Welcome to Math 161B - Applied Probability & Statistics II

Let me introduce myself: My name is Martina Bremer, my office is in Duncan Hall (215) and the best way to reach me is usually by e-mail (martina.bremer@sjsu.edu). I will try my best to reply to you within 24 hours (36 hours on weekends).

For prerequisites, grading policies and exam information please refer to the course syllabus. The schedule as well as many other materials will be available throughout the semester on our course website:

http://www.math.sjsu.edu/∼bremer/Teaching/Math161B/

This website is also the place to go to find the homework assignments (and later their solutions), quizzes, review material for exams, my office hour schedule and more. Your homework, quiz, project, and exam grades will be posted on Canvas.

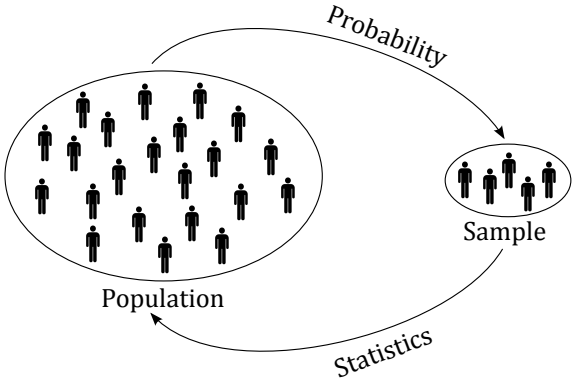
If you should ever find yourself falling behind in this course please come and see me as soon as possible so that we can discuss all the options you have of catching up. Math and Statistics classes tend to be very sequential. If you miss a section you will likely have difficulties following the subsequent sections.

The Goals of this Course

This course will provide an introduction to the most important statistical *methods*. Statistical methods are based on probability models. For each new situation (clinical trial, physics experiment, survey etc.) that generates data, we have to choose an appropriate probability model and assure that the assumptions of the model are satisfied. If they are, then that allows us to make deductions, estimations, and predictions.

What exactly is the difference between Probability and Statistics?

Probability is the science of studying the make-up of samples randomly drawn from a known population. Statistics is the science of drawing conclusions about the population from a sample.



Example: In the casino game of craps players roll two dice and bet on the outcomes. In the context of this game, phrase a relevant question that can be answered using probability. What are the assumptions you have to make in this case?

A statistician on the other hand, would use data to answer a question. Describe what the data looks like in the context of a craps game. Phrase a question that can be answered using the data. Think about how you would answer the question.

The Big Picture

Applied statistics is often utilized to learn more about a *population*. The word *population* is used in the abstract sense, it may mean people in a country, but also cells in an organism, cars that pass a certain landmark during a given day, goods produced by a certain manufacturer today, etc. If not every member of the population can be observed (*survey*), then a subset (*sample*) is chosen from the population. On each individual in the sample, one or more variables are observed and recorded (*data*).

The two main purposes of statistical data analysis are:

- Descriptive Statistics: Reporting data in numerically or graphically condensed form. For example, the sample mean or median can be used to represent a "typical" value of the data set.
- Inference: Fit a probabilistic model to the observations that accounts for random effects in the sample and represents the entire population. Use the data to draw conclusions about the model parameters (hypothesis tests, confidence intervals) and thus about the population.

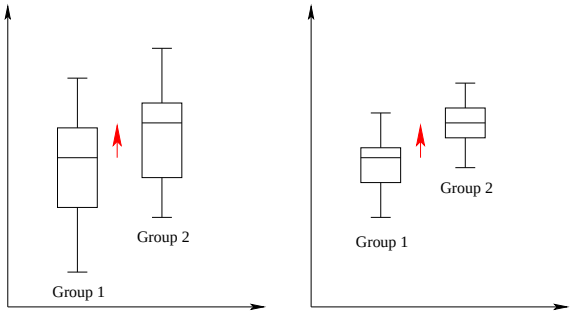
While descriptive statistics is an important first step in every data analysis project, this course will focus mostly on statistical inference methods. The course will teach you how to correctly apply these methods to given data sets and draw meaningful conclusions. Maybe more importantly, the course will increase your statistical literacy and help you recognize inaccurate or misleading statistical analyses.

> **Mark Twain** (??)
> There are three kinds of lies: lies, damned lies, and statistics.

The Main Topics

The two most important methods that you will learn about in this course are regression analysis and ANOVA. ANOVA stands for ANalysis Of VAriance. These two methods are connected (in ways which we will study later in the course), but for now let's take a look at their differences.

ANOVA: Suppose that you have randomly selected a sample from the population that you want to study. For simplicity, assume that you have collected data on two variables for each individual in the sample. One variable is *categorical* and assigns each individual into a group (e.g., healthy and diseased patients, plants or animals of different breeding strains, different makes of cars etc.). The other variable is *quantitative* (e.g., blood pressure of each patient, growth under drought conditions, time until transmission failure, etc.). It is to be expected that even within a group (e.g., healthy patients) not every individual will exhibit the same value of the quantitative variable (e.g., have the same blood pressure). There will likely be variation within the groups and possibly among the groups.



QUESTION: Does the quantitative variable takes on fundamentally different values in the groups defined by the categorical variable? Yes/No?

To answer this question, the idea behind ANOVA is to model the *response* (i.e., the quantitative variable) as a function of group means and errors. We will compare the variances of the errors *within* groups to the difference in means *among* groups. If the variance among groups is large compared to the variances within groups, we will decide that the groups are, indeed, fundamentally different. Depending on the context this may mean that we conclude that: the disease has an effect on blood pressure, a particular plant breed is more drought resistant than another, one car maker builds better transmissions than another.

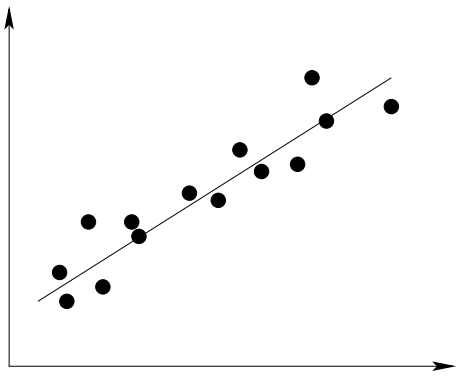
NOTE: Building up on this ANOVA model, we will also investigate scenarios in which individuals are categorized by more than one categorical predictor variable (e.g., male and female healthy and diseased patients and their blood pressure). It will then be possible to ask further reaching questions such as: Does the disease influence blood pressure in men and women differently?

Regression: Suppose (again) that you have randomly selected a sample from the population that you want to study. This time, you have collected two (or more) *quantitative* variables on each individual. And you want to study how the values of one (or several) of these variables (*predictors*) influence another variable (*response*).

Examples:

- How do the four C's (cut, color, clarity and carat) influence the price of diamonds?
- How do carbon emissions and other environmental variables influence global mean temperature over time?
- Can wages, asset holdings and the age of a worker be used to predict the average number of hours the person works in a week?

In a regression model, the response variable is written as a linear function of the predictor variables plus an *independent error term*. Statistical inference includes estimating the values of the model parameters, providing confidence intervals for them and conducting hypothesis tests on them. An important part of this analysis is to take a careful look at the error terms. Are they really (as the model demands) independent and identically distributed with a very specific distribution?



NOTE: After a regression model is fit, it allows us to make predictions about future observations and to study relationships between variables in the model. Building up on the basic regression models we will discuss topics such as variable selection: which predictors are important for the response? Are there maybe some predictors that should better be left out of the model? What criterion should we use to answer this question?

Recap:

	Response	Predictor(s)
ANOVA model	quantitative	categorical
Regression model	quantitative	quantitative