Inference for numerical data

Complete all **Exercises**, and submit answers to **Questions** on the Coursera platform.

Getting Started

Load packages

Let's load the necessary packages for this week's lab:

```
library(statsr)
library(dplyr)
library(ggplot2)
```

As usual, the data set and analysis functions will be provided by the `statsr` package and we will be using `dplyr` and `ggplot2` for manipulating and visualizing the data.

The data

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. These data contain information on both the expectant mothers and their children. We will be working with a random sample of the complete data set. For those of you who took the Inferential Statistics course as part of the Statistics with R specialization should recognize this as the same data set used in the Inference for numerical data lab where we used frequentist inference methods to explore these data.

You can load the `nc` data set into our workspace using the `data` function once the `statsr` package is loaded.

```
data(nc)
```

This data set consists of 1000 observations on 13 different variables, some categorical and some numerical. The definition of each variable is as follows:

variable	description
`fage`	father's age in years.
`mage`	mother's age in years.
`mature`	maturity status of mother.
`weeks`	length of pregnancy in weeks.
`premie`	whether the birth was classified as premature (premie) or full-term.
`visits`	number of hospital visits during pregnancy.
`marital`	whether mother is `married` or `not married` at birth.
`gained`	weight gained by mother during pregnancy in pounds.
`weight`	weight of the baby at birth in pounds.
`lowbirthweight`	whether baby was classified as low birthweight (`low`) or not (`not low`).
`gender`	gender of the baby, `female` or `male`.
`habit`	status of the mother as a `nonsmoker` or a `smoker`.
`whitemom`	whether mom is `white` or `not white`.

EDA and Bayesian Inference - `weight`

As a first step in the analysis, we should take a look at the variables in the dataset and how R has encoded them. The most straight forward way of doing this is using the `str` command:

```
str(nc)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1000 obs. of  13 variables:
##  $ fage          : int  NA NA 19 21 NA NA 18 17 NA 20 ...
##  $ mage          : int  13 14 15 15 15 15 15 15 16 16 ...
##  $ mature        : Factor w/ 2 levels "mature mom","younger mom": 2 2 2 2 2 2 2 2 2 2 ...
##  $ weeks         : int  39 42 37 41 39 38 37 35 38 37 ...
##  $ premie        : Factor w/ 2 levels "full term","premie": 1 1 1 1 1 1 1 2 1 1 ...
##  $ visits        : int  10 15 11 6 9 19 12 5 9 13 ...
##  $ marital       : Factor w/ 2 levels "married","not married": 1 1 1 1 1 1 1 1 1 1 ...
##  $ gained        : int  38 20 38 34 27 22 76 15 NA 52 ...
##  $ weight        : num  7.63 7.88 6.63 8 6.38 5.38 8.44 4.69 8.81 6.94 ...
##  $ lowbirthweight: Factor w/ 2 levels "low","not low": 2 2 2 2 2 1 2 1 2 2 ...
##  $ gender        : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 2 2 2 1 ...
##  $ habit         : Factor w/ 2 levels "nonsmoker","smoker": 1 1 1 1 1 1 1 1 1 1 ...
##  $ whitemom      : Factor w/ 2 levels "not white","white": 1 1 2 2 1 1 1 1 2 2 ...
```

As you review the variable summaries, consider which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

1. How many of the 13 variables are categorical?
   5
   6
   7
   8

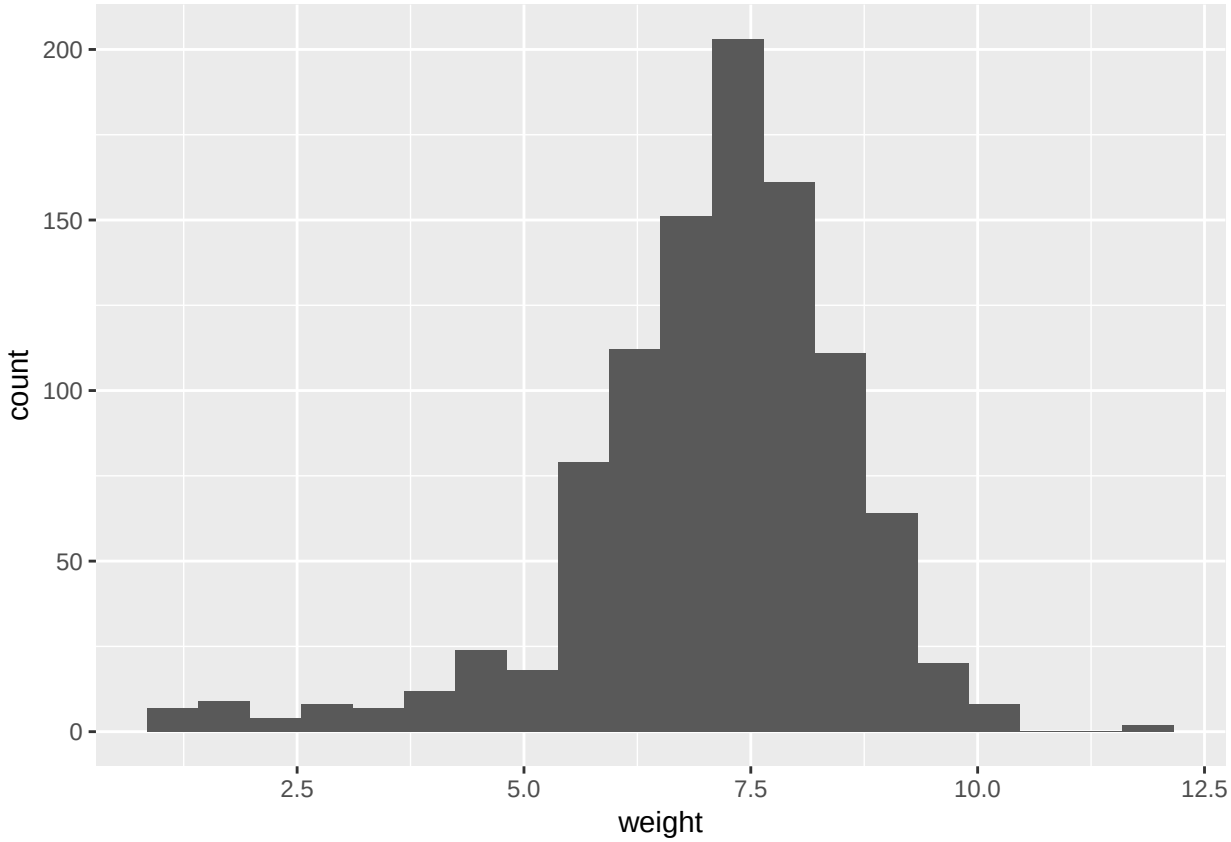
We will start with analyzing the weight of the babies at birth, which are contained in the variable `weight`.

Exercise: Using visualization and summary statistics, describe the distribution of weight of the babies at birth.

```
# type your code for the Exercise here, and Knit
summary(nc$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.380   7.310   7.101   8.060  11.750
```

```
ggplot(data = nc) +
  geom_histogram(aes(x = weight), bins = 20)
```



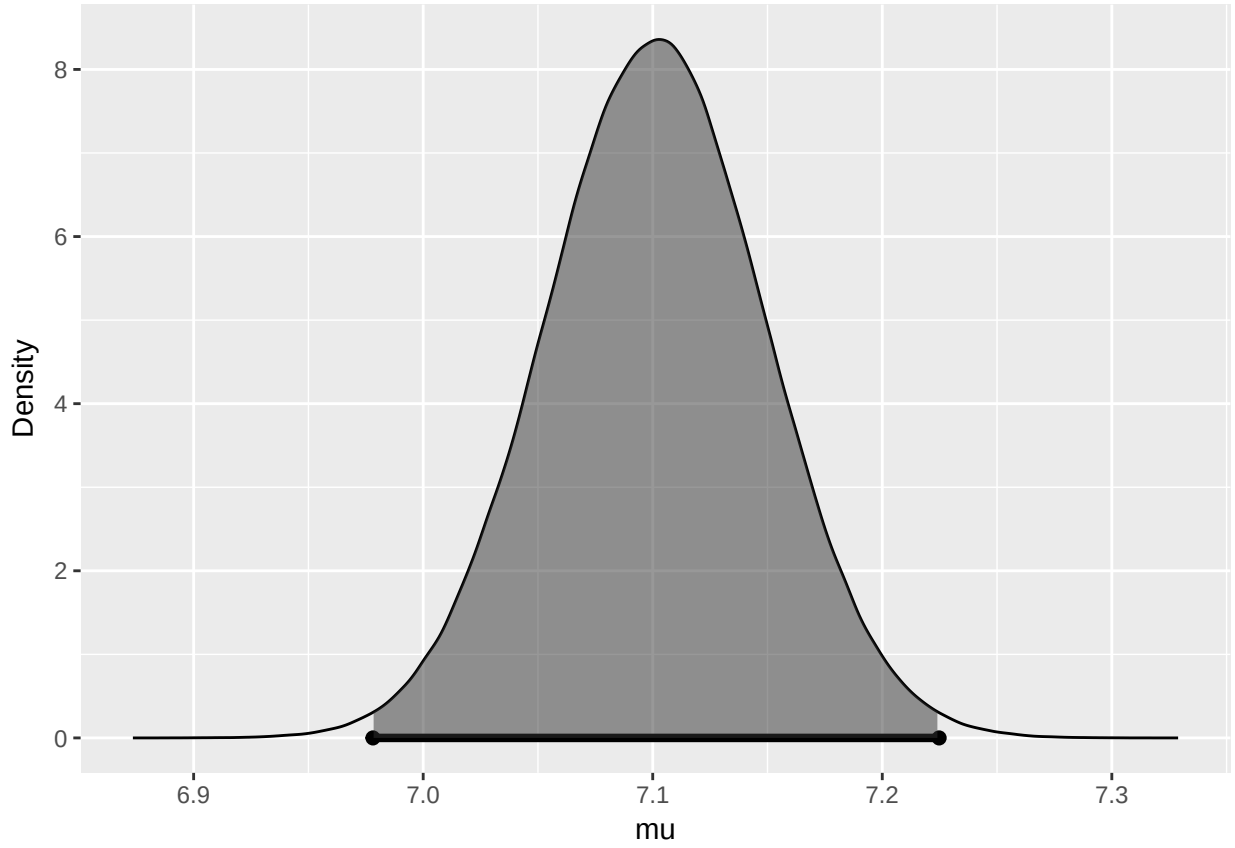
2. Which of the following best describes the distribution of `weight`?
   Left skewed
   Right skewed
   Uniformly distributed
   Normally distributed

As with the frequentist approach we use these data to perform basic inference on μ the average weight of all babies born in North Carolina. To do this we will use the `bayes_inference` function which will allow us construct credible intervals and calculate Bayes factors for a variety of different circumstances.

In order to construct a credible interval we must first provide the data, `weight` in this case, and then indicate that we want a credible interval (`type="ci"`) for a mean (`statistic="mean"`).

```
bayes_inference(y = weight, data = nc, statistic = "mean", type = "ci", cred_level = .99)
```

```
## Single numerical variable
## n = 1000, y-bar = 7.101, s = 1.5089
## (Assuming improper prior: P(mu, sigma^2) = 1/sigma^2)
##
## 99% CI: (6.9781 , 7.2248)
##
## Post. mean   = 7.101
## Post. median = 7.1011
## Post. mode   = 7.1026
```



The credible level for the interval can be specified using the `cred_level` argument.

3. Which of the following corresponds to the 99% credible interval for the average birth weight of all children born in North Carolina?
(7.00 , 7.19)
(6.98 , 7.22)
(6.94 , 7.26)
(6.94 , 7.27)

We can also conduct a Bayesian hypothesis test by calculating a Bayes factor, let's test to see if the average birth weight in North Carolina is significantly different from 7 pounds.

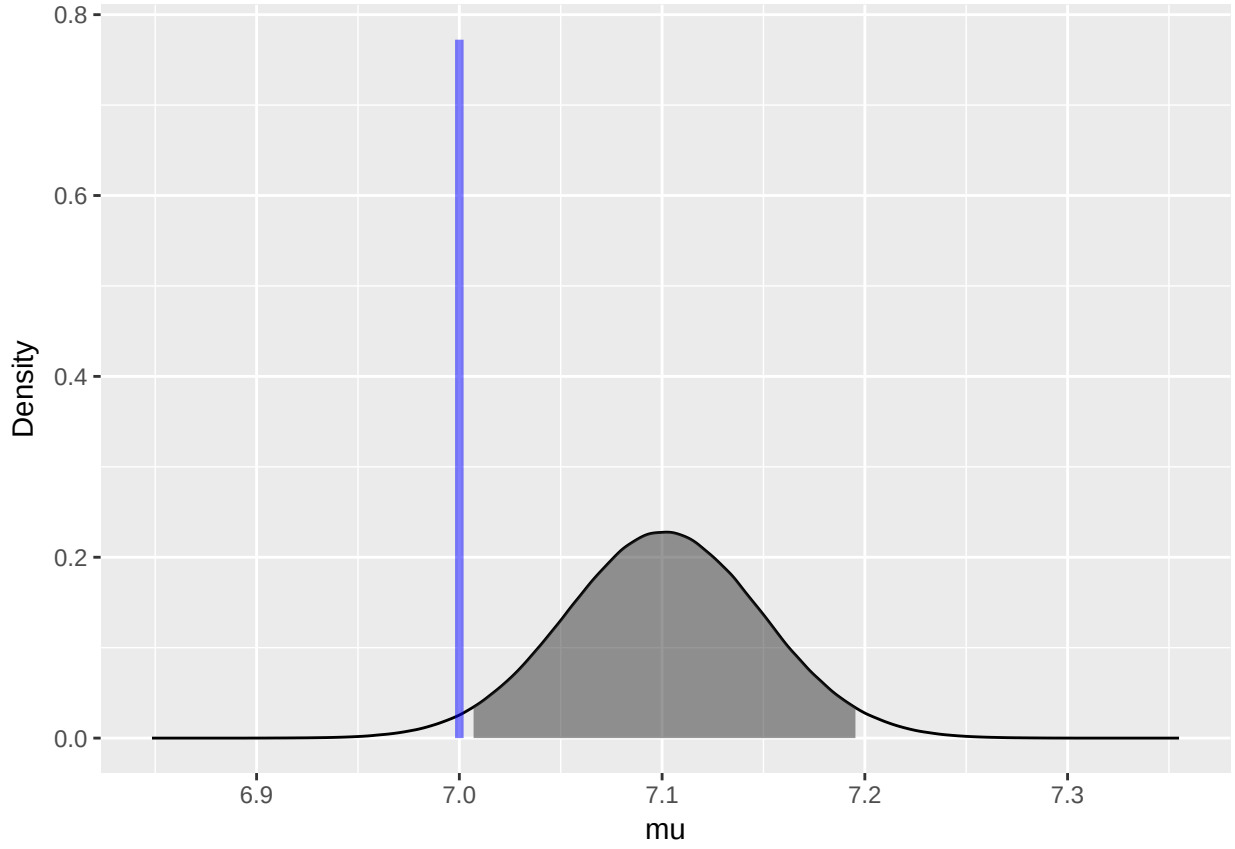
$$H_1 : \mu = 7$$
$$H_2 : \mu \neq 7$$

To conduct this hypothesis test we will again use the `bayes_inference` function but this time specify `type="ht"`, will then also need to provide the `null` and `alternative` arguments which define the null value (7) and the type of alternative hypothesis (`"twosided"`).

```
bayes_inference(y = weight, data = nc, statistic = "mean", type = "ht", null = 7,
                alternative = "twosided")
```

```
## Single numerical variable
## n = 1000, y-bar = 7.101, s = 1.5089
## (Assuming improper prior: P(mu, sigma^2) = 1/sigma^2)
##
## Hypotheses:
```

```
## H1: mu = 7
## H2: mu != 7
##
## Priors:
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 3.3915
## P(H1|data) = 0.7723
## P(H2|data) = 0.2277
```



4. Based of Jeffrey's scale for interpretation of a Bayes factors how should be describe the evidence in favor of H_1 from the results above?
   Not worth a bare mention
   Positive
   Strong
   Very Strong

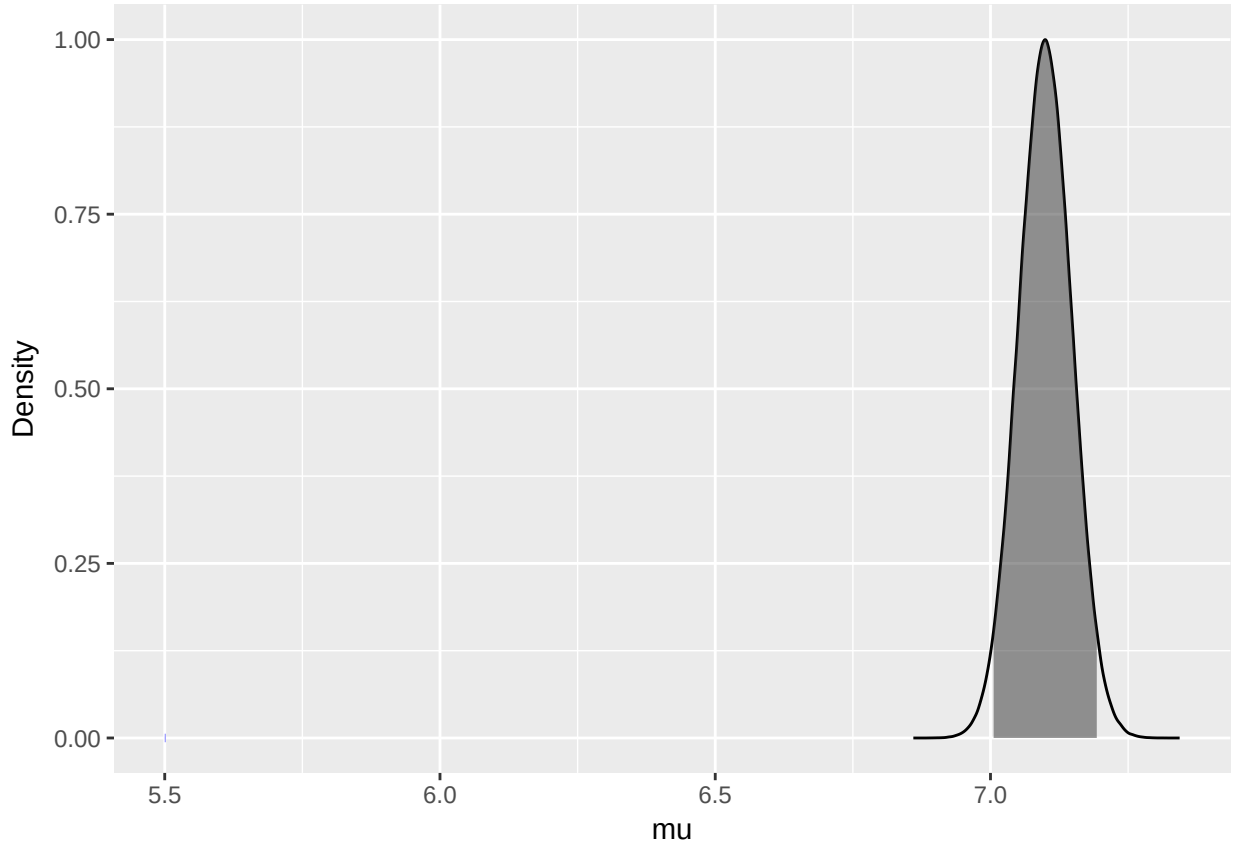
The graphical results show the magnitude of $P(H_1 \mid data)$ with the blue line and the posterior of μ given H_2 (scaled by $P(H_2 \mid data)$) via the black curve. The 95% credible interval for $\mu|data, H_2$ is given in grey.

Exercise: In the US low birth is defined as being less than 2500 grams ($\approx$ 5.5 lbs). Use the `bayes_inference` function to assess if the average birth weight in North Carolina is significantly different from this value. (The answer here should be obvious, but make sure that the Bayes factor result conforms with your intuition.)

```
# type your code for the Exercise here, and Knit
bayes_inference(y = weight, data = nc, statistic = "mean", type = "ht", null = 5.5,
```

```
                alternative = "twosided")
```

```
## Single numerical variable
## n = 1000, y-bar = 7.101, s = 1.5089
## (Assuming improper prior: P(mu, sigma^2) = 1/sigma^2)
##
## Hypotheses:
## H1: mu = 5.5
## H2: mu != 5.5
##
## Priors:
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H2:H1] = 9.42506e+161
## P(H1|data) = 0
## P(H2|data) = 1
```



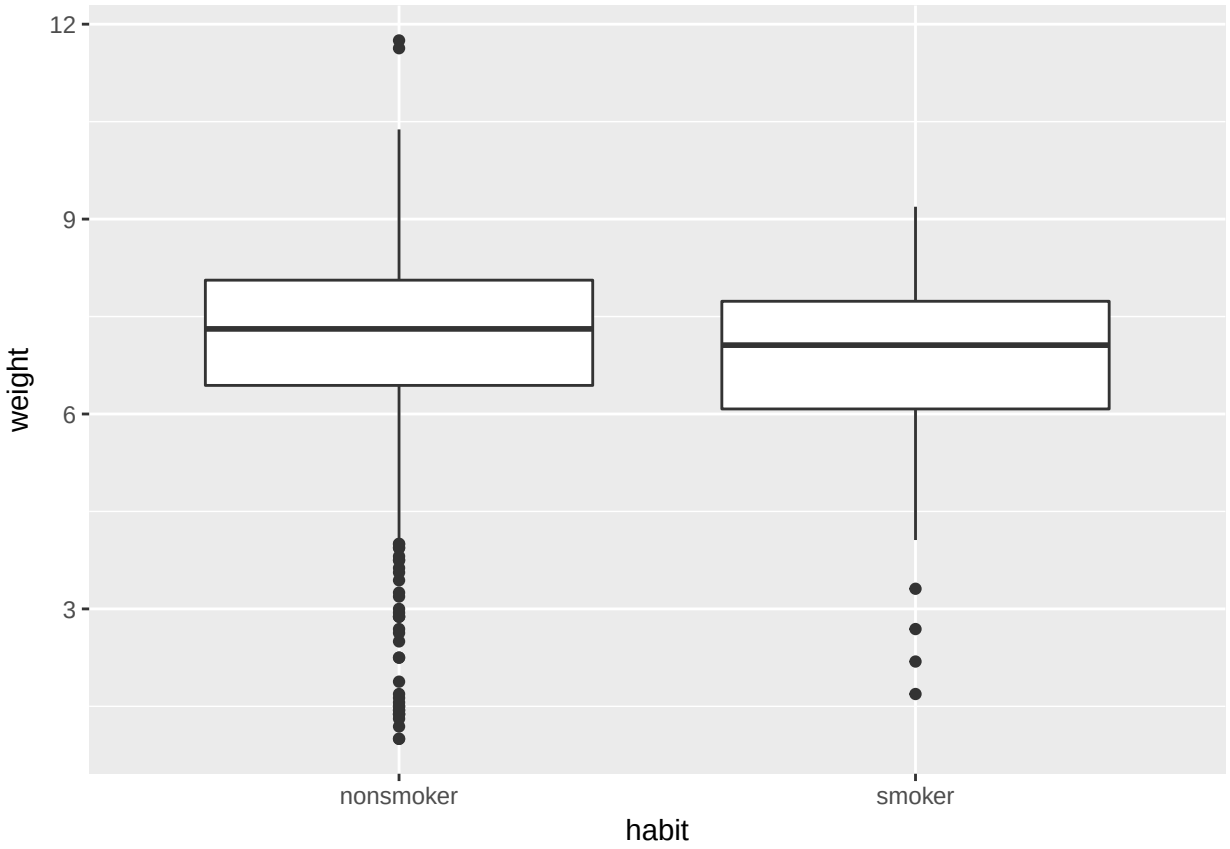
Inference for two means

Next, let's consider if the mother's smoking habit has any clear effect on the child's birth weight. Here we will use the variable `habit` to distinguish between smoking and non-smoking mothers. As with any analysis, a visualization is a good place to start, and will give us a better understanding of the data.

Exercise: Construct a side-by-side boxplot of `habit` and `weight` and compare the two distributions.

5. Which of the following is *false* about the relationship between habit and weight?
   Median birth weight of babies born to non-smoker mothers is slightly higher than that of babies born to smoker mothers.
   Range of birth weights of babies born to non-smoker mothers is greater than that of babies born to smoker mothers.
   Both distributions are extremely right skewed.
   The IQRs of the distributions are roughly equal.

```
# type your code for Question 5 here, and Knit
ggplot(data = nc[!is.na(nc$habit), ]) +
  geom_boxplot(aes(x = habit, y = weight))
```

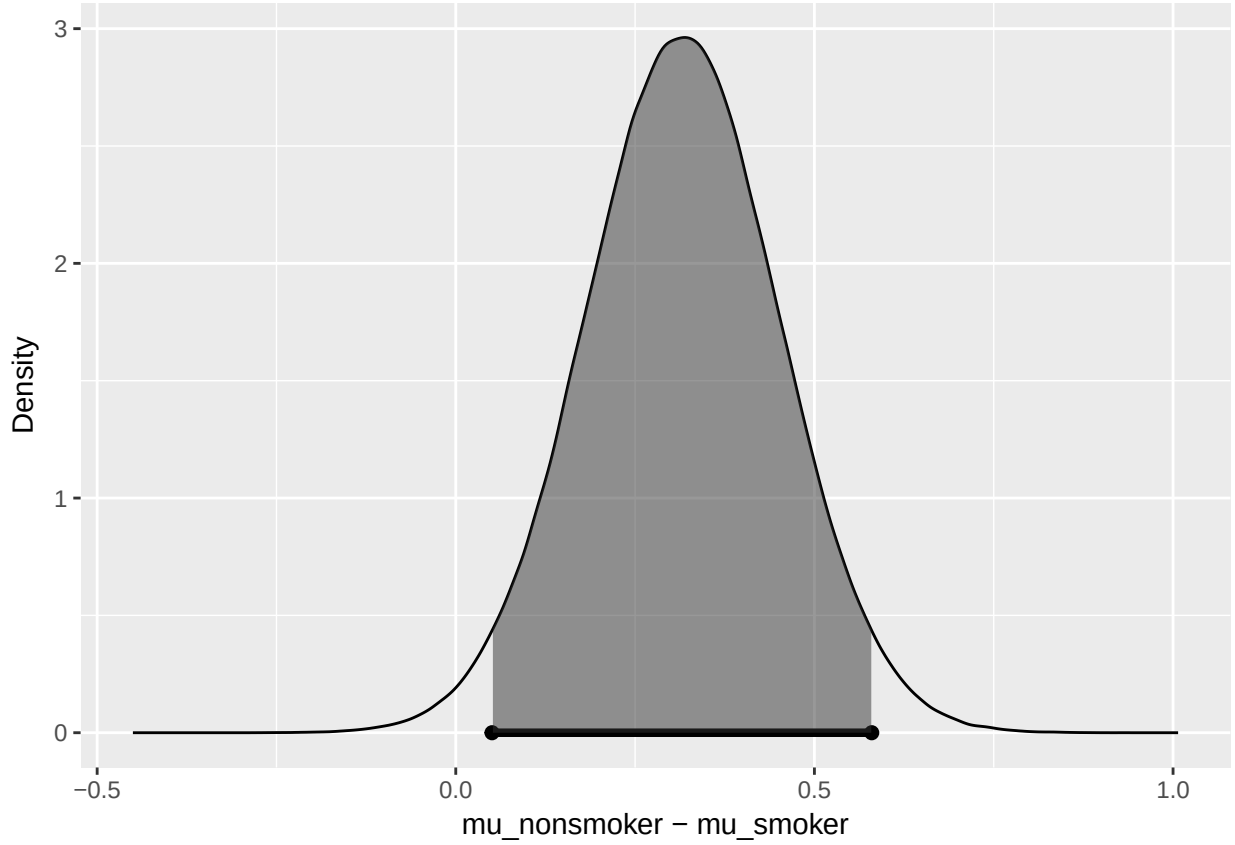


As before we can use the `bayes_inference` function to either construct a credible interval and or calculate a Bayes factor. The calls will be identical to the single mean case except now we will provide `habit` as a predictor variable (argument `x`). Note we also change `null=0`, since we are interested in comparing if the means of the two groups are equal or not.

```
bayes_inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ci")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## (Assuming independent Jeffrey's priors for mu and sigma^2)
##
## 95% Cred. Int.: (0.0505 , 0.5799)
##
## Post. mean   = 0.3154
## Post. median = 0.3156
```

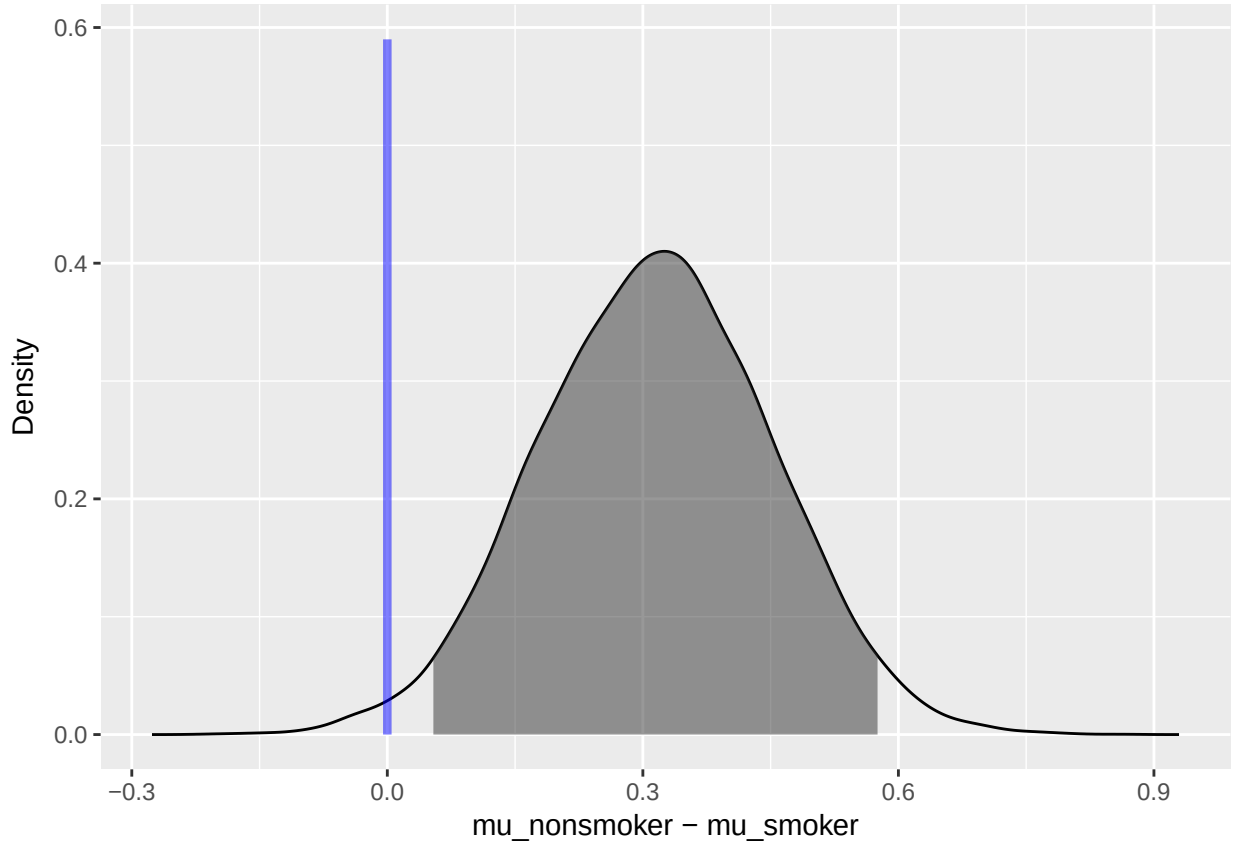
```
## Post. mode   = 0.3198
```



6. Based on the credible interval is there evidence that smoking reduces birth weight?
   Yes
   No

```
bayes_inference(y = weight, x = habit, data = nc, statistic = "mean", type = "ht", null = 0,
                alternative = "twosided")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, y_bar_nonsmoker = 7.1443, s_nonsmoker = 1.5187
## n_smoker = 126, y_bar_smoker = 6.8287, s_smoker = 1.3862
## (Assuming intrinsic prior on parameters)
## Hypotheses:
## H1: mu_nonsmoker  = mu_smoker
## H2: mu_nonsmoker != mu_smoker
##
## Priors:
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 1.4387
## P(H1|data) = 0.5899
## P(H2|data) = 0.4101
```



7. Based on the Bayes factor calculated above, how strong is evidence against H_1?
   Not worth a bare mention
   Positive
   Strong
   Very Strong

Inference for proportions

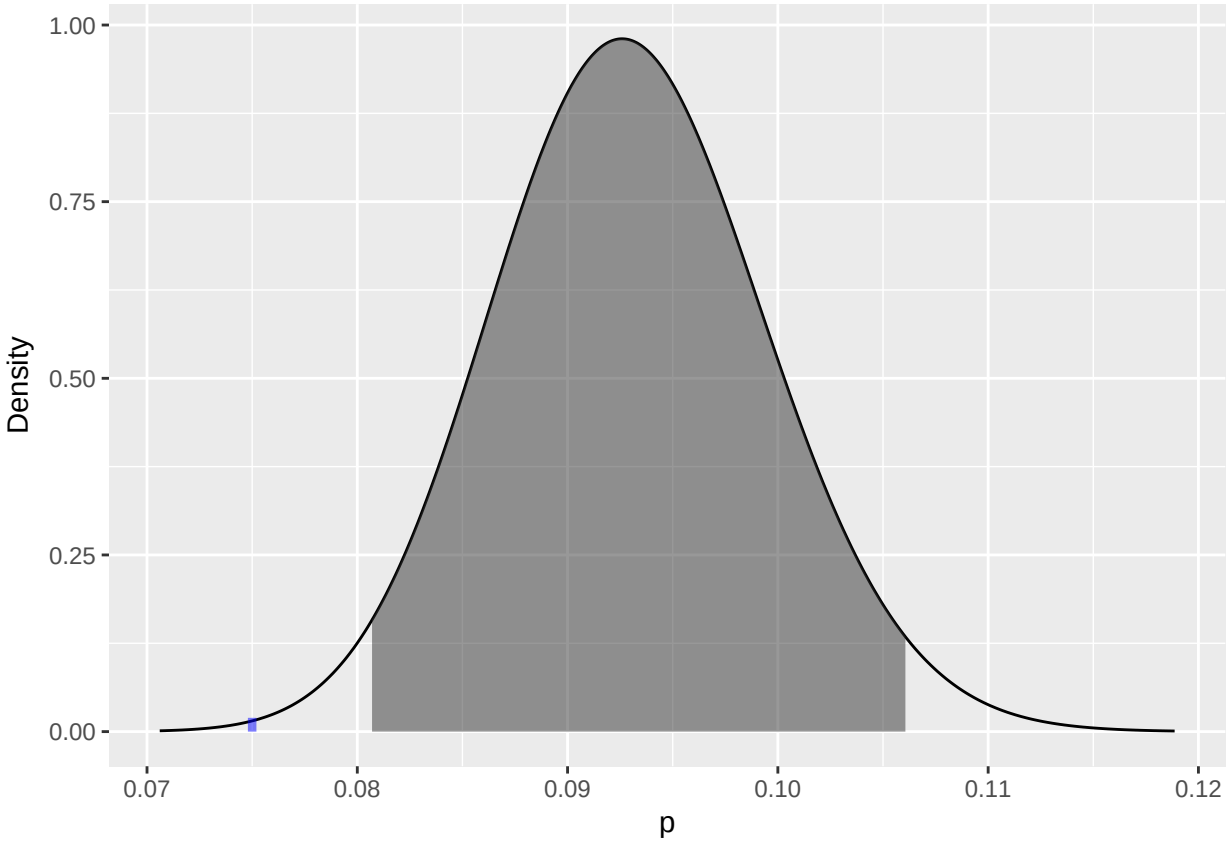
Finally, we can also conduct Bayesian inference when our outcome variable of interest is categorical. As with the frequentist inference function the only change is to specify a y argument that is categorical (with only two levels) and then specify which of its levels is the "success" using the `success` argument and finally change the `statistic` of interest to `"proportion"`.

For example if we want to test if only 7.5% of births in North Carolina are considered low birth weight we can calculate the Bayes factor using the following code:

```
bayes_inference(y = lowbirthweight, data = nc, success = "low", statistic = "proportion", type = "ht",
                null = 0.075, alternative = "twosided", beta_prior = c(75,925))
```

```
## Single categorical variable, success: low
## n = 1000, p-hat = 0.111
##
## Hypotheses:
## H1: p = 0.075
## H2: p != 0.075
##
## Priors:
```

```
## P(p) ~ Beta(a=75,b=925)
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H2:H1] = 50.3154
## P(H1|data) = 0.0195
## P(H2|data) = 0.9805
```



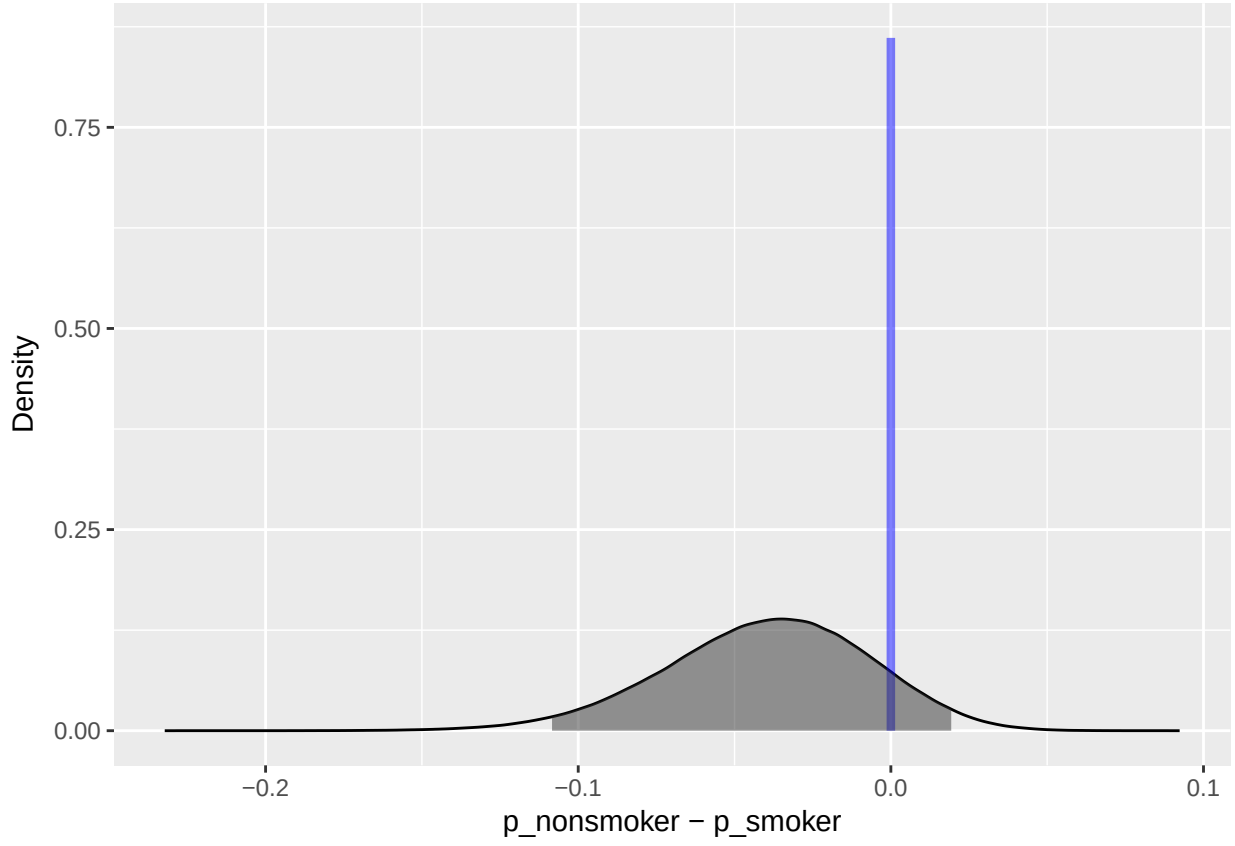
8. How would the Bayes factor above change if we were to increase the prior probability of H_2?
   Get bigger
   Get smaller
   Stay the same
9. How would the Bayes factor above change if we were to change the prior of p to be $Beta(75, 925)$?
   Get bigger
   Get smaller
   Stay the same

Using what you have learned so far, conduct a Bayesian inference procedure to evaluate whether these data provide evidence for or against smoking being associated with low birth weight and smoking being associated with premature birth.

10. These data provide _____ evidence _____ smoking affecting the chance of low birth weight.
    weak; for
    strong; for
    weak; against
    strong; against

```r
# type your code for Question 10 here, and Knit
bayes_inference(y = lowbirthweight, x = habit, data = nc, statistic = "proportion", type = "ht",
                null = 0, alternative = "twosided", success = "low", beta_prior1 = c(.5, .5),
                beta_prior2 = c(.5, .5))
```

```
## Response variable: categorical (2 levels, success: low)
## Explanatory variable: categorical (2 levels)
## n_nonsmoker = 873, p_hat_nonsmoker = 0.1054
## n_smoker = 126, p_hat_smoker = 0.1429
## Hypotheses:
## H1: p_nonsmoker =  p_smoker
## H2: p_nonsmoker != p_smoker
##
## Priors:
## P(p_nonsmoker) ~ Beta(a=0.5,b=0.5)
## P(p_smoker) ~ Beta(a=0.5,b=0.5)
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 6.2055
## P(H1|data) = 0.8612
## P(H2|data) = 0.1388
```

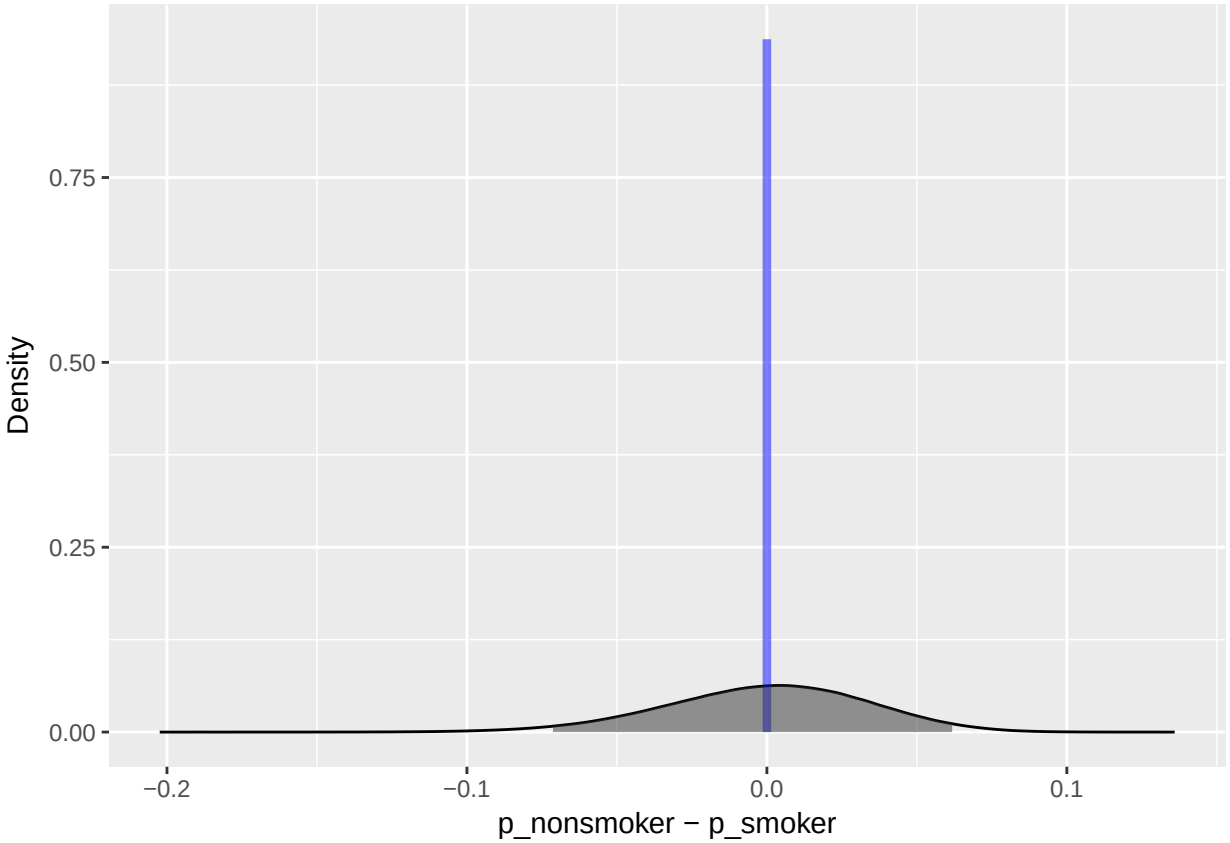


11. These data provide _____ evidence _____ smoking affecting the chance of premature birth.

weak; for
strong; for
weak; against
strong; against

```
# type your code for Question 11 here, and Knit
bayes_inference(y = premie, x = habit, data = nc, statistic = "proportion", type = "ht", null = 0,
                alternative = "twosided", success = "premie", beta_prior1 = c(.5, .5),
                beta_prior2 = c(.5, .5))
```

```
## Response variable: categorical (2 levels, success: premie)
## Explanatory variable: categorical (2 levels)
## n_nonsmoker = 872, p_hat_nonsmoker = 0.1525
## n_smoker = 126, p_hat_smoker = 0.1508
## Hypotheses:
## H1: p_nonsmoker =  p_smoker
## H2: p_nonsmoker != p_smoker
##
## Priors:
## P(p_nonsmoker) ~ Beta(a=0.5,b=0.5)
## P(p_smoker) ~ Beta(a=0.5,b=0.5)
## P(H1) = 0.5
## P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 14.8741
## P(H1|data) = 0.937
## P(H2|data) = 0.063
```


0.75
0.50
0.25
0.00
Density
−0.2
−0.1
0.0
0.1
p_nonsmoker − p_smoker