140.651 - Lab 12

HW5 Problem 1

Problem 1. A laboratory experiment found that in a random sample of 20 frog eggs having aquaporins, 17 exploded when put into water.

a. Plot and interpret the posteriors for p assuming a beta prior with parameters $(2, 2)$, $(1, 1)$ and $(.5, .5)$.

b. Calculate and interpret the credible interval for each of the beta prior parameter esttings. Note that the R package `binom` may be of use.

a. Plot and interpret the posteriors for p assuming a beta prior with parameters (2, 2), (1, 1) and (.5, .5).

- Suppose $p \sim Beta(\alpha, \beta)$ and our data $X \sim Binom(n, p)$.
- Posterior distribution

$$f(p|\alpha, \beta, X) \propto \text{Prior} \times \text{Likelihood} = f(p|\alpha, \beta) \cdot f(X|p)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \cdot \binom{n}{X} p^X (1-p)^{n-X}$$

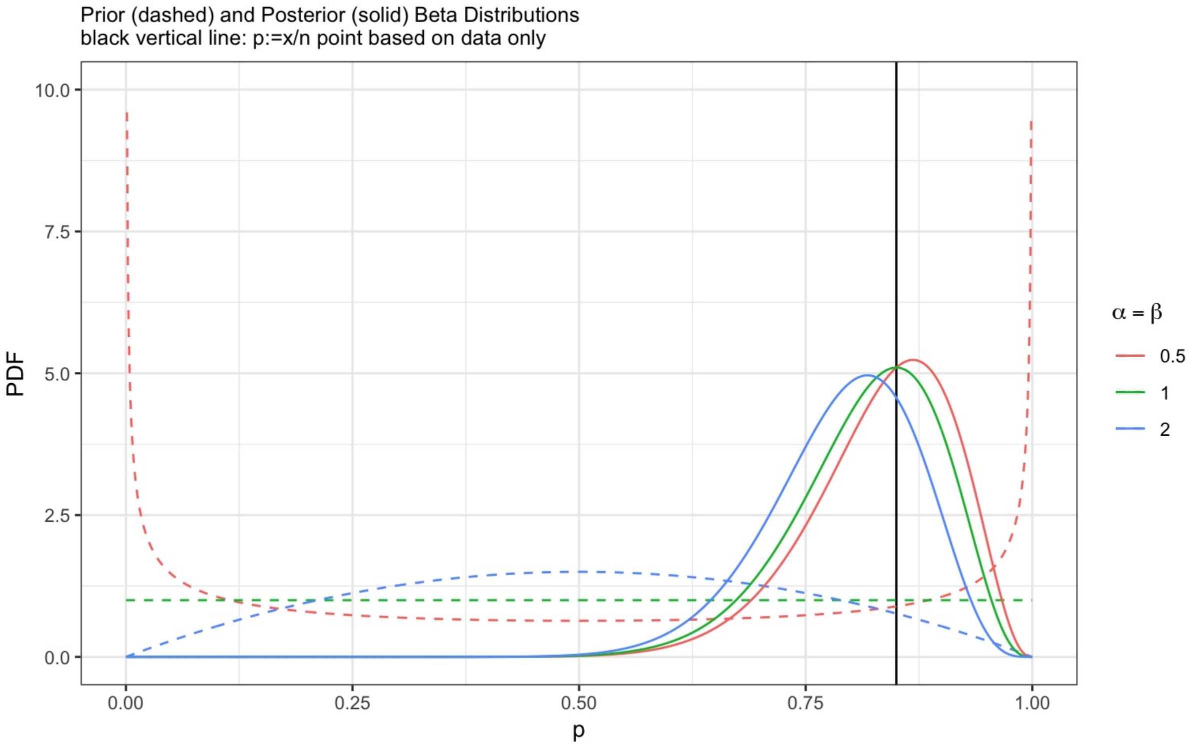
$$\propto p^{\alpha+X-1}(1-p)^{(n-X)+\beta-1}$$

- The posterior follows a $Beta(\alpha + X, (n-X) + \beta)$.

```
library(dplyr)
library(ggplot2)
library(latex2exp)
```

```r
n <- 20
x <- 17

## Grid of p values
p <- seq(0.0001, 1 - 0.0001, length = 1000)

## List of prior distribution parameter
beta.params <- list(c(2, 2), c(1, 1), c(0.5, 0.5))

## lapply: apply over each element of list `beta.params`
## here: param - vector from `beta.params` list
##       p - additional parameter we pass to the function
##
## Calculate priors for different p
priors <- lapply(beta.params, function(param, p)
  dbeta(p, param[1], param[2]), p = p)

## Calculate posteriors for different p
posteriors <- lapply(beta.params, function(param, p)
  dbeta(p, param[1]+ x, n - x + param[2]), p = p)

# Translate into data.frame and plot
data.frame(p = rep(p, 3),
           param = rep(c(2, 1, 0.5), each = 1000),
           prior = unlist(priors),
           posterior = unlist(posteriors)) %>%
  ggplot(aes(x = p, color = factor(param))) +
  geom_vline(xintercept = x/n) +
  geom_line(aes(y = prior), linetype = 2) +
  geom_line(aes(y = posterior)) +
  labs(title = "Prior (dashed) and Posterior (solid) Beta Distributions\nblack vertical line: p:=x/n point base
d on data only",
       x = "p",
       y = "PDF",
       color = TeX('$\\alpha = \\beta$')) +
  theme_bw() +
  theme(plot.title = element_text(size = 10)) +
  scale_y_continuous(limits = c(0, 10))
```

```
## Warning: Removed 2 rows containing missing values (geom_path).
```

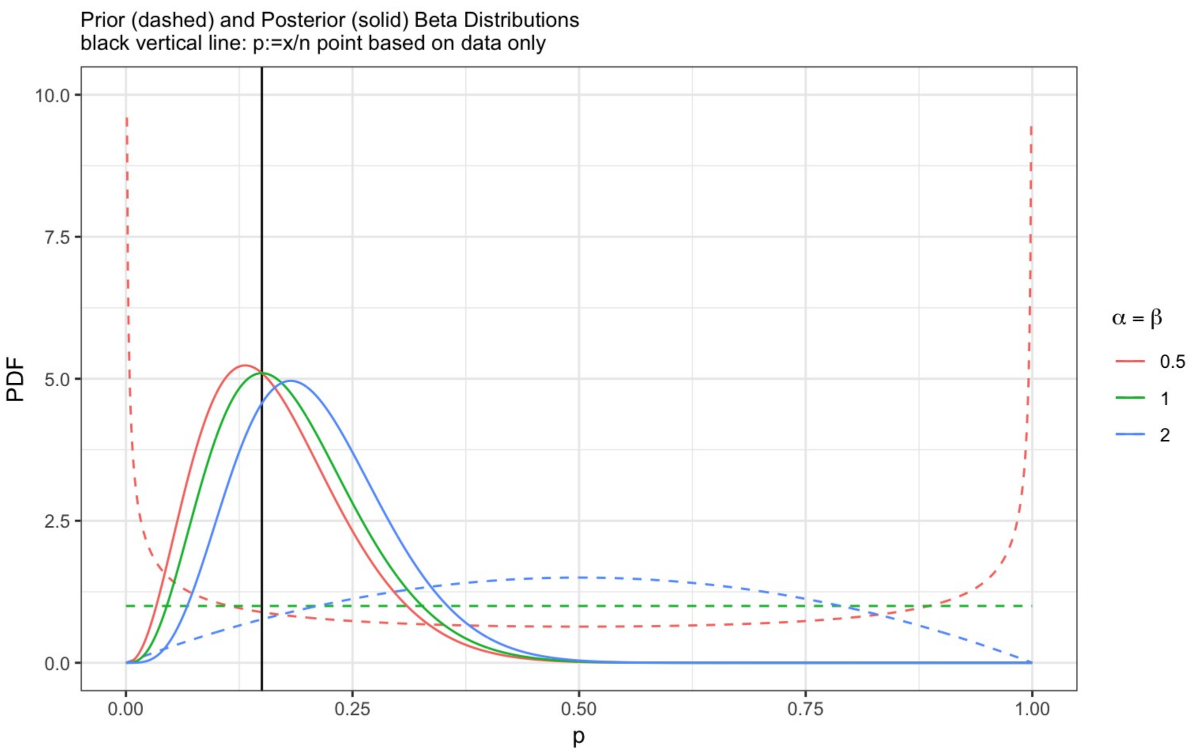


- Note that when $\alpha = \beta = 1$, the distribution follows a uniform distribution, and the posterior is thus just the likelihood.
- The posterior pools information from both the prior and the likelihood to derive a probability distribution for the value of p (generally: the posterior mean is a mixture of the MLE mean estimator and the prior mean).

For comparison only: what if data says `x = 3`?

```
n <- 20
x <- 3  ## only this piece of code chunk has changed

p <- seq(0.0001, 1 - 0.0001, length = 1000)

beta.params <- list(c(2, 2), c(1, 1), c(0.5, 0.5))
priors <- lapply(beta.params, function(param, p)
  dbeta(p, param[1], param[2]), p = p)
posteriors <- lapply(beta.params, function(param, p)
  dbeta(p, param[1]+ x, n - x + param[2]), p = p)

# Translate into data.frame and plot
data.frame(p = rep(p, 3),
           param = rep(c(2, 1, 0.5), each = 1000),
           prior = unlist(priors),
           posterior = unlist(posteriors)) %>%
  ggplot(aes(x = p, color = factor(param))) +
  geom_vline(xintercept = x/n) +
  geom_line(aes(y = prior), linetype = 2) +
  geom_line(aes(y = posterior)) +
  labs(title = "Prior (dashed) and Posterior (solid) Beta Distributions\nblack vertical line: p:=x/n point base
d on data only",
       x = "p",
       y = "PDF",
       color = TeX('$\\alpha = \\beta$')) +
  theme_bw() +
  theme(plot.title = element_text(size = 10)) +
  scale_y_continuous(limits = c(0, 10))
```

```
## Warning: Removed 2 rows containing missing values (geom_path).
```



Hints: how to tell what is the posterior distribution if one does not immediately recognize the kernel?

- Compare with tables of likelihood, conjugate prior and posterior available online, e.g. this, table on p. 2 (https://www2.stat.duke.edu/courses/Fall11/sta114/conjug.pdf)
- Google “likelihood A, prior B, what is posterior”

b. Calculate and interpret the credible interval for each of the beta prior parameter settings. Note that the R package `binom` may be of use.

For each of the credible intervals below, the probability that p is within the interval is 95%.

```
library(binom)

n <- 20
x <- 17

# 95% credible interval for a Beta(2,2) prior
binom.bayes(x, n, prior.shape1 = 2, prior.shape2 = 2)[c("lower", "upper")]
```

```
##       lower     upper
## 1 0.6314496 0.9378542
```

$ lower

```
# 95% credible interval for a Beta(1,1) prior
binom.bayes(x, n, prior.shape1 = 1, prior.shape2 = 1)[c("lower", "upper")]
```

```
##       lower     upper
## 1 0.6599475 0.9591231
```

```
# 95% credible interval for a Beta(0.5,0.5) prior
binom.bayes(x, n, prior.shape1 = 0.5, prior.shape2 = 0.5)[c("lower", "upper")]
```

```
##       lower     upper
## 1 0.6773164 0.9698094
```

Recall:

- (From Lecture 13, slide 10): Bayesian statistics posits a prior distribution on the parameter of interest. All inferences are then performed on the distribution of the parameter given the data, called the posterior distribution.
- (From Lecture 13, slide 25): A Bayesian credible interval is the Bayesian analog of a confidence interval. A 95 % credible interval, $[a, b]$ would satisfy $P(p \in [a, b] | x) = .95$.

Hence:

- Interpretation (e.g. for prior $Beta(0.5, 0.5)$): **Given the data, the probability that p is between 0.68 and 0.97 is 95%.**

HW5 Problem 3

Problem 3. Forced expiratory volume FEV is a standard measure of pulmonary function. We would expect that any reasonable measure of pulmonary function would reflect the fact that a person's pulmonary function declines with age after age 20. Suppose we test this hypothesis by looking at 10 nonsmoking males ages 35-39, heights 68-72 inches and measure their FEV initially and then once again 2 years later. We obtain this data.

Person	Year 0 FEV (L)	Year 2 FEV (L)	Person	Year 0 FEV (L)	Year 2 FEV (L)
1	3.22	2.95	6	3.25	3.20
2	4.06	3.75	7	4.20	3.90
3	3.85	4.00	8	3.05	2.76
4	3.50	3.42	9	2.86	2.75
5	2.80	2.77	10	3.50	3.32

a. Preform and interpret the relevant test. Give the appropriate null and alternative hypotheses. Intepret your results, state your assumptions and give a P-value.

b. A large test comparing the two-year decline in non-smokers of a different age. Perform a sample size calculation to detect a change in FEV over two years at least as large as that detected for males age 35-39. Use the data above a for any relevant constants that you might need.

a. Preform and interpret the relevant test. Give the appropriate null and alternative hypotheses. Intepret your results, state your assumptions and give a P-value.

Let D_i denote the difference in FEV between Year 2 and Year 0 for individual i , and assume that $D_i \overset{iid}{\sim} N(\mu_D, \sigma^2)$. Then, we want to test,

$$H_0 : \mu_D = 0$$
$$H_A : \mu_D \neq 0$$

Since n is small, we can use the t-test to test our hypothesis.

```
fev0 <- c(3.22, 4.06, 3.85, 3.50, 2.80,
          3.25, 4.20, 3.05, 2.86, 3.50)
fev2 <- c(2.95, 3.75, 4.00, 3.42, 2.77,
          3.20, 3.90, 2.76, 2.75, 3.32)
diff <- fev2 - fev0
t.test(diff)
```

default alternative $\neq 0$

```
##
##  One Sample t-test
##
## data:  diff
## t = -3.0891, df = 9, p-value = 0.01295
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.25465006 -0.03934994
## sample estimates:
## mean of x
##    -0.147
```

For $\alpha = 0.05$, we reject the null and conclude that there may be a change in FEV between Year 2 and Year 0 for individuals in this population. The 95% confidence interval is given by $(-0.25, -0.04)$ which does not contain 0, further verifying that the difference in FEV between the two years is not zero.

b. A large test comparing the two-year decline in non-smokers of a different age. Perform a sample size calculation to detect a change in FEV over two years at least as large as that detected for males age 35-39. Use the data above a for any relevant constants that you might need.

- Suppose we want a two-sided test for the change in FEV.
- And 80% power and $\alpha = 0.05$ (**must specify desired power and α level**).
- Estimate the mean and standard deviation of the 2-year FEV difference from the current study and use `power.t.test` for sample size claculation.

```
power.t.test(delta = mean(diff),
             sd = sd(diff),
             power = 0.8,
             sig.level = 0.05,
             alternative = "two.sided",
             type = "one.sample")
```

$H_A: \mu_D \neq 0$

```
##
##      One-sample t test power calculation
##
##               n = 10.30858
##           delta = 0.147
##              sd = 0.1504844
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Since n is integer valued, we find we need a sample size of 11 individuals to achieve the desired power for an $\alpha = 0.05$ test.

Note: Alternatively, you can treat the data as paired data to get the same results since the FEV at Year 0 is matched with the FEV at Year 2 for a given individual. However, the data is *not* of two unpaired samples.

Book Chapter 18 Problem 10 (page 452)

Problem 10. Here we would like to compare the theoretical and observed power using simulations.

a. Simulate 100000 times $n = 10$ $N(0,1)$ random variables and $n = 10$ $N(\mu, 1)$ variables, where $\mu = 0, 0.01, \ldots, 1$ and calculate the percent of rejections of a size $\alpha = 0.05$ two sided t-test; hint, you can simulate $N(\mu, 1)$ simulatneously using

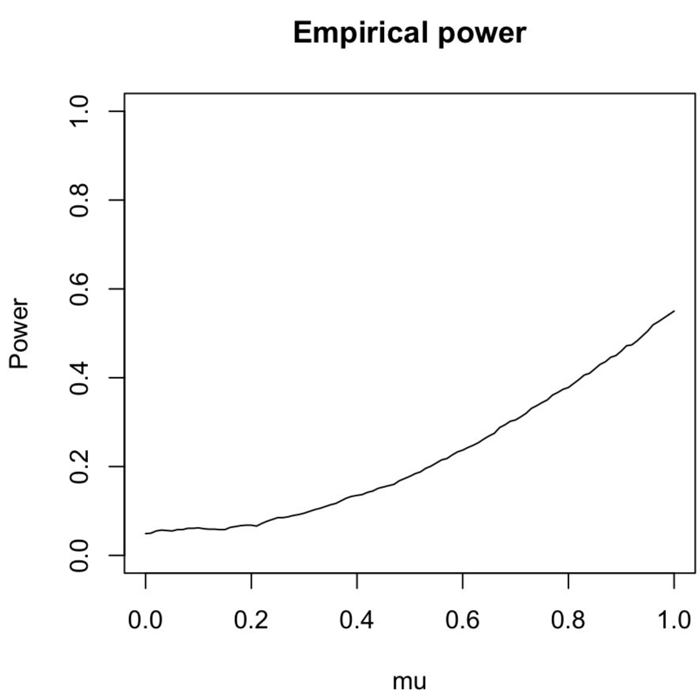
```
nsim=100000
mu=seq(0,1,by=0.01)
sim_vec=mu+rnorm(nsim)
```

b. Compare the proportion of rejections with the theoretical probability of rejection

a. Empirical power

Note: for the sake of code execution time, we use here $N = 1,000$ instead of $100,000$ as in problem description.

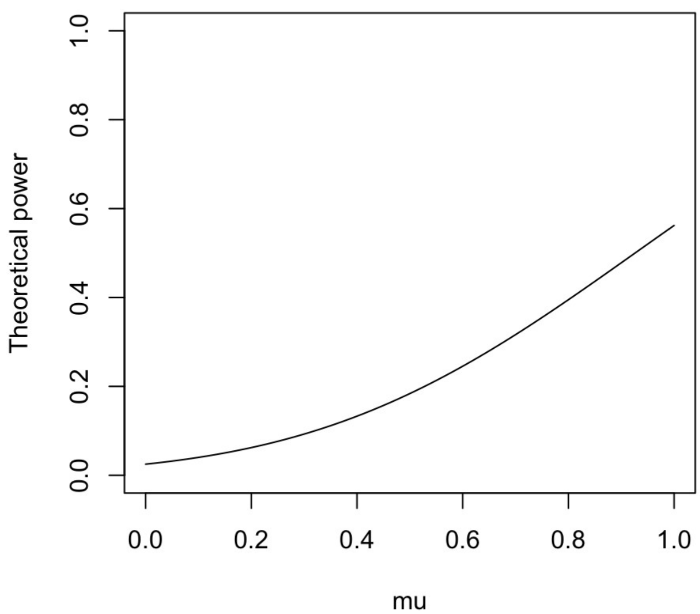
```r
set.seed(34534)

nsim = 1000
samp_null = matrix(rnorm(n = 10 * nsim), nrow = nsim)  ## one 10-element sample per row
error_alt = matrix(rnorm(n = 10 * nsim), nrow = nsim)

## Define grid of \mu (true differences)
mu = seq(0, 1, by=0.01)

## Object to store future power values
pwr = vector(length = length(mu))

## Iterate over true difference \mu values
for (i in 1:length(mu)){

  ## For particular true difference \mu value, perform N = 1000 t.tests and store p-value
  p.value.vec <- sapply(1:nsim, function(x){
    t.test(samp_null[x, ], mu[i] + error_alt[x, ], var.equal = TRUE)$p.value
  })

  ## Power: correctly rejecting the null hypothesis when it is false
  ## (here: rejecting the null at alpha=0.05 happens when p-value is smaller than 0.05)
  pwr[i] = mean(p.value.vec < 0.05)
}

plot(mu, pwr, xlab="mu", ylab="Power", main="Empirical power", type="l",
     xlim = c(0,1), ylim = c(0,1))
```



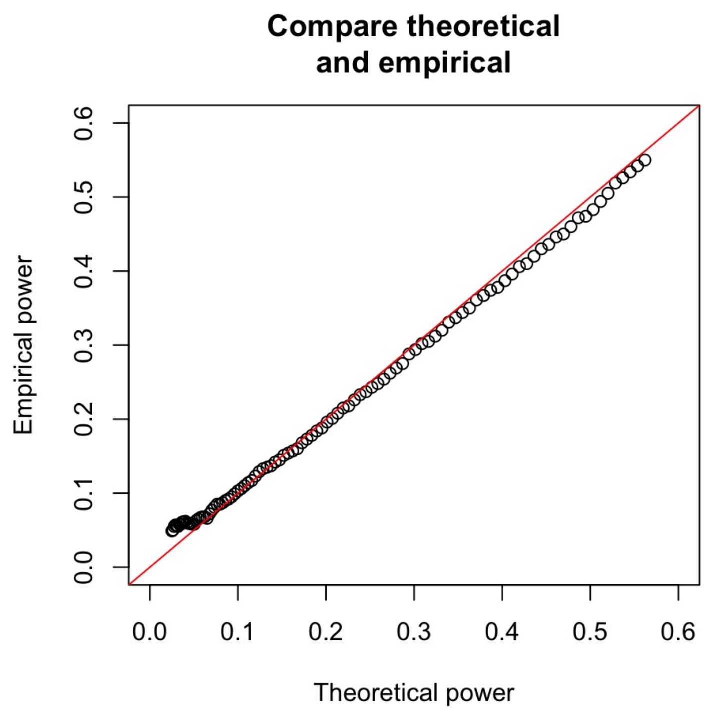
b. Theoretical power

- We obtain theoretical power with the use of `power.t.test` for each true difference μ value we considered

```
pwr.theo = power.t.test(n = 10, delta = mu, type = "two.sample")$power

plot(mu, pwr.theo, xlab="mu", ylab="Theoretical power", main="", type="l",
     xlim = c(0,1), ylim = c(0,1))
```

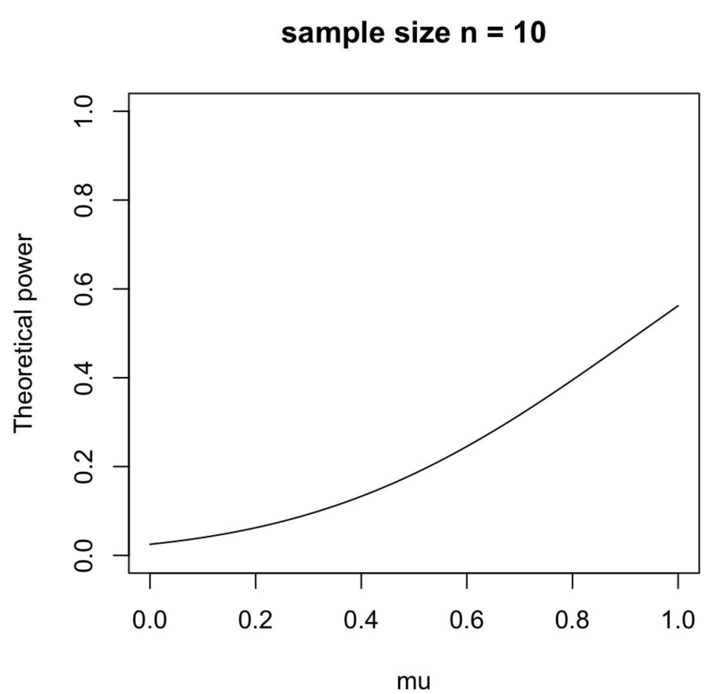



```
plot(pwr.theo, pwr, xlab="Theoretical power", ylab = "Empirical power", main="Compare theoretical\nand empirica
l",
     xlim = c(0,0.6), ylim = c(0,0.6))
abline(a=0, b=1, col="red")
```



[Additional] Theoretical power, but with larger sample size `n=100`

```
pwr.theo2 = power.t.test(n = 100, delta = mu, type = "two.sample")$power

plot(mu, pwr.theo, xlab="mu", ylab="Theoretical power", main="sample size n = 10", type="l",
     xlim = c(0,1), ylim = c(0,1))
```



```
plot(mu, pwr.theo2, xlab="mu", ylab="Theoretical power", main="sample size n = 100", type="l",
     xlim = c(0,1), ylim = c(0,1))
```