# An Introduction to Statistics for Life Sciences



PROF. DR. AKIN PALA

# Today



- Syllabus
- Class procedures
- Introductions

- **Homework** :
- Successfully make it back to school tomorrow.

- **Statistics quote of the day:**
- '*Statistics is the grammar of science*'. Karl Pearson.

# Syllabus



- **Homework:** I intend to incorporate SAS computing assignments and class examples.
- We will be using the SAS system.
  - I encourage you to learn as much as possible about computing routines for the calculations done in the course.

1. Register for SAS OnDemand for Academics, free SAS on Demand program https://odamid.oda.sas.com/SASODARegistration/
2. Select SAS Studio
3. Course enrollment link: https://odamid.oda.sas.com/SASODAControlCenter/enroll.html?enroll=c5b5e9e8-e54a-4dd1-bd9d-851cae965053
4. For more information about SAS OnDemand for Academics, including step-by-step registration instructions, visit the following site: http://support.sas.com/ondemand/
5. Run SAS on https://odamid.oda.sas.com/SASStudio/

# Syllabus



- You will use turnitin program to return your homework, please register at

http://turnitin.com

# Lectures



- The lectures files:

http://members.comu.edu.tr/akin/Lectures/lectures.html

# Syllabus



- Please feel free to discuss homework and prepare for tests with a study group. I strongly encourage you to take this “group” approach.
- However, each person should construct his/her own computer programs and summaries of the output.
- Similarities to other students and internet sources will be checked by the turnitin program
  - You will earn your score based on the original material.

# Statistics, the numbers



- The word **statistics has two meanings. In the more common usage, *statistics refers to numerical*** facts.
- The numbers that represent
  - Income of a family
  - Age of a student
  - Starting salary of a college graduate

are examples of statistics in this sense of the word.

# Statistics field



- The second meaning of *statistics refers to the field or the discipline of study*.
- *Statistics is the science of collecting, analyzing, presenting, and interpreting data, as* well as of making decisions based on such analyses.
  - This course is mainly focused on analyzing and interpreting data.

# Statistics



- Study of methods for
  - collecting,
  - summarizing
  - analyzing data

and for making inferences from data.

- Data are measurements, counts etc. from which conclusions can be drawn.
- Data are obtained from
  - Experiments
  - Surveys
  - Observations

# Source of Data



- Data may be obtained from
- Internal sources
  - Company personnel files, accounting records, routine records in a farm
- External sources
  - Turkish Statistical Institute, newspapers, books
    - http://www.turkstat.gov.tr
- Surveys
  - Turkey has a lot of political survey companies.
- Experiments and observations (behavior)
  - Most important data source for life sciences.

# History of Statistics, Early Ages



- Statistical methods date back at least to the 5th century BC.
- Early statistics started as a state arithmetic to assist a king/ruler who needed to know the wealth and # of his subjects for taxation or to wage a war, in addition to basing policy on demographic and economic data.
- Several centuries later there are written records of empirical probability in ship insurance.
- Games of chance such as gambling led to the theory of probability.

# History of Statistics

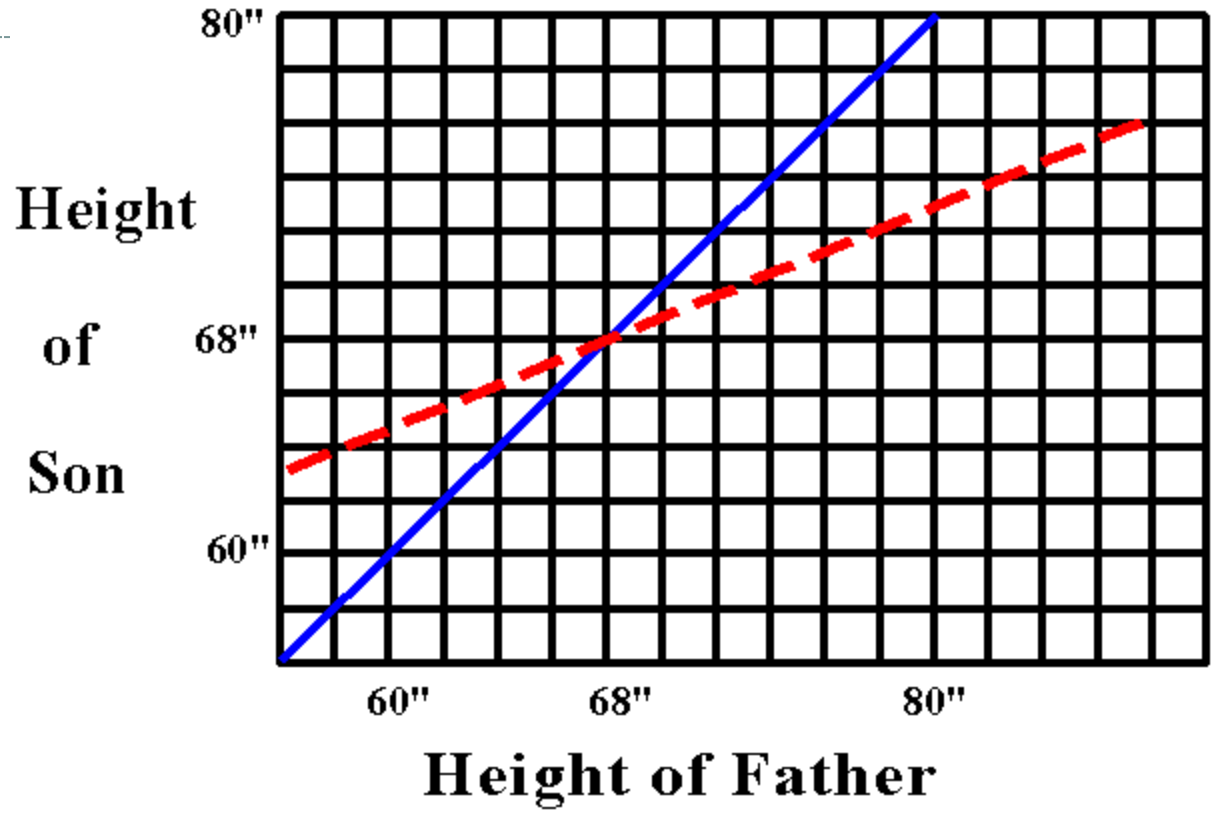


- Mendel had a biostatistical problem when he worked with plant hybrids.
- He worked with discrete characteristics, tall vs short, yellow peas vs. green peas etc.
- The scope of the discipline of statistics broadened in the early 19th century to include the collection and analysis of data in general.
- Today, statistics is widely employed in government, business, and natural and social sciences.

# History of Statistics



- The modern field of statistics emerged in the late 19th and early 20th century.
- Sir Francis Galton and Karl Pearson transformed statistics into a rigorous mathematical discipline used for analysis, not just in science, but in industry and politics as well.
- Galton introduced the concepts of standard deviation, correlation, regression.
- Galton studied inheritance of continuous characters, height in humans, intelligence in humans, etc. and applied these methods to those traits.




If the son's height were determined only by the father's height, the correlation should be that of the solid line. The dashed line is what is observed. Galton called this "regression to mediocrity."

# History of Statistics



- Pearson developed the Correlation coefficient.
- Galton and Pearson founded Biometrika as the first journal of mathematical statistics and biometry, and the latter founded the world's first university statistics department at University College London.
- Ronald Fisher coined the term "null hypothesis".
- Sir Ronald Fisher wrote textbooks defining the academic discipline in universities around the world.

# History of Statistics



- Fisher's most important publications were his 1916 paper “The Correlation between Relatives on the Supposition of Mendelian Inheritance” and his classic 1925 work “Statistical Methods for Research Workers”.
  - His paper was the first to use the statistical term, variance.
- Egon Pearson and Jerzy Neyman introduced the concepts of "Type II" error, power of a test and confidence intervals in the 1930s.
- Statistics continues to be an area of active research, for example on the problem of how to analyze Big data.

# Statistics Today



- College students from almost all fields of study are required to take at least one course in statistics.
  - A new way of thinking is introduced, thinking in terms of uncertainties, or chance.
- The study of statistical methods has taken on a prominent role in the education of students from a variety of backgrounds and academic pursuits.
- It is heavily used in academic research for analysis of results, regardless of the science field.
  - Computers and statistical software packages have enlarged the role of statistics as a tool for empirical research.

# Statistics is a Tool



- Statistics is a science fields itself.
  - It is a science if you don't have to add the word "science" to it ☺
- For all other fields, statistics is a tool for research.
- The research will be in genetics and molecular biology.
- It is the field of research, not the tools that must supply the "why" of the research problem.
- Sometimes this is overlooked and users are tempted to forget that they have to think
  - Statistics cannot think for researchers.
- Statistics can help design experiments and evaluate the resulting numerical data objectively.

# Statistics, general definition



- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.
- Statistics deals with everything related to data including the experimental design regarding planning of data collection.
  - Mistakes can be averted before it is too late.
  - A huge mistake: a single observation per group ☺
    - No within group variation (no error)

# Types of Statistics



- Theoretical or mathematical statistics deals with the development, derivation, and proof of statistical theorems, formulas, rules, and laws.
- Applied statistics involves the applications of those theorems, formulas, rules, and laws to solve real-world problems.

# Types of Statistics



- We are concerned with applied statistics and not with theoretical statistics.
- By the time you finish this class, you will have learned how to think statistically and how to make educated guesses.
- Applied statistics can be divided into two areas:
  - *descriptive statistics*
  - *inferential statistics*

# Descriptive statistics Terminology



- Descriptive statistics consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.
- Test scores of students enrolled in a Genetics class.
  - The whole set of numbers that represents the scores of students is called a **data set,**
    - The scores is the variable
  - The name/number of each student is called an element, member or experimental unit, the students are called **subjects**.
  - The score of each student is called an **observation**
- Normally, data sets are very large and we need summaries, tables and summary measures
  - Averages, standard deviations

to make sense of all the info.

# Inferential Statistics Terminology



- The collection of all elements of interest is called a **population**
  - **In the previous example, all students.**
- The selection of a number of elements from this population is called a **sample**
  - **Sample of students we have taking the class.**

# Inferential Statistics Terminology



- Statistics mainly deals with making decisions, inferences, predictions, and forecasts about populations based on results obtained from samples.
- A sample that represents the characteristics of the population as closely as possible is called a **representative sample**
  - I sampled blondes, but most of them have dark hair and skin? It is not a representative sample.
  - I sampled geniuses but some of them are dumb as a brick ☺ Then it is not a representative sample.

# Inferential Statistics Terminology



- If we want to find the milk yield of a typical Holstein cow, we may select 100 Holstein cows, find their lactation yield, and make a decision based on this information.
  - Run an association study, find a lac SNP that results in higher lactation milk yield. Animals with that SNP are expected to produce more milk per lactation.
- The area of statistics that deals with such decision-making procedures is referred to as **inferential statistics.**

# Probability



- Probability is a link between descriptive and inferential statistics.
- Gives a measurement of the likelihood that a certain outcome will occur.
- Probability makes inferences about the occurrence or nonoccurrence of an event under uncertain conditions.

# Inference from Data



- Inference from data is uncertain.
- A scientist must reason from particular cases to wider generalities, which is uncertain inference.
- This process enables us to disprove incorrect hypotheses, but does not let us prove correct hypotheses.
  - Reject the null hypothesis,
  - Failed to reject the null hypothesis.
    - We could not find out what the factors making up the differences are, and we call them error, or chance.
- The role of statistics is to quantify this uncertainty.
  - Why uncertain? Why only prove beyond a reasonable doubt?

# Inference from Data



- Why not be right all or very close to being right all the time?
- The drawback is the cost.
- Cost may rise because of
  - increased sample size,
  - penalty of a wrong decision,
  - the vagueness of the inference necessary to include the correct answer.

# Inference from Data



- Inductive inference
  - Specific → general conclusions
  - If all 57 crows we have are black, all crows must be black
- Deductive inference
  - If all crows are black, and the bird on the tree is black, then the bird must be a crow.
- Fun fact: A group of crows is called a murder, and not a herd.

# Inference from Data



- Every day we make decisions that may be personal, business related, or of some other kind.
- Usually these decisions are made under conditions of uncertainty.
  - Many times, the situations or problems we face in the real world have no precise or definite solution.
- Statistical methods help us make scientific and intelligent decisions in such situations.

# Inference from Data



- Decisions made by using statistical methods are called *educated guesses*.
- *Decisions made without using statistical (or scientific)* methods are *pure guesses and, hence, may prove to be unreliable*.
  - *For example, opening* a large store in an area with or without assessing the need for it may affect its success.

# Inference from Data



- Specific → General
- Sample → Population
- Population→ set of all measurements of interest.
  - Preferences of every voter in elections in Turkey
    - Not the people. People are the experimental units.
  - Yield obtained on every field in Turkey if this variety of wheat were planted.
- Sample→ set of measurements that are observed, a subset of population.
  - Preferences of 300 voters interviewed in a survey.
  - Yield from plots at 6 experimental stations.

# Sample→Population



- Census: a sample consisting of the whole population.
- If we knew the whole population, then there would be no need for inference statistics, we could just calculate means, frequencies etc.

# Sample→Population

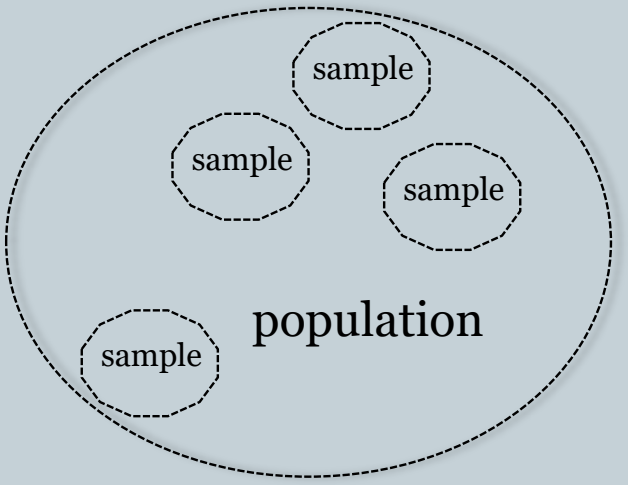


- Currently we are mostly using statistics to reach conclusions about the population of our samples.
- If you milk Holstein cows 6 times a day, they produce more milk.
- This conclusion is reached with about 100 cows, but could be applied to the majority of the population, provided the environmental conditions (feed, comfort, cleanliness) are sufficient.

# Uncertainty



- We conclude based on info obtained from sample.
- Why the uncertainty?
- There are many possible samples to be drawn from population, we may have drew just one of them.
  - Therefore sample size and variation is important.




The larger the sample size, the more closely the sample distribution looks like the normal distribution.

# Measure of Uncertainty



- In a poll of 1200 voters, 52 % favor candidate A
  - Margin of error (from confidence interval) in this poll is 3 %
- The voters for A is likely between 49 % and 55 %

# Qualitative/Quantitative Variables



- Qualitative (or categorical or attribute) data
  - can be separated into different categories that are distinguished by some nonnumeric characteristics.
  - Examples:
    - gender (male/female) of professional athletes
    - nationality of people in a room.
  - Numerical measurement is not possible.
- Quantitative data can further be distinguished between discrete and continuous types.
  - A variable that can be measured numerically is called a quantitative variable.
  - The data collected on a quantitative variable are called quantitative data.

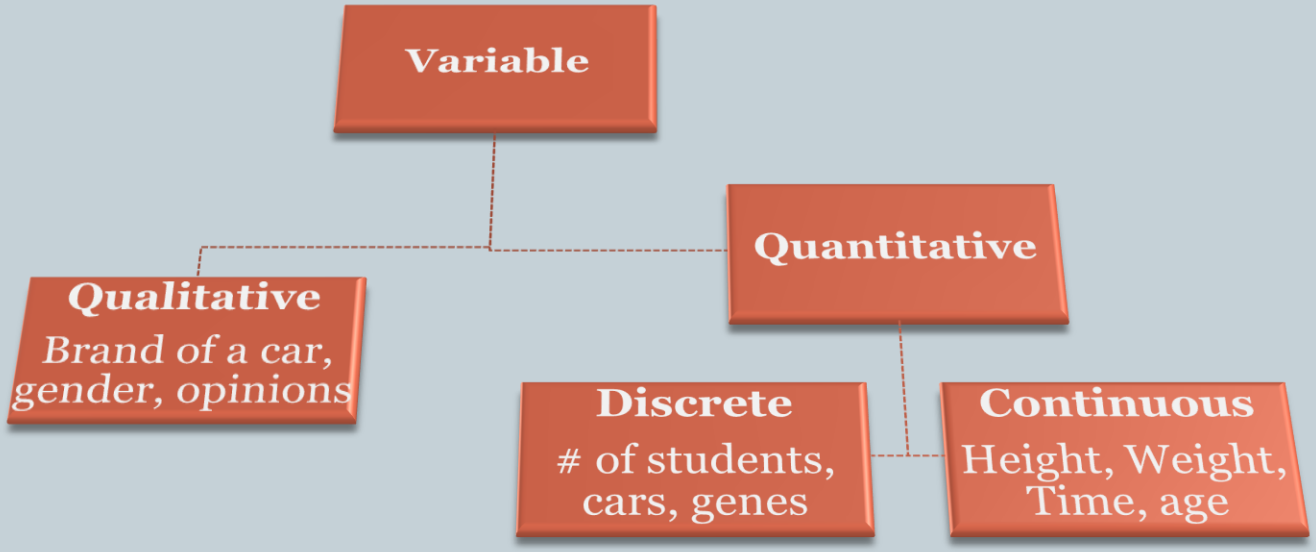
# Discrete/Continuous



- Discrete=discontinuous variables
  - The number of petals on a flower
  - Presence of horns on goats (zero or one)
  - Litter size in mice (baby mouse: pup, pinkie)
  - Number of enzymes
- Continuous variables
  - Surface area of petals on a flower
  - Surface area of loin eye
  - Height
  - Weight
    - Oz: ounce, lbs: pound (Roman pound=libra)
  - Amount of an enzyme

# Qualitative/Quantitative Variables





# Levels of Measurement



- Another way to classify data is to use levels of measurement.

# Definitions



- Nominal level of measurement
- Characterized by data that consist of names, labels, or categories only.
- The data cannot be arranged in an ordering scheme (such as low to high)
  - Example: survey responses yes, no, undecided

# Definitions



- Ordinal level of measurement
- Involves data that may be arranged in some order, but differences between data values either cannot be determined or are meaningless
  - Example: Course grades A, B, C, D, or F
    - There is nothing in between in this, unless you have B+, B- etc.

# Definitions



- Interval level of measurement
- Like the ordinal level, with the additional property that the difference between any two data values is meaningful.
- However, there is no natural zero starting point (where none of the quantity is present)
  - Example: Years 1000, 2000, 1776, and 1492
  - Year zero is not really a starting point, it is still a year, BC years.
  - Earth is 4.543 billion years old $\pm$ 50 million years.
  - So we don't know the exact year.

# Definitions



- Ratio level of measurement
- The interval level modified to include the natural zero starting point
  - where zero indicates that none of the quantity is present.
- For values at this level, differences and ratios are meaningful.
  - Example:  Prices of college textbooks
    - $0 represents no cost

# Levels of Measurement



- Nominal - categories only
- Ordinal - categories with some order
- Interval - differences but no natural starting point
- Ratio - differences and a natural starting point
  - I prefer quantitative-qualitative distinction.

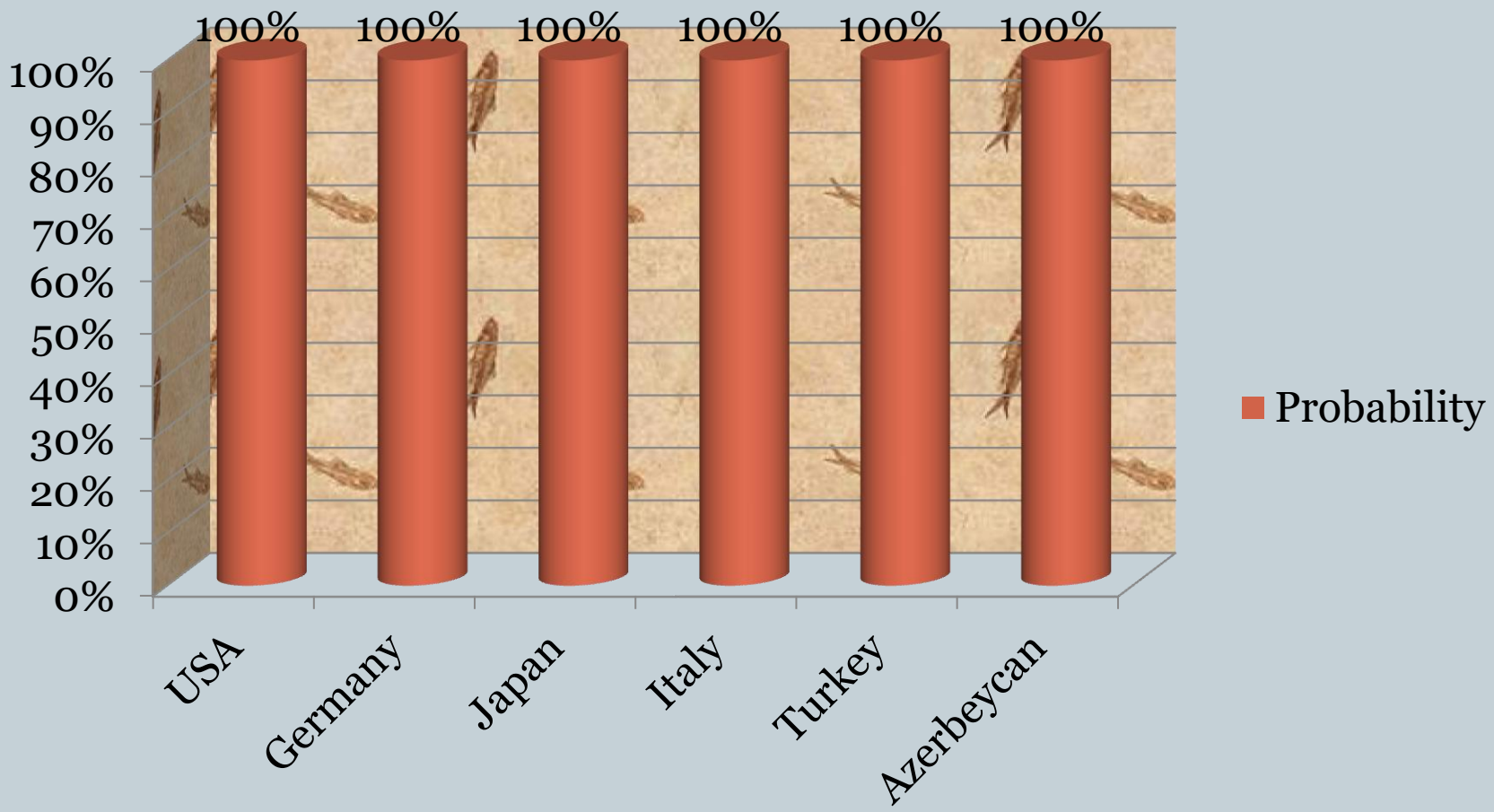
# Homework



- What is the probability that you will get an A this semester?

# Probability of Death by Country





# Homework 1



Lets make a data set, you don't have an obligation to provide your height ☺

| Subj | Gender | Age | City of Origin | Height |
|---|---|---|---|---|
| 1 | Male | 21 | izmir | 190 |
| 2 | Male | 21 | denizli | 178 |
| 3 | Male | 20 | manisa | 170 |
| 4 | Male | 20 | edirne | 185 |
| 5 | Male | 20 | tokat | 178 |
| 6 | Female | 20 | izmir | 162 |
| 7 | Female | 20 | kars | 168 |
| 8 | Female | 19 | halifax | 172 |
| 9 | Female | 21 | rize | 156 |
| 10 | Female | 20 | mugla | 165 |

Calculate averages of these numbers and make comments on the differences of the averages of gender and age. Optional: You can calculate interactions and make comments as well.