Comparing Two Populations

STAT 251, Supplement 3

Overview

Review

Two Populations

Review

Beta-Binomial

If the likelihood, $f(y|\theta)$, is the Binomial(n, θ) distribution *and* the prior distribution, $\pi(\theta)$, is the $Beta(a, b)$ distribution, *then* the posterior distribution, $\pi(\theta|y)$, is the $Beta(a + y, b + n - y)$ distribution.

$$Y|\theta \sim Binomial(n, \theta) \quad \text{and} \quad \theta \sim Beta(a, b)$$
$$\Rightarrow \theta|(Y = y) \sim Beta(a + y, b + n - y).$$

Two Populations

Comparing Two Populations

Suppose that we wish to compare two populations with respect to the proportion that have a particular characteristic.

Examples:

- $\theta_1 =$ proportion of vaccinated adults (population 1) that get the flu this winter; $\theta_2 =$ proportion of unvaccinated adults (population 2) that get the flu this winter.

Comparing Two Populations

Suppose that we wish to compare two populations with respect to the proportion that have a particular characteristic.

Examples:

- $\theta_1 =$ proportion of vaccinated adults (population 1) that get the flu this winter; $\theta_2 =$ proportion of unvaccinated adults (population 2) that get the flu this winter.
- $\theta_1 =$ proportion of BYU graduate students (population 1) that are currently employed; $\theta_2 =$ proportion of BYU undergraduate students (population 2) that are currently employed.

Comparing Two Populations

Suppose that we wish to compare two populations with respect to the proportion that have a particular characteristic.

Examples:

- $\theta_1 =$ proportion of vaccinated adults (population 1) that get the flu this winter; $\theta_2 =$ proportion of unvaccinated adults (population 2) that get the flu this winter.
- $\theta_1 =$ proportion of BYU graduate students (population 1) that are currently employed; $\theta_2 =$ proportion of BYU undergraduate students (population 2) that are currently employed.
- $\theta_1 =$ proportion of millenials (population 1) with a landline; $\theta_2 =$ proportion of senior citizens (population 2) with a landline.

Comparing Two Populations

Among the possibilities for making inference on two population proportions, natural options are

- Determine $Pr(\theta_1 < \theta_2, \text{given y's})$
- Determine $Pr(\theta_1 > \theta_2, \text{given y's})$
- Determine a posterior credible interval for $\theta_1 - \theta_2$. If it does not contain 0, this is evidence that $\theta_1 \neq \theta_2$. If it does contain 0, it is plausible that $\theta_1 = \theta_2$.

Each summary above depends on the posterior distribution of $\theta_1 - \theta_2$. The distribution of a difference in random variables is nontrivial to determine.

Assumptions, part 1

Suppose that we follow our tradition of having a beta prior on θ_1 (i.e., $\pi_1(\theta_1) = Beta(a_1, b_1)$), and a (possibly different) beta prior on θ_2 (i.e., $\pi_2(\theta_2) = Beta(a_2, b_2)$).

We will assume that the priors are independent (prior beliefs about θ_1 are independent of prior beliefs about θ_2, and vice versa). That is, we assume that $\pi_1(\theta_1|\theta_2) = \pi_1(\theta_1)$ and also, $\pi_2(\theta_2|\theta_1) = \pi_2(\theta_2)$.

Assumptions, part 2

Let y_1 denote the number of observed successes from a sample of n_1 observations from the first population.

Let y_2 denote the number of observed successes from a sample of n_2 observations from the second population.

We will assume $y_1|\theta_1 \sim \mathit{Binomial}(n_1, \theta_1)$ and $y_2|\theta_2 \sim \mathit{Binomial}(n_2, \theta_2)$.

Finally, we will assume that $y_1|\theta_1$ is independent of $y_2|\theta_2$.

Under these assumptions . . .

- The posterior distribution for $\theta_1|y_1$ is the $Beta(a_1^\star = a_1 + y_1, b_1^\star = b_1 + (n_1 - y_1))$ distribution.
- The posterior distribution for $\theta_2|y_2$ is the $Beta(a_2^\star = a_2 + y_2, b_2^\star = b_2 + (n_2 - y_2))$ distribution.
- $\theta_1|y_1$ and $\theta_2|y_2$ are independent.

So we have convenient forms for the parameters individually. But what about the distribution of $\theta_1 - \theta_2$?

The distribution for this difference of two quantities is not something that is particularly convenient for us to work with. Thus, we will turn to *Monte Carlo* methods.

Monte Carlo, our good friend!

Although possible to mathematically derive the posterior distribution of $\theta_1 - \theta_2$ and associated summaries thereof, this is well beyond the course prerequisites.

We can get simulation-based estimates. If the simulation is based on a large enough simulated sample, the estimates should be close to the actual values. A sample size of 5000 is sufficiently large for this problem.

First sample from the posterior distribution of $\theta_1 - \theta_2$, then compute estimates from the sample values.

Monte Carlo inference on $\theta_1 - \theta_2$

Recall that with our assumptions on the prior and the data, the posterior distributions for $\theta_1|y_1$ and $\theta_2|y_2$ were each beta distributions, and that they were independent of each other.

This allows us to employ the following procedure:

1. Take a random sample of size J from the posterior distribution of $\theta_1|y_1$.
2. Take a random sample, also of size J, from the posterior distribution of $\theta_2|y_2$.
3. Compute the J pairwise differences (first θ_1- first θ_2, second θ_1- second θ_2, etc.).
4. Obtain the desired summary (mean, median, variance, credible interval) from the J pairwise differences.

Example: Suppose $\theta_1|y_1 \sim Beta(75, 123)$, $\theta_2|y_2 \sim Beta(91, 53)$

```r
set.seed(8001)
theta1s <- rbeta(6000,75,123)
theta2s <- rbeta(6000,91,53)
diffs <- theta1s-theta2s
diffs[1:4] # Display the first four values
```

```
## [1] -0.2502185 -0.1997302 -0.2037435 -0.2648161
```

```r
mean(diffs)  ## Estimate of E(theta1-theta2)
```

```
## [1] -0.2535755
```

```r
#median(diffs)  ## Estimated median of (theta1-theta2)
```

We estimate that θ_1 is 0.25 less than θ_2.

Monte Carlo Estimated Credible Intervals

```
## 95% credible interval for theta1-theta2
quantile(diffs,c(.025,.975))

##       2.5%      97.5%
## -0.3540742 -0.1495734

## 98% credible interval for theta1-theta2
quantile(diffs,c(.01,.99))

##         1%        99%
## -0.3736450 -0.1314648
```

Neither credible interval contains 0, so we are very sure $\theta_1 \neq \theta_2$.
We are very sure θ_2 exceeds θ_1 by at least 0.13.

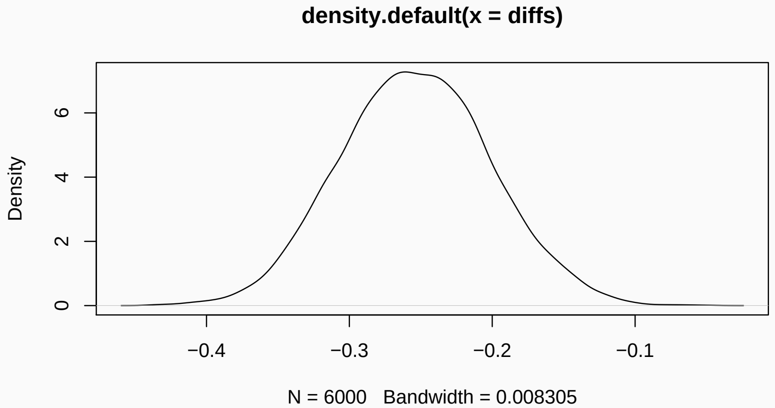
Monte Carlo Estimates of Probability

```
## Estimate of Posterior Pr(theta1 > theta2)
mean(diffs>0)

## [1] 0

## Estimated Posterior Probability that theta2
# exceeds theta1 by at least 0.2
mean(diffs <= -0.2)

## [1] 0.8461667
```

Estimated Posterior Density of $\theta_1 - \theta_2$

Using the *density* function in R, we can use the simulated values of $\theta_1 - \theta_2$ to estimate the pdf of the posterior.

```
plot(density(diffs))
```



Estimated Posterior Density of $\theta_1 - \theta_2$

```r
t1s <- rbeta(1000000, 75,123)
t2s <- rbeta(1000000,91,53)
plot(density(t1s-t2s),
  xlab=expression(theta[1]-theta[2]),
  main=expression(paste("Estimated Posterior Density of ",
                  theta[1], "-", theta[2]), sep=""))
```