Descriptive statistics

Center for Data Research and Analytics

Descriptive Research Questions

- Research questions determine what kind of study design and statistical method will be used. Research questions could be descriptive or analytical
- The descriptive research question commonly starts with who, which, what, and when, not the **Why**
- The primary objective of these research is to explore the **parameters**, the characteristics of the population(s)
  - What is the prevalence of cancer in Bangladesh
  - What is the prevalence of Parkinson's disease among the Bangladeshi geriatric population
  - Which students more likely to suffer anxiety disorder in the Rajshahi university. In this research question, researcher is looking for a subgroup where anxiety disorder is higher

Descriptive research questions are to gathering information about a population and generating hypothesis.

Study design

- Study design tells you how will you answer to a research question about a population
- This controls all the activities to answer a specific research question.
- Study design provides the strength to the evidence collected. The stronger the evidence in favor of the answer, the more powerful the recommendations are

Suppose you want to how many red/different colored candies does the Jar have?

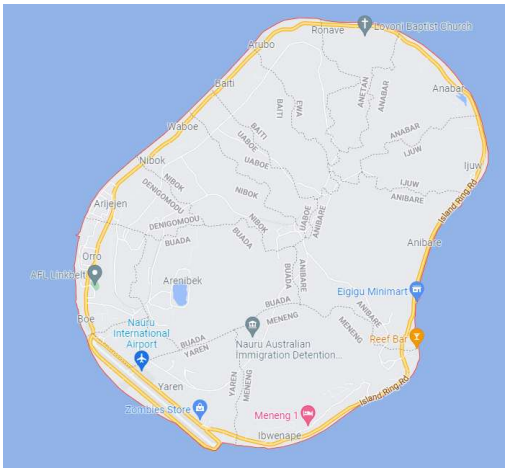
Open the jar, take each of the cady, give an id, determine the color and record *them until finish*. At the end calculate how many different colors you have.

This is your own study design

Study design establishes the degree of reliability of your research and recommendations

Research Problem

- **Nauru** is an independent island country
- Area: 8 Sq. miles
- Population: 11,000



Next couple of slides we will discuss three standard study designs that is useful descriptive research.

We want to know the seroprevalence of COVID-19. What would be the study design

Case reporting

- Doctors/public health researchers report the sporadic cases in journals and news papers.
- For multiple cases, doctors reports the case series
- Case reporting primarily explore the patterns in infection to find the intervention.
- The number of cases and different characteristics of disease are reported to aware the community. Statistical involvement is less in this design.

SAGE Open Medical Case Reports

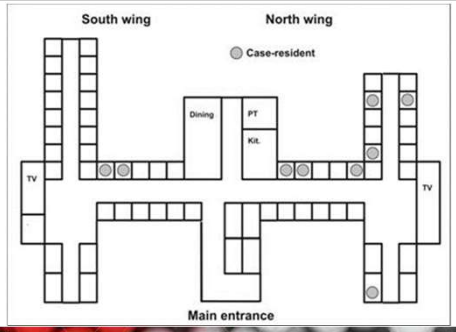
A peer-reviewed open access journal published by SAGE

Case reporting is the preliminary state of a disease research in public health.

Case series

Outbreak Investigation

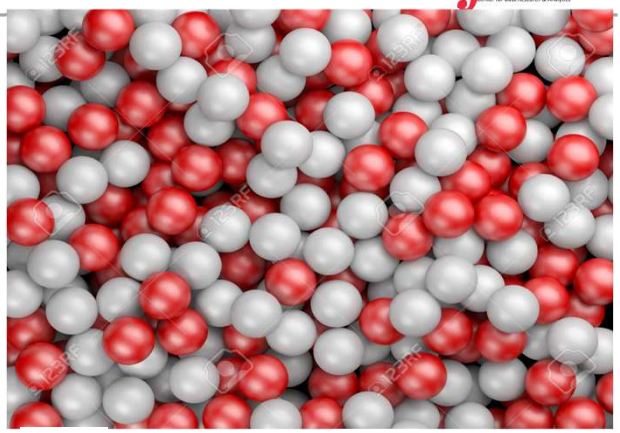
- When cases are happening more than the expected level in a population; we define that as an outbreak
- Find cases systematically and record information (demographic, clinical, risk factors, diagnosis) and look at the similarity and dissimilarity among the cases.
- Sometimes setup the surveillance to track the epidemic
- Outbreak investigation describes the patient population only



The purpose of outbreak investigation is to understand the pattern of the disease by summarizing and visualizing the data.

Cross-Sectional Study

- When disease is frequent and stable in the community, we prefer CSS to describe the population
- This method investigates all subjects in the population at a certain point of time
- Apply the rigorous statistical method to summarize the data
- Investigates the subjects at risk with this study design simultaneously.
- A CSS may be repeated multiple time to captures the time dependency of outcomes.



= 2,000 = 9000

The cross-sectional study design is good for those characteristic of the population which rarely changes with time

Proportions

- Proportions are a part of the whole set (population).
- Denominator is the number of subjects in a set and numerator is the number of subjects have specific characteristics in the set.
- It is as if the mean of binary variables.
- Often defined as risk or probability or relative frequency

- In a population of size 10, 6 were found smokers.
  - Frequency (smoker) = 6
  - Relative frequency = 6/10
- In another population of size 100, 40 were smokers
  - Frequency = ?
  - Proportion = ?

Frequencies are not comparable, but proportions are

Proportion is comparable across the populations

Prevalence/Incidence

- Disease prevalence and incidence both represent proportions of a population were diseased at a certain time.
  - Define a time scale for the start and end and event time
  - Time can be measured as the age of subjects
  - The time from exposure to a specific risk factor,
  - Calendar time
  - Time of diagnosis
- The numerator and denominators are defined differently in prevalence and incidence estimates (in the next slide)

For example

- Today could be the start of time or
- Each subject's time starts with the date of birth
- Time starts with the first diagnosis of a disease

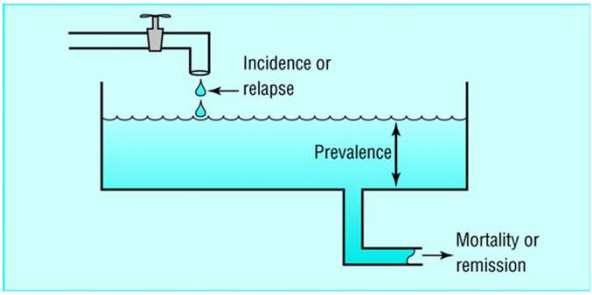
Time plays a central role in prevalence/incidence estimates.

Prevalence vs incidence

- **The prevalence**

$$P_t = \frac{\#\ of\ diseases\ at\ time\ t}{Total\ \#\ of\ subjects\ were\ at\ risk\ at\ time\ t}$$

Prevalence can me measured at a specified point or in an interval on the time scale.



- **The incidence**

$$I_t = \frac{\#\ of\ \textbf{New cases}\ of diseases\ by\ the\ end\ of\ time\ interval}{Total\ \#\ of\ subjects\ were\ at\ risk\ at\ the\ beginig\ of\ a\ time\ interval}$$

The incidence, also called cumulative incidence

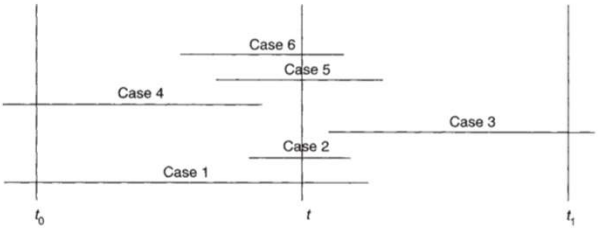
Both statistics measures the risk of disease at a certain time point in a populations

Prevalence calculation

- What is the prevalence at time t
  - # of person at risk at time
  - # of infected person at time t
- What is the interval prevalence during time $[t_0, t_1]$
  - # of person at risk during the time interval
  - # of infected person during the time interval
- The prevalence

$$P_t = \frac{\#\ of\ diseases\ at\ time\ t}{Total\ \#\ of\ subjects\ were\ at\ risk\ at\ time\ t}$$

Let us do it together



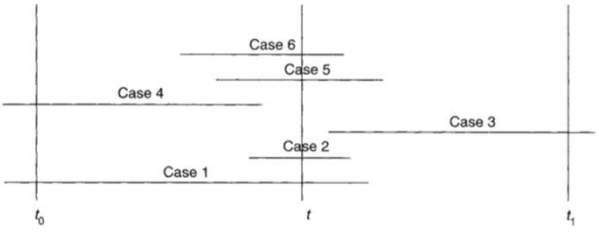
Six cases of disease in a population of, 100 individuals are represented. Lines represent the duration of disease.

Incidence/Cumulative Incidence Calculation

- What is the cumulative incidence in the time interval $[t_0, t_1]$.
- Total number of subject at the beginning of the time interval
- Number of new cases in the time interval
- The Incidence

$$I_t = \frac{\#\ of\ \textbf{\textit{New cases}}\ of diseases\ by\ the\ edn\ of\ time\ interval}{Total\ \#\ of\ subjects\ were\ at\ risk\ at\ the\ beginig\ of\ a\ time\ interval}$$

Let use do it together

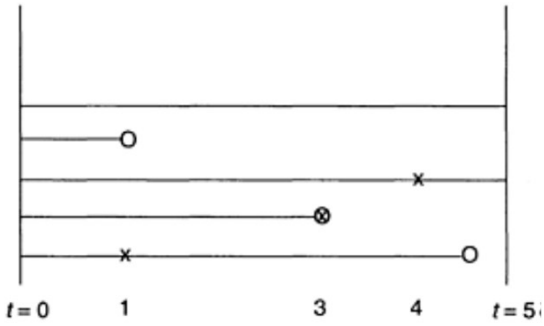


Six cases of disease in a population of, 100 individuals are represented. Lines represent the duration of disease.

Incidence rate

- The Incidence rate
- $IR = \frac{\#\ of\ New\ Case}{Total\ amount\ of\ person{-}tim\ \ at\ risk}$

- Calculation of the total amount of time at risk for the disease accumulated by the entire population over the time interval commonly measured in person time.
- cumulative incidence does not consider the amount of time at risk



Population of 5; the symbols represent: O, death; x, incident case of disease. Here, lines represent time alive

IR during *t*[0 , 5] is 3/(5+1+4+3+1)=3/14=0.21 cases/year

Incidence rate is the incidence per unit of time or person or person-time

Parkinson's Disease (PD) in Asia

Results from a study among the geriatric's

- The prevalence PD in Asia ranged from 35.8 to 68.3 per 100, 000
- The standardized incidence rates were 8.7 per 100 000 person-years in door-to-door surveys and
- 6.7 to 8.3 per 100 000 person-years in record-based surveys.

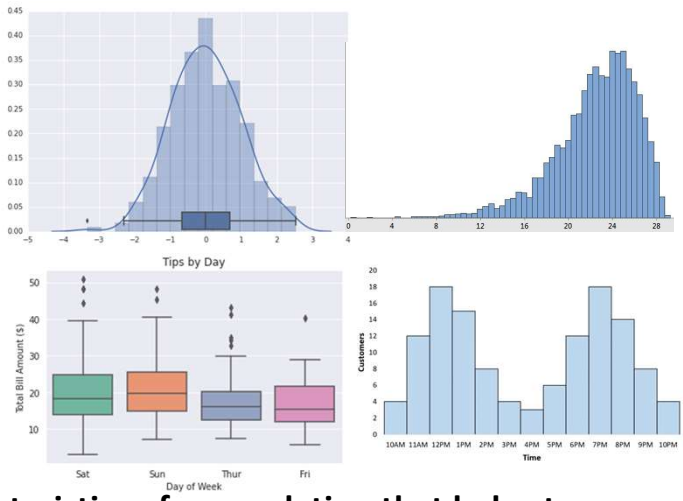
Reference:

Muangpaisan, W., et al. (2009). "Systematic review of the prevalence and incidence of Parkinson's disease in Asia." Journal of epidemiology: 0909290109-0909290109.

Let us interpret together

Descriptive statistics

- Familiar descriptive statistics for continuous variables are mean, median, standard deviation, quartile.
- Relevant informative charts are boxplot and histogram that captures the descriptive statistics
- Conditional means and boxplot gives the dependency pattern between the variables

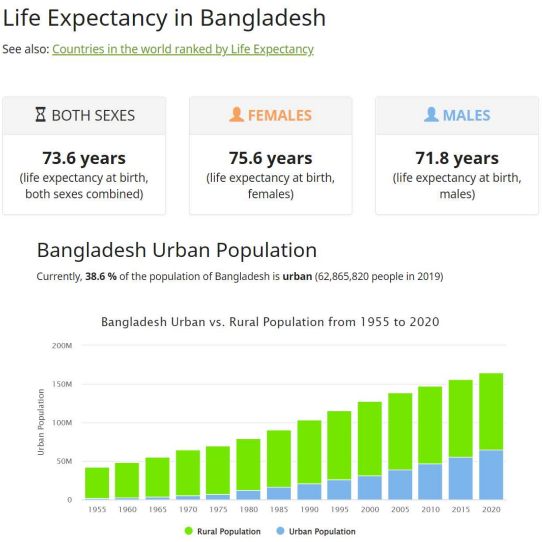


Descriptive statistics provides characteristics of a population that helps to make decisions

Conditional Mean/proportion

- Crude way to define, statistics in subgroups of the populations.
- Bangladeshi people survive 73.6 years (life expectancy). Males survive 71.8 years and females 75.6 years.
  - LE = 73.6 Y independent of sex
  - LE/male = 71.8
  - LE/Female = 75.6
- The second graph
  - Population/time and place
- A difference between groups can be translated as association between outcome and group variable.

We look at each sub-groups for the parameters



Regression

Consider the example on the right, we find two different life expectancy in females and males. We define as:

- Life expectance depends on gender.
- Life expectancy = f(gender)
- Model: LE = a + b gender + random error.
  - On an average the errors are zero

E(LE) = a + b gender

Interpretation:

- E(LE) = Population Average.
- a = E(LE/male or female) and
- b = difference of E(LE) between the gender.

Life Expectancy in Bangladesh

See also: Countries in the world ranked by Life Expectancy

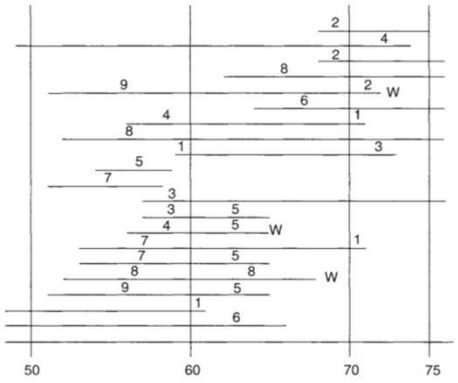
BOTH SEXES	FEMALES	MALES
73.6 years (life expectancy at birth, both sexes combined)	**75.6 years** (life expectancy at birth, females)	**71.8 years** (life expectancy at birth, males)

If we perform a regress LE over gender with females as reference group, then we would get
a = 75.6 and b = -3.8

Regression provides all conditional measurements of an output in the population

What if we remove the gender variable from the equation. What the regression will return?

Proportion, Incidence, prevalence, and IR



Schematic showing onset of disease at different ages in population of 20 individuals.

- Proportion of disease at age 60
- The incidence proportion for the disease between the ages 50 and 60 (assume that the disease is chronic so that those with the disease are no longer at risk).
- The incidence proportion between
  - ages 60 and 70 and
  - ages 70 and 75.
- The incidence rate for the intervals
  - ages 50 to 60,
  - ages 60 to 70, and
  - ages 70 to 75.
- Comment on your findings.

Submit your response in https://github.com/jaynal83/sphr

The figure illustrates observations on 20 individuals in a study of a disease D. The time (horizontal) axis represents age. The lines, one per individual, represent the evolution of follow-up: the left end point is the start of follow-up, the right endpoint indicates the age at onset of D except in cases of withdrawal (W). One example is, the first individual's (lowest horizontal line just above the axis line) follow-up started before his 50th birthday and developed disease D early in his 67th year, that is just after turning 66. The number with each line represents how many years passed within that the interval where the endpoint belongs.