Probabilities

Suppose we are playing a simple collectable card game (e.g. like Hearthstone, or Magic the Gathering). In this game, each player has a card deck which contains 30 cards (with no duplicate cards). At the start of this game, both players shuffle their decks. Then the player going first draws five cards, and the player going second draws six cards. After this, the game starts, and players alternate turns.

Each player draws an additional card at the start of their turn. So, for example, after their third turn player one should have drawn eight cards in total (the 5 cards they started with, plus another three cards over three turns). Player two should have drawn nine cards in total after their third turn.

For the following questions, suppose there is a special combination of five cards, and if you have those five cards in your hand you instantly win the game.

Question 1.a

What is the probability that the first player will draw this combination on their first turn and win the game immediately? What about the second player?

The probability that the first player will draw this combination on their first turn and win the game immediately is as below:

The total number of combinations to draw out five cards in a deck of 30 cards is:

$$\binom{30}{5}$$

The chances for winning the first run where by the players draws all five victory cards is

$$\binom{5}{5} = 1$$

So:

$$Pr(Win) = \frac{1}{\binom{30}{5}}$$

$$\binom{30}{5} = \frac{30!}{5!(30! - 5!)} = \frac{30! \times 29! \times 28! \times 27! \times 26!}{5! \times 4! \times 3! \times 2! \times 1!} = 142506$$

Hence for player 1:

$$Pr(Win) = \frac{1}{142506}$$

The number of combinations for the first player to win the first turn is:

$$\frac{1}{142506}$$

The probability that the second player winning the game immediately is as below:

The total number of combinations to draw out six cards in a deck of 30 cards is:

$$\binom{30}{6}$$

The chances for winning the first run where by the players draws all five victory cards is as below where 5 cards is selected from the victory cards and one card is selected from the remaining 25 cards:

$$\binom{5}{5}\binom{25}{1} = 25$$

So:

$$Pr(Win) = \frac{25}{\binom{30}{6}}$$

$$\binom{30}{6} = \frac{30!}{6!(30! - 6!)} = \frac{30! \times 29! \times 28! \times 27! \times 26! \times 25!}{6! \times 5! \times 4! \times 3! \times 2! \times 1!} = 593775$$

Hence for player 1:

$$Pr(Win) = \frac{25}{593775}$$

The number of combinations for the first player to win the first turn is:

$$\frac{1}{23751}$$

Question 1.b

What is the probability that the five cards required for victory are all at the bottom of a player's deck (i.e. they are the last five cards in their deck)?

The total number of combinations to draw out five cards in a deck of 30 cards is:

$$\binom{30}{5}$$

The chances that the five cards required for victory are all at the botton of a player's deck is

$$\binom{25}{25} = 1$$

So:

$$Pr(Win) = \frac{1}{\binom{30}{5}}$$

$$\binom{30}{5} = \frac{30!}{5!(30! - 5!)} = \frac{30! \times 29! \times 28! \times 27! \times 26!}{5! \times 4! \times 3! \times 2! \times 1!} = 142506$$

Hence the probability is:

$$Pr(FiveCardsatBottomofplayer'sdeck) = \frac{1}{142506}$$

Question 1.c

Suppose a player has drawn 15 cards from their deck. What is the probability that all of the cards in the winning combination are still in their deck?

Please give your answer as a fraction

Given that a player has drawn 15 cards from their deck, there is 15 cards left in the deck. In total, there are 30 cards and 15 cards drawn out which can be written as:

$$\binom{30}{15} = \frac{30!}{15!(30! - 15!)} = 155117520$$

In a favourable outcome (where all 5 of the winning cards are still in the deck), there are 25 cards left with 15 cards to select. This can be written as:

$$\binom{25}{15} = \frac{25!}{15!(25! - 15!)} = 3268760$$

Hence, the probability that all of the cards in the winning combination are still in the deck:

$$= \frac{FavourableOutcome}{TotalSampleSpace}$$

$$= \frac{3268760}{155117520}$$

Question 1.d

Suppose a player has drawn n cards from their deck, where n is between 0 and 30. What is the probability that **all** of the cards in the winning combination is still in their deck (i.e. that they have not drawn any piece of the winning combination yet) **in terms of** n?

You might like to verify your previous answers with your formula, but this is not required.

The total sample space will be:

$$\binom{30}{n} = \frac{30!}{n!(30! - n!)}$$

The total favourable outcome (where all 5 of the winning cards are still in the deck), there are 25 cards left with n cards to select. This can be written as:

$$\binom{25}{n} = \frac{25!}{n!(25! - n!)}$$

Hence, the probability that all of the cards in the winning combination are still in the deck:

$$= \frac{FavourableOutcome}{TotalSampleSpace}$$

$$= \frac{\frac{25!}{(25!-n!)n!}}{\frac{30!}{(30!-n!)n!}}$$

$$= \frac{25!}{30!} \times \frac{(30! - n!)}{(25! - n!)}$$

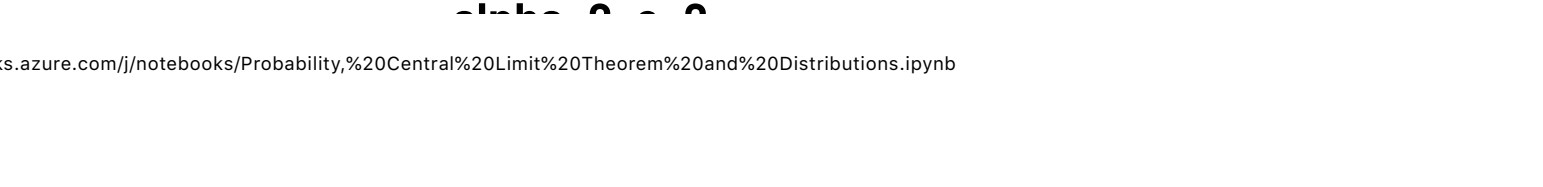
Question 2 - PDFs and Expectations

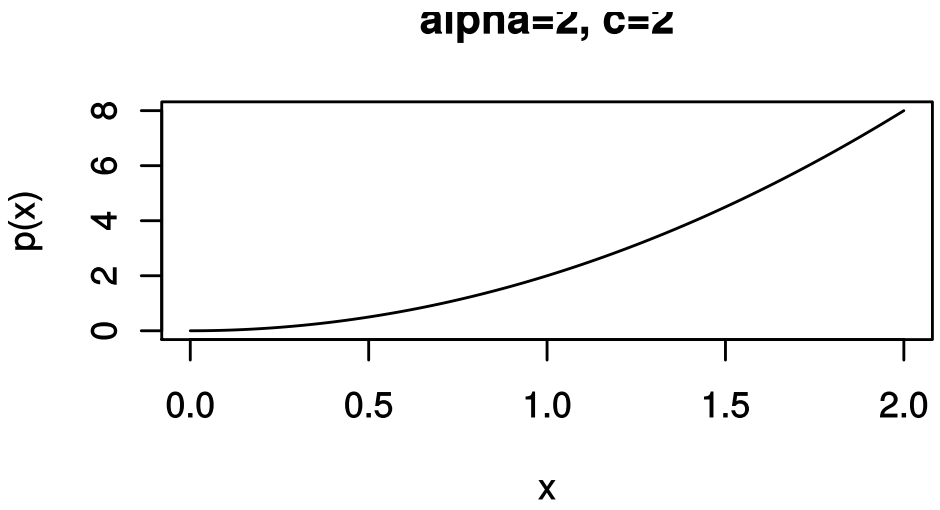
Suppose we have defined a probability density function for a random variable X as follows:

$$p(x) = \begin{cases} cx^2 & 0 \leq x \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

Notice that our PDF has two constants, c and α. α is a parameter, and c is a coefficient which we will carefully choose so the integral of cx^2 between 0 and α (with respect to x) is equal to 1.

In [3]:
```
library(repr)


alpha = 2
c = 2


x <- seq(0, alpha, 0.01)

# Reference: https://blog.revolutionanalytics.com/2015/09/resizing-plots-in-the-r-kernel-for-ju
options(repr.plot.width=5, repr.plot.height=3)

plot(
    x,
    c * x^2,
    "l",
    main=sprintf("alpha=%s, c=%s", alpha, c),
    xlab="x",
    ylab="p(x)"
)
```





Question 2.a

Suppose $\alpha = 1$. Find the value of c which would cause the integral of $p(x)$ from 0 to α with respect to x to be equal to 1. That is, find c such that

$$\int_0^1 cx^2 dx = 1$$

$$\left[\frac{1}{3}cx^3\right]_0^1 = 1$$

Substituting values x = 1 and 0:

$$\frac{1}{3}c(1)^3 - \frac{1}{3}c(0)^3 = 1$$

$$\frac{1}{3}c = 1$$

$$c = 3$$

Question 2.b

Find the value of c for a general value of α (you can do this in a way similar to how you answered question 2.a). That is, find c such that

$$\int_0^{\alpha} cx^2 dx = 1$$

$$F(\alpha) - F(0)$$

$$\left[\frac{1}{3}cx^3\right]_0^{\alpha} = 1$$

Subtituting values x = α and 0:

$$\frac{1}{3}c(\alpha)^3 - \frac{1}{3}c(0)^3 = 1$$

$$\frac{1}{3}c\alpha^3 = 1$$

$$c\alpha^3 = 3$$

$$c = \frac{3}{\alpha^3}, \alpha \neq 0$$

Question 2.c

Suppose c = 3 and α = 1. Find $E(X)$, the expected value of our variable X.

$$\begin{aligned} E(x) = \mu &= \int_0^1 xp(x) \\ &= \int_0^1 3x^3 dx \\ &= 3\int_0^1 x^3 dx \\ &= \left[\frac{3x^4}{4}\right]_0^1 \end{aligned}$$

Substituting values x = 0 and 1:

$$\begin{aligned} &= \left(\frac{3(1)^4}{4} - \frac{3(0)^4}{4}\right) \\ &= \frac{3}{4} \end{aligned}$$

Question 2.d

Suppose c = 3 and α = 1. Find $Var(X)$, the variance of our variable X.

$$\begin{aligned}\text{Var}(x) &= \int_0^1 (x-\mu)^2 p(x)dx \\ &= \int_0^1 3\left(x-\frac{3}{4}\right)^2 x^2 dx \\ &= 3\int x^4 dx - \frac{9}{2}\int x^3 dx + \frac{27}{16}\int x^2 dx\end{aligned}$$

Substituting values x = 0 and 1:

$$\begin{aligned}&= \left[\frac{3x^5}{5} - \frac{9x^4}{8} + \frac{9x^3}{16}\right]_0^1 \\ &= \left[3x^3\left(\frac{16x^2-30x+15}{80}\right)\right]_0^1 \\ &= 3(1)^3\left(\frac{16(1)^2-30(1)+15}{80}\right) \\ &= \frac{3}{80}\end{aligned}$$

Question 3 - Distributions

Suppose we are given the following information:

- You are modelling the number of people visiting a particular doctor's office within a day, with the hope of identifying a disease outbreak in the local area of the doctor
- It is known that, **on an average day**, 30 patients will see this doctor, with a **standard deviation** of 3 patients per day

Question 3.a

Describe a model you might use to model the number of patients on a given day (there might be more than one choice, so pick one and justify it). Also give the parameters of this model based on the given information.

YOUR ANSWER HERE

A Gaussian (normal) distribution is a suitable model to use as it is usually used to represent real-valued random variables. Given that it has a standard deviation of 3 which is relatively high, this indicates that the data is spread out over a large range of values. The values will converge into a normal distribution as the degree of skewness to the right decreases as the rate of occurence decreases. Therefore, we can say it can be modelled by the normal distribution.

The parameters of this model would be as below if X number of patients on a given day is normally distributed:

$$X \sim N(\mu, \sigma^2)$$

$$X \sim N(30, 9)$$

Question 3.b

On one particular day, 45 patients visit the doctor. Considering the model you developed in your answer to the previous question, do you think that this number of patients in a given day is cause for alarm? Use calculations to back up your answer by determining the probability of seeing 45 or more patients in a given day.

Using Gaussian(normal) distribution, the distribution can be converted to standard normal.We need to firstly calculate the z score to find the probability of seeing 45 of more patients in a given day.

$$Z = \frac{x - \mu}{\sigma}$$

$$= \frac{45 - 30}{3}$$

$$= 5$$

$$P(X > 45) = 1 - P(Z < 5)$$

$$= 1 - 0.999999713348428$$

$$= 2.86651572 \times 10^{-}7$$

Given that the probability of obtaining more than 45 patients is low, hence it is alarming to have exactly 45 patients visit the doctor on one particular day

Question 4 - Maximum Likelihood Estimation of Parameters

Suppose we are developing a new plant treatment which will (hopefully) improve crop yields. We have a dataset which contains weights for two candidate treatments, as well as a control group (which receives neither of the candidate treatments).

For this question, we will use a dataset which is built into R - the PlantGrowth dataset.

This dataset was originally released alongside the following paper:

Dobson, A. J. (1983) An Introduction to Statistical Modelling. London: Chapman and Hall.

In [4]:
```
#Cell is run to load the datasets!

# We are splitting the dataset into three groups, and we will find statistics for each group
control.dataset <- PlantGrowth[PlantGrowth$group == "ctrl","weight"]
treatment1.dataset <- PlantGrowth[PlantGrowth$group == "trt1","weight"]
treatment2.dataset <- PlantGrowth[PlantGrowth$group == "trt2","weight"]

control.dataset
treatment1.dataset
treatment2.dataset
```

4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14

4.81 4.17 4.41 3.59 5.87 3.83 6.03 4.89 4.32 4.69

6.31 5.12 5.54 5.5 5.37 5.29 4.92 6.15 5.8 5.26

Question 4.a

Suppose we want to create models for the weight of each group. You think a normal distribution would be suitable for this purpose, but a colleague has suggested that you should use a binomial distribution instead. Someone else proposed using a uniform distribution instead.

For both the binomial and uniform distributions, explain whether they would be a good choice (justifying your answer).

Also justify why using the normal distribution is a good choice here.

Answer

Binomial Distribution:

Binomial distribution is a discrete probability distribution used when there are only two type of events which is success or failure. Success and failure are mutually exclusive and cannot occur the same time. The distribution assumes a finite number of trials, n. For this case, to create models for the weight of each group, binomial distribution would not be suitable as weight is a continuous variable which means that any two possible values of the variable are an infinite number of other possible values. Therefore, it is not possible to create a model with binomial distribution as the model requires discrete variables and finite number of trials.

Uniform Distribution:

Uniform distribution is when every outcome is likely to occur. Given that weight is a continuous variable and it is never uniform, the uniform distribution is not suitable for modelling weight.

Normal Distribution:

Normal Distribution (Gaussian) is a continuous probability distribution. This is suitable for creating models for the weight of each group as it takes on uncountably infinite number of possible values.

Question 4.b

Suppose, rather than modelling the weights directly, we instead want to model the probability that a plant will grow to weigh over 6 units of weight, for each of the three treatments we are testing (treatments 1 and 2, and the control). Suggest a model that would be suitable for this purpose, and justify your choice.

Answer

A binomial distribution would be suitable for this case as we would like to measure two types of events which is success or failure. Success is when the plant grows to weigh over 6 units of weight where as failure would be vice versa. Each trial is independent with a total number of n identical trials conducted. The probability of success and failure is the same for all trials as trials are identical. In addition, it has a series of bernoulli trials of 10 in each of the 3 datasets(treatments 1 and 2, and the control) which is a finite number of trials.

Question 4.c

After considering our answers to questions 4.a and 4.b, we have decided to model the weights directly (i.e. we will use the model discussed in question 4.a, not 4.b). To do this, we will create three models: one for each of the three groups. We will use normal distributions to model each of the three groups, and then compare the estimated means of each group.

We now have to decide how we will calculate our estimates of the mean (μ) and standard deviation (σ) of each of our datasets. One approach is to use the maximum likelihood method, where we wish to maximize the likelihood of the data given the parameters μ and σ (that is, we wish to find the values of μ and σ which cause $P(x|\mu,\sigma)$ to be maximized). Note that maximizing something is the same as maximizing the log of that thing, because log (for any base > 1) is "monotonically increasing"- that is, if $a > b, \log(a) > \log(b)$. We're actually going to maximize the log-likelihood below.

A colleague of yours seems to think that maximizing the log-likelihood is the same as minimizing the mean absolute error. Another colleague disagrees, saying that they are misremembering and the likelihood is the same as minimizing the mean squared error. Yet another colleague seems to believe that we minimize the negative log-likelihood by minimizing the log-cosh loss (since they both have the word "log" in them; you are not convinced by this argument).

We are going to experiment with this using code, in the hope of finding the truth. We are going to use R's optimize() function to do this (look in R's documentation if you'd like to learn more about how optimize() works). The code using the optimize() function has already been provided, and the log_likelihood_fn is defined two cells down, right under the cell where you will

write your code for this question. **Note that we are using the built-in R estimate for the standard deviation, and are only estimating the mean using our four different methods.**

You will need to write functions to calculate the mean squared error (mse_fn), mean absolute error (mae_fn), and logcosh error (logcosh_fn). The code two cells down then uses R's optimize() function to calculate the value of μ which maximizes (for log-likelihood) or minimizes (for mse, mae and logcosh) the error function being optimized. This code has been provided for you; you need to write three functions in the cell below. Each function takes three arguments:

- mu, the value of the mean we are using to calculate our error
- sigma, the value of the standard deviation we are using to calculate our error
- x, a vector containing the data we are calculating the error on

The three functions are:

mse_fn
Returns $\frac{1}{n}\sum_{p\in x}(p-\mu)^2$, i.e. the mean squared error

mae_fn
Returns $\frac{1}{n}\sum_{p\in x}|p-\mu|$, i.e. the mean absolute error

logcosh_fn
Returns $\frac{1}{n}\sum_{p\in x}\log(\cosh(p-\mu))$, i.e. the mean logcosh error

For log_likelihood_fn, we used the following formula for a normal distribution (if you'd like to see a proof, you can do so at https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood) (https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood)):

$$L(x|\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{p\in x}(p-\mu)^2$$

Note that the notation $\sum_{p \in x}$ indicates that we are calculating the sum of the value inside the summation for every element p in the set x (i.e. we are summing together the calculated values for each of the elements in our dataset). E.g. if $x = (1, 2, 3)$, $\sum_{p \in x}(x-1)^2 = 0^2 + 1^2 + 2^2 = 5$

Once you've finished your functions, you should run the cell underneath them (the one containing the definition for log_likelihood_fn). This will print a table containing the estimates for the mean of each of our three datasets, using each of our four methods.

In [5]:
```
# creating function for mse function
mse_fn <- function (mu, sigma, x) {
    n <- length(x)
    term1 <- sum((x - mu)^2) / (n)
    return (term1)
}

#creating function for mae function
mae_fn <- function (mu, sigma, x) {
    n <- length(x)
    term1 <- sum(abs(x - mu)) / (n)
    return (term1)
}

# creating function for logcosh function
logcosh_fn <- function (mu, sigma, x) {
    n <- length(x)
    term1 <- sum(log(cosh(x - mu))) / (n)
    return (term1)
}
```

In [6]:
```
log_likelihood_fn <- function (mu, sigma, x) {
    # You need to fill this function in!
```

```
    # You need to fill this function in.
    n <- length(x)
    term1 <- -n * log(2 * pi) / 2
    term2 <- -n * log(sigma * sigma) / 2
    term3 <- -sum((x - mu)^2) / (2 * sigma * sigma)

    return (term1 + term2 + term3)
}

label <- c(
    'control',
    'treatment1',
    'treatment2'
)

# Maximize the log likelihood function
likelihood.control.mu <- optimize(f=log_likelihood_fn, interval=seq(-10, 10, 0.01),
                                  sigma=sd(control.dataset),
                                  x=control.dataset, maximum=TRUE)$maximum
likelihood.treatment1.mu <- optimize(f=log_likelihood_fn, interval=seq(-10, 10, 0.01),
                                     sigma=sd(treatment1.dataset), x=treatment1.dataset,
                                     maximum=TRUE)$maximum
likelihood.treatment2.mu <- optimize(f=log_likelihood_fn, interval=seq(-10, 10, 0.01),
                                     sigma=sd(treatment2.dataset), x=treatment2.dataset,
                                     maximum=TRUE)$maximum

likelihood <- c(
    likelihood.control.mu,
    likelihood.treatment1.mu,
    likelihood.treatment2.mu
)

# Maximize the mean squared error function
mse.control.mu <- optimize(f=mse_fn, interval=seq(-10, 10, 0.01),
                           sigma=sd(control.dataset),
                           x=control.dataset, maximum=FALSE)$minimum
mse.treatment1.mu <- optimize(f=mse_fn, interval=seq(-10, 10, 0.01)
```

```
mse.treatment1.mu <- optimize(f=mse_fn, interval=seq(-10, 10, 0.01),
                              sigma=sd(treatment1.dataset),
                              x=treatment1.dataset, maximum=FALSE)$minimum
mse.treatment2.mu <- optimize(f=mse_fn, interval=seq(-10, 10, 0.01),
                              sigma=sd(treatment2.dataset),
                              x=treatment2.dataset, maximum=FALSE)$minimum

mse <- c(
    mse.control.mu,
    mse.treatment1.mu,
    mse.treatment2.mu
)

# Maximize the mean absolute error function
mae.control.mu <- optimize(f=mae_fn, interval=seq(-10, 10, 0.01),
                           sigma=sd(control.dataset),
                           x=control.dataset, maximum=FALSE)$minimum
mae.treatment1.mu <- optimize(f=mae_fn, interval=seq(-10, 10, 0.01),
                              sigma=sd(treatment1.dataset),
                              x=treatment1.dataset, maximum=FALSE)$minimum
mae.treatment2.mu <- optimize(f=mae_fn, interval=seq(-10, 10, 0.01),
                              sigma=sd(treatment2.dataset),
                              x=treatment2.dataset, maximum=FALSE)$minimum

mae <- c(
    mae.control.mu,
    mae.treatment1.mu,
    mae.treatment2.mu
)


# Maximize the logcosh error function
logcosh.control.mu <- optimize(f=logcosh_fn, interval=seq(-10, 10, 0.01),
                               sigma=sd(control.dataset),
                               x=control.dataset, maximum=FALSE)$minimum
logcosh.treatment1.mu <- optimize(f=logcosh_fn, interval=seq(-10, 10, 0.01),
                                  sigma=sd(treatment1.dataset),
```

```
                                    sigma=sd(treatment1.dataset),
                                    x=treatment1.dataset, maximum=FALSE)$minimum
logcosh.treatment2.mu <- optimize(f=logcosh_fn, interval=seq(-10, 10, 0.01),
                                    sigma=sd(treatment2.dataset),
                                    x=treatment2.dataset, maximum=FALSE)$minimum

logcosh <- c(
    logcosh.control.mu,
    logcosh.treatment1.mu,
    logcosh.treatment2.mu
)

df <- data.frame(label, likelihood, mae, mse, logcosh)
df
```

A data.frame: 3 x 5

label	likelihood	mae	mse	logcosh
<fct>	<dbl>	<dbl>	<dbl>	<dbl>
control	5.032	5.148585	5.032	5.022639
treatment1	4.661	4.527325	4.661	4.606462
treatment2	5.526	5.389061	5.526	5.513414

Question 4.d

One of your colleagues in Question 4.c is correct; which one does it appear to be based on our calculations in the previous question? Prove this colleague correct using algebra (you only have to prove them correct; you don't have to disprove the other two).

Note: Although this colleague is correct in this particular case, what they suggest is not always the case for every possible model; but it is true for the Normal distribution, which we are using here.

ANSWER HERE

The colleague which stated that maximising log likelihood is the same as minimizing mean squared error is correct. By doing so, the formula will be derived into a sample mean. In order to proof this, we can do partial derivative of log likelihood equation with respect to μ . The only term we are differentiating is the last term of the MLE equation which is the same as the MSE equation if constant is taken out.

$$L(x|\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{p\in x}(p-\mu)^2$$

$$\frac{dL(x|\mu,\sigma)}{d\mu} = -0 - 0 - \frac{1}{2\sigma^2}\frac{dL(x|\mu,\sigma)}{d\mu}\sum_{p\in x}(p-\mu)^2$$

$$\frac{dL(x|\mu,\sigma)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{p\in x}-2(p-\mu)$$

$$= \frac{1}{\sigma^2}\sum_{p=x}(p-\mu)$$

We can then set it to 0 to find the μ once we have the partial derivative

$$\frac{dL(x|\mu,\sigma)}{d\mu} = 0$$
$$\frac{1}{\sigma^2}\sum_{p\in x}(p-\mu) = 0$$
$$\sum_{p\in x} p - n\mu = 0$$
$$\sum_{p\in x} p = n\mu$$
$$\frac{1}{n}\sum_{p\in x} p = \mu$$

Question 4.e

Given your maximum likelihood estimates for the mean of each population (and keeping in mind that we have a very small number of samples for each group), which treatment appears to work best?

Note: In later weeks we will learn how to use confidence intervals and hypothesis testing to more rigorously analyse the differences between the different groups, but for now we will just rely on the point estimates (which is not ideal and can lead to misleading results, but we will do this for now).

YOUR ANSWER HERE

Treatment 2 seems to be the most effective given that the maximum likelihood estimates for the mean of each of the population is the highest mean weight of the crops in this case.

Question 5 - Central Limit Theorem

Suppose our company is trialling a new production method for phone cases, based on 3D printing. 3D printing can a volatile process, and the company has decided to accept the fact that there will be a certain proportion of failures out of the total number of 3D prints.

However, before committing to the new process, management would like to estimate the probability of failure by printing a number of phone cases. They have asked you how many cases they should print to ensure they have a reasonably good idea of the probability of failure.

The engineers developing the new production method assure management that the probability of failure is somewhere between 1% and 20%, but they are unwilling to make any guarantees beyond this without testing the method first.

Question 5.a

We will model this problem with a binomial distribution. Justify why the binomial distribution is a good choice for this problem.

A binomial distribution would be suitable for this case as we would like to measure success or failure as the company would like to find out if there will be a certain proportion of failures of the total number of 3D prints. In addition, each trial is independent with a total number of n identical trials conducted. The probability of success and failure is the same for all trials as trials are identical. This fulfills the criteria for binomial distribution.

Question 5.b

Suppose that we are considering three potential failure probabilities:

- $\theta = 0.01$
- $\theta = 0.05$
- $\theta = 0.2$

We also are considering three potential sizes for our test production run (i.e. the number of phone cases we will print in our test run):

- $n = 50$
- $n = 200$
- $n = 800$

For each combination of failure probability and number of cases printed, calculate the limiting distribution for the sample mean.

From central limit theorem, the sample mean is approximately distributed as $\frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately distributed as $N(\mu, \frac{1}{n}\sigma^2)$ for large n with a rule of thumb n > 30. Given that it is a binomial distribution, success and failures need to be defined. In this example we will try to approximate binomial to normal distribution given that most distribution with large sample size will converge to normal distribution. As failure rate is measured, the success rate for this case would be the number of failures with probability of success as θ

Mean and variance in binomial distribution are measured with $\mu = n\theta$ and $\sigma^2 = n\theta(1-\theta)$ with n sample size. We can calculate μ and σ^2 for each combination of potential failure probabilities and sample size.

At θ = 0.01 and all the sample size (n = 50, n = 200, n = 800), we have:

$$\mu = 50(0.01) = 0.5$$

$$\mu = 200(0.01) = 2$$

$$\mu = 800(0.01) = 8$$

$$\frac{\sigma^2}{n} = \frac{50(0.01)(1-0.01)}{50} = 0.0099$$

$$\frac{\sigma^2}{n} = \frac{200(0.01)(1-0.01)}{200} = 0.0099$$

$$\frac{\sigma^2}{n} = \frac{800(0.01)(1-0.01)}{800} = 0.0099$$

At θ = 0.05 and all the sample size (n = 50, n = 200, n = 800), we have:

$$\mu = 50(0.05) = 2.5$$

$$\mu = 200(0.05) = 10$$

$$\mu = 800(0.05) = 40$$

$$\frac{\sigma^2}{n} = \frac{50(0.05)(1-0.01)}{50} = 0.475$$

$$\frac{\sigma^2}{n} = \frac{200(0.05)(1-0.01)}{200} = 0.475$$

$$\frac{\sigma^2}{n} = \frac{800(0.05)(1-0.01)}{800} = 0.475$$

At θ = 0.2 and all the sample size (n = 50, n = 200, n = 800), we have:

$$\mu = 50(0.2) = 10$$

$$\mu = 200(0.2) = 40$$

$$\mu = 800(0.2) = 160$$

$$\frac{\sigma^2}{n} = \frac{50(0.2)(1-0.01)}{50} = 0.16$$

$$\frac{\sigma^2}{n} = \frac{200(0.2)(1-0.01)}{200} = 0.16$$

$$\frac{\sigma^2}{n} = \frac{800(0.2)(1-0.01)}{800} = 0.16$$

Hence, after calculating the μ and σ^2 we have 9 limitting distribution for the sample mean as follow:

$$\overline{X} \sim N(0.5, 0.0099)$$
$$\overline{X} \sim N(2, 0.0099)$$
$$\overline{X} \sim N(8, 0.0099)$$
$$\overline{X} \sim N(2.5, 0.475)$$
$$\overline{X} \sim N(10, 0.475)$$
$$X \sim N(40, 0.475)$$
$$\overline{X} \sim N(10, 0.16)$$
$$\overline{X} \sim N(40, 0.16)$$
$$X \sim N(160, 0.16)$$

Question 5.c

Verify the results obtained

In [8]:
```
theta <- c(0.01,0.05,0.2)
n <- c(50,200,800)
# calculating mu
mu <- sapply(n, function(x) x*theta)
mu

# calculating variance
variance <- sapply(n, function(x) theta*(1 - theta))
variance
```

A matrix: 3 x 3 of type dbl

0.5	2	8
2.5	10	40
10.0	40	160

A matrix: 3 x 3 of type dbl

0.0099	0.0099	0.0099
0.0475	0.0475	0.0475
0.1600	0.1600	0.1600

Question 5.d

For each of the sample sizes and potential failure probabilities listed above, we now know the theoretical distribution by the Central Limit Theorem (we calculated this in Questions 6.b and 6.c). However, management is still not convinced and have asked us to develop a simulation which will experimentally demonstrate our calculations were correct.

Write a function below, called simulate, which returns an R list containing the point estimates of the mean and standard deviation of the limiting distribution for the mean from 50,000 simulations. This function takes two parameters (n and theta) and returns the maximum likelihood estimate for the mean and standard deviation of the limiting distribution based purely on the generated sample, in the form of an R list.

In [9]:
```
simulate <- function(n, theta) {
    Muvalues = rbinom(50000,n,theta)
    Sigmavalues = c(mean(Muvalues),sd(Muvalues))
}
```

In [10]:
```
for (theta in c(0.01, 0.05, 0.2)) {
    for (n in c(50, 200, 800)) {
        print(sprintf("theta = %s, n = %s", theta, n))
        print(simulate(n, theta))
    }
}
```

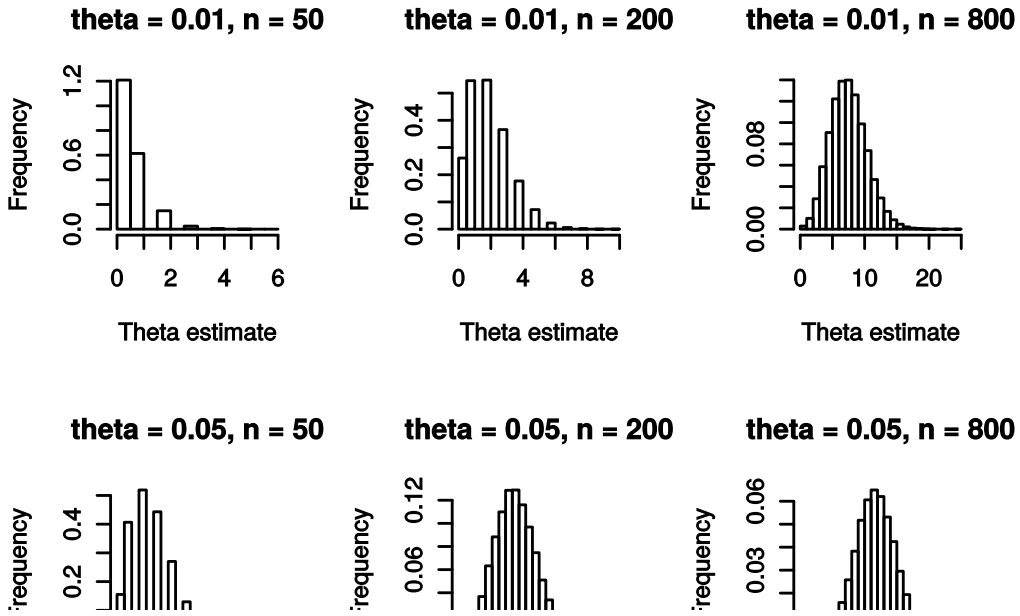
```
[1] "theta = 0.01, n = 50"
[1] 0.4979200 0.7012314
[1] "theta = 0.01, n = 200"
[1] 2.001900 1.404185
[1] "theta = 0.01, n = 800"
[1] 7.99862 2.81212
[1] "theta = 0.05, n = 50"
[1] 2.504960 1.542499
[1] "theta = 0.05, n = 200"
[1] 10.008360  3.084108
[1] "theta = 0.05, n = 800"
[1] 40.008000  6.158387
[1] "theta = 0.2, n = 50"
[1] 10.025660  2.832843
[1] "theta = 0.2, n = 200"
[1] 40.001440  5.665873
[1] "theta = 0.2, n = 800"
[1] 159.9393  11.2951
```

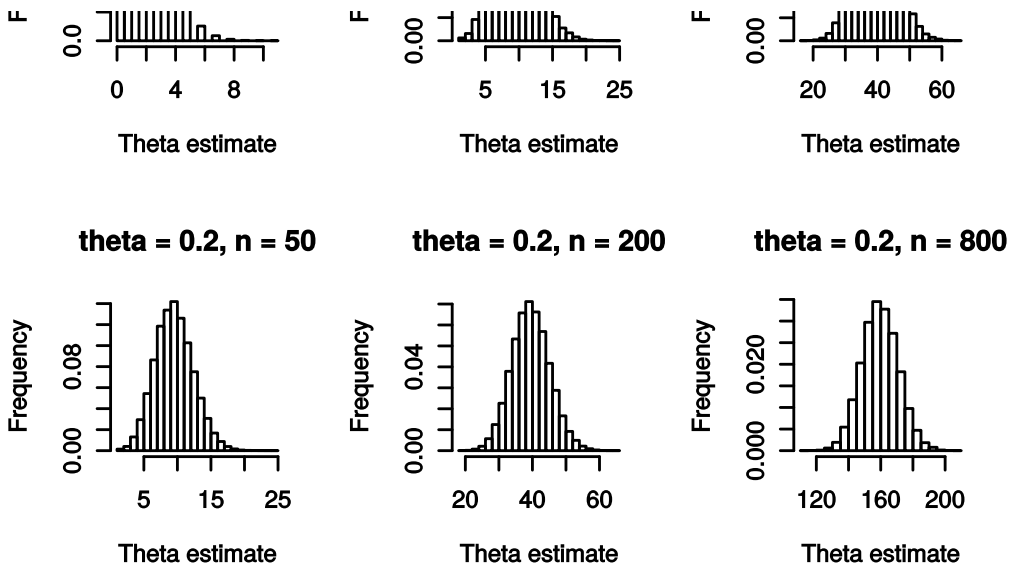
Question 5.e

We're presenting our findings to management; they have asked us to provide visualisations for our results. For each failure probability discussed above (0.01, 0.05 and 0.2) and for each potential sample size discussed above (50, 200, and 800), produce a histogram plot of the maximum likelihood estimates of the failure probability (calculated 50,000 times through 50,000 simulations).

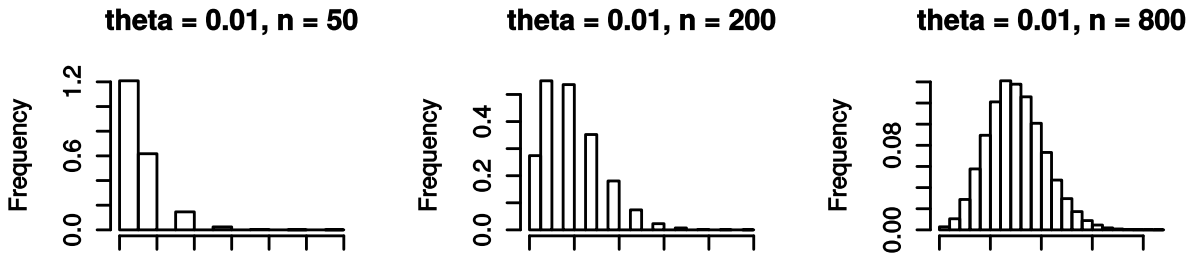
In [11]:
```
make.plot <- function(n, theta) {
    Muvalues = rbinom(50000,n,theta)
    hist(Muvalues, main = sprintf("theta = %s, n = %s", theta, n),
         breaks = 20,xlab = 'Theta estimate',ylab = 'Frequency',
         freq = FALSE)
}
```

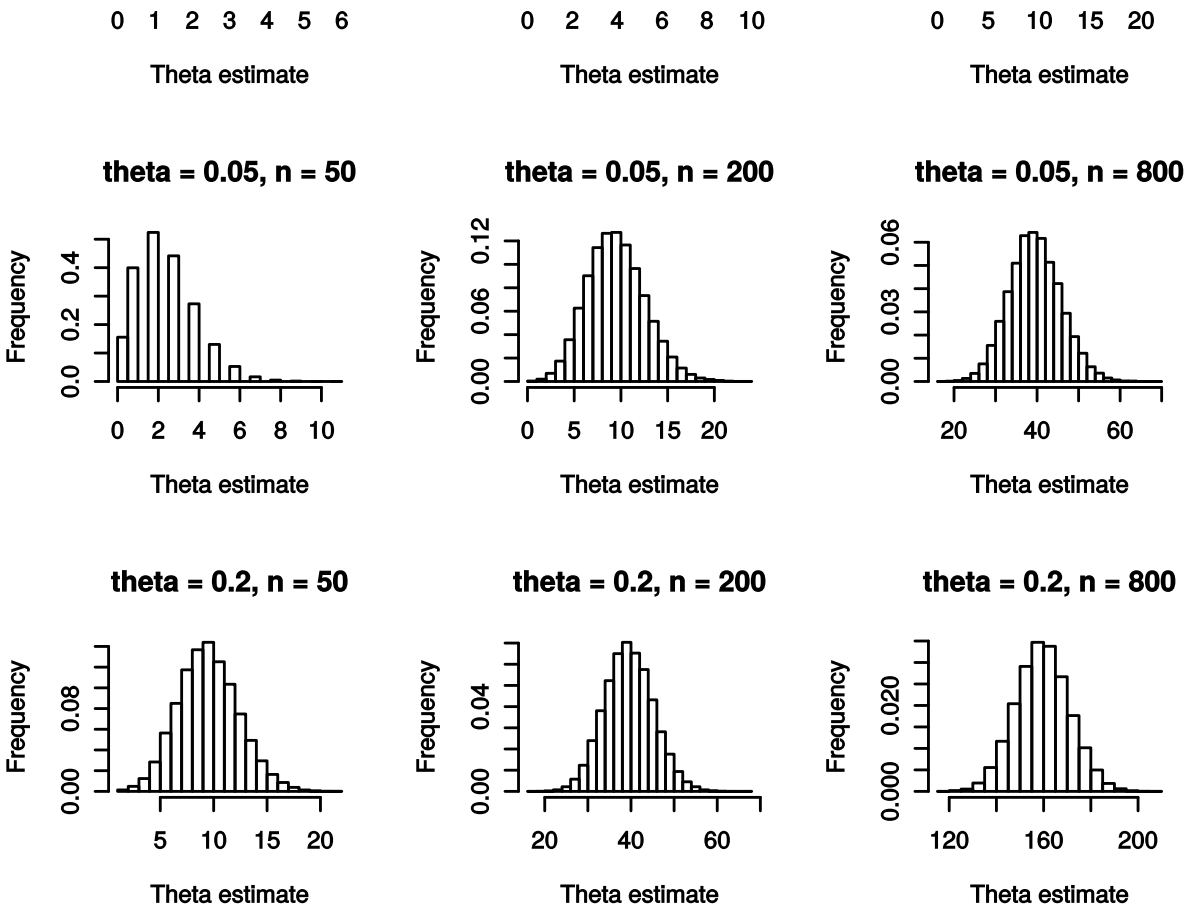
In [12]:
```
par(mfrow=c(3, 3))

options(repr.plot.width=5, repr.plot.height=6)

for (theta in c(0.01, 0.05, 0.2)) {
    for (n in c(50, 200, 800)) {
        make.plot(n, theta)
    }
}
```





In [13]:
```
par(mfrow=c(3, 3))

options(repr.plot.width=6, repr.plot.height=6)

for (theta in c(0.01, 0.05, 0.2)) {
    for (n in c(50, 200, 800)) {
        make.plot(n, theta)
    }
}
```





Question 5.f

Management has asked us recommend how many tests they should run. Based on all the information we have computed, do you recommend 50, 200, 800, or even more tests than that? Justify your answer using relevant calculations and/or by referring to the above plots.

YOUR ANSWER HERE

Based on all the information we have computed, I would recommend 800 tests. This is so as we can see that it is more normally distributed with a larger dataset based on the graph above.