single point) of e.g. $E[X]$, $Var[X]$.
- **For a dataset $(x_1, \ldots, x_n)$**, the probability that **a realization of the random variable X** is a member within an $(1-\alpha)100\%$ credibility interval for X (which is also random interval) equals **approximately** $(1-\alpha)$.

- The **width of the confidence interval decreases**, as the sample size increases whereas **the width of the credibility interval does not.**



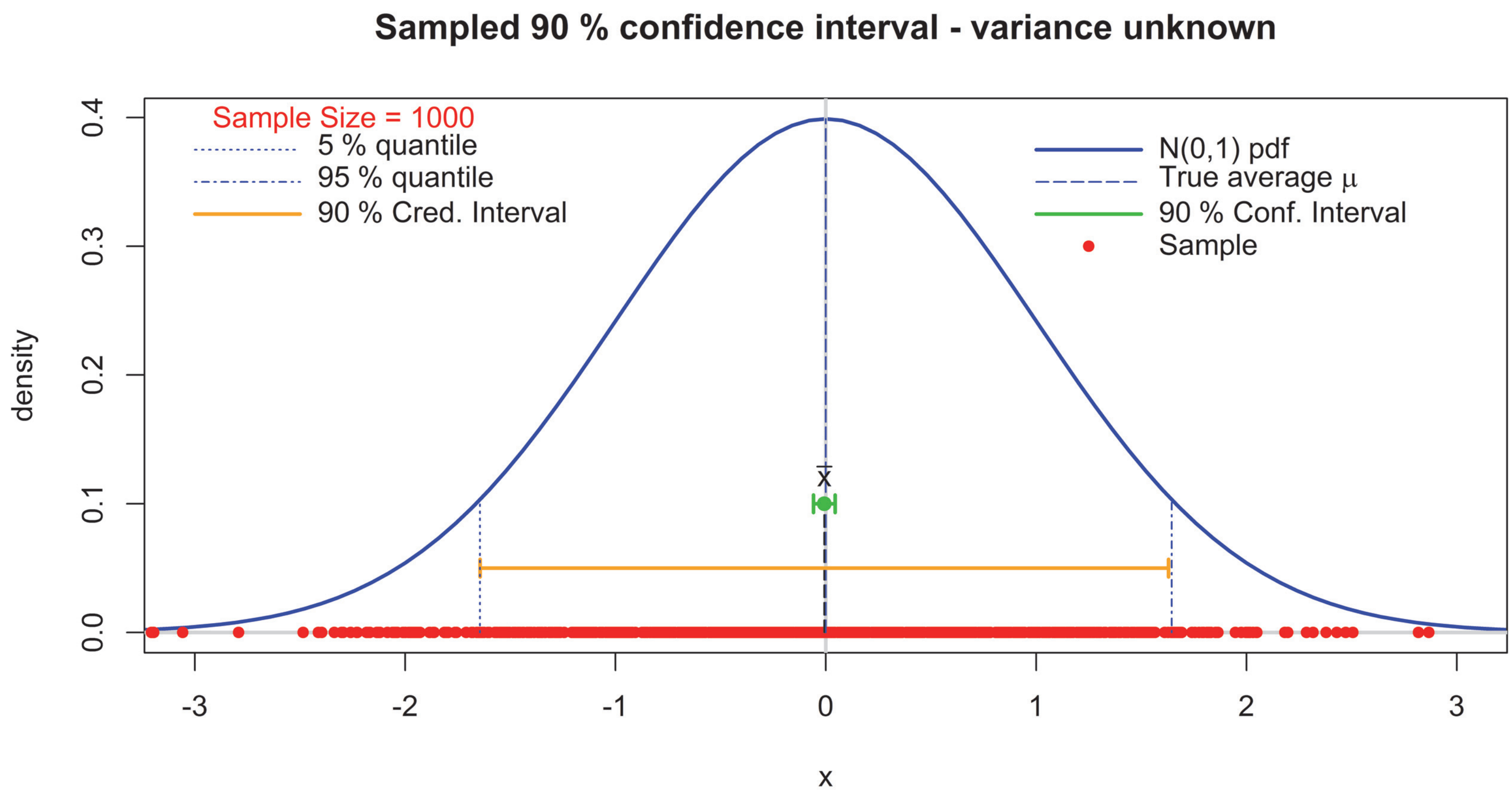
Analysis in file "Norm_Conf_Cred_Var_UnKnown.R"

- The **width of the confidence interval decreases**, as the sample size increases whereas **the width of the credibility interval does not.**



Analysis in file "Norm_Conf_Cred_Var_UnKnown.R"

- The **width of the confidence interval decreases**, as the sample size increases whereas **the width of the credibility interval does not.**



Analysis in file "Norm_Conf_Cred_Var_UnKnown.R"