Question 1 – Magic of the CLT

Let $X_i \sim Bin(5, 0.1)$ for $i = 1, 2, \ldots$ independently distributed. Let $S_n = \sum_{i=1}^{n} X_i$.

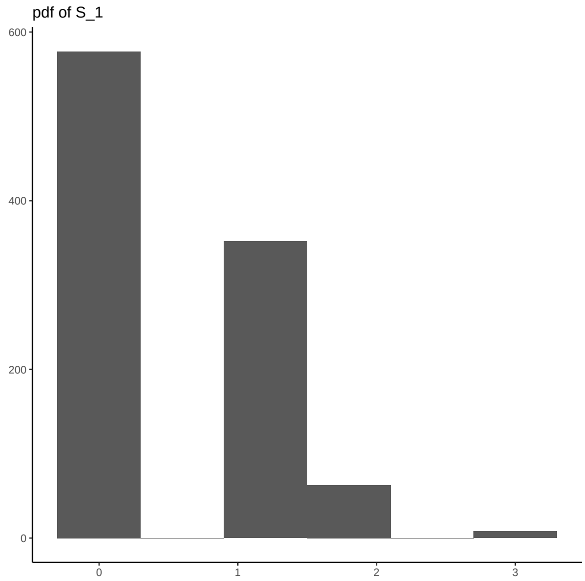
(a) Use R to plot S_1, S_{10}, S_{30} and comment on your plots

Solution: First we need to generate the random variables in the same way as shown in lectures.

```
> M <- matrix (0 ,100 ,1000)
> M[1,] <- rbinom(1000,5,0.1)
> for (i in  2:100){
+         M[i,] <- M[i-1,] + rbinom(1000,5,0.1)
+ }
```

Now we plot the histogram of S_1

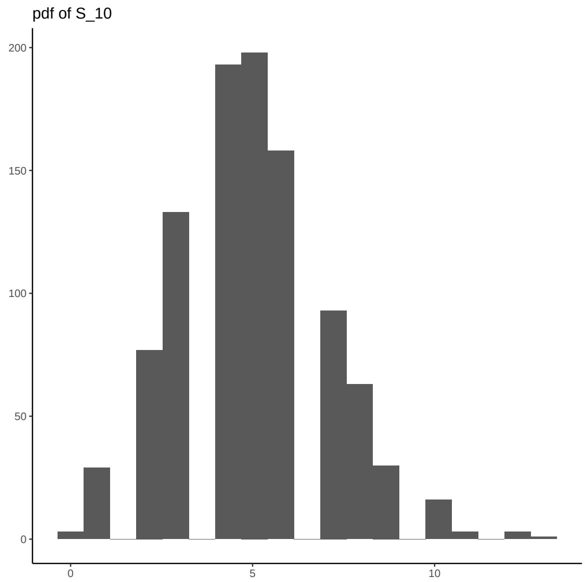
```
> library(ggplot2)
> ggplot()+geom_histogram(aes(M[1,]),bins=6)+labs(title='pdf of S_1',x='',y='')+theme_c
```



Looking at the plot above, we see that there is a right skew to the plot and peak at 0. The distribution follows the binomial curve.

Now plotting S_{10}

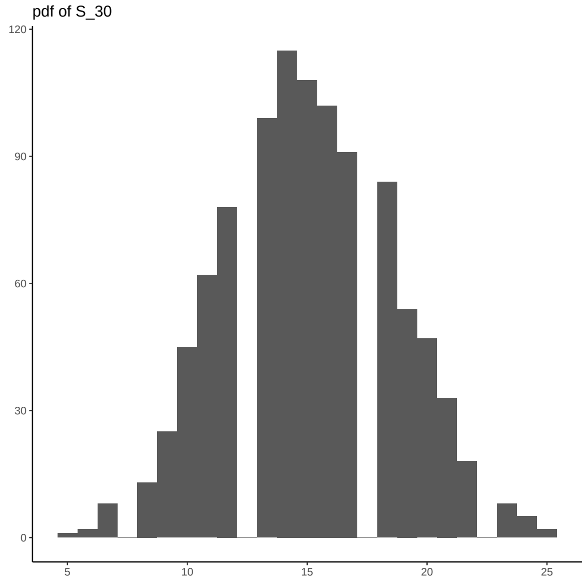
```
> ggplot()+geom_histogram(aes(M[10,]),bins=19)+labs(title='pdf of S_10',x='',y='') + th
```



We can see that this plot has become more symmetrical. There is still a little right skew, and the peak is around 5 and quite thin.

Plotting S_{30} we get

```
> ggplot()+geom_histogram(aes(M[30,]),bins=25)+labs(title='pdf of S_30',x='',y='') + th
```



This plot shows the distribution is getting closer to a symmetric bell shape. The mean has moved up to 15 while the range is now 25. The skewness is no longer evident.

(b) What is (approximately) the distribution of S_{100}?

Solution: As the number of samples increased above the distribution moved closer to the Normal distribution. The peak slowly gets thinner and the tails are longer. Due to the Central Limit Theorem we can conclude that S_{100} is approximately the Normal distribution.

Question 2 – Testing Errors

A textile fibre manufacturer is investigating a new drappery yearn, which the company claims has a thread elongation of 12kg with a known standard deviation of 0.5kg. The company wishes to test the claim of the mean, believing the standard deviation to be true.
The hypothesis H_0 is $\mu = 12$ and the alternative is chosen to be $H_1 : \mu < 12$ with a rejection region of $\mathcal{R} = \{\bar{x} < 11.5\}$. Suppose 4 samples of specimens are taken for the testing.

(a) What is the type I error probability α?

Solution:
To calculate the type I error we need to work out the probability of getting a value less than 11.5 given the hypothesised mean is $\mu = 12$. To calculate this we have

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{11.5 - 12}{\frac{0.5}{\sqrt{4}}} \\ &= -2 \end{aligned}$$

Looking up this probability,

$$P(Z < z) = 0.0227501$$

So $\alpha = 0.0227501$.

(b) Find the type II error probability β, if the true mean is in fact 11.25kg.

Solution:
The type II error is when the null hypothesis is retained where the alternative hypothesis is true. Here we calculate the probability of the mean being being greater than 11.5. To calculate this

$$\begin{aligned} z &= \frac{11.5 - 11.25}{\frac{0.5}{\sqrt{4}}} \\ &= 1 \end{aligned}$$

Looking up the this probability again.

$$P(Z > z) = 0.1586553$$

So $\beta = 0.1586553$.

Question 3 – Confidence

Let X_i be identically and independently distributed with unknown mean μ and known variance $\sigma = 0.2$. Suppose 16 samples are taken and 3.3 is its sample mean. Find the following probabilities:

(a) $P\left(|\bar{x} - \mu| \leq 1.5\frac{0.2}{\sqrt{16}}\right)$

Solution:
To calculate this probability we first need to rearrange the equation to standardise the random variable.

$$\begin{aligned} P\left(|\bar{x} - \mu| \leq 1.5\frac{0.2}{\sqrt{16}}\right) &= P\left(\frac{|\bar{x} - \mu|}{\frac{0.2}{\sqrt{16}}} \leq 1.5\right) \\ &= P(|Z| \leq 1.5) \end{aligned}$$

Now we need to find the probability between these ends, so we rewrite this as

$$\begin{aligned}
P\left(|\bar{x}-\mu| \leq 1.5\frac{0.2}{\sqrt{16}}\right) &= P(-1.5 \leq Z \leq 1.5) \\
&= P(Z \leq 1.5) - P(Z \leq -1.5) \\
&= 0.9331928 - 0.0668072 \\
&= 0.8663856
\end{aligned}$$

(b) Suppose now N samples are taken and $\bar{x}$ is its sample mean. Determine the probability that the distance between $\bar{x}$ and μ is at most $1.8\frac{\sigma}{\sqrt{N}}$.

Solution:
Following a similar process to the previous part. So rearranging the equation to standardise the random variable

$$\begin{aligned}
P\left(|\bar{x}-\mu| \leq 1.8\frac{\sigma}{\sqrt{N}}\right) &= P\left(\frac{|\bar{x}-\mu|}{\frac{\sigma}{\sqrt{N}}} \leq 1.8\right) \\
&= P(|Z| \leq 1.8) \\
&= P(Z \leq 1.8) - P(Z \leq -1,8) \\
&= 0.9640697 - 0.0359303 \\
&= 0.9281394
\end{aligned}$$

Question 4 – P-Values

For the hypothesis test $H_0 : \mu = 7$ against $H_1 : \mu \neq 7$ and variance known, calculate the P-values for the test statistic $z = 2.05$.
Solution:
Here the P-value is

$$P(|X| > 7) = P(|Z| > 2.05)$$

Here we just need to look up the probability that $P(|Z| > 2.05)$. So

$$\begin{aligned}
P(|Z| > 2.05) &= P(-2.05 < Z > 2.05) \\
&= P(Z < -2.05) + P(Z > 2.05) \\
&= 0.0201822 + 0.0201822 \\
&= 0.0403644
\end{aligned}$$

So the P-value here is 0.0404, so there is moderately significant evidence that the mean is not equal to 7.

Question 5 – Hypothesis Testing

The life in hours of a battery is known to be approximately normally distributed, with standard deviation $\sigma = 1.25$ hours. A random sample of 10 batteries has a mean life of $\bar{x} = 40.5$ hours.

(a) Is there evidence to support the claim that battery life exceeds 39.5 hours? Use $\alpha = 0.05$.

Solution: Setting up the hypothesis test we have the hypotheses

$H_0 :$ Battery life is equal to 39.5 hours

$H_1 :$ Battery life is greater than 39.5 hours

To calculate the P-value here we calculate the following probability.

$$\begin{aligned} P\left(X \geq 40.5\right) &= P\left(Z \geq \frac{40.5 - 39.5}{\frac{1.25}{\sqrt{10}}}\right) \\ &= P(Z \geq 2.5298221) \end{aligned}$$

looking this up in the standard normal tables

$$= 0.005706$$

So the P-value is 0.0057, which shows we have strong evidence to support that Battery life is greater than 39.5 hours.

(b) What sample size would be required to ensure that β (Type II error) does not exceed 0.12 if the true mean life is 40.8 hours?

Solution: First we need to find the value of z to find the probability of getting a probability of 0.12, and this should be $P(Z < z)$ as our hypothesis is greater than.

$$\begin{aligned} P(Z < z) &= 0.12 \\ z &= -1.1749868 \end{aligned}$$

Now, we have to work out the value of N so the test statistic is equal to this value. First working back from the standardisation and rearranging.

$$\begin{aligned} \frac{40.5 - 40.8}{\frac{1.25}{\sqrt{N}}} &= -1.1749868 \\ -0.3 &= -1.1749868\frac{1.25}{\sqrt{N}} \\ \sqrt{N} &= 1.1749868\frac{1.25}{0.3} \\ N &= (4.8957783)^2 \\ &= 23.9686452 \end{aligned}$$

We now round this value of N to the next integer. So the sample size requires is 24.

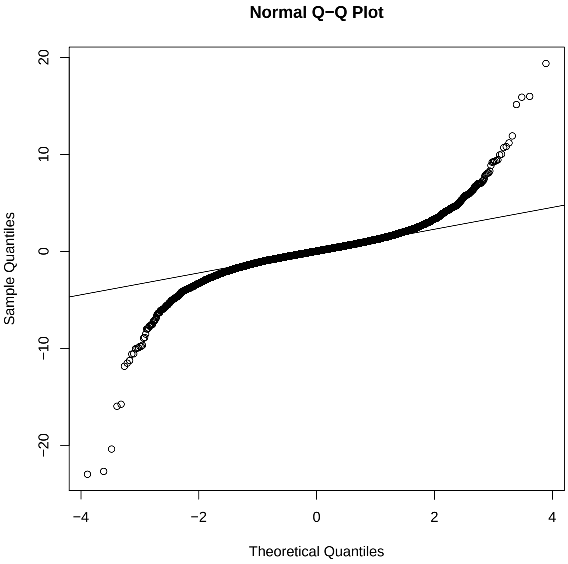
Question 6 – Explore the Student's t-Distribution

Let us use R to become familiar with the Student's t-distribution. For that you may have to use some of the following (R-inbuilt) functions:

```
> rt(n, df)     # random generator for distribution
> ?rt           # information on rt and its inputs
```

Take 10^4 samples of a Student's t-distribution with 3 degrees of freedom. Use a *qnorm* plot to test the relation to a normal distribution. Comment on your result.
Solution: Here we generate the random values from the t-Distribution and plot the Quantile-Quantile plot against the normal distribution.

```
> StudentSample=rt(10^4,3)
> qqnorm(StudentSample)
> qqline(StudentSample)
```



Looking at this plot we see that in the centre of the plot that the generated sample follows the ideal line closely. At the tails the data diverges from the idea line. This is expected as the t-Distribution has thicker tails than the normal distribution.

Question 7 – t-Test

A report based on a study conducted in 2003, reported on the body temperature (in Fahrenheit) of 25 females. The values are given in *5-7.csv*. Test the hypothesis $H_0 : \mu = 98.6$ versus $H_1 : \mu \neq 98.6$, using $\alpha = 0.05$.
Solution: First we need to read in the data from the file provided

```
> temps <- read.csv("5-7.csv")
```

To be able to calculate the t-test we need to calculate the mean and sample standard deviation.

```
> mean(temps$Temp.in.F)

[1] 98.264

> sd(temps$Temp.in.F)

[1] 0.4820788
```

The hypothesis test above is two-sided

$$\begin{aligned} P(|X| > 98.6) &= P\left(|T| > \frac{98.264 - 98.6}{\frac{0.4820788}{\sqrt{25}}}\right) \\ &= P\left(|T| > -3.4849072\right) \\ &= 0.0019124 \end{aligned}$$

So the P-value is 0.0019 so there is strong evidence to support the hypothesis that the mean body temperature is different $98.6°F$

Question 8 – Applied t-Test

The sodium content of fifteen 300-gram boxes of organic cornflakes was determined. The data (in milligrams) is contained in *5-8.csv*. Can you support a claim that mean sodium content of this brand of cornflakes is higher than 115 milligrams? (use $\alpha = 0.05$, state your hypothesis clearly and make a conclusion.)
Solution: First we need to read in the data from the file provided

```
> cornflakes <- read.csv("5-8.csv",header=FALSE)
```

Then we need to calculate the mean and standard deviation of the data.

```
> mean(cornflakes$V1)

[1] 114.4407

> sd(cornflakes$V1)

[1] 0.7701435
```

For the hypothesis test our hypotheses are

H_0 : The mean sodium content of this brand of cornflakes is 115 milligrams.

H_1 : The mean sodium content of this brand of cornflakes is greater than 115 milligrams.

To calculate the P-value we need to find the probability of obtaining a result greater than 115 milligrams based on our sample. We do this as follows

$$\begin{aligned} P(X \geqslant 115) &= P\left(T_{14} \geqslant \frac{114.441 - 115}{\frac{0.770}{\sqrt{15}}}\right) \\ &= P(T_{14} \geqslant -2.813) \\ &= 0.9930868 \end{aligned}$$

So the P-value is 0.9931. Therefore we see that there is inconclusive evidence to support that the mean sodium content of this brand of cornflakes is greater than 115 milligrams. So we retain the null hypothesis and conclude that the mean sodium content is not significantly different from 115 milligrams.