9.S918: Statistical Inference for Brain and Cognitive Sciences, Pset 2
due 3 March 2025

24 Feb 2025

1 Conditional Independence

In our coverage of exchangeability and conditional exchangeability, we made use of the fact that for random variables X, Y, Z, if $X \perp Y|Z$ then $P(Y|Z, X) = P(Y|Z)$. **Task:** Prove that this is true based on the definition of conditional independence we gave in class, namely that if $X \perp Y|Z$ then $P(X, Y|Z) = P(X|Z)P(Y|Z)$.

2 Computing counterfactual risks from observational data

You set up a simple experiment in which each participant logs on to your experiment's website, reads a brief article, and then reports whether they would share the article on social media. (Regardless of what they report, your website doesn't let them actually share it.) Your experimental materials are two articles with similar content but different framings—one positive-valence and one negative-valence, the two experimental CONDITIONS—and your scientific question is whether the positive-valence or the negative-valence article is more likely to be shared. You intended to randomize the assignment of experimental condition to participant, but you accidentally deployed an alpha-testing version of your experiment, in which the participant is first presented with two options—"For a good mood" and "For a bad mood". Based on their selection they are then presented with the positive version of the article (if they selected "For a good mood") or negative version of the article (if they selected "For a bad mood"), and then after reading it they reported whether they would share the article.

You ran the experiment on a lot of participants—4000!—and here is what the results look like:

Article Valence	Participant response: share the article?	Count
Positive	Yes	672
Positive	No	608
Negative	Yes	1040
Negative	No	1680

You are concerned that because the assignment of article to participant was not randomized, it may affect your ability to draw sound scientific inferences from these data. But when your PI finds out the situation, they point out that the effect size is substantial—a 62% rate of sharing positive-valence articles, versus only a 48% rate of sharing negative-valence articles—and the sample size is so big that the estimate of this effect size is fairly precise. Surely these data will be usable! **Task 1:** use the logic and definitions provided by the potential-outcomes framework (e.g., counterfactual risks, average causal effect, consistency, exchangeability, positivity, identifiability) to explain to your PI the problems in using the 62% and 48% rates observed in the data directly to characterize the causal effect of article valence on participants' self-reported proclivities to share the article.

After you have made this case to your PI, they respond, "Is there nothing to be done?" You then remember that you included in the experiment a detailed survey that, though brief, allows you to determine with essentially perfect accuracy whether each participant is an optimist or a pessimist. When you add the results of this survey to your data, you wind up with the resulting counts:[1]

Participant Outlook	Article Valence	Participant response: Share the article?	Count
Optimist	Positive	Yes	1664
Optimist	Positive	No	896
Optimist	Negative	Yes	480
Optimist	Negative	No	160
Pessimist	Positive	Yes	16
Pessimist	Positive	No	144
Pessimist	Negative	Yes	128
Pessimist	Negative	No	512

Most participants are optimists, and a strong correlation between participant outlook and choice of article valence is evident—optimists tend to choose positive articles, and pessimists tend to choose negative articles. **Task 2:** Let us provisionally assume that your measurement of participant outlook "fully captures" each participant's proclivity to choose the positively- vs. negatively-valenced article: specifically, that once outlook is taken into account, the participant's counterfactual proclivities to share each of the articles provides no further information as to which article they would choose to read. What is this notion of "fully captures" in the terminology of the potential outcomes framework that we have covered? Does participant outlook, under this assumption, help us use these data to answer the

[1] You can download this table as a CSV file at https://rlevy.github.io/quantitative-inference-spring-2025/assets/assignments/pset_2/article_sharing_experiment.csv.

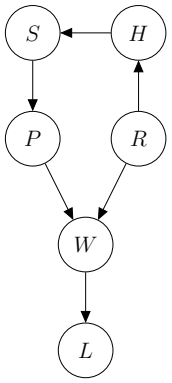
scientific question we posed? If so, explain how, and then answer the scientific question, computing the relevant quantities. If not, explain why not.

Task 3: based on common sense and whatever expertise you may have, critique this assumption that participant outlook "fully cpatures" article choice proclivity.

3 Identifying conditional independencies in (causal) Bayes nets using d-separation

Below is a modification of a classic example of a (causal) Bayes net widely used in the AI literature (e.g., **russell-norvig:2016-artificial-intelligence**; **pearl:2009-causality**), involving the sprinklers, rain, and whether one's lawn gets wet. In this version of the scenario, the sprinklers are automatically controlled by a humidity-sensitive sensor: twice a week, they will go off if the average humidity over the preceding three days has dropped below a certain threshold set by the homeowner. Rain not only increases the humidity but also gets the lawn and footpath wet. If the footpath gets wet, it becomes slippery. For good measure, the homeowner has a separate hygrometer that measures and logs the humidity near the sensor, to make sure the sensor is working properly.

Variable	**Meaning**
R	Whether it **R**ained recently
H	Recent **H**umidity
S	Whether the sprinkers' humidity **S**ensor detected a need to water the lawn
P	Whether the s**P**rinklers went off
W	Whether the lawn and footpath to the front door got **W**et
L	Whether the footpath is s**L**ippery



Task: We will be interested in what conditional independencies among variables in the network hold, given various types of information. For this exercise, when I ask, "Are A and B conditionally independent given C?", you may find that a useful way to think about that question is: suppose I already know C, and based on it I have some beliefs about A. If I additionally observe B, can it further change my beliefs about A? If yes, then A and B are **not** conditionally independent given C.

1. Are R and P conditionally independent, given no additional information? Answer both based on your intuition regarding the scenario and on the semantics of the Bayes net.

2. The homeowner is out of town but reads the hygrometer's log for the past three days (the hygrometer logs to the cloud). Are R and P conditionally independent now?

3. After reading the hygrometer's log, the homeowner gets a text message from the cat sitter, who reports slipping on the footpath on the way to the house. Are R and P conditionally independent now?

4. For each of the preceding three questions, suppose the sensor has broken so that it randomly sets off the sprinkler with 50% probability every three days (and the homeowner knows it). How would you represent this as an INTERVENTION in the language of causal Bayes nets? What does your post-intervention causal Bayes net look like? Does this intervention change the answers to any of the above three questions? Why?

5. One thing that you may find dissatisfying about the structure specified in our causal Bayes net is the directionality of the edge between H and R: although rain can cause increases in humidity, it is surely also the case that factors leading to an increase in humidity on the ground can also lead to rain. How might you modify the causal Bayes net to take this into account? Does this modification change the answer to any of the previous three questions?

4 Paired versus unpaired t-tests

For purposes of this problem, by "t-test" I mean a classic frequentist t-test; the next problem will cover Bayesian t-tests.

Recall that the paired and unpaired two-sample t-tests both test the null hypothesis that two means are the same, but the underlying assumptions are different: whereas the unpaired t-test assumes that the two samples are each iid normally distributed and are independent of each other (conditional on no additional information), the paired t-test assumes that each sample involves measurements from the same set of individuals or units in a single population and assumes that the difference between the two measurements is iid normally distributed among individuals.

Task: answer the following questions:

1. It is sometimes stated that the paired t test is more powerful than the unpaired t test. Of course, there is an uninteresting way in one test can be more powerful than another: if you set the α level (NOMINAL false positive rate) higher for test A than test B, then test A can easily have higher power than test B. But this is not what is meant when it's said that the paired t test is more powerful than the unpaired t test. State the more interesting—and more useful—sense in which the paired test is more powerful than the unpaired test. Why would the paired t-test be the more powerful of the two?

2. The file

https://rlevy.github.io/quantitative-inference-spring-2025/assets/assignments/pset_2/t-test-dataset.tsv

contains a dataset in which each row is a unit, each column is an experimental condition, and the cells are measurements from the corresponding unit–condition combination. Apply paired and unpaired t-tests to the dataset. Which test gives a "more significant" result (i.e., a p-value closer to zero)? How does what you find relate to the generalization stated in part 1 of this problem?

3. A frequentist statistical test is called CONSERVATIVE in a particular setting if, for a particular α level of statistical significance, the actual rate of Type I error (incorrectly rejecting H_0 when it is true) is **lower** than α in that setting, ANTICONSERVATIVE if the actual rate of Type I error is **higher** than α in that setting. For this part of the problem, you will use Monte Carlo to generate hypothetical paired-samples datasets and look at the (anti)conservativity of paired and unpaired tests on these hypothetical datasets. Assume that the two measurements from each individual come from a BIVARIATE NORMAL distribution—this is a joint distribution on two random variables $\langle X_1, X_2 \rangle$ with means $\langle \mu_1, \mu_2 \rangle$ and COVARIANCE MATRIX $\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, where σ_1 is the standard deviation of X_1, σ_2 is the standard deviation of X_2, and ρ is the correlation between X_1 and X_2.[2] Set $\sigma_1 = \sigma_2$ and look at the shapes of histograms of p-values for both paired and unpaired t-tests as a function of the correlation coefficient $-1 \leq \rho \leq 1$. What do you see? Explain your findings.

5 The Bayesian t-test

Over the past 15 years or so there has been a movement to supplant frequentist methods with Bayesian methods, including replacing hypothesis testing within the Neyman–Pearson paradigm with Bayesian hypothesis testing using Bayes factors. For the t-test, an influential proposal is due to **rouder-etal:2009-bayesian-t-tests<empty citation>**. Their one-sample Bayesian t-test assumes that the observations are iid normally distributed with mean μ and standard deviation σ, and for the "alternative hypothesis" H_1 places a prior on these parameters in the following way. We define the EFFECT SIZE δ as the ratio of the mean to the standard deviation: $\delta = \mu/\sigma$. The prior for the "alternative" hypothesis H_1 is then specified on σ and δ as follows

$$p(\sigma^2) = \frac{1}{\sigma^2} \qquad \text{(also known as the JEFFREYS PRIOR)}$$

$$\delta \sim t_1 \qquad \text{(also known as the CAUCHY DISTRIBUTION)}$$

[2] The bivariate normal distribution is a special case of the multivariate normal distribution, which you can access in `R` using the `mvtnorm` library's `*mvnorm()` functions. For example, the following call takes 100 iid samples from the bivariate normal distribution with $\langle \mu_1, \mu_2 \rangle = \langle 0, 0 \rangle$ and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$:

```
rmvnorm(100,c(0,0),matrix(c(1,0,0,1),2,2))
```

The resulting samples are provided in a 2×100 matrix.

The null hypothesis H_0 is identical to the above except $\delta = 0$.

It turns out that the Bayes Factor for this model comparison can be computed fairly straightforwardly, with just a single numeric approximation of an integral in one dimension (see **rouder-etal:2009-bayesian-t-tests**, Equation 1). An implementation can be found in `R`'s `BayesFactor` package, using the `ttestBF()` function with the argument `rscale="wide"`. For this problem, you can use this function from `R` or any language that allows calls to `R` functions (e.g., using the Python `rpy2` package). **Note:** this function returns the "raw" Bayes Factor $BF_{10} = \frac{P(H_1)}{P(H_0)}$, but for this problem please use the log-Bayes Factor $\log \text{BF}_{10} = \log \frac{P(H_1)}{P(H_0)}$. (After you finish the problem, it's worth re-doing it with raw Bayes Factor to demonstrate that it's easier to see the relevant patterns with $\log \text{BF}_{10}$.)

Task: Answer the following questions:

1. Use Monte Carlo simulation to estimate and plot the distribution of (i) p-values; and (ii) Bayes Factors; when H_0 is true, for different values of N, including at least $N \in \{10, 100, 1000\}$ and σ, including at least $\sigma = 1$ and $\sigma = 10$. What do you notice? Explain what you see.

2. Now plot the Bayes Factor against the p value for each of the combinations of N and σ that you tried, together with values of μ including at least 0 and σ. What do you see? **Want a challenge?** Consult Equation 1 of **rouder-etal:2009-bayesian-t-tests** and use it to explain the patterns that you see.

3. Is it possible for the same dataset to yield a frequentist t-test outcome of $p < 0.05$ but a $\log \text{BF}_{10} < k_0$ for some $k_0 < 0$ (i.e. the Bayes Factor favors H_0)? What about the opposite result: a t test outcome of $p < 0.05$ but a $\log \text{BF}_{10} > k_1$ for some $k_1 > 0$? In each case that is possible, what is the most extreme possible value of k (i.e., small values of k_0 or large values of k_1) that you can find? Provide some interpretation of your results.