Session 3: Comparison of 2 proportions

Exercise 3.1: Burns

Burns: Estimating the difference between two proportions. We have carried out a clinical trial to assess whether patients recover faster from hypodermic burns when one uses the conventional treatment or the experimental treatment. The number of patients treated in both treatments is the same, 40. The nurses have already used both treatments, so they have some prior information. In fact, they believe that using the conventional treatment the probability to get better in five days is mainly between 0.4 and 0.8. On the other hand they believe that using the experimental treatment the probability to get better will be mainly between 0.6 and 0.9.

The ultimate goal is to see if the improvement probability is better with the experimental treatment, p_E than with the conventional p_C.

The Bayesian models (statistical model + prior) for each treatment will be:

- **Conventional treatment**:

$$Y \sim Binomial(n = 40, p_C)$$

$$p_C \sim Beta(\alpha_C, \beta_C)$$

- **Experimental treatment**:

$$Y \sim Binomial(n = 40, p_E)$$

$$p_E \sim Beta(\alpha_E, \beta_E)$$

Remember:

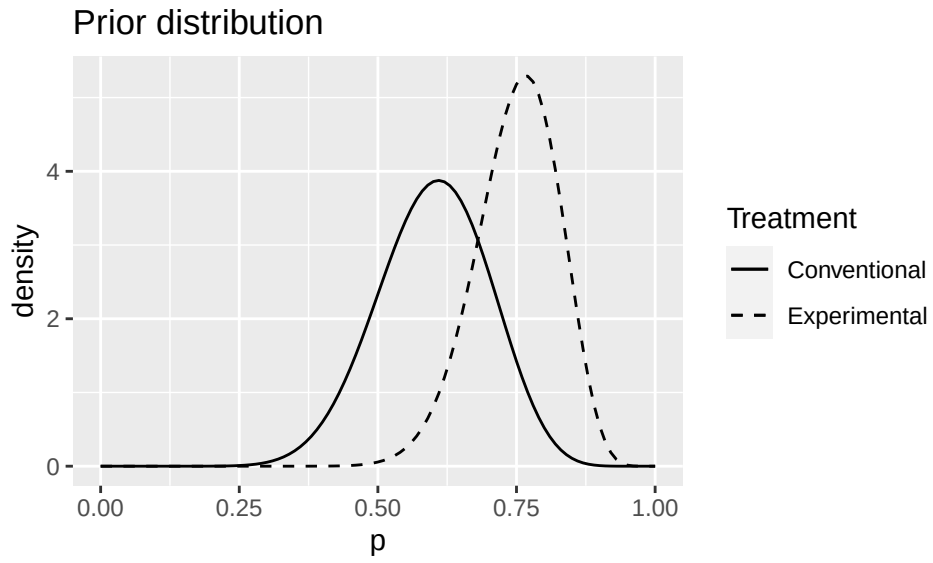
$$E[p] = \frac{\alpha}{\alpha + \beta}$$

$$Var[p] = \frac{\alpha\beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)}$$

a) Choose a priori distribution for every treatment according to the statement

```
library(ggplot2)
library(dplyr)
library(tidyr)

#http://www.wolframalpha.com/
#solve(a/(a+b)=, sqrt((ab)/((a+b)^2*(a+b+1)))=)
```

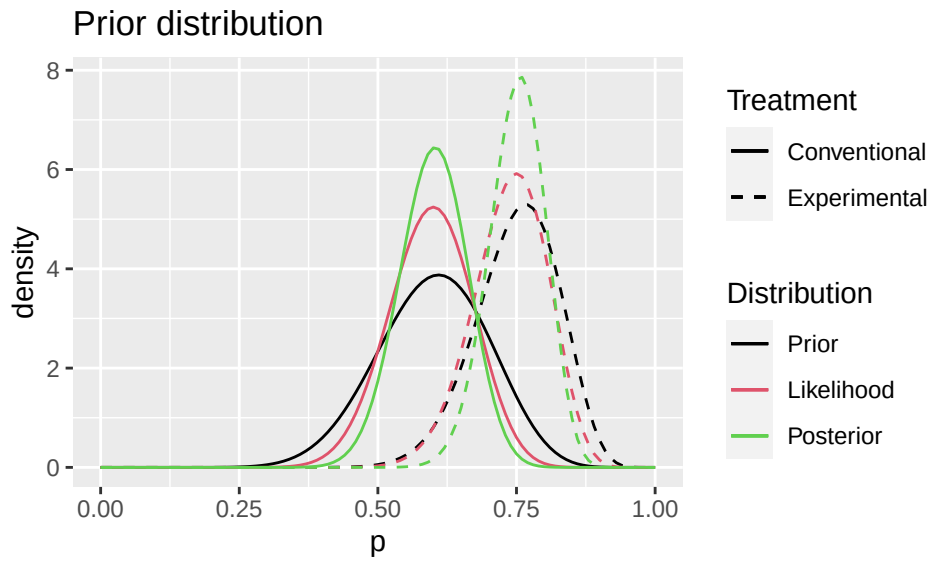
```r
## Prior distributions
delta_p <- 0.01
p <- seq(0, 1, delta_p)

# Conventional prior: 0.4 < p < 0.8
prior_par_c <- c(alpha = 69/5, beta = 46/5)
prior_c <- dbeta(p, prior_par_c[1], prior_par_c[2])

# Experimental prior: 0.6 < p < 0.9
prior_par_e <- c(97/4, beta = 97/12)
prior_e <- dbeta(p, prior_par_e[1], prior_par_e[2])

# Plot prior comparison
df <- tibble(p, prior_c, prior_e)
ggplot(df) +
  geom_line(aes(x = p, y = prior_c, linetype = "c")) +
  geom_line(aes(x = p, y = prior_e, linetype = "e")) +
  ggtitle("Prior distribution") +
  ylab("density") +
  scale_linetype_manual(name = "Treatment",
    values = 1:2, labels = c("Conventional", "Experimental"))
```



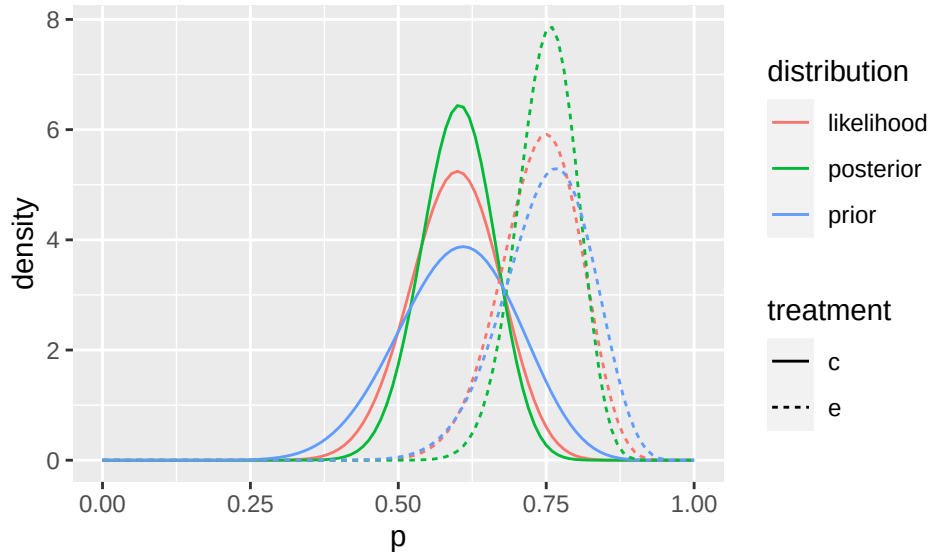
b) Draw the prior distribution, the posterior distribution and the likelihood

```r
# function for every treatment in the same graph.

## Data
n <- 40
y_c <- 24
y_e <- 30


## Likelihood distributions
```

```r
likelihood_c <- dbinom(y_c, n, p)
likelihood_e <- dbinom(y_e, n, p)

## Function to standardize
std_dist <- function(x, delta) {
  x / (sum(x) * delta)
}


## Posterior distributions
# b.1) Grid solution:
# Standardize likelihoods
# Calculate products
# Posteriors: standardize products
df <- df %>%
  mutate(likelihood_c = std_dist(likelihood_c, delta_p),
    likelihood_e = std_dist(likelihood_e, delta_p),
    product_c = prior_c * likelihood_c,
    product_e = prior_e * likelihood_e,
    posterior_c = std_dist(product_c, delta_p),
    posterior_e = std_dist(product_e, delta_p))


# b.2) Analytical solution (conjugate)
posterior_par_c <- prior_par_c + c(y_c, n - y_c)
posterior_c <- dbeta(p, posterior_par_c[1], posterior_par_c[2])

posterior_par_e <- prior_par_e + c(y_e, n - y_e)
posterior_e <- dbeta(p, posterior_par_e[1], posterior_par_e[2])


# b.3) Plots
# Option 1:
ggplot(df) +
  geom_line(aes(x = p, y = prior_c, linetype = "c", col = "1")) +
  geom_line(aes(x = p, y = prior_e, linetype = "e", col = "1")) +
  geom_line(aes(x = p, y = likelihood_c, linetype = "c", col = "2")) +
  geom_line(aes(x = p, y = likelihood_e, linetype = "e", col = "2")) +
  geom_line(aes(x = p, y = posterior_c, linetype = "c", col = "3")) +
  geom_line(aes(x = p, y = posterior_e, linetype = "e", col = "3")) +
  ggtitle("Prior distribution") +
  ylab("density") +
  scale_linetype_manual(name = "Treatment",
    values = 1:2, labels = c("Conventional", "Experimental")) +
  scale_color_manual(name = "Distribution",
    values = 1:3, labels = c("Prior", "Likelihood", "Posterior"))
```



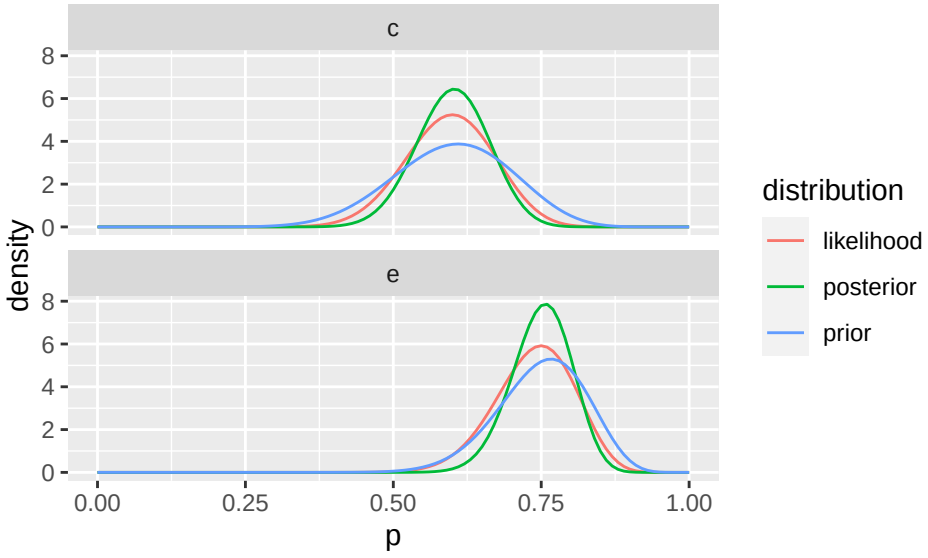

```
# Option 2:
df_long <- df %>%
    select(-c(product_c, product_e)) %>%
    gather(distribution, density, -p) %>%
    separate(distribution, into = c("distribution", "treatment"), sep = "_")

    ggplot(df_long) +
    geom_line(aes(x = p, y = density, linetype = treatment, col = distribution))
```



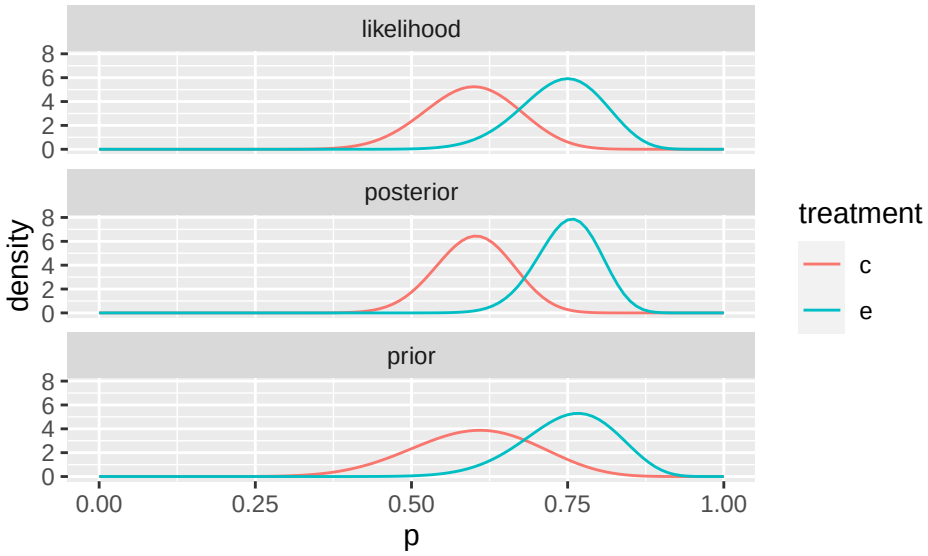
Exercise:

- Instead of using 2 aesthetics (linetype and color), use facet_wrap to compare distributions.

```
# Facet by treatment
    ggplot(df_long) +
    geom_line(aes(x = p, y = density, col = distribution)) +
    facet_wrap(~ treatment, ncol = 1)
```




```
# Facet by distribution
    ggplot(df_long) +
    geom_line(aes(x = p, y = density, col = treatment)) +
    facet_wrap(~ distribution, ncol = 1)
```



c) Draw the posterior distribution of the difference between rates of improvement.

$H_1 : p_E > p_C$
$H_2 : p_E \leq p_C$

It's equivalent to:

$H_1 : p_E - p_C > 0$
$H_2 : p_E - p_C \leq 0$

Defining a new variable:

$H_1 : \gamma > 0$
$H_2 : \gamma \leq 0$

So, we can calculate the probability: $P[H_1|y_C, y_E] = P[\gamma > 0|y_C, y_E]$

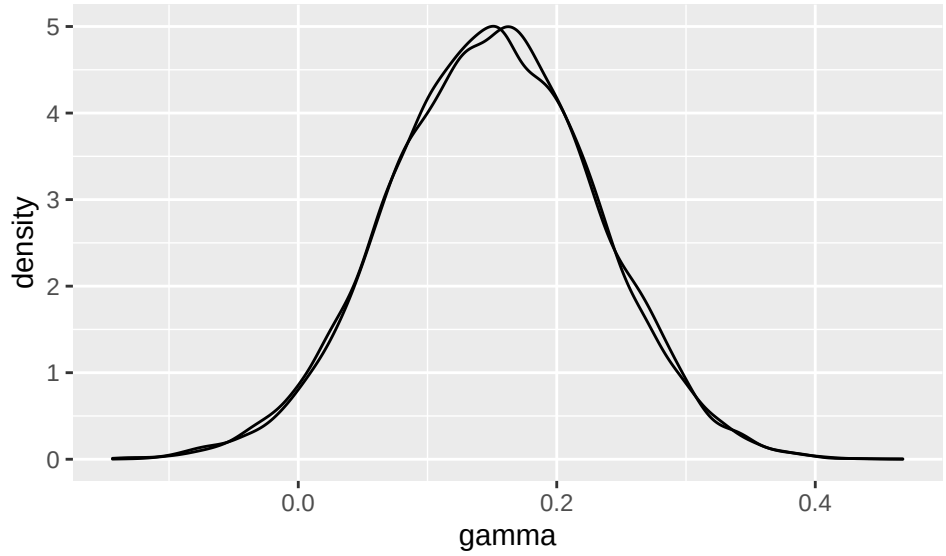
Sometimes, doing analytical calculations can be difficult. That's why we're going to do it by simulation.

If we simulate n_sim values from the a posteriori distribution of p_E and p_C, we can calculate the difference between the two to get a sample of the variable γ.

```
# New variable: gamma = difference between rates of improvement

# c.1) Simulate from grid solution:
n_sim <- 10000

df %>%
    select(p, posterior_c, posterior_e)
```

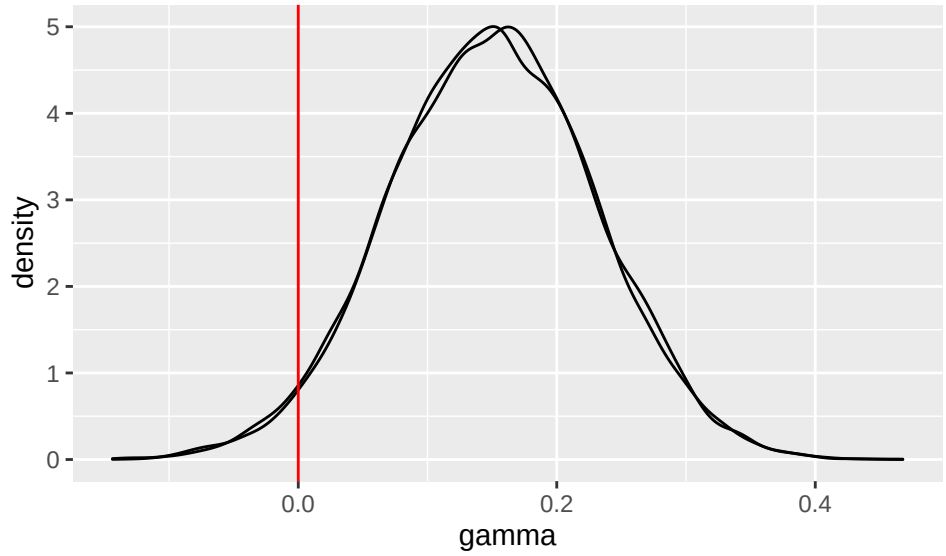
```
# A tibble: 101 x 3
       p posterior_c posterior_e
   <dbl>       <dbl>       <dbl>
 1  0       0           0
 2  0.01    7.89e-56    1.80e-89
 3  0.02    7.39e-45    1.62e-73
 4  0.03    1.74e-38    3.24e-64
 5  0.04    5.37e-34    1.22e-57
 6  0.05    1.53e-30    1.48e-52
 7  0.06    9.74e-28    2.03e-48
 8  0.07    2.19e-25    6.21e-45
 9  0.08    2.29e-23    6.32e-42
10  0.09    1.34e-21    2.78e-39
# ... with 91 more rows
```

```
post_sim_c <- sample(p, size = n_sim, replace = TRUE, prob = df$posterior_c)
post_sim_e <- sample(p, size = n_sim, replace = TRUE, prob = df$posterior_e)
gamma <- post_sim_e - post_sim_c


# c.2) Simulate from analytical solution (conjugate)
n_sim <- 10000
post_sim_c <- rbeta(n_sim, posterior_par_c[1], posterior_par_c[2])
post_sim_e <- rbeta(n_sim, posterior_par_e[1], posterior_par_e[2])
gamma_2 <- post_sim_e - post_sim_c

# c.3) Plot

fig <- ggplot() +
  geom_density(data = tibble(gamma), aes(gamma)) +
  geom_density(data = tibble(gamma_2), aes(gamma_2))
fig
```



d) Compute the probability that the probability to improve using the experimental treatment is larger than using the conventional treatment.

```
fig +
  geom_vline(xintercept = 0, color = "red")
```

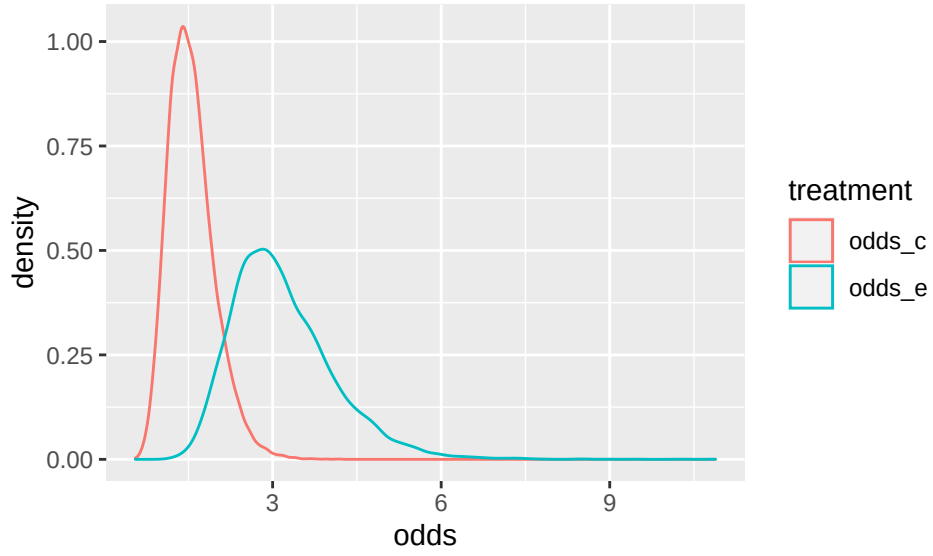



```
mean(gamma > 0)
```

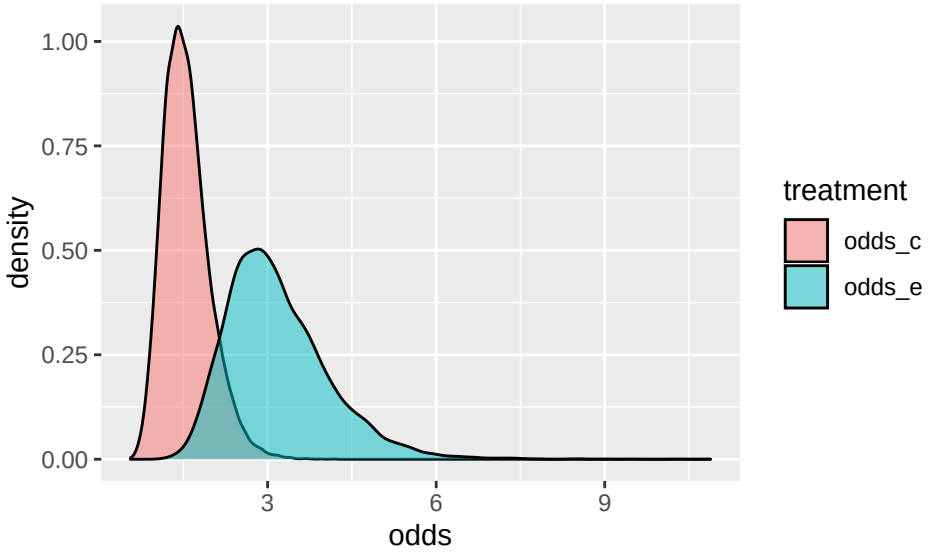
```
[1] 0.9652
```

e) Compute and draw the posterior distribution for the Odds Ratio and give a 95% credible interval for it. Interpret the result.

```r
# odds_ratio = p / (1-p)
odds_c <- post_sim_c / (1 - post_sim_c)
odds_e <- post_sim_e / (1 - post_sim_e)

# 95% credible interval
ci95_c <- quantile(odds_c, c(0.025, 0.975))
ci95_e <- quantile(odds_c, c(0.025, 0.975))

# Plot
odds <- tibble(
  odds_c, odds_e) %>%
    gather(treatment, odds)

ggplot(odds) +
  geom_density(aes(odds, col = treatment))
```




```r
ggplot(odds) +
  geom_density(aes(odds, fill = treatment), alpha = 0.5)
```



Interpretation:
- p = 0.75: probability to get better
- odds = 0.75/0.25 = 3: the odds of get better are 3 to 1 (we expect 3 successes for every 1 failure)

Optional exercise:

- Redo the calculations using a normal priori