DOCTOR 2020 | JU

COMMUNITY MEDICINE

WRITER : **2018**

CORRECTOR : **Farah mherat**

DOCTOR: **Hamza**

INTRODUCTION INTO BIOSTATISTICS

What is Biostatistics? It is a word of 2 parts:

1-**Bio** = from biology (which indicates that this science is concerned with living things – mainly humans-).

2 Statistics= Art of prediction.

So, in other words, **biostatistics** can be defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine.

It is a growing field with applications in many areas of biology including epidemiology, medical sciences, health sciences, educational research and environmental sciences.

- **How is statistics the art of prediction?**

Here is how: Suppose that you are a physician or a researcher and you have a question in your head ,and you wonder whether or not there is a relationship between 2 factors (2 variables), for example let's say that you are a dermatologist and you noticed that the **skin cancer** proportions had been increasing among women in the last 10 years ,and you noticed at the same period of time- in the last 10 years- that also more women were using **sunblock** , in your head you will wonder if there is a relationship between these two things! Each one of these 2 things is called a variable; because a woman can or cannot be with skin cancer, and can or cannot be using a sunblock.

Ok, so the first thing to do in order to find an answer for this question in your head (whether or not there is a relationship), is that you go to the **literature** (which is the studies that have been published before you by other researchers), try to find an answer in the literature, if you didn't find any answer, then you are the person who we are looking for to answer this question; because no one had done this research question before you in Jordan for example! -hurray, congrats! -, so you have to go and do a research study in order to answer this research question.

Now, let's suppose that there are 4 million women in Jordan, is it possible to go to each one of them and ask each one if she has skin cancer or not ?! and if she use sunblock or not?! This will take a very long period of time, and that's going to be so hard and requires a great effort, and it can be so **difficult to implement**, or **impossible** sometimes. Instead, you have just to collect a sample from the population, but this sample is **not** collected **haphazardly**, it is collected using certain rules and there are several techniques in sampling, that we will discuss in this sheet.

Ok, you collected the sample, let's say 1000 women from those 4 million women, you ask each one of these 1000 the two questions, and their answers represent your **data**, but wait a minute! These data in their **primary simple form** are meaningless and **don't make sense**! Instead, you explain, organize, and summarize the data. So, for example you say that the **average** is 30 for example. That is much **more meaningful.**

This is the first step in biostatistics, which is called **descriptive statistics** (which is basically the process of organizing, summarizing, and describing the sample characteristics, using the lowest number of words and numbers, e.g. using average, standard deviation, upper and lower limit, quantiles, percentiles, … etc.).

So far, have you answered the research question by descriptive statistics **only**?! Not yet, you have just described the data, for example, you say that the percentage of skin cancer among women in Jordan is 12%, and the users of sun blocks among them are 17%. But, you didn't see whether there is a **relationship** between the 2 factors, **not on the sample level, but on the population level.**

To do so, here comes **the second and most important step of biostatistics** which is called the **inferential statistics,** here comes **the art of prediction**. Based on the numbers that I have from my sample and the description of it, I use the **magic of biostatistics** to come up with conclusions and results regarding the bigger population (the 4 million women in the example above).

You have to keep in mind that you can **never** be 100% sure about your conclusion regarding the bigger population, but you can still be as high as 95% confident, which is a very good number. So, at the end of your study, you can say something like this: (We are 95% confident that there is a statistically significant association between using sunblock and developing skin cancer), this is a very important finding; because based on it you can provide your health education, your recommendations, and your patients. Thus, you added a very important piece to our knowledge, you added a very good understanding of this phenomenon.

Now**, how can we move from a sample composed just of couple of thousands to a population with 4 million?** using **inferential statistics.**

- What are the types of inferential statistics? Chi-square, T-test, One-Way ANOVA …etc. will be discussed later on.

-By using inferential statistics, you can predict the answer for a whole population.

Concerns of Biostatistics:

1.Biostatistics is concerned with collection, organization, summarization, and analysis of data.

2.We seek to draw inferences about a body of data when only a part of the data is observed.

Purposes of Statistics:

1.To describe and summarize information thereby reducing it to smaller, more meaningful sets of data.

2. To make predictions or to generalize about occurrences based on observations. Inferential statistics

3.To identify associations, relationships or differences between the sets of observations.

Data:

Data are numbers which can be measurements or can be obtained by counting. Biostatistics is concerned with the interpretation of the data and the communication of information about the data.

Populations and Samples:

A population: is the collection or set of all of the values that a variable may have. The entire category under consideration.

A sample: is a part of a population. The portion of the population that is available, or to be made available, for analysis.

Example: Studying the self-esteem and academic achievement among college students.
• Population: all student who are enrolled in any college level.

• Sample: students' college at the University of Jordan.

Sampling

How can I go to a large population and reduce it in a small sample? And how can I be confident that this small sample represent the large population? This is the **art of sampling.**

•**Sampling**: the process of selecting portion of the population. Or in other words, it is the selection of a number of study units/ subjects from a defined population.

•**Representativeness**: the key characteristic of the sample is close to the population. (how much does the sample represent the population).

We want the sample to be small enough so it's manageable and large enough so it's representativeness

•**Sampling bias**: excluding any subject without any scientific rational. Or excluding any subject not based on the major inclusion and exclusion criteria.

Neglect a significant or a considerable part of the original population

Questions to Consider:

1.(Reference population): to whom are the results going to be applied?

2.(Study population): What is the group of people from which we want to draw a sample?

3.(Sample Size): How many people do we need in our sample?

4.(Sampling Method): How will these people be selected?

I will choose my sample from the sampling frame which is smaller than study population

---

The population which I want to make a conclusion/inference to it at the end of my research study is called the **reference population**; because I am generalizing my findings to them, I refer to them when I report my findings. Reference population is usually in millions or hundreds of thousands (e.g. the 4 million Jordanian women in the example at the beginning of this sheet), so you will be **never** able to access it.

You have to go deeper and deeper to end up with a good sample.

-In Jordan, we have 12 governorates, distributed in 3 regions (North, Middle, South), and in each governorate there are several cities, villages, camps…etc. And all of them contain women.

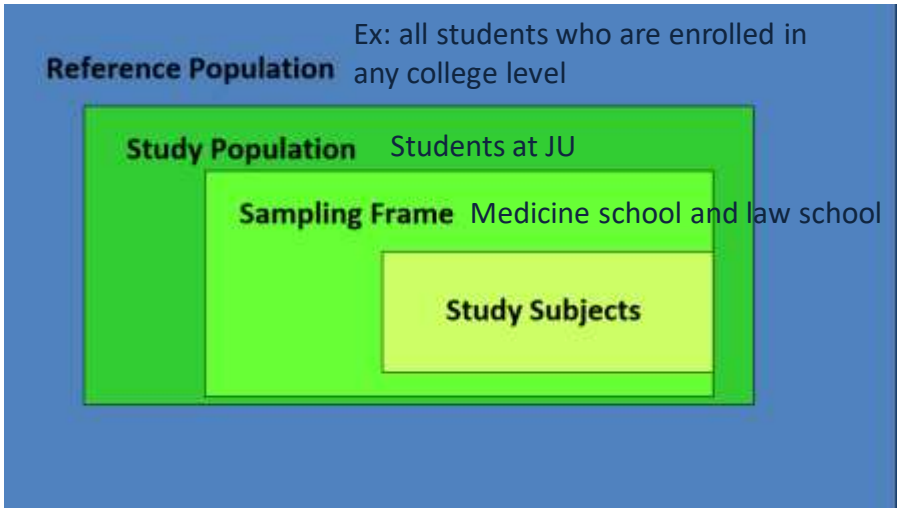
I select the governorates or cities or villages where I want to go, in order to apply my study, e.g. I select one governorate from the north, one from the middle, and one from the south. So, I shrink it down. This is called the **study population.**

But, it doesn't make sense that I go to all the women in each governorate or city which I had selected in order to ask them my study questions! Yes, it is narrower, but still too broad! I have to shrink it down more, so I will think about going to primary healthcare centers in each city that I had selected, and see the women there, and take my sample from them. These places like PHC centers where I will **physically** be there to collect my data is called the **sampling frame** (also called **accessible population**).

•Sampling frame is a listing of all the units that compose the study population, or it is the listing of all elements of a population, i.e., a list of all medical students at the university of Jordan, 2014-2016.

• Sampling depends on the sampling frame.

Still in the accessible population there are thousands, but I don't want all these thousands to be in my sample, I only need one thousand for example. So, I select one of the several sampling techniques, and from those thousands, I choose the required one thousand. These 1000 women who end up in my sample is called the **study subjects** (also called **sample members**). Note that sample size is not chosen randomly, there are specific ways to select my sample size to answer any type of research question.



•**Element**: The single member of the population (population element or population member are used interchangeably).

element= sample member = subject (the three words have the same meaning in **quantitative research**). But, in (**qualitative research** they are called **participants**).

Types of Sampling Methods (sampling techniques)

By using sampling techniques, we can enter the sampling frame and select the certain number of subjects needed for the study.

There are 2 main types:

1-Probability sampling techniques(methods):

The techniques in which we guarantee an equal chance for everybody in the sampling frame to be selected into the sample.

-Involves the random selection process to select a sample from members or elements of a populations, to ensure that each unit of the sample is chosen on the basis of chance.

-All units of the study population should have an equal or at least a known chance of being included in the sample.

- Requires a sampling frame (Listing of all study units).

equal chance= random selection

- ❖ There are many types of probability sampling, and we will discuss **five** in this sheet, which are:

•Simple Random Sampling. • Systematic sampling. • Stratified sampling. • Cluster sampling. • Multistage sampling.

2 Nonprobability sampling techniques:

Sometimes you are **not able** to or **not interested** in guaranteeing and equal chance for everybody in the sampling frame to be selected into the sample. Instead we do something called nonprobability sampling. It used most in the case of having a specific criteria of the people you need in your research (ex: people with a rare disease

not able to= because you don't have enough time and enough resources for example.

not interested= because of the nature of your research question.

In many scenarios, guaranteeing an equal chance for each member in the sampling frame to be selected to the sample is **not possible** or is **not desired**. Most frequent reasons for use of nonprobability samples involve **convenience** and the **desire** to use available subjects.

Nonprobability sampling: the sample elements are chosen from the population by nonrandom methods. More likely to produce a **biased sample** than the random sampling. This restricts the **generalization** of the study findings.

- ❖ We will talk about **four** types of nonprobability sampling, which are:

• Convenience sampling. • Snowball sampling. • Quota sampling. • Purposive sampling.

Probability Sampling Methods

1 Simple Random Sampling:

This is the simplest of probability sampling.

Steps:

a. Make a numbered list of all units in the population.

b. Decide on the sample size.

c.Select the required number of sampling units using the lottery method or a random number table.

How can I ensure that the selection process of the sample was 100% random?

There are some websites and books that give you something like this (see the table below), basically lists of random numbers and tell you which numbers to choose (these are totally random and they have no system or order).

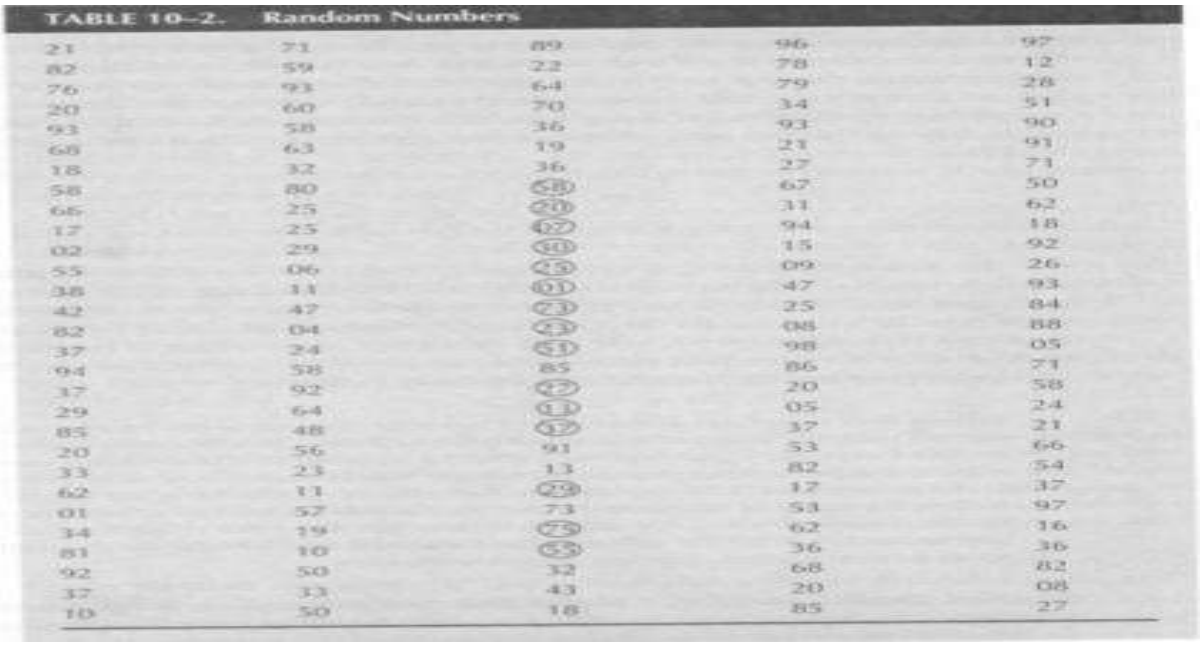


TABLE 10–2. Random Numbers

21	71	89	96	97
82	59	22	78	12
76	93	64	79	28
20	60	70	34	51
93	58	36	93	90
68	63	19	21	91
18	32	36	27	71
58	80	58	67	50
66	25	20	31	62
17	25	07	94	18
02	29	80	15	92
55	06	43	09	26
38	11	01	47	93
42	47	73	25	84
82	04	49	08	88
37	24	61	98	05
94	58	85	86	71
37	92	47	20	58
29	64	11	05	24
85	48	97	37	21
20	56	91	53	66
33	23	13	82	54
62	11	28	17	37
01	57	73	53	97
34	19	75	62	16
81	10	63	36	36
92	50	32	68	82
37	13	43	20	08
10	50	18	85	27

For example, I came to a class which contains 100 students and I want to choose 10 only for my sample, so I make a list which contains the students' names and beside each name there is a specific number from 1 to 100. I hide the names and keep only the numbers. Then from the 100 number available, I randomly choose 10 numbers. Then I take the 10 randomly chosen students into my sample.

The problem is that sometimes the 10 randomly chosen students can be from the same category or kind, e.g. they all could be females, or all could be Jordanians, or all could be young. Sometimes this is not desired. Maybe the sample you got doesn't representativeness

So, even simple random sampling is not ideal enough, so there are other types that works better in this case.

2- Systematic Sampling (or Systematic Random Sampling):

- Individuals are chosen at **regular intervals** from the sampling frame.
- Ideally we randomly select a number to tell us the **starting point**.

e.g. every 5th household or every 10th women attending ANC.

For example, I want to choose 10 students from 100 randomly, but I don't want them all to be from the front, or from the middle, or from the back. So, in order to achieve better randomness, I take the first 10 and I close my eyes and pick one of them **randomly**, let's suppose I picked the student number 7, so what I have to do is that I skip 10 students and pick the student number 17, skip 10, pick the student number 27, skip 10, pick the student number 37,47, 57, 67, 77, 87,97.

So I will end up with 10 students randomly selected, but you **guarantee** that all the places in the class have been covered.

There is an **equal chance for all**; because in the beginning I choose the first one randomly.

In this way we are guaranteeing an equal chance for everybody in the sampling frame to be selected even better.

- Sampling fraction = $\frac{Sample\ size}{Study\ population}$
- Interval size = $\frac{Study\ population}{Sample\ size}$

3-Stratified Sampling:

Sometimes your sampling frame is way too more complex than just a section in a classroom, it could be a whole university, so we need more sophisticated probability sampling techniques, one of them is stratified sampling.

•If we have some study units with different characteristics which we want to include in the study, then the sampling frame needs to be divided into **strata** according to these characteristics.

•**Ensures that proportions of individuals with certain characteristics** in the sample will be **the same** as those in the whole study population.

•Random or systematic samples of predetermined sample size will have to be obtained from each stratum based on a sampling fraction for each stratum.

For example, the University of Jordan has 50 thousand students, and I need only 1000 student for my sample, but I don't have lists with all the 50 thousand students in the university, or I could have, but that is too huge and broad.

So why don't I divide the university into 3 strata for example (A stratum for the Medical colleges, and one for the Scientific colleges, and another one for Humanities Colleges).

Then, for each stratum, I do a simple random sampling.

By this, I ensure that all the disciplines which share the same general characteristics will be represented enough in my sample.

-Stratified sampling works very nicely when you have variable sampling frames.

4-Cluster Sampling:

Another alternative way. Instead of dividing something into strata and then doing simple random sampling for each one, why not to take ready **groups or clusters** into my sample from the beginning.

For example, the number of students in the faculty of medicine is 4000, and you want only 200 of them to be in your sample. At the same time, the students in the faculty are already divided into sections of equal number of students, let's say 50 students per section, so why not to pick randomly any 4 sections, including all the students in each one. Note that all the students in the 4 randomly picked sections (clusters) must be included in the sample.

Conclusions about Cluster Sampling:

• Selection of study units (clusters or groups) instead of the selection of individuals.

• All subjects/units in the cluster who meet the criteria will be sampled.

• Clusters often geographic units. e.g. schools, villages, etc.

• Usually used in interventional studies. e.g. assessing immunization coverage

• **Advantages:**

a- sampling frame is not required in this case.

b-Sampling study population scattered over a large area.

5-Multistage Sampling:

Let's suppose you are conservative even more and you want really a sophisticated way to make your sample, so you go and combine **more than 2** of the previously mentioned types. To make sure that the sample is random enough and every subject has an equal chance

For example, you make stratified sampling and inside it you make cluster sampling,

Or you make cluster sampling and in each cluster you make systematic sampling.

So, you make **three types or more** of probability sampling techniques in the same process, that is why it's called multistage sampling.

Disadvantages: longer and more complicated and needs lot of time.

Conclusions about Multistage Sampling:

• Involves more than one sampling method.

• Is therefore carried out in phases.

• Does not require an initial sampling frame of whole population.

• NEED TO KNOW SAMPLING FRAME OF CLUSTERS E.G. PROVINCES.

• Require sampling frames of final clusters.

•Applicable to community based studies e.g. interviewing people from different villages selected from different areas, selected from different districts, provinces.

Nonprobability sampling techniques

Remember that we use them when we are not able to or not interested in guaranteeing and equal chance for everybody in the sampling frame to be selected into the sample, with nonrandom selection. And we use them in general when our study is about the community (Because I don't have lists with their names).

1- Convenience sampling (Accidental or incidental sampling):

Side note: it is convenience not convenient.

For example, I have 100 students in the class, and I want to pick 10 of them, I enter the class and choose the first 10 in front of me (as convenient as that).

It does **not guarantee an equal chance** for everybody but we don't care because maybe we don't have enough time.

It is **not random**, because there is a very high risk for **bias**; as the students who usually tend to sit in the back have different certain characteristics than those who usually tend to sit in the front. Also, students who were absent at that day of sampling had no chance.

•People may or may not be typical of the population, no accurate way to determine their representativeness.

• Most frequently used in health research.

• **Advantages**: Saves time and money.

2-Snowball sampling:

A method by which the study subjects assist in obtaining other potential subjects (networking).

Useful in topics of research where the subjects are reluctant to make their identity known, Drug users, Aids patients, etc.

It is also useful with subjects with rare diseases. For example, I want a sample of women who have children with **autism**, it is very difficult to find because it is too rare. So what I do is that I will try to find only one woman whose child has autism and collect the data from her. Then, I ask her whether she knows other women whose children have autism as well, most probably that she knows, and she will guide you to other women whose children have the same case, and one by one, the sample of interest will get bigger until we reach the sample size that we want.

Very useful in rare diseases

-It is called snowball sampling because it begins very small just like a snowball, then as it rolls it becomes bigger.

3-Quota sampling: This method used when you afraid that a certain portion from your sampling frame will be under presented or estimated

-Used during elections to guarantee the **representation of minorities.**

-Also it is very important in some researches like medical researches.

For example, during the first half of the 20th century (1900-1950), all the Physicians and researchers in the USA were doing researches and producing new knowledge, science, medications, and recommendations, based on these researches. But, all the samples taken were composed of white people only, ignoring all minorities including black people, Arabs, Latinos, Asians, and Native Americans …etc. They were not represented in any of these samples. So, the drugs and researches were only working for white people, and that is **ethically unacceptable.** The majority of medical researchers say true / false for white people(major part)but not necessary true/false for black people so the Quota sampling solved this problem

In order to solve this problem, and during the second half of the 20th century, they started saying that when it comes to **population studies**, we have **to increase the chance of minorities** in being represented in the samples. (We have to give them a quota in the samples).

Now, let's say we want to pick 1000 American to represent all the Americans, for a hypertension study for example. And despite that Blacks in USA don't compose more than 15% of the total American population, but we pick the sample in which: one third of it are white people, one third black people, and one third Latinos. As a result, when I do a conclusion at the end, I ensure that this conclusion works really, not only for the Whites, not only for the majority, but for everybody!

Ok, someone may ask that the whites represent about 80% of Americans, so doesn't it make sense if they represent most of the sample? No; because we are afraid that there are physiological differences between Whites & Blacks, due to the **race**. So, quota works perfectly here.

The question is: Does increasing the minority chance to be selected in the sample here impact the final outcome?

Right, but **we take the risk to impact the outcome positively, instead of taking the risk to ignore the minority negatively**. So, it is for the best to increase the chance of minority to be selected.

-We know that it is sometimes thought to be counterintuitive or provoking, but this is something we use in **ethnicity and racial studies**, because it is very useful & important in some situations.

-In the 1970s, they had published a study which said that a certain dietary intake doesn't influence the blood pressure. But actually, it doesn't influence blood pressure among Whites ONLY. Unfortunately, it increases blood pressure significantly among Blacks. And that was **ethically wrong.**

The important question now: What is the difference between the quota sampling in the nonprobability sampling techniques, and between the stratified sampling in the probability sampling?

The main difference is the percentages, in the **stratified** sampling we respect **the real percentages** (if 30% of the university is from the medical stratum, then I take 30% of my sample to be from the medical stratum. If 50% of the university is from the medical stratum, then I take 50% of my sample from the medical stratum …etc.).

However, in **quota** sampling, we **maximize** someone chances, and **minimize** another one chances, **on purpose.** In order to know if there is an association or a relationship statistically.

-In Jordan, we had started to give a quota for refugees for example, because we don't want to end up with a conclusion that works on Jordanians only! there are other nationalities living with us on the same land!

-**Note**: If we do for each quota a separated study or sampling, then it is a **multistage sampling**.

Some more information about Quota Sampling:

•In quota sampling, the sample is selected by convenience (e.g. the first 50% of males and 50% of females).

• A mean for securing potential subjects from these strata.

•In a quota sampling variables of interest to the researcher (include subject attributes), such as age, gender, educational background, are included in the sample.

4- Purposive sampling (handpicking, judgmental):

•Subjects are chosen because they are typical or representative of the accessible population, or because they are experts (more knowledgeable) in the field of research topic.

• Qualitative researchers use Purposive sampling.

For example: let's suppose that I have taken a random sample of 100 Jordanians, then I asked them a very **specific question**, let's say I asked them about their opinion of the impeachment of the American president Donald Trump, maybe 95% of Jordanians will say that they don't know what you are talking about, because they have no knowledge about it, or no opinion on it, or they don't care; **so that will be waste of my time and their time, and that will be misleading.**

Thus, on purpose, I go to Journalists, politicians, students of strategic studies and political sciences, then I ask them that question.

So, when you purposefully target a certain group; because they have certain characteristics, and collect your data from this small group because of the nature of your research question or the nature of your study, this is called **purposive sampling.**

The question here is: **After finishing my study -which I have used purposive sampling in it- and ending up with a conclusion, can I do a generalization for it?**

The answer is yes, I can, which is unexpected, but that is how it works, remember it is a nonprobability type at the end!

Simply, I end with a result like this: (Jordanians are with the impeachment of the American president), and I don't say that: (Only Jordanian Journalists are with the impeachment) for example.

And here comes the art of prediction, I predict that if all the Jordanians were experts in a specific field, then they would have answered the same answer as the experts.

-**Finally, depending on the sampling technique that I use, the power of my results will be determined.**

-**What do we prefer to become really strong and powerful in our results?**

Any **probability sampling technique**. But, in the real life, you will realize that most of the researchers settle down to just convenience sampling. But also, the findings that came up from the convenience sampling are not always that bad, they can lead us to the right direction.

THE END

Smile, and have a beautiful day!!!