

SC.203 - Scientific Method

Lecture 5 - Introduction to Design of Experiments and Data Analysis

Prof. DUONG Nguyen Vu

Professor, Director

John von Neumann Institute, Vietnam National University HCM

Assoc. Prof. TRAN Minh-Triet

Faculty of Information Technology

University of Science - Vietnam National University HCM

- Introduction to controlled experiments, providing:
 - Practical rules rather than theoretical explanations,
 - Necessary notions to get acquainted to some basic elements of design of experiments,
 - Commonly used statistical tests, namely Analysis of Variance (ANOVA),
 - Brief description of non-parametric alternatives to ANOVA,
 - Simple examples entailing only one independent variable, but more than two levels, allowing to plan and perform quite complex exploratory experiments.

Obligations and Recommendations

- The researcher has responsibility towards the subjects involved in the experiment.
- The subjects have the right to privacy, confidentiality, and the right to be informed about the nature of the research, and must be treated with respect.
- Any subject has the right to withdraw from an experiment at any time.
- Any explanation that is required by the subjects shall be provided (compatibility with the purposes of the experiment, i.e. do not reveal the hypotheses beforehand).
- Sometimes the subjects might feel frustrated if they feel that their performance was not good enough. The researcher must stress that the main goal of the research is to assess the performance or to evaluate some properties of some tools/algorithms/automations in general. Must make sure that every participant feels comfortable.
- It is a good practice to obtain an informed consent from the participants.
- The researcher shall be as neutral as possible, do not reveal expectations about the experiments either explicitly or implicitly.

Other Obligations

- Other obligations of the researcher:
 1. Perform a risk/benefits evaluation (e.g. possible effects on the participants, etc.)
 2. Do not make any alteration on the data. If so, then explain what was done and why (e.g. how and why some participants' scores were removed.)
 3. Errors identified *a posteriori* to the research have to be acknowledged to the readers.
 4. Avoid plagiarism and do not write other people's work as if it was own.
 5. Anybody who contributed to a research should be acknowledged in all written materials.
 6. Share the data.

Planning an Investigation

- Some advices on how to structure and plan a research:
 1. *Review past studies in the same area* (literature review). The knowledge of what was done before on the same topic is essential to every research.
 2. *Defining a specific topic*. The review of the state of the art, the specific problems and topics of interests can be identified.
 3. *Define the hypotheses*. In general terms, a research should be guided by some hypotheses. This is mostly true when we want to perform experiments: the aim of an experiment is to test a hypothesis.
 4. *Pilot studies*. It is a good habit to run some preliminary trials, in order to check if the design and the procedures are appropriate and to verify if the equipments are properly working.

Variables and Conditions

- Experiment allows the manipulation of some factors in order to measure the possible effects of this manipulation.
- **Variables** are measurable properties of a certain event. A variable can be:
 - **Independent**, i.e. causing a change of another variable, or
 - **Dependent**, i.e. affected by a dependent variable.
 - **Irrelevant**, i.e. can unpredictably affect the outcomes of an experiment.
- Example: “test whether a new teaching method would help young pupils in the study of math.”
 - A simple design: two groups of pupils: one learning with new method, and one with traditional scheme.
 - independent variable manipulated here is the **teaching method** and has two levels: *new* vs. *traditional* method.
 - Dependent variable: scores obtained by the two groups of pupils in an exam.

Validity

- **Validity** refers to the conclusions that a researcher can establish concerning the causal relationship between the independent and the dependent variable. There are four types of validity that need to be taken into account: *internal* validity, *construct* validity, *external* validity, and *statistical* validity.

Validity

- **Internal validity**, refers to the internal logic of the relationship identified: the relationship has to be clear, meaning that no irrelevant variable had, in someway, caused the effect (i.e. modified the dependent variable).
- **External validity** refers to the generalizability of the results of a research to other subjects, settings, etc.

Validity

- **Construct validity**, refers to the extent to which what it is manipulated and measured it is really what is needed to support a theory.
 - Example: taking again the example of the pupils learning math, the researcher should evaluate the adequacy and the strength of the teaching method and also consider whether measuring the math scores is the most appropriate way to show that the method was really effective, etc.).
- **Statistical validity**, this type of validity implies that the causal relationship (between and independent variable and dependent variable) discovered should not be casual.

Measurements and Scales

- The notion of variable is closely related to the notion of measurement, a rather obvious statement. “Measuring” means assigning numerical values to an event or to an object according to a certain rule.
- It is common to distinct four scales of measurement that differ in the manner (the rule) the values are assigned to events or objects.
 - Nominal scales. In this scale, the numbers are labels allowing the classification of objects and events according to some categories (e.g. female=1, male=2, or vice versa).
 - Ordinal scales. This scale allows the ranking of the events or objects according to an order (e.g. from “very sweet”=10, to “very bitter” =1, etc.). This type of scale (e.g. the ordinal position) does not provide any information about the differences between its composing elements; in other words, one cannot say that “very sweet” is ten times sweeter than “very bitter”.

Measurements and Scales

- Interval scales. With the interval scale, instead, the differences among the values have meanings, since it is assumed that there are equal intervals between the values.
 - A classical example of interval scale is the temperature. The difference between 30° and 20° Celsius is the same as between 40° and 50° .
- Ratio scales. Similarly to the interval scale, it scale gives information concerning the magnitude of the differences between the measured events, but the ratio scale gives additional information: it is provided with a true zero point.
 - Taking again the example of the temperature: would it be meaningful to state that: 20° is twice as warm as 10° ? If the same example is provided with centimetres, then the statement is provided with sense.

ELEMENTS OF STATISTICS OFTEN USED IN DATA ANALYSIS

Prof. Dr. Vu Duong

*EUROCONTROL Experimental Center &
University of Science - Vietnam National University HCM*

CONTENTS

- Module 1: Data Organization
- Module 2: Descriptive Statistics
- Module 3: Inferential Statistics & Hypothesis Testing
- Module 4: Parametric Tests

DATA ORGANIZATION

Two methods of organizing data:

- Frequency distributions,
- Graphs.

Frequency distribution – A table in which all of the scores are listed along with the frequency with which each occurs.

Graphs – most common graphs are bar graphs, histograms, and frequency polygons (line graphs).

FREQUENCY DISTRIBUTIONS

- Example 1: Exam scores of this class (100 students).
 1. List the scores following names of students,
 2. Sort the scores from lowest to highest
 3. For each score:
 - count the **frequency** f (num of occurrences), add in one column.
 - calculate the **relative frequency** rf (ratio of f over number of samples), fill in one additional column.

CLASS INTERVALS

- Frequency distribution is a way of presenting data that makes their patterns easier to see.
- With large datasets, we can group the scores and create a **class interval frequency distribution** by combining individual scores into categories, or intervals, and list them along with the frequency of scores in each interval.
- Rule of thumb when creating class intervals is to have between 10 and 20 categories. (Hinkle et al., 1998)
- A quick method of calculating the width of the interval is to subtract the lowest score to the highest, and then divide the result by the number of intervals we would like.

CLASS INTERVALS

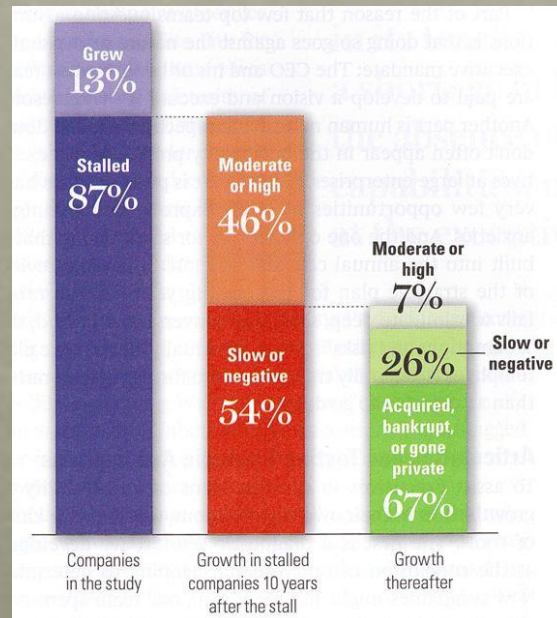
Class interval frequency distribution – A table in which the scores are grouped into intervals, and listed along with the frequency of scores in each interval.

GRAPHS

- “A picture is worth a thousand words.” ☺
- Pictorial representations can be used to represent data. The choice depends on the type of data collected, and what the researcher hopes to emphasize or illustrate.
- Graphs typically have two coordinated axes. The y -axis is usually shorter than the x -axis. (typically 60-75%)

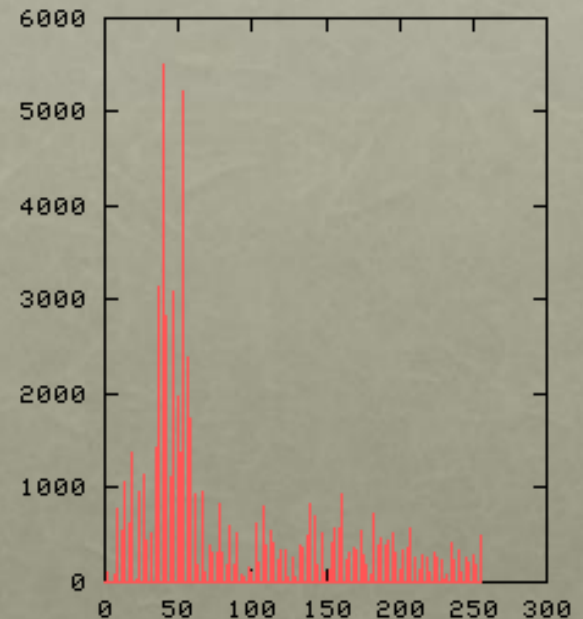
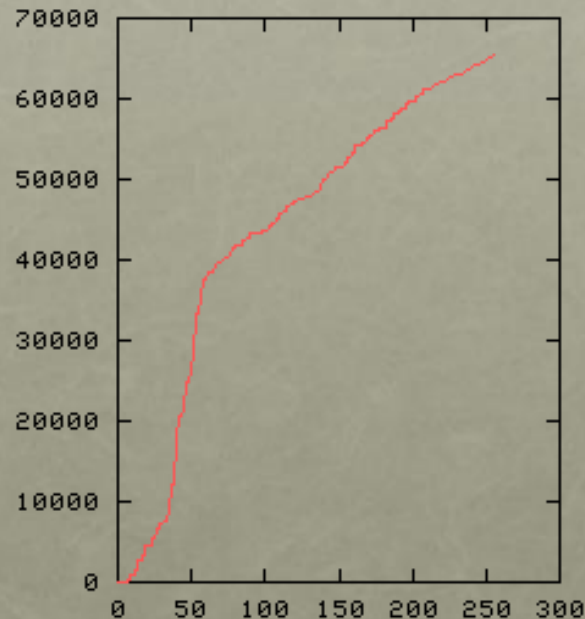
BAR GRAPHS

- **Bar graph** – A graphical representation of a frequency distribution in which vertical bars are centered above each category along the x -axis and are separated from each other by a space, indicating that the levels of the variable represent distinct, unrelated categories.



HISTOGRAMS

- **Histogram** – A graphical representation of a frequency distribution in which vertical bars centered above scores on the x -axis touch each other to indicate that the scores on the variable represent related, increasing values.

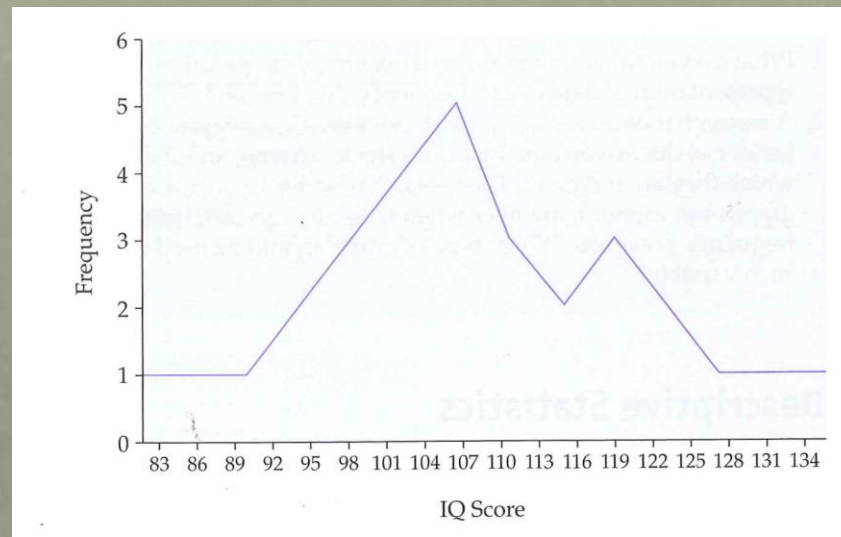
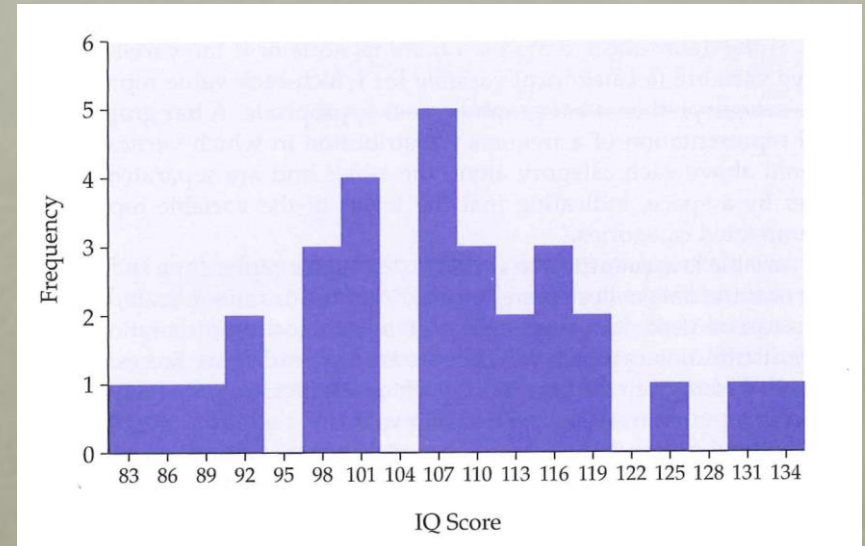
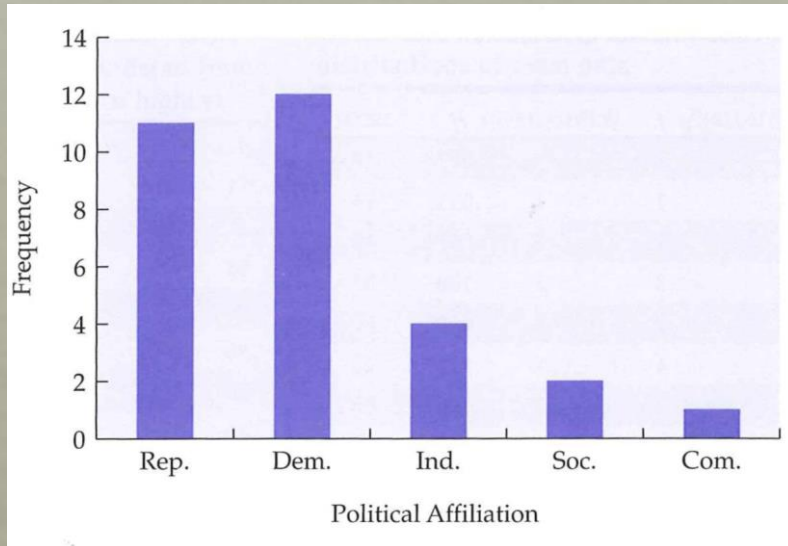


USAGE

Bar graphs and histograms are frequently confused.

- If the variable is a **qualitative** variable, or if the data collected are on a nominal scale, then a bar graph is most appropriate.
- If the variable is **quantitative**, or if the data collected are ordinal, interval, or ratio in scale, then a histogram can be used.
 - In both cases, the the height of each bar indicates the frequency for that level of the variable (on the x -axis).
 - The spaces on a bar graph indicate not only the qualitative differences among the categories, but also that the order of values of the variables on the x -axis is arbitrary (can be placed in any order).
 - The contiguous bars in a histogram indicate not only the increasing quantity of the variable, but also that the values of the variable have a definite order that cannot be changed.

FREQUENCY POLYGON



MORE DEFINITIONS

- **Qualitative variable** – A categorical variable for which each value represents a discrete category.
- **Quantitative variable** – A variable for which the scores represent a change in quantity.
- **Frequency Polygon** – A line graph of the frequencies of individual scores.
- **Cumulative Histogram** – Histogram in which frequencies are cumulated with increasing frequency intervals .

Data Organization

IN REVIEW

TYPES OF ORGANIZATIONAL TOOLS

	Frequency Distribution	Bar Graph	Histogram	Frequency Polygon
Description	A list of all scores occurring in the distribution along with the frequency of each	A pictorial graph with bars representing the frequency of occurrence of items for qualitative variables	A pictorial graph with bars representing the frequency of occurrence of items for quantitative variables	A line graph representing the frequency of occurrence of items for quantitative variables
Use with.	Nominal, ordinal, interval, or ratio data	Nominal data	Typically ordinal, interval, or ratio data; most appropriate for discrete data	Typically ordinal, interval, or ratio data; most appropriate for continuous data

CRITICAL THINKING CHECK 1

1. What do you think might be the advantage of a graphical representation of data over a frequency distribution?
2. A researcher observes driving behavior of a roadway, noting the gender of drivers, the types of vehicles driven, and the speeds at which they are traveling. Which type of graph should be used to describe each variable (gender, types of vehicles, speeds)?

DESCRIPTIVE STATISTICS

Descriptive statistics are Numerical measures that describe a distribution by providing information on:

- the central tendency of the distribution,
- the width of the distribution, and
- the shape of the distribution

MEASURES OF CENTRAL TENDENCY

- A measure of **central tendency** is a representative number that characterizes the “*middleness*” of an entire set of data.
- The three measures of central tendency: **mean, median, mode**.

Mean – arithmetic average of a distribution.

Median – middle score in a distribution after the scores have been arranged from highest to lowest or versa.

Mode – score in a distribution that occurs with the greatest frequency.

MEAN

- Most commonly used measure of central tendency. Mathematically, this is



where:

- μ represents the symbol for the population mean;
- Σ represents the symbol for “the sum of”;
- x_i represents the individual scores i ; and
- N represents the number of scores in the distribution.

MEDIAN

- Median is used when *mean* might not be representative of a distribution.
- Example of mean salary of a company of 25 employees is circa. 100K\$, but the distribution is biased by the salary of the CEO at 1.8M\$. The mean is not representative of the central tendency of the distribution.

EXAMPLE

SALARY	FREQUENCY	fx
\$ 15,000	1	15,000
20,000	2	40,000
22,000	1	22,000
23,000	2	46,000
25,000	5	125,000
27,000	2	54,000
30,000	3	90,000
32,000	1	32,000
35,000	2	70,000
38,000	1	38,000
39,000	1	39,000
40,000	1	40,000
42,000	1	42,000
45,000	1	45,000
1,800,000	<u>1</u>	<u>1,800,000</u>
$N = 25$		$\Sigma X = 2,498,000$

MODE

- Mode – the score that occurs at with the greatest frequency.
 - In a distribution where all scores occur with equal frequency, such distribution has no mode.
 - In another distribution, several scores occur with equal frequency, such distribution may have 2-mode (bimodal), 3-mode (tri-modal), or n -modes.
 - The mode is the only indicator of central tendency that can be used with nominal data.

SUMMARY

IN REVIEW

Measures of Central Tendency

TYPES OF CENTRAL TENDENCY MEASURES

	Mean	Median	Mode
Definition	The arithmetic average	The middle score in a distribution of scores organized from highest to lowest or lowest to highest	The score occurring with greatest frequency
Use with	Interval and ratio data	Ordinal, interval, and ratio data	Nominal, ordinal, interval, or ratio data
Cautions	Not for use with distributions with a few extreme scores		Not a reliable measure of central tendency

CRITICAL THINKING CHECK 2

1. In the example of traffic observation (gender, types of vehicle, speeds). What is an appropriate measure of central tendency to calculate for each type of data?
2. If one driver was driving at 100km/h (25km/h faster than anyone else), which measure of central tendency would you recommend against using?

MEASURES OF VARIATION

- **Measure of variation** –a number that indicates the degree to which scores are either clustered or spread out in a distribution.
 - The three measures of variation: **range, standard deviation, variance.**
- **Range** – the difference between the lowest and the highest scores in a distribution. Usually reported with the mean of the distribution.
 - **Standard deviation** – the average difference between the scores in a distribution and the mean or central point of the distribution, or, more precisely, the square root of the averaged squared deviation from the mean.
 - **Variance** – the standard deviation squared.

RANGE

- Provides information concerning the difference in the spreads of the distributions.
- This simple measure of variation only the highest and lowest scores enter the calculation, and all other scores are ignored.
- Thus the range is easily distorted by one unusually high or low score in a distribution.

STANDARD DEVIATION σ

- More sophisticated measure of variation, uses all the scores in the distribution in its calculation.

- Most commonly used.

Standard: average, normal, usual.

Deviation: diverge, move away from, digress.

Standard Deviation: the average movement away from something, e.g., the center of the distribution – the mean.

- It's the average distance of all the scores in the distribution from the mean or central point in the distribution – or the square root of the averaged squared deviation from the mean.

CALCULATING σ

- To calculate the average distance of all the scores from the mean of the distribution, we (1) first determine how far each score is from the mean (this is the deviation, or difference score); then (2) average these scores.

- In other words:

1 calculate for each score:

2 sum the squared deviation scores:

3 divide the sum by the total number of scores and take the square root of that number.

EXAMPLE

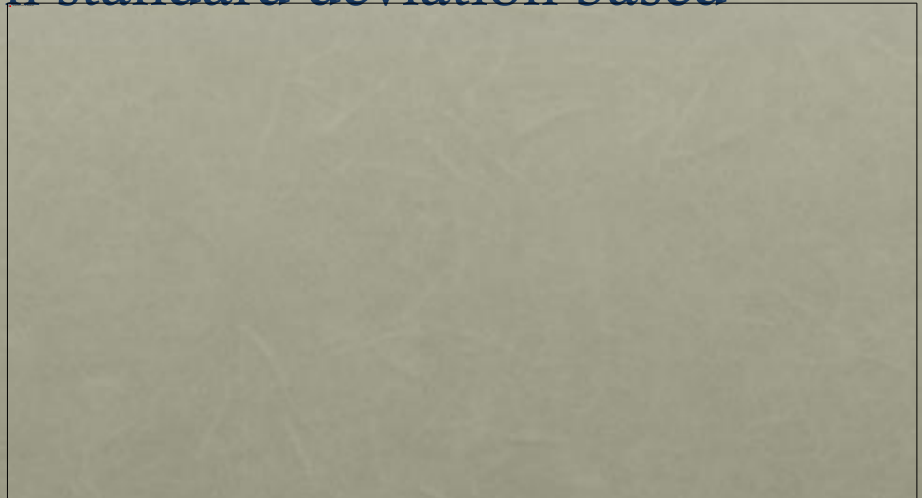
X	$X - \mu$	$(X - \mu)^2$	f	$f(X - \mu)^2$
45	-29.00	841.00	1	841.00
47	-27.00	729.00	1	729.00
54	-20.00	400.00	1	400.00
56	-18.00	324.00	1	324.00
59	-15.00	225.00	1	225.00
60	-14.00	196.00	2	392.00
63	-11.00	121.00	1	121.00
65	-9.00	81.00	1	81.00
69	-5.00	25.00	1	25.00
70	-4.00	16.00	1	16.00
74	0.00	0.00	3	0.00
75	1.00	1.00	1	1.00
76	2.00	4.00	1	4.00
77	3.00	9.00	1	9.00
78	4.00	16.00	2	32.00
80	6.00	36.00	1	36.00
82	8.00	64.00	2	128.00
85	11.00	121.00	1	121.00
86	12.00	144.00	1	144.00
87	13.00	169.00	1	169.00
90	16.00	256.00	1	256.00
92	18.00	324.00	1	324.00
93	19.00	361.00	1	361.00
94	20.00	400.00	1	400.00
95	21.00	441.00	1	441.00
			$N = 30$	$5,580.00 = \sum (X - \mu)^2$

σ RAW-SCORE FORMULA



SAMPLING THE POPULATION

- If you are using sample data to estimate the population standard deviation, then the standard deviation formula must be slightly modified.
- The modification provides what is called unbiased estimator of the population standard deviation based on sample data.
- The modified formula is



UNBIASED ESTIMATOR

- Notice that the symbol for unbiased estimator is s whereas the symbol for the sample standard deviation is S .
- Main difference is the denominator $N-1$ rather than N . The reason is that the standard deviation within a small sample may not be representative of the population; i.e., there may not be as much variability in the sample as there actually is in the population. We therefore divide by $N-1$ because dividing a smaller number increases the standard deviation and thus provides a better estimate of the population σ .

VARIANCE σ^2

- Variance = standard deviation squared.
 - for a population is σ^2 ;
 - for a sample is S^2 ;
 - for the unbiased estimator is s^2 ;
- The variance tells us the degree to which each individual score deviates from the mean of the distribution. It can also be calculated starting from the sum of the squared deviates, obtained subtracting each score from the mean and squaring each value. Then all the values are summed up; this sum is then divided by the number of scores.
- Variance is not appropriate with ordinal or nominal data.

CRITICAL THINKING CHECK 3

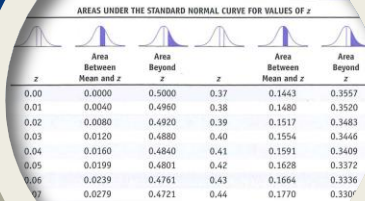
1. For a distribution of scores, what information does a measure of variation convey that a measure of central tendency does not?
2. Today's weather report included information on the normal rainfall for this time of the year. The amount of rain that fell today was 1.5cm above normal. To decide whether this is an abnormal high, you need to know that the standard deviation for rainfall is 0.75cm.
 - a) What would you conclude about how normal the amount of rainfall was today?
 - b) Would your conclusion be different if the standard deviation were 2cm rather than 0.75cm?

ANSWERS 3

1. A measure of variation tells us about the spread of the distribution, i.e., are the scores clustered closely about the mean, or are they spread over a wide range?
2. Rainfall:
 - a) The amount of rainfall for the indicated day is 2 standard deviations above the mean: well above the average.
 - b) If the standard deviation was 2 rather than 0.75, then the amount of rainfall for the indicated day would be less than 1 standard deviation above the mean: above average but that's it!

TYPES OF DISTRIBUTIONS

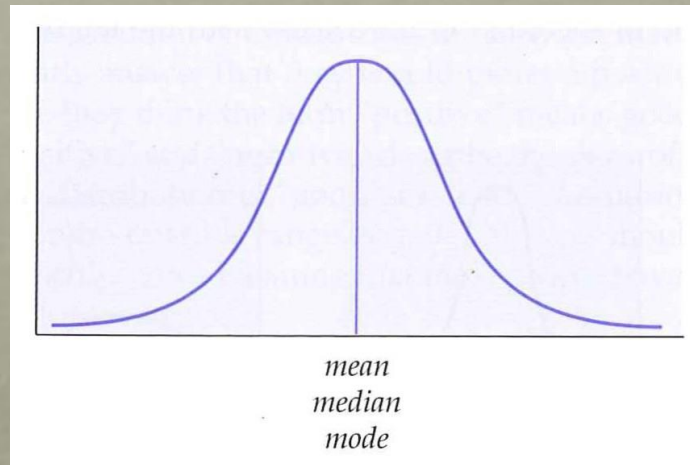
AREAS UNDER THE STANDARD NORMAL CURVE FOR VALUES OF z



z	Area Between Mean and z	Area Beyond z	z	Area Between Mean and z	Area Beyond z
0.00	0.0000	0.5000	0.37	0.1443	0.3557
0.01	0.0040	0.4960	0.38	0.1480	0.3520
0.02	0.0080	0.4920	0.39	0.1517	0.3483
0.03	0.0120	0.4880	0.40	0.1554	0.3446
0.04	0.0160	0.4840	0.41	0.1591	0.3409
0.05	0.0199	0.4801	0.42	0.1628	0.3372
0.06	0.0239	0.4761	0.43	0.1664	0.3336
0.07	0.0279	0.4721	0.44	0.1700	0.3300

NORMAL DISTRIBUTIONS

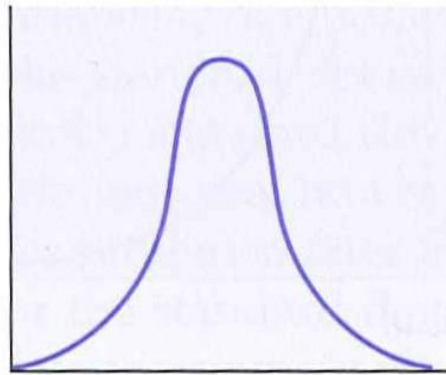
- Normal distribution – when a distribution of scores is very large, it often tends to approximate a pattern called a *normal distribution*. When plotted as a frequency polygon, it forms a symmetrical, bell-shaped pattern often called a *normal curve*.



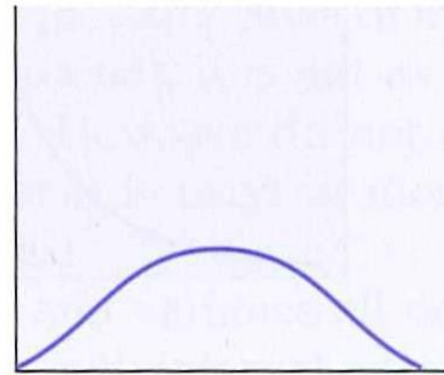
- Normal distribution – A theoretical frequency distribution that is (i) bell-shaped symmetrical , (ii) the mean, median, and mode are equal, (iii) uni-modal, (iv) most observations are at the center of the distribution, and (v) when σ is plotted on the x-axis, the percentage of scores falling between the mean and any point on the x-axis is the same.

KURTOSIS

- **Kurtosis** – how flat or peaked a normal distribution is.
- **Mesokurtic** – Normal curves that have peaks of medium height and distributions that are moderate in breadth.
- **Leptokurtic** – normal curves that are tall and thin, with only a few scores in the middle of the distribution having a high frequency.
- **Platykurtic** – normal curves that are short and relatively more dispersed.



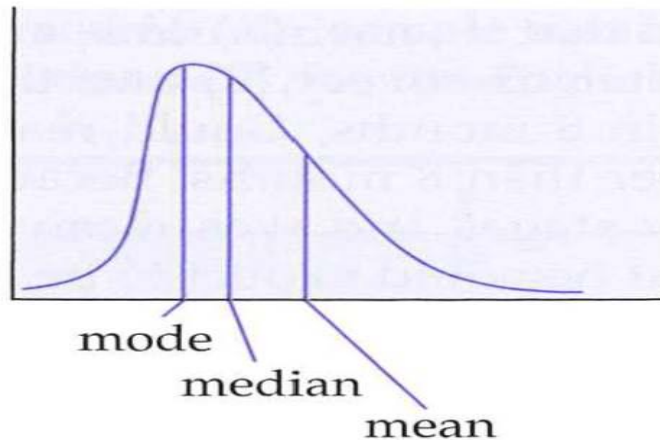
Leptokurtic



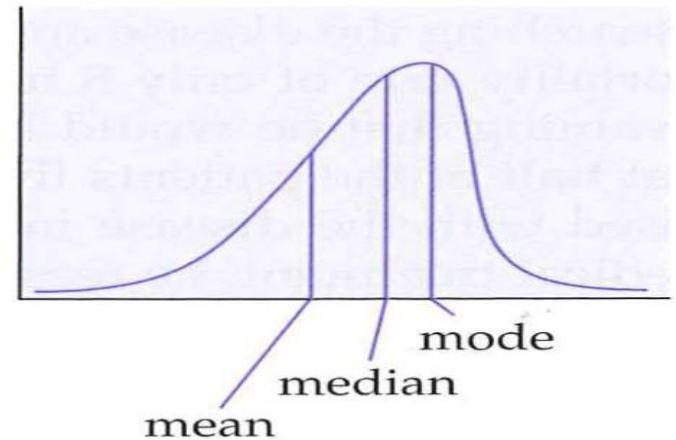
Platykurtic

SKEWED DISTRIBUTIONS

- **Positively skewed distribution** – a distribution in which the peak is to the left of the center point and the tail extends toward the right, or in the positive direction.
- **Negatively skewed distribution** – a distribution in which the peak is to the right of the center point and the tail extends toward the right, or in the negative direction.



positively skewed distribution



negatively skewed distribution

Z-SCORES

- Z-score –a number that indicates how many standard deviation units a raw score is from the mean of a distribution.

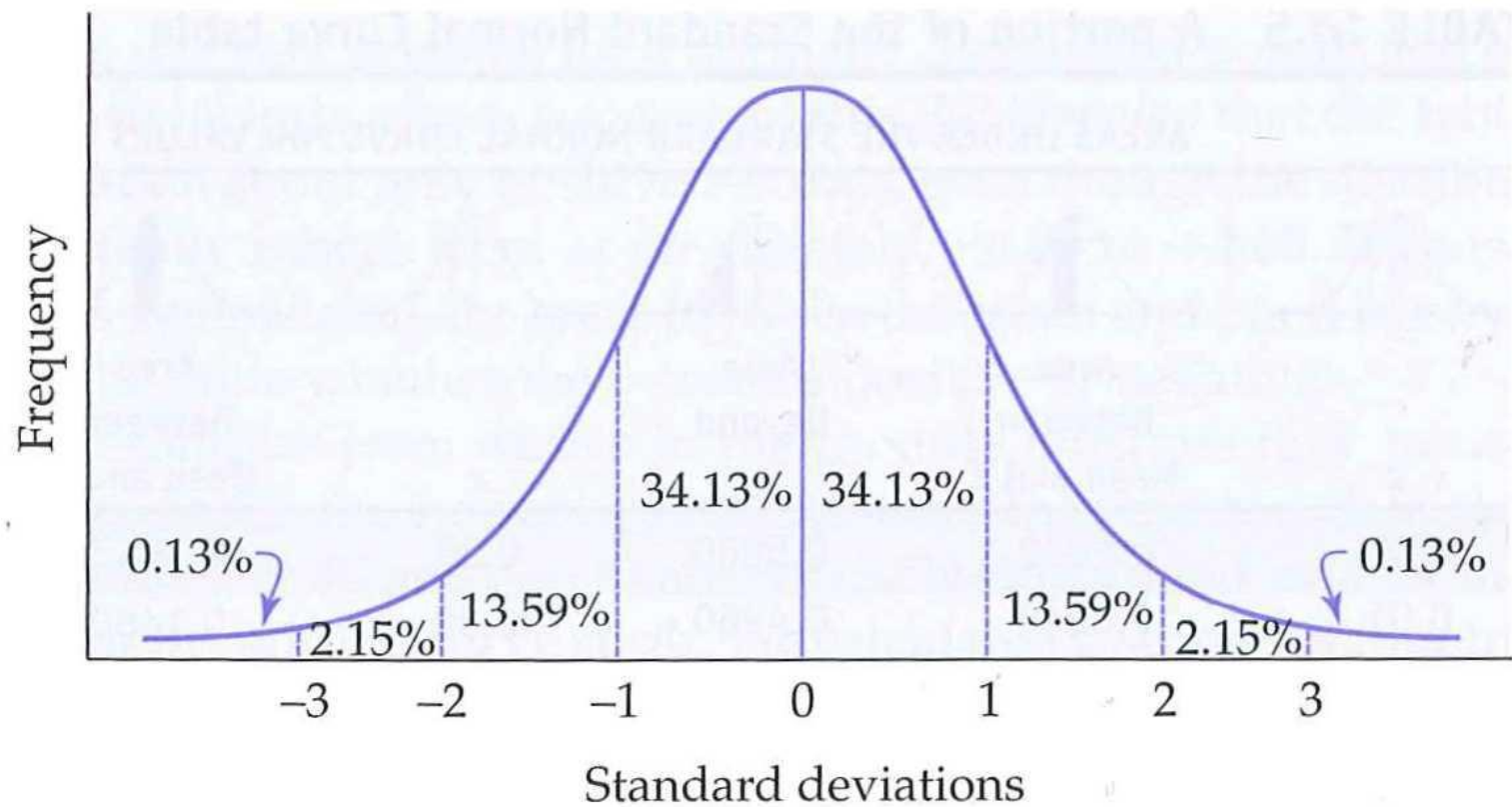


- Z-score is used for comparative purposes between a single score to the set.

STANDARD NORMAL DISTRIBUTION

- Standard normal distribution – has the mean of 0 and standard deviation of 1. It is a theoretical distribution defined by a specific mathematic formula for comparative purposes.
- It provides information about the proportion of scores that are higher or lower than any other score in the distribution, as well as the probability of occurrence of score that is higher or lower than any other score in the distribution.

STANDARD DEVIATIONS

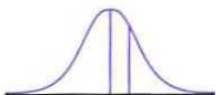
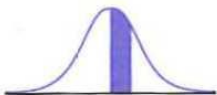
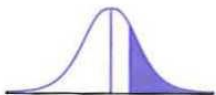

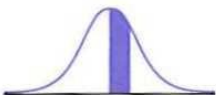
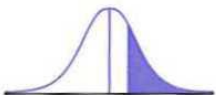


PERCENTILE RANK

- Symmetrical property of standard normal distribution helps determine *probabilities*.
 - Probability of a score that falls above the mean is equal to the proportion in that area, i.e., $p = 0.50$
- Standard normal curve can also be used to determine an individual's **percentile rank** – the percentage of scores equal to or below a given raw score, or the percentage of scores the individual's score is higher than.
- From the individual z-score, percentile rank is calculated as its corresponding “*Area between Mean and z*” added by 0.50.
 - If z-score is +1.27, the “*Area between Mean and z*” of +1.27 is 0.8980 (see the Statistical Tables), added by 0.50, the percentile rank is being in the 89.9th percentile.
 - If z-score is negative, we use the “*Area Beyond z*”

STATISTICAL TABLES

AREAS UNDER THE STANDARD NORMAL CURVE FOR VALUES OF z

					
z	Area Between Mean and z	Area Beyond z	z	Area Between Mean and z	Area Beyond z
0.00	0.0000	0.5000	0.37	0.1443	0.3557
0.01	0.0040	0.4960	0.38	0.1480	0.3520
0.02	0.0080	0.4920	0.39	0.1517	0.3483
0.03	0.0120	0.4880	0.40	0.1554	0.3446
0.04	0.0160	0.4840	0.41	0.1591	0.3409
0.05	0.0199	0.4801	0.42	0.1628	0.3372
0.06	0.0239	0.4761	0.43	0.1664	0.3336
0.07	0.0279	0.4721	0.44	0.1770	0.3300

RAW SCORE FROM PERCENTILE RANK?

- If the individual is at the 75th percentile, we can deduce that the “Area between Mean and z ” is 0.25 by subtracting 0.50, and the corresponding z -score in the lookup table is 0.67
- If the mean $\mu = 90$ and standard deviation $\sigma = 15$, given z , we can calculate x_i from $z = (x_i - \mu) / \sigma$.
- In this case $x_i = z \cdot \sigma + \mu = 0.67 \times 15 + 90 = 100.05$

USING STANDARD NORMAL DISTRIBUTIONS

- Standard normal distribution is useful for determining how a single score compares with a population, or samples of scores, and also for determining probabilities and percentile ranks.
- Knowing how to use the proportions under the standard normal curves increases the information we can derive from a single score.

IN REVIEW

Distributions

TYPES OF DISTRIBUTIONS

	Normal	Positively Skewed	Negatively Skewed
Description	A symmetrical, bell-shaped, unimodal curve	A lopsided curve with a tail extending toward the positive or right side	A lopsided curve with a tail extending toward the negative or left side
z-score transformations applicable?	Yes	Yes	Yes
Percentile ranks and proportions under standard normal curve applicable?	Yes	No	No

CRITICAL THINKING CHECK 4

1. Concept sketch:

- On one graph, draw two distributions with the same mean but different standard deviations.
- Draw a second set of distributions on another graph with different means but the same standard deviation.

2. Why is it not possible to use the proportions under the standard normal curve with skewed distributions?

CRITICAL THINKING CHECK 4

3. Students at VNU consume an average of 7 coffees per day with a standard deviation of 2.5,
- a. What proportion of students consume an amount equal to or greater than 6 coffees per day?
 - b. What proportion of students consume an amount equal to or greater than 8.5 coffees per day?
 - c. What proportion of students consume an amount between 6 to 8.5 coffees per day?
 - d. What is the percentile rank for an individual who consumes 5.5 coffee per day?
 - e. How many coffee would an individual at the 75th percentile drink per day?

CRITICAL THINKING CHECK 4

4. Based on what you have learned about z-scores, percentile ranks, and the area under the standard normal curve, fill in the missing information in the following table representing performance on an exam that is normally distributed with $\bar{X} = 55$ and $S = 6$.

	<i>X</i>	<i>z-Score</i>	<i>Percentile Rank</i>
John	63		
Ray		-1.66	
Betty			72

CORRELATION COEFFICIENTS

- **Correlation method** – a method that assesses the degree of relationship between two variables, e.g. weight and height.
- **Correlation coefficient** – a measure of the degree of relationship between two sets of scores. It varies between $[-1, +1]$ i.e., from weak to strong correlation.
- **Pearson product-moment correlation coefficient** - usually referred to as Pearson's r , most commonly used when both variables are measured on an interval or ratio scale.
- **Coefficient of determination (r^2)** – a measure of the proportion of the variance in one variable that is accounted for by another variable; calculated by squaring the correlation coefficient. Example, it tells us how much of the variation in height is accounted for the variation in weight. (r^2) is typically reported as the percentage of the variance in height can be accounted for the variance in weight.

CORRELATION COEFFICIENTS

- **Spearman's rank-order correlation coefficient** – the one used when one or more variables are measured on an ordinal (ranking) scale.
- **Point-biserial correlation coefficient** – the one used when one of the variables is measured on a dichotomous nominal scale (only 2 possible values, e.g. gender) and the other is measured on an interval or ratio scale
- **Phi coefficient** – the one used when both measured variables are dichotomous and nominal.

CORRELATION COEFFICIENTS

IN REVIEW

Correlation Coefficients

TYPES OF COEFFICIENTS

	Pearson	Spearman	Point-Biserial	Phi
Type of data	Both variables must be interval or ratio	Both variables are ordinal (ranked)	One variable is interval or ratio and one variable is nominal and dichotomous	Both variables are nominal and dichotomous
Correlation reported	$\pm 0.0-1.0$	$\pm 0.0-1.0$	$\pm 0.0-1.0$	$\pm 0.0-1.0$
r^2 Applicable?	Yes	Yes	Yes	Yes

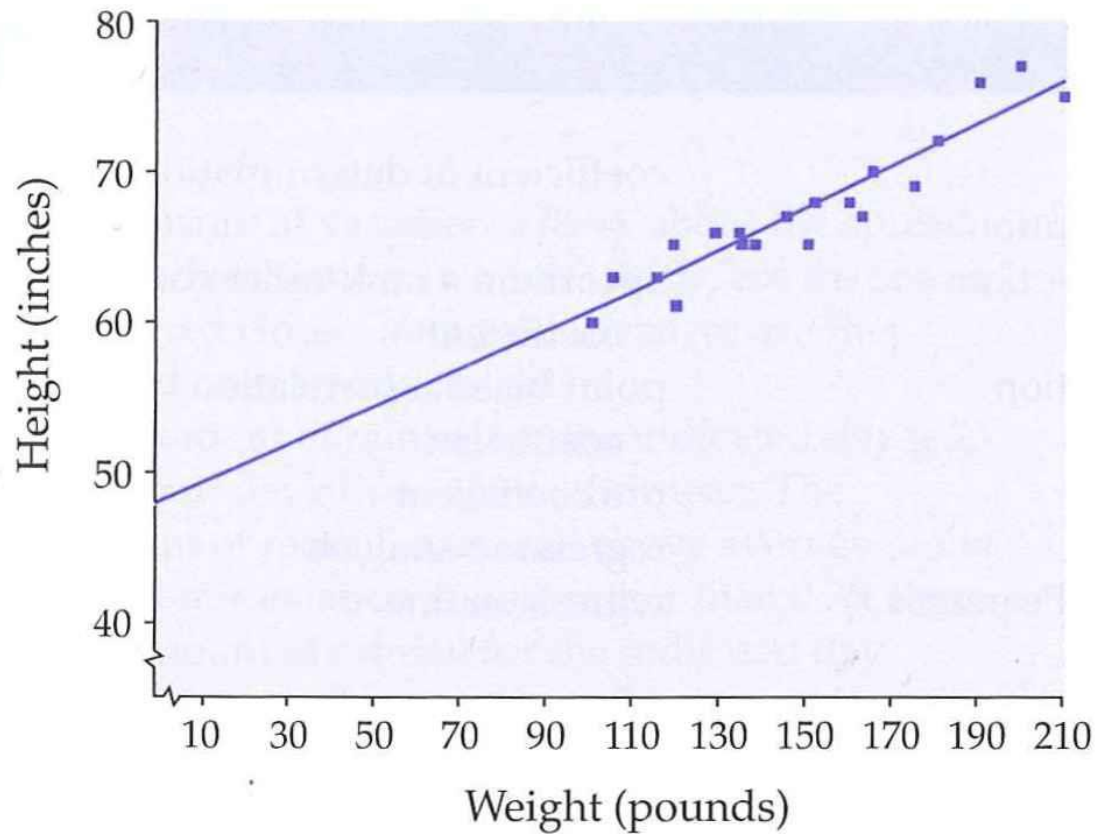
CRITICAL THINKING CHECK 5

1. In a recent study, researchers were interested in determining the relationship between gender and the amount of time spent studying for a group of college students. Which correlation coefficient should be used to assess this relationship?

REGRESSION ANALYSIS

- **Regression analysis** – a procedure that allows to predict an individual score on one variable based on knowing one or more other variables.
 - Regression analysis involves determining the equation for the best fitting line for a data set. This equation is based on the equation for representing a line $y=mx+b$, where m is the slope of the line and b is the y -intercept (the point where the line crosses the y -axis).
 - For a linear regression analysis, the formula is essentially the same: $Y' = bX + a$ where Y' is the predicted value on the Y variable, b is the slope of the line, and a is the y -intercept.
 - **Regression line** – the best fitting straight line drawn through the center of a scatter plot that indicates the relationship between the variables.

REGRESSION ANALYSIS



INFERENCE STATISTICS

Prof. Dr. Vu Duong

*EUROCONTROL Experimental Center &
University of Science - Vietnam National University HCM*

HYPOTHESIS TESTING

- **Hypothesis Testing** – the process of determining whether a hypothesis is supported by the results of a research study.
- **Inferential Statistics** – procedures for drawing conclusions about a population based on data collected from a sample.

EXAMPLE

- To examine the relationship between the type of after-school program attended by a child and the child's intelligence level.
- The researcher is interested in whether students who attend after-school programs that are academically orientated (math, writing, computer use) score higher on an intelligence test than students who do not attend such programs.
- The hypothesis might be that children in academic after-school programs have higher IQ scores than children in the general population.
- Because most intelligence tests are standardized with a mean score (μ) of 100, and a standard deviation (σ) of 15, the students in academic after-school programs must score higher than 100 for the hypothesis to be supported.
 - Two groups of students: Group A (GA), and Group B (GB).
 - Two programs: Program Academic (PA), Program Standard (PS)

NULL & ALTERNATIVE HYPOTHESES

- Statistical techniques are much better at demonstrating that something is not true.
- The logical route is to propose exactly the opposite of what we want to demonstrate to be true, and then disprove or falsify that hypothesis: Null hypothesis.
- In our example:
 - Null hypothesis (H_0) – the hypothesis predicting that no difference exists between the groups being compared.
 $H_0: \mu_{GA} = \mu_{GB}$
 - Alternative hypothesis (H_a) – the hypothesis that the researcher wants to support, predicting that a significant difference exists between the group to be compared. $H_a: \mu_{GA} > \mu_{GB}$
 - When we use inferential statistics, we are trying to reject H_0 , which means that H_a is supported.

ONE- AND 2-TAILED HYPOTHESIS TESTS

- One-tailed hypothesis (directional hypothesis) – an alternative hypothesis in which the researcher predicts the direction of the expected difference between the groups (e.g., $\mu_{GA} > \mu_{GB}$).
 - In this case, the appropriate null hypothesis shall be $H_0: \mu_{GA} \leq \mu_{GB}$.
- Two-tailed hypothesis (non-directional hypothesis) – an alternative hypothesis in which the researcher predicts that the groups being compared differ, but does not predict the direction of the difference.
 - $H_0: \mu_{GA} = \mu_{GB}$
 - $H_a: \mu_{GA} \neq \mu_{GB}$

TYPE I AND II ERRORS

IN HYPOTHESIS TESTING

- **Type I error** – an error in hypothesis testing in which the null hypothesis is rejected when it is true.
- **Type II error** – an error in hypothesis testing in which there is a failure to reject the null hypothesis when it is false.

The Researcher's Decision	THE TRUTH (UNKNOWN TO THE RESEARCHER)	
	H_0 is true	H_0 is false
Reject H_0 (say it is false)	Type I error	Correct decision
Fail to reject H_0 (say it is true)	Correct decision	Type II error

STATISTICAL SIGNIFICANCE AND ERRORS

- **Statistical significance** – an observed difference between two descriptive statistics (such as means) that is unlikely to have occurred by chance.
 - Say a difference is statistically significant at the .05 (or the 5%) level (also known as the .05 alpha level) means that a difference - as large or larger than what we observed- between the sample and the population could have occurred by chance only 5 times or less out of 100.
 - In other words, the likelihood that this result is due to chance is small. The alpha level, is the probability of making a Type I error. Set at .05, i.e. Type I error occurs as often as 5 times out of 100.
 - We can set alpha level to .01 to reduce the risk of producing Type I error, but in doing so, we increase the risk of making Type II error. An example would be the rate of false alarms in a bag scanner: reducing false alarms would increase the rate of omitting desired detected objects.

SUMMARY

Two-tailed or
nondirectional test

An alternative hypothesis stating that a difference is expected between the groups, but there is no prediction as to which group will perform better or worse

The mean of the sample is different from or unequal to the mean of the general population

One-tailed or
directional test

An alternative hypothesis stating that a difference is expected between the groups, and it is expected to occur in a specific direction

The mean of the sample is greater than the mean of the population, or the mean of the sample is less than the mean of the population

Type I error

The error of rejecting H_0 when we should fail to reject it

Equivalent to a “false alarm,” saying that there is a difference between the groups when in reality there is none

Type II error

The error of failing to reject H_0 when we should reject it

Equivalent to a “miss,” saying that there is not a difference between the groups when in reality there is

Statistical significance

When the probability of a Type I error is low (.05 or less)

The difference between the groups is so large that we conclude it is due to something other than chance

CRITICAL THINKING CHECK 6

1. A researcher hypothesizes that children in the South weigh less (because they spend more time outside) than the national average. Identify H_0 and H_a . Is this a one- or two-tailed test?
2. A researcher collects data on children's weights from a random sample of children in the South and concludes that children living there weigh less than the national average. The researcher, however, does not realize that the sample includes many children who are small for their age and that in reality there is no difference in weight between children in the South and the national average. What type of error is the researcher making?
3. If a researcher decides to use the .10 level rather than the conventional .05 level of significance, what type of error is more likely to be made? Why? If the .01 level is used, what type of error is more likely? Why?

DEGREE OF FREEDOM

- Degree of Freedom (DOF) = Freedom to vary.
 - Let's imagine to have two dishes A and B and that we need to place some beans into the dishes, given the condition that the total sum of the beans to be placed into the two dishes is 10.
 - At the beginning of this process we are free to put any number of beans into the first dish (e.g. 4); but then we won't be free to place any number of beans into the dish B; this choice is constrained by the fact that the total sum has to be 10 (necessarily we have 6 beans left). This means that we have only 1 DOF.
 - Now, let's assume that we have four dishes (A, B, C and D), and that we need to fill them with 20 beans. We are free to place any amount of beans into the dishes A, B and C (e.g. 5, 3 and 6), but the number of beans for the dish C is constrained: there are 6 beans available for that dish. In this case we have 3 DOF.

- More Example:
 - Now, let's take a sample of 4 subjects, and consider a situation in which only the number of subjects (4), the scores of three subjects (e.g. 2, 3 and 6) and the sum of the scores (20) are known.
 - The score of the fourth subject is also predictable. This means that this fourth score has no freedom to vary, and that there are 3 DOF.
- The DOF is an important concept to many statistical tests. It is necessary to calculate the appropriate DOF.

PARAMETRIC STATISTICS

- **Parametric test** – a statistical test that involves making assumptions about estimates of population characteristics, or parameters.
 - This assumptions typically involve knowing the mean μ and standard deviation σ of the population and that the distribution is normal.
 - Parametric statistics are generally used with interval or ratio data.
- **Non-parametric test** – a statistical test that does not involve the use of any population parameters – μ and σ are not needed, and the distribution doesn't have to be normal.

INDEPENDENT-GROUPS T-TEST

- Independent-groups t test – a parametric inferential test for comparing sample means of two independent groups of scores.
 - It compares the means of two different samples of participants, and indicate the difference:
 - If they are so similar, we may conclude they are likely coming from the same population,
 - If they are so different, we can conclude they likely represent two different populations.
- Let t_o be the value of the t test, X_A and X_B be the means of the two groups of n_A and n_B participants of which s_A^2 and s_B^2 are the respective variances:



INTERPRETING T-TEST

- Supposing we've got $t_o = 4.92$.
 1. determine degree of freedom (DOF), which for independent-group t-tests are $(n_A - 1) + (n_B - 1) = \text{eg. } 18$.
 2. alternative hypothesis was one-tailed and $\alpha = .05$
 3. look up Table B.2 "Critical Values for the t Distribution" and find critical value $t_c = 1.734$
 4. value t_o falls beyond t_c . Thus the null hypothesis is rejected, and the alternative hypothesis is supported. The result is reported as " $t(18) = 4.92, p < .05$ "

EFFECT SIZE & COHEN'S d

- **Effect size** – the proportion of variance in the dependent variable that is accounted for by the manipulation of the independent variable. It indicates how big a role the conditions of the independent variable play in determining scores on the dependent variable.
 - Thus it is an estimate of the effect of the independent variable, regardless the sample size. The larger it is, the more consistent is the influence of the independent variable. ie, the more knowing the conditions of the independent variable improves the accuracy in predicting the scores on the dependent variable.
- **Cohen's d** – an inferential statistic for measuring effect size.



ASSUMPTIONS OF T-TESTS

- The assumptions of the independent-groups t test are:
 - The data are interval-ratio scale.
 - The underlying distribution are bell-shaped.
 - The observations are dependent.
 - Homogeneity of variance: if we could compute the true variance of the population represented by each sample, the variances in each population would be the same.
- If any of these assumptions is violated, it is appropriate to use another statistic.
 - If the scale of measurement is not interval-ratio or if the distribution is not bell-shaped, then it may be more appropriate to use a non-parametric statistic.
 - If the observations are not independent, then it is appropriate to use a statistic for within- or matched-participants designs.

CRITICAL THINKING CHECK 7

1. How is effect size different from significance level? In other words, how is it possible to have a significant result yet a small effect size?
2. How does increasing the sample size affect a t test? Why does it affect a t test in this manner?
3. How does decreasing variability affect a t test? Why does it affect a t test in this manner?

ANOVA

ANALYSIS OF VARIANCE

BETWEEN- & WITHIN-SUBJECT DESIGN

- The terms «a simple design», mentioning the fact that the researcher had selected two groups of pupils, one receiving the new teaching method and the other one receiving a traditional method. This means, for example, that ten students are allocated to one experimental condition, and ten students to the other condition. In this case we have a *between-subjects design*.
- Another way to tackle the variability of the subjects is assigning the same subjects to different conditions. This is a *within-subjects design*. Also this design implies some problems, mostly related to carry-over or learning effects.

DESIGN OF EXPERIMENTS

When using different groups of subjects it is necessary to bear in mind that there could be a variability linked to the subjects (e.g. some students remarkably good at maths).

In order to deal with this type of problems, all the subjects have to be assigned randomly to each condition.

- ***Random assignment means that every subject must have equal chances to end up in one of the conditions. Random assignment is essential. Following non-random strategies may lead to biased results.***

Counterbalance the order of the treatment.

- ***In our example, the treatment corresponds to the type of teaching method (e.g. treatment A and B = method new and traditional).***
- ***So, having twenty pupils as subjects, half of the subjects should receive first the treatment A and then the treatment B; and the other half, should receive first the treatment B and then the treatment A.***
- ***For more than two condition a Latin Square might be used (i.e. giving a treatment so that the order of presentation is different for each subject).***

MULTI-GROUP EXPERIMENT

- Considering the comparison of performance between a group of students learning Maths after-school (GA); a group learning Physics (GC); and a group learning nothing (GB).
 - Previous t-test is not appropriate because it only compares two groups. If we do three experiments, we need to compare three t tests to determine the differences.
 - The problem is that using multiple tests inflates the Type I error rate, eg. The chance of making a Type I error on one test is .05, the overall chance of making such an error increases as more tests are conducted: $[1 - (1 - \alpha)^c]$ in this case $c=3$, overall alpha is .14

BONFERRONI ADJUSTMENT

- We can adjust the single alpha level so that to achieve an overall alpha level that is acceptable.
 - **Bonferroni adjustment** – setting a more stringent alpha level for multiple tests to minimize Type I errors.
 - However, reducing local alpha level would increase the likelihood of Type II errors.
 - Shall use a single statistical test that compares all groups: e.g. *Analysis of Variance* (ANOVA), Kruskal-Wallis or Chi Square.

TESTS FOR MULTI-GROUPS

- **ANOVA** – an inferential statistical tests for comparing the means of three or more groups by analyzing the variance in a study.
- Non-parametric analyses with ordinal data: the **Kruskal-Wallis** analysis of variance.
- Non-parametric analyses with nominal data: the **chi-square** test.

ONE-WAY RANDOMIZED ANOVA

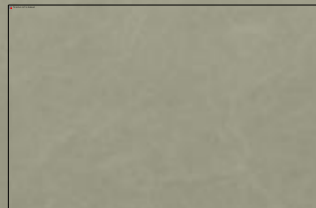
- **One-way randomized ANOVA** – an inferential statistical test for comparing the means of three or more groups using a between-subject design and one independent variable.
 - The term **randomized** indicates that subjects are randomly assigned to conditions;
 - The term **one-way** indicates that the design uses only one independent variable.

MEASUREMENTS IN ANOVA

- **Grand mean** – the mean performance across all subjects in a study. The mean of all the means.
- **Error variance** – the amount of variability among the scores caused by chance or uncontrolled variables (such as individual differences between subjects).
 - Can be estimated by looking at the amount of variability *within* each condition. **Within-group variance** – the variance within each condition; an estimate of the population error.
 - **Between-groups variance** is an estimate of systematic variance – an estimate of the effect of the independent variable *and* error variance.
- **F-ratio** – the ratio of between-group variance to within-group variance.

F - RATIO

- Thus, the F ratio compares the variance between the levels of an independent variable (i.e. difference between groups due the independent variable) and the variance within the levels of that variable (or error, since it cannot be explained).
 - If H_0 is true, the estimated variances between the means are more or less the same (F approximating 1) and thus the independent variable did not produce any effect on the dependent variable.
- In order to calculate the F ratio, for every source of variance (between and within), the Mean Squares (MS) have to be calculated. The MS are obtained dividing the Sum of the Squared Deviates (SS) by the appropriate degree of freedom.



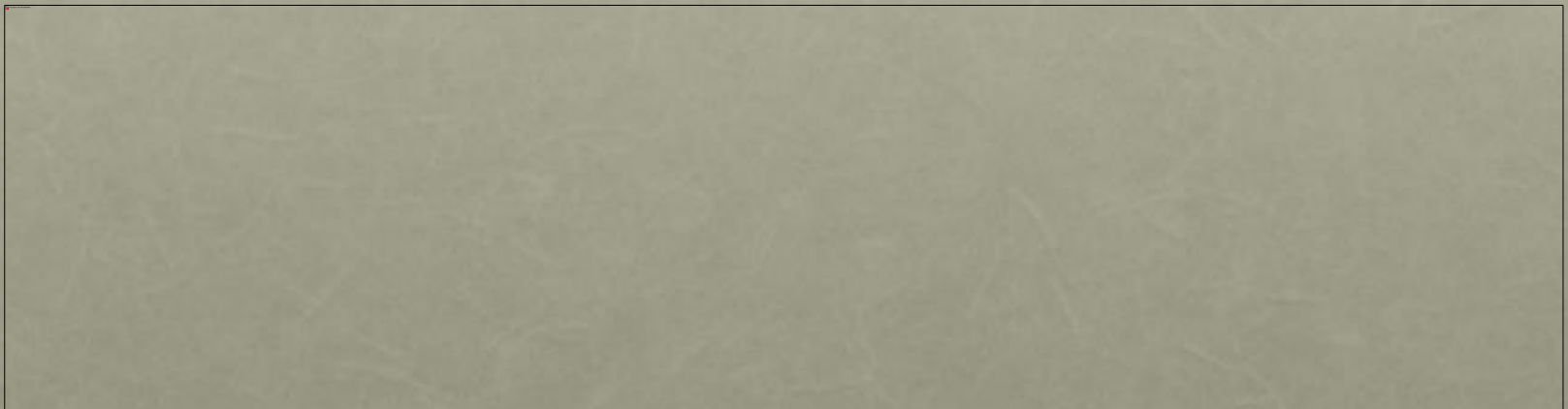
SUM OF SQUARES

- **Total sum of squares (SS)**– the sum of the squared deviations of each score from the grand mean: $SS_{total} = \sum (x_i - X_G)^2$; X_G is the grand mean.
- **Within-groups sum of squares (SS_{wg})** – the sum of the squared deviations of each score from the its group mean. $SS_{wg} = \sum (x_i - X_g)^2$; X_g is the group mean.
- **Between-groups sum of squares (SS_{bg})** – the sum of the squared deviations of each group's mean from the grand mean, multiplied by n , the number of subjects in each group. $SS_{bg} = \sum [(X_g - X_G)^2 n]$;
- **Mean square (MS)** – an estimate of either total variance, variance between group, or variance within groups. MS is obtained by dividing SS to appropriate *df* (degree of freedom).

ONE WAY BETWEEN-SUBJECTS ANOVA

- The One Way between-subjects ANOVA split the variance into two sources: (i) between groups variance, and (ii) within groups variance (error)

Sources of variance	SS	df	MS	F	p (alpha)
Between groups	SS_{bg}	df_{bg}	SS_{bg} / df_{bg}	MS_{bg} / MS_{wg}	...level...
Within groups (or Error)	SS_{wg}	df_{wg}	SS_{wg} / df_{wg}		
Total	SS_{tot}	df_{tot}			



ANOVA SUMMARY TABLE

Appendix: Statistical Supplement: ANOVA summary table using computational formulas

SOURCE	df	SS	MS	F
Between groups	$k - 1$	$\sum \left[\frac{(\sum x_g)^2}{n_g} \right] - \frac{(\sum x)^2}{N}$	$\frac{SS_b}{df_b}$	$\frac{MS_b}{MS_w}$
Within groups	$N - k$	$\sum \left[\sum x_g^2 - \frac{(\sum x_g)^2}{n_g} \right]$	$\frac{SS_w}{df_w}$	
Total	$N - 1$	$\sum x^2 - \frac{(\sum x)^2}{N}$		

INTERPRETING F

- When the F value is obtained, we have to bear in mind two things:
 1. Which *alpha* level was chosen (e.g. .05);
 2. The degrees of freedom of our F (according to the previous example we have $F_{2,27}$, where 2 is df_{bg} and 27 is df_{wg}).
- At this point, given the degrees of freedom 2 and 27, we can compare the calculated F value, with the one reported in the appropriate Statistical Table where the critical values (for any given *alpha* level) are provided.
 - Usually, books of statistical analysis report the Statistical Tables.
- If the calculated F value is equal to or larger than the F value of the table, then the null hypothesis can be rejected and the experimental hypothesis can be accepted.
- But, if the calculated F is smaller than the one indicated in the table, we can and accept the null hypothesis and reject the experimental hypothesis.
 - Checking the tables is somehow old-fashioned nowadays. Statistical packages report the exact *alpha* value for every test performed.

ASSUMPTIONS OF ONE-WAY RANDOMIZED ANOVA

- Similar to those for the t test for independent groups:
 - Data are on an interval-ratio scale;
 - Underlying distribution is normally distributed;
 - Variances among populations being compared are homogeneous.

INTERPRETING ANOVA

- The ANOVA tested a “two-tailed” hypothesis, it checks for overall differences among conditions and gives only a “preliminary” result, which has to be further analysed (i.e. with three conditions, a directional hypothesis is only one among of the possible divergent hypotheses).
- If the ANOVA identified a significant difference among the means of the three groups, then it is possible to proceed to pair-wise comparisons, comparing one by one all the three conditions (e.g. 1 vs. 2; 1 vs. 3; 2 vs. 3), by using other statistical tests (an example of appropriate post-hoc test is the HSD **Tukey**).

If the test allowed discovering a significant difference among conditions, pair-wise comparisons testing directional hypotheses can be carried out.

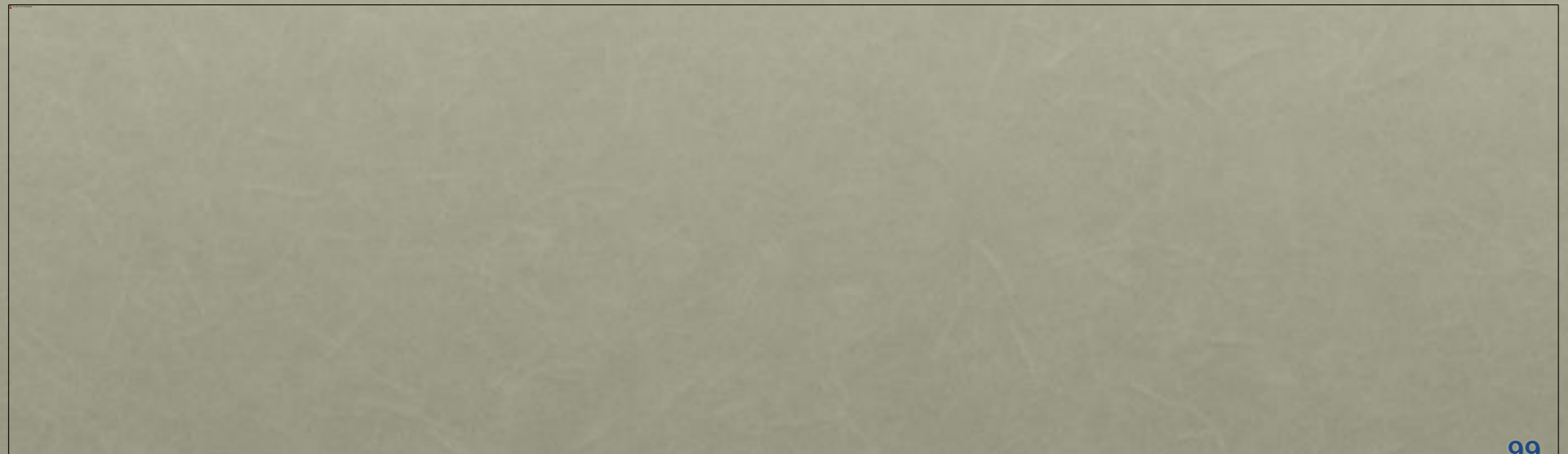
Post-hoc test – when used with an ANOVA, a means of comparing all possible pairs of groups to determine which ones differ significantly from each other.

Tukey’s Honestly Significance Difference – a post-hoc test used with ANOVA for making all pairwise comparisons when conditions have equal n .

ONE WAY WITHIN-SUBJECTS ANOVA

- This type of ANOVA can be used to seek differences between more than two levels of one independent variable administered to the same subjects.
 - The same general logic explained in the previous section applies, since the total variation in the data is attributable to two sources: the differences between the groups and the differences within the groups.
 - However, this last source of variance is further split. More specifically, this type of ANOVA allows distinguishing the part of the SS_{wg} that can be due to individual differences. This part (called SS_{subj}) is removed from the analysis and the remaining SS_{error} is used as the measure of unexplained variation.

CALCULATING



SUMMARY

One-Way Randomized ANOVA

IN REVIEW

CONCEPT	DESCRIPTION
Null hypothesis (H_0)	The independent variable had no effect—the samples all represent the same population.
Alternative hypothesis (H_a)	The independent variable had an effect—at least one of the samples represents a different population than the others.
F -ratio	The ratio is formed when the between-groups variance is divided by the within-groups variance.
Between-groups variance	This estimate of the variance of the group means about the grand mean includes both systematic variance and error variance.
Within-groups variance	This estimate of the variance within each condition in the experiment is also known as error variance, or variance due to chance.
Eta-squared (η^2)	This measure of effect size is the variability in the dependent variable attributable to the independent variable.
Tukey's post hoc test	This test is conducted to determine which conditions in a study with more than two groups differ significantly from each other.

TWO-WAY ANOVA

- Now, let's imagine that we performed an experiment aiming to explore whether there is a difference (for example, in terms of reaction times) between three different interaction metaphors to be used with a three-dimensional environment.
- For the study, a within-subjects design was used, that is, the same subjects performed the tasks with all the three metaphors. The study was exploratory, so we decided to choose an alpha level of .05.
- Then, we carried out the ANOVA to see if there was any significant difference in the performance with the three metaphors, and we obtained the F value.
- The calculated F appeared smaller than the critical F reported in the table, at the level of significance chosen. Thus, we have to reject the experimental hypothesis: no significant difference is found among the three metaphors. Since we cannot accept the experimental hypothesis, more detailed analyses with pair-wise comparisons will not be performed.

NON PARAMETRIC TESTS: THE RANKING

- Non parametric tests allow to perform statistical analyses when the ANOVA conditions are not present.
- One of the differences between parametric and non parametric tests is that non parametric tests compare the subjects' performance not using the original scores but the ranks of the scores.
- For instance, if we have a group of scores, they have to be ranked in order to determine which score is higher or lower. Usually the rank is assigned so that the smallest score goes first (the smallest score is 3, which is ranked 1; the biggest score is 12, which is ranked 7).
 - The ranking is very easy, being all the scores different. But when we have the same value for more than one score (i.e. tied scores), then the average of the ranks have to be assigned. In other words, having three scores of 1 (1, 1, 1), the rank value should be: $1+2+3/3=2$.
- There are different ways to assign ranks depending on the type of design.

- For between-subjects designs the ranks have to be assigned as if they were a single set of ranks.
 - For example, if we compare the scores of two groups, the ranks will be:



- For within-subjects design the way to assign ranks to scores is a bit more complicated and varies depending on the test.

THE KRUSKAL-WALLIS TEST

- This test corresponds to the between-subjects ANOVA previously described and should be used when the requirements for the ANOVA are not met.
 - As an example, we have three groups (A, B, and C), each with 10 controllers who are engaged in a simple task, like rating the easy of use of three types of radar displays (according to a scale so that 1 means “very bad” and 10 means “very good”).
 - Our hypothesis is that there is a difference among the means of the ratings of the displays. The null hypothesis is that there is no difference. For this study we decided to set .05 as the decision criterion.

FRIEDMAN TEST

- This test should be used for within-subjects designs entailing more than two conditions.
 - If the within-subjects ANOVA cannot be used because the requirements for a parametric test are not satisfied (e.g. the scale is not an interval scale, and/or the distribution does not approximate the normal, etc.), the Friedman test is an adequate choice.
- Suppose that we want to compare three interaction metaphors and that we ask to a group of 8 controllers to rate each metaphor using a scale from 1 to 5.
- Since it is a within-subjects design, every controller takes part to three different conditions, one for each metaphor. We decided to set an alpha level of .05.

FREIDMAN TEST

- Also for the Freidman test the original scores should be ranked. But, differently from the Kruskal-Wallis, this time, the scores ranked entail each subject across each condition (to use a “graphical explanation”, the ranking is done “horizontally” for each subject).
- For example, the scores of the subject 1 are ranked (in the order) 1, 3 and 2. The scores of the subject 2 are ranked: 1, 3 and 2, etc.



FREIDMAN TEST



CRITICAL THINKING CHECK 8

1. Of the following four F -ratios, which appears to indicate that the independent variable had an effect on the dependent variable?

1.25/1.11 0.91/1.25 1.95/0.26 0.52/1.01

2. The following ANOVA summary table represents the results from a study of the effects of exercise on stress. There were three conditions in the study: a control group, a moderate exercise group, and a high exercise group. Each group had 10 participants, and the mean stress levels for each group were control = 75.0, moderate exercise = 44.7, and high exercise = 63.7. Stress was measured using a 100-item stress scale, with 100 representing the highest level of stress. Complete the ANOVA summary table, and determine whether the F -ratio is significant. In addition, calculate eta-squared and Tukey's HSD, if necessary.

ANOVA summary table

<i>Source</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Between		4,689.27		
Within		82,604.20		
Total				

Summary

- As a summary to this chapter, the main steps required to start and perform an experiment are:
 1. Formulate the hypothesis
 2. Select the independent and the dependent variables
 3. Check for irrelevant variables
 4. Design the experiment (type of design, counterbalancing, etc.)
 5. Check the scale and think about the possible options of statistical tests
 6. Check the validity
 7. Perform a pilot test
 - Get a “general impression” and perform some checks (e.g. validity; appropriateness of the variables; irrelevant variables; suitability and functioning of the equipments; correctness of the recordings; etc.)
 - If required, refine the experiment and carry out additional pilot tests

Summary (2)

8. Define the α level
9. Collect the participants (at least 10 for each condition)
10. Use the random assignment to allocate the subjects to each condition
11. Counterbalance (when required)
12. Collect the data
13. Analyze the data with descriptive statistics
14. Check the distribution and the homogeneity of variance
15. Perform the statistical analysis as it was planned (using the most appropriate statistical test)
16. Accept / reject the null hypothesis, respecting the chosen α level

Summary (3)

- The last important step is the analysis of the results. When reporting the results of the experiment, the researcher should explain what happened and why, that is, finding an explanation of the results; eventually, also comparing the findings with the results found in the literature.
- Moreover, any other possible explanation for the results should be explored and reported.
- As a matter of fact, sometimes, an experiment is not sufficient to address completely the topics that it aimed to tackle and definitive conclusions cannot be drawn. The analysis of the results can provide some hints and ideas to perform a more refined experiment, to test new hypotheses, or to address the problems according to a different point of view.

Recommended Bibliography

- Baird, D.C. (1988) *Experimentation : An Introduction to Measurement Theory and Experimental Design*, 2nd Ed., Prentice Halls.
- Brodsky, B. E., Darkhovsky, B. S., (2000), *Non parametric statistical diagnosis*, Kluwer Academic Publishers, The Netherlands.
- Das, M. et Giri,(1989) N. *Design and Analysis of Experiments*, 2nd Ed., John Wiley & Sons..
- Dean, A., Voss, D., (2000), *Design and analysis of experiments*, Springer, NY.
- Frigon, L., Mathews, D., (1997), *A practical guide to experimental design*, John Wiley and Sons Inc., NY.
- Gooding, D., Pinch, T., et Schaffer, S. (Editors) (1989). *The Use of Experiment*. Cambridge University Press.
- Hicks, C. R., Turner, K. V., (1999), *Fundamental concepts in the design of experiments* Fifth Edition, Oxford University Press, NY
- Myers, J. L. (1966). *Fundamentals of Experimental Design*. Allyn & Bacon.
- Scheaffer, R. et al. (1986). *Elementary Survey Sampling*, 3rd Ed., Duxbury Press.
- Wilson, E. Bright, Jr. (1990). *An Introduction to Scientific Research*. Dover (McGraw-Hill 1952, 1980).
- Wickens, C. D., Gordon, S. E., Liu, Y., (1997), *An introduction to human factors engineering*, Addison Wesley, NY.