

Worksheet 02 Group 2 Exercise 4

Andrea Staub

Emanuel Mauch

Holly Vuarnoz

Jan Hohenheim (that's me!)

Sophie Haldemann

```
library(tidyverse)
library(glue)
library(ggthemes)

theme_set(theme_solarized_2())
```

Exercise 4

Per exercise 3:

$$m|y_{1:n} \sim \mathcal{N}\left(\frac{kn\bar{y} + \lambda\mu}{kn + \lambda}, \frac{1}{kn + \lambda}\right)$$

In our the code:

$$k := k$$

$$u := \mu$$

$$l := \lambda$$

With the given data:

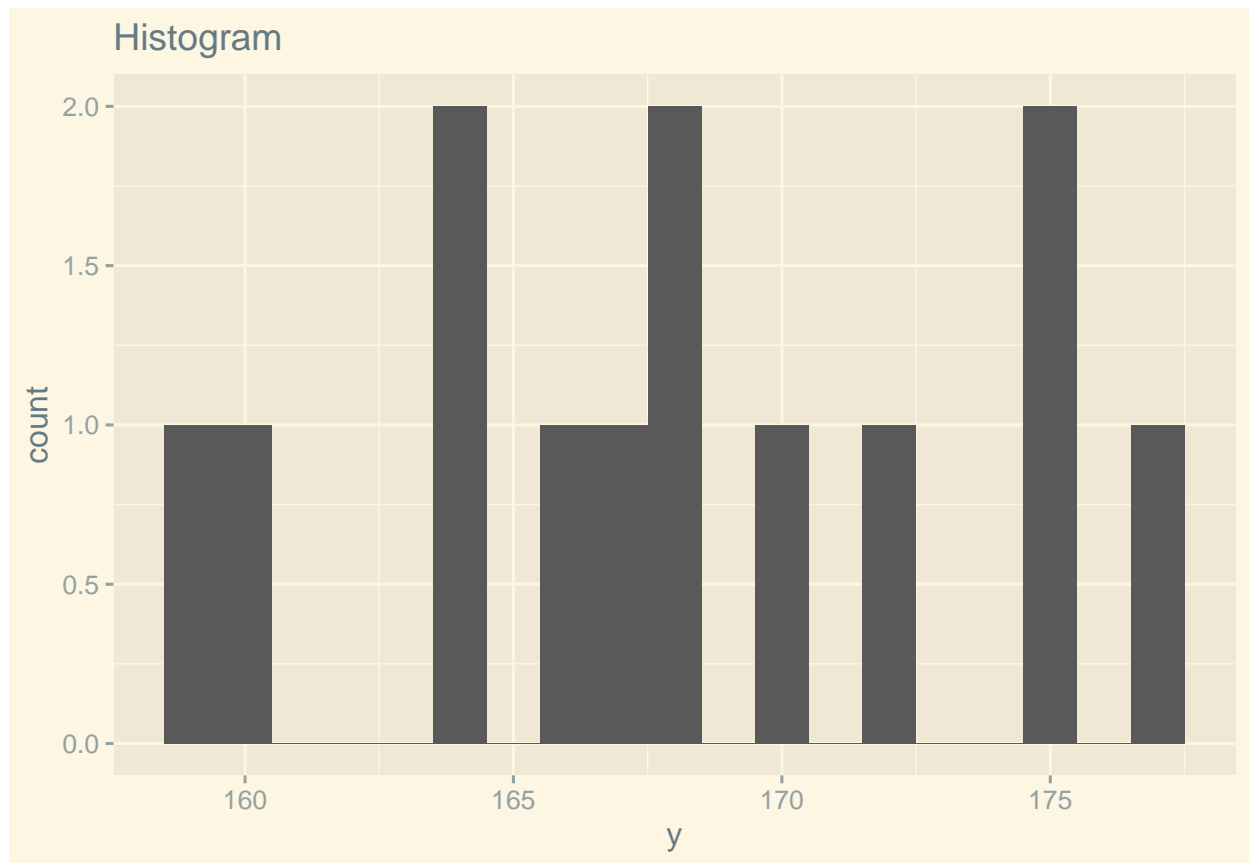
```
dat <- tibble(y = c(166, 168, 168, 177, 160, 170, 172, 159, 175, 164, 175, 167, 164))
k <- 1/900
u <- 161
l <- 1/70
```

a)

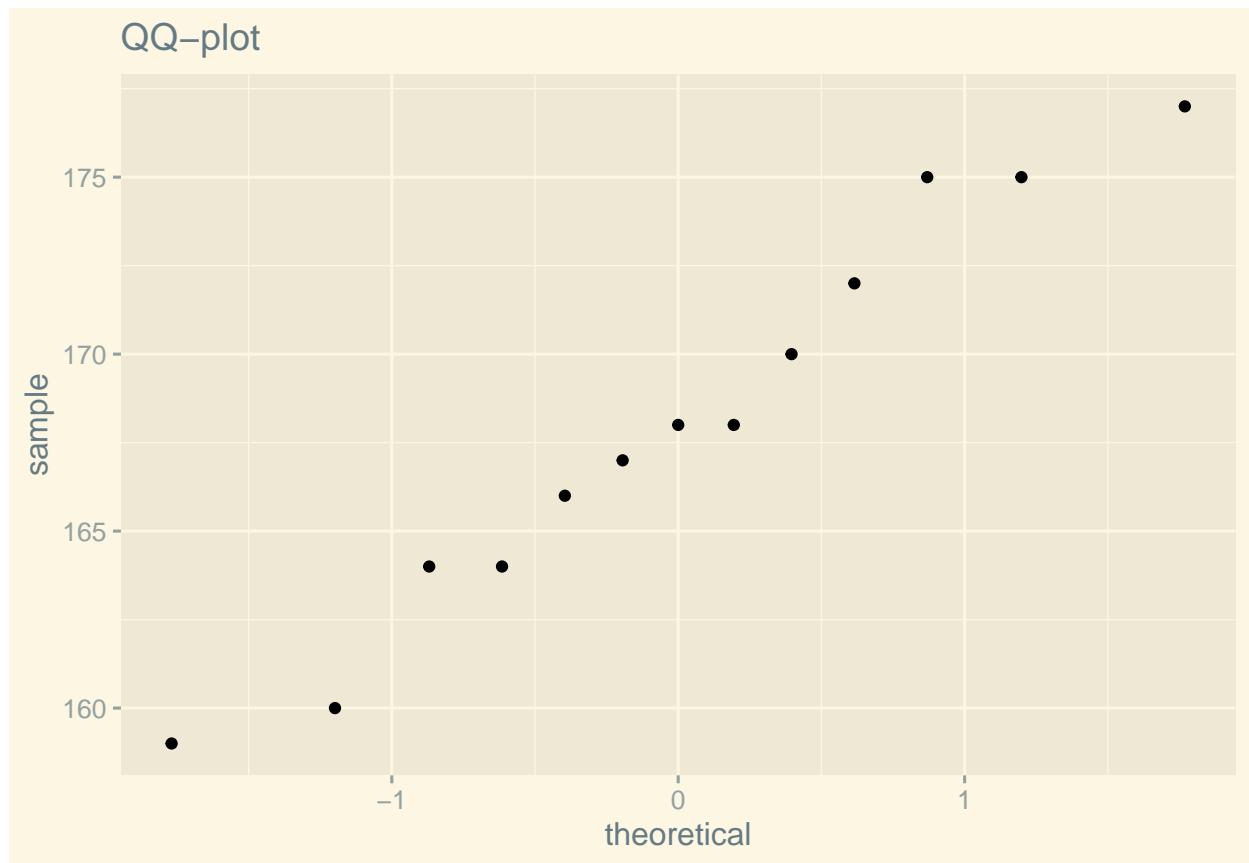
```
dat |> summary()
```

```
##           y
##  Min.      :159.0
##  1st Qu.:164.0
##  Median :168.0
##  Mean    :168.1
##  3rd Qu.:172.0
##  Max.    :177.0
```

```
dat |>  
  ggplot(aes(x = y)) +  
  geom_histogram(binwidth = 1) +  
  ggtitle("Histogram")
```

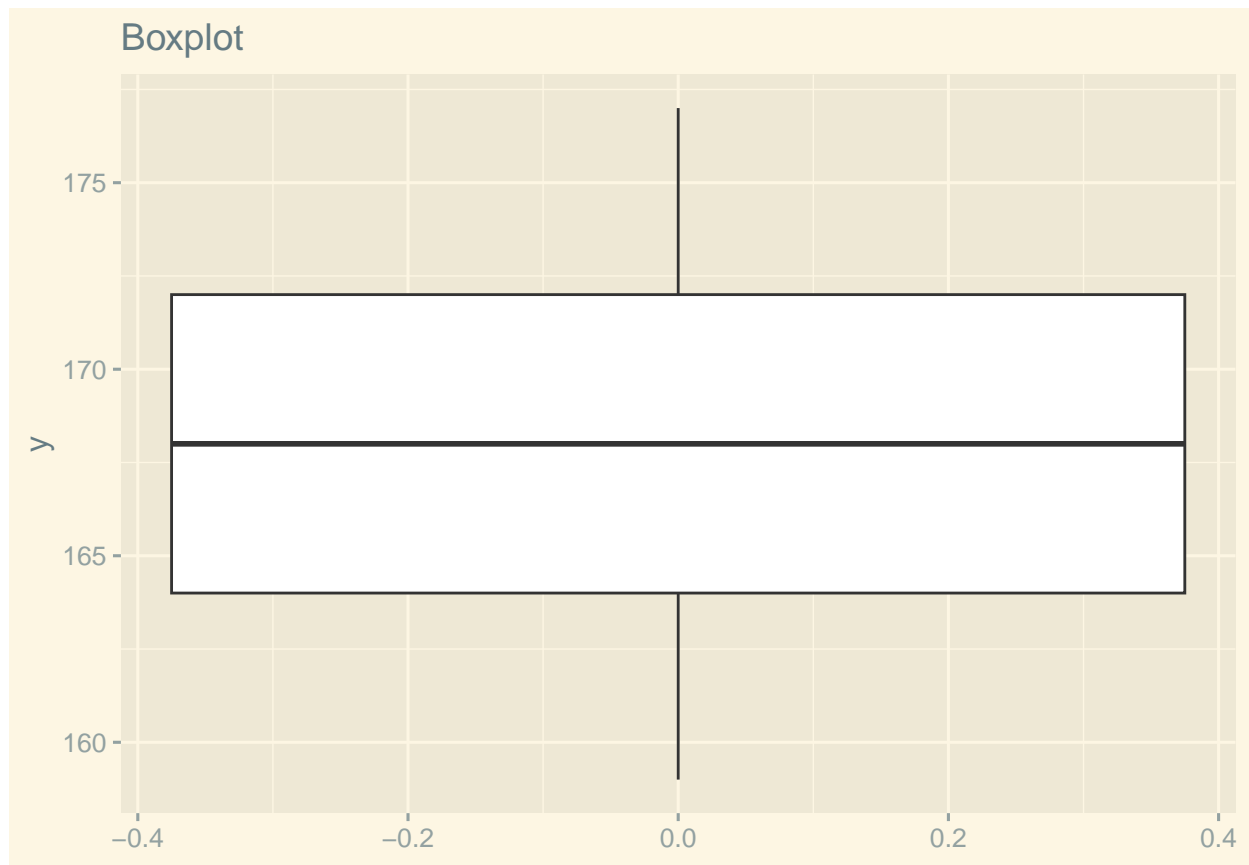


```
dat |>  
  ggplot(aes(sample = y)) +  
  stat_qq() +  
  ggtitle("QQ-plot")
```



The QQ plot tells us that the data is probably normally distributed

```
dat |>  
  ggplot(aes(y = y)) +  
  geom_boxplot() +  
  ggtitle("Boxplot")
```



The boxplot shows that there are no obvious outliers.

```

alpha <- 0.05 / 2 # two tailed
n <- length(dat$y)
mean <- mean(dat$y)
std_dev <- sd(dat$y)
std_err <- std_dev / sqrt(n)
t <- qt(1 - alpha, df = n - 1)
ci <- mean + c(-1, 1) * t * std_err
ci

```

```
## [1] 164.6722 171.4816
```

```

# or, alternatively
confint(lm(y ~ 1, data = dat))

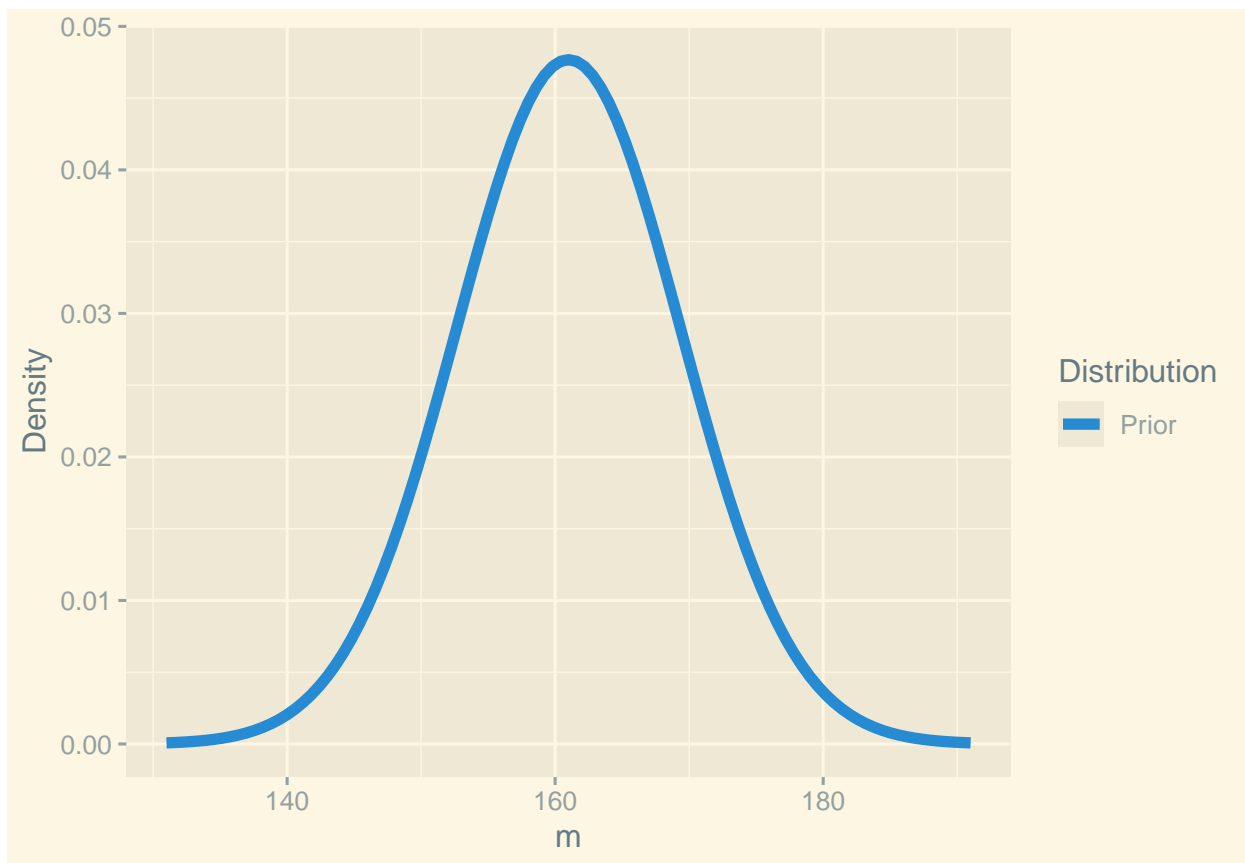
```

```
##                2.5 %    97.5 %
## (Intercept) 164.6722 171.4816
```

The correct frequentist interpretation of the confidence interval is tricky. Strictly, the 95% confidence interval says that if we were to sample from the same population many times and calculated the 95% confidence interval each time, 95% of these intervals would contain the true mean. The more commonly used interpretation is that the true mean is probably between 165 cm and 171 cm.

b)

```
ggnorm <- function(mean, sd, Distribution) {  
  stat_function(  
    fun = dnorm,  
    n = 101,  
    args = list(mean, sd),  
    linewidth = 2,  
    aes(color = Distribution)  
  )  
}  
  
std_dev <- 1/sqrt(1)  
  
tibble(x = u + 30 * c(-1, 1)) |>  
  ggplot(data = , aes(x)) +  
  ggnorm(u, std_dev, "Prior") +  
  labs(  
    x = "m",  
    y = "Density"  
  ) +  
  scale_colour_solarized()
```



```

pretty <- function(n) {
  # round to 2 decimal places
  n |>
    round(2) |>
    format(nsmall = 2)
}
glue("Prior mean: {pretty(u)} cm")
glue("Prior standard deviation: {pretty(std_dev)} cm")
glue("Prior median: {pretty(u)} cm")
ci_prior <- qnorm(c(0.025, 0.975), u, std_dev)
glue("95% CrI: [{pretty(ci_prior[1])} cm, {pretty(ci_prior[2])} cm]")
glue("(95% CI for data: [{pretty(ci[1])} cm, {pretty(ci[2])} cm])")
glue("P[X > 200] = {1 - pnorm(200, u, std_dev)} cm")

## Prior mean: 161.00 cm
## Prior standard deviation: 8.37 cm
## Prior median: 161.00 cm
## 95% CrI: [144.60 cm, 177.40 cm]
## (95% CI for data: [164.67 cm, 171.48 cm])
## P[X > 200] = 1.57039256865055e-06 cm

```

Our prior distribution tells us that the mean is probably around 161 cm and probably not below 144 cm or above 178 cm. It is extremely unlikely that the mean is above 200 cm.

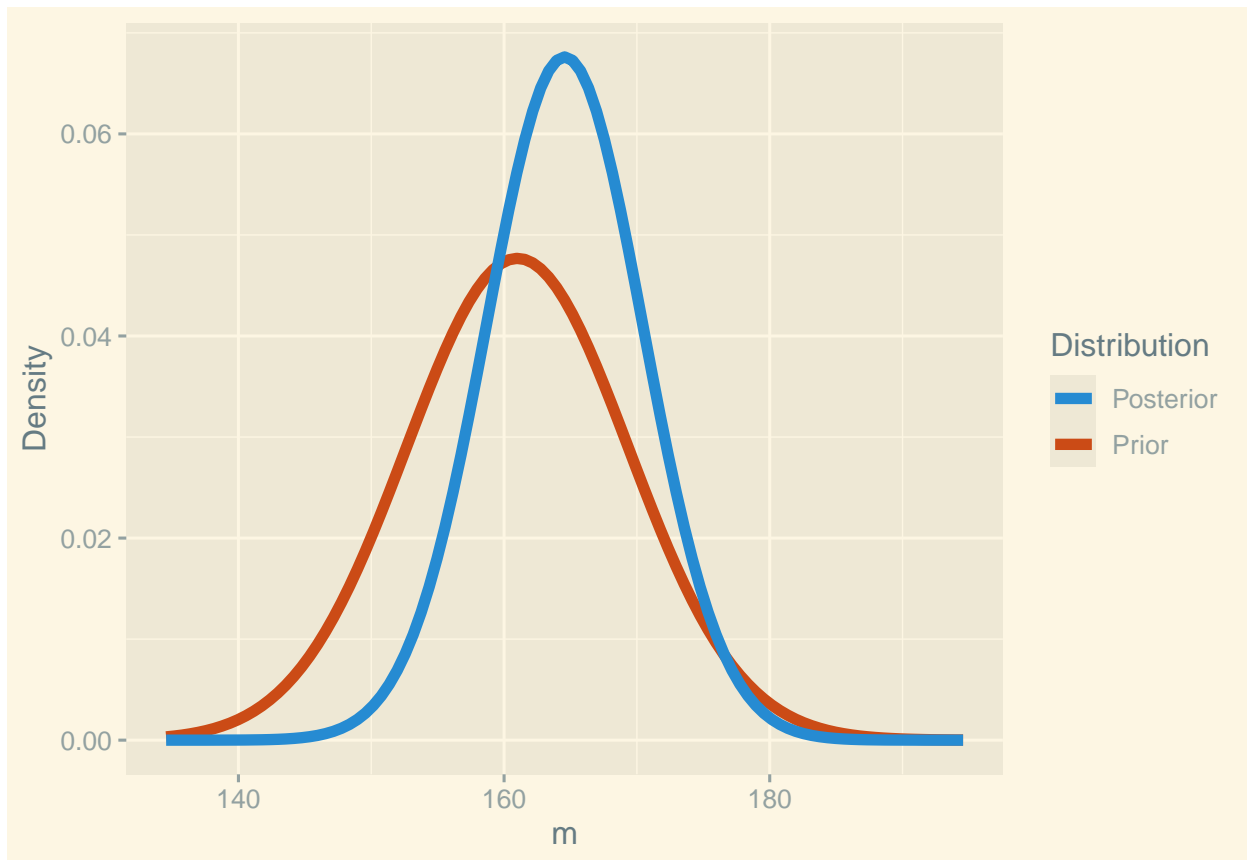
c)

$$m|y_{1:n} \sim \mathcal{N}(\frac{kn\bar{y} + \lambda\mu}{kn + \lambda}, \frac{1}{kn + \lambda})$$

```
u_prior <- u
std_dev_prior <- std_dev
mean_y = mean(dat$y)
u_posterior <- (k * n * mean_y + l * u_prior) / (k * n + l)
std_dev_posterior <- 1 / sqrt(k * n + l)
```



```
tibble(x = u_posterior + 30 * c(-1, 1)) |>
  ggplot(data = , aes(x)) +
  ggnorm(u_prior, std_dev_prior, "Prior") +
  ggnorm(u_posterior, std_dev_posterior, "Posterior") +
  labs(
    x = "m",
    y = "Density"
  ) +
  scale_colour_solarized()
```



```

glue("Prior -> Posterior")
glue("mean: {u_prior} -> {pretty(u_posterior)} cm")
glue("standard deviation: {pretty(std_dev_prior)} -> {pretty(std_dev_posterior)} cm")
glue("median: {pretty(u_prior)} -> {pretty(u_posterior)} cm")
ci_prior <- qnorm(c(0.025, 0.975), u_prior, std_dev_prior)
ci_posterior <- qnorm(c(0.025, 0.975), u_posterior, std_dev_posterior)
glue("95% CrI: [{pretty(ci_prior[1])}] cm, {pretty(ci_prior[2])}] cm" -> [{pretty(ci_posterior[1])}] cm, {pretty(ci_posterior[2])}] cm)

## Prior -> Posterior
## mean: 161 -> 164.56 cm
## standard deviation: 8.37 -> 5.90 cm
## median: 161.00 -> 164.56 cm
## 95% CrI: [144.60 cm, 177.40 cm] -> [152.99 cm, 176.12 cm]

```

Based on our data, we shifted our mean estimate from 161 cm to 164.6 cm. We now know that the true mean is probably between 153 cm and 176 cm. As we have more data to work with, our uncertainty expressed by the standard deviation has decreased, and thus our 95% CrI is narrower.