# Biostatistics
# Unit Five

Dr. Mahmoud Alhussami

# Proportions and Probabilities

- We often interpret proportions as probabilities. If the **proportion** with a disease is 1/10 then we also say that the **probability** of getting the disease is 1/10, or 1 in 10.
- Proportions are usually quoted for **samples** - probabilities are almost always quoted for **populations.**

# Workers Example

| Smoking | Workers | Cases | Controls |
|---|---|---|---|
| No | Yes | 11 | 35 |
| | No | 50 | 203 |
| Yes | Yes | 84 | 45 |
| | No | 313 | 270 |

- For the cases:
  - Proportion of exposure=84/397=0.212 or 21.2%
- For the controls:
  - Proportion of exposure=45/315=0.143 or 14.3%

# Prevalence

Disease Prevalence = the proportion of people with a given disease at a given time.

disease prevalence =

$$\frac{\text{Number of diseased persons at a given time}}{\text{Total number of persons examined at that time}}$$

Prevalence is usually quoted as per 100,000 people so the above proportion should be multiplied by 100,000.

# Interpretation

$$\Pr evalence = \frac{Cases(old + new)}{Total}$$

Old = duration of the disease

New = speed of the disease

# Sensitivity and Specificity

- Sensitivity and specificity are terms used to describe the effectiveness of screening tests. They describe how good a test is in two ways - finding false positives and finding false negatives

- **Sensitivity** is the Proportion of diseased who screen positive for the disease

- **Specificity** is the Proportion of healthy who screen healthy

# Sensitivity and Specificity

| | Condition Present | Condition Absent |
|---|---|---|
| Test Positive | True Positive (TP) | False Positive (FP) |
| Test Negative | False Negative (FN) | True Negative (TN) |

- Test Sensitivity (Sn) is defined as the probability that the test is positive when given to a group of patients who have the disease.
  - Sn= (TP/(TP+FN))x100.
  - It can be viewed as, 1-the false negative rate.
- The Specificity (Sp) of a screening test is defined as the probability that the test will be negative among patients who do not have the disease.
  - Sp = (TN/(TN+FP))X100.
  - It can be understood as 1-the false positive rate.

# Positive & Negative Predictive Values

- The positive predictive value (PPV) of a test is the probability that a patient who tested positive for the disease actually has the disease. PPV = (TP/(TP+FP))X 100.
- The negative predictive value (NPV) of a test is the probability that a patient who tested negative for a disease will not have the disease. NPV = (TN/(TN+FN))X100.

# The Efficiency

- The efficiency (EFF) of a test is the probability that the test result and the diagnosis agree.
- It is calculated as:

$$\mathrm{EFF} = ((\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}))\ \mathrm{X}\ 100$$

# Example

- A cytological test was undertaken to screen women for cervical cancer.

| | Test Positive | Test Negative | Total |
|---|---|---|---|
| Actually Positive | 154 (TP) | 225 (FP) | 379 |
| Actually Negative | 362 (FN)<br>516 (TP+FN) | 23,362 (TN)<br>23587(FP+TN) | 23,724 |

- Sensitivity =?
- Specificity = ?

# Relative Risk

- **Relative risks** are the ratio of risks for two different populations (ratio=a/b).

$$\text{Relative Risk} = \frac{\text{disease incidence in group 1}}{\text{disease incidence in group 2}}$$

- If the risk (or proportion) of having the outcome is 2/10 in one population and 1/10 in a second population, then the relative risk is: (2/10) / (1/10) = 2.0

- A relative risk >1 indicates increased risk for the group in the numerator and a relative risk <1 indicates decreased risk for the group in the numerator.

# Odd’s Ratio and Relative Risk

- **Odds ratios** are better to use in case-control studies (cases and controls are selected and level of exposure is determined retrospectively)

- **Relative risks** are better for cohort studies (exposed and unexposed subjects are chosen and are followed to determine disease status - prospective)

# Odd’s Ratio and Relative Risk

- When we have a two-way classification of exposure and disease we can approximate the relative risk by the odds ratio

| | | Disease | | |
|---|---|---|---|---|
| Exposure | | Yes | No | |
| | Yes | A | B | A+B |
| | No | C | D | C+D |

- Relative Risk=A/(A+B) divided by C/(C+D)
- Odd’s Ratio= A/B divided by C/D = AD/BC

# Case Control Study Example

- Disease: Pancreatic Cancer
- Exposure: Cigarette Smoking

| | | | | |
|---|---|---|---|---|
| Exposure | Disease | | | |
| | | Yes | No | |
| | Yes | 38 | 81 | 119 |
| | No | 2 | 56 | 58 |

- Relative Risk= (38/119)/(2/58)=9.26
- Odd's Ratio= (38/81)/(2/56)=(38*56)/(2*81) =13.14

# Criteria for Selection of a Data-Collection Instrument

- Practicality of instrument: cost and appropriateness for the study population.
- Reliability: consistency and stability, measured by the use of corelational procedures: correlation coefficient (0 and 1.0) between two sets of scores or between the ratings of two judges.
- Validity. The degree to which an instrument measures what it is supposed to measure.

# Reliability

- The consistency and accuracy with which an instrument measures an attribute
- Reliability assessments involve computing a reliability coefficient
  - Most reliability coefficients are based on correlation coefficients

# Three Aspects of Reliability Can Be Evaluated

- Stability: extent to which an instrument yields the same results on repeated administrations.
- Internal consistency: extent to which all the instrument’s items are measuring the same attribute.
- Equivalence: estimates of interrater or interobserver reliability are obtained.

# Stability

- The extent to which scores are similar on two separate administrations of an instrument
- Evaluated by test–retest reliability:
  - Requires participants to complete the same instrument on two occasions
  - A correlation coefficient between scores on first and second administration is computed
  - Appropriate for relatively enduring (stable) attributes (e.g., self-esteem)

# Internal Consistency

- The extent to which all the instrument’s items are measuring the same attribute
- Evaluated by administering instrument on one occasion
- Appropriate for most multi-item instruments
- Evaluation methods:
  - Split-half technique
  - Coefficient alpha.
  - Example: all items measure depression if one item measuring guilt then it is not internally consistent.

# Equivalence

- The degree of similarity between alternative forms of an instrument or between multiple raters/observers using an instrument.
- Inter-rater/inter-observer reliability:
  - **Inter-rater - consistency of 2 raters performance (.90).**
  - **Intra-rater - consistency of 1 rater's performance (.90).**
- **Alternate forms (parallel forms) - construct 2 tools using the same outcomes, administer both tools to same group of subjects on same day and test for significant difference in scores**
- Most relevant for structured observations
- Assessed by comparing observations or ratings of two or more observers (interobserver/interrater reliability)
- Small number of categories is desired, the kappa statistic is often used (a metric that compares an **Observed Accuracy** with an **Expected Accuracy,** random chance).
  - **Example: using two different forms of questions**

# Reliability Coefficients

- Represent the proportion of true variability to obtained variability:

$$r = \frac{V_T}{V_o}$$

- Should be at least .70; .80 preferable
- Can be improved by making instrument longer (adding items)

# Validity of the instrument

- **The degree to which an instrument measures what it is supposed to be measuring.**
- **The greater the validity of an instrument the more confidence one can have that the instrument will obtain data that will answer the research questions or test the research hypotheses.**

# Types of validity

- **Face validity.**
- **Content validity.**
- **Criterion validity.**
- **Construct validity.**

# Face validity

- A brief and hasty examination of an instrument.
- Refers to whether the instrument looks as though it is measuring the appropriate construct.
- Based on judgment of experts in the content area, no objective criteria for assessment.

# Content validity

- The degree to which an instrument has an appropriate sample of items for the construct being measured.
- Concerned with the scope or range of items used to measure the variable, i.e. number and type of items to measure the concept.
- Evaluated by expert evaluation, via the content validity index (CVI)

# Criterion validity

- The degree to which the instrument correlates with an external criterion
- Validity coefficient is calculated by correlating scores on the instrument and the criterion
- Two types of criterion-related validity :
  - Concurrent
  - Predictive

# Construct validity

- The degree to which an instrument measures the construct that is supposed to measure.
- Concerned with the questions:
  - What is this instrument really measuring?
  - Does it adequately measure the construct of interest?

# Methods of Assessing Construct Validity

- Known-groups technique
- Relationships based on theoretical predictions
- Multitrait–multimethod matrix method (MTMM)
- Factor analysis

# Known-Groups Technique

- Assesses contrast validity.
- In this procedure, the instrument is administered to groups hypothesized to differ on the critical attribute because of a known characteristic.
- It is a method to support construct validity and provided when a test can discriminate between a group of individuals known to have a particular trait and a group who do not have the trait.
- Assess controlled versus uncontrolled blood pressure.

# Relationships based on Theoretical Predictions

- It involves testing hypothesized relationships on the basis of theory or prior research.
- A researcher might reason as
  - According to the theory, construct X is positively related to construct Y.
  - Instrument A is a measure of construct X; instrument B is a measure of construct Y.
  - Scores on A & B are correlated positively, as predicted.
  - Therefore, it is inferred that  A & B are valid measure of X & Y.

# Multitrait–Multimethod Matrix Method

Builds on two types of evidence:

- Convergence
- Discriminability (Divergent)

# Convergence

- Evidence that different methods of measuring a construct yield similar results
- Convergent validity comes from the correlations between two different methods measuring the same trait

# Discriminabililty

- Evidence that the construct can be differentiated from other similar constructs

- Discriminant validity assesses the degree to which a single method of measuring two constructs yields different results

# Statistical Inference involves:

- Estimation
- Hypothesis Testing

Both activities use sample statistics (for example, $\overline{X}$) to make inferences about a population parameter (μ).

# Estimation

- Estimation can take two forms:
  - Point estimation: involves calculating a single statistic to estimate the parameter. E.g. mean and median.
    - Disadvantages: they offer no context for interpreting their accuracy and a point estimate gives no information regarding the probability that it is correct or close to the population value.
  - Interval estimation: is to estimate a range of values that has a high probability of containing the population value .

# Interval Estimation

- For example, it is more likely the population height mean lies between 165-175cm.
- Interval estimation involves constructing a confidence interval (CI) around the point estimate.
- The upper and lower limits of the CI are called confidence limits.
- A CI around a sample mean communicates a range of values for the population value, and the probability of being right. That is, the estimate is made with a certain degree of confidence of capturing the parameter.

# Confidence Intervals around a Mean

- 95% CI = (mean $\underline{+}$ (1.96 x SEM)
- This statement indicates that we can be 95% confident that the population mean lies between the confident limits , and that these limits are equal to 1.96 times the true standard error, above and below the sample mean.
- E.g. if the mean = 61 inches, and SEM = 1, What is 95% CI.
  - Solution: 95% CI = (61 $\underline{+}$ (1.96 X 1))
    95% CI = (61 $\underline{+}$ 1.96)
    95% CI = $59.04 \leq \mu \leq 62.96$
- E.g. if the mean = 61 inches, and SEM = 1, What is 99% CI.
  - Solution: 99% CI = (61 $\underline{+}$ (2.58 X 1))
    99% CI = (61 $\underline{+}$ 2.58)
    99% CI = $58.42 \leq \mu \leq 63.58$

# Types of Statistical Inference

- Hypothesis testing:
  - Hypothesis testing is a second approach to inferential statistics.
  - Hypothesis testing involves using sampling distributions and the laws of probability to make an objective decision about whether to accept or reject the null hypothesis.
  - The sample may deviate from the defined population’s true nature by certain amount.
  - This deviation is called sampling error.
  - Drawing the wrong conclusion is called an error of inference.
  - There are two types of errors of inference defined in terms of the null hypothesis:
    - Type I error
    - Type II error

# Sample Size Determination

- Is the act of choosing the number of observations or replicates to include in a statistical sample.
- The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample.

# Sample Size

How large should a sample Be?

- Factors to be considered in deciding the size of the sample:
  - Homogeneity of the population
  - The degree of precision desired by the researcher
  - The type of sampling procedure that will be used

# Sample Size

- Large sample sizes may be needed in the following instances:
  - Many uncontrolled variables are present. i.e., inability to control for age
  - Small differences are expected in members of the population on the variable of interest
  - The population must be divided into subgroups
  - Dropout rate among subjects is expected to be high
  - Statistical test are used that require minimum sample sizes.

# Sampling Error and Sampling Bias:

- Sampling error: the difference between data obtained from a random sample and the data obtained that would be obtained if an entire population were measured.
- Error is not under the researcher’s control and caused by chance

- Sampling bias: is the bias that is caused by the researcher when the samples are not carefully selected (not a matter of chance)
- Example: selection from the telephone directory but this record has some people missing form the register for some reasons

# Sample Size

- Before using the sample size calculator, there are two terms that you need to know:
  - **Confidence Interval**
  - **Confidence Level**
- **Confidence interval** (also called margin of error) is the plus-or-minus figure usually reported. E.G. 10, 20, 30.
- The **Confidence level** tells you how sure you can be. It is expressed as a percentage and represents how often the true percentage of the population who would pick an answer lies within the confidence interval. E.G. 95%, 99%

# Example

- If you asked a sample of 1000 people in a city which brand of cola they preferred, and 60% said Brand A, you can be very certain that between 40 and 80% of all the people in the city actually do prefer that brand, but you cannot be so sure that between 59 and 61% of the people in the city prefer the brand.
- The wider the confidence interval you are willing to accept, the more certain you can be that the whole population answers would be within that range.

# Factors that Affect Confidence Intervals

- Sample size: The larger your sample size, the more sure you can be that their answers truly reflect the population. This indicates that for a given confidence level, the larger your sample size, the smaller your confidence interval.
- Percentage: Your accuracy also depends on the percentage of your sample that picks a particular answer. If 99% of your sample said "Yes" and 1% said "No," the chances of error are remote, irrespective of sample size. However, if the percentages are 51% and 49% the chances of error are much greater.
- Population size

# Sample Size

Too Big:

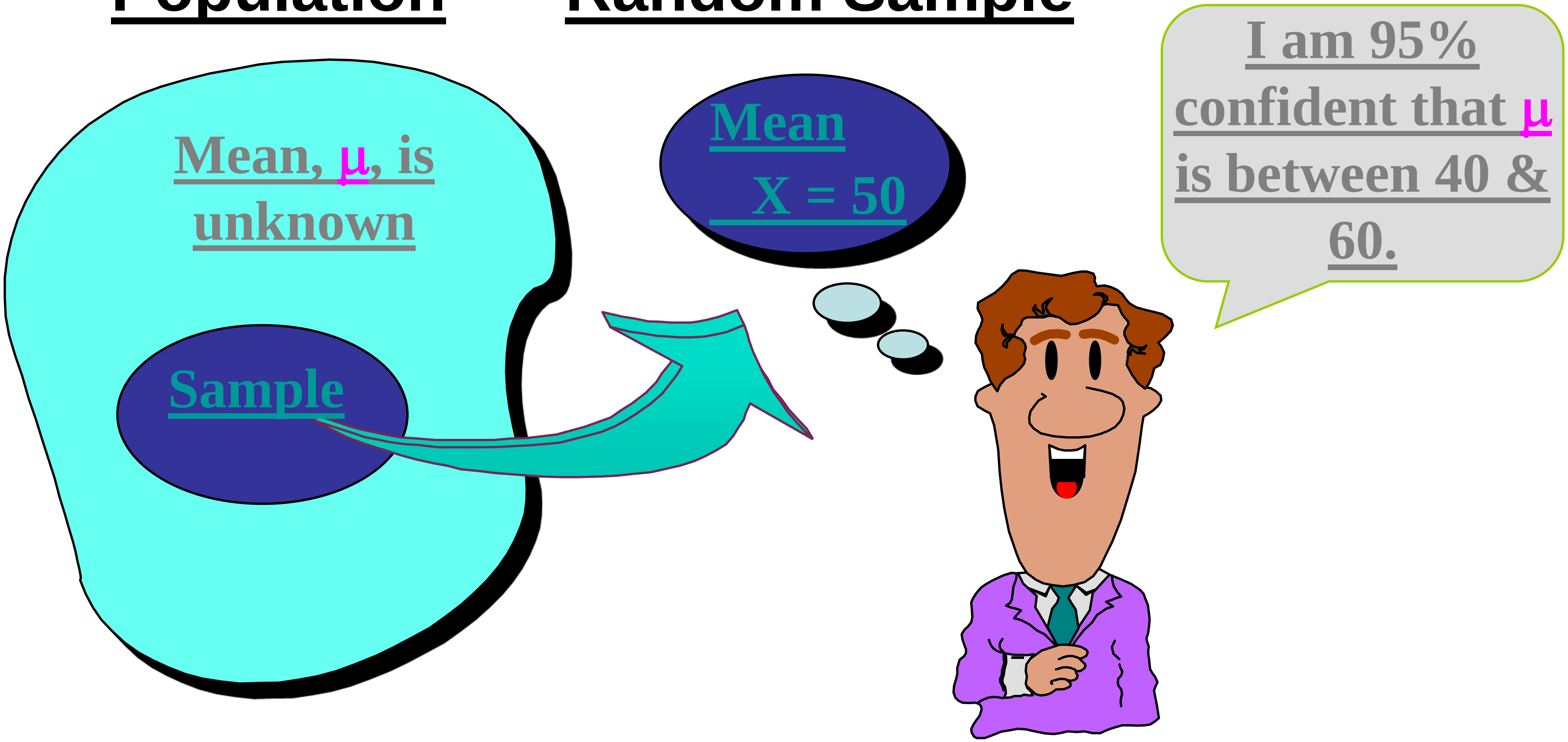
- Requires too much resources

Too Small:

- Won’t do the job

# Estimation Process

**Population** **Random Sample**



Mean, μ, is unknown

Sample

Mean $\bar{X} = 50$

I am 95% confident that μ is between 40 & 60.

# Methods of Sample Size Determinattion

- **Estimation of means.**
- **Estimation of Proportions.**
- **Power Tables for Effect Size**
- **Power of a Statistical Test (**G* power)

# Example: Sample Size for Mean

- **What sample size is needed to be 90% confident of being correct within ± 5? A pilot study suggested that the standard deviation is 45.**

$$n = \frac{Z^2\sigma^2}{\text{Error}^2} = \frac{1645^2 \; 45^2}{5^2} = 219.2 \cong 220$$

**Round Up**

# Example: Sample Size for Proportion

- What sample size is needed to be within ± 5 with 90% confidence? Out of a population of 1,000, we randomly selected 100 of which 30 were defective.

$$n = \frac{Z^2 p(1-p)}{error^2} = \frac{1.645^2(.30)(.70)}{.05^2} = 227.3$$

$\cong$ 228

Round Up

# Power Tables for Effect Size

- **Power Tables for Effect Size d** (from Cohen 1988, pg. 55). Cohen's *d* is defined as the difference between two means divided by a standard deviation for the data.
- **Power Tables for Effect Size r** (from Cohen 1988, pg. 102). is a statistical concept that measures the strength of the relationship between two variables on a numeric scale.
- Power_Tables.pdf

# G*POWER

- G*POWER is a FREE program that can make the calculations a lot easier

  http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

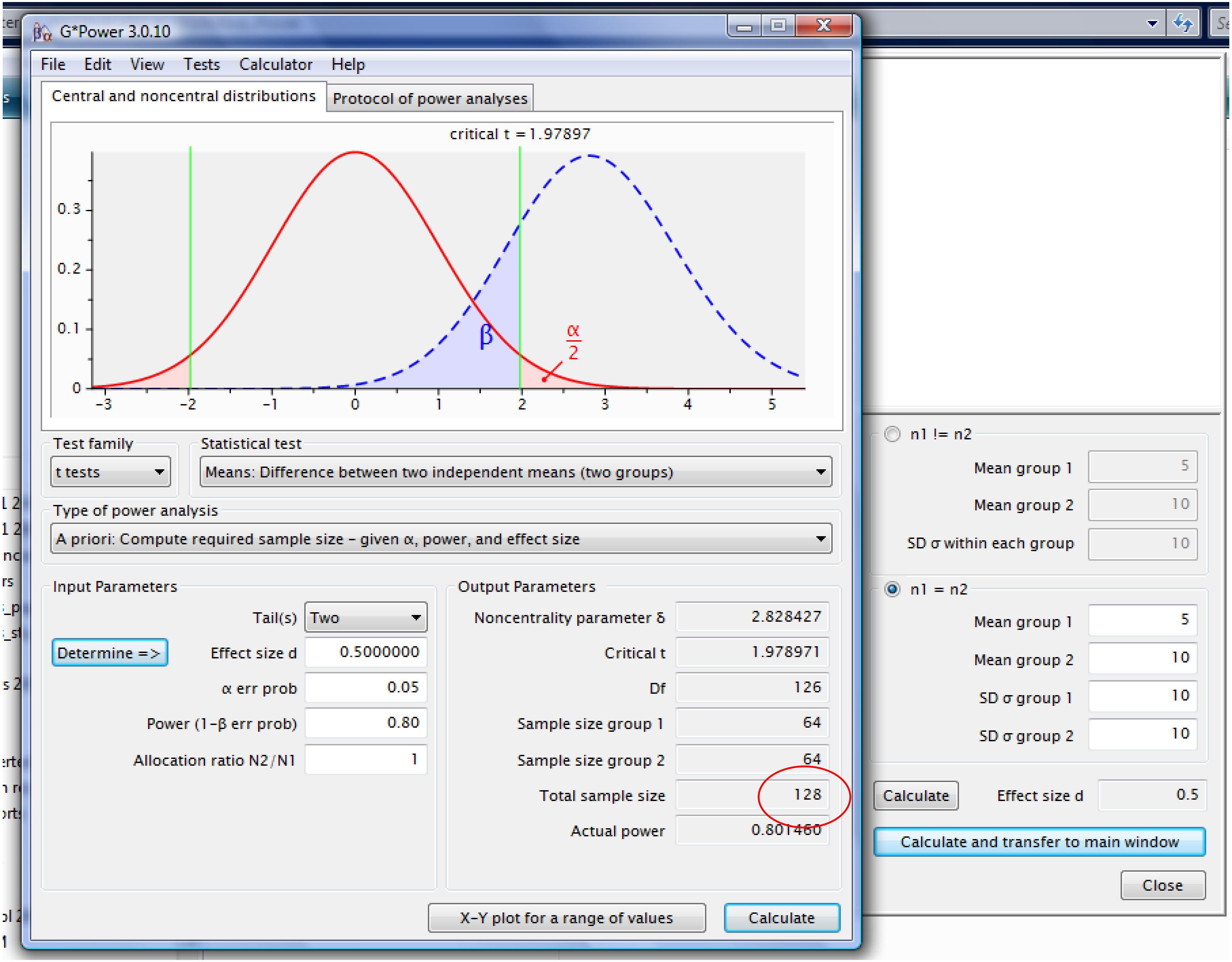
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

G*Power computes:

- power values for given sample sizes, effect sizes, and alpha levels,
- sample sizes for given effect sizes, alpha levels, and power values
- suitable for most fundamental statistical methods
- GPower 3.1.lnk

# power

- **power** is:
  - the probability of correctly rejecting a ***false*** null hypothesis
  - the probability that the study will yield significant results ***if the research hypothesis is true***
  - the probability of ***correctly identifying a true*** alternative hypothesis


G*Power 3.0.10
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
critical t = 1.97897
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
4
5
β
α/2
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

| | |
|---|---|
| Tail(s) | Two |
| Effect size d | 0.5000000 |
| α err prob | 0.05 |
| Power (1-β err prob) | 0.80 |
| Allocation ratio N2/N1 | 1 |

Output Parameters

| | |
|---|---|
| Noncentrality parameter δ | 2.828427 |
| Critical t | 1.978971 |
| Df | 126 |
| Sample size group 1 | 64 |
| Sample size group 2 | 64 |
| Total sample size | 128 |
| Actual power | 0.801460 |

○ n1 != n2

| | |
|---|---|
| Mean group 1 | 5 |
| Mean group 2 | 10 |
| SD σ within each group | 10 |

● n1 = n2

| | |
|---|---|
| Mean group 1 | 5 |
| Mean group 2 | 10 |
| SD σ group 1 | 10 |
| SD σ group 2 | 10 |

Effect size d 0.5

# Power of a Statistical Test

- The result of using computer program G* power (Faul et al., 2007) showed that the required sample size was 128 participants. This figure was arrived at by using compromised $\beta=0.80$, $\alpha = 0.05$ (2-tailed) and medium effect size.