Introduction to Statistics

LAKSHIKA TENNAKOON- MD, MPHIL
TRAUMA, ACUTE CARE AND CRITICAL CARE SURGERY
STANFORD UNIVERSITY
EMAIL: LAKSHIKA@STANFORD.EDU

January-2022

Objectives

To support, Medical Students in their Research Interests and Statistics.

To provide Methodological and Analytical support for Research Projects.

Why study Statistics?

- Data are everywhere!
- Statistical techniques are used to make many decisions that affect our lives.
- No matter what your career, you will make professional decisions that involve data.

- The science of collectiong, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.
- Statistical analysis – used to manipulate summarize, present, and investigate data in a meningful way.

Statistical Data

- The collection of data that are relevant to the problem being studied is commonly the most difficult, expensive, and time-consuming part of the entire research project
- Statistical data are usually obtained by counting or measuring items
  - **Primary data** are collected specifically for the analysis/project.
  - **Secondary data** have already compiled and are available for statistical analysis (NIS, NEDS, KIDS…).
- A **variable** is an item of interest that can take on many different values.

Sampling

- A sample should have the same characteristics

as the population it is representing.

- Sampling can be:

with replacement: a member of the population may be chosen more than once (picking the candy from the bowl).

without replacement: a member of the population may be chosen only once (lottery ticket).

Sampling Methods

Sampling methods can be:

- **random:** each member of the population has an equal chance of being selected **eg: Randomized controlled trials.**

- **nonrandom**

The actual process of sampling causes **sampling errors**

For example, the sample may not be large enough or representative of the population.

- Factors not related to the sampling process cause non sampling errors.

Type of Statistics

Descriptive statistics – Methods of organizing, summarizing, and presenting data in an informative way

Inferential statistics – The methods used to determine something about a population on the basis of a sample

- Population –The entire set of individuals or objects of interest or the measurements obtained from all individuals of interest.
- Sample – A portion, or part, of the population of interest.

Descriptive Statistics

Collect data

- Survey, case-control study, cohort, RCT etc

Present data

- Tables, pie charts, and graphs

Summarize data

- Sample mean, SD,SE

Data…….

- Data categories:

Qualitative - data are measurements that each going into one of several categories (hair color, ethnic groups and other attributes of the population).

Quantitative - data are observations that are measured on a numerical scale (distance traveled to college, number of children in a family, etc.)

Qualitative data

Qualitative data are generally described by words or letters

They are not as widely used as quantitative data.

Qualitative data can be separated into two subgroups:

- **Dichotomous** (if it takes the form of a word with two options (gender - male or female)
- **Polynomic** (if it takes the form of a word with more than two options (**education :** primary school, secondary school and university).

Quantitative data

Quantitative data are always numbers and are the

result of counting or measuring attributes of a population

Quantitative data can be separated into two

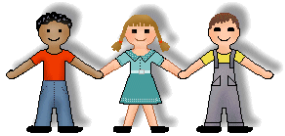
subgroups:

discrete (if it is the result of *counting* (the number of students of a given ethnic group in a class, the number of books on a shelf, …)

continuous (if it is the result of *measuring* (distance traveled, weight of luggage, …)

Data types

- **Variables**
  - **Scale**
    - **Continuous**
      Measurements
      takes any value
    - **Discrete:**
      Counts/ integers
  - **Categorical**
    - **Ordinal:**
      obvious order
    - **Nominal:**
      no meaningful order



Data Presentation

Frequency distribution – Frequency distributions are used to summarize large volumes of data values

tab gender,m

Indicator of gender

	Freq	Percent
Male	170,283	42.64
Female	228,948	57.33
Missing	96	0.02
Total	**399,327**	**100.00**

Data Presentation…….

Charts, pie, and graphs

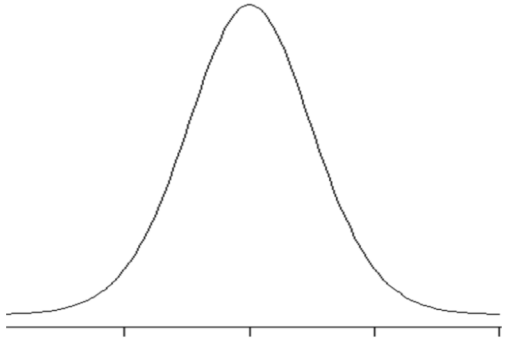
- Frequency distributions are good ways to present the essential aspects of data collections.
- Pictures are always more effective in displaying large data.
- Help to undestand the data distributions are normal or any abormality ?
- Help to identify key statistical tests to perform for the analysis!

Scale Data

If have scale data

To analyse it we often assume it follows a

Normal distribution

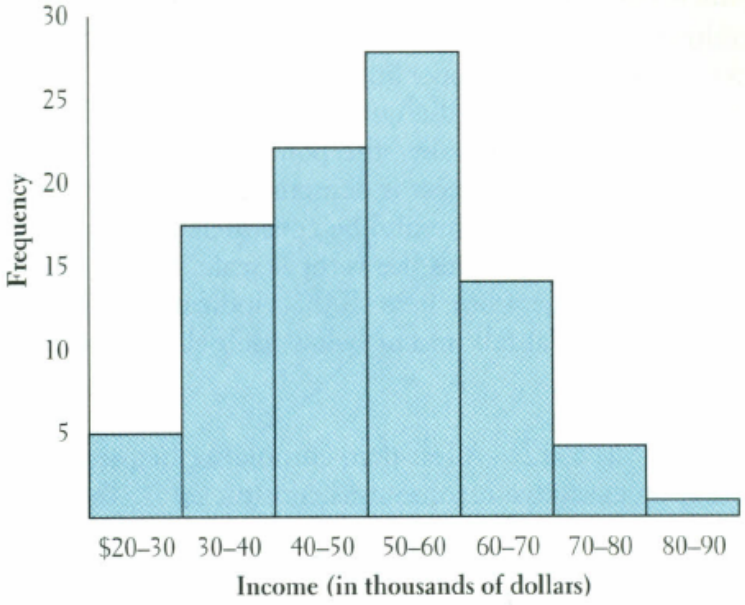


ormal

Histogram

- Frequently used to graphically present data.
- Is often used for interval and ratio data.
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes.

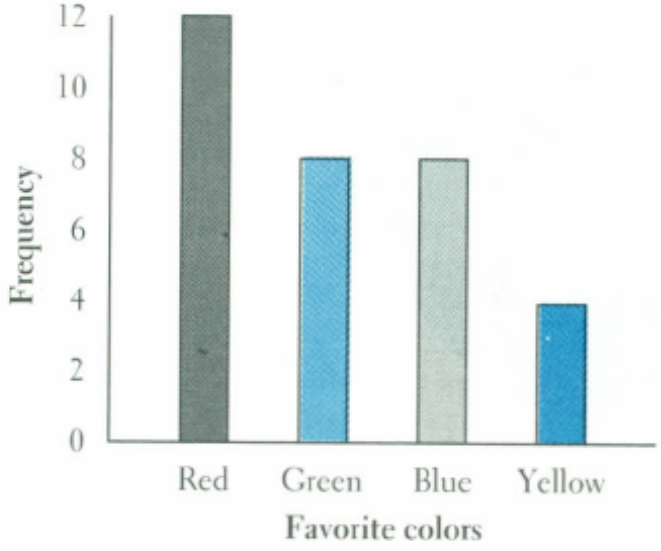
FIGURE 3.7 Histogram—Executive Incomes for the Sunrunner Corporation



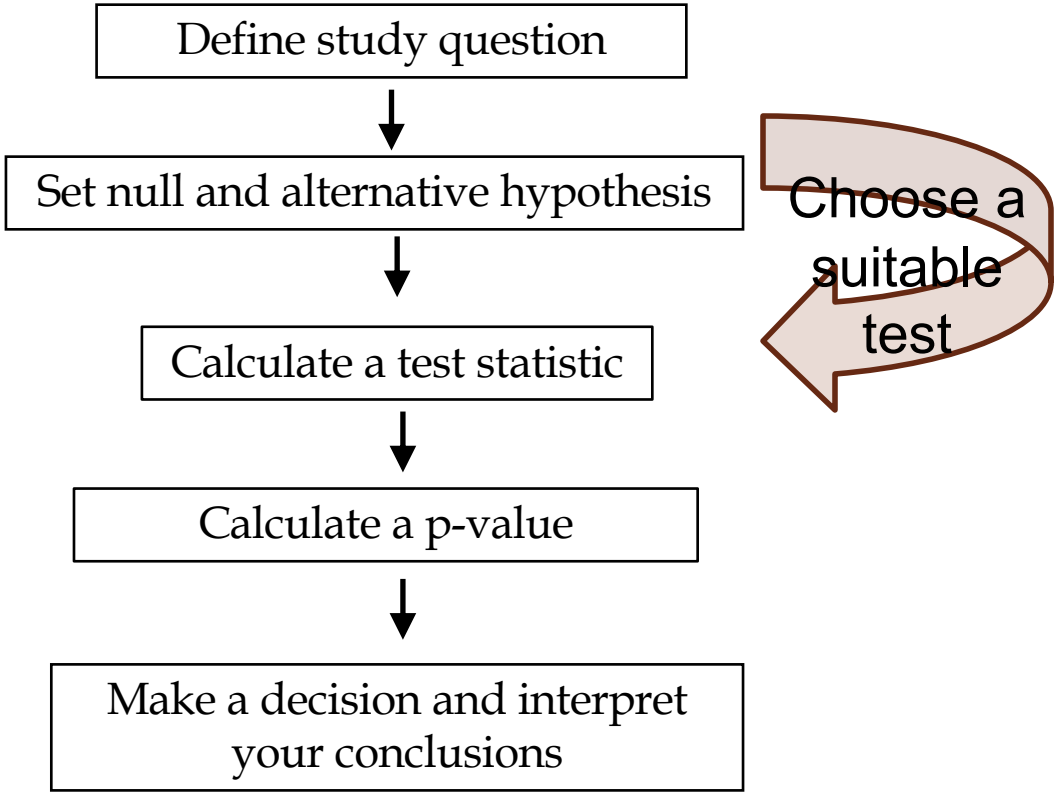
Bar Chart

- Another common method for graphically presenting nominal and ordinal scaled data
- One bar is used to represent the frequency for each category
- The bars are usually positioned vertically with their bases located on the horizontal axis of the graph
- The bars are separated, and this is why such a graph is frequently used for nominal and ordinal data

FIGURE 3.4 Bar Chart—Favorite Colors of 32 People



Steps to undertaking a Hypothesis test



Example: Titanic

The ship Titanic sank in 1912 with the loss of most of its passengers

809 of the 1,309 passengers and crew died

Death rate= 61.8%

Research question: Did class (of travel) affect survival?

Chi squared Test?

Null: There is NO association between class and survival

Alternative: There IS an association between class and survival

Class * Survived? Crosstabulation

Count

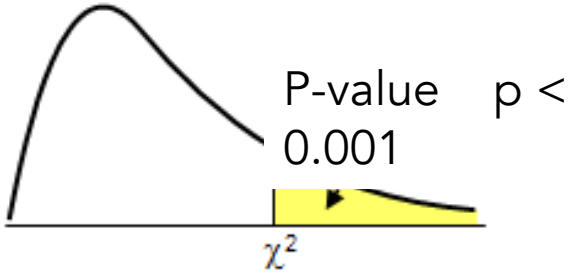
		Survived?		
		Died	Survived	Total
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1309

What's a p-value?

Probability of getting a test statistic at least as extreme as the one calculated **if the null is true**

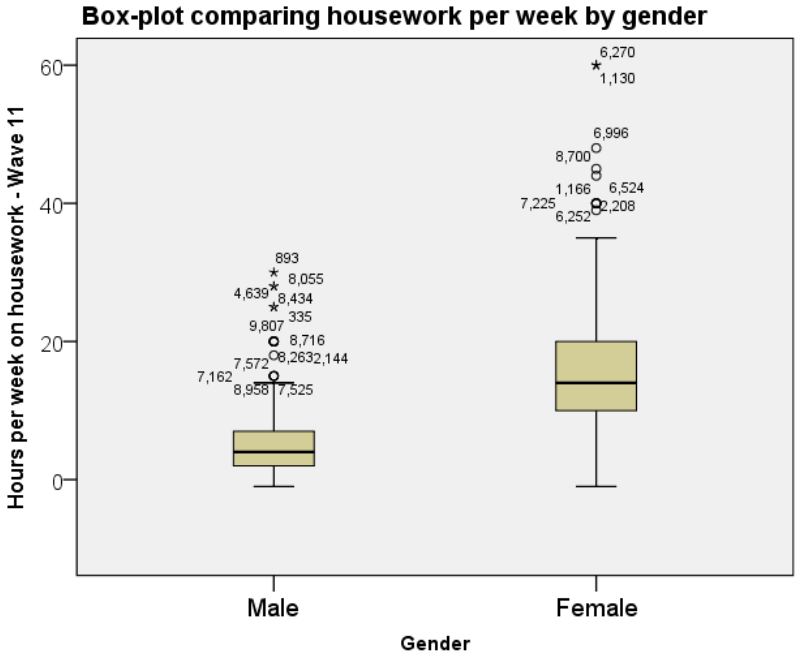
In Titanic example, the probability of getting a test statistic of 127.859 or above (**if the null is true**) is < 0.001

Distribution of test statistics



Summarizing Means

- Calculate summary statistics by group
- Look for outliers/ errors
- Use a box-plot or confidence interval plot



T-tests- Paired or Independent (Unpaired) Data?

T-tests are used to compare two population means

- **Paired data:** same individuals studied at two different times or under two conditions
  PAIRED T-TEST
- **Independent:** data collected from two separate groups INDEPENDENT SAMPLES T-TEST

What if the Assumptions are not met?

There are alternative tests which do not have these assumptions

Test	Check	Equivalent non-parametric test
Independent t-test	Histograms of data by group	Mann-Whitney
Paired t-test	Histogram of paired differences	Wilcoxon signed rank

Confidence Intervals

A range of values within which we are confident (in terms of probability) that the true value of a population parameter lies

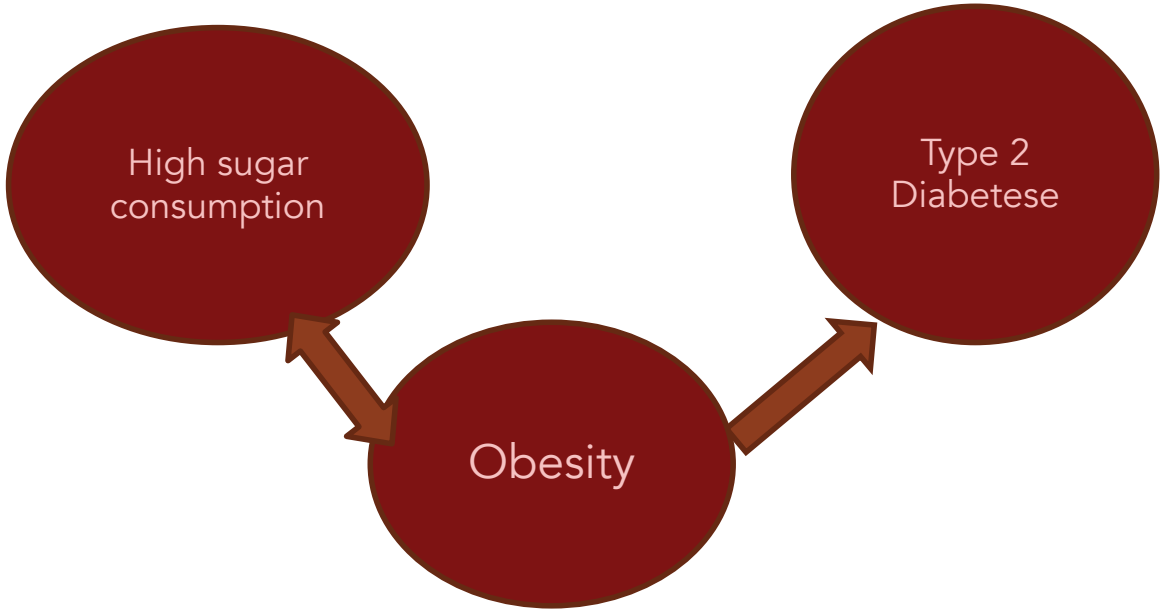
A 95% CI is interpreted as 95% of the time the CI would contain the true value of the pop parameter

i.e. 5% of the time the CI would fail to contain the true value of the pop parameter

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Triglyceride level at week 8 (mg/dl) - Triglyceride level at baseline (mg/dl)	-11.371	80.360	13.583	-38.976	16.233	-.837	34	.408

Confounding

Is there something else affecting both sugar consumption and diabetese?



Regression

Regression is useful when we want to

a) *look for significant relationships* between two variables

b) *predict* a value of one variable for a given value of the other

Logistic Regression

- Logistic regression has a binary dependent variable
- The model can be used to estimate probabilities
- **Example:** insurance quotes are based on the likelihood of you having an accident
- Dependent = Have an accident/ do not have accident
- Independents: Age , gender, occupation, marital status, annual mileage

Logistic regression	Number of obs = 1,028,436		
	LR chi2(26) = 9963.83		
	Prob > chi2 = 0.0000		

DIED	Odds Ratio	Std. Err.	z	P>z	[95% Conf.	Interval]
FEMALE	0.8881598	0.0082406	-12.78	0.000	0.8721545	0.9044589
_IPAY1_2	1.046217	0.0222354	2.13	0.034	1.003532	1.090718
_IPAY1_3	1.076135	0.0178496	4.42	0.000	1.041713	1.111695
_IPAY1_4	1.16044	0.0410167	4.21	0.000	1.08277	1.243681
_IPAY1_5	0.8331684	0.118226	-1.29	0.198	0.6308817	1.100317
_IPAY1_6	1.59524	0.0475147	15.68	0.000	1.504779	1.691139
_IAgecat_3	1.216527	0.0378275	6.3	0.000	1.144601	1.292974
_IAgecat_4	1.836373	0.0593259	18.81	0.000	1.723701	1.95641
_IAgecat_5	2.269361	0.0735235	25.29	0.000	2.129738	2.418138
_IAgecat_6	2.845507	0.0929346	32.02	0.000	2.669066	3.033612
_IRACE_2	0.9055029	0.0127098	-7.07	0.000	0.8809315	0.9307595
_IRACE_3	0.9728799	0.0171997	-1.56	0.12	0.9397465	1.007182

Choosing the Right Test

1) A clearly defined research question.

1) What is the dependent variable and what type of variable is it?

2) How many independent variables are there and what data types are they?

3) Are you interested in comparing means or investigating categorical relationships?

4) Do you have repeated measurements of the same variable for each subject? or one time measurement?

5) Do you like to conduct a logistic regression? Why?

Research Infrastructure: Trauma and Acute Care Surgery

- Research Ethics/Regulatory
- Administrative Databases

Acquisition of wide range of national and local databases
Development of data quality protocols
Guidance on Data Use Agreements (DUA)
Linking DUAs to HCUP
Designing documentations based on rules for specific databases
Cataloging variable variation for different years and databases
Creating standardized statistical coding systems
Communications with HCUP/AHRQ scientific team
Development of statistical coding using wide range of statistical software

- Research Platform to support Surgery Residents and Medical Students (SWAT)

Research support for residents to conduct projects including developing research questions, study design, sample size calculations, methodology, data processing and analysis.

- **Data Security/Repository**

Data management and storage of large number of research and clinical databases

- **Project Consultation/Mentorship**

Protocol development (research design, sample size calculations, data processing)
Guidance on selection of an appropriate database for individual project
Methodology and statistical analysis
Policies on authorship
Multiple coaching
Creating de-identified datasets for various projects
Flexible consultations on biostatistics and methodology
Data processing, statistical coding and data analysis
IRB reviews and submissions
Data collection, coding, processing, data quality checks and storage
Catalogue ICD-9/ICD-10 codes for specific projects
Screening and recruitment of study participants
Systematic reviews and meta-analysis
Manuscript writing
Protocols for conference presentations and practice