# Warsaw Summer School 2023, OSU Study Abroad Program

Variability

Standardized Distribution

# Variability, VR

**Variability, RV – also known as spread, width, or dispersion – describes how spread out or closely clustered a <u>set</u> of data is.**

**In the textbook the following measures are discussed: the range, the mean deviation, the variance, and the standard deviation. These are good for metric variables. However, we will start with non metric variables (nominal and ordinal).**

# Variability, VR

- **Nominal variables**

**A comparison of the observed frequency distribution with the uniform distribution (all categories represented by the same number of cases).**

**Index of dissimilarity:**

$$\mathbf{DISS} = \tfrac{1}{2} \sum | p_k - u_k | \text{ for } k = 1, 2, 3 \ldots n$$

**$p_k$ = percentage of observed cases in the category k**

**$u_k$ = percentage of cases in the category k under the uniform distribution**

# Variability, VR

- **Ordinal variables**

**A comparison of the values for all cases with the median value.**

**Absolute deviation from the median**

$$VM = \sum | x_{ij} – Md_j | / N$$

**where $x_{ij}$ refers to the value of the case i of the variable j and $Md_j$ refers to the median of variable j.**

# Variability, VR

**Metric variables (interval and ratio)**

**A comparison of the values for all cases with the mean value.**

**Variance**

$$\mathbf{VAR = s^2 = \sum ( x_{ij} - \overline{X}_{j} )^2 / N}$$

**where $x_{ij}$ refers to the value of the case i of the variable j and $X_j$ refers to the mean of variable j.**

# Variability, VR

**Standard deviation**

**Standard Deviation (s) =**

**the square root of the Variance (of $s^2$)**

$$s = \sqrt{s^2}$$

# Variability, VR

| Scale | Indexes of dissimilarity and diversity | Deviation from the median | Variance/ standard deviation |
|---|---|---|---|
| Nominal | Yes | No | No |
| Ordinal | Yes | Yes | No |
| Interval | Yes | Yes | Yes |

## Basic characteristics of the distribution

**Mean = $(\Sigma X_i) / N$**

**where $X_i$ means “the value for each case,” and Σ means “add all of these up” and i refers to all cases from i = 1 to i = N**

**Standard deviation**

**$S = \sqrt{\text{Var}} = \sqrt{s^2}$**

**$\text{VAR} = s^2 = \sum (X_i - \text{Mean of } X)^2 / N$**

# Z-scores

**Definition: a z-score measures the difference between a raw value (a variable value $X_i$) and the mean using the standard deviation of the distribution as the unit of measure.**

**Z score = (Value for a given case – Mean) / Standard deviation**

**Z score ($X_i$) = $X_i$ – Mean for X / $\sqrt{s^2}$**

# Z-scores

A z-score specifies the precise location of each value Xi within a distribution.

The sign of the z-score (+ or -) signifies whether the score is above the mean (positive) or below the mean (negative).

The numerical value of the z-score specifies the distance from the mean expressed in terms of the proportion of the standard deviation.

# Z-scores

## Raw-values and z-scores

1. Shape. The shape of the z-score distribution is the same as the shape of the raw-score distribution.

2. The mean. When raw scores are transformed into z-scores, the resulting z-score distribution will always have a mean of zero. This fact makes the mean a convenient reference point.

3. The standard deviation. When raw scores are transformed into z-scores, the resulting z-score distribution will always have a standard deviation of one.

# Shape

- **Three properties of the distribution’s shape (metric variables):**

- **MODALITY**
- **KURTOSIS**
- **SKEWNESS**

# Shape

**MODALITY: a number of frequency peaks.**

- **Unimodal: one clear frequency peak.**
- **Bimodal: two clear frequency peaks.**
- **Mutimodal: three or more clear frequency peaks.**

# Shape

## KURTOSIS: Peakedness of a distribution

- **Mesokurtosis** (a “normal” distribution): a moderate peakedness.
- **Platykurtosis** (a flat distribution): a fairly flat lump of values in the center.
- **Leptokurtosis** (a peaked distribution): a high peak of values in the center.

# Shape

**SKEWNESS: Departure from symmetry in a distribution.**

**The way skewness depends on the rightmost or leftmost extreme scores, called the tails of the distribution.**

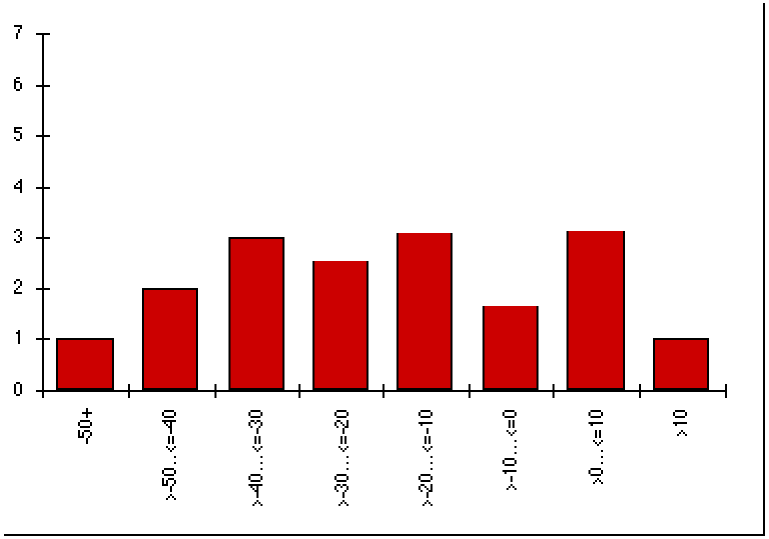
- **Zero skewness: the right tail and the left tail are symmetric.**
- **Positive skewness: the right tail contains extreme (far from symmetric) scores.**
- **Negative skewness: the left tail contains extreme (far from symmetric) scores.**

## CT = central tendency, VR = variability

- **How do CT and VR help us to assess the distribution's shape?**

**Modality: Comparison of the mode with the frequency of other values informs us whether we have one clear frequency peak, or two picks, or more than two picks.**

# Multimodal distribution



## Kurtosis

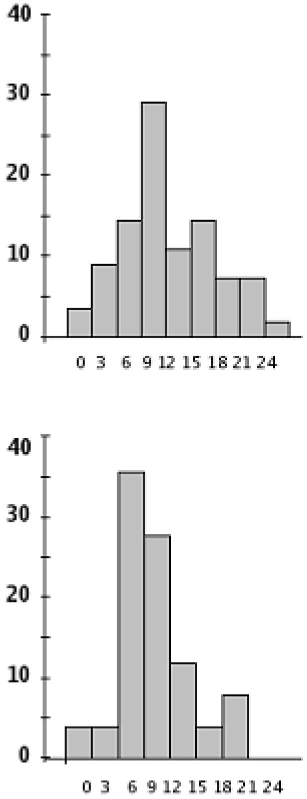
**Kurtosis: Comparison of the standard deviation (SD) with the mean (M) shows whether we have picked distribution of not:**

**SD >> M  = peaked distribution**

**SD << M = flat distribution**

**A >> B, A is many times greater than B**
**A << B, A is many times smaller than B**

# Kurtosis



# Skewness

- **Skewness:**

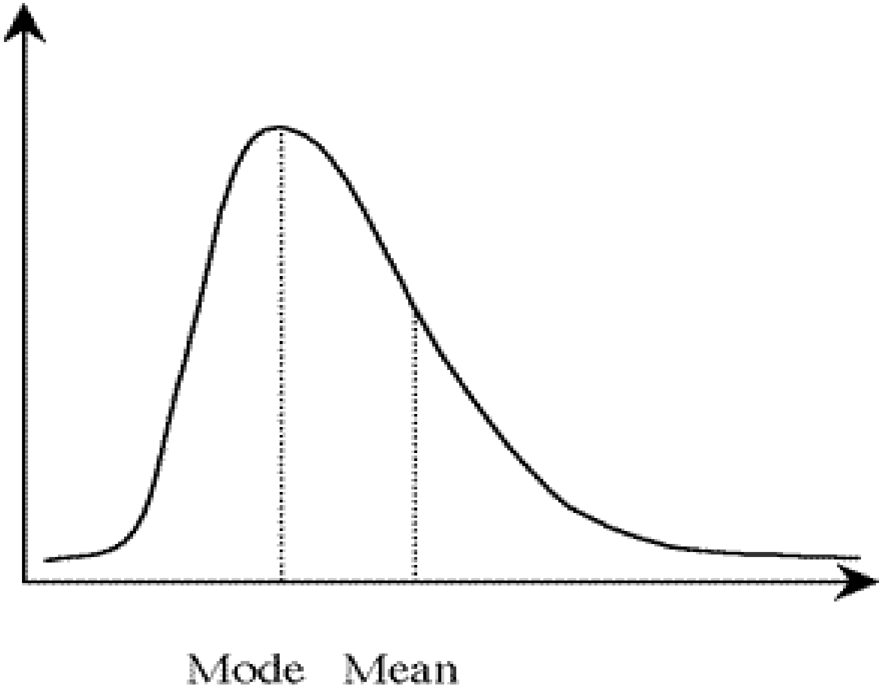
  **- Whenever Mean > Median > Mode → positively skewed distribution.**

  **- Whenever Mean < Median < Mode → negatively skewed distribution.**

  **- Whenever Mean = Median = Mode →**

  **normal distribution.**

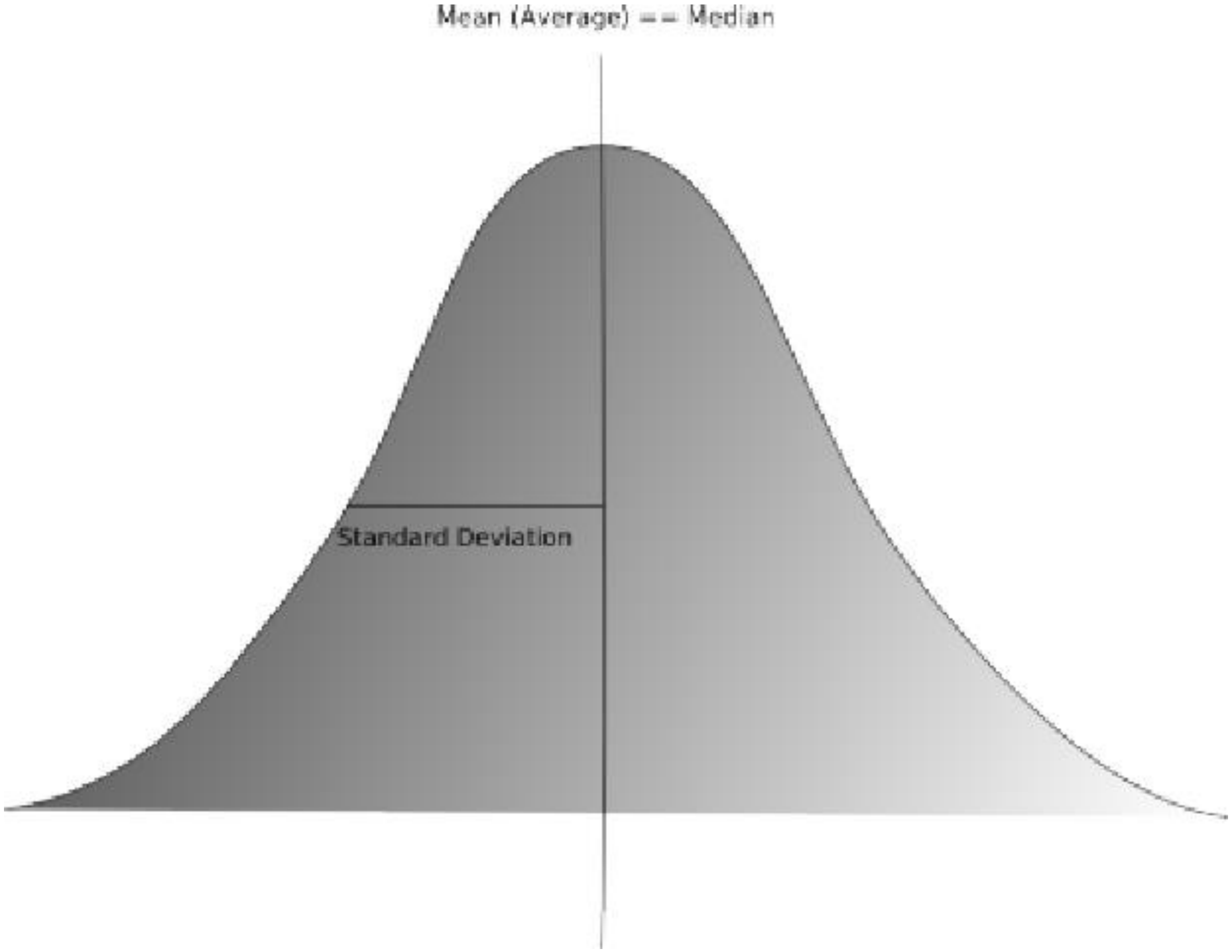
# Skewness



# Normal distribution

- **Normal distribution:**
  - **Unimodal**
  - **Mesokurtosic (a moderate peakedness)**
  - **Symetric (<u>zero skewness)</u>**

# Normal distribution



# Normal distribution

**Normal distribution** refers to the distribution in which the mean and standard deviation are given and the shape of the distribution is derived from a mathematical equation. This distribution is:

- (1) unimodal,
- (2) symmetric,
- (3) mesokurtic.

## 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

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

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

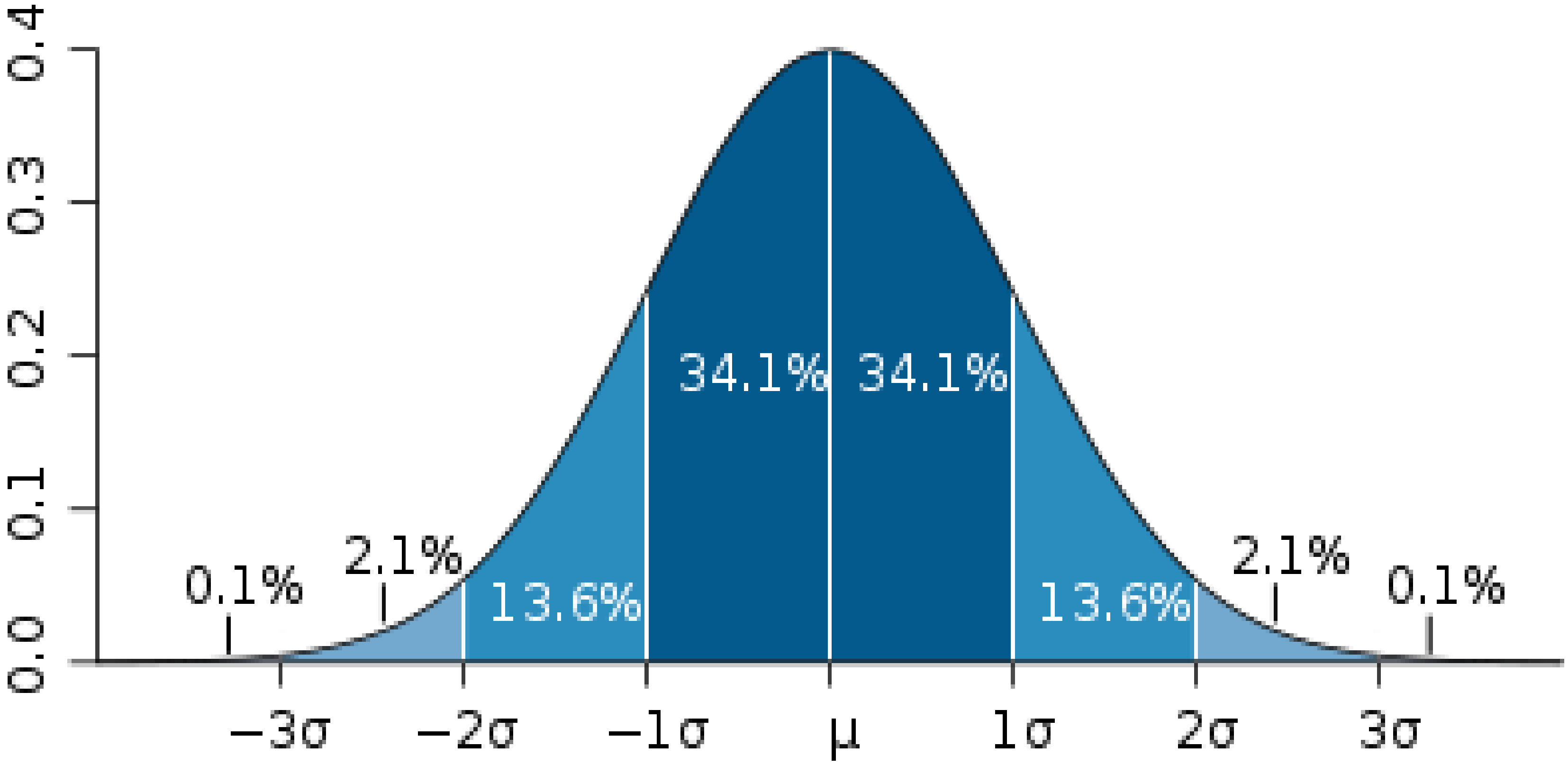
**(2) It provides a tool for deciding about errors that we might commit in testing hypotheses (in inferential statistics).**

## 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

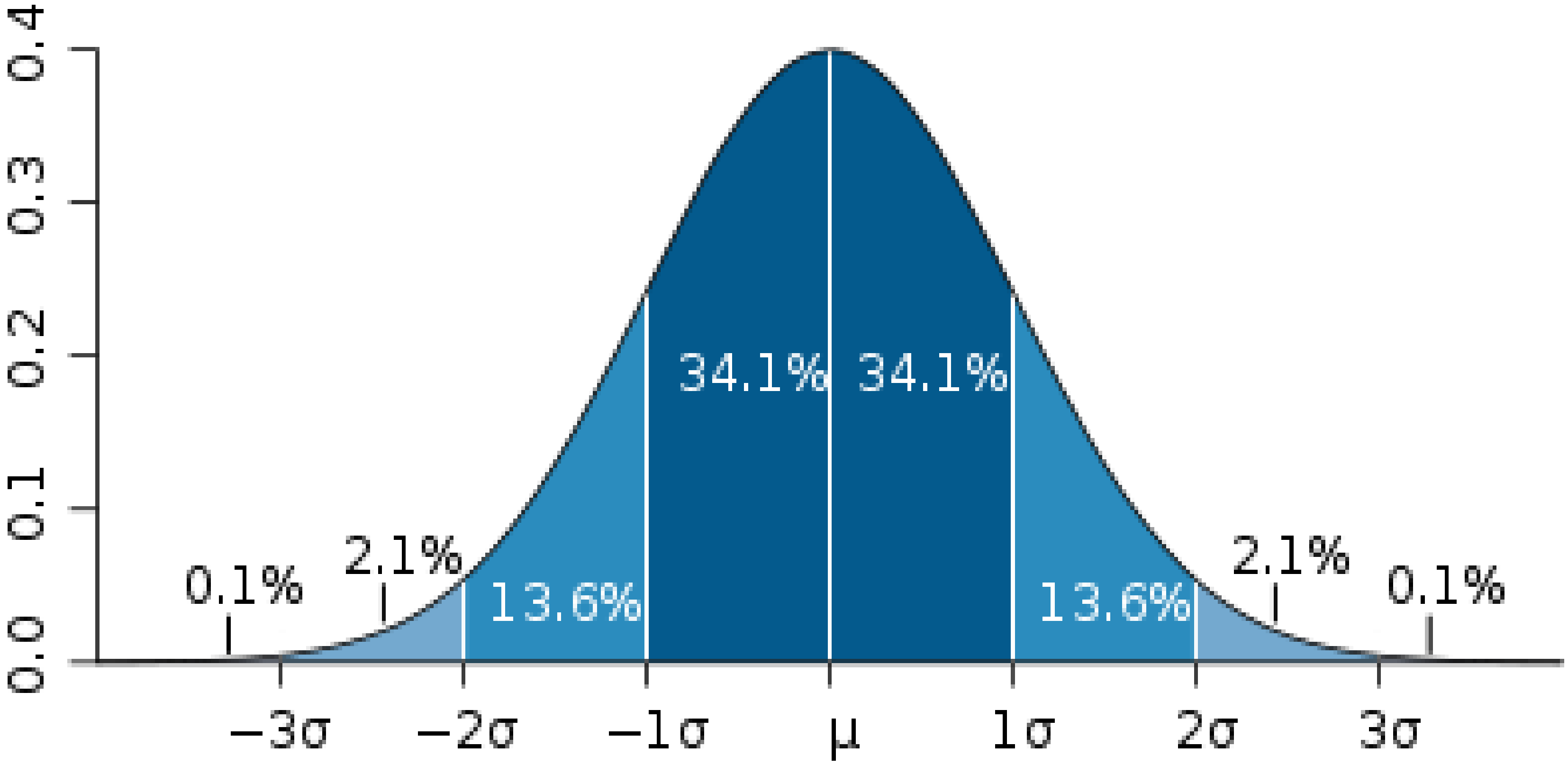
How far apart are two people A and B, where A = 115, and B = 99.

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 and 115 is one standard deviation 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 $X_K < X_L$.

# Z-scores

For finding points in the normal distribution the mean value and the standard deviation are crucial.

Z score = (Value – Mean) / Standard deviation

# Z-scores

- **Why is the knowledge about z-scores so important? For two reasons:**
- **- first, in evaluating individual scores we rely on deviations from the average;**
- **- second, in evaluating individual scores we want to take into account how the scores are spread out.**

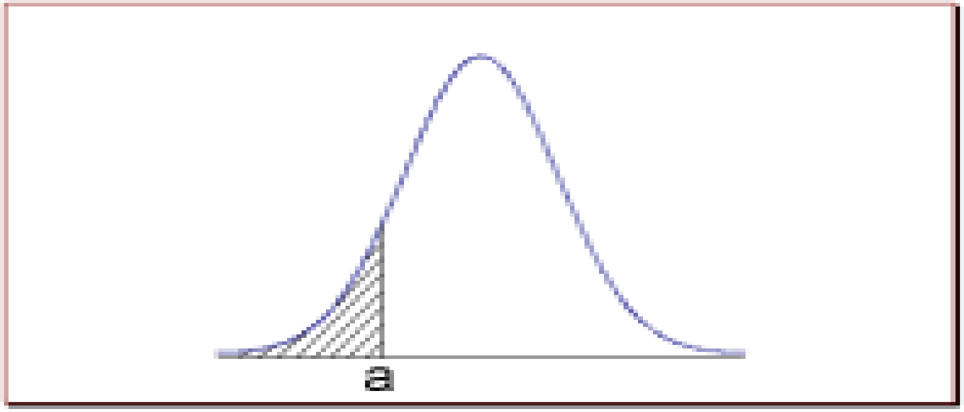
## Probability

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).**

**For probability distribution of the observed variable we use μ for the mean, and σ for standard deviation (s).**

**Standardized normal probability distribution is expressed by z-scores:**

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

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

**Appendix 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 |