02-24-00201
Probability and Statistics II

DR. AHMED TAYEL

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

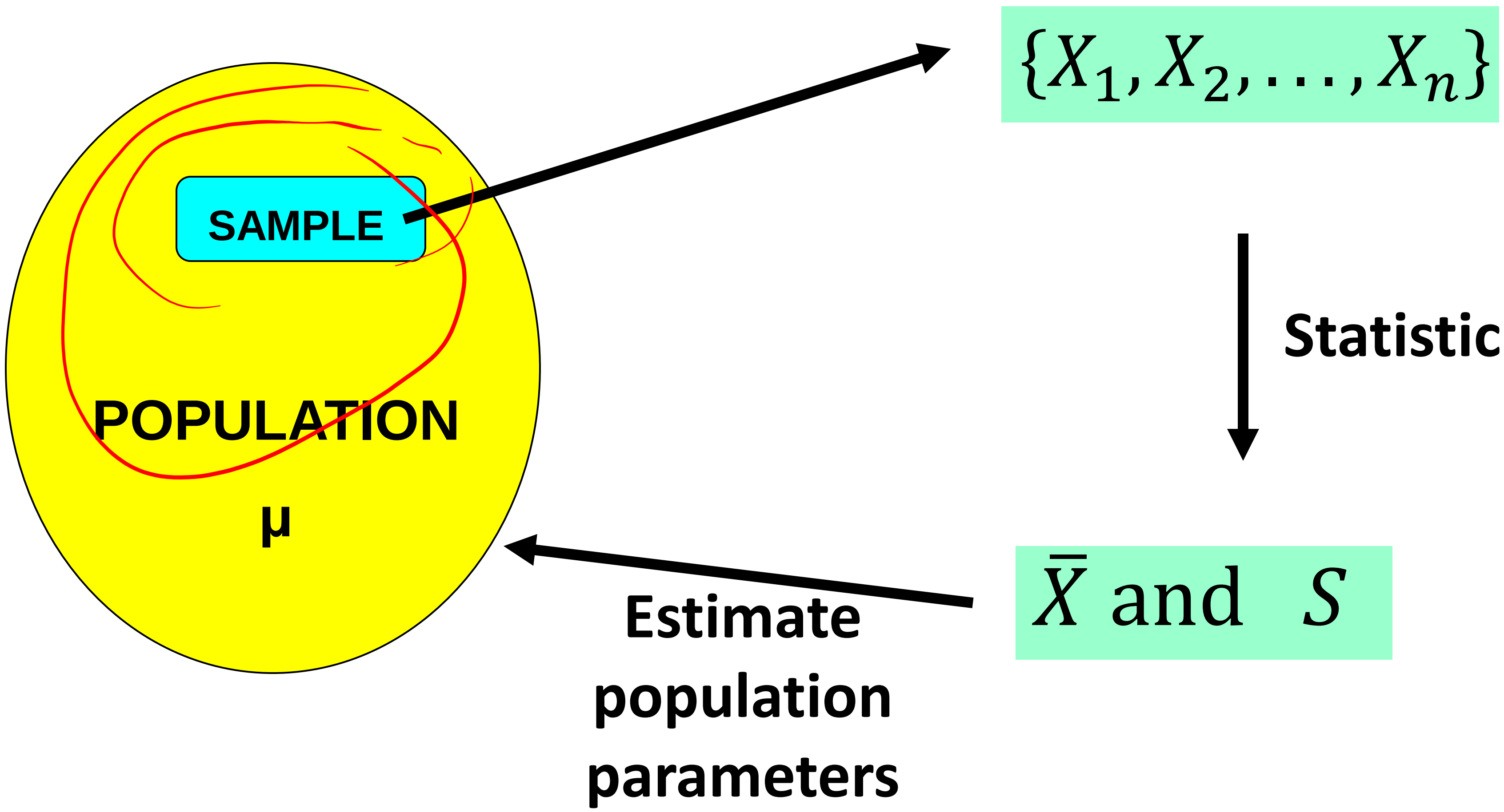
ahmed.tayel@alexu.edu.eg

Outline

1. Review.
2. Case of two populations.
3. Distributions derived from the Normal distribution.
    1. The Chi-squared distribution **[This lecture]**.
        1. Distribution of $\boldsymbol{s^2}$.
    2. The t-distribution **[Next lecture]**.
    3. The F-distribution **[Next lecture]**.

1. Review

The Sampling Process



Central Tendency in the Sample

Definition:

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

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

Variability in the Sample

Definition:

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

Estimate to σ^2

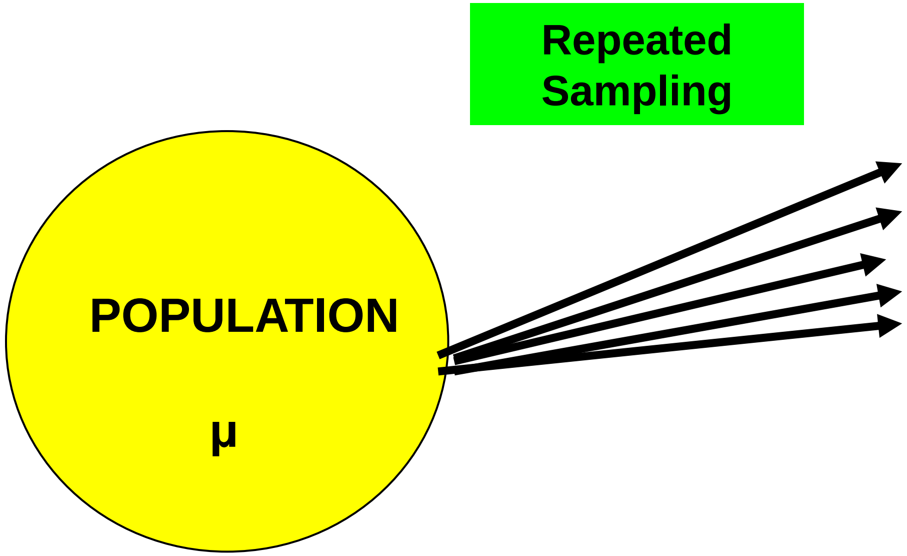
$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \ (\text{unit})^2$$

Definition:

The sample standard deviation is defined to be the statistic:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \quad \text{(unit)}$$

The Sampling Distribution

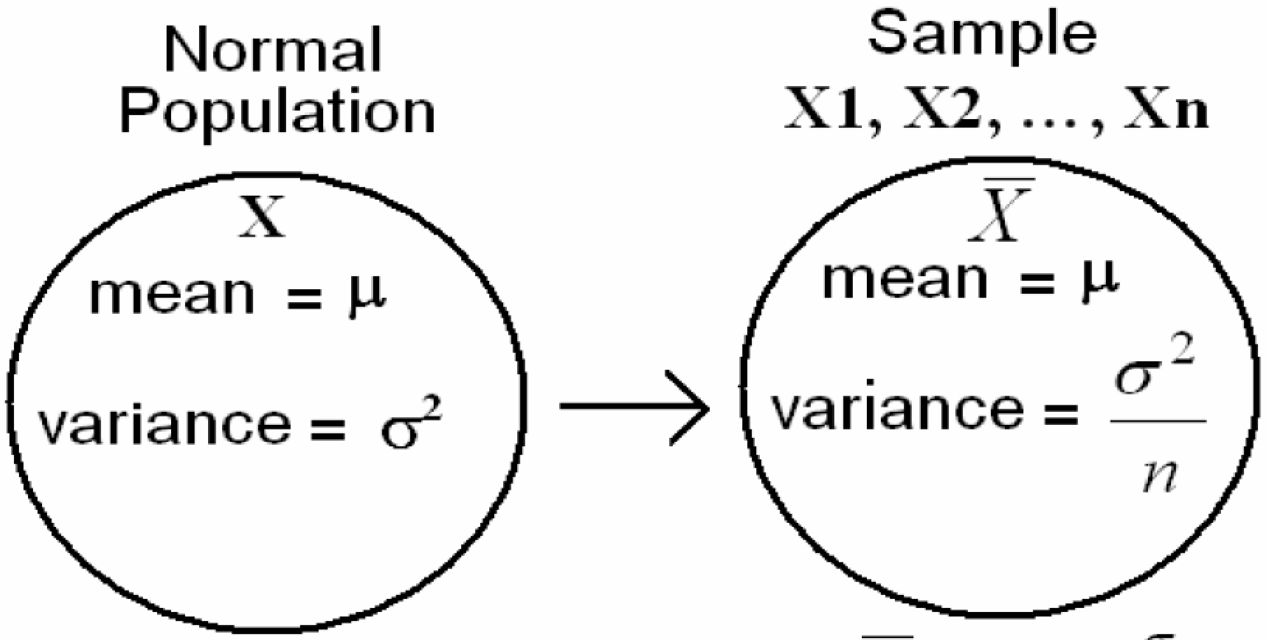


$\bar{X}$ is also a random variable

$$\left\{ \bar{x}_1,\ \bar{x}_2,\ \bar{x}_3,\ \dots,\ \bar{x}_N \right\}$$

Sampling Distribution

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



$X \sim N(\mu,\sigma)$

$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow$

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

What if the population is not normally distributed?

Theorem: (Central Limit Theorem)

If X_1, X_2, …, 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

$\boxed{n \geq 30}$

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

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 $X_1, X_2, \ldots, X_n$.
We have **three cases** for the distribution of $\bar{X}$

[1]

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

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

[2]

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

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

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

[3]

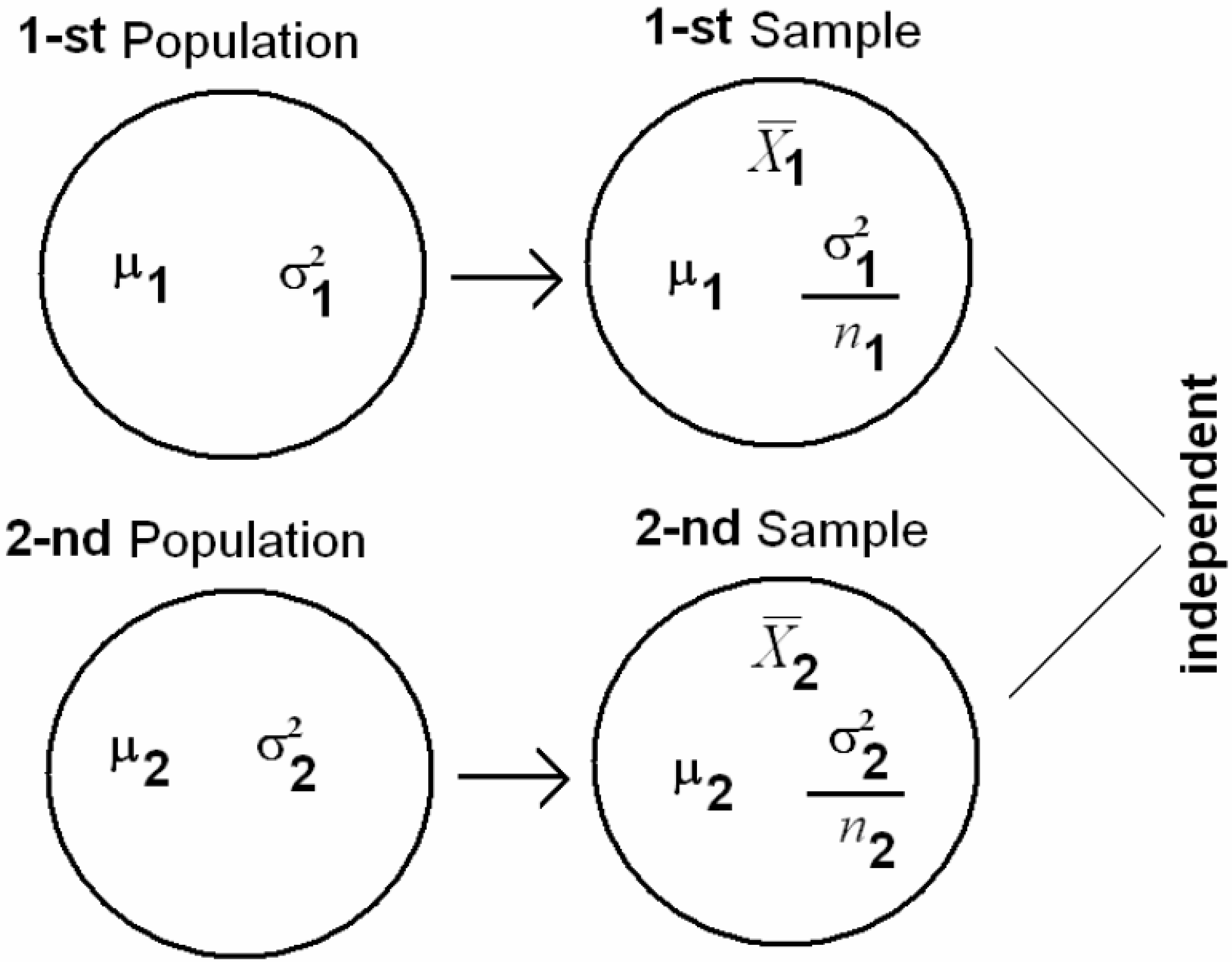
- $n < 30$
- σ^2 is unknown
- Next lecture

2. Case of two populations

Sampling Distribution of the Difference between Two Means

Suppose that we have **two** populations:

- 1-st population with mean μ_1 and variance $\sigma^2{}_1$
- 2-nd population with mean μ_2 and variance $\sigma^2{}_2$
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about $\mu_1 - \mu_2$.
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\bar{X}_1$ be the sample mean of the 1-st sample.
- Let $\bar{X}_2$ be the sample mean of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.


1-st Population
μ_1 σ_1^2
1-st Sample
$\bar{X}_1$
μ_1 $\frac{\sigma_1^2}{n_1}$
2-nd Population
μ_2 σ_2^2
2-nd Sample
$\bar{X}_2$
μ_2 $\frac{\sigma_2^2}{n_2}$
independent

Case of two populations

$Var(ax+b)$
$= a^2 Var(x)$

$Var(X-Y)$
$= Var(x) + (-1)^2 Var(y)$
$= Var\,x + Var(y)$

Since $\overline{X_1}$ and $\overline{X_2}$ are independent

$$E(\overline{X_1} - \overline{X_2}) = E(\overline{X_1}) - E(\overline{X_2}) = \mu_1 - \mu_2$$

$$V(\overline{X_1} - \overline{X_2}) = V(\overline{X_1}) + V(\overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If $\overline{X_1} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\overline{X_2} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$

Then

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\Leftrightarrow$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

What if the population is not normally distributed?

Theorem

If n_1 and n_2 are large, then the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal with mean $\quad n_1 \geq 30 \;$ and $n_2 \geq 30$

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\bar{X}_1 - \bar{X}_2 \sim \mathrm{N}(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$\Leftrightarrow$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Case of σ_1^2 and σ_2^2 are unknown

If $n_1 \geq 30$ and $n_2 \geq 30$, the central limit theorem (CLT) is still valid.

If we replace σ_1^2 , σ_2^2 by s_1^2 , s_2^2

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

$$\Leftrightarrow$$

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Note:

$$\sigma_{\bar{X}_1-\bar{X}_2}=\sqrt{\sigma^2_{\bar{X}_1-\bar{X}_2}}=\sqrt{\frac{\sigma_1^2}{\underset{\bar{x}_1}{n_1}}+\frac{\sigma_2^2}{\underset{\bar{x}_2}{n_2}}}\neq\sqrt{\frac{\sigma_1^2}{n_1}}+\sqrt{\frac{\sigma_2^2}{n_2}}=\frac{\sigma_1}{\sqrt{n_1}}+\frac{\sigma_2}{\sqrt{n_2}}$$

$$\neq \sigma_1+\sigma_2$$

Example

The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 years and standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6 years and standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a random sample of 49 tubes from manufacturer *B* ?

A

μ_1

σ_1

μ_2

σ_2

B

n_1 ≥ 30

$P(\bar{X}_1 \geq \bar{X}_2 + 1)$

n_2 ≥ 30

Solution:

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

We need to find the probability that the mean lifetime of manufacturer A is at least 1 year more than the mean lifetime of manufacturer B which is $P(\bar{X}_1 \geq \bar{X}_2 + 1)$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

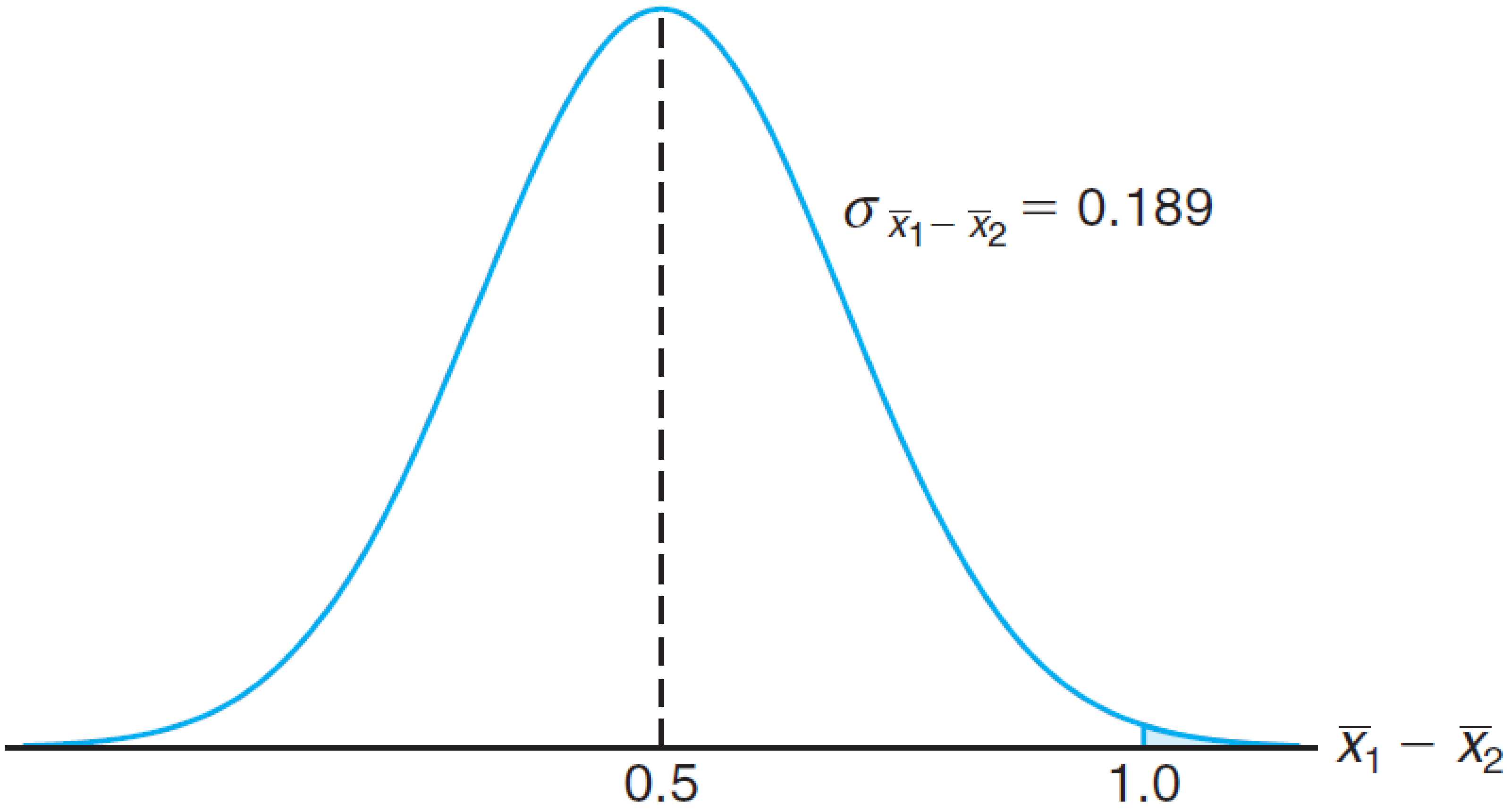
$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 6.5 - 6.0 = 0.5$$

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(0.9)^2}{36} + \frac{(0.8)^2}{49} = \boxed{0.03556}$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{0.03556} = 0.189$$

$$\bar{X}_1 - \bar{X}_2 \sim N(0.5, 0.189)$$


$\sigma_{\bar{x}_1 - \bar{x}_2} = 0.189$
$\bar{x}_1 - \bar{x}_2$
0.5
1.0

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

s.d.

σ

$$P(\overline{X}_1 \geq \overline{X}_2 + 1) = P(\overline{X}_1 - \overline{X}_2 \geq 1) \quad \frac{-0.5}{0.189} \sim N(0.5, 0.189)$$

$$\frac{-0.5}{0.189}$$

$$= P\left(\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{1 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

$$= P\left(Z \geq \frac{1 - 0.5}{0.189} \right)$$

$$= P(Z \geq 2.65)$$

$$= 1 - P(Z < 2.65)$$

$\Phi(2.65)$

$$= 1 - 0.9960$$

$$= 0.0040$$

3. Distributions derived from the Normal distribution

3.1 The Chi-squared distribution

Definition of Chi-squared distribution

independent & identically distributed

- If $Z_1, Z_2, \ldots, Z_k$ are i.i.d. $\sim N(0,1)$

Standard Normal #

- Let $U = Z_1^2 + Z_2^2 + \cdots + Z_k^2$
- Then $U \sim \chi_k^2$ "Chi-squared distribution with k degrees of freedom (df)"

Definition of Chi-squared distribution

- Ex: Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim N(\mu, \sigma^2)$
- Chi-squared are the sum of **standard** normal.
- $U = \underbrace{\left(\frac{X_1-\mu}{\sigma}\right)}_{z_1}^2 + \underbrace{\left(\frac{X_2-\mu}{\sigma}\right)}_{z_2}^2 + \cdots + \underbrace{\left(\frac{X_n-\mu}{\sigma}\right)}_{z_n}^2 \sim \chi_n^2$

Properties of Chi-squared distribution

- No –ve values of χ_k^2 "sum of square values"
- Not symmetric.
- For small df curve is skewed to the right.
- As df increases the shape becomes more symmetric → approximated to normal.

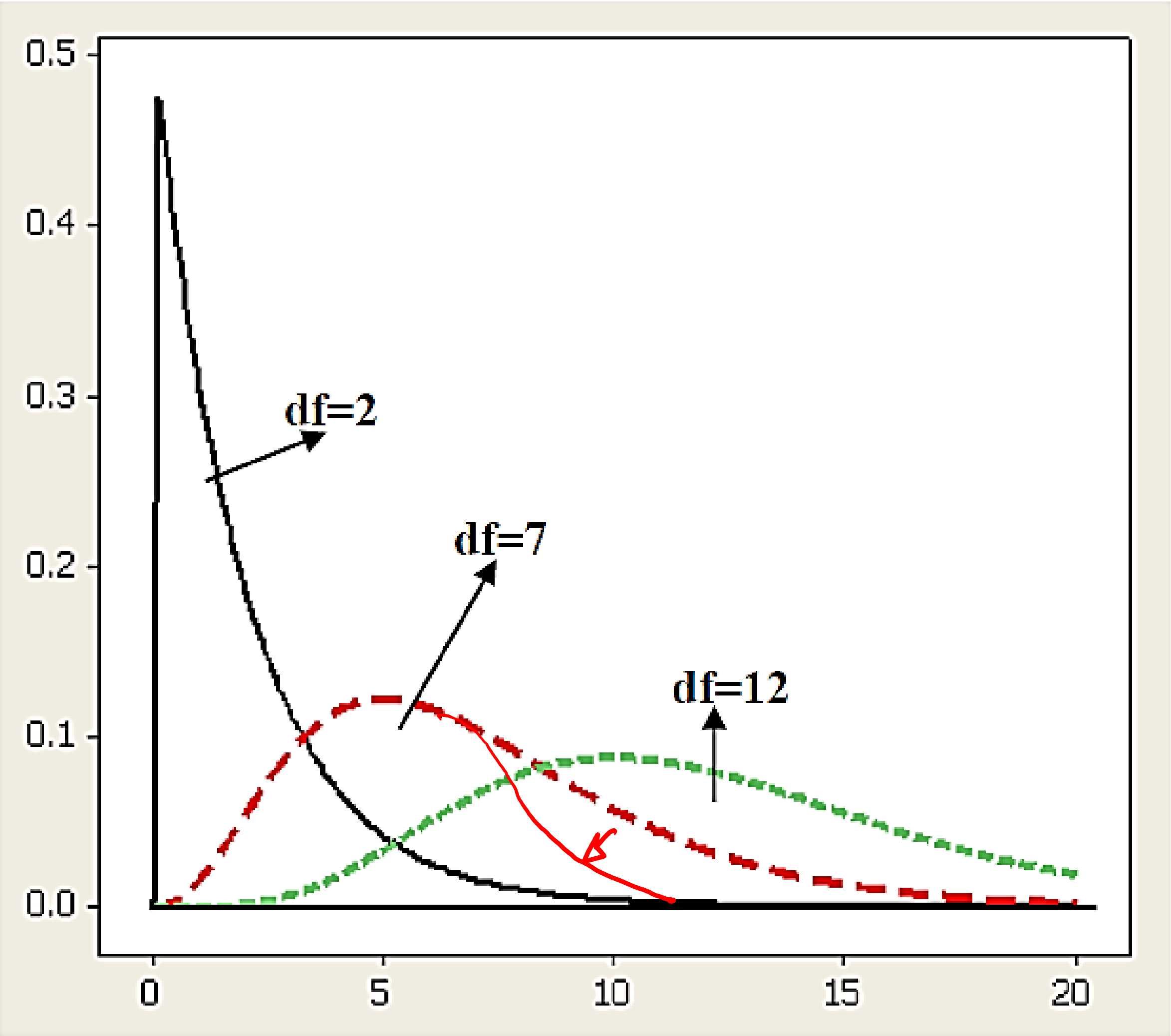


Fig. 1.1 χ_ν^2 distribution curves for various values of ν

Properties of Chi-squared distribution

$$\int_0^\infty x \, f_x(x)\, dx$$

- $E\left[\chi_k^2\right] = k$

- Proof not required

$$
\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^\infty x c x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx \\
&= c \int_0^\infty x^{n/2} \exp\left(-\frac{1}{2}x\right) dx \\
&= c\left\{\left[-x^{n/2} 2 \exp\left(-\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. + \int_0^\infty \frac{n}{2} x^{n/2-1} 2 \exp\left(-\frac{1}{2}x\right) dx\right\} \quad \text{(integrating by parts)} \\
&= c\left\{(0-0) + n \int_0^\infty x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx\right\} \\
&= n \int_0^\infty c x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx \\
&= n \cancel{\int_0^\infty f_X(x)\,dx} \quad 1 \\
&= n \quad \text{(integral of a pdf over its support equals 1)}
\end{aligned}
$$

Properties of Chi-squared distribution

- $V[\chi_k^2] = 2\,k$

- Proof not required

$\int_{}^{\infty} x^2 f_X(x)\, dx$

$$
\begin{aligned}
&\mathrm{E}[X^2] \\
&= \int_0^\infty x^2 f_X(x)\,dx \\
&= \int_0^\infty x^2 c x^{n/2-1} \exp\left(-\frac{1}{2}x\right)dx \\
&= c\int_0^\infty x^{n/2+1} \exp\left(-\frac{1}{2}x\right)dx \\
&= c\left\{\left[-x^{n/2+1}2\exp\left(-\frac{1}{2}x\right)\right]_0^\infty\right. \\
&\quad \left.+\int_0^\infty \left(\frac{n}{2}+1\right)x^{n/2}2\exp\left(-\frac{1}{2}x\right)dx\right\} \qquad \text{(integrating by parts)} \\
&= c\left\{(0-0)+(n+2)\int_0^\infty x^{n/2}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= c(n+2)\left\{\int_0^\infty x^{n/2}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= c(n+2)\left\{\left[-x^{n/2}2\exp\left(-\frac{1}{2}x\right)\right]_0^\infty\right. \\
&\quad \left.+\int_0^\infty \frac{n}{2}x^{n/2-1}2\exp\left(-\frac{1}{2}x\right)dx\right\} \qquad \text{(integrating by parts)} \\
&= c(n+2)\left\{(0-0)+n\int_0^\infty x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= (n+2)n\int_0^\infty c x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= (n+2)n\int_0^\infty f_X(x)\,dx \\
&= (n+2)n \qquad \text{(integral of a pdf over its support equals 1)} \\
&\mathrm{E}[X]^2 = n^2 \\
&\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\
&= (n+2)n - n^2 = n(n+2-n) = 2n
\end{aligned}
$$

Properties of Chi-squared distribution

Area under curve after χ^2 is α

- $\chi^2_{\alpha,k}$
- $P\left(U > \chi^2_{\alpha,k}\right) = \alpha$
- $F_{CDF}\left(\chi^2_{\alpha,k}\right) = 1 - \alpha$
- $\chi^2_{\alpha,k} = F^{-1}_{CDF}(1 - \alpha)$

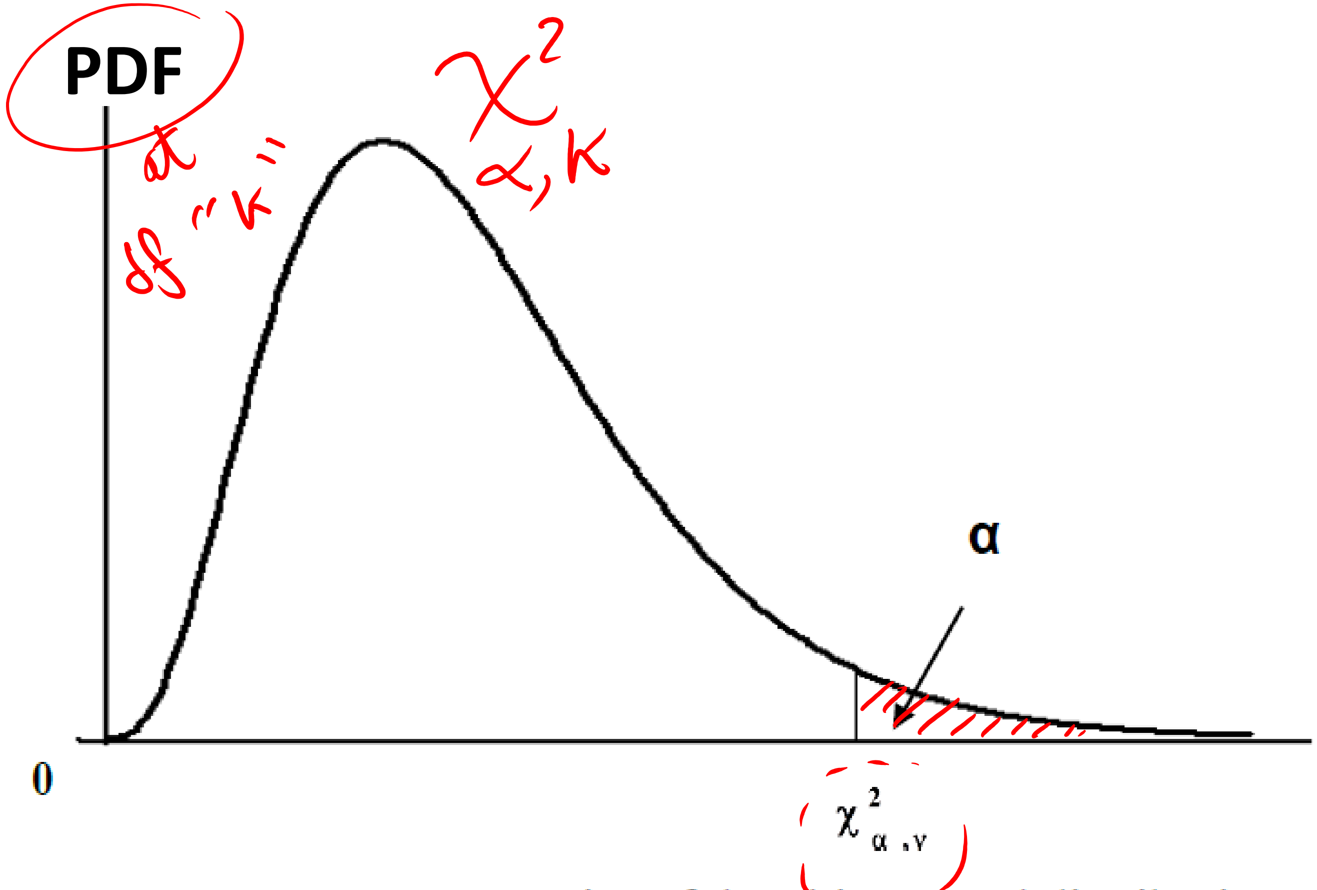


Fig. 1.2 Percentages point of the chi-squared distribution

Properties of Chi-squared distribution

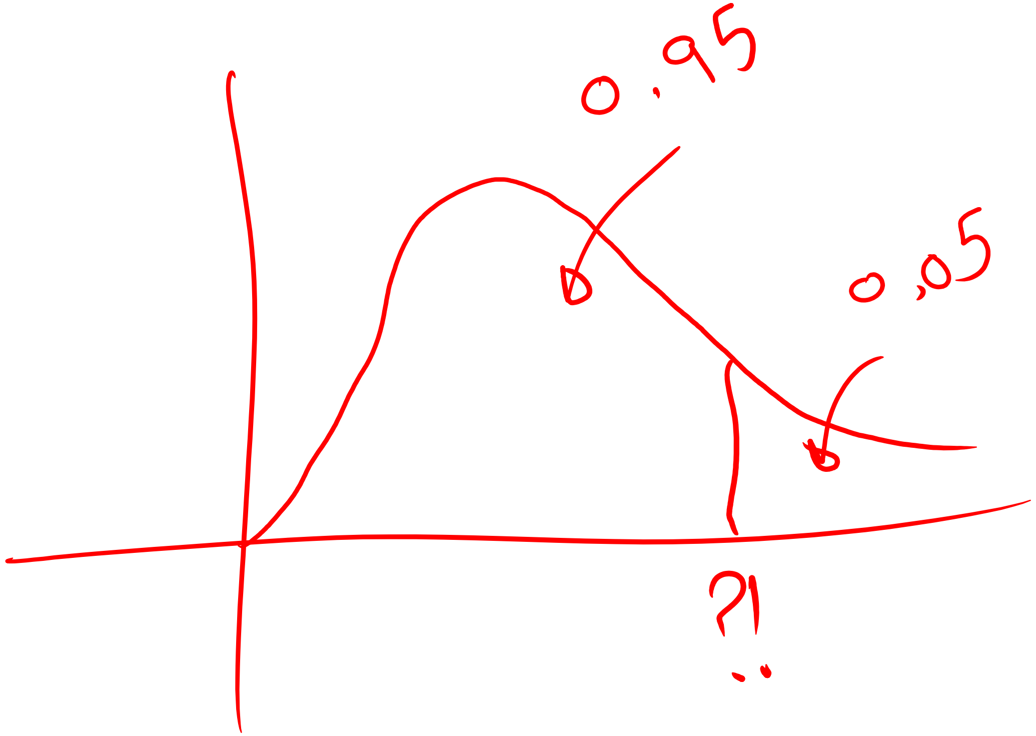
α : area after the pt.

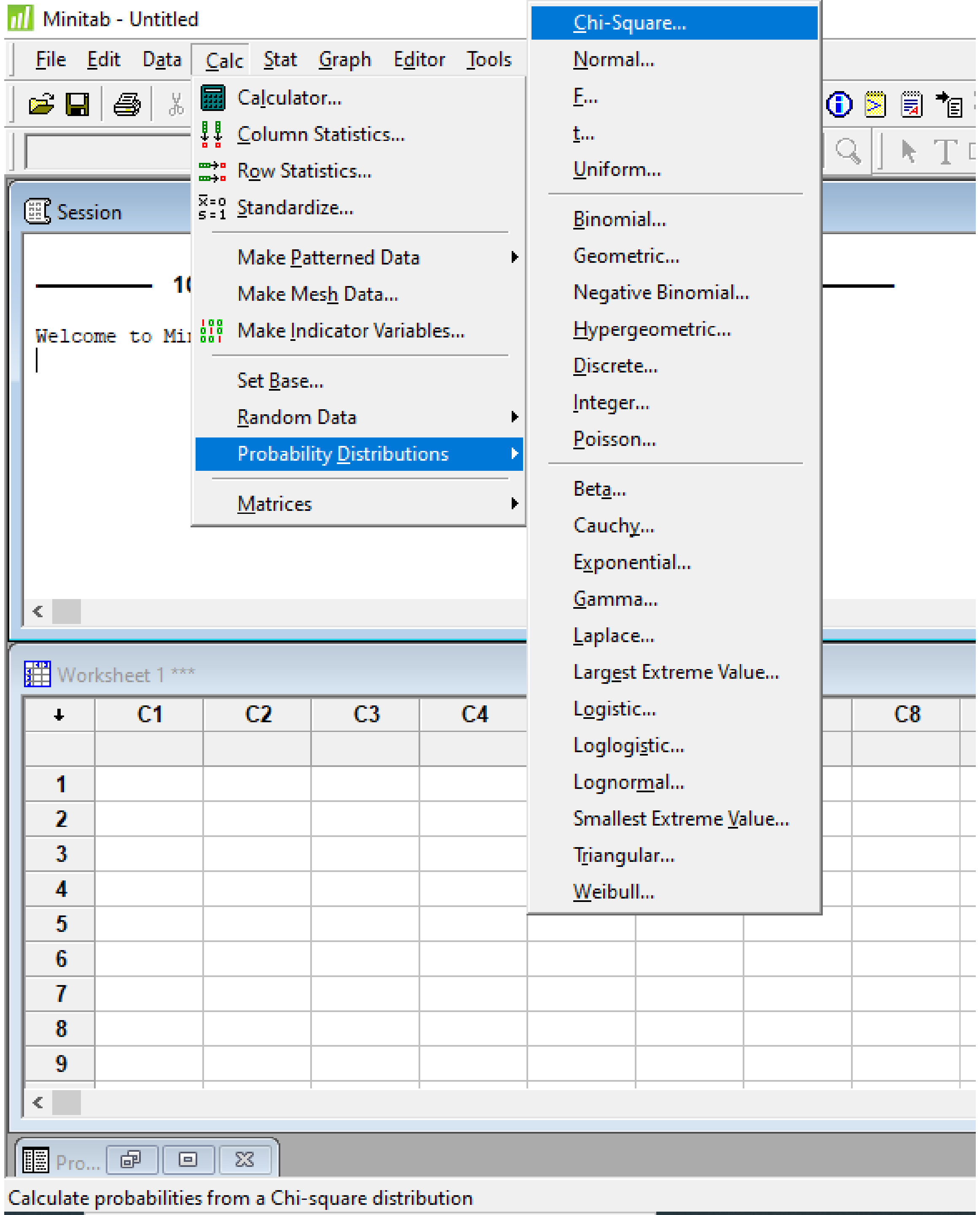
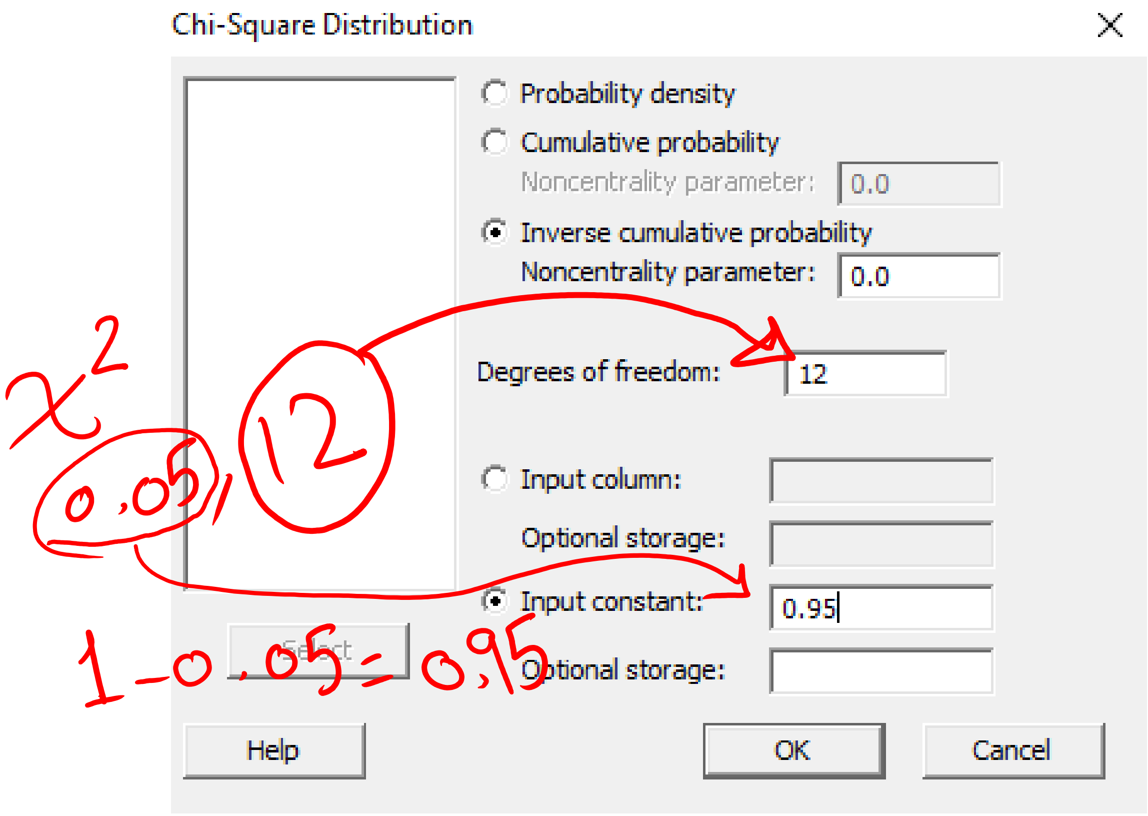
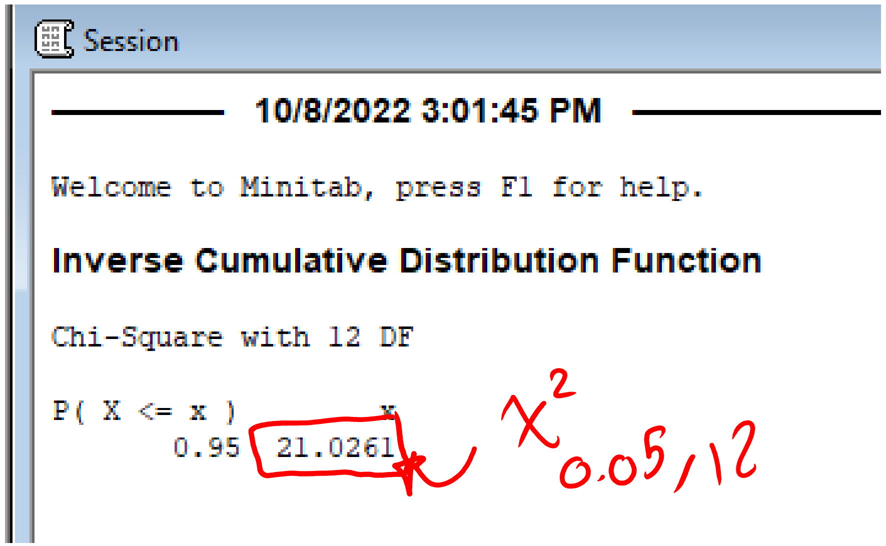
df

- Ex: $\chi^2_{0.05\,,\,12}$

CDF

- $F\left(\chi^2_{0.05\,,\,12}\right) = ?$ 0.95
- $\chi^2_{0.05\,,\,12} = ?$
- Using MINITAB or R




Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools
Calculator...
Column Statistics...
Row Statistics...
Standardize...
Make Patterned Data
Make Mesh Data...
Make Indicator Variables...
Set Base...
Random Data
Probability Distributions
Matrices
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...
Session
Welcome to Mi
Worksheet 1 ***
C1 C2 C3 C4 C8
1 2 3 4 5 6 7 8 9
Calculate probabilities from a Chi-square distribution

Chi-Square Distribution
Probability density
Cumulative probability
Noncentrality parameter: 0.0
Inverse cumulative probability
Noncentrality parameter: 0.0
Degrees of freedom: 12
Input column:
Optional storage:
Input constant: 0.95
Optional storage:
Select
Help OK Cancel
χ^2
0,05, 12
1-0,05 = 0,95

Session
10/8/2022 3:01:45 PM
Welcome to Minitab, press F1 for help.
Inverse Cumulative Distribution Function
Chi-Square with 12 DF
P(X <= x) x
0.95 21.0261
$\chi^2_{0,05,12}$

Properties of Chi-squared distribution

- Let $U_1, U_2, \ldots, U_n$ be i.i.d. RVs having Chi-squared distribution with $k_1, k_2, \ldots, k_n$ dfs.
- Then $V = \sum_1^n U_i \sim \chi^2_{k_1+k_2+\cdots+k_n}$

Properties of Chi-squared distribution

- Ex: $U_1 \sim \chi_3^2 \rightarrow U_1 = Z_1^2 + Z_2^2 + Z_3^2$
- $U_2 \sim \chi_5^2 \rightarrow U_2 = Z_4^2 + Z_5^2 + Z_6^2 + Z_7^2 + Z_8^2$
- $V = U_1 + U_2$

  $= Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + Z_6^2 + Z_7^2 + Z_8^2$
- $V \sim \chi_8^2$

3.1.1 Distribution of s^2

Distribution of s^2

$$\sum_{i=1}^{n} 5 = 5+5+5+\dots+5 \text{ to } n = 5n$$

sample variance def

- $s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$
- $\rightarrow (n-1)\, s^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2$ [Add and subtract $\boldsymbol{\mu}$]
- $\rightarrow (n-1)\, s^2 = \sum_{i=1}^{n}\left((X_i-\mu)-(\bar{X}-\mu)\right)^2$
- $= \sum_{i=1}^{n}(X_i-\mu)^2 + \underbrace{\sum_{i=1}^{n}\underbrace{(\bar{X}-\mu)^2}_{\text{No } i}}_{= n(\bar{X}-\mu)^2} - \underbrace{2\sum_{i=1}^{n}\left((X_i-\mu)\underbrace{(\bar{X}-\mu)}_{\text{No } i}\right)}$

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$= -2\,(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)$$

$\Sigma x_i - \Sigma \mu$

$$= -2\,(\bar{X}-\mu)\left(\cancel{\sum_{i=1}^{n}} X_i - n\,\mu\right)$$

$n\bar{X}$

$$= -2\,n\,(\bar{X}-\mu)^2$$

$$= \Delta - n(\bar{X}-\mu)^2$$

Distribution of s^2

- $(n-1)\,s^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n\,(\bar{X}-\mu)^2$

[Divide both sides by $\boldsymbol{\sigma}^2$]

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

- $$\frac{(n-1)\,s^2}{\sigma^2} = \underbrace{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}}_{= \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2} - \underbrace{\frac{n\,(\bar{X}-\mu)^2}{\sigma^2}}_{= \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2}$$

$\sim N(\mu, \sigma^2)$

$N(0,1)$ $\qquad$ $\sim N(0,1)$

$\sim \chi^2_n \quad - \quad \sim \chi^2_1$

Assuming $\bar{X}$ & X_i's are indep.

- $$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Distribution of s^2

Recall:

$$E\left[\chi_k^2\right] = k$$

$$V\left[\chi_k^2\right] = 2k$$

$E(aX+b) = aE(X)+b$

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$Var(aX+b) = a^2 Var(X) + 0$

$$E\left(\frac{(n-1)\,s^2}{\sigma^2} + 0\right) = n-1$$

$$\frac{\cancel{(n-1)}}{\sigma^2} E\left(s^2\right) = \cancel{n-1}$$

$$\boxed{E\left(s^2\right) = \sigma^2}$$

$$V\left(\frac{(n-1)\,s^2}{\sigma^2}\right) = 2\,(n-1)$$

$$\frac{(n-1)^{\cancel{2}}}{\sigma^4} V\left(s^2\right) = 2\,\cancel{(n-1)}$$

$$\boxed{V\left(s^2\right) = \frac{2\sigma^4}{n-1}}$$