5

Biostatistics

Doctor 2018 | Medicine | JU

◉Sheet ○Slides

Done By

Ameen + Alia

Contributed in the Scientific Correction

Ameen + Alia

Contributed in the Grammatical Correction

Ameen + Alia

Doctor

Hamzeh Aduraidi

In the previous lecture, we talked about probability. We said that probability can be estimated using tables with the most important one being the **z-table** which works perfectly when we have a **continuous** variable that is standardized to become **normally** distributed. How about when the variable is categorical or discrete?

In this case, we can use 2 other tables

Binomial table	Poisson table
For dichotomous variables i.e. when the discrete variable has 2 options only E.g. male/female	When the discrete variable has 3 or more categories

NOTE: we are not required to know how to estimate probabilities using these 2 tables, the probability would be given by the professor in the exam. We just have to know how to use the z-table.

Now, with unit 3 wrapped up let's start talking about a new topic.

Biostatistics includes 2 steps. The 1st is descriptive statistics & the 2nd is inferential statistics. In inferential statistics, we use descriptive statistics to answer our research question (to come up with a conclusion (*inference*) regarding the population). At the end, we make a decision which can be either to **keep or reject the null hypothesis**. And of course, we prefer to **reject** the null hypothesis if we can, because it opposes our own hypothesis.

We previously talked about descriptive statistics and now we'll be talking about *inferential statistics*. The latter includes **many** tests and to be able to choose the appropriate one, you need to do the following:

1- Determine the independent and the dependent variables.
2- After that, think about the dependent (outcome) variable. What is its level of measurement? Is it ***continuous*** (either interval/ratio)? Then use ***parametric statistics***. Is it ***discrete*** (either nominal/ordinal)? Then use ***nonparametric*** statistics.

Parametric statistics are the preferred type of inferential statistics, why?

1- **Continuous** outcome variables are more powerful and flexible in their conclusions. Whereas when outcome variables are *discrete* (like when the variable is just yes/no) conclusions are less flexible/satisfying.
2- Parametric methods do not *only* allow the researcher to study the effect of many independent variables on the dependent variable, but they also make the study of their interaction possible.

As a result, we always aspire to use parametric statistics. BUT to use parametric statistics some assumptions must be made & none of them can be neglected.

Parametric assumptions

- The observations must be independent.
- **Dependent** variable should be continuous.
- The observations must be drawn from normally distributed populations. *Extra info: normality can be tested by a software called SPSS,* if normal --> parametric statistical tests can be used.
- These populations must have the same variances. Homogeneity of variance can be tested by *Levene's test* (we'll talk about it in future lectures)
- The groups should be randomly drawn from normally distributed and independent populations e.g.
  Male X Female Pharmacist X Physician Manager X Staff
  NO OVERLAP
- The independent variable is categorical (discrete)
- Distribution for the two or more independent variables is normal.

And if one or more of these assumptions are broken, we have to settle for the less favorable option: non parametric statistics. So, for a statistical method to be classified as nonparametric, it must satisfy at least one of the following conditions.

- The method can be used with nominal data.
- The method can be used with ordinal data.
- The method can be used with interval or ratio data when no assumption can be made about the population probability distribution (in small samples).

Non parametric methods:

- Are often the *only way* to analyze nominal or ordinal data and draw statistical conclusions.
- Require no assumptions about the population probability distributions. Don't require large populations, or normally distributed populations, or continuous outcome variables or homogeneity of variance.
- Are often called distribution-free methods.
- Good for outliers.
- Non-parametric tests based on ranks of the data work well for ordinal data (data that have a defined order, but for which averages may not make sense).
- Have disadvantages: they lack the power and the flexibility that's found in parametric statistical methods.

- There is at least one nonparametric test equivalent to each parametric test. These tests fall into several categories:
  1- Tests of differences between groups (independent samples).
  2- Tests of differences between variables (dependent samples).
  3- Tests of relationships between variables (correlation)

Now we've got that creepy table from experimental epidemiology. We'll break it down into blocks to make it easier to understand. But before that, let's review some terms.

Independent variable: The cause. E.g. having obesity

Dependent variable: The outcome, it's called so because it depends on the independent variable. E.g. Developing chronic heart diseases

Now regarding groups of samples, they can be dependent and independent as well.

Independent samples: Totally different samples. E.g. 10 students from section 1 & 10 students from section 2, these are two independent samples.

Dependent samples: They're actually the same group of people, but tested at different times. E.g. a group of 10 students from section 3 measured their weight. After one month of giving them dietary pills, we measured their weight again. These are considered 2 observations/samples, but they are called dependent because the second is so similar to the first.

Number of samples and their type, dependent or independent?

Level of dependent variable

Level of Measurement	Sample Characteristics					Correlation
	1 Sample	2 Sample		K Sample (i.e., >2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	X^2	X^2	Macnarmar's X^2	X^2	Cochran's Q	
Rank or Ordinal		Mann Whitney U	Wilcoxin Matched Pairs Signed Ranks	Kruskal Wallis H	Friendman's ANOVA	Spearman's rho
Parametric (Interval & Ratio) Continuous	z test or t test	t test between groups	t test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				

So, in rows we have the sample characteristics. Samples can be 1 sample, 2 samples, or K (3 or more) samples.

And in columns we have the level of the dependent variable. We already know that it can be nominal, ordinal, or continuous (includes both interval and ratio).

Now let's take each case on its own, and things will get clear.

1-If we are studying a nominal dependent variable and we have 2 independent samples → We use independence Chi squared

2-If we are studying a nominal dependent variable and we have 2 dependent samples → We use 'Mcnarmers' Chi squared

3-If we are studying a nominal dependent variable and we have 3 or more independent samples → We use independence Chi squared as well

4-If we are studying a nominal dependent variable and we have 3 or more dependent samples → We use Cochran's Q

5-If we are studying an ordinal dependent variable and we have 2 independent samples → We use Mann Whitney U

6-If we are studying an ordinal dependent variable and we have 2 dependent samples → We use Wilcoxin Matched pairs

7-If we are studying an ordinal dependent variable and we have 3 or more independent samples → We use Kruskal Wallis H

8-If we are studying an ordinal dependent variable and we have 3 or more dependent samples → We use Friendman's ANOVA

9-If we are studying a continuous dependent variable and we have 2 independent samples → We use t-test between groups/independent t-test/student's t-test

10-If we are studying a continuous dependent variable and we have 2 dependent samples → We use t-test within groups/dependent t-test

11-If we are studying a continuous dependent variable and we have 3 or more independent samples → We use One way ANOVA

12-If we are studying a continuous dependent variable and we have 3 or more dependent samples → We use repeated measure ANOVA

**That was all what we need to know for Biostatistics. However, we'll mention some more notes for experimental epidemiology.*

1-We can have 1 sample. In this case, we compare our study sample to something standard, not to another sample. We use certain tests which are mentioned in the table under '1 sample'.

2-Correlation tests are used to detect if there's any relationship between two variables at all. These tests vary depending on levels of both variables. Such tests include:

a-Nominal + Nominal → Chi squared
b-Ordinal + Ordinal → Spearman's rho
c-Continuous + Continuous → Pearson's r
d-Nominal + Continuous → independent t-test
e-Ordinal + Continuous → Biserial

SHORT QUIZ

Which statistical test should be used to analyze data?

1) An environmental psychologist developed a program to decrease garbage in certain streets of University of Jordan. It was implemented in 2014, and the amount of litters collected (in kilograms) in the same streets was measured in 2015 and then again in 2017.

a-Dependent t-test

b-One way ANOVA

c-Chi squared

d-Kruskal Wallis H

2) A researcher wishes to determine whether cooperative learning, computer assisted learning, or self-paced learning provides the most academic achievement for students, given that achievement is resembled by grades (4 for A, 3 for B, 2 for C, 1 for fail).

a-Repeated measure ANOVA

b-One way ANOVA

c-Kruskal Wallis H

d-Student's t-test

ANSWERS

1-**A.** Amount of garbage is a **continuous** variable. Testing the **same** streets **twice** means 2 dependent samples. This is analyzed by dependent t-test.

2-**C.** The grades are **ordinal**. We have **three different** methods of learning which means 3 independent samples. This is analyzed by Kruskal Wallis H.

Good Luck!!