Kako odabrati pravi test za procjenu statističke značajnosti razlike između skupina?

Comparing groups for statistical differences: how to choose the right statistical test?

Marius Marusteri[1*], Vladimir Bacarea[2]

[1]Medical Informatics and Biostatistics Department, Academical Management & European Integration Department, University of Medicine and Pharmacy Targu Mures, Romania

[2]Research Methodology Department, University of Medicine and Pharmacy Targu Mures, Romania

Corresponding author*: msmarusteri@yahoo.com

Sažetak

Odabir pravog statističkog testa može ponekad predstavljati veliki izazov za početnika na polju biostatistike.

Ovaj članak opisuje postupak odabira testa za procjenu statističke značajnosti razlike između dvije ili više skupina. Potrebno je, primjerice, znati kojim tipom podataka raspolažemo (nominalni, ordinalni, intervalni/omjerni), kako su podaci organizirani, koliko je uzoraka/skupina i radi li se o zavisnim ili nezavisnim uzorcima. Također treba znati slijede li podaci iz populacije Gaussovu raspodjelu ili ne. Ključno je pitanje treba li se u slučaju kad su ispunjeni svi uvjeti primijeniti jednosmjerni ili dvosmjerni test, pri čemu je bitno napomenuti da drugi test ima jaču statističku snagu.

Ispravan pristup postupku odabira testa prikazan je u obliku pitanja i odgovora, kako bi se korisniku pružilo bolje razumijevanje osnovnog koncepta. Neki od neophodnih osnovnih koncepata su: statističko zaključivanje, statističko ispitivanje hipoteze, koraci neophodni za primjenu statističkog testa, parametrijski testovi nasuprot neparametrijskim te jednosmjerni nasuprot dvosmjernim testovima itd.

U završnom dijelu članka predložit ćemo algoritam za odabir testa, koji se temelji na ispravnom postupniku za izbor statističkog testa u svrhu statističke usporedbe jedne, dviju ili više skupina, kako bi pokazali praktičnu primjenu osnovnih koncepata.

Za neki drugi članak ostavit ćemo neke vrlo osporavane koncepte kao što su izrazito visoke ili izrazito niske vrijednosti i njihov utjecaj u statističkoj analizi, utjecaj vbrijednosti koje nedostaju itd.

Ključne riječi: statističko zaključivanje; statističko ispitivanje hipoteze; odabir statističkog testa

Abstract:

Choosing the right statistical test may at times, be a very challenging task for a beginner in the field of biostatistics.

This article will present a step by step guide about the test selection process used to compare two or more groups for statistical differences. We will need to know, for example, the type (nominal, ordinal, interval/ratio) of data we have, how the data are organized, how many sample/groups we have to deal with and if they are paired or unpaired. Also, we have to ask ourselves if the data are drawn from a Gaussian on non-Gaussian population. A key question is, if the proper conditions are met, should a one-tailed test or two-tailed test be used, the latter typically being the most powerful choice.

The appropriate approach is presented in a Q/A (Question/Answer) manner to provide to the user an easier understanding of the basic concepts necessary to fulfill this task. Some of the necessary fundamental concepts are: statistical inference, statistical hypothesis tests, the steps required to apply a statistical test, parametric versus nonparametric tests, one tailed versus two tailed tests etc.

In the final part of the article, a test selection algorithm will be proposed, based on a proper statistical decision-tree for the statistical comparison of one, two or more groups, for the purpose of demonstrating the practical application of the fundamental concepts.

Some much disputed concepts will remain to be discussed in other future articles, such as outliers and their influence in statistical analysis, the impact of missing data and so on.

Keywords: statistical inference; statistical hypothesis testing; statistical tests selection

Pristiglo: 31. listopada 2009. *Received: October 31, 2009*

Prihvaćeno: 30. studenog 2009 *Accepted: November 30, 2009*

Uvod

Kako bi se odabrao ispravan statistički test pri analizi podataka, potrebno je barem:

- dobro poznavati osnovne statističke termine i koncepte;
- poznavati nekoliko aspekata povezanih s podacima sakupljenim tijekom istraživanja (npr. tip podataka – nominalni, ordinalni, intervalni ili omjerni, kako su podaci organizirani, koliko ima skupina (obično su to ispitivana i kontrolna skupina), jesu li skupine uparene ili nisu (zavisni/nezavisni uzorci) te slijede li podaci iz uzoraka ili uzoraka populacije normalnu raspodjelu);
- dobro razumjeti cilj statističke analize;
- raščlaniti čitav statistički postupak na dobro strukturirani postupnik za odabir ispravnog testa koji slijedi algoritamski način, kako bi se izbjegle neke pogreške.

Pitanja i odgovori koji slijede prikazat će, korak po korak, termine i koncepte potrebne za ispravan odabir statističkog testa.

Pitanje 1: Koji su osnovni termini i koncepti potrebni?

Odgovor 1: Zaključivanje iz premise je čin ili proces izvođenja logičnog zaključka o posljedici iz premise.

Statističko zaključivanje iz premise ili statističko izvođenje zaključka obuhvaća primjenu statistike i (slučajnog) uzorkovanja, kako bi se donio zaključak o nekom nepoznatom aspektu neke statističke populacije (1,2).

Tu vrstu statistike treba razlikovati od deskriptivnih statističkih testova (3) kojima se opisuju glavna svojstva skupine podataka u kvantitativnom smislu (npr. primjena mjera središnjice kao što su srednja vrijednost, medijan, mod ili pokazatelja rasapa kao što su varijanca, standardna devijacija itd). Deskriptivni se statistički testovi razlikuju od inferencijskih/induktivnih statističkih testova po tome što za cilj imaju kvantitativno sažeti skup podataka, a ne donijeti informaciju o populaciji koju predstavljaju.

Uporabom inferencijskih statističkih testova pokušavamo izvući zaključak o populaciji iz njenog (slučajnog) uzorka ili govoreći općenitije, o nekom njenom slučajnom procesu tijekom određenog vremena, kao što se može vidjeti i na sljedećoj slici (Slika 1.).

Statističko zaključivanje iz premise može uključivati (3,4):

1. **Procjenu vrijednosti**, zajedno s korištenjem podataka iz uzorka kako bi se izračunala pojedinačna vrijednost koja treba služiti kao najbolja procjena za nepoznati (fiksni ili slučajni) parametar populacije (npr. relativni rizik (engl. *relative risk*, RR) = 3,72).
2. **Procjenu intervala** – primjena podataka iz uzorka za izračun intervala mogućih (i vjerojatnih) vrijednosti nekog nepoznatog parametra, suprotno procjeni vrijednosti, za koju je rezultat jedan broj (npr. 95%-tni in-

Introduction

In order to choose the right statistical test, when analyzing the data from an experiment, we must have at least:

- a decent understanding of some basic statistical terms and concepts;
- some knowledge about few aspects related to the data we collected during the research/experiment (e.g. what types of data we have - nominal, ordinal, interval or ratio, how the data are organized, how many study groups (usually experimental and control at least) we have, are the groups paired or unpaired, and are the sample(s) extracted from a normally distributed/ Gaussian population);
- a good understanding of the goal of our statistical analysis;
- we have to parse the entire statistical protocol in an well structured - decision tree /algorithmic manner, in order to avoid some mistakes.

The following questions and answers, will present, step by step, the terms and concepts necessary to realize this goal.

Question 1: What are the required basic terms and concepts?

Answer 1: Inference is the act or process of deriving a logical consequence conclusion from premises.

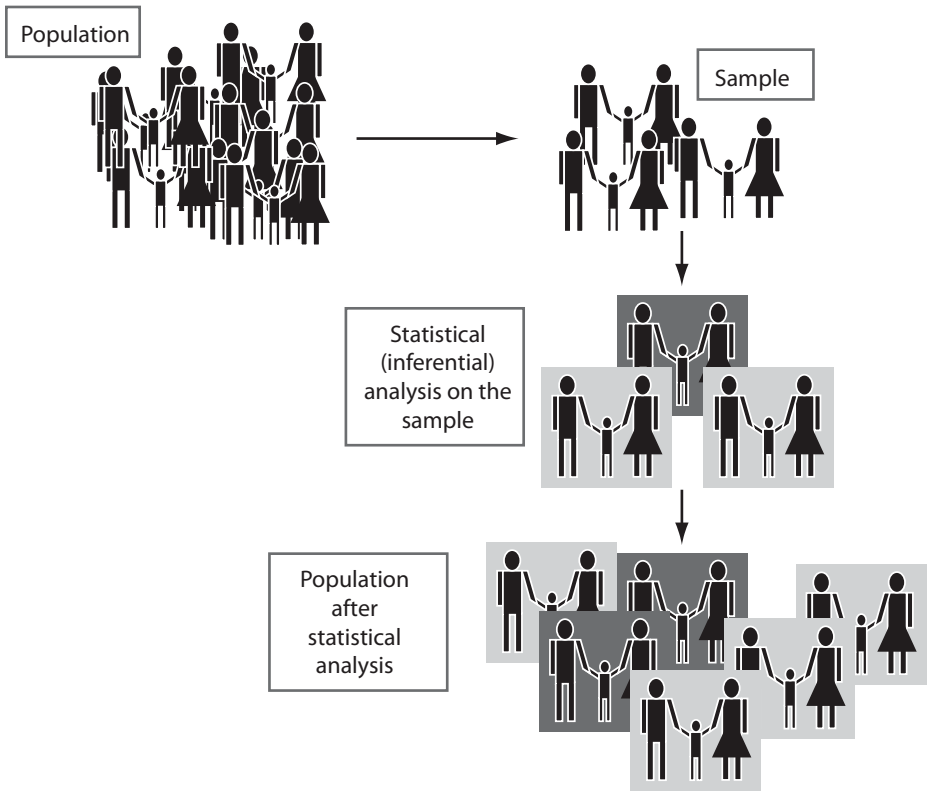
Statistical inference or statistical induction comprises the use of statistics and (random) sampling to make inferences concerning some unknown aspect of a statistical population (1,2).

It should be differentiated from descriptive statistics (3), which is used to describe the main features of data in quantitative terms (e.g. using central tendency indicators for the data – such as mean, median, mode or indicators of dispersion - sample variance, standard deviation etc). Thus, the aim of descriptive statistics is to quantitatively summarize a data set, opposed to inferential/inductive statistics, which is being used to support statements about the population that the data are thought to represent.

By using inferential statistics, we try to make inference about a population from a (random) sample drawn from it or, more generally, about a random process from its observed behavior during a finite period of time, as it can be seen in the following figure (Figure 1).

Statistical inference may include (3, 4):

1. **Point estimation**, involving the use of sample data to calculate a single value (also known as a statistic), which is to serve as a "best guess" for an unknown (fixed or random) population parameter (e.g. relative risk RR = 3.72).
2. **Interval estimation** is the use of sample data to calculate an interval of possible (or probable) values of



Slika 1. Primjena statističke analize na uzorku/uzorcima kako bi se donio zaključak o populaciji

Figure 1. Using statistical analysis on sample(s) to make inferences about a population

terval pouzdanosti (engl. *confidence interval*, CI) 95% CI za RR je 1,57-7,92).

Moramo razumjeti da je ponekad moguće rabiti oboje vrste procjene (i vrijednosti i intervala) kako bi se donijeli zaključci o parametru neke populacije uzetom iz njenog uzorka.

Ako definiramo istinitu vrijednost kao vrijednost dotične populacije koja bi se dobila idealnim mjerenjem bez pogrešaka bilo kakvog tipa, morat ćemo prihvatiti činjenicu da možda nikada nećemo znati koji parameter sadržava istinitu vrijednost populacije (4). Međutim, kombiniranjem ove dvije procjene možemo dobiti određeni stupanj pouzdanosti da će istinita vrijednost biti unutar tog intervala, čak i u slučaju da naš rezultat (procijenjena vrijednost) nije jednak istinitoj vrijednosti, kao što je prikazano na donjoj slici (Slika 2.).

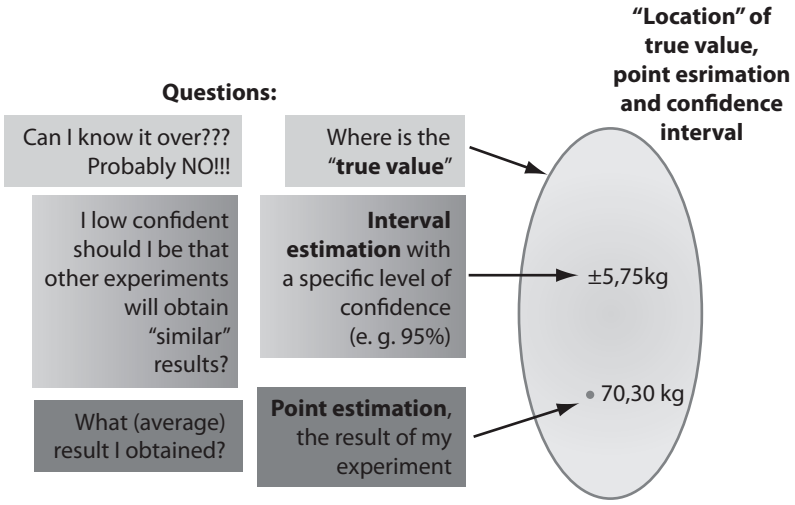
3. **Predviđanje/prognozu** - prognoziranje je postupak procjene u nepoznatim situacijama. Predviđanje je izjava ili tvrdnja da će se određeni događaj dogoditi u budućnosti te je to puno pouzdanije nego prognoziranje. Predviđanje je pojam vrlo sličan prognozi, no puno je općenitiji. Rizik i nesigurnost su ključni pojmovi kod predviđanja i prognoziranja.

an unknown population parameter, in contrast to point estimation, which is a single number (e.g. confidence interval 95% CI for RR is 1.57-7.92).

We have to understand that sometime it is possible to use both, point and interval estimation, in order to make inferences about a parameter of the population through a sample extracted from it.

If we define the "true value" as the actual population value that would be obtained with perfect measuring instruments and without committing error of any type, we will have to accept that we may never know the true value of a parameter of the population (4). But, using the combination of these two estimators, we may obtain a certain level of confidence, that the true value may be in that interval, even if our result (point estimation) is not necessarily identical with the true value, as is illustrated in the figure below (Figure 2).

3. **Prediction/forecast** - forecasting is the process of estimation in unknown situations. A prediction is a statement or claim that a particular event will occur in the future in more certain terms than a forecast, so prediction is a similar, but more general term. Risk and uncertainty are central to forecasting and prediction.



Slika 2. Koncept istinite vrijednosti, procijenjene vrijednosti i intervala pouzdanosti

Figure 2. The concept of true value, point estimation and confidence interval

4. **Statističko ispitivanje hipoteze** - zadnji, no time nikako nevažan i svakako najučestaliji način donošenja statističkog zaključka iz premise je statističko ispitivanje hipoteze. To je metoda donošenja statističkih odluka temeljem eksperimentalnih podataka i te se odluke gotovo uvijek donose pomoću takozvanih testova za ispitivanje nulte hipoteze.

   Nulta hipoteza (H_0) formalno opisuje neke aspekte statističkog ponašanja skupa podataka i taj se opis smatra valjanim ukoliko to ponašanje podataka nije proturječno nultoj hipotezi.

   Zbog toga se nultoj hipotezi suprotstavlja druga hipoteza, takozvana alternativna hipoteza (H_1). Statistički test u biti ispituje samo nultu hipotezu. Test ispitivanja nulte hipoteze ima oblik: „*Ne postoji (statistički značajna) razlika između skupina*" za ispitivanje razlike i „*Ne postoji povezanost*" za ispitivanje korelacije. Alternativna se hipoteza ne može potvrditi. Možemo samo odbaciti nultu hipotezu (u tom slučaju prihvaćamo alternativnu hipotezu) ili prihvatiti nultu hipotezu.

   Važno je shvatiti da većina statističkih protokola koji su u svakodnevnoj primjeni rabe jedan ili više testova za ispitivanje statističke hipoteze.

Pitanje 2: Zašto nam je potrebno statističko zaključivanje iz premise i njegov ključni princip – statističko ispitivanje hipoteze?

Odgovor 2: U kratkim crtama, zato što moramo na znanstveni način pokazati da je, primjerice, promatrana razlika između srednjih vrijednosti izmjerenih parametara tijekom pokusa na dva uzorka statistički značajna (4).

Statistički značajna razlika pojednostavljeno znači da postoji statistički dokaz te razlike; to ne znači da je razlika nuž-

4. **Statistical hypothesis testing** - last but not least, probably the most common way to do statistical inference is to use a statistical hypothesis testing. This is a method of making statistical decisions using experimental data and these decisions are almost always made using so-called "null-hypothesis" tests.

   The null hypothesis (H_0) formally describes some aspect of the statistical behavior of a set of data and this description is treated as valid unless the actual behavior of the data contradicts this assumption.

   Because of this, the null hypothesis is contrasted against another hypothesis, so-called "alternative hypothesis" (H_1). Statistical tests actually test the null hypothesis only. The null hypothesis test takes the form of: "*There is no difference among the groups*" for difference tests and "*There is no association*" for correlation tests. One can never "prove" the alternative hypothesis. One can only either reject the null hypothesis (in which case we accept the alternative hypothesis), or accept the null hypothesis.

   It is important to understand that most of the statistical protocols used in current practice include one or more tests involving statistical hypothesis.

Question 2: Why do we need statistical inference and its principal exponent – statistical hypothesis testing?

Answer 2: In a few words, because we need to demonstrate in a scientific manner that, for example, an observed difference between the means of a parameter measured during an experiment involving two samples, is "statistically significant" (4).

A "statistically significant difference" simply means there is statistical evidence that there is a difference; it does

no velika, važna ili značajna u smislu korisnosti pronalaska. To jednostavno znači da postoji mjerljiva vjerojatnost da pojedinačne vrijednosti iz uzorka dobro predstavljaju parametre populacije.

Uzmimo jedan primjer kako bi bolje razumjeli koncept. Uzeli smo dvije skupine ispitanika – ispitivana skupina je primila liječenje i prepisana im je modificirana prehrana, a kontrolna skupina je primila placebo i bila na regularnoj prehrani. Kod obje je skupine izmjerena i zabilježena tjelesna temperatura i težina. Rezultati pokusa prikazani su u tablici 1.

Ako pogledamo rezultate, mogli bismo temeljem algebarskog rezoniranja zaključiti da postoji veća razlika između srednjih vrijednosti za težinu nego između srednjih vrijednosti za tjelesnu temperaturu. No, kada primijenimo odgovarajući statistički test za usporedbu između srednjih vrijednosti (u ovom je slučaju odgovarajući test Studentov t-test za neparne (nezavisne) podatke) rezultat će biti iznenađujući. Jedina statistički značajna razlika jest ona između srednjih vrijednosti tjelesne temperature, što je sasvim oprečno našim očekivanjima temeljenim na općem iskustvu i znanju.

Postaje jasno da su nam statistički testovi (za ispitivanje statističke značajnosti) potrebni kako bismo mogli donijeti zaključak da je nešto postiglo ili nije postiglo „statistički značajnu razliku". Niti statističke niti znanstvene odluke ne smiju se pouzdano temeljiti samo na „onome što ljudsko oko vidi" ili na promatračevom „prethodno stečenom iskustvu"!

Mora se primijetiti da ispitivač ne može biti 100% siguran o promatranoj razlici, čak i kada je ona statistički značajna. Kako bi se promatrač mogao suočiti s nesigurnošću, u takvim se situacijama uvode dva komplementarna ključna koncepta inferencijske statistike: pouzdanost (npr. kao interval pouzdanosti) i razina značajnosti (engl. *significance level,* α ili alpha) (5).

Pojednostavljeno, razina značajnosti može se definirati kao vjerojatnost odluke o odbijanju nulte hipoteze kada je nulta hipoteza zapravo istinita (odluka poznata kao pogreška tipa I ili lažno pozitivna odluka. Najćešće korištene razine značajnosti su 5%, 1% i 0,1%, što empirijski odgovara razini pouzdanosti od 95%, 99% i 99,9%.

not mean the difference is necessarily large, important, or significant in terms of the utility of the finding. It simply means that there is a measurable probability that the sample statistics are good estimates of the population parameters.

For a better understanding of the concept, let's take an example. We took two samples of human subjects – a test sample, which received a treatment and a modified diet, and a control sample, which received placebo and a regular diet. For both samples, the body temperature and weight were recorded. The results from the experiment are presented in table 1.

If we will look at the results, based on "algebraic reasoning" we might say that there is a larger difference between the means of weight for those samples, than between the means of body temperature. But when we apply an appropriate statistical test for comparison between means (in this case, the appropriate test is "t-test for unpaired data"), the result will be surprising. The only statistically significant difference is between the means of body temperature, exactly the opposite conclusion that the one expected by our general knowledge and experience.

It becomes clear that we need statistical (significance) tests, in order to conclude that something has or hasn't achieved "statistical significance". Neither statistical nor scientific decisions can be reliably based on the judgment of "human eyes" or an observer's "previous experience(s)"!

It must be noted that the researcher cannot be 100% sure about an observed difference, even when statistically significant. To deal with the level of "uncertainty" in such situations, two, let's say "complementary", key concepts of inferential statistics are introduced: confidence (C) (e.g. as in confidence intervals) and significance level (α - alpha) (5).

In simple terms, significance level (α, or alpha), may be defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true (a decision known as a Type I error, or "false positive determination"). Popular levels of significance are 5%, 1% and 0.1%, empirically corresponding to a "confidence level" of 95%, 99% and 99.9%.

Tablica 1. Rezultati pokusa

Table 1. The results from the experiment

	Temperature control group (°C)	Temperature test group (°C)		Weight control group (kg)	Weight test group (kg)
Mean	37.0	37.9	**Mean**	80.0	85.0
	P-value	**Is P ≤ 0.05?**		**P-value**	**Is P ≤ 0.05?**
F test	0.049	Yes	**F test**	0.942	No
t-test	0.040	Yes	**t-test**	0.183	No

Kako bismo bolje razumjeli pojmove pouzdanosti i razine značajnosti, pogledajmo jedan općeniti primjer. Ako je procjenjena vrijednost nekog parametra P, s intervalom pouzdanosti [x, y] na razini pouzdanosti C, tada će svaka vrijednost izvan intervala [x, y] biti statistički značajno različita od P za razinu značajnosti $\alpha = 1 - C$, pod istim pretpostavkama raspodjele koje su se koristile pri izradi intervala pouzdanosti.

To znači, da ako u procjeni drugog parametra promatrana vrijednost bude manja od x ili veća od y možemo odbaciti nultu hipotezu. U tom slučaju nulta hipoteza glasi: „istinita vrijednost ovog parametra iznosi P", na razini značajnosti α; i obrnuto, ako procijenjena vrijednost drugog parametra leži unutar intervala [x, y], nećemo moći odbaciti nultu hipotezu koja kaže da je parametar jednak P.

Pitanje 3: Koje korake treba poduzeti za primjenu statističkog testa?

Odgovor 3:

1. Treba ispitati odgovarajuću nultu i alternativnu hipotezu.
2. Treba odabrati razinu značajnosti (označava se grčkim simbolom α (alfa). Često rabljene razine značajnosti su 5%, 1% i 0,1% što odgovara vrijednosti α (alfe) od 0,05, 0,01 i 0,001.
3. Izračunati odgovarajuću pojedinačnu vrijednost (S) prema ispravnoj matematičkoj jednadžbi testa.
4. Usporediti pojedinačnu vrijednost (S) s odgovarajućim kritičnim vrijednostima (engl. *critical value*, CV) (dobivenim iz statističkih tablica u standardnim slučajevima). U ovom se koraku može izračunati P vrijednost.
5. Odlučiti je li nulta hipoteza dokazana pa time i prihvaćena, ili je odbačena, a prihvaćena alternativna hipoteza. Pravilo u donošenju odluke jest da se nulta hipoteza odbaci ukoliko je $S > CV$ i obrnuto. U praksi to znači da ćemo, ako je $P \leq \alpha$, odbaciti nultu hipotezu; u ostalim ćemo je slučajevima prihvatiti (4).

Ako analizu načinimo suvremenim statističkim programima, računalo sâmo transparentno provodi korake 3 i 4, tako da odmah možemo dobiti P vrijednost te možemo izostaviti korak konzultiranja velikih statističkih tablica. Većina statističkih programa nudi izračunate rezultate pojedinih vrijednosti iz testa.

Konačno, primijenimo li neki statistički test kako bismo testirali naše podatke iz nekog pokusa, dobiti ćemo P vrijednost, koja definira vjerojatnost da ćemo dobiti takvu ili veću razliku pod uvjetom da je nulta hipoteza točna (6).

Ako se vratimo na naš primjer, prikazan u tablici 1., P vrijednost će nam dati odgovor na to pitanje (7). Ako su srednje vrijednosti populacija iz kojih potiču ta dva uzorka zaista iste, koja je vjerojatnost da ćemo ipak zaključiti

To a better understanding of these two terms, let's take a general example. If the point estimate of a parameter is P, with confidence interval [x, y] at confidence level C, then any value outside the interval [x, y] will be significantly different from P at significance level $\alpha = 1 - C$, under the same distributional assumptions that were made to generate the confidence interval.

That is to say, if in an estimation of a second parameter, we observed a value less than x or greater than y, we would reject the null hypothesis. In this case, the null hypothesis is: "the true value of this parameter equals P", at the α level of significance; and conversely, if the estimate of the second parameter lay within the interval [x, y], we would be unable to reject the null hypothesis that the parameter equaled P.

Question 3: What steps are required to apply a statistical test?

Answer 3:

1. The statement of relevant null and alternative hypotheses to be tested.
2. Choosing significance level (represented by the Greek symbol α (alpha). Popular levels of significance are 5%, 1% and 0.1%, corresponding to a value of 0.05, 0.01 and 0.001 for α (alpha).
3. Compute the relevant test's statistics (S), according with correct mathematical formula of the test.
4. Compare the test's statistic (S) to the relevant critical values (CV) (obtained from tables in standard cases). Here we may obtain so-called "P value".
5. Decide to either "fail to reject" the null hypothesis or reject it in favor of the alternative hypothesis. The decision rule is to reject the null hypothesis (H_0) if $S > CV$ and vice versa. Practically, if $P \leq \alpha$, we will reject the null hypothesis; otherwise we will accept it (4).

If we use modern statistical software, steps 3 and 4 are transparently done by the computer, so we may obtain directly the "P value", avoiding the necessity to consult large statistical tables. Most statistical programs provide the computed results of test statistics.

Finally when we apply a statistical significance test to data from an experiment, we obtain a so-called "P-value", which expresses the probability of having observed our results as extreme or even more extreme when the null hypothesis is true (6).

If we return to our example, illustrated in the table 1, the P value answers this question (7). If the populations from which those two samples were extracted, really did have the same mean, what is the probability of observing such a large difference (or larger) between sample means in an experiment when samples are of this size?

da postoji tako velika (ili veća) razlika između srednjih vrijednosti s tom veličinom uzorka?

Stoga, ako P vrijednost iznosi 0,04, to znači da postoji vjerojatnost od 4% da ćemo uočiti razliku koja je zaista tako velika u slučaju kad su srednje vrijednosti dviju populacija zapravo identične (nulta hipoteza je istinita). U ovom će nas slučaju gotovo mamiti zaključak da stoga, postoji vjerojatnost od 96% da ta promatrana razlika zapravo odražava istinitu razliku između populacija, a vjerojatnost da je to rezultat slučaja iznosi 4%. To je pogrešan zaključak. Ono što možemo reći je da bismo slučajnim uzorkovanjem iz identične populacije ustanovili razliku koja bi bila ista ili manja u 96% slučajeva, dok bismo veću razliku od opažene ustanovili samo u 4% slučajeva.

Odabir ispravnog statističkog testa. Što trebamo znati prije početka statističke analize?

Pitanje 4: Koje tipove podataka možemo dobiti tijekom ispitivanja?

Odgovor 4: Osnovni podaci dobiveni ispitivanjem mogu biti kvantitativni (numerički) ili kvalitativni (kategorički) podaci, obje skupine imaju nekoliko podtipova (4).

Kvantitativni (numerički) podaci mogu biti:

1. **Diskretni (diskontinuirani)** numerički podaci, samo u slučaju ako postoji konačan broj mogućih vrijednosti ili ako postoji prostor na brojevnom pravcu između svake dvije moguće vrijednosti (npr. iščitavanje vrijednosti sa zastarjelog živinog termometra).
2. **Kontinuirani** podaci čine ostatak numeričkih podataka koji se ne mogu smatrati diskretnima. To su tipovi podataka koji se obično povezuju s nekom vrstom naprednog mjerenja na instrumentima prema razvojnom stupnju struke.

Ono što je važnije jest da se podaci mogu mjeriti intervalnom ljestvicom ili onom omjernom. Za samu statističku analizu razlika između te dvije ljestvice mjerenja nije važna.

1. **Intervalna mjerna ljestvica** - podaci izraženi ovom mjernom ljestvicom nemaju apsolutnu nulu i ne možemo reći da je dvostruko veća brojčana vrijednost zaista dva puta veća. Primjerice, iako vrijednosti temperature izmjerene na Celzijusovoj ljestvici imaju jednake intervale između stupnjeva, 0°C nije apsolutna nula. Nula na Celzijusovoj ljestvici označava točku ledišta vode, no ne i totalnu odsutnost temperature. Nema smisla reći da je temperatura od 10°C dvostruko toplija 5°C.
2. **Omjerna mjerna ljestvica** - podaci izraženi omjernom mjernom ljestvicom imaju apsolutnu nulu. Na primjer, prilikom mjerenja duljine, nula predstavlja odsutnost duljine, a 10 metara je dvostruko dulje od 5 metara.

Thereby, if the P-value is 0.04, that means that there is a 4% chance of observing a difference as large as we observed when the two population means are actually identical (the null hypothesis is true). It is tempting to conclude, therefore, that there is a 96% chance that the difference we observed reflects a real difference between populations and a 4% chance that the difference is due to chance. This is a wrong conclusion. What we can say is that random sampling from identical populations would lead to a difference smaller than we observed (that is, the null hypothesis would be retained) in 96% of experiments and approximately equal to or larger than the difference we observed in 4% of experiments.

Choosing the right statistical test. What do we need to know before we start the statistical analysis?

Question 4: What type(s) of data may we obtain during an experiment?

Answer 4: Basic data collected from an experiment could be either quantitative (numerical) data or qualitative (categorical) data, both of them having some subtypes (4).

The quantitative (numerical) data could be:

1. **Discrete (discontinuous)** numerical data, if there are only a finite number of values possible or if there is a space on the number line between each 2 possible values (e.g. records from an obsolete mercury based thermometer).
2. **Continuous data**, that makes up the rest of numerical data, which could not be considered discrete. This is a type of data that is usually associated with some sort of advanced measurement using state of the art scientific instruments.

More importantly, the data may be measured at either an interval or ratio level. For the purposes of statistical analysis, the difference between the two levels of measurement is not important.

1. **Interval data** - interval data do not have an absolute zero and therefore it makes no sense to say that one level represents twice as much as that level if divided by two. For example, although temperature measured on the Celsius scale has equal intervals between degrees, it has no absolute zero. The zero on the Celsius scale represents the freezing point of water, not the total absence of temperature. It makes no sense to say that a temperature of 10 on the Celsius scale is twice as hot as 5.
2. **Ratio data** - ratio data do have an absolute zero. For example, when measuring length, zero means no length, and 10 meters is twice as long as 5 meters.

Oba tipa podataka (i intervalni i omjerni) mogu se koristiti u parametrijskim testovima.

Kvalitativni (kategorički) podaci mogu biti:

1. **Binarni (logički) podaci** – osnovni tip kategoričkih podataka (npr. pozitivno/negativno; prisutno/odsutno itd.).
2. **Nominalni podaci** - kod kompleksnijih kategoričkih podata, prvu (i najslabiju) razinu podataka predstavljaju nominalni podaci. Podaci nominalne razine dobiju se iz vrijednosti koje se razlikuju samo po nazivu. Ne postoji neka standardna shema poretka (npr. rumunjska, mađarska, hrvatska skupina ljudi itd.).
3. **Ordinalni (uredbeni) podaci** - slični su nominalnim podacima u tome da se podaci razlikuju prema nazivu, a od nominalnih podataka ih razlikuje shema stupnjevanja (npr. povremeni pušači, umjereni i teški pušači).

Pitanje 5: Kako se ti tipovi podataka mogu organizirati prije početka statističke analize?

Odgovor 5: Sirovi ili primarni podaci su još neobrađeni podaci sakupljeni na licu mjesta (4).

Primarni se podaci sakupljaju tijekom znanstvenog istraživanja te ih je potrebno prebaciti u format koji dozvoljava interpretaciju i analizu između varijabli.

Obično se podaci iz pokusa sakupljaju pomoću sistema za upravljanje bazama podataka (Microsoft Access, Oracle, MySQL ili čak namjenskih elektroničkih sistema pohrane zdravstvenih podataka ili tabličnih programa (kao što su Microsoft Excel ili OpenOffice Calc). U oba se slučaja podaci za istraživanje moraju prebaciti u program koji omogućuje rad s podacima kako bi se pripremili za statističku analizu. Moraju biti organizirani u tabličnom obliku s odgovarajućim brojem redova i stupaca, u formatu koji rabi većina statističkih programskih paketa.

Numerički se podaci mogu organizirati na dva načina, ovisno o zahtjevima statističkog programa koji se koristi:

1. **Indeksirani podaci** – imat ćemo barem dva stupca: jedan će stupac sadržavati brojeve zabilježene tijekom pokusa, a drugi će sadržavati grupirajuću varijablu. Na taj način, korištenjem samo dvaju stupaca tablice možemo zabilježiti podatke za velik broj uzoraka. Takav se pristup koristi u jakim i opsežnim statističkim programima kao što su SPSS (koji je razvio SPSS Inc., danas odjel u sklopu IBM), pa čak i besplatnim programima kao što su Epiinfo (koji je razvio Centar za kontrolu bolesti (Center for Disease Control), http://www.cdc.gov/epiinfo/downloads.htm) ili OpenStat (koji je razvio Bill Miller, http://statpages.org/miller/openstat/).
2. **Sirovi podaci** – podaci se organiziraju u specifičan stupac (ili red) za sve uzorke koje možemo imati. Iako ovaj pristup sa staništa početnika možda izgleda prirodniji i logičniji, relativno mali broj statističkih prog-

Both interval and ratio data can be used in parametric tests.

The qualitative (categorical) data could be:

1. **Binary (logical) data** - a basic type of categorical data (e.g. positive/negative; present/absent etc).
2. **Nominal data** - on more complex categorical data, the first (and weakest) level of data is called nominal data. Nominal level data is made up of values that are distinguished by name only. There is no standard ordering scheme to this data (e.g. Romanian, Hungarian, Croatian groups of people etc.).
3. **Ordinal (ranked) data** - the second level of categorical data is called ordinal data. Ordinal data are similar to nominal data, in that the data are distinguished by name, but different than nominal level data because there is an ordering scheme (e.g. small, medium and high level smokers).

Question 5: How could be these data types organized, before starting a statistical analysis?

Answer 5: Raw data is a term for data collected on source which has not been subjected to processing or any other manipulation (primary data) (4).

Thus, primary data is collected during scientific investigations, which need to be transformed into some format that allows interpretation and analysis between the variables.

Usually, the data from an experiment are collected using either a database managements system (Microsoft Access, Oracle, MySQL or even dedicated e-health record systems) or spreadsheet software (such as Microsoft Excel or OpenOffice Calc). In both cases, to be ready for statistical analysis, research data must be exported to a program that allows working with the data. They must be organized in a tabular (spreadsheet-like) manner using tables with an appropriate number of rows and columns, a format used by the majority of statistical packages.

If we have to deal with numerical data, those data can be organized in two ways, depending of the requirements of the statistical software we will use:

1. **Indexed data** – when we will have at least two columns: a column will contain the numbers recorded during the experiment and another column will contain the "grouping variable". In this manner, using only two columns of a table we may record data for a large number of samples. Such approach is used in well known and powerful statistical software, such as SPSS (developed by SPSS Inc., now a division of IBM) and even in free software like Epiinfo (developed by Center for Disease Control - http://www.cdc.gov/epiinfo/downloads.htm) or OpenStat (developed by Bill Miller, http://statpages.org/miller/openstat/).

2. Raw data – when data are organized using a specific column (row) for every sample we may have. Even if this

rama ga rabi (npr. MS Excel Statistics Add-in, OpenOffice Statistics ili Graphpad Instat and Prism, koje je razvio Graphpad Software Inc.).

Ako su naši zabilježeni podaci kvalitativni (kategorički), tablica primarnih podataka treba se sažeti u takozvanu tablicu kontingencije ili sadržajnosti. Tablica sadržajnosti je u svojoj osnovi format prikaza koji se primjenjuje za analizu i bilježenje povezanosti između dvije ili više kategorijskih varijabli. Uglavnom postoje dva tipa tablica sadržajnosti „2 x 2" (tablice s 2 reda i 2 stupca) i „N x N" (gdje je N > 2).

Pitanje 6: Koliko možemo imati uzoraka?

Odgovor 6: Ovisno o ustroju istraživanja postoje tri situacije (4,7):

- jedan uzorak;
- dva uzorka;
- tri ili više uzoraka.

U slučaju jednog uzorka postavlja se važno pitanje: kakav se statistički zaključak može donijeti, budući da je očigledno kako nisu ispunjeni uvjeti za usporedbu?

Iako izgleda kao da nema smisla, ipak se može napraviti barem nekakva statistička analiza. Primjerice, ako damo pirogeni lijek uzorku laboratorijskih životinja, moći ćemo napraviti usporedbe između srednje vrijednosti tjelesne temperature zabilježene tijekom pokusa i dobro poznate standardne vrijednosti za tu vrstu životinje, kako bismo pokazali je li razlika između tih vrijednosti statistički značajna i kako bi zaključili ima li lijek pirogeni učinak.

Ako u istraživanju imamo dva uzorka (što je najčešća situacija), sve što nam je činiti jest pratiti ispravni postupak inferencijske statistike kako bismo napravili ispravnu usporedbu između uzoraka.

Kod više od dva uzorka, statistička će se analiza činiti nešto kompliciranijom, no postoje statistički testovi s kojima se itekako mogu obraditi ovakvi podaci. Primjerice, možemo raditi usporedbe srednjih vrijednosti za sve uzorke u jednom trenutku koristeći analizu varijance (ANOVA test).

Također moramo imati na umu da postoje neki *post hoc* testovi koji se primjenjuju u drugom stupnju analize varijance u slučaju da je nulta hipoteza odbačena. Ti testovi mogu napraviti usporedbu između svakog para uzoraka iz pokusa.

Pitanje 7: Da li su uzorci zavisni (parni) ili nezavisni (neparni)?

Odgovor 7: Općenito, kad god je ispitanik u jednoj skupini povezan s ispitanikom u drugoj skupini, govorimo o parnim uzorcima.

Primjerice, u istraživanju majki i kćeri, uzorci su upareni, majka sa svojom kćerkom. Ispitanici u dva uzorka nisu nezavisni jedni od drugih. Za nezavisne uzorke je vjerojat-

approach may be considered more intuitive from the beginner's viewpoint, it is used by a relative small number of statistical software (e.g. MS Excel Statistics Add-in, OpenOffice Statistics or the very intuitive Graphpad Instat and Prism, developed by Graphpad Software Inc.).

If our recorded data are qualitative (categorical) data, the primary data table should be aggregated in a contingency table. A contingency table is essentially a display format used to analyze and record the relationship between two or more categorical variable. Basically, there are two types of contingency tables: "2 x 2" (tables with 2 rows and 2 columns) and "N x N" (where N > 2).

Question 6: How many samples may we have?

Answer 6: Depending on the research/study design, we may have three situations (4,7):

- one sample;
- two samples;
- three or more samples.

If we have only one sample, we may ask a pertinent question: what statistical inference could be made, because no obvious comparison terms seem to be available?

Even if it looks like a dilemma, still some statistical analysis may be done. For example, if we administer a pyrogenic drug to one sample of laboratory animals, we still may be able to make some comparisons between the mean of body temperature recorded during the experiment and a well-known "normal" value for that species of animals, in order to demonstrate if the difference between those values has "statistical significance" and to conclude if the drug has or has not some pyrogenic effects.

If there are two samples involved in the research (this is one of the most common situations), all we have to do is to follow the proper protocol of inferential statistics to make the convenient comparisons between samples.

When more than two samples are involved, the analysis seems to be a little more complicated, but there are statistical tests available, more than capable to deal with such data. For example, we can make comparisons of means for all samples in one instance using analysis of variance (ANOVA test).

Also, we have to know that some *post hoc* tests are available, used if the null hypothesis is rejected at the second stage of the analysis of variance and able to make comparison between each and every pair of samples from the experiment.

Question 7: Do we have dependent or independent samples/paired or unpaired groups?

Answer 7: In general terms, whenever a subject in one group (sample) is related to a subject in the other group (sample), the samples are defined as "paired".

nost da član populacije bude odabran kao uzorak potpuno neovisna o bilo kojem drugom odabranom članu, bilo da se radi o skupini tog člana ili nekoj drugoj skupini u istraživanju (7).

Parni se podaci mogu definirati kao vrijednosti koje se obično mjere u parovima i stoga se može očekivati da one više variraju između parova, nego između ispitanika unutar para. Ukoliko nisu postignuti ti uvjeti, u tom slučaju govorimo o neuparenim ili nezavisnim uzorcima.

Zašto je to važno? Postoje mnogi statistički testovi koji imaju različite verzije ukoliko se radi o parnim, odnosno neparnim uzorcima te imaju različit matematički pristup koji može dovesti do različitih rezultata. Primjerice, dobro poznati statistički test t-test koji se primjenjuje za usporedbu srednje vrijednosti između dva uzorka, ima različite verzije za parne/neparne uzorke: t-test za parne (zavisne) uzorke i t-test za neparne uzorke.

Stoga je odabir t-testa za parne uzorke (zavisne) umjesto onog za neparne (nezavisne) pogreška koja može dovesti do krivih rezultata/zaključaka u procesu statističkog zaključivanja iz premise.

Test za parne uzorke moramo odabrati u slučaju kada pokus slijedi jedan od ovih ustroja (7):

- kada mjerimo varijable prije i poslije intervencije kod svakog ispitanika;
- kada odabiremo ispitanike kao parove, uparene prema varijablama kao što su npr. dob, etnička skupina ili stupanj ozbiljnosti bolesti; jedan od parova bude liječen na jedan način; a drugi par na alternativni način;
- izvodimo laboratorijski pokus nekoliko puta, svaki puta s kontrolnim i ispitivanim uzorkom u duplikatu;
- mjerimo varijablu ishoda kod parova dijete/roditelj (ili bilo kojem sličnom paru).

Općenito govoreći, kad god očekujemo da će nam vrijednost u jednom uzorku biti bliža određenoj vrijednosti u drugom uzorku, nego što bi bila kod slučajno odabrane vrijednosti u drugom uzorku, moramo odabrati test za uparene podatke. U drugom slučaju odabiremo test za nezavisne uzorke.

Pitanje 8: Slijede li uzorkovani podaci normalnu/Gaussovu raspodjelu?

Odgovor 8: Ovisno o vrsti raspodjele, odabiremo parametrijske, odnosno neparametrijske testove.

Trebamo imati na umu da mnogi statistički testovi (npr. t-test, ANOVA i njene varijante) *a priori* pretpostavljaju da podaci uzorkovani iz populacije slijede Gaussovu (normalnu/zvonoliku) raspodjelu. Testovi koji slijede tu pretpostavku nazivaju se parametrijskim testovima, a njima se bavi parametrijska statistika (4).

Parametrijska statistika pretpostavlja da podaci slijede jedan tip raspodjele vjerojatnosti (npr. normalnu raspod-

For example, in a study of mothers and daughters, the samples are paired, a mother with her daughter. Subjects in the two samples are not independent of each other. For independent samples, the probability of a member of the population being selected is completely independent of any other subject being selected, either in the subject's own group or in any other group in the study (7).

Paired data may be defined as values which fall normally into pairs and can therefore be expected to vary more between pairs than within pairs. If such conditions aren't met, we will have to deal with unpaired or independent samples.

Why is this so important? Because there are many statistical tests that have different versions for paired/unpaired samples, with a different mathematical approach which may lead to different results. For example, a well-known statistical test, the t-test used for comparison of means between two samples, has different versions for paired/unpaired samples: paired (dependent) samples t-test and unpaired (independent) samples t-test.

Thereby, choosing a paired test (test for dependent samples) instead of an unpaired test (test for independent sample) is a mistake and may lead to wrong results/conclusions in the statistical inference process.

We have to choose a paired test when the experiment follows one of these designs (7):

- when we measure a variable before and after an intervention in each subject;
- when we recruit subjects as pairs, matched for variables such as age, ethnic group or disease severity - one of the pair gets one treatment; the other gets an alternative treatment;
- when we run a laboratory experiment several times, each time with a control and treated preparation handled in parallel;
- when we measure an outcome variable in child/parent pairs (or any other type of related pairs.

Broadly speaking, whenever we expect a value in one sample to be closer to a particular value in the other sample, than to a randomly selected value in the other sample, we have to choose a paired test, otherwise we choose an independent samples test.

Question 8: Are the data sampled from a normal/Gaussian distribution(s)?

Answer 8: Based on the normality of distributions, we chose parametric or nonparametric tests.

We should know that many statistical tests (e.g. t-tests, ANOVA and its variants), *a priori* assume that we have sampled data from populations that follow a Gaussian (normal/bell-shaped) distribution. Tests that follow this assumption, are called parametric tests and the branch of

jelu) i donosi zaključke o parametrima raspodjele. Međutim, kod mnogih populacija, kao i kod bioloških podataka, podaci ne slijede precizno Gaussovu raspodjelu. Gaussova se raspodjela širi u beskonačnost u oba smjera te tako uključuje i beskonačno negativne kao i beskonačno pozitivne brojeve, a biološki podaci su često po svojoj prirodi ograničeni u stupnjevanju. No, mnogi biološki podaci ipak slijede zvonoliku raspodjelu koja sliči Gaussovoj raspodjeli.

Stoga će ANOVA testovi, t-testovi i ostali statistički testovi ispravno ispitivati čak i u slučaju da je raspodjela samo približna Gaussovoj (posebno kod velikih uzoraka, npr. > 100 ispitanika) i ti se testovi rutinski primjenjuju na mnogim poljima znanstvenog istraživanja.

No u nekim situacijama, primjerice kada imamo mali uzorak (npr. < 10 ispitanika) ili kada kao varijablu rezultata imamo medicinski rezultat (npr. Apgar rezultat), primjena takvog testa, koji pretpostavlja da populacija slijedi normalnu raspodjelu, bez odgovarajućeg znanja o tom fenomenu, mogla bi rezultirati P vrijednošću koja bi navodila na pogrešan zaključak.

Iz tog razloga, druga grana statistike, neparametrijska statistika, nudi metode i testove nezavisne o raspodjeli podataka, koji se ne oslanjaju na pretpostavku da su podaci uzeti iz dâte raspodjele vjerojatnosti (u našem slučaju je to normalna raspodjela). Takvi se testovi nazivaju neparametrijski statistički testovi (4). Trebamo zapamtiti da gotovo svaki parametrijski statistički test ima odgovarajuću neparametrijsku inačicu.

Jedna od vjerojatno najtežih odluka tijekom statističkog protokola je koji test odabrati: parametrijski ili neparametrijski. Pitanje koje si možemo postaviti je: ako se neparametrijski testovi ne oslanjaju na pretpostavku da podaci koje obrađuju slijede normalnu raspodjelu, zašto ne primijeniti samo one tipove testova s kojima bi izbjegli pogrešku?

Kako bi razumjeli razliku između tih dvaju tipa testova moramo razumjeti daljnja dva osnovna statistička koncepta: robusnost (engl. *robustness*) i snaga statističkog testa (engl. *power*).

Robusni test se može upotrijebiti čak i kada neke od pretpostavki za izvođenje testa nisu zadovoljene. Neparametrijski testovi su robusniji od svojih parametrijskih inačica, primjerice mogu obraditi vrlo male uzorke, gdje su podaci daleko od normalne raspodjele.

Snaga statističkog testa je vjerojatnost da će taj test odbaciti nultu hipotezu, ako je alternativna hipoteza istinita (npr. da se neće napraviti pogreška tipa II). Kao što je autorica Ilakovac (6) prethodno detaljno opisala, pogreška tipa II je također poznata pod nazivom pogreška druge vrste, β pogreška ili lažno negativna, a definira se kao pogreška neisključivanja nulte hipoteze u slučaju kada je ona

statistical science that uses such tests is called parametric statistics (4).

Parametric statistics assume that data come from a type of probability distribution (e.g. normal distribution) and make inferences about the parameters of the distribution. However, many populations from which data are measured - and biological data are often in this category - never follow a Gaussian distribution precisely. A Gaussian distribution extends infinitely in both directions and so includes both infinitely low negative numbers and infinitely high positive numbers and biological data are often naturally limited in range. Still, many kinds of biological data do follow a bell-shaped that is approximately Gaussian.

Thus, ANOVA tests, t-tests and other statistical tests work well, even if the distribution is only approximately Gaussian (especially with large samples, e.g. > 100 subjects) and these tests are used routinely in many fields of science.

But in some situations, for example when we have to deal with small samples (e.g. < 10) or have as an outcome variable a medical score (e.g. Apgar Score), applying such a test that assumes that the population follows a normal distribution, without a proper knowledge of the phenomena, could result in a P-value that may be misleading.

For this reason, another branch of statistics, called nonparametric statistics, propose distribution-free methods and tests, which do not rely on assumptions that the data are drawn from a given probability distribution (in our case, the normal distribution). Such tests are named nonparametric statistical tests (4). We should be aware that almost every parametric statistical test has a correspondent nonparametric test.

Maybe one of the most difficult decisions when we go through a statistical protocol is to choose between a parametric or nonparametric test. A pertinent question we may ask is the following one: if the nonparametric tests do not rely on assumptions that the data are drawn from normal distribution, why not use only such type of tests, to avoid a mistake?

To understand the difference between these two types of tests, we have to understand two more basic concepts in statistics: robustness and power of a statistical test.

A robust statistical test is one that performs well enough even if its assumptions are somewhat violated. In this respect, nonparametric tests tends to be more robust than their parametric equivalents, for example by being able to deal with very small samples, where data are far to be normally distributed.

The power of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true (e.g. that it will not make a type II error). As already previously extensively reviewed by Ilakovac

zaista neistinita. S pojačanjem snage statističkog testa, opada vjerojatnost pogreške tipa II. Neparametrijski testovi su često robusniji, no obično imaju manju snagu testa. Drugim riječima, kod velikih uzorka može biti potrebno da se donose zaključci s istim stupnjem pouzdanosti (7).

Pitanje 9: Kada možemo primijeniti odgovarajući neparametrijski test?

Odgovor 9: Neparametrijski test trebamo zasigurno primijeniti u situacijama kao što su ove (7):

- Ako je varijabla ishoda ordinalan podatak ili rezultat s manje od dvanaestak kategorija (npr. Apgar rezultat). Jasno je da u tim slučajevima uzorak iz populacije ne može slijediti Gaussovu raspodjelu.
- Ako je uzorak premalen (< 10);
- Ako je nekoliko vrijednosti van ljestvice, previsoke ili preniske da bi se mjerile specijalnom mjernom tehnikom. Iako uzorak iz populacije slijedi normalnu raspodjelu, njegove podatke nije moguće analizirati parametrijskim testom (npr. t-testom ili ANOVA testom). Neparametrijski test se kod ove vrste podataka jednostavno može primjeniti, jer se neće oslanjati na pretpostavke da podaci slijede normalnu raspodjelu. Neparametrijski testovi bilježe izvorne podatke kao ordinalne podake. Izuzetno niskim i izuzetno visokim vrijednostima dodijele se vrijednosti stupnja i na taj se način neće narušiti analiza, kao što bi to bio slučaj kod primjene izvornih podataka s ekstremnim vrijednostima. U tom slučaju neće biti važno da neke vrijednosti nisu bile precizno izmjerene.
- Ako imamo dovoljno statističke pouzdanosti da je populacija daleko od one čiji podaci slijede normalnu raspodjelu. Mnoštvo testova koji ispituju normalnost raspodjele može ispitati prate li podaci iz uzorka normalnu raspodjelu.

Testovi za ispitivanje normalnosti raspodjele primjenjuju se za određivanje jesu li skupovi podataka dobro organizirani normalnom raspodjelom. Drugim riječima, kod ispitivanja statističke hipoteze, ti će testovi ispitati podatke prema nultoj hipotezi da se vidi slijede li oni normalnu raspodjelu.

Najčešći primjeri takvih testova su:

1. **D'Agostino-Pearsonov test** normalnosti raspodjele – koji izračunava iskošenost (engl. *skewness*) i spljoštenost (engl. *kurtosis*), kako bi izrazio koliko su podaci udaljeni od normalne raspodjele po pitanju asimetrije i oblika. Nadalje, on izračunava koliko se svaka od tih vrijednosti razlikuje od vrijednosti koja je očekivana u slučaju normalne raspodjele, te računa P vrijednost iz zbroja tih odstupanja. Taj je test normalnosti raspodjele višestruko upotrebljiv te ima veliku snagu (u usporedbi s nekim drugim testovima) pa ga stoga preporučuju neke suvremene statističke knjige.

(6), a Type II error is also known as an "error of the second kind", a β error, or a "false negative" and is defined as the error of failing to reject a null hypothesis when it is in fact not true. As power increases, the chances of a Type II error decrease. Nonparametric tests tend to be more robust, but usually they have less power. In other words, a larger sample size can be required to draw conclusions with the same degree of confidence (7).

Question 9: When may we choose a proper nonparametric test?

Answer 9: We should definitely choose a nonparametric test in situations like these (7):

- The outcome variable is a rank or score with fewer than a dozen or so categories (e.g. Apgar score). Clearly the population cannot be Gaussian in these cases.
- The same problem may appear when the sample size is too small (< 10 or so).
- When a few values are off scale, too high or too low to measure with a specific measurement technique. Even if the population is normally distributed, it is impossible to analyze the sample data with a parametric test (e.g. t-test or ANOVA). Using a nonparametric test with these kinds of data is easy because it will not rely on assumptions that the data are drawn from a normal distribution. Nonparametric tests work by recoding the original data into ranks. Extreme low and extreme high values are assigned a rank value and thus will not distort the analysis as would use of the original data containing extreme values. It won't matter that a few values were not able to be precisely measured.
- When we have enough "statistical confidence" that the population is far from normally distributed. A variety of "normality tests" are available to test the sample data for normal distribution.

Normality tests are used to determine whether a data set is well-modeled by a normal distribution or not. In other words, in statistical hypothesis testing, they will test the data against the null hypothesis that it is normally distributed.

The most common examples of such tests are:

1. **D'Agostino-Pearson normality test** – which computes the skewness and kurtosis to quantify how far from normality the distribution is in terms of asymmetry and shape. It then calculates how far each of these values differs from the value expected with a normal distribution, and computes a single P-value from the sum of these discrepancies. It is a versatile and powerful (compared to some others) normality test, and is recommended by some modern statistical books.
2. **Kolmogorov-Smirnov test** – used often in the past - compares the cumulative distribution of the data with

2. **Kolmogorov-Smirnovljev test** normalnosti raspodjele, koji se prije često upotrebljavao, uspoređuje kumulativnu raspodjelu podataka s očekivanom kumulativnom normalnom raspodjelom, a P vrijednost mu se temelji na najvećoj vrijednosti odstupanja, što baš i nije najosjetljiviji način da se procjeni normalnost raspodjele, stoga se smatra staromodnim.
3. pored ova dva testa postoji zaista velik broj ostalih testova koji ispituju normalnost raspodjele, kao što su: **Jarque-Bera test, Anderson-Darlingov test, Cramér-von-Misesov test, Lillieforsov test normalnosti** (adaptacija Kolmogorov-Smirnovljevog testa), **Shapiro-Wilkinsonov test, Shapiro–Francia test normalnosti** itd.).

Primjena testova za ispitivanje normalnosti raspodjele čini se jednostavnim načinom odlučivanja trebamo li se odlučiti za parametrijski ili neparametrijski test. No ona to nije, budući da trebamo paziti na veličinu uzor(a)ka prije no što primijenimo te testove. Za male uzorke (npr. < 15), testovi normalnosti baš i nisu korisni. Oni imaju malu snagu razlikovanja između populacije čiji podaci slijede Gaussovu raspodjelu i one čiji podaci ne slijede normalnu raspodjelu. Mali uzorci jednostavno ne sadržavaju dovoljno informacija da nam omoguće donošenje zaključka o obliku raspodjele za cijelu populaciju. Donja tablica sažima testove o kojima smo pisali na neposredan način (Tablica 2.).

Ako podaci ne slijede Gaussovu (normalnu) raspodjelu, možda ćemo moći pretvoriti vrijednosti kako bi tvorili Gaussovu raspodjelu (4). U ovom članku nećemo opisivati kako se to radi, no kao dobar primjer za mjerenja (numeričkih podataka) možemo spomenuti jednostavan način na koji se to može napraviti, a to je logaritamska pretvorba: nova vrijednosti = log (stara vrijednost).

U nekim slučajevima takav jednostavan pristup može nam omogućiti primjenu parametrijskog statističkog testa umjesto neparametrijskog.

the expected cumulative normal distribution, and bases its p-value simply on the largest discrepancy, which is not a very sensitive way to assess normality, thus becoming obsolete.
3. Beside of these two, there are a relatively large number of other normality tests, such as: Jarque-**Bera test, Anderson-Darling test, Cramér-von-Mises criterion, Lilliefors test for normality** (itself an adaptation of the Kolmogorov-Smirnov test), **Shapiro-Wilk test, the Shapiro–Francia test for normality** etc.

Using normality tests seems to be an easy way to decide if we will have to use a parametric or a non-parametric statistical test. But it is not, because we should pay attention to the size of the sample(s) before using such tests. For small samples (e.g. < 15), normality test are not very useful. They have little power to discriminate between Gaussian and non-Gaussian populations. Small samples simply do not contain enough information to let us make inferences about the shape of the distribution of the entire population. The table below will summarize the above discussion in a straightforward manner (Table 2).

If the data do not follow a Gaussian (normal) distribution, we may be able to transform the values to create a Gaussian distribution (4). It is not the subject of this paper how this could be done, but, as a good example, for measurements (numerical data) one simple way to do this is to use logarithmic transformation: new value = log (old value).

In some cases, such a simple approach may permit us to use a parametric statistical test instead of a nonparametric one.

Question 10: Shall we choose one-tailed or two-tailed tests?

Answer 10: Let's imagine that we design some studies/experiments to compare the height of young male adults (18-35 years) between various countries in the world (e.g

Tablica 2. Parametrijski nasuprot neparametrijskim testovima

Table 2. Parametric versus nonparametric tests

	Parametric tests	Nonparametric test	Shall we use a normality test?
Large samples (> 100)	**Robust.** P-value will be nearly correct, sometime even if population is fairly far from Gaussian population.	**Powerful.** If the population is Gaussian, the P-value will be nearly identical to the P-value we would have obtained from parametric test. With large sample sizes, nonparametric tests are almost as powerful as parametric tests.	**Useful.** We may use a normality test to determine whether the data are sampled from a Gaussian population.
Small samples (< 15)	**Not robust.** If the population is not Gaussian, the P-value may be misleading.	**Not powerful.** If the population is Gaussian, the P-value obtained from a nonparametric test may be higher than the P-value obtained from a parametric test.	**Not very useful.** It has little power to discriminate between Gaussian and non-Gaussian populations.

Pitanje 10: Hoćemo li primijeniti jednosmjeran (engl. *one-tailed test*) ili dvosmjeran (engl. *two-tailed test*)?

Odgovor 10: Zamislimo da imamo neka istraživanja/pokuse usporedbe visine kod mlađih muškaraca (18-35 godina) između raznih zemalja svijeta (npr. između Švedske i Južne Koreje i između Rumunjske i Bugarske). Tijekom statističke analize oblikovati ćemo nultu hipotezu H_0 (da ne postoji razlika između srednje vrijednosti visine između ta dva nezavisna uzorka) i alternativnu hipotezu H_1 za određeni statistički test. Recimo da podaci slijede Gaussovu raspodjelu i da je cilj provesti specifični test kako bi odredili treba li odbaciti nultu, a prihvatiti alternativnu hipotezu (u tom slučaju se primjenjuje t-test za neuparene/nezavisne uzorke).

No, postoje dvije različite vrste testova koji se mogu primijeniti (4,7).

Jednosmjerni test traži samo povećanje ili smanjenje (promjenu u jednom smjeru) kod parametra, dok dvosmjerni test traži bilo kakvu promjenu kod parametara (koja može biti bilo kakve vrste – povećanje ili smanjenje).

Kako bismo razumjeli taj koncept, trebamo definirati kritično područje na rubu raspodjele kod testa za ispitivanje hipoteze: skup svih ishoda koji će nas, ako se dogode, dovesti do odluke o odbacivanju nulte hipoteze i prihvaćanju alternativne hipoteze.

U jednosmjernom testu, postojat će samo jedno kritično će područje na rubu raspodjele (sivo polje na slici 3.). Ako vrijednost iz našeg uzorka leži u tom području, odbacit ćemo nultu hipotezu, a prihvatiti alternativnu. U dvosmjernom testu tražimo ili povećanje ili smanjenje, što znači da u tom slučaju postoje dva kritična područja na rubu raspodjele, kao što je vidljivo na slici 3.

between Sweden and South Korea and another, between Romania and Bulgaria). So, during a statistical analysis, a null hypothesis H_0 (e.g. there is not a difference between the heights mean for those two independent samples) and an alternative hypothesis H_1 for the specific statistical test has been formulated. Let's consider that the distribution is Gaussian and the goal is to perform a specific test to determine whether or not the null hypothesis should be rejected in favor of the alternative hypothesis (in this case t-test for unpaired/independent samples will be the relevant one).

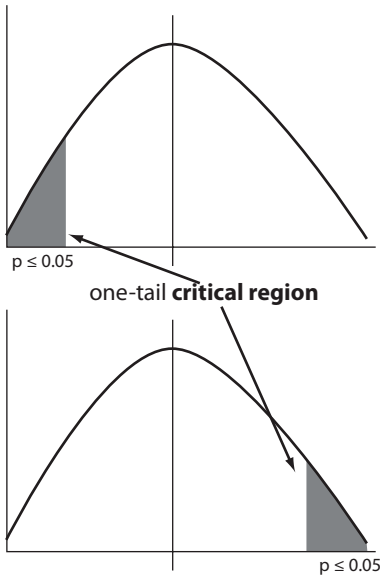
But there are two different types of tests that can be performed (4,7).

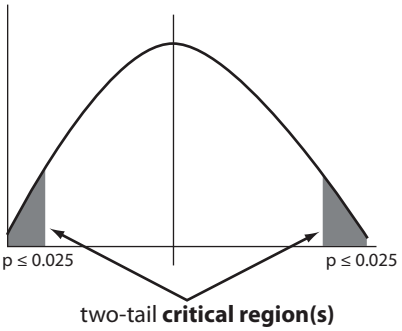
A one-tailed test looks only for an increase or a decrease (a one-way change) in the parameter whereas a two-tailed test looks for any change in the parameter (which can be any change - increase or decrease).

To understand this concept we have to define the critical region of a hypothesis test: the set of all outcomes which, if they occur, will lead us to decide to reject the null hypothesis in favor of the alternative hypothesis.

In a one-tailed test, the critical region will have just one part (the grey area in the figure below). If our sample value lies in this region, we reject the null hypothesis in favor of the alternative one. In a two-tailed test, we are looking for either an increase or a decrease. In this case, therefore, the critical region has two parts, as in figure 3.

When comparing two groups, we must distinguish between one- and two-tail P-values. The two-tail P-value answers this question: Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart (or further) as we obser-





Slika 3. Kritična područja na rubu raspodjele kod jednosmjernih i dvosmjernih testova

Figure 3. Critical regions in one-tailed and two-tailed tests

Kada uspoređujemo dvije skupine, moramo razlikovati između P vrijednosti dobivene jednosmjernim i dvosmjernim testom. P vrijednost dvosmjernog testa daje odgovor na pitanje: Pod pretpostavkom da je nulta hipoteza istinita, koja je vjerojatnost da će slučajno odabrani uzorci imati srednje vrijednosti toliko razdvojene (ili još više) kao što vidimo u ovom pokusu s bilo kojom skupinom koja ima veću srednju vrijednost?

Kako bi mogli interpretirati P vrijednost jednosmjernog testa, moramo prije početka sakupljanja podataka predvidjeti koja će skupina imati veću srednju vrijednost. P vrijednost jednosmjernog testa odgovara na pitanje: Pod pretpostavkom da je nulta hipoteza istinita, koja je vjerojatnost da ćemo kod slučajno odabranih uzoraka opaziti srednje vrijednosti toliko razdvojene (ili još više) kao što vidimo u ovom pokusu s određenom skupinom koja ima veću srednju vrijednost?

P vrijednost jednosmjernog testa prikladna je samo u slučaju kada nam prethodni podaci, fizička ograničenja i zdrav razum ukazuju da je razlika, ako ona uopće postoji, može biti samo u jednom smjeru. S druge strane, može nas zanimati samo rezultat u jednom smjeru. Primjerice, u slučaju da je razvijen novi lijek koji liječi stanje za koje postoji stari lijek, jasno je da će istraživači biti zainteresirani za nastavak istraživanja na novom lijeku, samo u slučaju ako djeluje bolje od starog lijeka. Nulta hipoteza će se prihvatiti ako novi lijek djeluje isto ili gore nego stari lijek.

Dakle, suštinsko je pitanje ovdje poznajemo li u dovoljnoj mjeri ustroj istraživanja, kako bismo znali može li se razlika dogoditi isključivo u jednom smjeru ili nas zanima razlika između skupina u oba smjera.

Prema tome, za jednosmjerni se test možemo odlučiti jedino u slučaju kada uspoređujemo srednje vrijednosti visina odraslih muškaraca između Švedske i Južne Koreje, jer nam zdrav razum i iskustvo govori da razlika, ako će je uopće biti, može biti samo u jednom smjeru (muškarci Šveđani bi trebali biti viši od muškaraca iz Južne Koreje).

Kada tu istu analizu napravimo za Rumunje i Bugare, takva pretpostavka možda neće biti istinita, što znači da ćemo odabrati dvosmjerni test.

Za izračun P vrijednosti jednosmjernim testom trebamo se odlučiti samo ako su dvije stvari istinite:

- prije nego što sakupimo podatke moramo moći predvidjeti koja će skupina imati veću srednju vrijednost;
- ako se dogodi da druga skupina ima veću srednju vrijednost – čak i ako je samo malo veća – tada moramo tu razliku pripisati slučaju.

Iz svih se tih razloga preporuča, posebno početnicima, da prije posegnu za ispravnim dvosmjernim testom umjesto jednosmjernog, osim ako imaju dobar razlog za odabir P vrijednosti jednosmjernog testa.

Pitanje 11: Što je cilj naše statističke analize?

ved in this experiment with either group having the larger mean?

To interpret a one-tail P-value, we must predict which group will have the larger mean before collecting any data. The one-tail P-value answers this question: Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart (or further) as observed in this experiment with the specified group having the larger mean?

A one-tail P-value is appropriate only when previous data, physical limitations or common sense tell us that a difference, if any, can only go in one direction. Or alternatively, we may be interested in a result only in one direction. For example, if a new drug has been developed to treat a condition for which an older drug exists. Clearly, researchers are only interested in continuing research on the new drug if it performs better than the old drug. The null hypothesis will be accepted if the new drug performs the same or worse than the older drug.

So, the real issue here is whether we have sufficient knowledge of the experimental situation to know that differences can occur in only one direction, or we are interested only in group differences in both directions.

In the light of these things, considering the above mentioned studies, we may choose a one-tail test only when we compare the heights mean of adult males between Sweden and South Korea, because our common sense and experience tell us that a difference, if any, can only go in one direction (the adult male Sweden citizens should be taller than the South Korean citizens).

When we make the same analysis for Romanian and Bulgarian citizens, this presumption may not be accurate, so we will have to choose a two-tailed test.

We should only choose a one-tail P-value when two things are true:

- first, we must have predicted which group will have the larger mean before we collect any data;
- if the other group ends up with the larger mean - even if it is quite a bit larger - then we must attribute that difference to chance.

For all these reasons, especially for beginners, choosing the right two-tailed test instead of a one-tailed test is recommended, unless we have a good reason to pick a one-tailed P-value.

Question 11: What is the goal of our statistical analysis?

Answer 11: When using basic statistical analysis, we may have, at most, three main goals (4,7):

1. To compare means (or medians) of the one, two or more groups/samples (e.g. is blood pressure higher in control than treated group(s)?).
2. To make some correlation, to look at how one or more independent variable(s) and one dependent variable

Odgovor 11: Kod osnovne statističke analize možemo imati najviše tri glavna cilja (4,7):

1. Usporediti srednje vrijednosti (ili medijane) jedne, dvije ili više skupine/uzoraka (npr. je li krvni tlak viši kod kontrolne skupine ili liječene skupine/liječenih skupina?).
2. Napraviti korelaciju, da se ustanovi kako se jedna ili više nezavisnih varijabli i jedna zavisna varijabla odnose međusobno (npr. kako težina i/ili dob utječu na krvni tlak).
3. Izmjeriti povezanost između jedne ili više nezavisnih varijabli (npr. epidemiološki čimbenici rizika) i jedne i više zavisnih varijabli (npr. bolesti). To je takozvana analiza tablica kontingencije ili sadržajnosti, gdje možemo promatrati kako su nezavisna varijabla/nezavisne varijable (npr. dim cigarete ili teški oblik pušenja) povezani s jednom ili više zavisnih varijabli (npr. rak pluća i njegovi razni oblici).

Iako postoje tri cilja, u ovom ćemo članku obraditi samo prvi cilj: usporedbu srednjih vrijednosti između jedne, dvije ili više skupina/uzoraka. Ovisno o tome koliko uzoraka imamo, naš će cilj biti dati znanstveni odgovor na sljedeća pitanja:

- Za jednu skupinu/jedan uzorak: izmjerili smo varijablu u tom uzorku i srednja vrijednost je drugačija od hipotetske (normalne) vrijednosti. Je li to poslijedica slučaja? Ili nam to govori da je promatrana razlika statistički značajna?
- Za dvije skupine/dva uzorka: izmjerili smo varijablu u dvije skupine i srednje vrijednosti (i/ili medijani) izgledaju kao da su različite. Je li to rezultat slučaja? Ili nam to govori da između skupina zaista postoji razlika?
- Za tri ili više skupina/uzoraka: izmjerili smo varijablu u tri ili više skupina i srednje vrijednosti (i/ili medijani) su različite. Je li to rezultat slučaja? Ili nam to govori da između skupina zaista postoji razlika? Između kojih skupina postoje razlike?

Kako bi se dao znanstveni odgovor na ova pitanja moramo usporediti srednje vrijednosti (medijane) onih skupina/uzoraka primjenom jednog od sljedećih statističkih testova (preporučujemo primjenu dvosmjernog testa, osim ako nemamo dobar razlog za odabir jednosmjernog testa) (Tablica 3.).

Poznavajući osnovne statističke pojmove i koncepte, postupak odabira statističkog testa iz gornje tablice vrlo je shvatljiv, ukoliko slijedimo algoritamski način te ispravno pratimo postupnik za izbor testa, kao što je ovaj prikazan na slici 4., kako bismo izbjegli pogreške tijekom postupka.

Zaključak

Postupak odabira ispravnog statističkog testa može biti problematičan zadatak, no dobro poznavanje i razumijevanje odgovarajućih statističkih pojmova i koncepata može pomoći u donošenju ispravne odluke.

relate to each other (e.g. how do weight and/or age affect blood pressure).

3. To measure association between one or more independent variables (e.g. epidemiological risk factors) and one or more dependent variables (e.g. diseases). This is so-called analysis of contingency tables, where we may look at how independent variable(s) (e.g. smoke or higher levels of smoking) are associated with one or more dependent variable(s) (e.g. lung cancer and its various forms).

Even if there are three goals, we will here discuss only the first goal: the means comparison between one, two or more groups/samples. Depending on how many samples we have, our goal will be to provide a scientific response to the following questions:

- For one group/sample: we've measured a variable in this sample and the mean is different from a hypothetical ("normal") value. Is this due to chance? Or does it tell us that the observed difference is a significant one?
- For two groups/samples: we've measured a variable in two groups, and the means (and/or medians) seem to be distinct. Is that due to chance? Or does it tell us the two groups are really different?
- For three or more groups/samples: we've measured a variable in three or more groups, and the means (and/or medians) are distinct. Is that due to chance? Or does it tell us the groups are really different? Which groups are different from which other groups?

To provide a scientific response for such questions, we have to compare means (medians) of those groups/samples using one of the following statistical tests (with the recommendation to use a two-tailed test, unless we have a good reason to pick a one-tailed test) (Table 3).

Now, knowing the basic terms and concepts, the tests selection process from the tests presented in the above table, can be very easy to understand if we shall think in an algorithmic manner, parsing the proper decision-tree, such as the one presented in the figure 4, to avoid any mistakes during the process.

Conclusion

The selection process of the right statistical test may be a difficult task, but a good knowledge and understanding of the proper statistical terms and concepts, may lead us to the correct decision.

We need, especially, to know what type of data we may have, how are these data organized, how many sample/groups we have to deal with and if they are paired or unpaired; we have to ask ourselves if the data are drawn for a Gaussian on non-Gaussian population and, if the proper

Tablica 3. Statistički testovi kojima se uspoređuju srednje vrijednosti (medijani) za jednu, dvije, tri ili više skupina/uzoraka

Table 3. Statistical tests that compare the means (medians) for one, two, three or more groups/samples

How many samples?	Paired/ Unpaired Dependent/ Independent samples?	All sample(s) are drawn from a normal distribution?/ Parametric (P) or non-parametric test (NP)?	Name of the statistical test	Observations
1 sample	One sample only	Yes/P	One sample t-test	
		No/NP	Wilcoxon rank sum test, One Sample Chi-Square test	
2 samples	Paired	Yes/P	Paired t-test	
		No/NP	Wilcoxon matched pairs test	
	Unpaired	Yes/P	Independent samples t-test	It assumes that the two samples have equal variance (in other words that the difference between the variance of the two samples has not statistical significance). The F test may be used to prove this assumption.
		Yes/P	Welch's corrected unpaired t-test	It assumes that those two samples have unequal variance. The F test may be used to prove this assumption.
		No/NP	Mann-Whitney U test	We may observe that is only one nonparametric test for unpaired data, instead of 2 tests for parametric data. This is happened because a nonparametric tests will not rely on assumptions that the data are drawn from a normal distribution, thus the use of variance become meaningless
3 or more samples	Paired	Yes/P	Repeated-measures one-way ANOVA	We will present here only the simple form of analysis of variance (ANOVA), not the two-way or multifactorial ANOVA. **Some *post hoc* tests are available, able to make comparison between each and every pair of samples from the experiment.**
		No/NP	Friedman's test	*Post hoc* tests are available, able to make comparison between each and every pair of samples from the experiment.
	Unpaired	Yes/P	One-way ANOVA	*Post hoc* tests are available
		No/NP	Kruskal-Wallis test	*Post hoc* tests are available

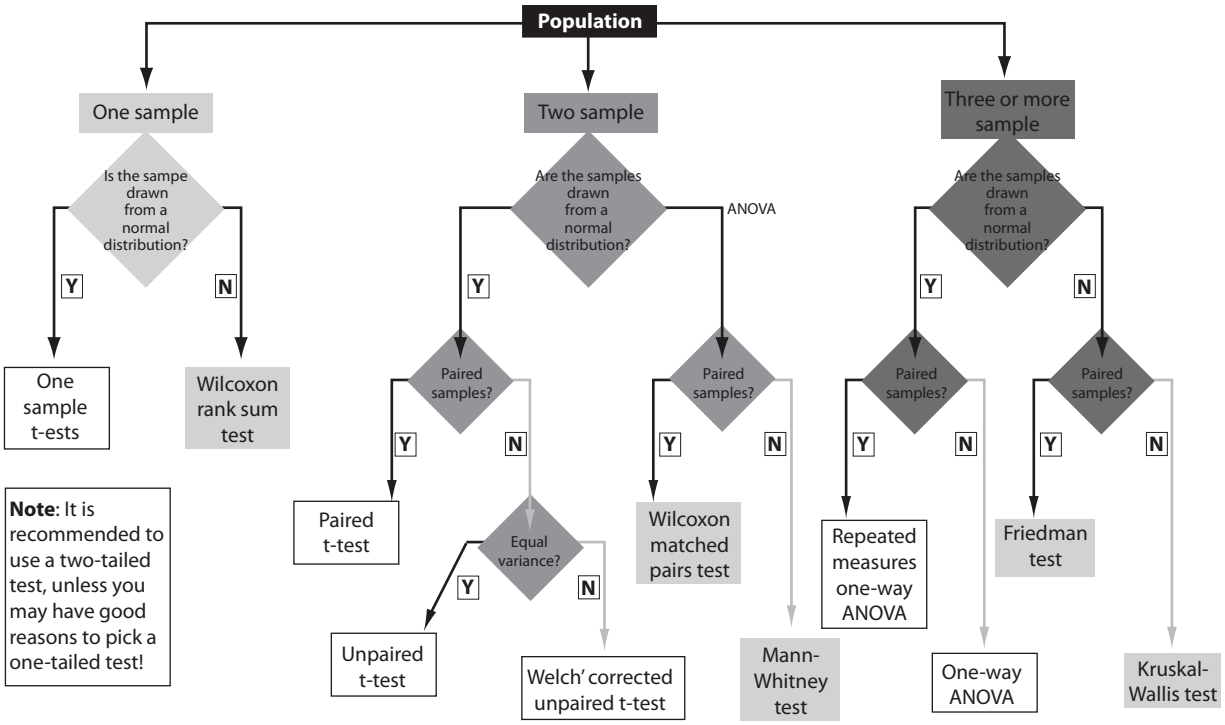
Posebno je potrebno znati s kojim tipom podataka raspolažemo, kako su ti podaci organizirani, koliko skupina/uzoraka imamo i jesu li podaci parni (zavisni) ili neparni (nezavisni); moramo si postaviti pitanje slijede li podaci iz populacije Gaussovu raspodjelu ili ne, te ukoliko postoje uvjeti za to, hoćemo li odabrati jednosmjerni test (nasuprot dvosmjernom testu koji je obično preporučen izbor).

Temeljem takvih informacija možemo slijediti statistički postupnik za izbor ispravnog testa prema algoritamskom

conditions are met, to choose an one-tailed test (versus the two-tailed one, which is, usually, the recommended choice).

Based on such kind of information, we may follow a proper statistical decision-tree, using an algorithmic manner able to lead us to the right test, without any mistakes during the test selection process.

Even if we didn't discussed here the means comparison when two or more factors are involved (e.g. bifactorial



Slika 4. Proces odabira ispravnog statističkog testa

Figure 4. The selection process for the right statistical test

načelu, koji bi nas trebao moći dovesti do ispravnog testa bez pogrešaka tijekom postupka odabira.

Čak i ako u ovom članku nismo govorili o usporedbi srednjih vrijednosti kada su uključena dva ili više čimbenika (npr. bifaktorijalna ANOVA) ili ostala dva glavna cilja statističkog zaključivanja (analiza tablica kontingencije ili sadržajnosti i korelacijska/regresijska analiza), algoritamskim bi principom i u takvim slučajevima morali moći odabrati ispravni statistički test.

Za neki drugi članak ostavit ćemo neke vrlo osporavane koncepte kao što su izrazito visoke ili izrazito niske vrijednosti i njihov utjecaj u statističkoj analizi, utjecaj vbrijednosti koje nedostaju itd.

ANOVA) or the other two main goals of statistical inference (analysis of contingency tables and correlation/regression analysis), the algorithmic manner would be useful also in such cases, using the same approach to choose the right statistical test.

Still, some much disputed concepts will remain to be discussed in other future articles, such as outliers and their influence in statistical analysis, the impact of the missing data and so on.

Literatura/References

1. *Cox DR. Principles of statistical inference. Cambridge University Press, 2006.*
2. *McHugh ML. Standard error: meaning and interpretation. Biochem Med 2008;18:7-13.*
3. *Slavkovic A. Analysis of Discrete Data. Available at: http://www.stat.psu.edu/online/courses/stat504/01_overview/index.html. Accessed: October 24, 2009.*
4. *Marusteri M. [Notiuni fundamentale de biostatistica:note de curs]/ Fundamentals in biostatistics:lecture notes. University Press Targu Mures, 2006. (in Romanian)*
5. *Simundic AM. Confidence interval. Biochem Med 2008;18:154-61.*
6. *Ilakovac V. Statistical hypothesis testing and some pitfalls. Biochem Med 2009;19:10-6.*
7. *Motulsky HJ. GraphPad Prism - Statistics Guide. GraphPad Software Inc., San Diego California USA, 2007, Available at: www.graphpad.com. Accessed: October 24, 2009.*