Assignment 3

anonymous

1 General information

I did not use AI for solving this exercise.

2 Inference for normal mean and deviation (3 points)

Loading the library and the data.

```
data("windshieldy1")
# The data are now stored in the variable `windshieldy1`.
# The below displays the data:
windshieldy1
```

```
[1] 13.357 14.928 14.896 15.297 14.820 12.067 14.824 13.865 17.447
```

The below data is **only for the tests**, you need to change to the full data `windshieldy1` when reporting your results.

```
windshieldy_test <- c(13.357, 14.928, 14.896, 14.820)
```

2.1 (a)

```
n <- length(windshieldy1)
samples_mean <- mean(windshieldy1)
samples_variance <- var(windshieldy1)
cat(paste("Sample Mean:", samples_mean, "\nSample Variance:", samples_variance, "\n"))
```

```
Sample Mean: 14.6112222222222
Sample Variance: 2.17315294444444
```

The model likelihood follows a normal distribution:

$$p(y|\mu, \sigma^2) \propto N(\mu, \sigma^2)$$

The prior distribution is as follows:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

In the given case, the prior distribution for $\mu|y$ is a Student's t-distribution with $n-1$ degrees of freedom, centered at the sample mean $\bar{y}$ and with a scale parameter $\frac{s^2}{n}$:

$$p(\mu|y) = t_{n-1}(\bar{y}, \frac{s^2}{n}) = t_8(14.6112, 0.2414556)$$

Finally, the posterior, as derived in the BDA3 book, is proportional to:

$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right)$$

Where:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

So, as computed before, in the given case:

$$s^2 = 2.1731$$

$$\bar{y} = 14.6112$$

2.2 (b)

Keep the below name and format for the functions to work with `markmyassignment`:

```
# Useful functions: mean(), length(), sqrt(), sum()
# and qtnew(), dtnew() (from aaltobda)

mu_point_est <- function(data) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    samples_mean = mean(data)
    return(samples_mean)
}

p_estimate <- mu_point_est(windshieldy1)
cat("Point Estimate:", p_estimate, "\n")
```

```
Point Estimate: 14.61122
```

```
mu_interval <- function(data, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.

    limit_1 <- (1 - prob) / 2
    limit_2 <- limit_1 + prob
```

```
    samples_mean <- mean(data)
    samples_variance <- var(data)
    n <- length(data)

    lower <- qtnew(limit_1, df = n - 1, mean = samples_mean,
                   scale = sqrt(samples_variance) / sqrt(n))
    upper <- qtnew(limit_2, df = n - 1, mean = samples_mean,
                   scale = sqrt(samples_variance) / sqrt(n))
    results <- c(lower, upper)
    return(results)

}

interval <- mu_interval(windshieldy1)
cat("95% Posterior Interval:", interval, "\n")
```

```
95% Posterior Interval: 13.47808 15.74436
```

The point estimate, represented as 14.6112222, corresponds to the expected mean value.

The 95% posterior interval, denoted as 13.4780812, 15.7443633, signifies the range within which the mean is anticipated to be located.

```
mu_pdf <- function(data, x){
    # Compute necessary parameters here.
    # These are the correct parameters for `windshieldy_test`
    # with the provided uninformative prior.
    n = length(data)
    df = n - 1
    location = mu_point_est(data)
    samples_variance  <- var(data)
    scale = sqrt(samples_variance) / sqrt(n)
    # Use the computed parameters as below to compute the PDF:
    dtnew(x, df, location, scale)
}

x_interval = mu_interval(windshieldy1, .999)
lower_x = x_interval[1]
upper_x = x_interval[2]
x = seq(lower_x, upper_x, length.out=1000)
plot(
    x, mu_pdf(windshieldy1, x), type="l",
    xlab=TeX(r'(average hardness $\mu$)'),
    ylab=TeX(r'(PDF of the posterior $p(\mu|y)$)')
)
```

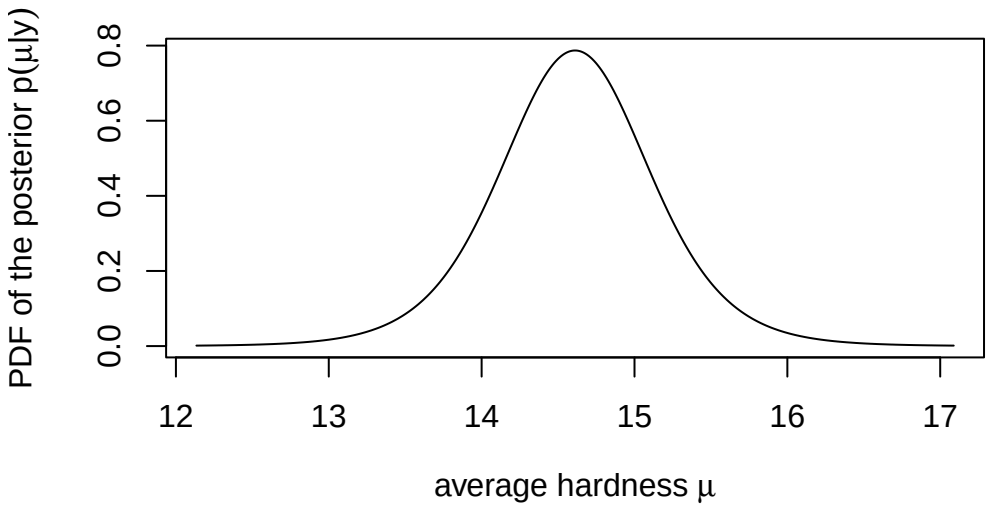


Figure 1: PDF of the posterior $p(\mu|y)$ of the average hardness μ

2.2.1 The posterior predictive distribution:

The posterior predictive distribution for a future observation, $\tilde{y}$, can be written as a mixture,$p(\tilde{y}|y) = \int\int p(\tilde{y}|\mu,\sigma^2,y)p(\mu,\sigma^2|y)d\mu d\sigma^2$. The first of the two factors in the integral is just the normal distribution for the future observation given the values of (μ,σ^2), and does not depend on y at all. To draw from the posterior predictive distribution, first draw μ,σ^2 from their joint posterior distribution and then simulate $\tilde{y} \sim \mathcal{N}(\mu,\sigma^2)$.

In fact, the posterior predictive distribution of $\tilde{y}$ is a t distribution with location $\bar{y}$, scale $\left(1+\frac{1}{n}\right)^{\frac{1}{2}} s$, and $n-1$ degrees of freedom. This analytic form is obtained using the same techniques as in the derivation of the posterior distribution of μ. Specifically, the distribution can be obtained by integrating out the parameters μ and σ^2 according to their joint posterior distribution. We can identify the result more easily by noticing that the factorization $p(\tilde{y}|\sigma^2,y) = \int p(\tilde{y}|\mu,\sigma^2,y)p(\mu|\sigma^2,y)d\mu$ leads to $p(\tilde{y}|\sigma^2,y) = \mathcal{N}(\tilde{y}|\bar{y},(1+\frac{1}{n})\sigma^2)$, which is the same, up to a changed scale factor, as the distribution of $\mu|\sigma^2,y$. For the deriving steps, you can check page 66 of BDA3 book.

2.3 (c)

Keep the below name and format for the functions to work with `markmyassignment`:

```
# Useful functions: mean(), length(), sqrt(), sum()
# and qtnew(), dtnew() (from aaltobda)

mu_pred_point_est <- function(data) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    samples_mean = mean(data)
    return(samples_mean)
}

p_predictive_estimate <- mu_pred_point_est(windshieldy1)
cat("Point Predictive Estimate:", p_predictive_estimate, "\n")
```

```
Point Predictive Estimate: 14.61122
```

```r
mu_pred_interval <- function(data, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    samples_mean <- mean(data)
    samples_var <- var(data)
    n <- length(data)
    limit_1 <- (1 - prob) / 2
    limit_2 <- limit_1 + prob
    pred_1 <- qtnew(limit_1, n-1, samples_mean,
                    scale = sqrt( 1 + 1 / n ) * sqrt(samples_var))
    pred_2 <- qtnew(limit_2, n-1, samples_mean,
                    scale = sqrt( 1 + 1 / n ) * sqrt(samples_var))
    result <- c(pred_1, pred_2)

    return(result)
}

interval_predictive <- mu_pred_interval(windshieldy1)
cat("95% Posterior Predictive Interval:", interval_predictive, "\n")
```

```
95% Posterior Predictive Interval: 11.02792 18.19453
```

```r
mu_pred_pdf <- function(data, x){
    # Compute necessary parameters here.
    # These are the correct parameters for `windshieldy_test`
    # with the provided uninformative prior.
    n = length(data)
    df = n - 1
    location = mu_pred_point_est(data)
    samples_variance  <- var(data)
    scale = sqrt(1 + 1 / n) * sqrt(samples_variance)
    # Use the computed parameters as below to compute the PDF:
    dtnew(x, df, location, scale)
}

x_interval = mu_pred_interval(windshieldy1, .999)
lower_x = x_interval[1]
upper_x = x_interval[2]
x = seq(lower_x, upper_x, length.out=1000)
plot(
    x, mu_pred_pdf(windshieldy1, x), type="l",
    xlab=TeX(r'(new hardness observation $\tilde{y}$)'),
    ylab=TeX(r'(PDF of the posterior predictive $p(\tilde{y}|y)$)')
)
```

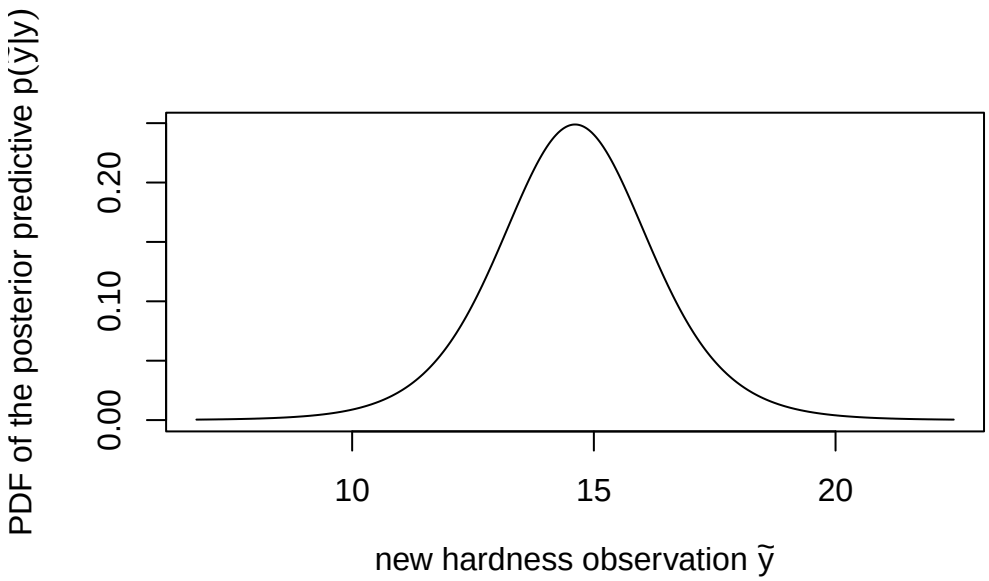


Figure 2: PDF of the posterior predictive $p(\tilde{y}|y)$ of a new hardness observation $\tilde{y}$

3 Inference for the difference between proportions (3 points)

3.1 (a)

The noninformative prior can be expressed as:

$$p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

Given that the observations follow a Binomial model, the likelihood is defined as:

$$p(y|\theta) \propto \binom{n}{y} \theta^y (1-\theta)^{n-y} = Binomial(n, y)$$

The posterior distribution can be derived as follows:

$$p(\theta|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$$

Now, let's apply this to the control group, where 39 out of 674 individuals died. The likelihood is as following:

$$p(y|\theta) \propto Binomial(674, 39)$$

Given the prior as :

$$Beta(1, 1)$$

The posterior distribution is calculated as:

$$p(\theta|y) \propto Beta(1 + 39, 1 + 674 - 39) = Beta(40, 636)$$

Similarly, for the treatment group, where 22 out of 680 individuals died.

The likelihood is given by:

$$p(y|\theta) \propto Binomial(680, 22)$$

Given the same prior as:

$Beta(1, 1)$.

The posterior distribution is calculated as:

$$p(\theta|y) \propto Beta(1 + 22, 1 + 680 - 22) = Beta(23, 659)$$

3.2 (b)

Given the two posterior distributions, the samples can be extracted:

```
set.seed(4711)
ndraws = 10000
p0 <- rbeta(ndraws, 40, 636)
p1 <- rbeta(ndraws, 23, 659)
```

Keep the below name and format for the functions to work with `markmyassignment`:

```
# Useful function: mean(), quantile()

posterior_odds_ratio_point_est <- function(p0, p1) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    p2 <- (p1 / ( 1 - p1 ))/( p0 / (1 - p0))
    e <- mean(p2)
    return(e)
}

point_estimate <- posterior_odds_ratio_point_est(p0, p1)
cat("Point Estimate of Posterior Odds Ratio:", point_estimate, "\n")
```

```
Point Estimate of Posterior Odds Ratio: 0.5709403
```

```
posterior_odds_ratio_interval <- function(p0, p1, prob = 0.95) {
    # Do computation here, and return as below.
    # This is the correct return value for the test data provided above.
    p2 <- (p1 /( 1 - p1 ))/(p0 /(1 - p0))
    limit_1 <- (1-prob)/2
    limit_2 <- limit_1 + prob
    lower <- quantile(p2, probs = limit_1)
    upper <- quantile(p2, probs = limit_2)
    result <- c(lower, upper)
    return(result)
}

posterior_interval <- posterior_odds_ratio_interval(p0, p1, 0.95)

cat("95% Posterior Odds Ratio Interval:",
```

```
    posterior_interval[1], "-", posterior_interval[2], "\n")
```

```
95% Posterior Odds Ratio Interval: 0.3209808 - 0.9284053
```

3.3 (c)

The selected prior, represented as Beta$(1, 1)$, is classified as a noninformative prior, indicating its minimal impact on the posterior distribution. In this context, as we analyze the way the posterior is formulated:

$$p(\theta|y) \propto Beta(\theta|\alpha + y, \beta + n - y)$$

It becomes evident that when the sample size (n) and the number of positive outcomes (y) are sufficiently large, the influence of the chosen prior becomes negligible. In this specific example, with both α and β set to 1, and n being 674 with y equaling 39, the prior's impact is inconsequential and does not alter the overall trend of the posterior distribution. Assuming a prior distribution of $Beta(\alpha = 0.06, \beta = 0.54)$ obtained from a logistic regression, there are lower odds of mortality if using a beta-blocker. In fact, the probability that the beta-blockers reduced the odds of dying when compared to no using them is 95%.

```
#B(0.06,0.54)
set.seed(4711)
ndraws = 10000
p0 <- rbeta(ndraws, 39.06, 653.54)
p1 <- rbeta(ndraws, 22.06, 658.54)
```

```
# Useful function: mean(), quantile()

point_estimate <- posterior_odds_ratio_point_est(p0, p1)
cat("Point Estimate of Posterior Odds Ratio:", point_estimate, "\n")
```

```
Point Estimate of Posterior Odds Ratio: 0.5767649
```

```
posterior_interval <- posterior_odds_ratio_interval(p0, p1, 0.95)
cat("95% Posterior Odds Ratio Interval:",
    posterior_interval[1], "-", posterior_interval[2], "\n")
```

```
95% Posterior Odds Ratio Interval: 0.3206591 - 0.9440159
```

```
#B(1,2)
set.seed(4711)
ndraws = 10000
p0 <- rbeta(ndraws, 40, 660)
p1 <- rbeta(ndraws, 23, 637)
```

```
# Useful function: mean(), quantile()

point_estimate <- posterior_odds_ratio_point_est(p0, p1)
cat("Point Estimate of Posterior Odds Ratio:", point_estimate, "\n")
```

```
Point Estimate of Posterior Odds Ratio: 0.6131031
```

```r
posterior_interval <- posterior_odds_ratio_interval(p0, p1, 0.95)
cat("95% Posterior Odds Ratio Interval:",
    posterior_interval[1], "-", posterior_interval[2], "\n")
```

```
95% Posterior Odds Ratio Interval: 0.3428568 - 0.9939253
```

4 Inference for the difference between normal means (3 points)

Loading the library and the data.

```r
data("windshieldy2")
# The new data are now stored in the variable `windshieldy2`.
# The below displays the first few rows of the new data:
head(windshieldy2)
```

```
[1] 15.980 14.206 16.011 17.250 15.993 15.722
```

4.1 (a)

In the scenario where the sample standard deviations, σ_1 and σ_2, are unknown, we adopt a noninformative prior distribution defined as:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

The likelihood function for this setup is expressed as:

$$p(y|\mu, \sigma^2) \propto \sigma^{-n} exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

Here, s^2 is computed as:

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Now, the joint posterior distribution can be formulated as:

$$p(\mu, \sigma^2|y) \propto \sigma^{-n}\sigma^{-2} exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

Of particular interest in this context is the marginal posterior distribution for μ, which simplifies to:

$$p(\mu|y) \propto [1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}]^{-n/2}$$

Remarkably, this distribution corresponds to a t-distribution:

$$p(\mu|y) \propto t_{n-1}(\bar{y}, s^2|n)$$

.

4.2 (b)

```
# Useful functions: mean(), length(), sqrt(), sum(),
# rtnew() (from aaltobda), quantile() and hist().

n1 <- length(windshieldy1)
sample_mean_win1 <- mean(windshieldy1)
sample_var_win1 <- var(windshieldy1)

n2 <- length(windshieldy2)
sample_mean_win2 <- mean(windshieldy2)
sample_var_win2 <- var(windshieldy2)

cat("Sample Sizes (n):\n")
```

```
Sample Sizes (n):
```

```
cat("windshieldy1:", n1, "\n")
```

```
windshieldy1: 9
```

```
cat("windshieldy2:", n2, "\n")
```

```
windshieldy2: 13
```

```
cat("\nSample Means:\n")
```

```
Sample Means:
```

```
cat("windshieldy1:", sample_mean_win1, "\n")
```

```
windshieldy1: 14.61122
```

```
cat("windshieldy2:", sample_mean_win2, "\n")
```

```
windshieldy2: 15.82108
```

```
cat("\nSample Variances:\n")
```

```
Sample Variances:
```

```r
cat("windshieldy1:", sample_var_win1, "\n")
```

```
windshieldy1: 2.173153
```

```r
cat("windshieldy2:", sample_var_win2, "\n")
```

```
windshieldy2: 0.7614481
```

```r
s_n_1 <- sample_var_win1^2 / n1
s_n_2 <- sample_var_win2^2 / n2

cat("\ns^2/n:\n")
```

```
s^2/n:
```

```r
cat("windshieldy1:", s_n_1, "\n")
```

```
windshieldy1: 0.5247326
```

```r
cat("windshieldy2:", s_n_2, "\n")
```

```
windshieldy2: 0.04460024
```

For the two data sets:

$$p(\mu_1|y_1) = t_8(14.6112222, 0.5247326)$$

$$p(\mu_2|y_2) = t_{12}(15.8210769, 0.0446002)$$

```r
ndraws = 10000
p1 <- rtnew(ndraws, n1, sqrt(1 + 1 / n1) * sqrt(sample_var_win1))
p2 <- rtnew(ndraws, n2, sqrt(1 + 1 / n2) * sqrt(sample_var_win2))
p_d <- p1 - p2
```

```r
posterior_d_interval<-function(p, prob){
  limit_1 <- (1 - prob) / 2
  limit_2 <- limit_1 + prob
  a <- quantile(p, probs = limit_1)
  b <- quantile(p, probs = limit_2)
  result <- c(a, b)
  return(result)
}
```

```r
posterior_interval_d <- posterior_d_interval(p_d, 0.95)
cat("Posterior Interval (95% confidence):\n")
```

```
Posterior Interval (95% confidence):
```

```r
cat("Lower Limit:", posterior_interval_d[1], "\n")
```

```
Lower Limit: -2.547429
```

```r
cat("Upper Limit:", posterior_interval_d[2], "\n")
```

```
Upper Limit: 3.830725
```

The -2.5474293, 3.8307253 are the limits for the 95% posterior interval.

```r
posterior_d_point_est <- function(p) {
  est <- mean(p)
  return(est)
}

point_d_estimate <- posterior_d_point_est(p_d)

cat("Posterior Point Estimate:\n")
```

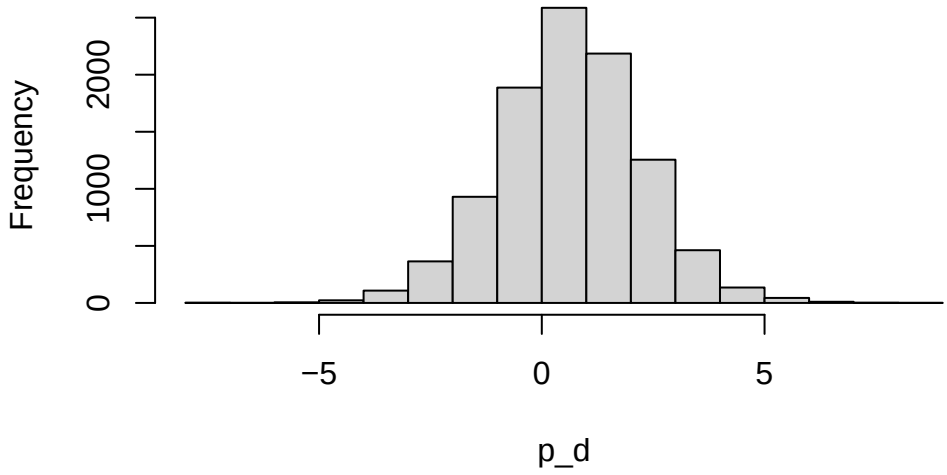
```
Posterior Point Estimate:
```

```r
cat("Point Estimate:", point_d_estimate, "\n")
```

```
Point Estimate: 0.6525621
```

The point estimate is 0.6525621.

```r
hist(p_d, main = "Histogram of 10000 samples from the difference of means")
```



4.3 (c)

The probability that the means are the same is zero. This is a matter of how the problem is defined. It is a hypothesis testing problem, so the null hypothesis consists in assuming that the computations from the data sets are true, while the alternative hypothesis can be that the subtraction of the means is either greater than zero or smaller than zero or that the subtraction of the means is different than zero. For this reason, the concept of the means taking the same value is not considered in the distribution. Note that the probability of a single point is always zero. Thus, we have $p(\mu_2 - \mu_1 = 0) = 0$.