Practice Exercises: Lesson 1.1

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro. https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

Exercises

1.1 Migraine and acupuncture, Part I. A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.[2]

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

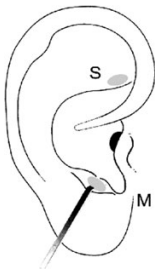


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

(a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

(b) What percent were pain free in the control group?

(c) In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

(d) Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

1.2 Sinusitis and antibiotics, Part I. Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. The distribution of responses is summarized below.[3]

		Self-reported improvement in symptoms		
		Yes	No	Total
Group	Treatment	66	19	85
	Control	65	16	81
	Total	131	35	166

(a) What percent of patients in the treatment group experienced improvement in symptoms?

(b) What percent experienced improvement in symptoms in the control group?

(c) In which group did a higher percentage of patients experience improvement in symptoms?

(d) Your findings so far might suggest a real difference in effectiveness of antibiotic and placebo treatments for improving symptoms of sinusitis. However, this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients in the antibiotic and placebo treatment groups that experience improvement in symptoms of sinusitis?

[2] G. Allais et al. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

[3] J.M. Garbutt et al. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

Exercises

1.3 Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[10]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.4 Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[11]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.5 Cheaters, study components. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. The study's findings can be summarized as follows: "Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls."[12]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

[10] B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

[11] J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

[12] Alessandro Bucciol and Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". In: *Journal of Economic Psychology* 32.1 (2011), pp. 73–78.

1.6 Stealers, study components. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken.[13]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

1.7 Migraine and acupuncture, Part II. Exercise 1.1 introduced a study exploring whether acupuncture had any effect on migraines. Researchers conducted a randomized controlled study where patients were randomly assigned to one of two groups: treatment or control. The patients in the treatment group received acupuncture that was specifically designed to treat migraines. The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. What are the explanatory and response variables in this study?

1.8 Sinusitis and antibiotics, Part II. Exercise 1.2 introduced a study exploring the effect of antibiotic treatment for acute sinusitis. Study participants either received either a 10-day course of an antibiotic (treatment) or a placebo similar in appearance and taste (control). At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. What are the explanatory and response variables in this study?

1.9 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[14]

(a) How many cases were included in the data?

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

(c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Photo by Ryan Claussen (http://flic.kr/p/6QTcuX) CC BY-SA 2.0 license

1.10 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[15]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

(b) How many participants were included in the survey?

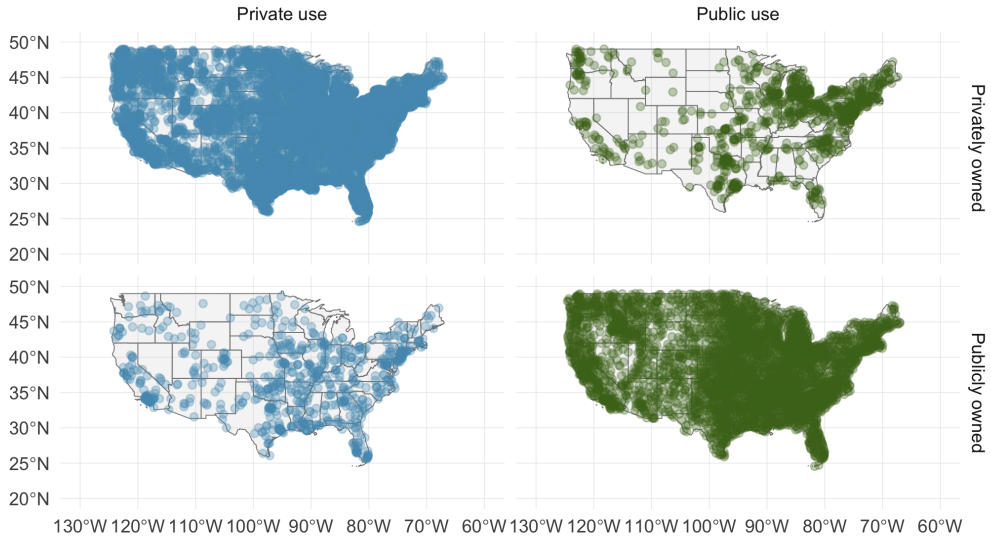
(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

[13] P.K. Piff et al. "Higher social class predicts increased unethical behavior". In: *Proceedings of the National Academy of Sciences* (2012).

[14] R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[15] National STEM Centre, Large Datasets from stats4schools.

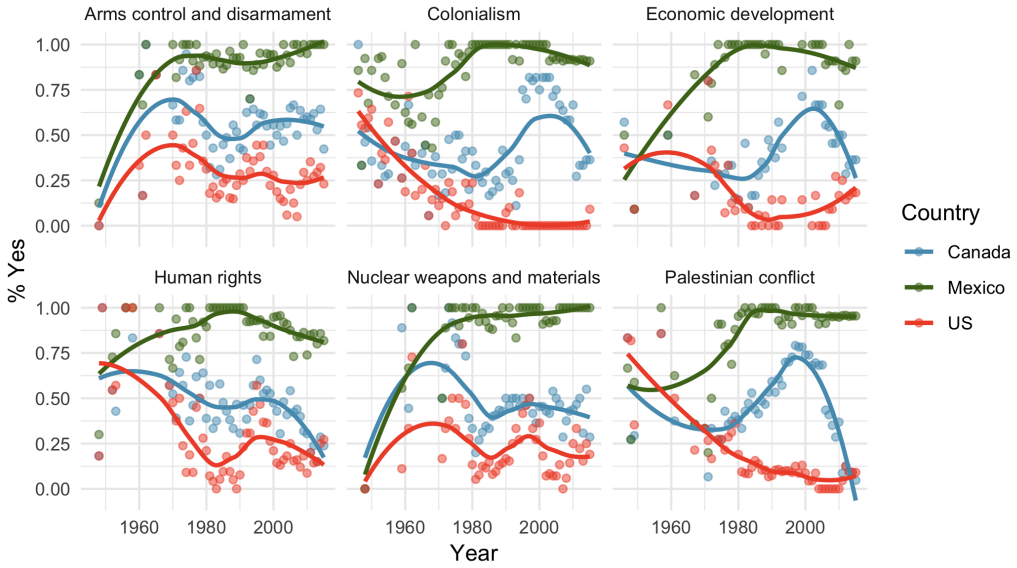
1.11 US Airports. The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.[16]



(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

1.12 UN Votes. The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.[17]



(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

[16] Federal Aviation Administration, www.faa.gov/airports/airport_safety/airportdata_5010.

[17] David Robinson. *unvotes: United Nations General Assembly Voting Data*. R package version 0.2.0. 2017. URL: https://CRAN.R-project.org/package=unvotes.

1.42 Screens, teens, and psychological well-being. In a study of three nationally representative large-scale data sets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being.[35]

(a) What type of study is this?

(b) Identify the explanatory variables.

(c) Identify the response variable.

(d) Comment on whether the results of the study can be generalized to the population, and why.

(e) Comment on whether the results of the study can be used to establish causal relationships.

1.43 Stanford Open Policing. The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.[36] The following is an excerpt from a summary table created based off of the data collected as part of this project.

County	State	Driver's race	No. of stops per year	% of stopped cars searched	% of stopped drivers arrested
Apaice County	Arizona	Black	266	0.08	0.02
Apaice County	Arizona	Hispanic	1008	0.05	0.02
Apaice County	Arizona	White	6322	0.02	0.01
Cochise County	Arizona	Black	1169	0.05	0.01
Cochise County	Arizona	Hispanic	9453	0.04	0.01
Cochise County	Arizona	White	10826	0.02	0.01
...	...	...	...	...	...
Wood County	Wisconsin	Black	16	0.24	0.10
Wood County	Wisconsin	Hispanic	27	0.04	0.03
Wood County	Wisconsin	White	1157	0.03	0.03

(a) What variables were collected on each individual traffic stop in order to create to the summary table above?

(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

1.44 Space launches. The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).[37]

	1957 - 1999		2000 - 2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	-	-	5	65

(a) What variables were collected on each launch in order to create to the summary table above?

(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

[35] Amy Orben and AK Baukney-Przybylski. "Screens, Teens and Psychological Well-Being: Evidence from three time-use diary studies". In: *Psychological Science* (2018).

[36] Emma Pierson et al. "A large-scale analysis of racial disparities in police stops across the United States". In: *arXiv preprint arXiv:1706.05678* (2017).

[37] JSR Launch Vehicle Database, A comprehensive list of suborbital space launches, 2019 Feb 10 Edition.