9.S918: Statistical Inference for Brain and Cognitive Sciences, Pset 2
due 26 April 2024

19 April 2024

1 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/statistical-inference-spring-2024/assets/assignments/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 μ_1, μ_2 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. 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.

2 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 et al. (2009). 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

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

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

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 et al., 2009, 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 $\log \mathrm{BF}_{10} = \log \frac{P(H_1)}{P(H_0)}$. (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 $\log \mathrm{BF}_{10}$.)

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 Rouder et al., 2009 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 $\log \mathrm{BF}_{10} < k_0$ 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 $\log \mathrm{BF}_{10} > k_1$ 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.

References

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.