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

19 April 2024

Note: you don't have to do both problems 4 and 5; one is enough.

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/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.[1] 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-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

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

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

[1] 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.

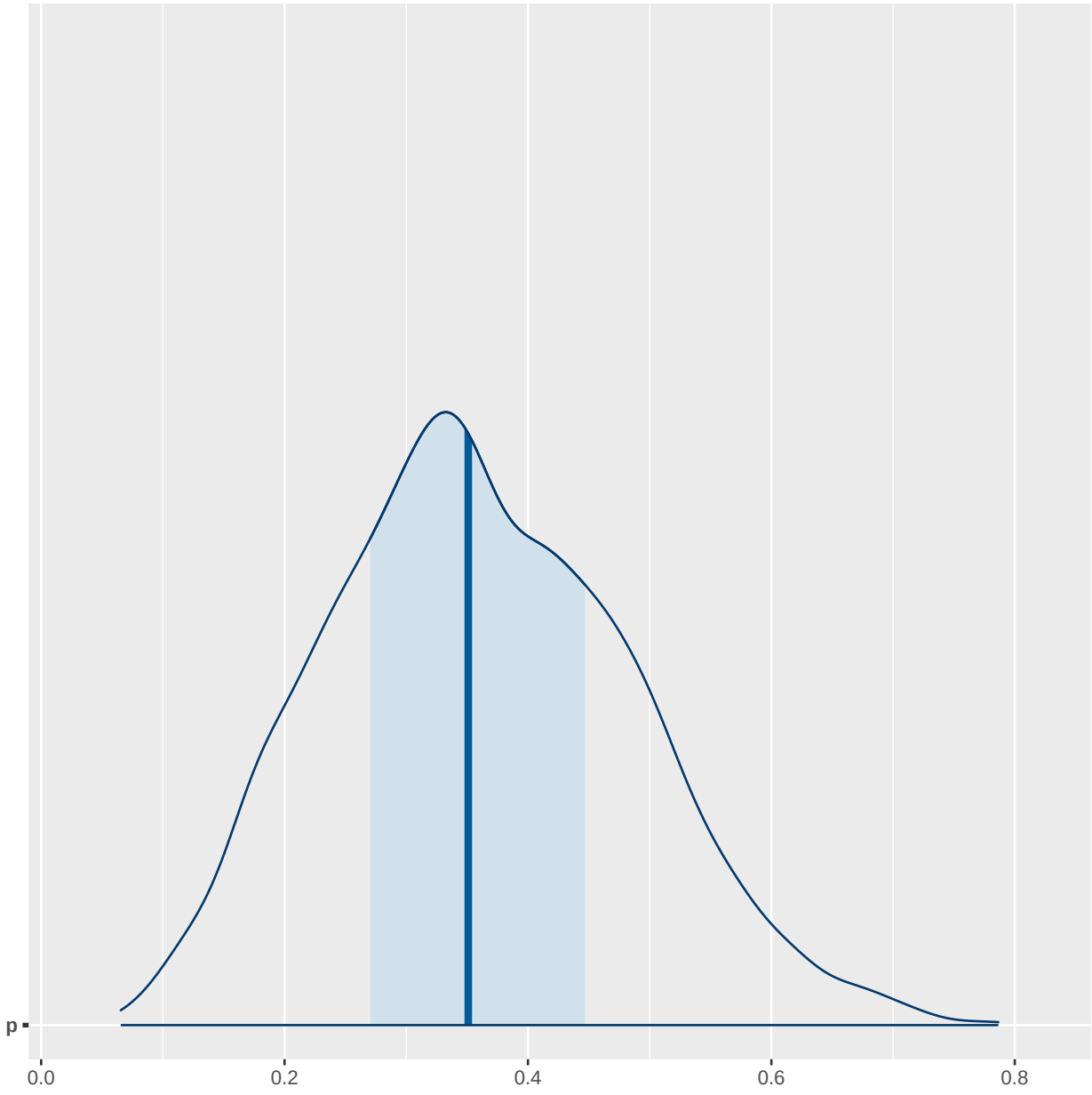
3 Using Stan to estimate Bayesian posteriors for binomial data

Install Stan on your computer (https://mc-stan.org/). The following `R` code creates a string variable called `beta_binomial_model_code` whose value is Stan program implementing a standard beta–binomial model using a (2,2) beta prior, runs the program on binomial data consisting of $n = 10, r = 3$, prints summary statistics of the fitted model and shows a density plot of the estimated posterior on the binomial success parameter p. (Note that for brevity I've turned off the messages showing progress of the HMC sampler; to bring that back, remove the `refresh = 0` argument from the `stan()` call.)

```
library(rstan)
library(bayesplot)

beta_binomial_model_code <- "
data {
  int<lower = 0> n;  // Total number of trials
  int<lower = 0> r;  // number of successes
}
parameters {
  real<lower = 0, upper = 1> p;
}
model {
  // Prior:
  p ~ beta(2,2);
  // Likelihood
  r ~ binomial(n,p);
}
"

n <- 10
r <- 3

set.seed(1) # for reproducibility
posterior <- stan(model_code = beta_binomial_model_code,
                  data = list(r = r, n = n),
                  iter = 2000,
                  warmup = 1000,
                  refresh = 0)
```

```
print(posterior)

## Inference for Stan model: anon_model.
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##       mean se_mean   sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
## p     0.36    0.00 0.12   0.14  0.27  0.35  0.45  0.62  1434    1
## lp__ -9.64    0.02 0.72 -11.82 -9.81 -9.37 -9.18 -9.13  1641    1
##
## Samples were drawn using NUTS(diag_e) at Mon Apr 22 21:51:06 2024.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
```

```
## convergence, Rhat=1).

mcmc_areas(posterior, pars="p")
```



Task:

1. Run this Stan model using the provided binomial data $n = 10, r = 3$. Confirm that your results match those provided in the example above.

2. Extract the samples from the posterior $P(p|r, n)$ from the resulting fitted Stan model—in `rstan` you can do this by applying the `extract()` function to the fitted model, and extracting the element named `p`: in `R`, the call would be `extract(posterior)$p`. Compute the mean and standard deviation of these posterior samples to confirm that they match the posterior mean and posterior standard deviation presented in the model summary.

3. Although Stan ran just fine on this example, we actually didn't need Stan here, because the beta distribution is CONJUGATE to the binomial distribution, so the posterior is also a beta distribution. What are the parameters of the beta posterior for this example? Plot the beta posterior and compare it to the estimated posterior you got from Stan. Is the posterior as estimated by Stan reasonably close to the exact beta posterior?

4 Frequentist confidence intervals and Bayesian posterior credible intervals for normally distributed data.

1. Recall that for a size-N sample of iid normally distributed data, the standard way of constructing a $(1-\alpha)$ frequentist confidence interval on the underlying mean is given by $\bar{x} \pm \frac{sQ_{N-1}(1-\frac{\alpha}{2})}{\sqrt{N}}$ where $\bar{x}$ is the sample mean, $Q_{t_{N-1}}()$ is the QUANTILE FUNCTION for the t distribution with $N-1$ degrees of freedom, and s is the sample standard deviation. Remember, $\frac{s}{\sqrt{N}}$ is the STANDARD ERROR OF THE MEAN. **Question:** there is a rule of thumb that two times the standard error of the mean is the rough width of a 95% confidence interval for the population mean. Explain how this rule is justified. What circumstances could make it less justified?

2. The file

   https://rlevy.github.io/statistical-inference-spring-2024/assets/assignments/pset_2/normal-data.tsv

   contains a size $N = 30$ sample of iid normally distributed data. Write a Stan program that puts an UNINFORMATIVE PRIOR—for present purposes, a prior that is relatively flat in the plausible range of model parameters—and run Stan on this dataset to estimate a sample-based posterior on the model parameters. Compare the symmetric 95% Bayesian credible interval on the population mean with the standard 95% frequentist confidence interval as constructed in part 1 of this problem. Describe the similarities and/or differences. Now try the same thing with a smaller sample—just the first $N = 4$ observations from the dataset. Are the Bayesian and frequentist CIs more or less similar to one another as the sample size increases? Why?

5 Diagnosing Stan model convergence issues with a multimodal prior & posterior

This problem is a recapitulation of an example I covered briefly in class on April 18.

Suppose we receive a coin from a factory that produces two types of coins in equal quantities: bent coins that are much more likely to land heads than tails; and bent coins that are much more likely to land tails than heads. You formalize this as a bimodal mixture of two equally weighted beta priors: one with $\langle \alpha_1, \alpha_2 \rangle = \langle \frac{1}{5}\alpha, \frac{4}{5}\alpha \rangle$ and the other with $\langle \alpha_1, \alpha_2 \rangle = \langle \frac{4}{5}\alpha, \frac{1}{5}\alpha \rangle$, for some α. The file

https://rlevy.github.io/statistical-inference-spring-2024/assets/assignments/pset_2/binom_mixture.stan

expresses this model for $\alpha = 25$.

Without looking at the coin, you flip it an even number of times n, and it turns up heads $\frac{n}{2}$ times and tails $\frac{n}{2}$ times. You want to estimate the posterior on the binomial success parameter p. (**Want a challenge?** You can actually compute this posterior analytically. Do so. But you should still do the rest of the problem as it will give you familiarity with signs of poor MCMC convergence.)

Questions/Tasks:

1. Multimodal posteriors are much harder than unimodal posteriors to explore fully, because MCMC has a hard time transitioning from the vicinity of one mode to the vicinity of another. Hamiltonian Monte Carlo, which Stan uses, is often better than other MCMC methods, but it still has a hard time. Using your knowledge of the beta–binomial model, explain what relative settings of α and n will make it easier versus harder to adequately explore the posterior using MCMC.

2. To diagnose whether an MCMC exploration of the posterior has converged reasonably well, one will often run multiple chains in parallel, initializing each one in a different part of the parameter space. Why would this be useful?

3. Estimate the posterior using Stan, varying α, n, the number of iterations you run the model for (using the `iter=` argument of `stan()`), and potentially the number of chains. Inspect your runs using the `traceplot()` function on your fitted models. What warning diagnostics does Stan give when there are signs that the Markov chains have not converged? What helps make those warning diagnostics go away? Inspect the estimated posteriors using the `mcmc_areas()` function. Given the prior and the data, do the estimated posteriors look reasonable when Stan has issued warnings? How about when you have made the warnings go away by changing the prior, data, and/or algorithmic choices such as number of iterations? posteriors look reasonable