

Set 15 - Comparing probabilities

STAT 401 (Engineering) - Iowa State University

February 27, 2017

One probability

Consider the model $Y \sim \text{Bin}(n, \theta)$ or, equivalently, $Y_i \stackrel{\text{ind}}{\sim} \text{Ber}(\theta)$. We have discussed a number of statistical procedures to draw inferences about θ :

- pvalue for a hypothesis test, e.g. $H_0 : \theta = \theta_0$,
- confidence interval for θ ,
- credible interval for θ ,
- posterior model probability, e.g. $p(H_0|y)$, and
- probability statements, e.g. $P(\theta < \theta_0|y)$.

Now, we will consider what happens when you have multiple θ s.

Two probabilities

Consider the model

$$Y_i \sim \text{Bin}(n_i, \theta_i)$$

for $i = 1, 2$ and you are interested in the relationship between θ_1 and θ_2 .

We can perform the following statistical procedures:

- pvalue for a hypothesis test, e.g. $H_0 : \theta_1 = \theta_2$,
- confidence interval for $\theta_1 - \theta_2$,
- credible interval for $\theta_1 - \theta_2$,
- posterior model probability, e.g. $p(H_0|y)$, and
- probability statements, e.g. $P(\theta_1 < \theta_2|y)$.

where $y = (y_1, y_2)$.

Data example

Suppose you have two manufacturing processes and you are interested in which process has the larger probability of being within the specifications.

So you run the two processes and record the number of successful products produced:

- Process 1: 135 successful products out of 140 attempts
- Process 2: 216 successful products out of 230 attempts

In R, you can code this as two vectors:

```
successes = c(135,216)
attempts  = c(140,230)
```

or, better yet, as a data.frame:

```
d = data.frame(process = factor(1:2),
               successes = successes,
               attempts  = attempts)
```

Pvalues and confidence intervals

Because there is no indicating that you expect one of the two manufacturing processes to have a higher probability, you should perform a two-sided hypothesis test and calculate a two-sided confidence interval.

```
(pt <- prop.test(d$successes, d$attempts))
```

2-sample test for equality of proportions with continuity correction

data: d\$successes out of d\$attempts

X-squared = 0.67305, df = 1, p-value = 0.412

alternative hypothesis: two.sided

95 percent confidence interval:

-0.02417591 0.07448647

sample estimates:

	prop 1	prop 2
	0.9642857	0.9391304

```
pt$p.value; pt$conf.int
```

```
[1] 0.4119914
```

```
[1] -0.02417591 0.07448647
```

```
attr("conf.level")
```

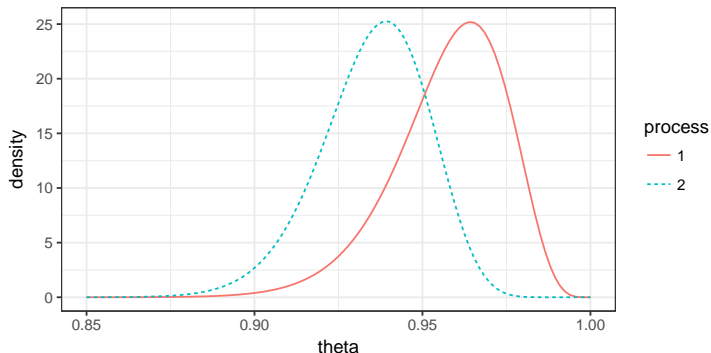
```
[1] 0.95
```

Posteriors

```
posterior <- function(d) {
  data.frame(theta = seq(.85,1,length=101)) %>%
    mutate(density = dbeta(theta, 1+d$successes, 1+d$attempts-d$successes))
}

dp <- d %>% group_by(process) %>% do(posterior(.))

ggplot(dp, aes(x=theta, y=density, color=process, linetype=process, group=process)) +
  geom_line() + theme_bw()
```



Credible interval for the difference

To obtain statistical inference on the difference, we draw samples from the posterior and then calculate the difference:

```
n      <- 1e5
theta1 <- rbeta(n, 1+d$success[1], 1+d$attempts[1] - d$success[1])
theta2 <- rbeta(n, 1+d$success[2], 1+d$attempts[2] - d$success[2])
diff   <- theta1 - theta2

# Bayes estimate for the difference
mean(diff)

[1] 0.02235018

# Estimated 95% equal-tail credible interval
quantile(diff, c(.025,.975))

      2.5%      97.5%
-0.02489203  0.06739588

# Estimate of the probability that theta1 is larger than theta2
mean(diff > 0)

[1] 0.83609
```

Multiple probabilities

Now, let's consider the more general problem of

$$Y_i \sim \text{Bin}(n_i, \theta_i)$$

for $i = 1, 2, \dots, J$ and you are interested in the relationship amongst the θ_j .

We can perform the following statistical procedures:

- pvalue for a hypothesis test, e.g. $H_0 : \theta_i = \theta$ for all i ,
- confidence interval for $\theta_i - \theta_j$ for a specified i and j ,
- credible interval for $\theta_i - \theta_j$ for a specified i and j ,
- posterior model probability, e.g. $p(H_0|y)$, and
- probability statements, e.g. $P(\theta_i < \theta_j|y)$ for a specified i and j .

Data example

Suppose you have two manufacturing processes and you are interested in which process has the larger probability of being within the specifications.

So you run the two processes and record the number of successful products produced:

- Process 1: 135 successful products out of 140 attempts
- Process 2: 216 successful products out of 230 attempts
- Process 3: 10 successful products out of 10 attempts

In R, you can code this as two vectors:

```
successes = c(135,216,10)
attempts  = c(140,230,10)
```

or, better yet, as a data.frame:

```
d = data.frame(process = factor(1:3),
               successes = successes,
               attempts  = attempts)
```

Pvalues and confidence intervals

Because there is no indicating that you expect one of the two manufacturing processes to have a higher probability, you should perform a two-sided hypothesis test and calculate a two-sided confidence interval.

```
prop.test(d$successes, d$attempts)
```

```
Warning in prop.test(d$successes, d$attempts): Chi-squared approximation may be incorrect
```

```
3-sample test for equality of proportions without continuity correction
```

```
data: d$successes out of d$attempts
X-squared = 1.6999, df = 2, p-value = 0.4274
alternative hypothesis: two.sided
sample estimates:
  prop 1    prop 2    prop 3
0.9642857 0.9391304 1.0000000
```

```
prop.test(d$successes[c(1,3)], d$attempts[c(1,3)])$conf.int
```

```
Warning in prop.test(d$successes[c(1, 3)], d$attempts[c(1, 3)]): Chi-squared approximation may be incorrect
```

```
[1] -0.10216886  0.03074029
attr(,"conf.level")
[1] 0.95
```

An alternative test

```
d$failures <- d$attempts - d$successes  
chisq.test(d[c("successes", "failures")])
```

Warning in chisq.test(d[c("successes", "failures")]): Chi-squared approximation may be incorrect

Pearson's Chi-squared test

```
data: d[c("successes", "failures")]  
X-squared = 1.6999, df = 2, p-value = 0.4274
```

```
chisq.test(d[c("successes", "failures")], simulate.p.value = TRUE)
```

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

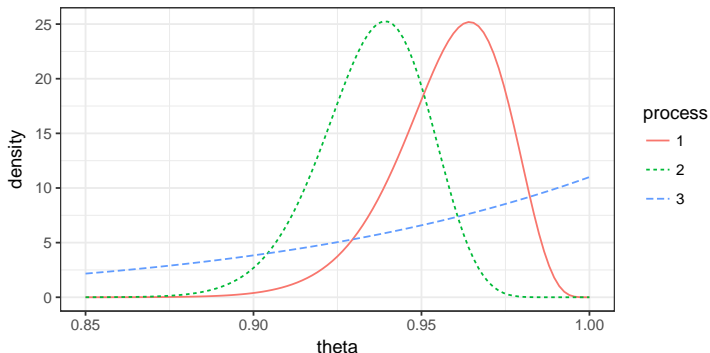
```
data: d[c("successes", "failures")]  
X-squared = 1.6999, df = NA, p-value = 0.4158
```

Posteriors

```
posterior <- function(d) {
  data.frame(theta = seq(.85,1,length=101)) %>%
    mutate(density = dbeta(theta, 1+d$successes, 1+d$attempts-d$successes))
}

dp <- d %>% group_by(process) %>% do(posterior(.))

ggplot(dp, aes(x=theta, y=density, color=process, linetype=process, group=process)) +
  geom_line() + theme_bw()
```



Credible interval for the difference

To compare the probabilities, we draw samples from the posterior and compare them.

```
posterior_samples <- function(d) {
  data.frame(
    rep = 1:1e5,
    name = paste0("theta", d$process),
    theta = rbeta(1e5, 1+d$successes, 1+d$attempts-d$successes),
    stringsAsFactors = FALSE)
}

draws <- d %>% group_by(process) %>% do(posterior_samples(.)) %>% ungroup() %>%
  select(-process) %>% tidyr::spread(name, theta)

# Estimate of the comparison probabilities
draws %>%
  summarize('P(theta1>theta2|y)' = mean(draws$theta1 > draws$theta2),
            'P(theta1>theta3|y)' = mean(draws$theta1 > draws$theta3),
            'P(theta2>theta3|y)' = mean(draws$theta2 > draws$theta3)) %>%
  gather(comparison, probability)

# A tibble: 3  2
  comparison probability
  <chr>         <dbl>
1 P(theta1>theta2|y) 0.84002
2 P(theta1>theta3|y) 0.63235
3 P(theta2>theta3|y) 0.48610
```