Warsaw Summer School 2023, OSU Study Abroad Program

Probability and the Normal Curve

Samples and Populations

Probability distribution

- **A probability distributtion is directly analogous to a frequency distribution, except that is based on probability theory. In a probablity distribution , we specify the possible values of a variable and the probabilities associated with each value.**

Probability distribution

For probability distribution we use μ for the mean, and σ for standard deviation (like for a population).

The idea of standard scores integrates our knowledge of central tendency (μ) and variability (σ).

Normal distribution

The normal distribution has been known by many different names: *the law of error*, *the law of facility of errors*, or *Gaussian law*.

<Carl Friedrich Gauss, 1794>

The name "normal distribution" was coined by Galton. The term was derived from the fact that this distribution was seen as typical, common, *normal*.

<Francis Galton, 1875>

Normal distribution

- **Normal distribution:**
  - **Unimodal**
  - **Mesokurtosic (a moderate peakedness)**
  - **Symetric (zero skewness)**

Normal distribution

Why is the normal distribution important in science? There are two main reasons:

(1) It provides a tool for analyzing data (in descriptive statistics)

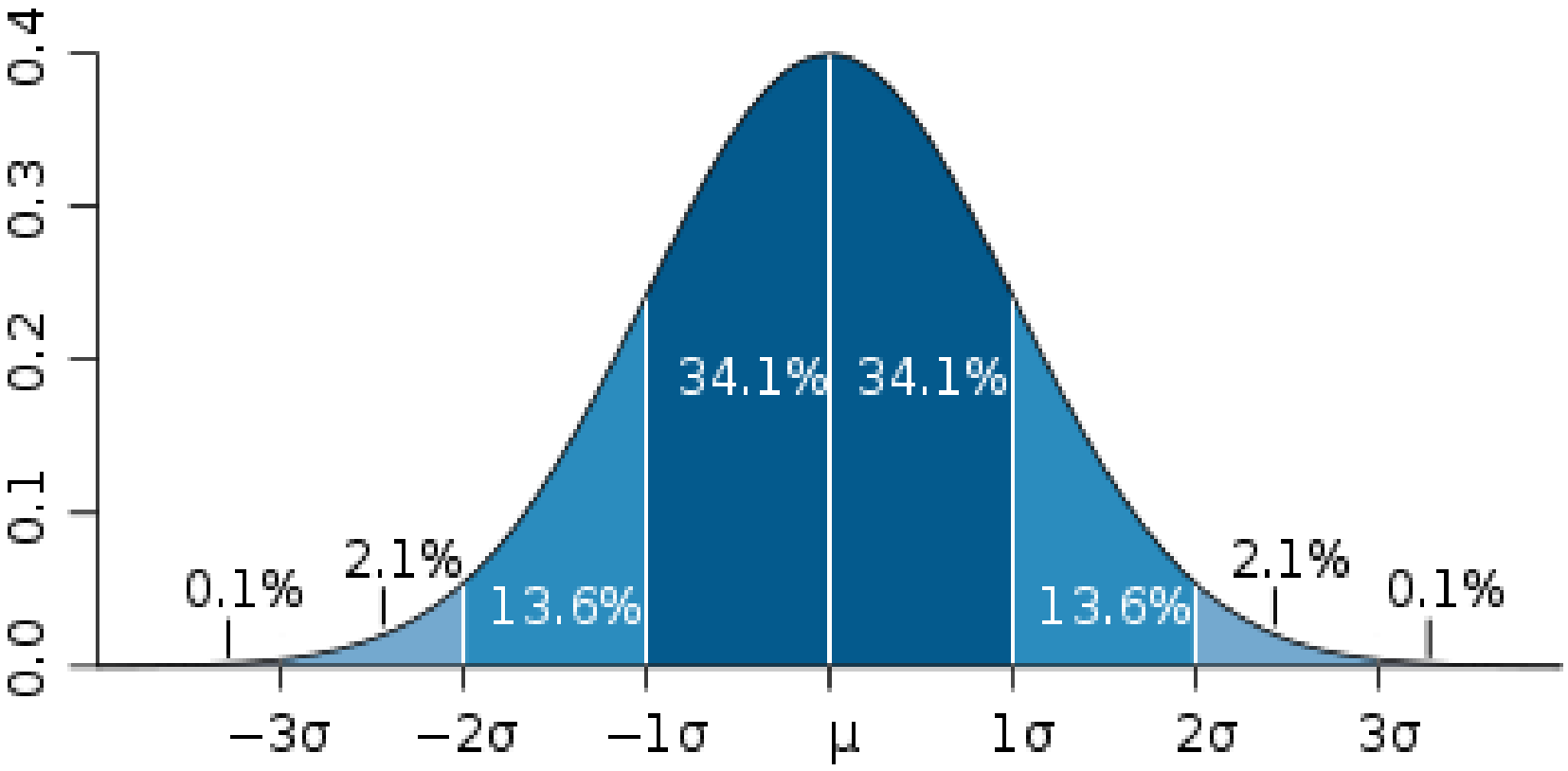
(2) It provides a tool for deciding about errors that we might commit in testing hypotheses (in inferential statistics).

Normal distribution

Every normal curve (regardless of its mean or standard deviation) conforms to the following rule:

- **About 68% of the area under the curve falls within 1 standard deviation of the mean.**
- **About 95% of the area under the curve falls within 2 standard deviations of the mean.**
- **About 99.7% of the area under the curve falls within 3 standard deviations of the mean.**

Normal distribution



Normal distribution

If we know that a given metric variable is normally distributed, we know also a lot about the place of particular values in this distribution.

Let assume that we measure IQ of the large population and we obtain the distribution that it is unimodal, symmetric, mesokurtic.

The results are that the mean value = 100, and the standard deviation = 15

Normal distribution

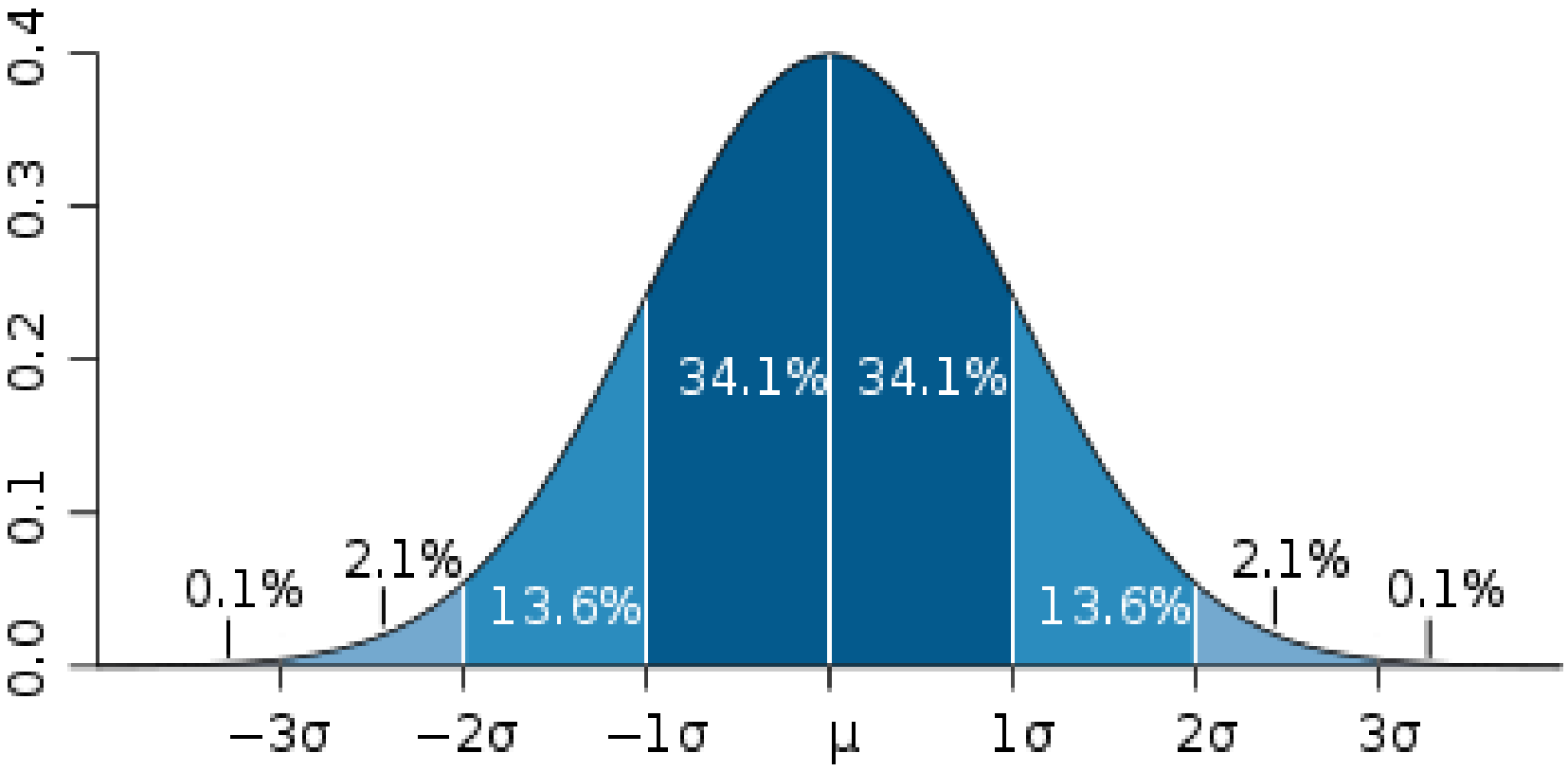
How far apart are two people A and B, where A = 115, and B = 99.

The difference, 16 points, tell us only little about the meaning of this result. The important information is how many people (in terms of percentage of all) have the values between 99 and 115

Note: 99 is close to the mean (mean =100) and 115 is one standard deviation (sd = 15) above the mean.

Let's look at the graph.

Normal distribution



Normal distribution

Consider a person C who also differ from B by one standard deviation but in plus, C = 130. What percentage of people is between them?

For a given normal distribution, for any pair of people K, L, we can say what percentage of people is between them, i.e., has values of the variable $X_K > X_M < X_L$, for $X_K < X_L$.

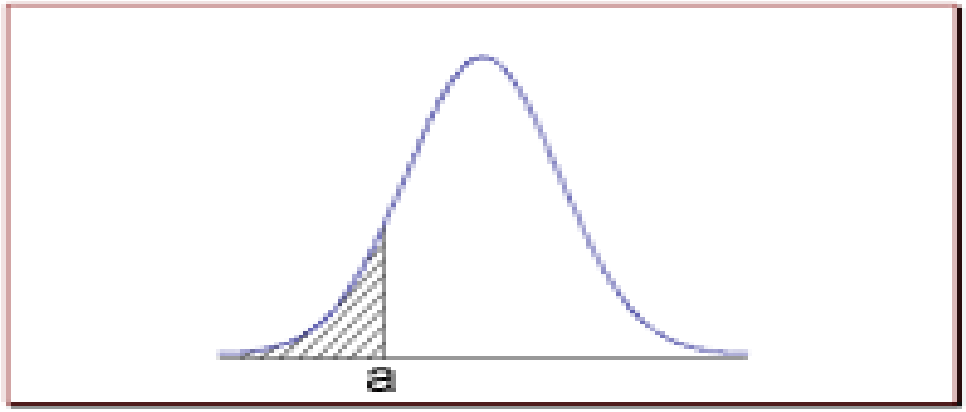
Probability

The second use of the normal curve deals with deciding about errors that we might commit in testing hypotheses.

It is about probabilities of committing errors.

Two important properties:

The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity (non-shaded area).



The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity (shaded area).

Probability distribution

<u>Standardized normal probability distribution</u> is expressed by z-scores:

$$z = (X_i - \mu) / \sigma$$

Appendix B shows the proportion of the area above and below the z-score.

- Column A = z-score

- Column B = area between mean and z (proportion)

- Column C = area beyond z (proportion)

Note:

Column B + C = .5000

A	B	C
.00	.00	50.00
.50	19.15	30.85
1.00	34.13	15.87
1.65	45.05	4.95
1.96	47.50	2.50
2.00	47.72	2.28
2.57	49.49	.51
3.00	49.87	.13

Population

A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population

Populations and samples

Sample frame (or sampling space): a group of data points that represent all possible outcomes; the registry from which we draw a sample.

From a statistical point of view, any population could be a sample frame (or sampling space).

The differences between “population” and “sampling frame”

An ideal sampling frame. Qualities

- **all units have a logical, numerical identifier**
- **all units can be found - their contact information is present**
- **the frame is organized in a logical, systematic fashion**
- **the frame has additional information about the units that allow the use of more advanced sampling frames**
- **every element of the population of interest is present in the frame**
- **every element of the population is present *only once* in the frame**
- **no elements from outside the population of interest are present in the frame**

Populations and sampling frames

Population and census

Sample frames are usually constructed that they provide a material for drawing a sample or samples.

E.g. population = adult men and women, aged 21-65 in Columbus area, estimated at 500,000

Sampling framework = a registry of people obtained by door-to-door survey on the randomly chosen streets, N=80,000

Population

In statistics, populations are treated as:

- **- large**
- **- unobtainable**
- **- hypothetical**

Then, we need samples, drawn from the sampling frames.

Sample

A sample: subgroup or a part of the population -- subset of the sampling frame.

Random sampling: process of selecting a sample such that the selection of one observation is independent of the selection of any other observation.

Random sample

A good random sample must satisfy two requirements:

(1) For a given draw, each individual in the sample space (population) must have an equal chance of being selected.

(2) For a number of consecutive draws, there must be constant probability for each individual in the sample space (population) to be selected.

Sampling and sampling distribution

In each sample we obtain specific sample characteristics (statistics), such as the mean and SD

Theoretically, we can have many samples. Therefore, we could have many sample statistics.

The concept of sampling distribution.

Sampling distribution

A sampling distribution is a distribution of statistics obtained by selecting all the possible samples of a specific size from a population. It is a theoretical distribution showing the frequency of occurrence of values of some statistics computed for all possible samples of size N sampled with replacement from a given sample space (population).

- **Components of this definition in reverse order:**

Sampling distributions

(1) Sampling distributions are based on all possible samples of fixed size.

(2) Sampling distributions do not convey information about the scores of individual cases; they deal with measures computed for samples.

(3) Sampling distributions are frequency (proportion) distributions. They depict how often various values of some statistic occur.

(4) Sampling distributions are theoretical distributions in the sense that we normally never see them.

The sampling distribution of the mean

- **The sampling distribution of the mean is a distribution of sample means.**

- **The Central Limit Theorem essentially implies that the mean of the sampling distribution of the mean equals the mean of the population and that the standard error of the mean equals the standard deviation of the population divided by the square root of N as the sample size gets infinitely larger. In addition, the sampling distribution of the mean will approach a normal distribution.**

The importance of the central limit theorem

- **The importance of the central limit theorem to statistical thinking cannot be overstated. Most of hypothesis testing and sampling theory is based on this theorem. In addition, it provides a justification for using the normal curve as a model for many naturally occurring phenomena.**
- **If a trait, such as intelligence, can be thought of as a combination of relatively independent events, in this case both genetic and environmental, then it would be expected that trait would be normally distributed in a population.**

Sampling Distribution of Means

The Distribution of the Population

Mean

$$\mu = (\Sigma X_i)/ N$$

Standard deviation

$$\sigma = \sqrt{\Sigma (X - \tilde{X})^2 / N}$$

Sampling Distribution of Means

The sampling distribution of means is a frequency of means for all samples

Characteristics of a Sampling Distribution of Means

(A) Shape. The distribution of sample means will be normal if either one of the following two conditions is satisfied:

- **- the population from which the samples are selected is normal,**
- **- the size of the sample is relatively large (N=30).**

(B) Central Tendency. The mean of the distribution of sample means will be identical to the mean of the population.

(C) Variability. The standard deviation of a sampling distribution of means is smaller than the standard deviation of the population.

Standardized Normal Curve

The Sampling Distribution of Means as a Standardized Normal Curve

- **z-scores for mean values**
- $z = (\tilde{\tilde{X}} - \mu) / \sigma x$

Standard Error of the Mean

Standard error of the mean = standard deviation of the mean.

- $\sigma_x = \sigma / \sqrt{N}$

Confidence intervals

Confidence limits are defined as follows:

- **CL = $\mu - z\,\sigma_x$ and $\mu + z\,\sigma_x$**

where z is a z-score for the chosen probability level (95%)

- **σ_x - is the standard error of μ.**

The more confident you want to be that the limits include μ, the larger the z value. In practice we take:

- **z = 1.96 for 95% of confidence**
- **z = 2.58 for 99% of confidence**

Level of confidence

- **$\alpha = 1 -$ level of confidence**
- **For 95% level of confidence (0.95) $\alpha = 0.05$**
- **For 99% level of confidence (0,99) $\alpha = 0.01$**
- **$p \leq 0.05$**
- **$p \leq 0.01$**

Example

Example

- $\tilde{\tilde{X}} = 4$
- $s = 2$
- $n = 100$

- $\sigma_x = \sigma / \sqrt{N}$ thus, $2/\sqrt{100} = 2/10 = .20$
- CL95 = $4 - (1.96)(.20)$ & $4 + (1.96)(.20)$

  3.61 4.39

Sample proportions

Standard error of the sample proportions:

- $\sigma_p = \sqrt{pq/N}$ where $q = 1 - p$
- $CL = p - \sigma_p Z$ and $p + \sigma_p Z$

Example

Example

- **N = 50**
- **p = .40**
- **q = .60**

- $\sigma_p = \sqrt{(.40)(.60)/\ 50} = .069$
- **CL95= .40 - 1.96(.069) and .40 + 1.96(.069)**
  1.96(.069) =.14

95% CL for .40 and n = 50 is

- **from .26 to .54**

But for n = 5,000 95% CL for .40 are from 39 to 41

$\bar{x}$

Standard error of Mean(I) - Mean(II)

Two possibilities:

A1 Standard error for $\bar{X} - \mu$ with known σ

$$\sigma_x = \sigma / \sqrt{n}$$

A2 Standard error for $\bar{X} - \mu$ without known σ

$$s_x = s / \sqrt{n - 1}$$

$\bar{x}$

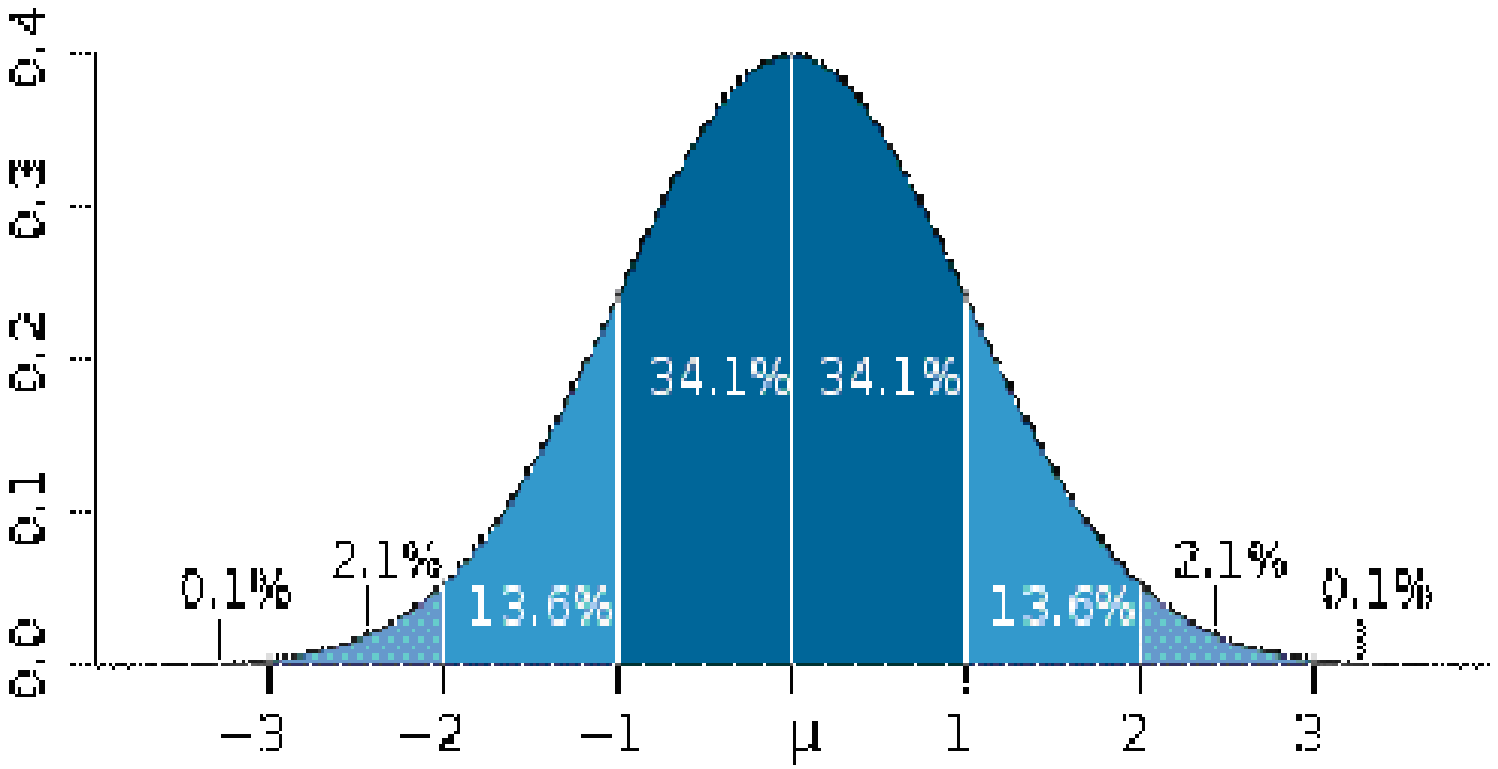
Standard error of Mean(I) - Mean(II)

A1 Standard error for $\bar{X} - \mu$ with known σ

$\sigma_x = \sigma / \sqrt{n}$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$$

z-test distribution



Example

$\mu = 40 \quad \sigma = 5 \qquad \bar{X} = 42 \qquad N = 100$

1. Assumptions

2. State your hypotheses: H0: $\bar{X} = \mu$, H1: $\bar{X} \neq \mu$ (H1: $\bar{X} \neq 40$)
3. Sampling distribution & the test statistic (here z-test)
4. Alpha level (95%)
5. Test statistic

(Is the difference 42 - 40 simply due to sampling error).

$$\text{z-test} = (42 - 40) / (5/\sqrt{100}) = 2 / .5 = 4$$

6. Decision & interpret results

σ unknown

This example refers to a one-sample z-test: We compare one sample to the population information given, to see if the sample mean (x) is different enough from the population mean (μ) to say that the sample is distinct from the population.

Often the population standard deviation (σ) is not provided. Then, we cannot use the z-test because we do not know that the sampling distribution is actually z-normal. All we have is sample information (x, s) a given population mean (μ).

$\bar{x}$

Standard error of Mean(I) - Mean(II)

A1 Standard error for $\bar{X} - \mu$ with known σ

$\sigma_x = \sigma / \sqrt{n}$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

A2 Standard error for $\bar{X} - \mu$ with unknown σ

$s_x = s / \sqrt{n-1}$

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

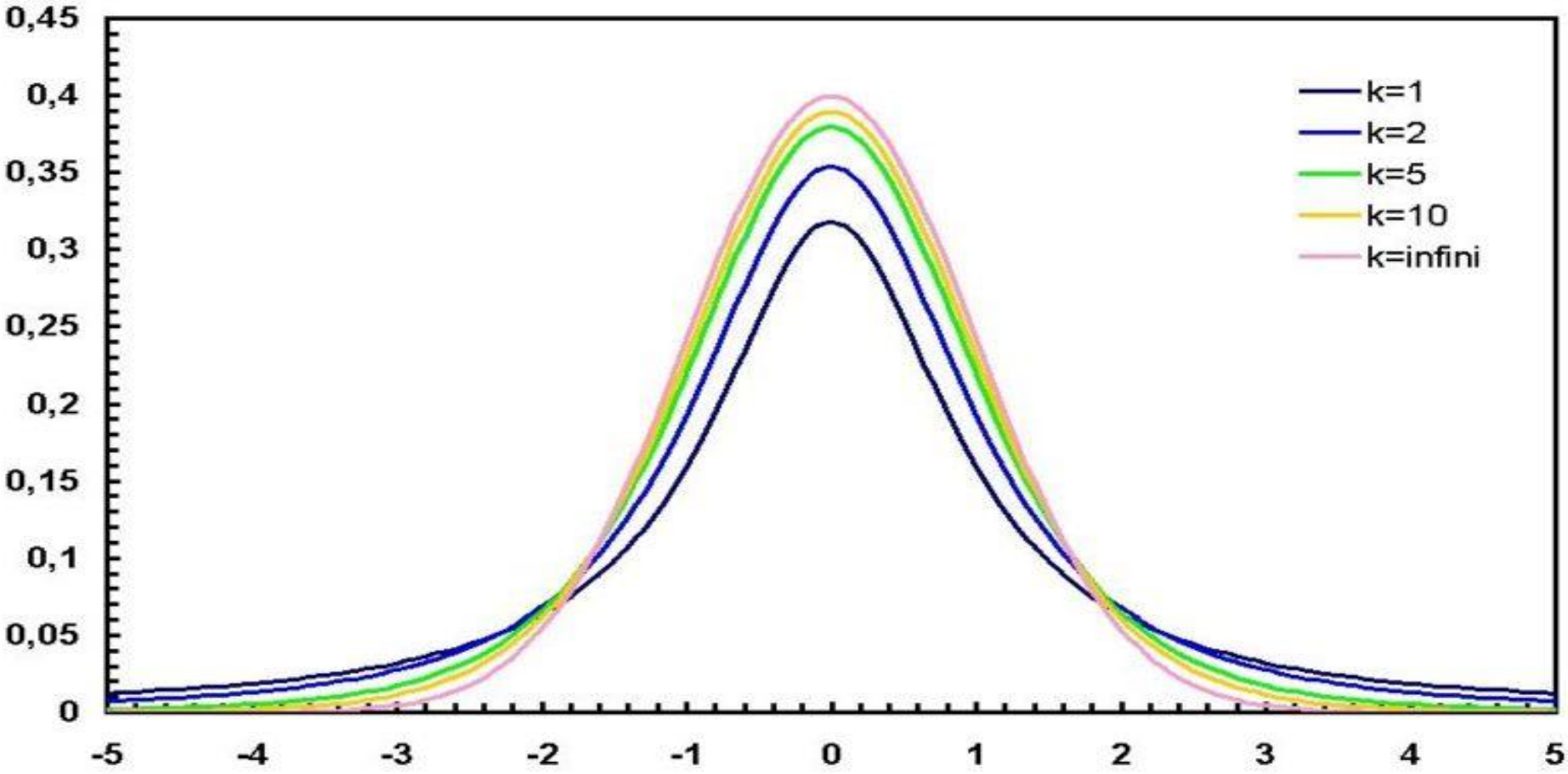
t-distribution

1. **T-distribution varies with degrees of freedom of the sample**

2. **For small n, it is flatter than z-distribution, although also unimodal symmetric, mesokurtosic**

3. **For large n, t-distribution and z-distribution converge**

t-distribution



t-distribution