

02-24-00201

Probability and Statistics II

DR. AHMED TAYEL

Department of Engineering Mathematics and Physics, Faculty of
Engineering, Alexandria University

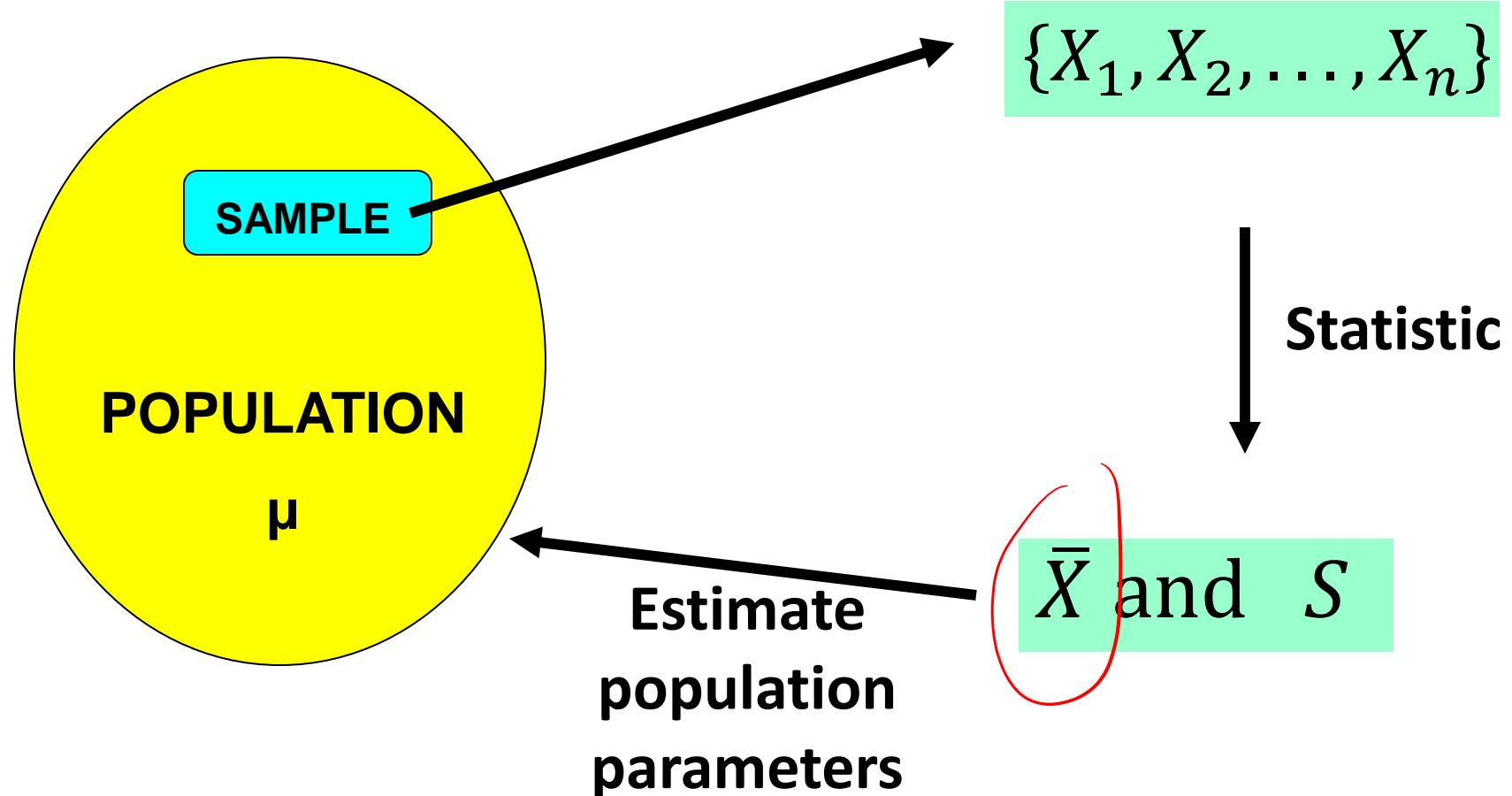
ahmed.tayel@alexu.edu.eg

Outline

1. Review.
2. Distributions derived from the Normal distribution.
 1. The Chi-squared distribution [**Previous lecture**].
 1. Distribution of s^2 .
 2. The t-distribution [**This lecture**].
 3. The F-distribution [**This lecture**].

1. Review

The Sampling Process



Central Tendency in the Sample

Definition:

If X_1, X_2, \dots, X_n represents a random sample of size n , then the sample mean is defined to be the statistic:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

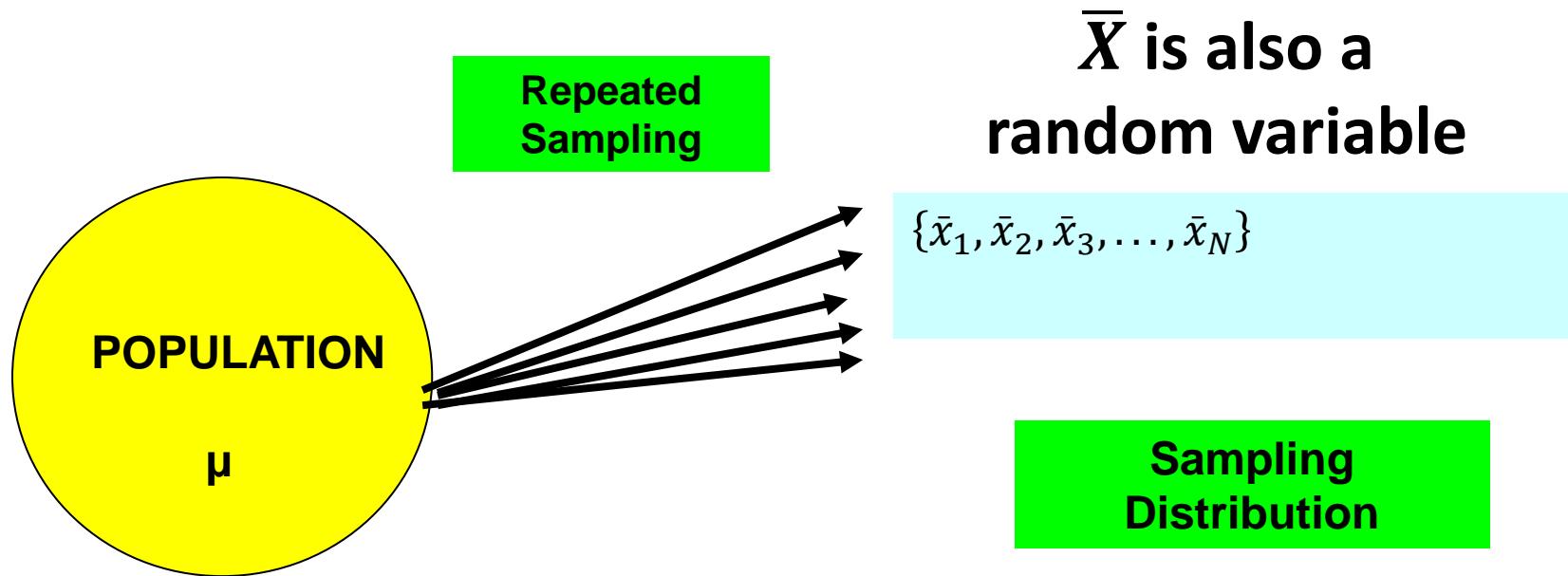
Variability in the Sample

Definition:

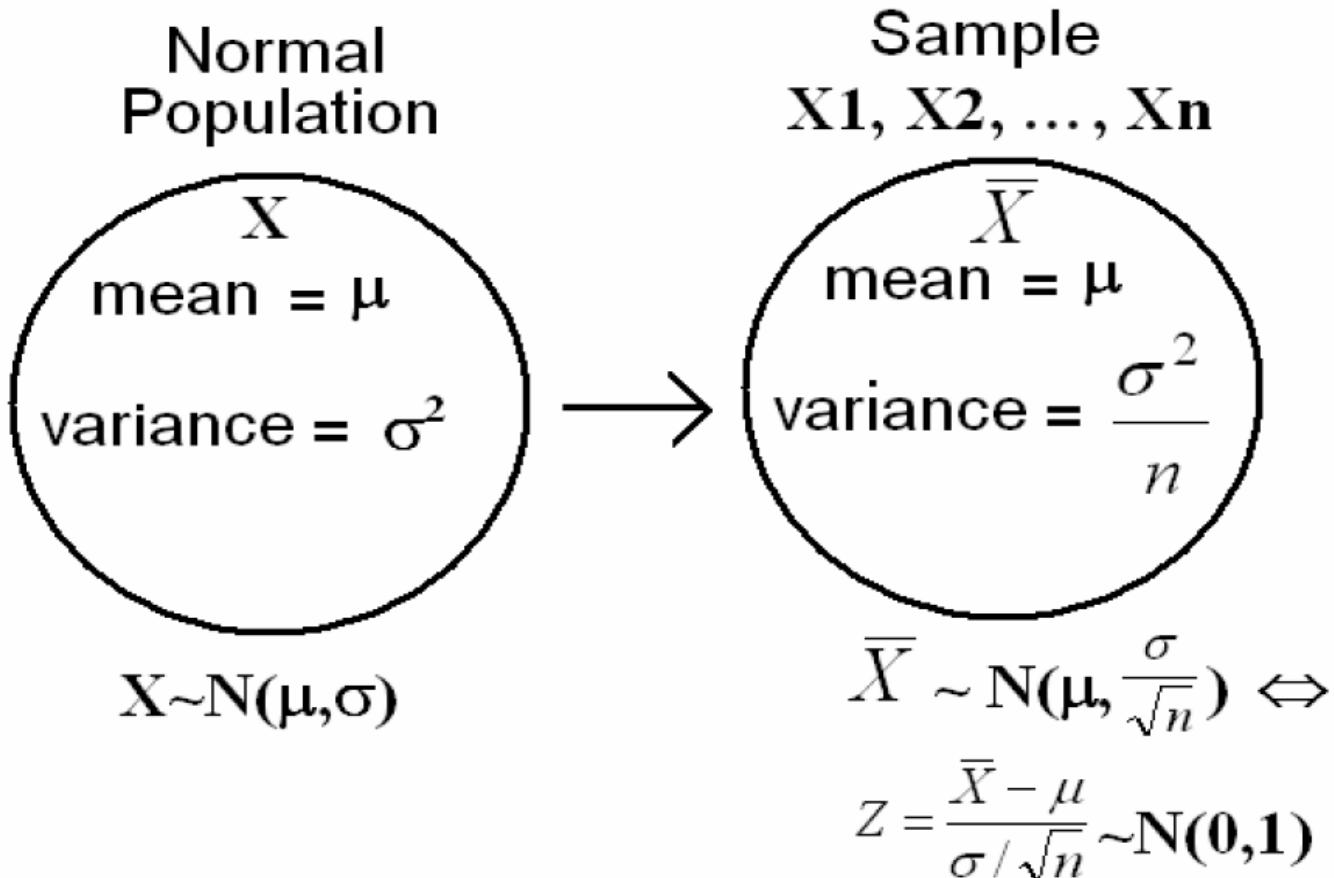
If X_1, X_2, \dots, X_n represents a random sample of size n , then the sample variance is defined to be the statistic:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1} \text{ (unit)}^2$$

The Sampling Distribution



- If X_1, X_2, \dots, X_n is a random sample of size n from $N(\mu, \sigma)$, then $\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}})$ or $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.
- $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$



What if the population is not normally distributed?

Theorem: (Central Limit Theorem)

If X_1, X_2, \dots, X_n is a random sample of size n from any distribution (population) with mean μ and finite variance σ^2 , then, if the sample size n is large, the random variable $n \geq 30$

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

is approximately standard normal random variable, i.e.,

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1) \text{ approximately.}$$

Case of σ^2 is unknown

If $n \geq 30$, the central limit theorem (CLT) is still valid.

If we replace σ^2 by s^2

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim N(0,1)$$

Summary

Given the random sample X_1, X_2, \dots, X_n .

We have **three cases** for the distribution of \bar{X}

[1]

- Sample taken from a normal population
- σ^2 is known

$$\bar{X} \sim \text{Norm}\left(\mu, \frac{\sigma^2}{n}\right)$$

[2]

- $n \geq 30$ and sample taken from any distribution

$$\bar{X} \sim \text{Norm}\left(\mu, \frac{\sigma^2}{n}\right)$$

[3]

- Sample taken from a normal population
- $n < 30$
- σ^2 is unknown
- This lecture
- t-distribution

$$\bar{X} \sim \text{Norm}\left(\mu, \frac{s^2}{n}\right)$$

2. The t-distribution

- Large sample
- s is a good estimate to σ
- does not vary much from sample to sample

σ
unknown
Normal population

- Small sample
- s^2 fluctuates considerably from sample to sample

$n \geq 30$

$n < 30$

By CLT

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

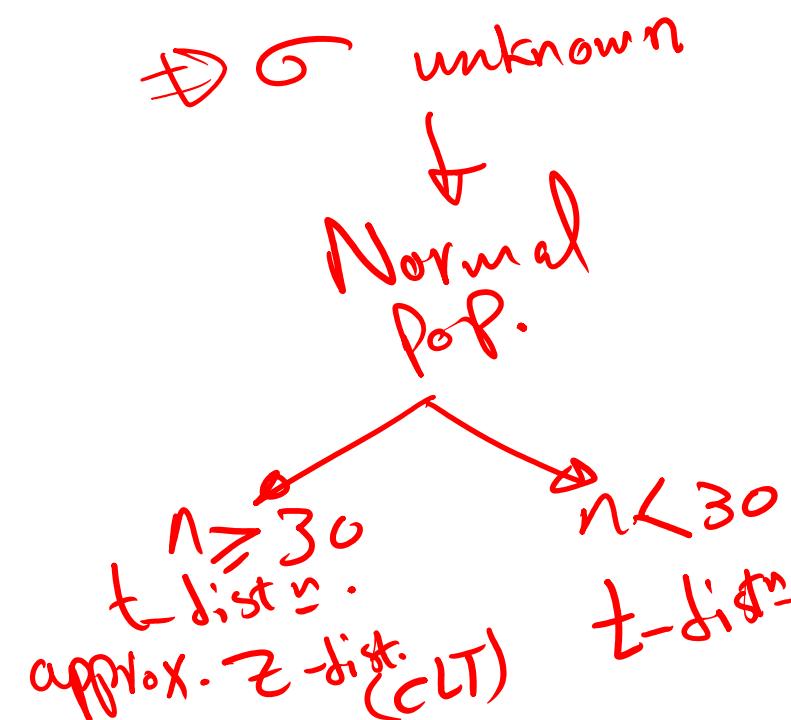
$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim ??$$

Student's t Distribution

Student's t Distribution

- Let $Z \sim N(0,1)$ $V \sim \chi^2_k$ Z, V independent
- Define: $T = \frac{Z}{\sqrt{V/k}}$ $\rightarrow t$ distribution with k degrees of freedom
- Application: $X_1, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$
 - $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma} \right) \sim N(0,1)$
 - $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ Z, V Independent
 - $T = \frac{\sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma} \right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \sqrt{n} \left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \right) = \left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \right) \sim t(n-1)$

t distribution with $n - 1$ degrees of freedom



Student's t Distribution

If the population standard deviation, σ , is **unknown**, replace σ with the sample standard deviation, s . If the population is normal, the resulting statistic:

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

Population must be s^2 , normal

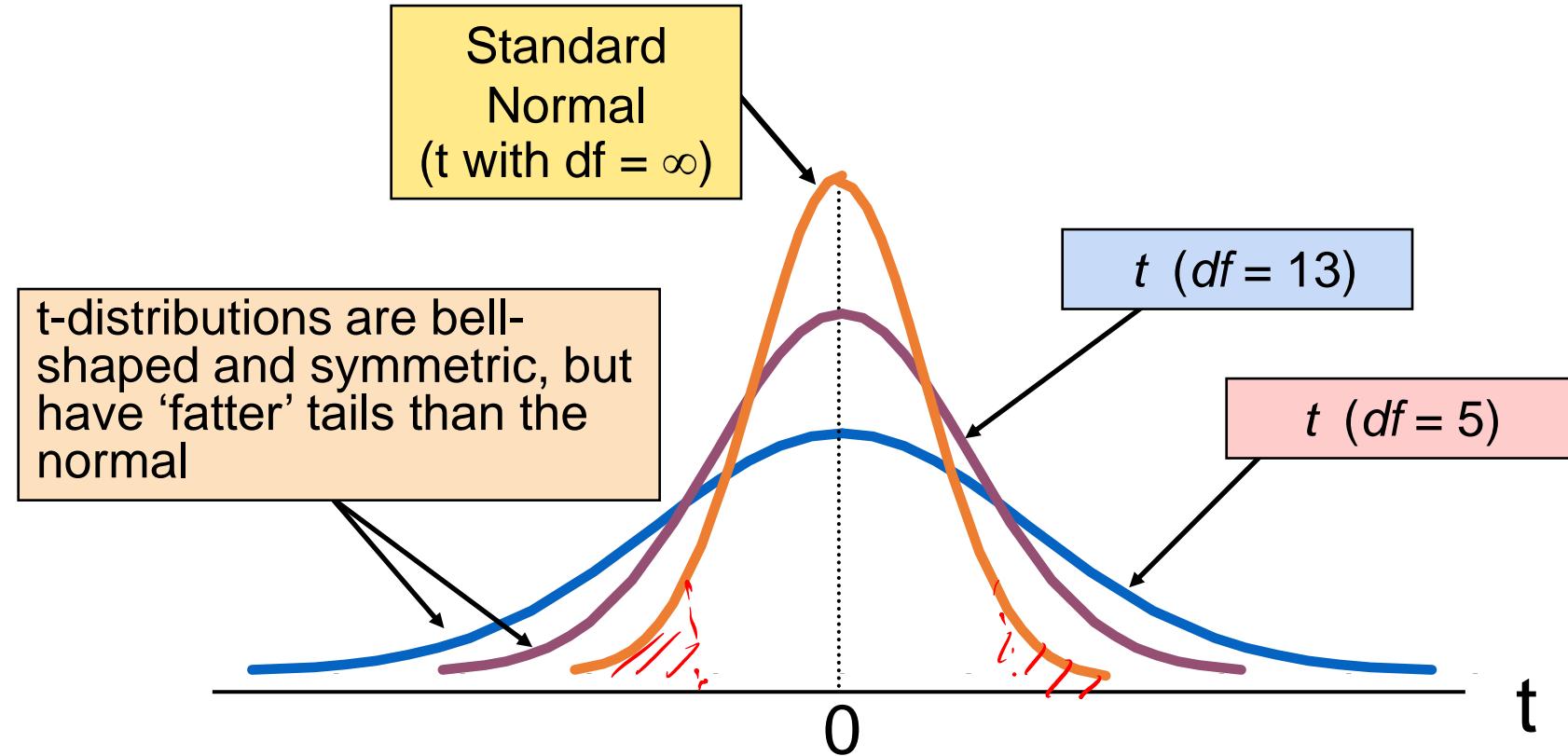
Normal

$$T = \frac{\bar{Z}}{\sqrt{\sum_{k=1}^n x_k^2 / n}}$$

has a **t distribution with $(n - 1)$ degrees of freedom.**

Student's t Distribution

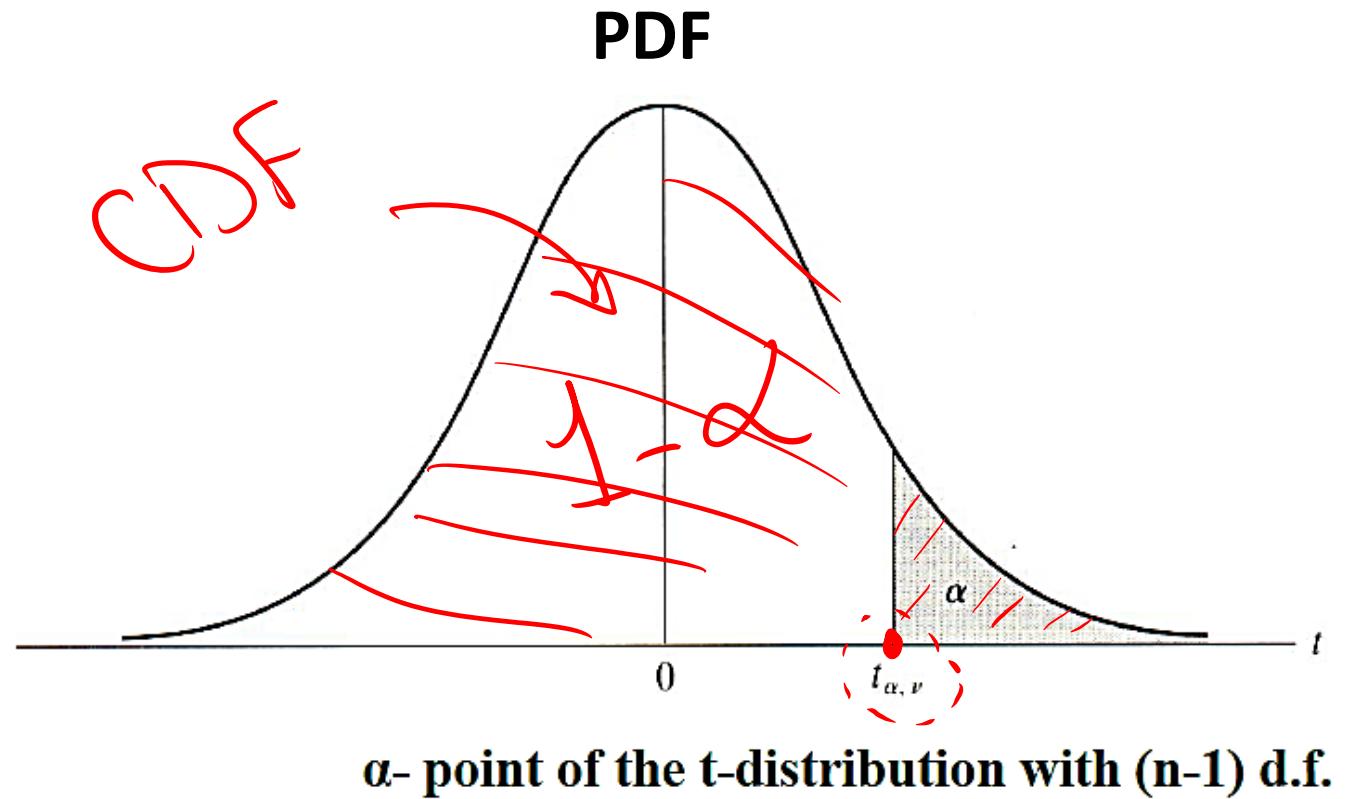
- The expected value of t is 0.
- The variance of t is greater than 1, but approaches 1 as the number of degrees of freedom increases.
- When the sample size is small (<30) we use t distribution.



Student's t Distribution

area under PDF after the point $t_{\alpha, k}$

- $t_{\alpha, k}$
- $P(U > t_{\alpha, k}) = \alpha$
- $F_{CDF}(t_{\alpha, k}) = 1 - \alpha$
- $t_{\alpha, k} = F_{CDF}^{-1}(1 - \alpha)$



Student's t Distribution

Area of half the shape

Degrees of freedom

- $t_{0.5, 5} = 0$

- $t_{0.025, 5} = 2.57 \rightarrow t_{0.975, 5} = -2.57$

• $F(t_{0.025, 5}) = 1 - 0.025 = 0.975$

- $\Pr(t_{\alpha_1} < T < t_{\alpha_2}) = F(t_{\alpha_2}) - F(t_{\alpha_1})$

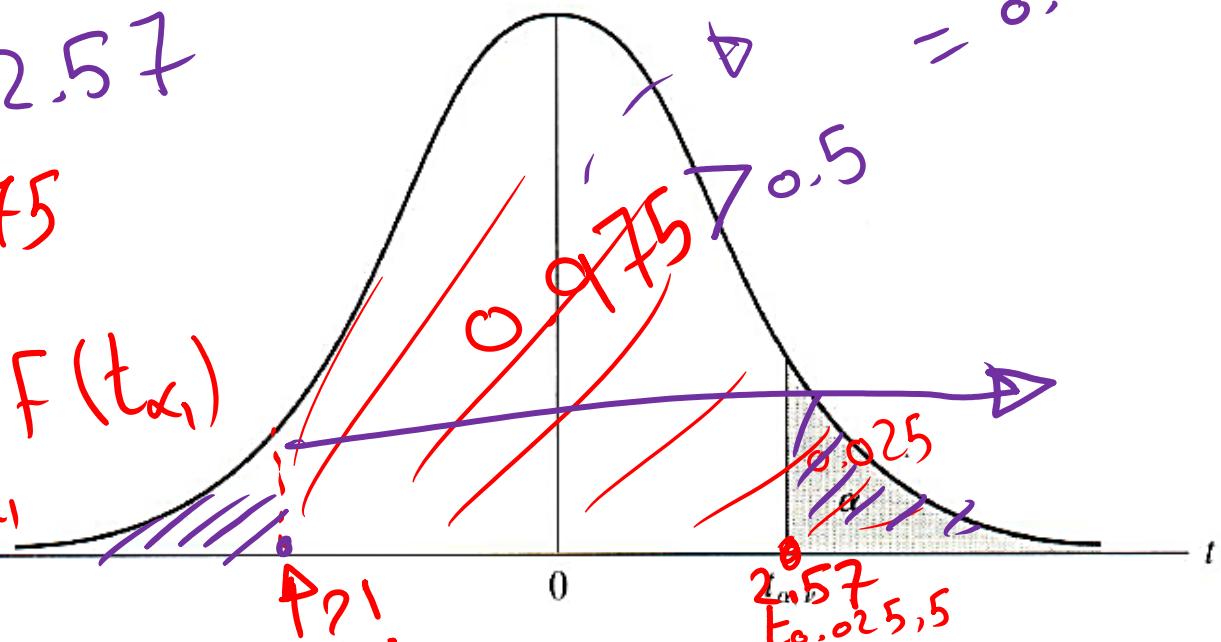
- $t_{0.025, 5} = ?$

$$= (1 - \alpha_2) - (1 - \alpha_1) = 1 - \alpha_2 - 1 + \alpha_1 = \alpha_1 - \alpha_2$$

- Using MINITAB or R

Similar to the std. normal

PDF



a-point of the t-distribution with (n-1) d.f.
-2.57

The figure shows a Minitab session window with the following details:

- File**: Minitab - Untitled
- Calc** menu open, showing options like Chi-Square..., Normal..., F..., t... (highlighted), Uniform..., Binomial..., Geometric..., Negative Binomial..., Hypergeometric..., Discrete..., Integer..., Poisson..., Beta..., Cauchy..., Exponential..., Gamma..., Laplace..., Largest Extreme Value..., Logistic..., Loglogistic..., Lognormal..., Smallest Extreme Value..., Triangular..., and Weibull...).
- Session** pane: Welcome to Minitab
- Worksheet 1 ***** pane: A graph showing several probability density functions (PDFs) for different distributions. The x-axis is labeled "df" with values 10 and 5. The y-axis has labels 1 through 9. A green curve is labeled "df 10" and a purple curve is labeled "df 5". A blue line is labeled "0.025" and "2.571". A yellow line is labeled "0.025" and "2.571".
- t Distribution** dialog box open:
 - Probability density option is not selected.
 - Cumulative probability option is not selected.
 - Inverse cumulative probability option is selected.
 - Noncentrality parameter: 0.0
 - Degrees of freedom: 5 (highlighted with a red arrow)
 - Input column: (empty)
 - Optional storage: (empty)
 - Input constant: 0.975 (highlighted with a red arrow)
 - Optional storage: (empty)
- Session** output pane:

10/16/2022 1:44:54 PM

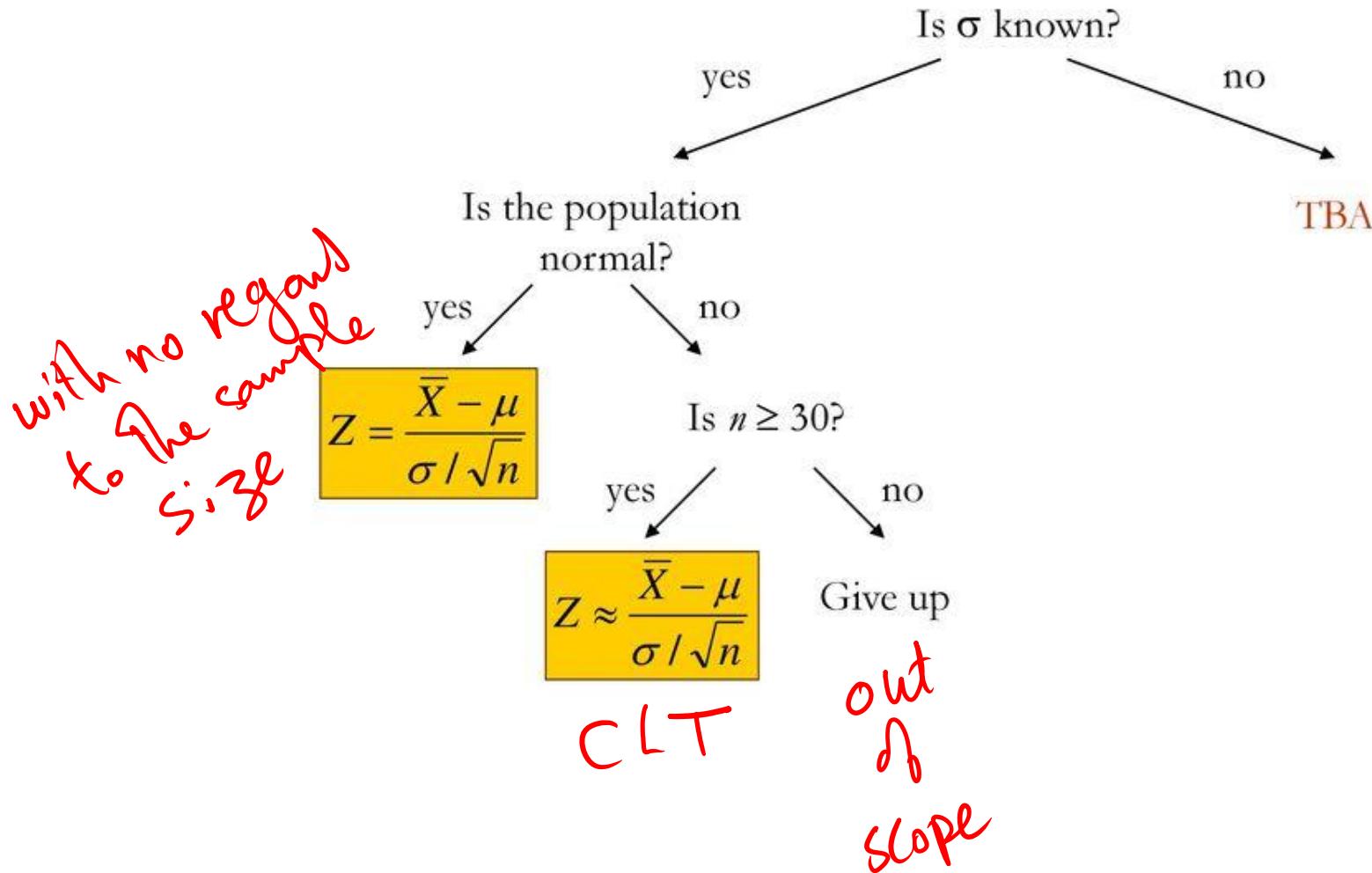
Welcome to Minitab, press F1 for help.

Inverse Cumulative Distribution Function

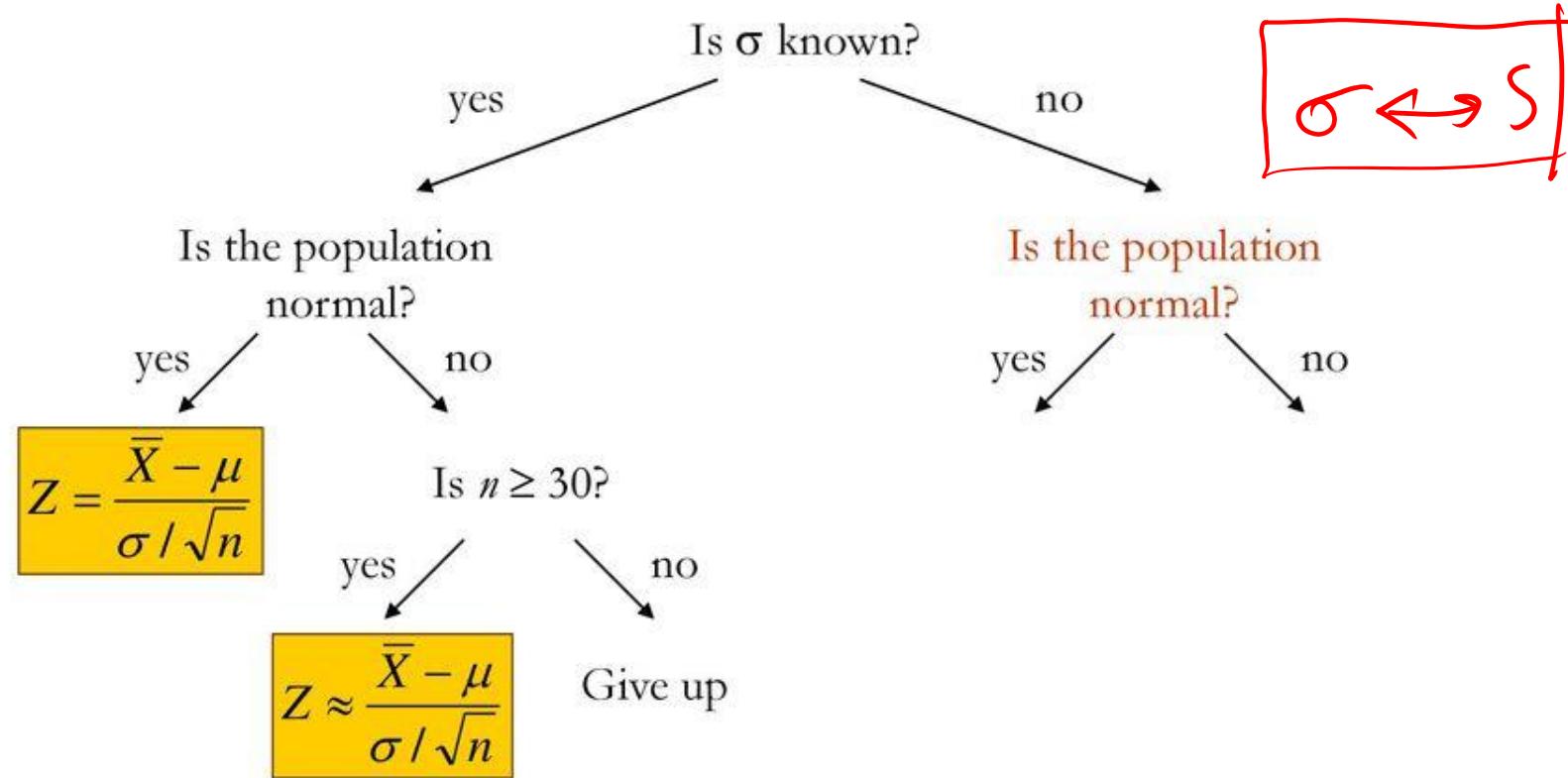
Student's t distribution with 5 DF

P(X <= x)	x
0.975	2.57058

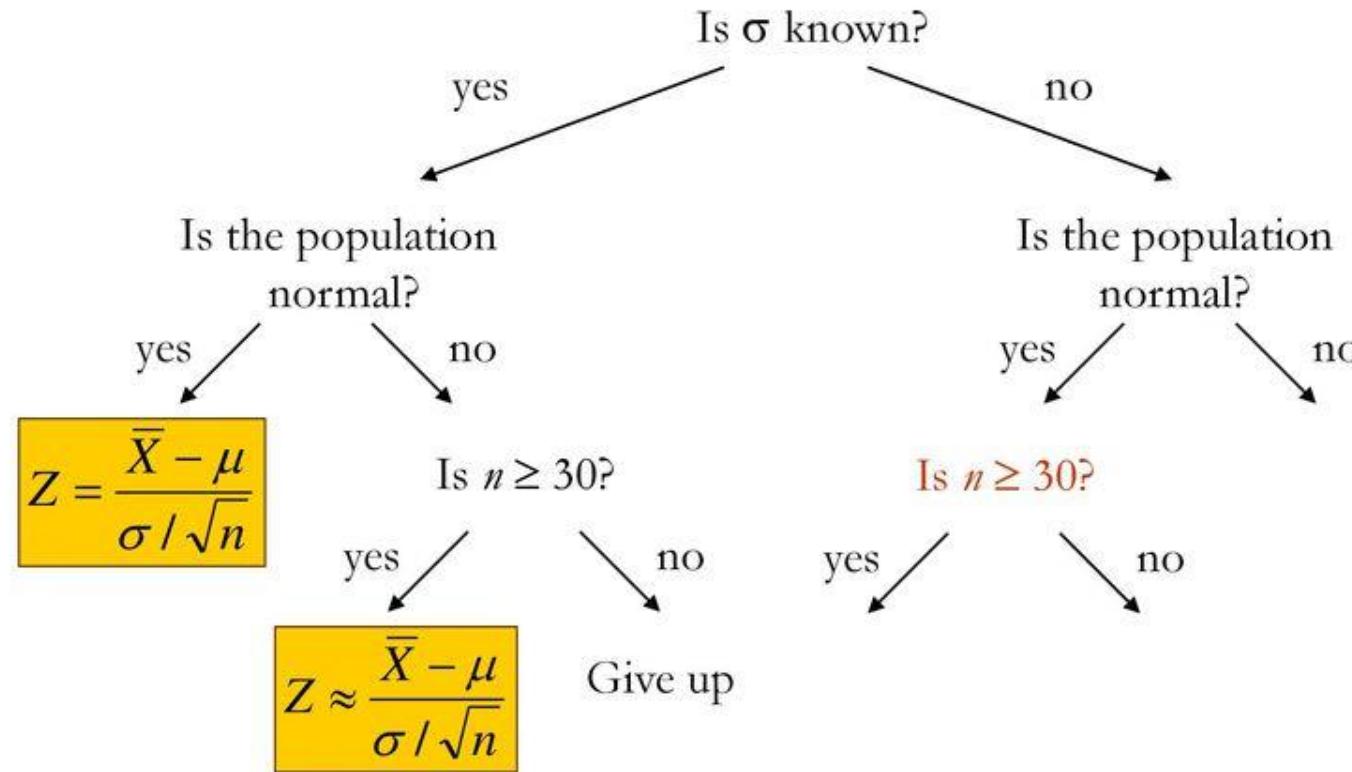
The Decision Tree



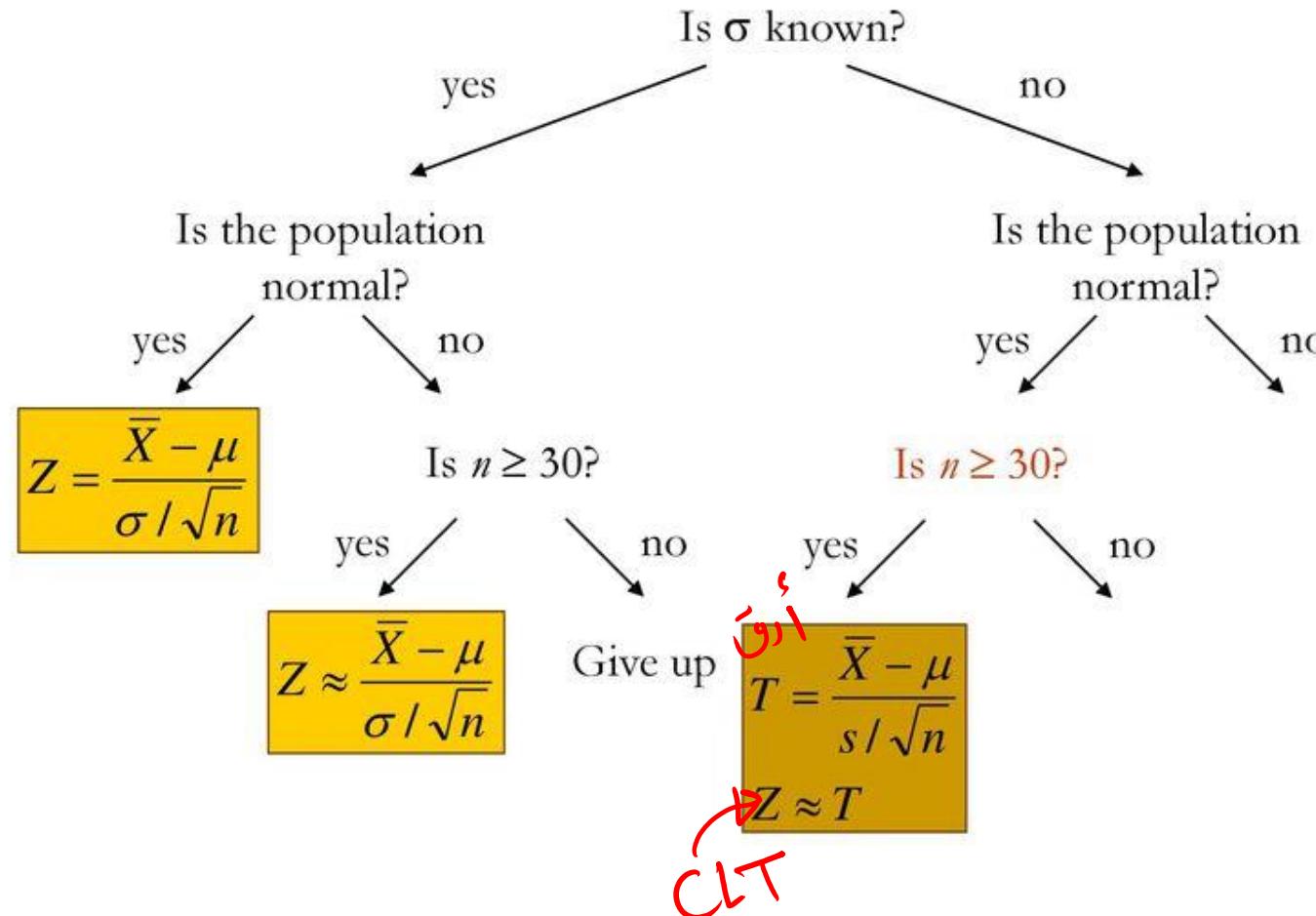
The Decision Tree



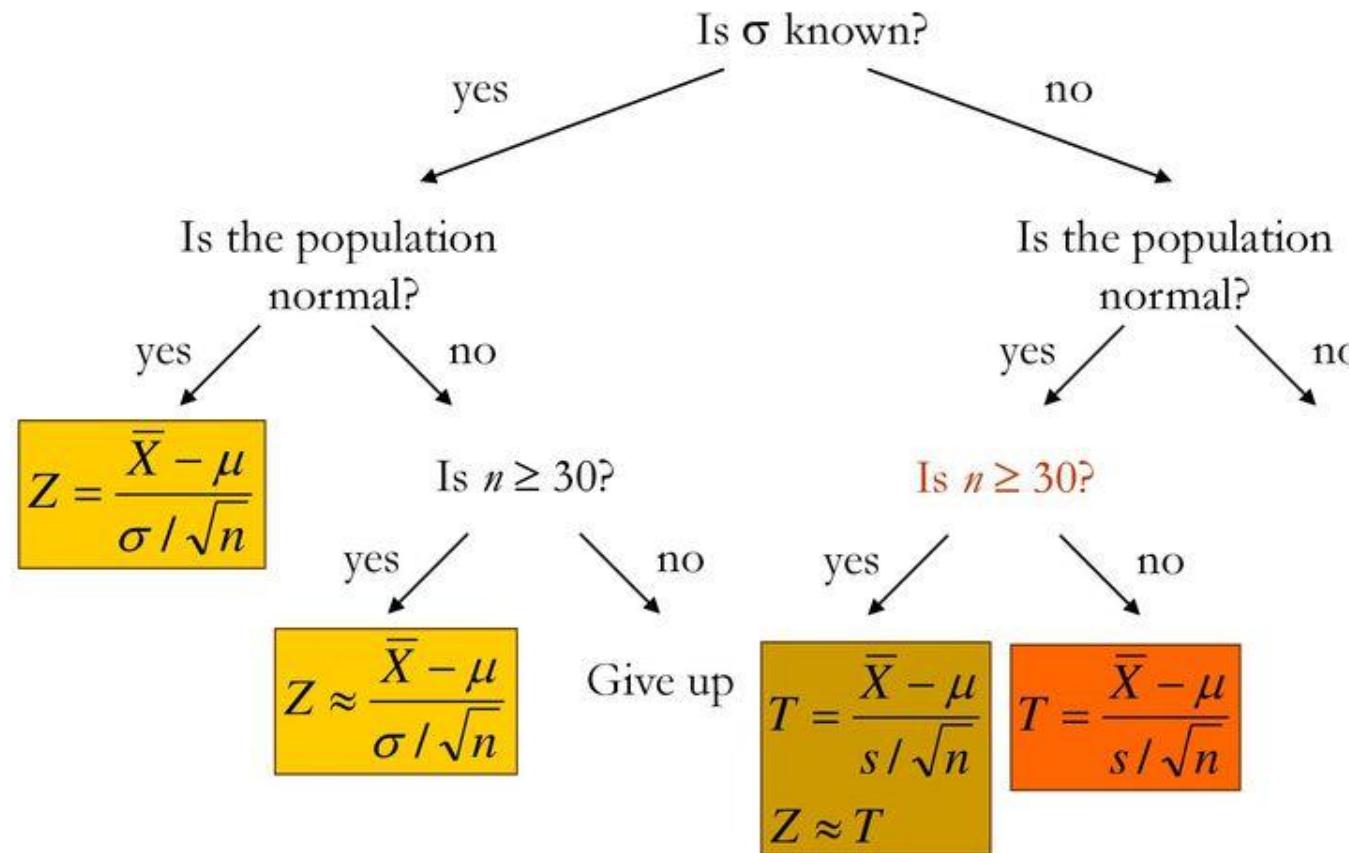
The Decision Tree



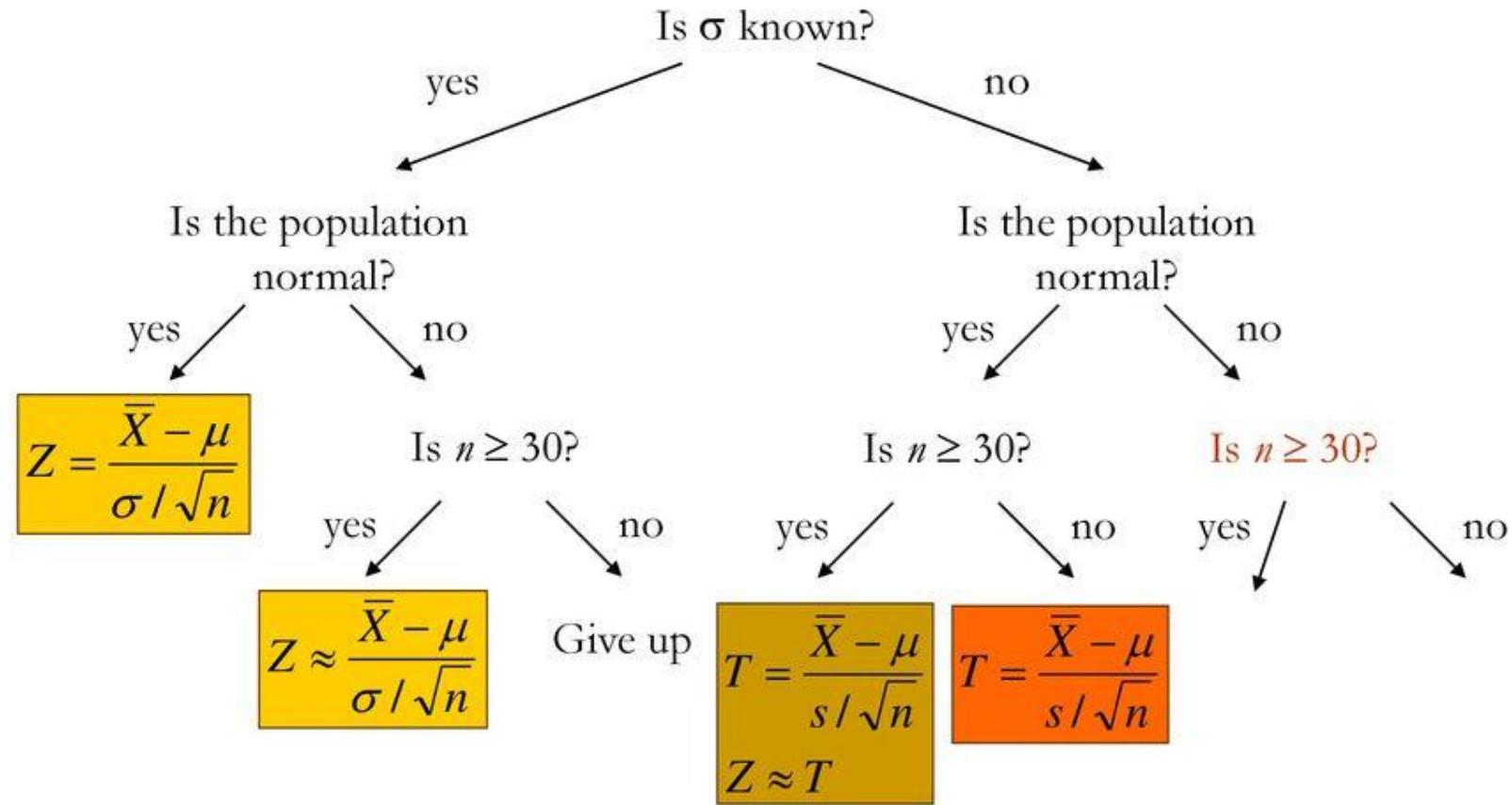
The Decision Tree



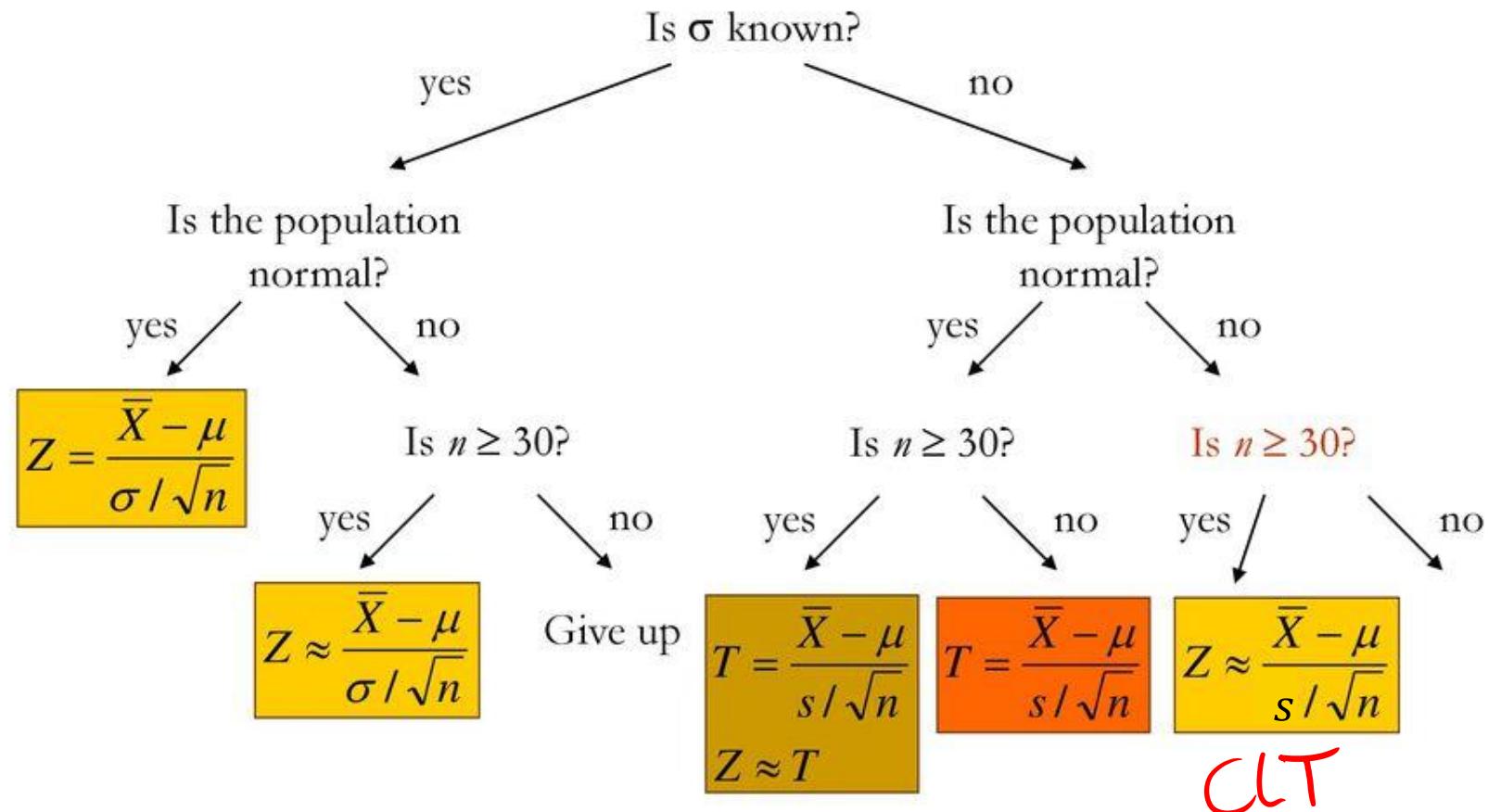
The Decision Tree



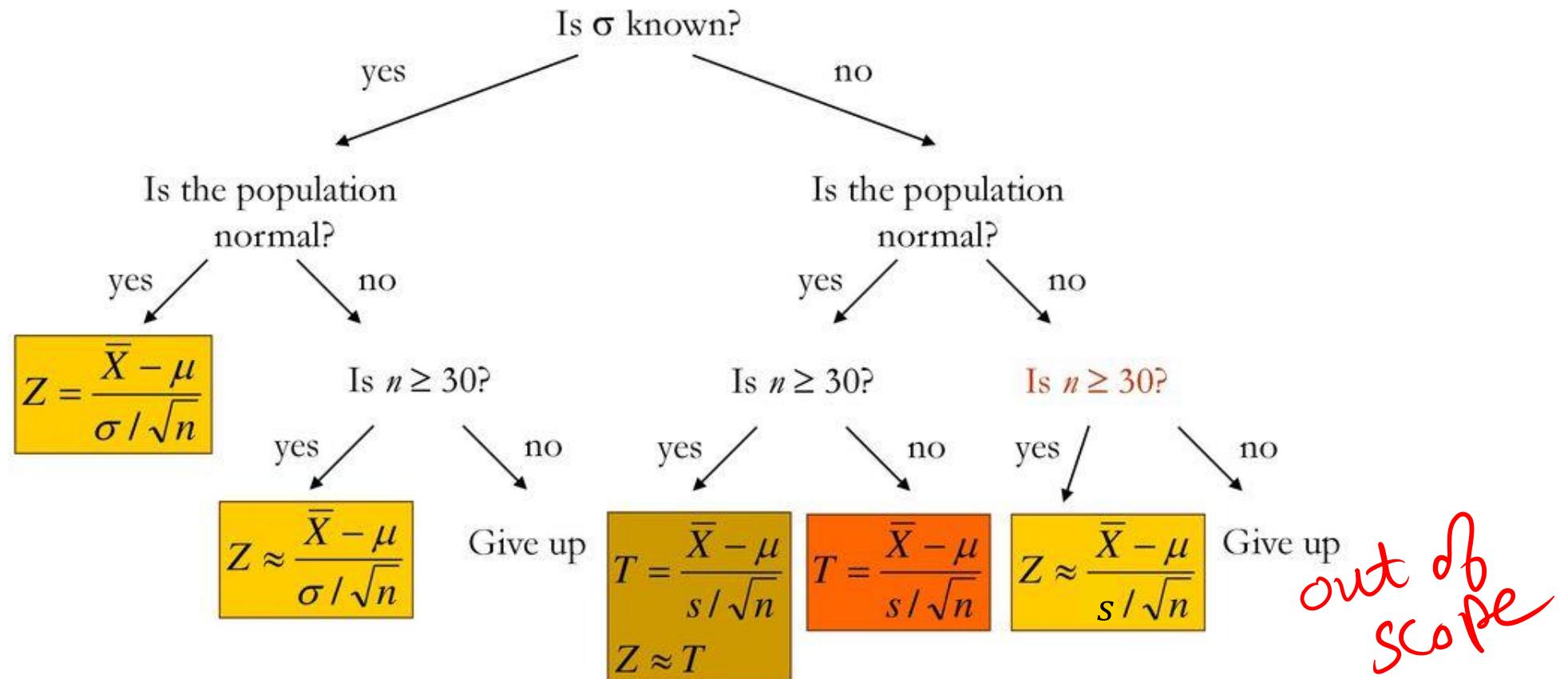
The Decision Tree



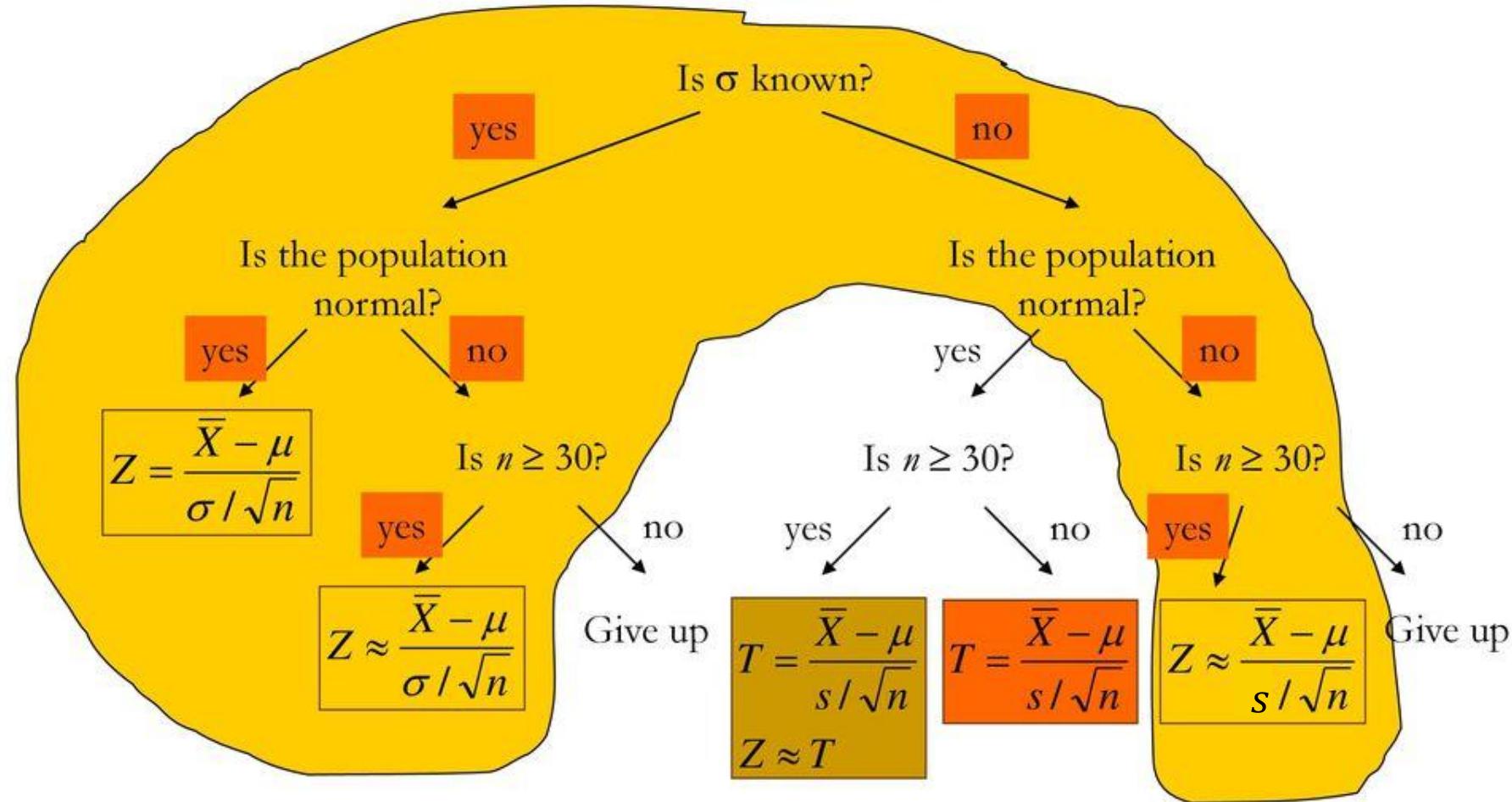
The Decision Tree



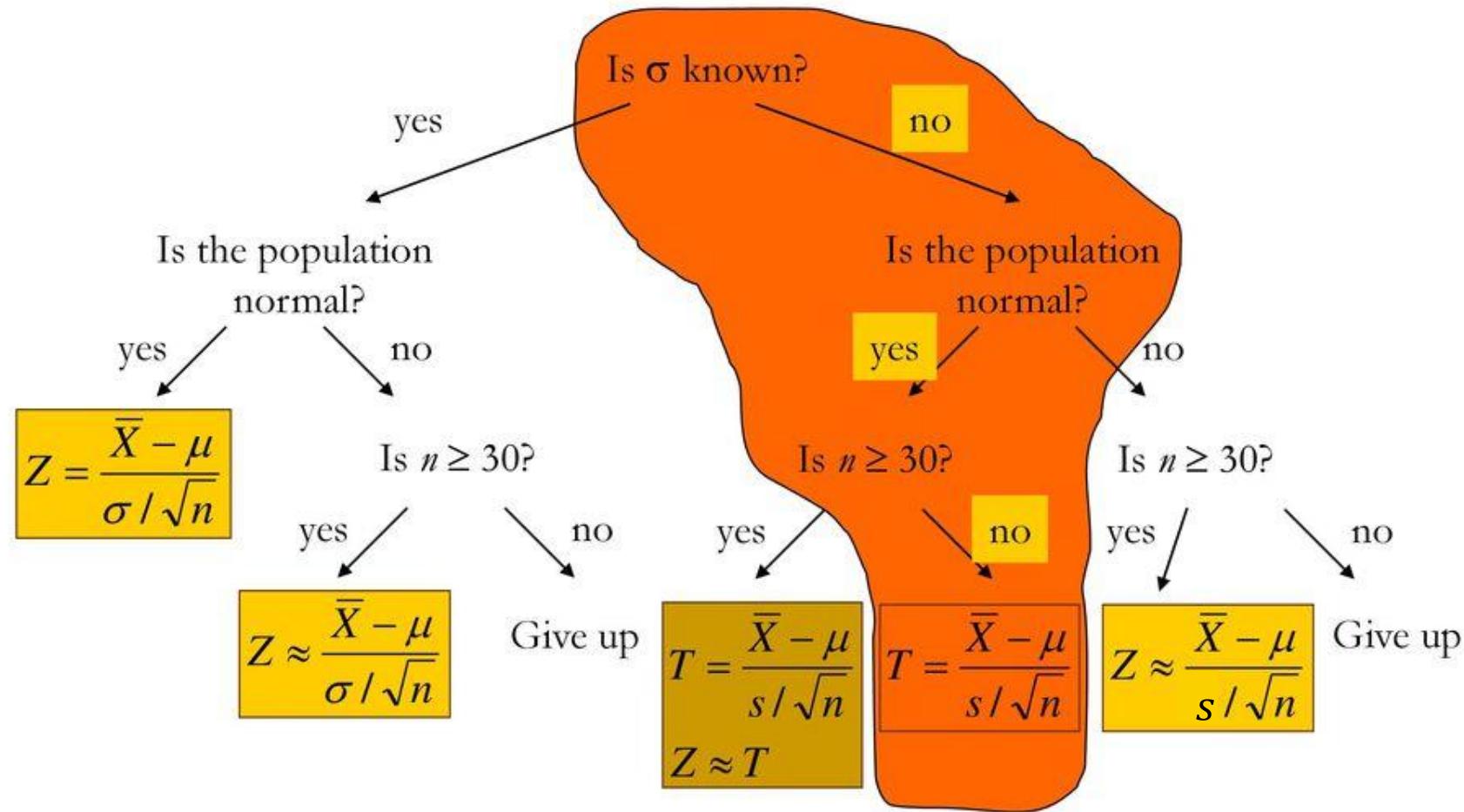
The Decision Tree



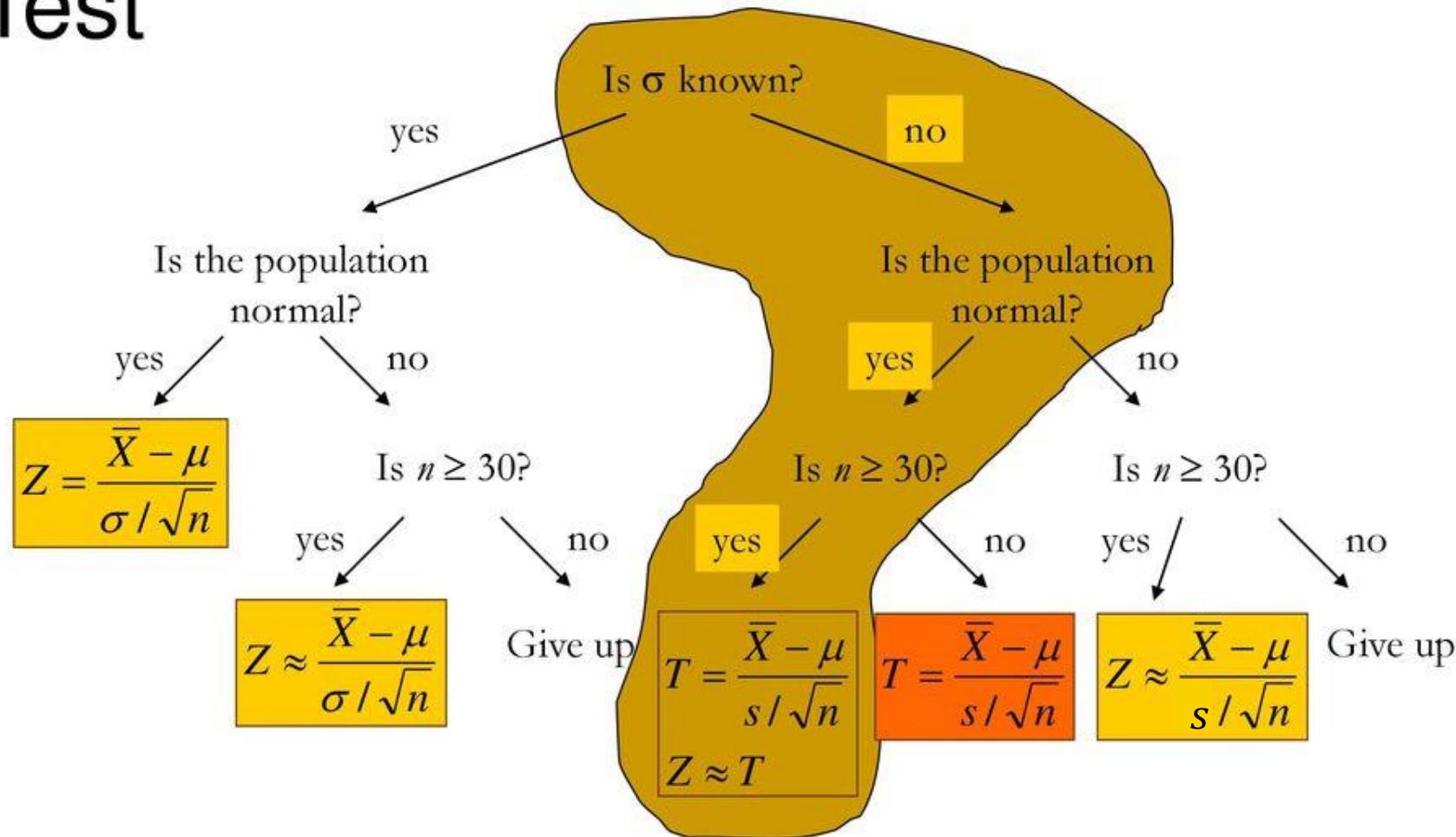
When to Use Z-Test



When to Use T-Test

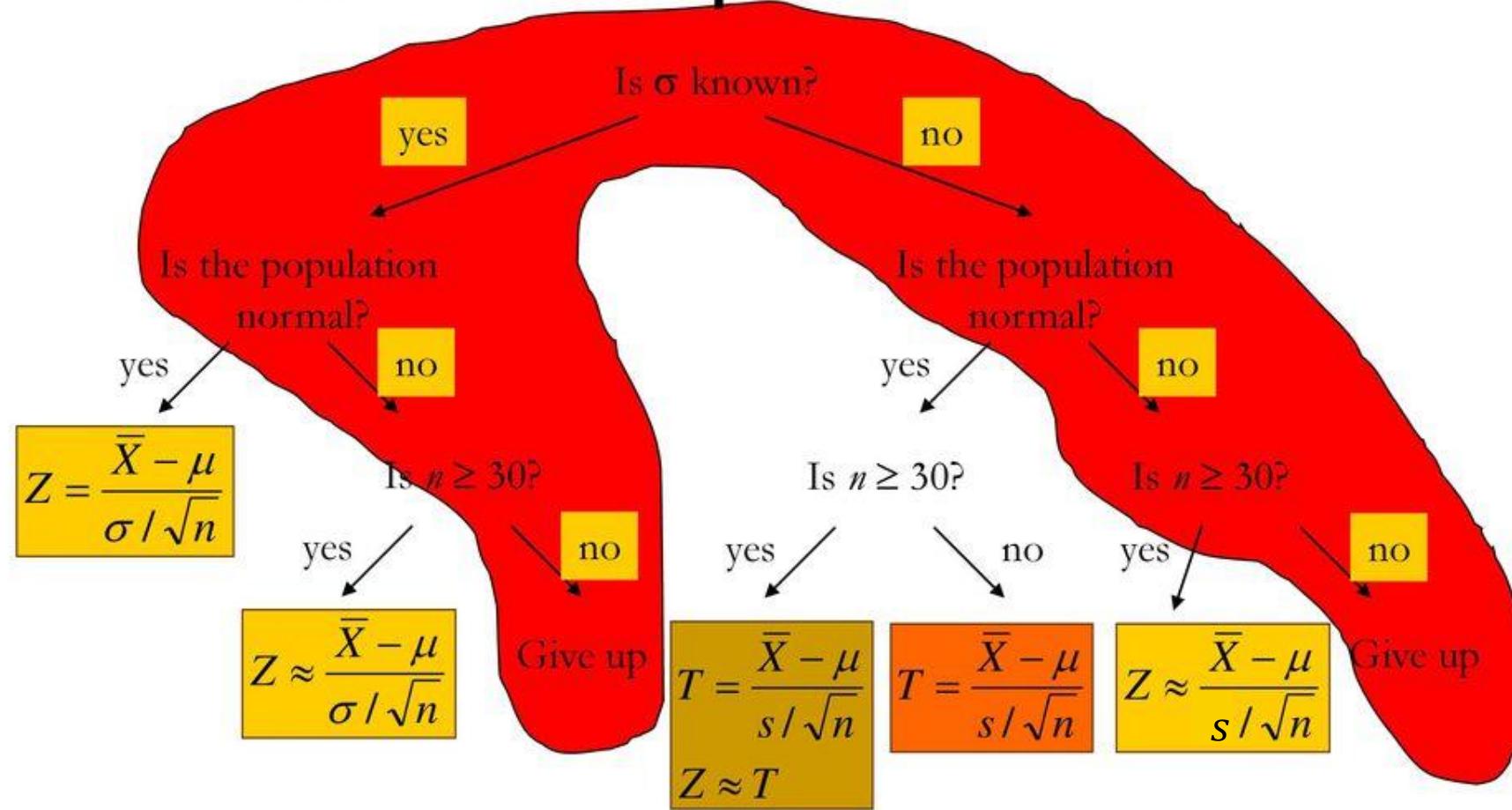


When to Use Either Z-Test or T-Test



When to Give Up

out of scope



3. The F-distribution

F - Distribution

- $V \sim \chi^2_{k_1}$, $W \sim \chi^2_{k_2}$, V, W independent
- $F = \frac{V/k_1}{W/k_2} \rightarrow F$ distribution with (k_1, k_2) degrees of freedom
- Application ($\{X_i\}$ and $\{Y_j\}$ independent):

$$X_1, \dots, X_{n_1} \sim iid N(\mu_1, \sigma_1^2) \quad Y_1, \dots, Y_{n_2} \sim iid N(\mu_2, \sigma_2^2)$$

- $V = \frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ $W = \frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ V, W independent

$$F = \frac{V/k_1}{W/k_2} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

Compare Variance
of two Population

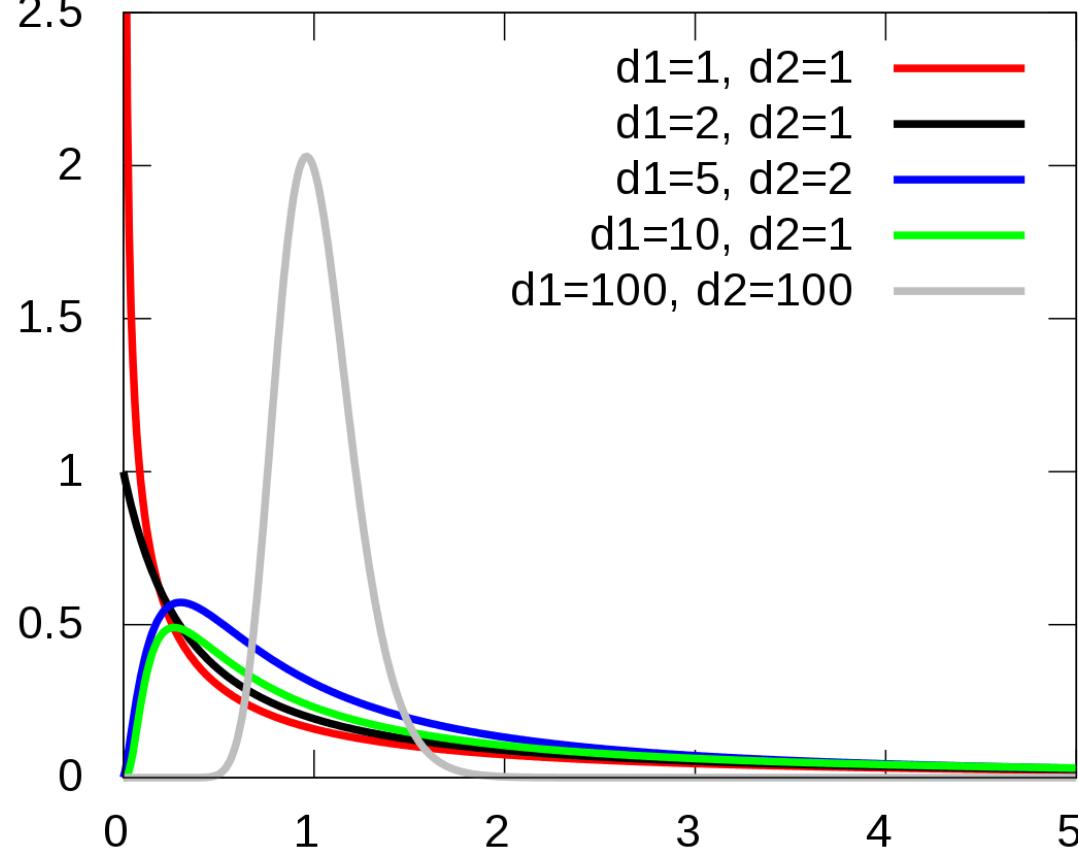
F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom

F - Distribution

- No -ve value
- Not symmetric

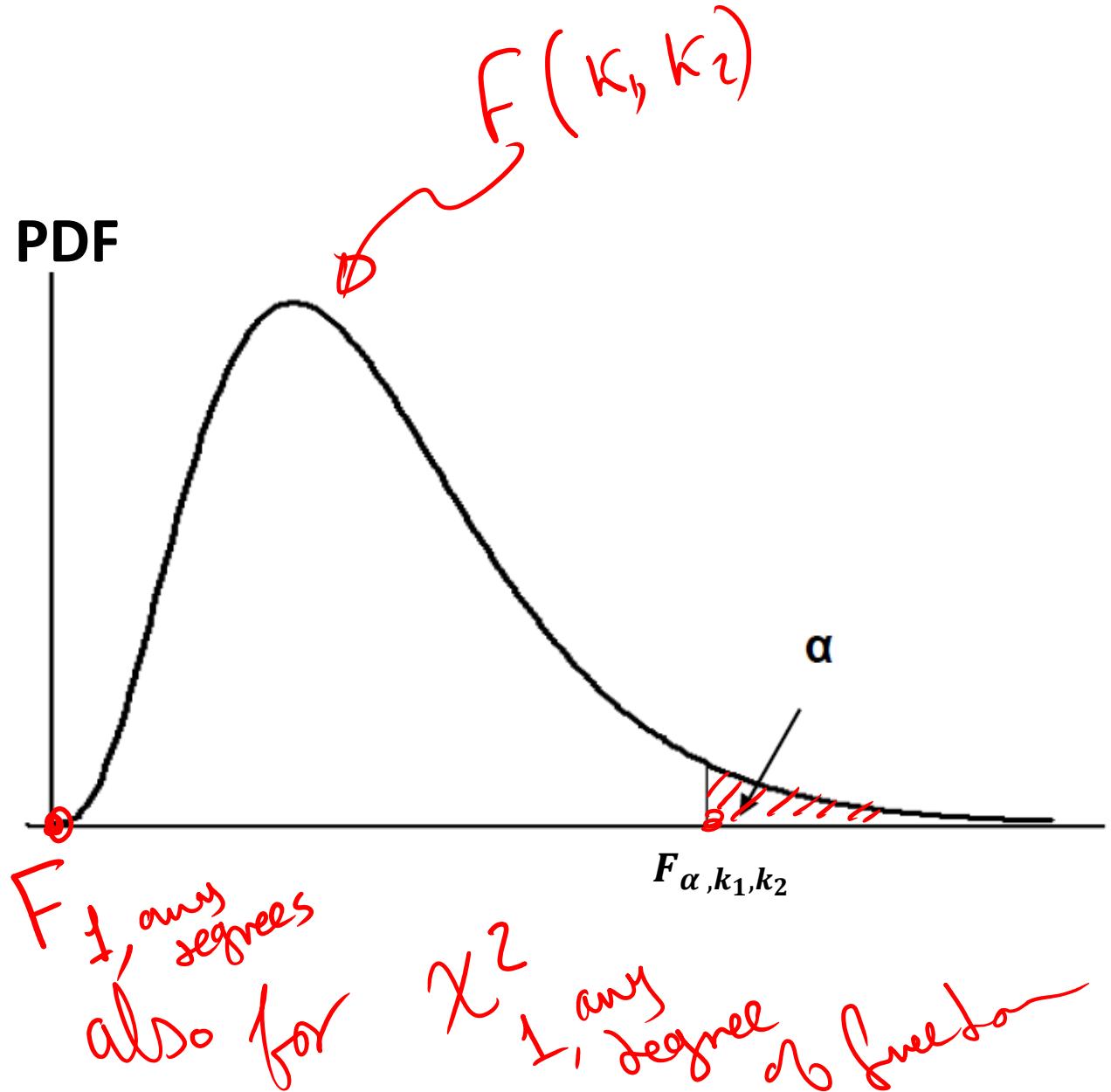
$\chi^2_{K_1}$ \rightarrow x_{val}
 χ^2_2 \rightarrow x_{val}
 $\chi^2_{K_2}$ \rightarrow x_{val}

PDF



F - Distribution

- F_{α, k_1, k_2}
- $P(U > F_{\alpha, k_1, k_2}) = \alpha$
- $F_{CDF}(F_{\alpha, k_1, k_2}) = 1 - \alpha$
- $F_{\alpha, k_1, k_2} = F_{CDF}^{-1}(1 - \alpha)$
- $F_{\alpha, k_1, k_2} = ?$
- Using MINITAB or R



The figure shows the Minitab software interface with the following details:

- Minitab - Untitled** window title.
- File Edit Data Calc Stat Graph Editor Tools** menu bar.
- Session** workspace on the left.
- Worksheet 1 ***** workspace on the left.
- Calc** menu open, showing options: Chi-Square..., Normal..., F..., t..., Uniform..., Binomial..., Geometric..., Negative Binomial..., Hypergeometric..., Discrete..., Integer..., Poisson..., Beta..., Cauchy..., Exponential..., Gamma..., Laplace..., Largest Extreme Value..., Logistic..., Loglogistic..., Lognormal..., Smallest Extreme Value..., Triangular..., Weibull... The **F...** option is highlighted.
- F Distribution** dialog box open:
 - Probability density
 - Cumulative probability
 - Inverse cumulative probability

Noncentrality parameter: 0.0

Numerator degrees of freedom: 6 (circled with red arrow)

Denominator degrees of freedom: 10 (circled with red arrow)

Input column: _____

Optional storage: _____

Input constant: 0.05 (circled with red arrow)

Optional storage: _____

OK Cancel
- Session** output window at the bottom:
 - 10/16/2022 2:26:00 PM
 - Welcome to Minitab, press F1 for help.
 - Inverse Cumulative Distribution Function
 - F distribution with 6 DF in numerator and 10 DF in denominator
 - P(X <= x) x
 - 0.05 0.246308 (circled with red bubble)

Handwritten annotations in red:

- A box around $F^{-1}(0.95, 6, 10)$ with arrows pointing to "Area after the pt" and "0.95".
- A box around $F^{-1}(0.05)$ with arrows pointing to "Area after the pt" and "0.05".
- Red arrows point from the circled "df's" in the dialog box to the "Numerator degrees of freedom" and "Denominator degrees of freedom" fields.
- Red arrows point from the circled "Input constant" to the "0.05" field in the dialog box.
- Red arrows point from the circled "0.246308" in the session output to the "x" value in the output table.

Ex: Consider the four independent random variables X, Y, U and V such that $X \sim N(0, 16)$, $Y \sim N(5, 4)$, $U \sim \chi^2(4)$ and $V \sim \chi^2(16)$.
 State the distribution of each of the following variables

$$\begin{aligned} E(X+2Y) &= E(X) + 2E(Y) = 0 + 2 \cdot 5 = 10 \\ V(X+2Y) &= V(X) + 4V(Y) = 16 + 4(4) = 32 \end{aligned}$$

$\rightarrow \sim N(10, 32)$

$$\begin{aligned} E(2X-Y) &= 2E(X) - E(Y) = 2(0) - (5) = -5 \\ V(2X-Y) &= 4V(X) + V(Y) = 4(16) + (4) = 68 \end{aligned}$$

$\sim N(-5, 68)$

A Linear Combination of Normally Distributed Random Variables

Let X_1, X_2, \dots, X_n be independent, normally distributed random variables with expected values $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectively and let a_1, a_2, \dots, a_n be constants.

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

$$\begin{aligned} E(ax+by+c) &= aE(X) + bE(Y) + c \\ V(ax+by+c) &= a^2 V(X) + b^2 V(Y) \end{aligned}$$

Ex: Consider the four independent random variables X, Y, U and V such that $X \sim N(0,16)$, $Y \sim N(5,4)$, $U \sim \chi^2(4)$ and $V \sim \chi^2(16)$.
 State the distribution of each of the following variables

*Sum of the squares of
indep. Std normal*

$$\frac{X^2}{16} + \frac{(Y-5)^2}{4}$$

$\frac{X^2}{16}$ \rightarrow Z_1^2
 $\frac{(Y-5)^2}{4}$ \rightarrow Z_2^2 ($Z_1^2 + Z_2^2$)
 $\frac{Y-5}{2}$ \rightarrow Z_2

$\frac{X-0}{4}$ \rightarrow Std normal

$\sim \chi^2(2)$
 $\delta_{\chi^2(2)} = t(16)$

$$\frac{X}{\sqrt{V}}$$

$\frac{X-0}{\sqrt{V/16}} = \frac{X/4}{\sqrt{V/16}} = \frac{X}{\sqrt{V}}$
 \rightarrow should be divided by its degree of freedom

\leftarrow should be Std. normal

$$\frac{4U}{V}$$

$$\frac{U/4}{\sqrt{V/16}} \sim F(4, 16)$$

$$\frac{4U}{\sqrt{V}} \sim F(4, 16)$$