Erika Anderson, Maison Beckley, Catie Cleveland, Nick Gammal, Olivia KOhler, Shannon Maguire

STAT 313

Homework 1

1. What do you believe are similarities and differences between Statistics and Mathematics?

|  |  |
| --- | --- |
| Differences | Similarities |
| * Statistics addresses probability of outcomes * Statistics makes predictions, math gives definite solutions and answers * Statistics involves observing patterns, interpreting trends, deciphering relationships within and between datasets while math looks less at varying data trends and probable relationships and more at known, defined relationships that always hold true | * Both use numbers and equations to represent relationships between things in the real world. * Both are programs in the math department * Both are commonly used as tools for science that help us understand the world around us. |

1. From your perspective(s), what is Statistics? Is there a difference between “statistics” and “Statistics”? If so, what?

Statistics to me is the sampling and collecting data from populations, analyzing and interpreting the data in a reasonable way in order to understand relationships and trends. Statistics is a tool that allows us to represent the real world in an abstract way. This is important for solving problems. I would say that “Statistics” encompasses the entire subject, whereas “statistics” refers to the outcomes of statistical tests that have been performed. So, in Statistics, a common statistic to calculate is the t-statistic.

“Statistics” includes the entire practice of collecting and analyzing data whereas “statistics” are measurements or characteristics of a sample that describes that data. It describes the entire process of collecting and recording data, running the analysis and interpreting the results to help make decisions on real world issues. Generally, when someone says, “look at the statistics” they are referring to the results of the analysis and what that may mean for decision making based on those results.

1. Just because a variable has numeric values, does not mean it is a numeric variable. What are the different types of variables that can appear in a dataset? How can you determine if a variable is numerical versus categorical?

Categorical and numerical data can appear in different datasets. In categorical variables, there are certain sections or categories that observational units are assigned to. There are two types of categorical variables. Categorical nominal variables have no specific order of listing. Categorical ordinal variables, on the other hand, have intrinsic ordering when they are listed. An example of a categorical nominal variable would be if students were assigned to groups 1-4. In this case there are four groups that students must fall into. This would be a nominal categorical variable, because it does not matter what order the groups are in. Alternatively, as stated, an ordinal categorical variable would have a natural order. For example, to get to group 3, one would have to have advanced from 1 to 2, then from 2 to 3.

Numerical variables are the second type of variables in a dataset. However, just because a variable consists of numbers *does not* mean it is numerical. Numerical variables must be capable of being added, subtrated, divided, and averaged in a way that is reasonable. For example, similar to above, if students were assigned to groups 1-4, it would not give any statistical information to average numbers 1,2,3, and 4 together. Therefore, this data would not be numerical, but categorical as stated above.

Numerical data, similar to categorical data, also has two types. Continuous variables are a type of numerical data in which there is no gap between the possible values, and belong to the real number line. Continuous variables also must fall between two values so they can be quantified. Otherwise, without a starting or ending point, they would be unquantifiable and extend into infinity. For example, say you are looking at a data set of the temperature in Morro Bay. Between the lowest temperature recorded and the highest, there is an infinite number of outcomes, and we have basically a 0% chance of randomly selecting any given value. The second type of numerical variables are discrete numerical variables. Discrete numerical variables increase or decrease by whole-number values. For example, if you were to count students in a class, there would (hopefully) be no such thing as half of a student.

1. What are the differences between observational studies and experiments?

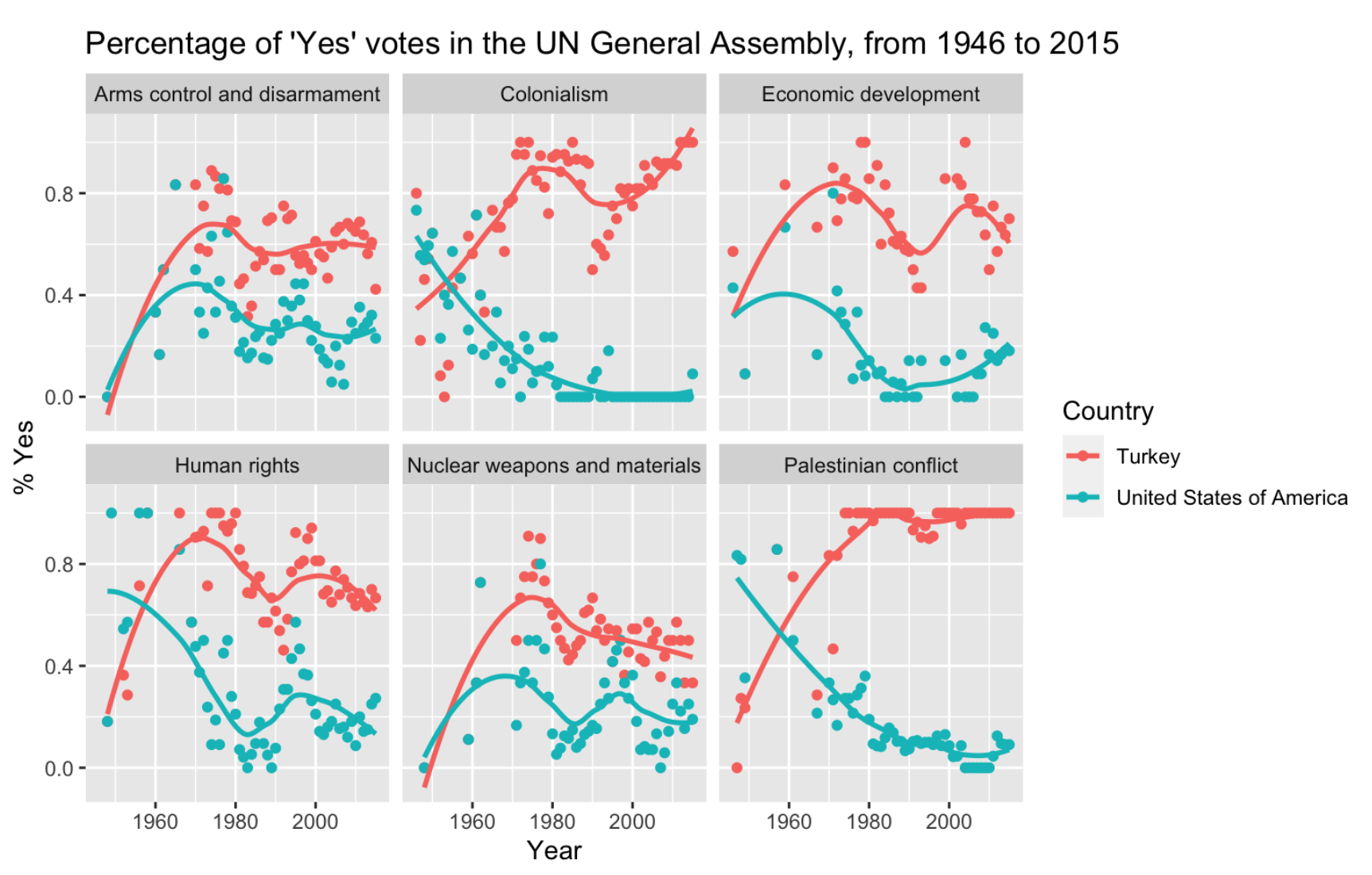
Observational studies do not manipulate the observational units, while experiments do manipulate experimental units. Watching bird flight patterns would be observational, but injecting apple trees with hormones would be experimental. The process for each of these would be as such: for observational studies, the order would be to observe the population, collect data, and analyze the data, and present the trends and relationships of interest. For an experimental study, the order would be to identify a population, manipulate the population in a desired way so there is an experimental treatment group(s) and a control, record the data, analyze it, and present the trends and relationships of interest. The control group within an experimental study is used as a reference point to quantify the effects of the treatment group on the observational unit. In an observational study, the observational units are not changed or manipulated in a way that would influence the data outcome. An observational study usually involves the collection of information using a cohort of samples or individuals in order to test hypotheses. Mostly, observational studies provide insights into whether two variables are associated, but they do not usually give evidence of causation that changing one of the variables definitely changes the other because they do not control for other factors possibly affecting the observational units. On the other hand, experiments seek to elucidate *causation,* not just association, between variables. Experiments, while still not 100% certain, test whether variables are associated and causally affect each other. Experiments use the idea of explanatory and response variables to test whether a change in explanatory variables elicits a change in response variables. An example of a randomized experiment would be testing the effectiveness of a plant fertilizer. A sample of plants would be randomly assigned into the control and treatment groups, and the fertilizer would only be administered to the treatment group with the control group receiving a placebo.

It’s important to note that even in controlled experiments, we cannot with 100% certainty declare causation. When we observe differences in the treatment group compared to the control group, we can only determine that there is a correlation between the variables. This is largely because most experiments have other variables that cannot be controlled. In an ideal world we could control every aspect of the experiment, but when working with people, plants, animals, etc. there are too many external factors that can influence your results.

1. What are different methods for sampling from a population? Why would a researcher choose a stratified random sample instead of a simple random sample?

A stratified random sample might be a more useful representation of a population than a simple one. For example, if a doctor is testing the efficacy of a drug, they might want to stratify the sample by separating biological sexes, or age groups. This is important because there are known differences in hormone levels among opposite sexes and those hormones can vary with age. Different bodies might react to the drug differently. A simple random sample gives every individual in a population an equal chance of being selected in the sample, but it may have a biased outcome by chance. A stratified random sample may be better in this case because it assigns the population into groups and randomly samples each group, resulting in a possibly more randomized and less biased-by-chance sample.

Another method of sampling from a population is cluster sampling where researchers can create clusters of the entire population, often geographically, whether that be clustering entire towns or schools, and then randomly sample from *entire* clusters within the population. The difference between a random cluster sample and a stratified random sample is that cluster samples are often easier to perform because your sample does not pull from all clusters. Instead, you sample a subset of the clusters. The disadvantage of a cluster sample is that it may be exclusionary, biased, or un-representative of the population.

1. Describe what you see in the plot below. 

The above data for Turkey and the United States are visualized as scatterplots with associated trendlines. Several of the subplots seem to have an inverse relationship to each other. This can be seen noticeably in the “Palestinian conflict” and somewhat in the “Colonialism” subplot. The relationships between year and percentage of ‘Yes’ votes seem to have a stronger correlation among the Colonialism and Palestinian Conflict voting patterns. Weaker associations are seen in the other topics.

The association in all of these plots is between the percentage of ‘yes’ votes by country in the UN General Assembly and the year those votes took place. What differs are the trendlines of each country’s votes on specific topics. In some cases, such as the Palestinian Conflict, Turkey’s and the USA’s voting patterns are inverses of each other, with an increasingly negative (more ‘no’ votes) trend for the USA and an increasingly positive (more ‘yes’ votes) trend in Turkey. Conversely, some topics such as arms control and disarmament have been voted on similarly by the two countries in recent history. This graphic helps the viewer visualize some of the complexity of international government at the UN.