Problem Set 3

Solutions

Notes

Other students who I worked with on this assignment (if any):

Introduction

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please generate your solutions in R markdown and upload a knitted pdf document to Gradescope. Please put your name in the "author" section in the header.

The questions in this problem set use material from the slides on parameter estimation.

Context for Questions 1-3

We have seen that count k of successes from n Bernoulli trials is modeled as $Binomial(\text{size} = n, \text{probability} = p)$ then the maximum likelihood estimate of p equals $\frac{k}{n}$. Suppose this is repeated M times and k_1 successes are observed in n_1 Bernoulli trials, k_2 successes are observed in n_2 Bernoulli trials, and so on through k_M successes in n_M Bernoulli trials. The goal to find the maximum likelihood estimate of p if these are modeled as samples from $Binomial(\text{size} = n_1, \text{probability} = p)$, $Binomial(\text{size} = n_2, \text{probability} = p)$, and so on through $Binomial(\text{size} = n_M, \text{probability} = p)$.

Question 1

Consider the likelihood $L(k_1, k_2, ...k_n)$ function for $\{k_1, k_2, ...k_n\}$ as outcomes from M independent binomial distributions $Binomial(\text{size} = n_1, \text{probability} = p), Binomial(\text{size} = n_2, \text{probability} = p), ...Binomial(\text{size} = n_M, \text{probability} = p)$.

Question 1.1

Please give the likelihood function $L(k_1, k_2, ...k_n)$.

Your answer here:

$$
\begin{aligned}
L(p|k_1, k_2, .., k_m, n_1, n_2, ..., n_m) &= \prod_{i=1}^{m} f(p, n_i|k_i) \\
&= \prod_{i=1}^{m} \binom{n_i}{k_i} p^{k_i} (1-p)^{n_i - k_i}
\end{aligned}
$$

Question 1.2

Please give the log of the likelihood function as a sum of terms of the form $\log\left[\binom{n_i}{k_i} p^{k_i}(1-p)^{n_i-k_i}\right]$

Your answer here:

$$
\begin{aligned}
\log(L(p|k_1, k_2, .., k_m, n_1, n_2, ..., n_m)) &= \log\Big(\prod_{i=1}^{m} \binom{n_i}{k_i} p^{k_i}(1-p)^{n_i-k_i}\Big) \\
&= \sum_{i=1}^{m} \log\Big(\binom{n_i}{k_i} p^{k_i}(1-p)^{n_i-k_i}\Big) \\
&= \sum_{i=1}^{m} \Big[\log\binom{n_i}{k_i} + \log(p^{k_i}) + \log((1-p)^{n_i-k_i})\Big] \\
&= \sum_{i=1}^{m} \Big[\log\binom{n_i}{k_i} + k_i \log(p) + (n_i - k_i)\log(1-p)\Big]
\end{aligned}
$$

Question 2

Question 2.1

Please give the derivative with respect to p of $\sum_{i=1}^{M}\Big[\log\binom{n_i}{k_i} + k_i\log(p) + (n_i - k_i)\log(1-p)\Big]$.

Your answer here:

$$
\begin{aligned}
\frac{\partial \log(L)}{\partial p} &= \sum_{i=1}^{m}\Big[0 + k_i\frac{1}{p} + (n_i - k_i)\frac{1}{1-p}\Big] \\
&= 0 + \sum_{i=1}^{m} k_i\frac{1}{p} + \sum_{i=1}^{m}(n_i - k_i)\frac{1}{1-p} \\
&= \frac{1}{p}\sum_{i=1}^{m} k_i - \frac{1}{1-p}\sum_{i=1}^{m}(n_i - k_i)
\end{aligned}
$$

Question 2.2

Please give the value of p that maximizes $\sum_{i=1}^{M}\Big[\log\binom{n_i}{k_i} + k_i\log(p) + (n_i - k_i)\log(1-p)\Big]$.

Your answer here:

$$\begin{aligned}
\frac{1}{p}\sum_{i=1}^{m} k_i - \frac{1}{1-p}\sum_{i=1}^{m}(n_i - k_i) &= 0 \\
\frac{\sum_{i=1}^{m} k_i}{p} &= \frac{\sum_{i=1}^{m}(n_i - k_i)}{1-p} \\
\frac{\sum_{i=1}^{m} k_i}{p} &= \frac{\sum_{i=1}^{m}(n_i - k_i)}{1-p} \\
(1-p)\sum_{i=1}^{m} k_i &= p\sum_{i=1}^{m}(n_i - k_i) \\
\sum_{i=1}^{m} k_i - p\sum_{i=1}^{m} k_i &= p\sum_{i=1}^{m} n_i - p\sum_{i=1}^{m} k_i \\
\sum_{i=1}^{m} k_i &= p\sum_{i=1}^{m} n_i \\
\hat{p} &= \frac{\sum_{i=1}^{m} k_i}{\sum_{i=1}^{m} n_i}
\end{aligned}$$

Question 3

If the M samples $\{k_1, k_2, ...k_n\}$ from M independent binomial distributions $Binomial(\text{size} = n_1, \text{probability} = p), Binomial(\text{size} = n_2, \text{probability} = p), ...Binomial(\text{size} = n_M, \text{probability} = p)$ are viewed as $\sum_{i=1}^{M} k_i$ successes in $\sum_{i=1}^{M} n_i$ independent Bernoulli trials with probability of success equal to p, what is the maximum likelihood estimate of p.

Your answer here: As compared to problems 1-2, now we have only a single observation, so we look for the maximum likelihood estimate by maximizing the likelihood of p given one data point drawn from a binomial distribution with an observed $x = \sum_{i=1}^{n} k_i$ successes and $n = \sum_{i=1}^{m} n_i$ trials.

So the likelihood function is:

$$L\Big(p|\sum_{i=1}^{m} k_i, \sum_{i=1}^{m} n_i\Big) = \binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i} p^{\sum_{i=1}^{m} k_i}(1-p)^{\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i}$$

The log likelihood function is:

$$\begin{aligned}
\log\Big(L\Big(p|\sum_{i=1}^{m} k_i, \sum_{i=1}^{m} n_i\Big)\Big) &= \log\Big(\binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i} p^{\sum_{i=1}^{m} k_i}(1-p)^{\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i}\Big) \\
&= \log\binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i} + \log\Big(p^{\sum_{i=1}^{m} k_i}\Big) + \log\Big((1-p)^{\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i}\Big) \\
&= \log\binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} k_i} + \Big(\sum_{i=1}^{m} k_i\Big)\log(p) + \Big(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i\Big)\log(1-p)
\end{aligned}$$

Now we take the partial derivative of the log likelihood function with respect to p:

$$\begin{aligned}\frac{\partial \log(L)}{\partial p} &= 0 + \Big(\sum_{i=1}^{m} k_i\Big)\frac{1}{p} - \Big(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i\Big)\frac{1}{1-p} \\ &= \Big(\sum_{i=1}^{m} k_i\Big)\frac{1}{p} - \Big(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i\Big)\frac{1}{1-p}\end{aligned}$$

Now set the derivative equal to zero and solve for p:

$$\begin{aligned}\Big(\sum_{i=1}^{m} k_i\Big)\frac{1}{p} - \Big(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i\Big)\frac{1}{1-p} &= 0 \\ \frac{\sum_{i=1}^{m} k_i}{p} &= \frac{\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i}{1-p} \\ (1-p)\sum_{i=1}^{m} k_i &= p\Big(\sum_{i=1}^{m} n_i - \sum_{i=1}^{m} k_i\Big) \\ \sum_{i=1}^{m} k_i - p\sum_{i=1}^{m} k_i &= p\sum_{i=1}^{m} n_i - p\sum_{i=1}^{m} k_i \\ \sum_{i=1}^{m} k_i &= p\sum_{i=1}^{m} n_i \\ \hat{p} &= \frac{\sum_{i=1}^{m} k_i}{\sum_{i=1}^{m} n_i}\end{aligned}$$

So in the end the maximum likelihood estimate for $\hat{p}$ is the same when looking at maximizing the joint likelihood of m observations with a series of successes $k_1, k_2, ..., k_m$ and a series of number of trials $n_1, n_2, ...n_m$ versus thinking of that data as one observation represented by the sum of m random variables with successes $\sum_{i=1}^{m} k_i$ and number of trials $\sum_{i=1}^{m} n_i$.

Question 4

Context

The code below generates a sample, `samp1`, of size 10,000 from the $Binomial(\text{size} = 20, \text{probability} = 0.5)$ distribution and a sample, `samp2` of size 10,000 from the $Binomial(\text{size} = 50, \text{probability} = 0.3)$ distribution.

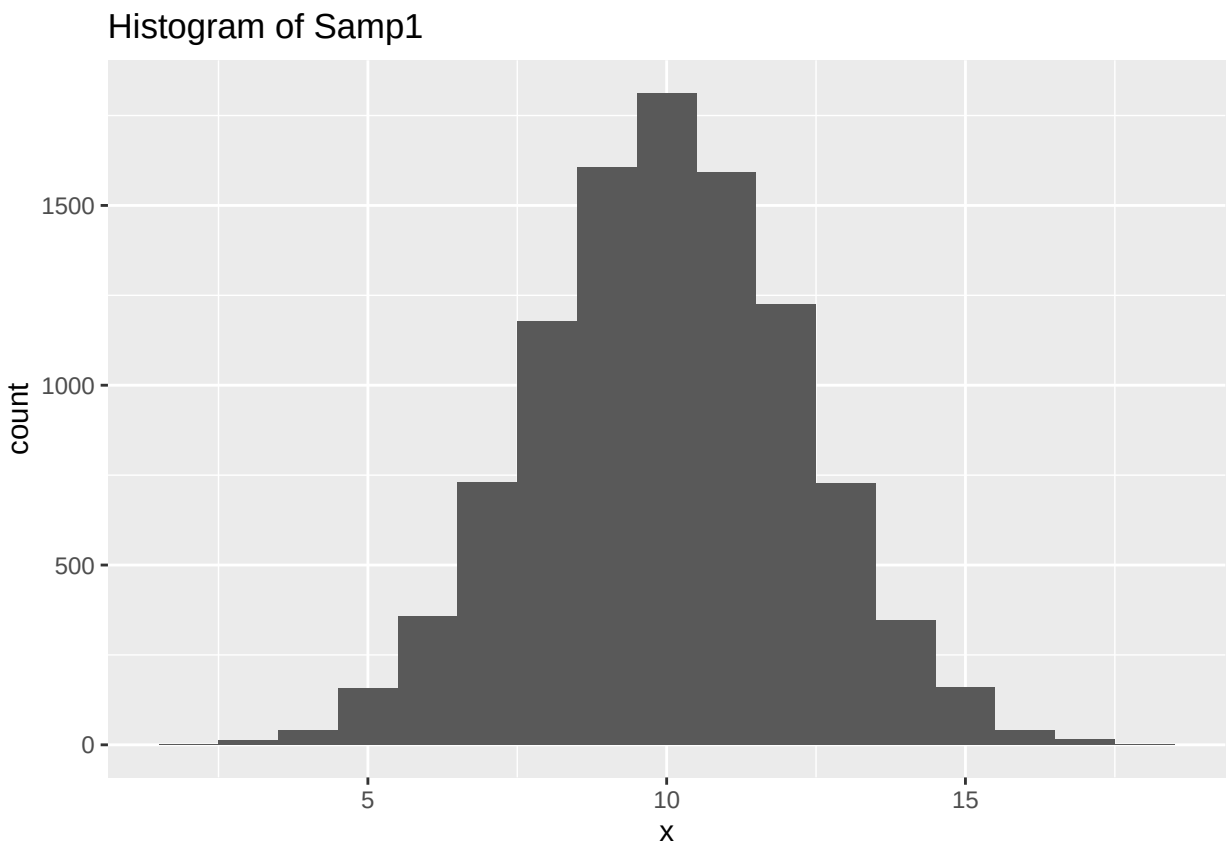
```
set.seed(12345)
samp1<-rbinom(10000,20,.5)
dat1<-data.frame(x=samp1)
samp2<-rbinom(10000,50,.25)
dat2<-data.frame(x=samp2)
```

Question 4.1

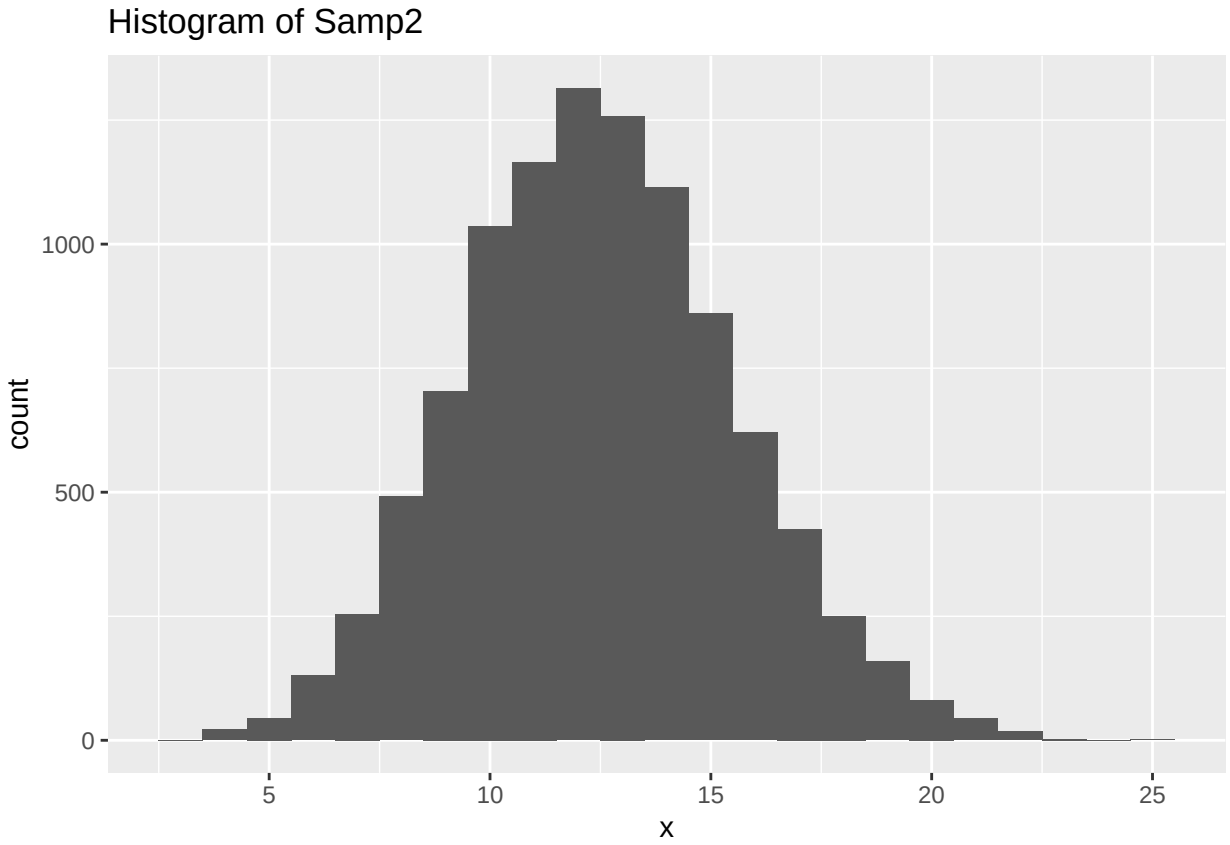
Please display separate histograms of `samp1` and `samp2` with binwidth equal to 1.

Your code and plots here:

```r
ggplot(dat1, aes(x = x)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of Samp1")
```




```r
ggplot(dat2, aes(x=x))+
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of Samp2")
```



Question 4.2

Treating `samp1` and `samp2` as samples from Normal distributions $Normal(\mu_1, \sigma_1^2)$ and $Normal(\mu_2, \sigma_2^2)$, please give maximum likelihood estimates of μ_1, σ_1^2, μ_2, and σ_2^2.

Your answer here:

Option 1: MLE estimators by hand

Likelihood function:

$$\begin{aligned} L(\mu_1, \sigma_1^2 | x_1, x_2, .., x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_1^2}} \cdot \exp\{-(\frac{x_i - \mu_1)^2}{2\sigma_1^2}\} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma_1^2}}\Big)^n \prod_{i=1}^{n} \exp\{-(\frac{x_i - \mu_1)^2}{2\sigma_1^2}\} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma_1^2}}\Big)^n \exp\{\sum_{i=1}^{n} -(\frac{x_i - \mu_1)^2}{2\sigma_1^2}\} \end{aligned}$$

Log likelihood function:

$$\begin{aligned}
\log(L) &= \log\left(\left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^n \exp\{\sum_{i=1}^{n} -\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\}\right) \\
&= \log\left(\left(\frac{1}{\sigma_1^2 2\pi}\right)^{n/2}\right) + \log\left(\exp\{\sum_{i=1}^{n} -\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\}\right) \\
&= \frac{n}{2}\log\left(\frac{1}{\sigma_1^2 2\pi}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu_1)^2}{2\sigma_1^2} \\
&= \frac{n}{2}\log(1) - \frac{n}{2}\log(\sigma_1^2 2\pi) - \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{2\sigma_1^2} \\
&= -\frac{n}{2}\log(\sigma_1^2 2\pi) - \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{2\sigma_1^2}
\end{aligned}$$

Finding μ_1:

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \mu_1} = 0 + 2\cdot\frac{\sum_{i=1}^{n}(x_i-\mu_1)}{2\sigma_1^2} &= 0 \\
\frac{\sum_{i=1}^{n}(x_i-\mu_1)}{\sigma_1^2} &= 0 \\
\sum_{i=1}^{n}(x_i-\mu_1) &= 0 \\
\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\mu_1 &= 0 \\
\sum_{i=1}^{n} x_i - n\mu_1 &= 0 \\
\hat{\mu_1} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Finding σ_1^2:

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \sigma_1} = -\frac{n}{2}\cdot 2\cdot\frac{\sigma_1 2\pi}{\sigma_1^2 2\pi} + 2\cdot\frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{2\sigma_1^3} &= 0 \\
-\frac{n}{\sigma_1} + \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{\sigma_1^3} &= 0 \\
n &= \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{\sigma_1^2} \\
\hat{\sigma_1^2} &= \frac{\sum_{i=1}^{n}(x_i-\mu_1)^2}{n}
\end{aligned}$$

μ_2 and σ_2^2 will follow the same pattern.

Now calculating values from the data using these MLE estimate formulas:

```
sq_diff <- function(x,mu){(x-mu)^2}

mu1 <- sum(samp1)/length(samp1)
mu1
```

```
## [1] 10.0043
```

```
sigma1_sq <- sum(sapply(samp1, sq_diff, mu=mu1))/length(samp1)
sigma1_sq
```

```
## [1] 4.940282
```

```
mu2 <- sum(samp2)/length(samp2)
mu2
```

```
## [1] 12.4697
```

```
sigma2_sq <- sum(sapply(samp2, sq_diff, mu=mu2))/length(samp1)
sigma2_sq
```

```
## [1] 9.414482
```

Option 2: MLE estimators numerically using nlm

```
neg.loglike <- function(theta, sample){
  result <- -sum(log(dnorm(sample, mean=theta[1], sd=theta[2])))
  return(result)
}

mle_samp1 <- nlm(neg.loglike, c(10,10), sample=samp1)
mu1 <- mle_samp1$estimate[1]
mu1
```

```
## [1] 10.0043
```

```
sigma1_sq <- (mle_samp1$estimate[2])^2
sigma1_sq
```

```
## [1] 4.940277
```

```
mle_samp2 <- nlm(neg.loglike, c(10,10), sample=samp2)
mu2 <- mle_samp2$estimate[1]
mu2
```

```
## [1] 12.46969
```

```
sigma2_sq <- (mle_samp2$estimate[2])^2
sigma2_sq
```

```
## [1] 9.414478
```

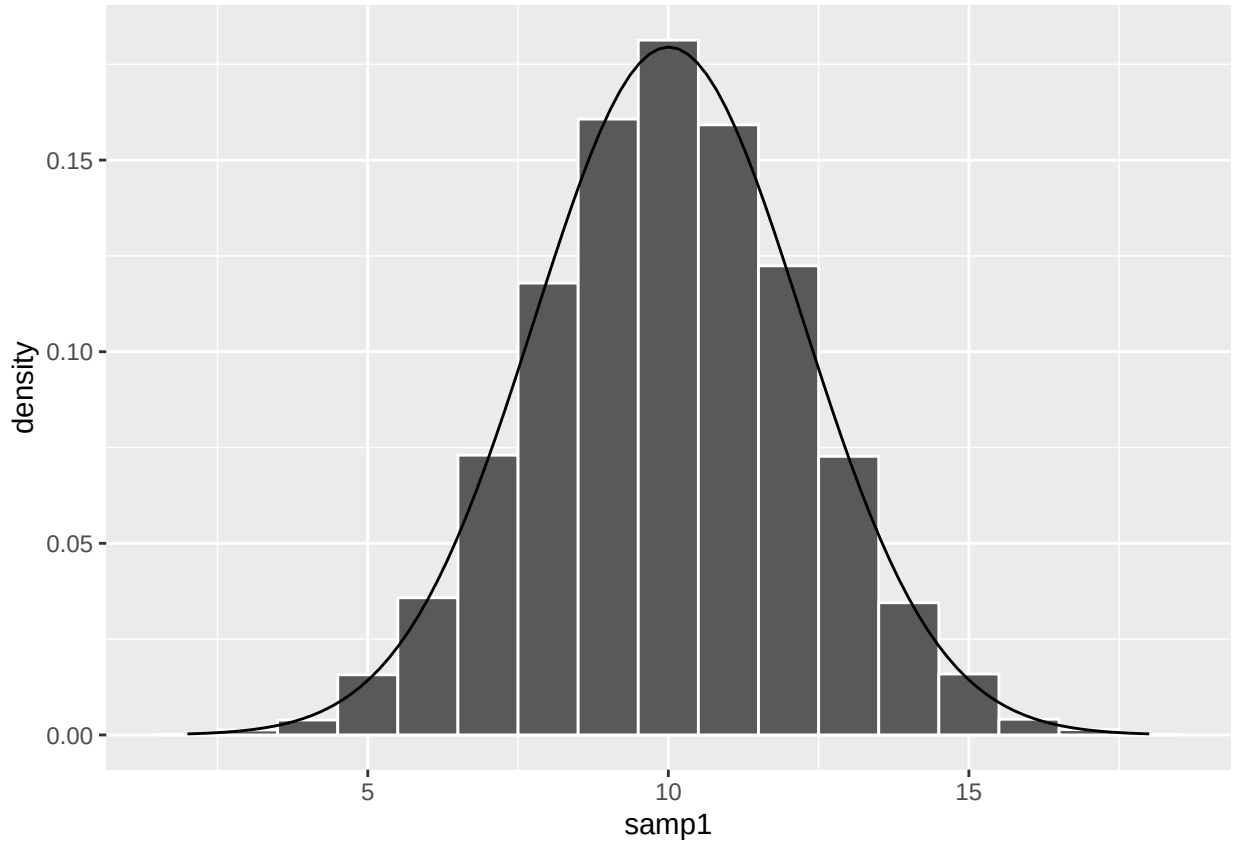
Question 4.3

The plotting methods from `continuous_probability_distributions_2_4_2.Rmd` and practice problem set 2 may be useful here.

For `samp1` please display the density histogram with density curve for $Normal(\mu_1, \sigma_1^2)$ superimposed.

Your code and plots here:

```
dat1 |>
  ggplot(aes(x = samp1)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, color = "white") +
  stat_function(fun = dnorm, args = list(mean = mu1, sd = sqrt(sigma1_sq)))
```

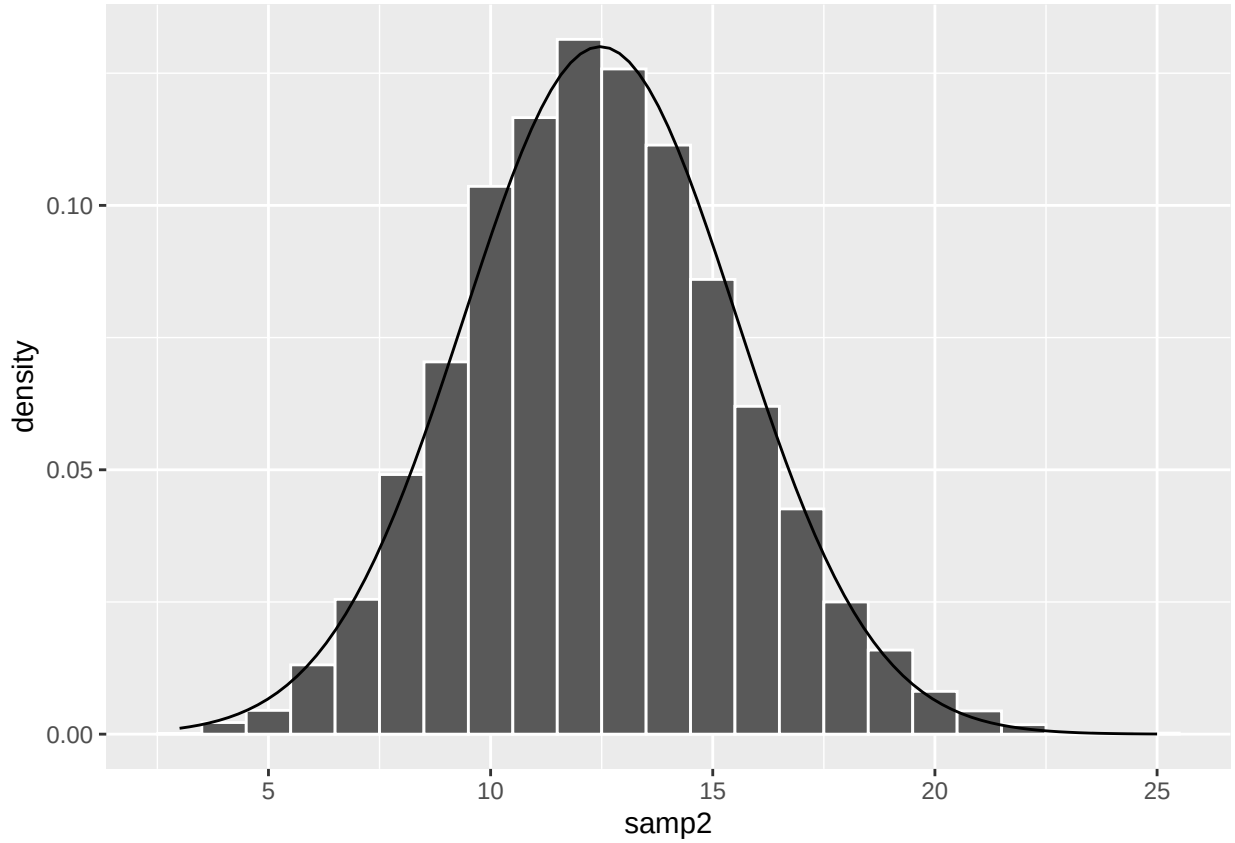


Question 4.4

For `samp2`: please display the density histogram with density curve for $Normal(\mu_2, \sigma_2^2)$ superimposed.

Your code and plots here:

```
dat2 |>
  ggplot(aes(x = samp2)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, color = "white") +
  stat_function(fun = dnorm, args = list(mean = mu2, sd = sqrt(sigma2_sq)))
```



Question 6

Let us consider $y_1, \ldots, y_n$ as samples from Normal distributions $Normal(\mu_1 = mx_1 + b, \sigma^2)$, ... , $Normal(\mu_n = mx_n + b, \sigma^2)$ where the input value x_i is a given constant, and the observations come in pairs (y_i, x_i). Based on the graph produced by the code chunk below answer these questions:

Question 6.1

Do you agree with the statement: "the line represents the fact that the peak of the normal distribution (the peak represents the average value) changes with each value of the input x_i and is given by the line $y_i = m \times x_i + b$." Why or why not? If not please write a revised statement.

Your answer here: Yes. The mean of the distribution generating the value of sample y_i will be centered at mean given by $mx_i + b$. When x_i changes, the peak of the distribution used to generate y_i will also change. And the peak of the distribution will fall at the points along the provided line on the graph.

Question 6.2

Do you think that every observation would lie exactly on the line? Why or why not? Would it be more likely that an observation would lie close to the line or far away from the line for a given (fixed) value of the input x_i?

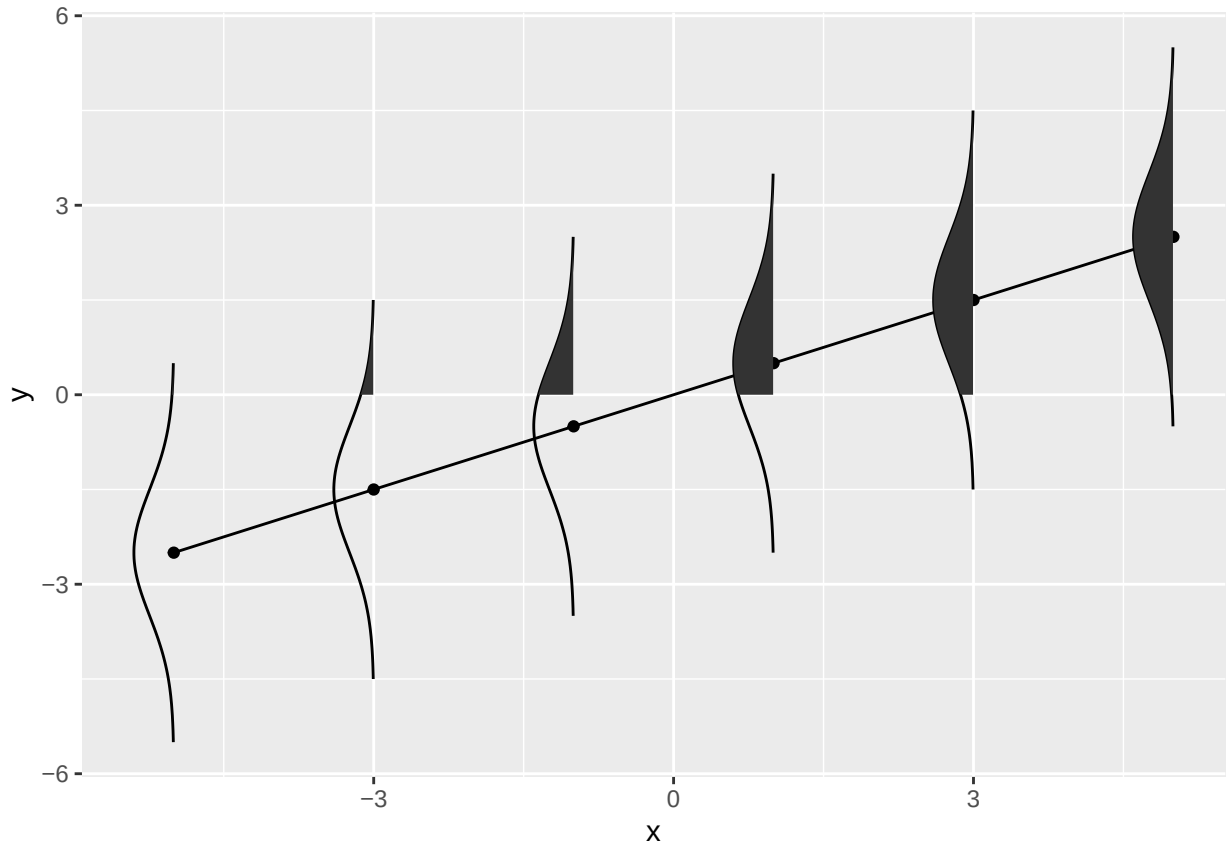
```r
x <- seq(1, 11, 2)
y <- x*0.5

x <- x - mean(x)
y <- y - mean(y)

df <- data.frame(x, y)

# For every row in `df`, compute a rotated normal density centered at `y` and shifted by `x`
curves <- lapply(seq_len(NROW(df)), function(i) {
  mu <- df$y[i]
  range <- mu + c(-3, 3)
  seq <- seq(range[1], range[2], length.out = 100)
  data.frame(
    x = -1 * dnorm(seq, mean = mu) + df$x[i],
    y = seq,
    grp = i
  )
})
# Combine above densities in one data.frame
curves <- do.call(rbind, curves)


ggplot(df, aes(x, y)) +
  geom_point() +
  geom_line() +
  # The path draws the curve
  geom_path(data = curves, aes(group = grp)) +
  # The polygon does the shading. We can use `oob_squish()` to set a range.
  geom_polygon(data = curves, aes(y = scales::oob_squish(y, c(0, Inf)),group = grp))
```



Your answer here: No, every point y_i will not lie directly along the line. The line represents the peak of the distribution from which a given point y_i is sampled - each y_i is sampled from a normal distribution with mean given by $mx_i + b$. However, these y_i are samples from a normal distribution with variance σ^2. That means they are random variables sampled from a probability distribution and could come from any point along the distribution with probability greater than 0. So unless the variance σ^2 was equal to zero, each individual point is going to deviate to some degree from the mean of the distribution it was sampled from. However, points are more likely to lie close to the line, since points towards the center of the distribution are more likely to occur than points at the extremes.

Question 7

Let us consider $y_1, \ldots, y_n$ as samples from Normal distributions $Normal(\mu_1 = mx_1 + b, \sigma^2), \ldots, Normal(\mu_n = mx_n + b, \sigma^2)$. Provide the work to show that the likelihood function $L(y_1, ..., y_n)$ is given by

$$(2\pi\sigma^2)^{-n/2} e^{\sum_{i=1}^{n} -\frac{(y_i-\mu_i)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} e^{\sum_{i=1}^{n} -\frac{(y_i-(mx_i+b))^2}{2\sigma^2}}$$

Your answer here: The probability density function for an individual point y_i would be given by

$$f(y_i; \mu_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\{-\frac{1}{(y_i - \mu_i)^2} 2\sigma^2\}$$

For simplification purposes we can also move around the square root in the first term to write this function as:

$$f(y_i; \mu_1, \sigma) = (2\pi\sigma^2)^{-1/2} \cdot \exp\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\}$$

Then the joint likelihood function is the product of the individual likelihood functions and simplify:

$$\begin{aligned} L(y_1, y_2, .., y_n) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \cdot \exp\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\} \\ &= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\{\sum_{i=1}^{n} -\frac{(y_i - \mu_i)^2}{2\sigma^2}\} \end{aligned}$$

Since $\mu_i = mx_i + b$, we can substitute in for μ_1 and write this as:

$$L(y_1, y_2, .., y_n) = (2\pi\sigma^2)^{-n/2} \exp\{\sum_{i=1}^{n} -\frac{(y_i - (mx_i + b))^2}{2\sigma^2}\}$$

Question 8

In the previous problem the mean is always changing for each observation y_i let us assume that the mean value $\mu_i = mx_i + b$ where x_i is an input variable. Substitute the formula $\mu_i = mx_i + b$ into the likelihood equation in 3.1.2 and calculate the partial derivatives $\frac{dL}{dm}$ and $\frac{dL}{db}$. Solve for m and b respectively.

Your answer here: First we take the log of the likelihood function to make it easier to find the MLE estimates for m and b

$$\begin{aligned} \log(L(y_1, y_2, ..., y_n)) &= \log\Big((2\pi\sigma^2)^{-n/2} \exp\{\sum_{i=1}^{n} -\frac{(y_i - (mx_i + b))^2}{2\sigma^2}\}\Big) \\ &= \log\Big((2\pi\sigma^2)^{-n/2}\Big) + \log\Big(\exp\{\sum_{i=1}^{n} -\frac{(y_i - (mx_i + b))^2}{2\sigma^2}\}\Big) \\ &= \Big(-\frac{n}{2}\Big)\log(2\pi\sigma^2) + \sum_{i=1}^{n} -\frac{(y_i - (mx_i + b))^2}{2\sigma^2} \\ &= \Big(-\frac{n}{2}\Big)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - b)^2 \\ &= \Big(-\frac{n}{2}\Big)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i b + x_i^2 m^2 - 2y_i x_i m + 2x_i bm + b^2) \\ &= \Big(-\frac{n}{2}\Big)\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + b\frac{\sum_{i=1}^{n} y_i}{\sigma^2} - m^2\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2} + m\frac{\sum_{i=1}^{n} y_i x_i}{\sigma^2} - bm\frac{\sum_{i=1}^{n} x_i}{\sigma^2} - b^2\frac{n}{2\sigma^2} \end{aligned}$$

First we take the derivative with respect to b and set it equal to 0 to find the MLE estimate for b:

$$\begin{aligned} \frac{\partial L}{\partial b} &= \frac{\sum_{i=1}^{n} y_i}{\sigma^2} - \frac{m\sum_{i=1}^{n} x_i}{\sigma^2} - 2b\frac{n}{\sigma^2} \\ nb &= \sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i \\ b &= \frac{\sum_{i=1}^{n} y_i}{n} - m\frac{\sum_{i=1}^{n} x_i}{n} \\ \hat{b} &= \bar{y} - m\bar{x} \end{aligned}$$

Noting that the average of all the y_i is given by $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and similiarly the average of all x_i is given by $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Next take the derivative with respect to m and set it equal to 0, then plug in the MLE estimate we found for b:

$$\begin{aligned}
\frac{\partial L}{\partial m} &= -2m\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\sum_{i=1}^n y_i x_i}{\sigma^2} - \frac{b\sum_{i=1}^n x_i}{\sigma^2} = 0 \\
-m\frac{\sum_{i=1}^n x_i^2}{\sigma^2} &+ \frac{\sum_{i=1}^n y_i x_i - b\sum_{i=1}^n x_i}{\sigma^2} = 0 \\
\frac{m\sum_{i=1}^n x_i^2}{\sigma^2} &= \frac{\sum_{i=1}^n y_i x_i - b\sum_{i=1}^n x_i}{\sigma^2} \\
m\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i - b\sum_{i=1}^n x_i \\
m\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i - (\bar{y} - m\bar{x})\sum_{i=1}^n x_i \\
m\sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i - \bar{y}\sum_{i=1}^n x_i + m\bar{x}\sum_{i=1}^n x_i \\
m\Big(\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i\Big) &= \sum_{i=1}^n y_i x_i - \bar{y}\sum_{i=1}^n x_i \\
m &= \frac{\sum_{i=1}^n y_i x_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} \\
m &= \frac{\sum_{i=1}^n y_i x_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} \cdot \frac{\frac{1}{n}}{\frac{1}{n}} \\
m &= \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\frac{\sum_{i=1}^n x_i}{n}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}\frac{\sum_{i=1}^n x_i}{n}} \\
\hat{m} &= \frac{\frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}
\end{aligned}$$

Question 9

From the previous problem we know the maximum likelihood estimate for the slope and intercept are given by:

$$\frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2}$$

and the intercept is

$$\bar{y} - m\bar{x}$$

Uncomment b and fill in the formula for the estimate of b. Explain why the points do not lie exactly on the (regression) line, the estimates are not exactly the same as the expected value slope of 2 and intercept of 1. Note that you should get the same estimates for the slope and intercept as provided by the built in lm() command.

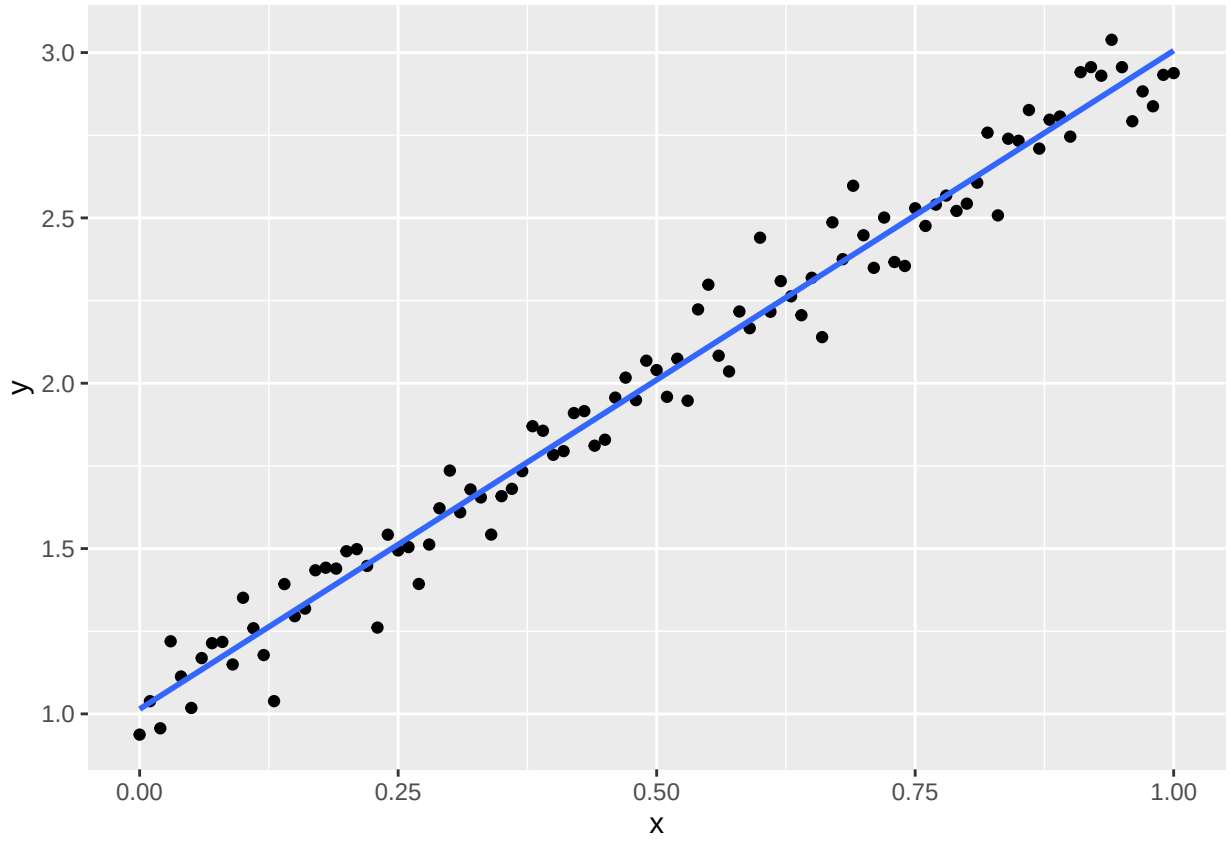
Your answer and code here:

```
set.seed(1)
x = seq(0,1,by=.01)
```

```r
y = 2*x+1+rnorm(length(x),0,.1)

ggplot(data=data.frame(x,y),aes(x=x,y=y)) + geom_point() + geom_smooth(method=lm, se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
m = (mean(y*x)-mean(y)*mean(x))/(mean(x^2)-mean(x)^2)
print(m)
```

```
## [1] 1.991375
```

```r
#Uncomment the two lines below and fill in the correct equation.
b = mean(y)-m*mean(x)
print(b)
```

```
## [1] 1.014479
```

```r
#Compare your answer with the built in R command
linear = lm(y~x)
linear$coefficients
```

```
## (Intercept)           x
##    1.014479    1.991375
```

These values do not lie exactly on the regression line because each point y_i is drawn from a normal distribution that only has mean given by $mx_i + b$ - and when you draw a random variable from a normal distribution you are not necessarily going to draw an observation from exactly the mean of that distribution. The MLE estimates for the mean and intercept of this distribution are not exactly at $m = 2$ and $b = 1$ as we would expect, since we have only sampled some points. If we had an infinite number of points, we would expect that the MLE estimates would be exactly right, but with any finite sample, you're going to get some amount of deviation due to random chance.

Question 10

Data and Background

This data set lists the individual observations for 934 children in 205 families in 1886 recorded by Galton. One question by Galton is concerning the relation between heights of parents and their offspring. It is from this investigation that Galton coined the phrase: "regression towards the mean."

Remark: According to Dominique Aubert-Marson (in the article Sir Francis Galton: the founder of eugenics), "Not only was Sir Francis Galton a famous geographer and statistician, he also invented "eugenics" in 1883. Eugenics, defined as the science of improving racial stock, was developed from a new heredity theory, conceived by Galton himself, and from the evolution theory of Charles Darwin, transposed to human society by Herbert Spencer. Galton's eugenics was a program to artificially produce a better human race through regulating marriage and thus procreation."

Question 10.1

In the code chunk below create a linear model with the input variable being 'midparentHeight' and 'childHeight' as the output variable. According to this model how does a one unit increase of the variable 'midparentHeight' affect 'childHeight'? What is the slope estimate and what would a slope of zero imply about the relationship between the two variables? Print the linear regression coefficients in the code chunk below as well.

Your answer and code here:

```
attach(GaltonFamilies)
head(GaltonFamilies)
```

```
##   family father mother midparentHeight children childNum gender childHeight
## 1    001   78.5   67.0           75.43        4        1   male        73.2
## 2    001   78.5   67.0           75.43        4        2 female        69.2
## 3    001   78.5   67.0           75.43        4        3 female        69.0
## 4    001   78.5   67.0           75.43        4        4 female        69.0
## 5    002   75.5   66.5           73.66        4        1   male        73.5
## 6    002   75.5   66.5           73.66        4        2   male        72.5
```

```
# ?lm  provides help page for lm function
model = lm(childHeight ~ midparentHeight, data = GaltonFamilies)
print(model)
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = GaltonFamilies)
```

```
##
## Coefficients:
##     (Intercept)  midparentHeight
##         22.6362           0.6374
```

```
summary(GaltonFamilies$midparentHeight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   64.40   68.14   69.25   69.21   70.14   75.43
```

A 1 inch increase in midparentHeight would be expected to increase childHeight by 0.6374 inches (the value given by the coefficient on midparentHeight in the linear model). A slope of 0 would imply no relationship between the parent's height and the child's height.

Question 10.2

In the code chunk below we calculate a confidence interval for the slope of the regression equation shown in the previous question. What does the confidence interval imply about the slope and relationship between the variables? Is it possible that there is a negative or no relationship between the variable 'midparentHeight' and 'childHeight' based on the interval?

Your answer and code here:

```
confint(model)
```

```
##                      2.5 %     97.5 %
## (Intercept)     14.2659135 31.0065676
## midparentHeight  0.5164552  0.7582666
```

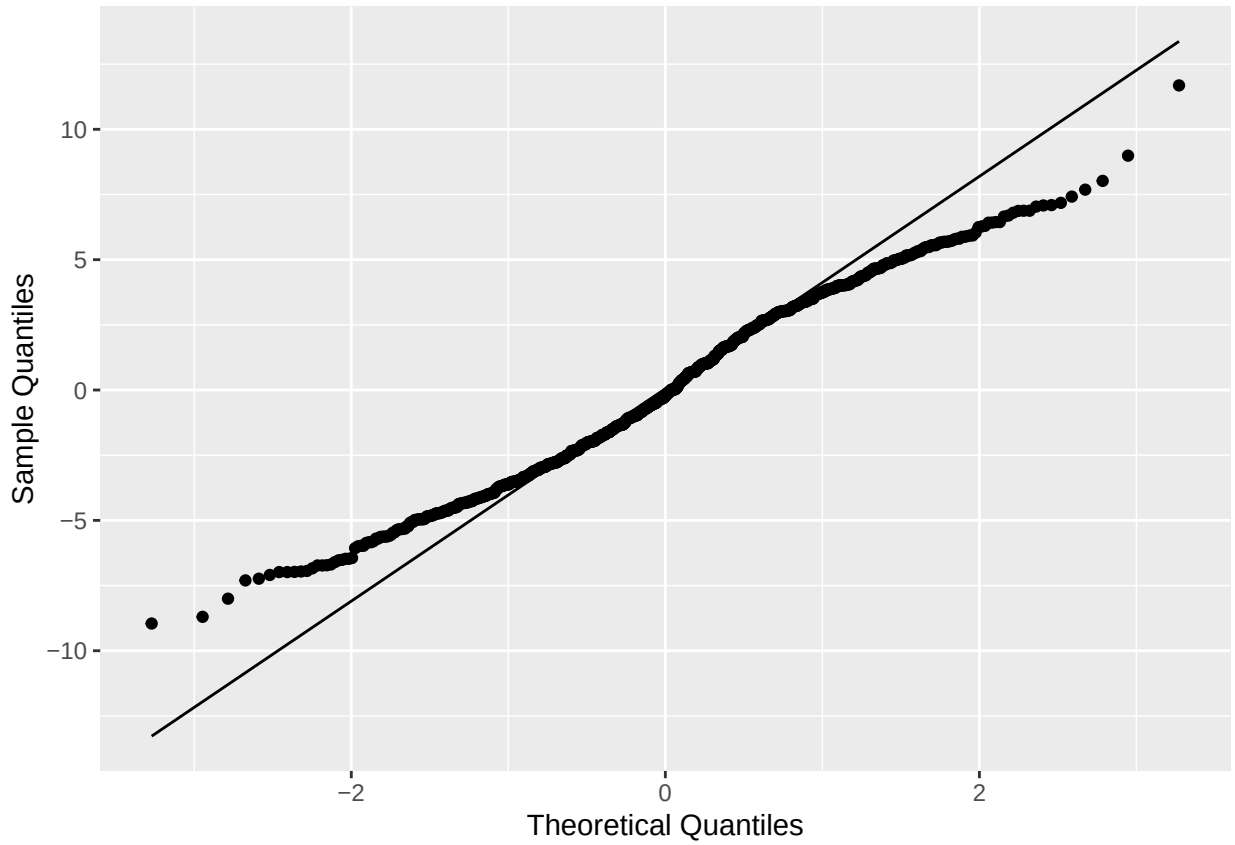
A 95% confidence interval for the coefficient on midparentHeight ranges from 0.52 to 0.76. At the 95% significance level, we can conclude that midparentHeight does have a positive relationship on childHeight. So we can be at least 95% confident that there is not a negative or null relationship between the variables midparentHeight and childHeight.

Question 10.3

One of the main assumptions of a linear regression model is that the error terms are normally distributed. The errors for a specified linear model and set of data is called the residuals. Assess the normality of the residuals by making a qqplot and performing the shapiro test on the residuals. Comment on if you believe the residuals are normal or not.

Your answer and code here:

```
#Make qqplot with qq line below
residuals <- data.frame(res = model$residuals)
residuals |>
  ggplot(aes(sample = res))+
  geom_qq() +
  geom_qq_line() +
  labs(x = "Theoretical Quantiles",
       y = "Sample Quantiles")
```




```
#Perform shapiro test below
#model$residuals
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.98714, p-value = 2.636e-07
```

The qqplot shows that the residuals do not appear to be normally distributed. The qqplot divides the residuals into quantiles with the smallest residuals on the left in the lowest quantiles and the largest residuals at the right of the graph in the largest quantiles. If these residuals were normally distributed, we would expect the quantiles with the smallest residuals to be plotted lower on the y-axis to lie on the diagonal line, which would mean observing more smaller values amongst our residuals. We would also expect to see the quantiles with the largest residuals plotted higher on the y-axis to also lie on the line, which would mean observing more larger values amongst our residuals.

We can verify what we observe on the graph through a Shapiro-Wilk test of the residuals. The null hypothesis for our Shapiro-Wilk test is that the data (the residuals) are being drawn from a normal distribution. Our p-value for this test is extremely small 2.6×10^{-7}, giving us extremely strong evidence in support of rejecting the null hypothesis and accepting the alternate hypothesis that the data was not drawn from a normal distribution. We have a very high degree of confidence that these residuals did not come from a normal distribution.

Not required for full credit, but some additional details on interpretation of the plot: The qqplot suggests that the residuals may be distributed a bit less like a bell curve and a bit more like a flatter/uniform

distribution. There may be fewer residuals at the extremes/tails of the distribution (very large or very small residuals) and so therefore a larger number of residuals with values across a range of numbers in the middle of the distribution. We can confirm this intuition from the qqplot by looking at a histogram of the residuals:

```
hist(model$residuals, breaks=30)
```