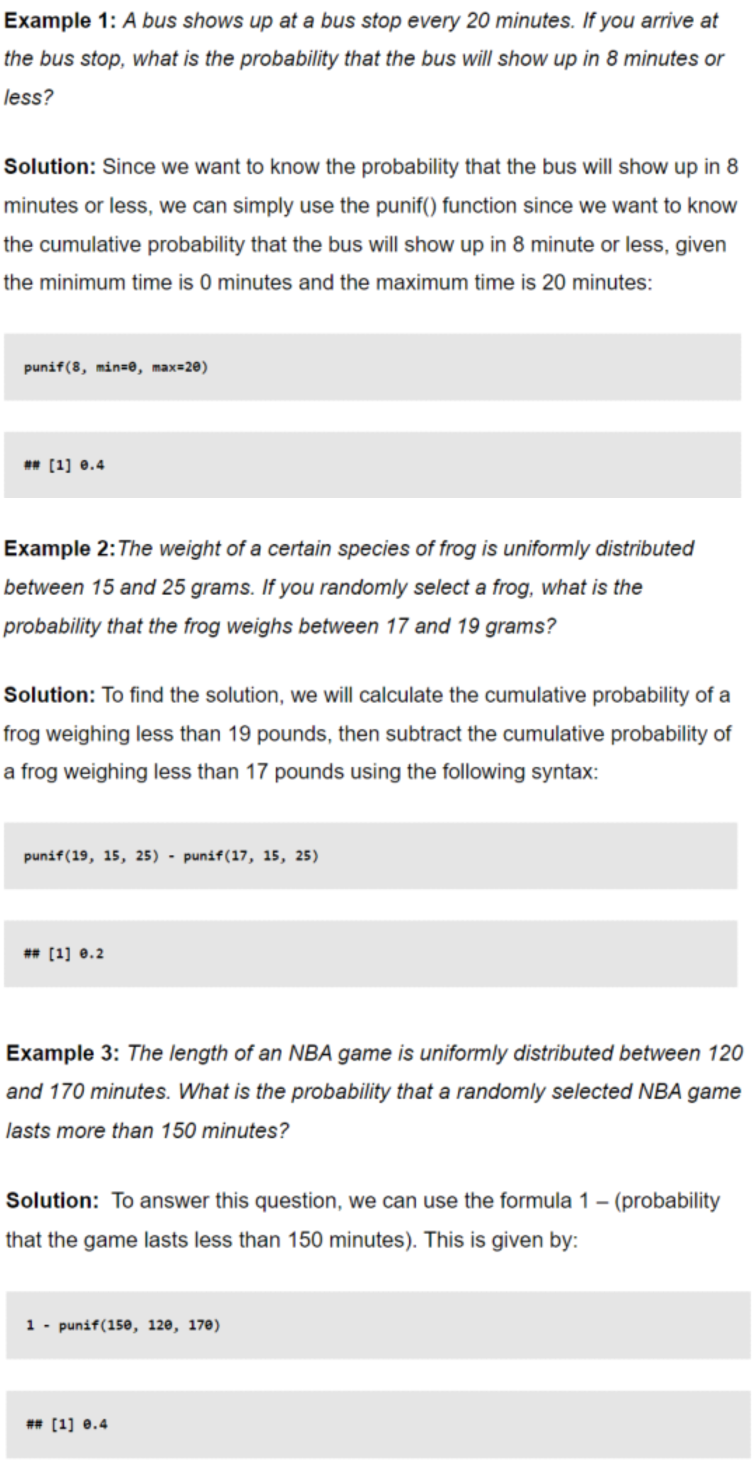
Xác suất và thống kê với ngôn ngữ R

1. Uniform distribution



Example 1: *A bus shows up at a bus stop every 20 minutes. If you arrive at the bus stop, what is the probability that the bus will show up in 8 minutes or less?*

Solution: Since we want to know the probability that the bus will show up in 8 minutes or less, we can simply use the punif() function since we want to know the cumulative probability that the bus will show up in 8 minute or less, given the minimum time is 0 minutes and the maximum time is 20 minutes:

```
punif(8, min=0, max=20)
```

```
## [1] 0.4
```

Example 2:*The weight of a certain species of frog is uniformly distributed between 15 and 25 grams. If you randomly select a frog, what is the probability that the frog weighs between 17 and 19 grams?*

Solution: To find the solution, we will calculate the cumulative probability of a frog weighing less than 19 pounds, then subtract the cumulative probability of a frog weighing less than 17 pounds using the following syntax:

```
punif(19, 15, 25) - punif(17, 15, 25)
```

```
## [1] 0.2
```

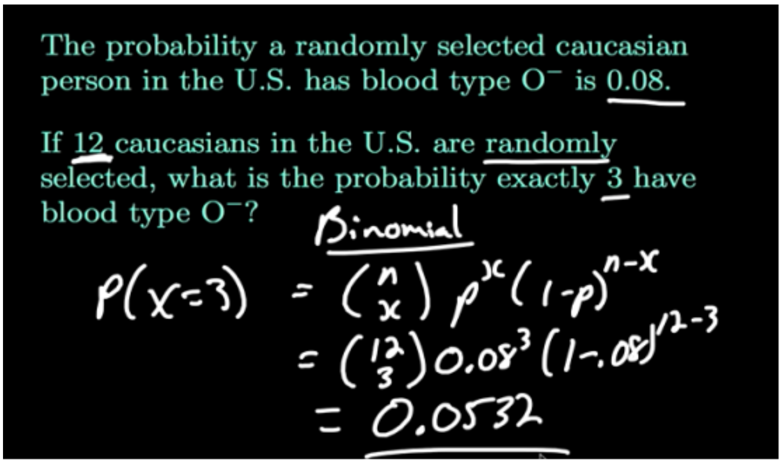
Example 3: *The length of an NBA game is uniformly distributed between 120 and 170 minutes. What is the probability that a randomly selected NBA game lasts more than 150 minutes?*

Solution: To answer this question, we can use the formula 1 – (probability that the game lasts less than 150 minutes). This is given by:

```
1 - punif(150, 120, 170)
```

```
## [1] 0.4
```

2. Binomial distribution

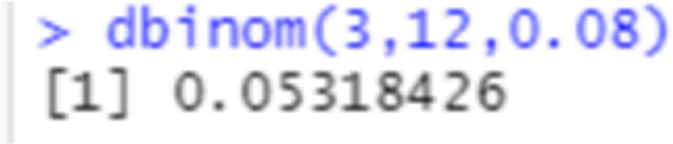


The probability a randomly selected caucasian person in the U.S. has blood type O$^-$ is 0.08.

If 12 caucasians in the U.S. are randomly selected, what is the probability exactly 3 have blood type O$^-$?

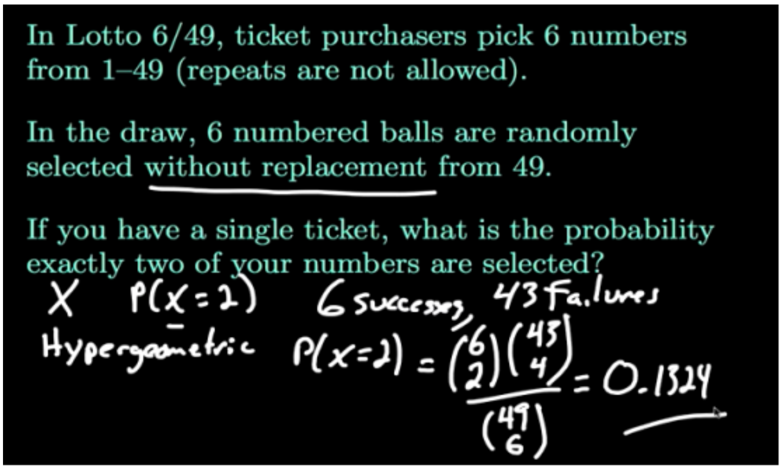
Binomial

$$P(X=3) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \binom{12}{3} 0.08^3 (1-.08)^{12-3}$$
$$= 0.0532$$



```
> dbinom(3,12,0.08)
[1] 0.05318426
```

3. Hypergeometric distribution



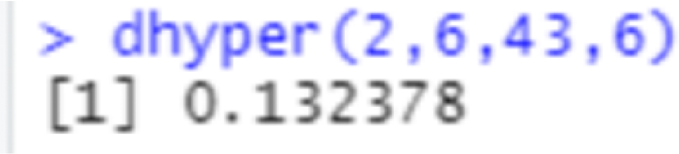
In Lotto 6/49, ticket purchasers pick 6 numbers from 1–49 (repeats are not allowed).

In the draw, 6 numbered balls are randomly selected without replacement from 49.

If you have a single ticket, what is the probability exactly two of your numbers are selected?

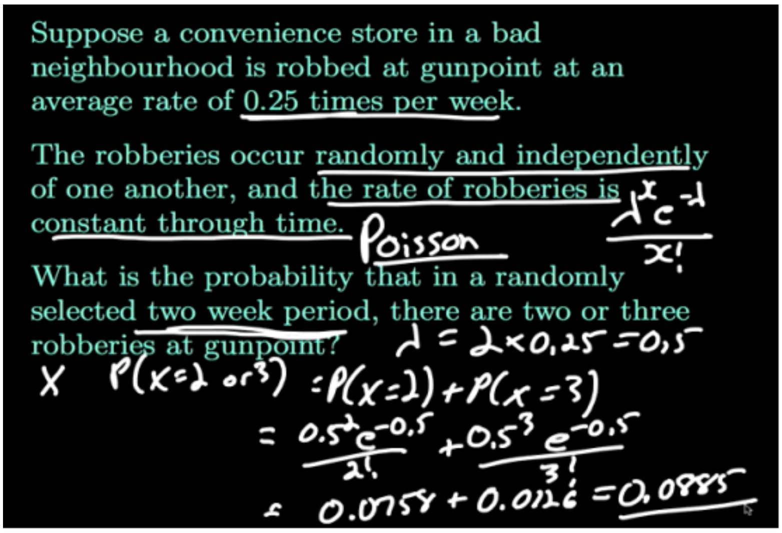
X $P(X=2)$ 6 successes, 43 Failures

Hypergeometric $P(X=2) = \frac{\binom{6}{2}\binom{43}{4}}{\binom{49}{6}} = 0.1324$



```
> dhyper(2,6,43,6)
[1] 0.132378
```

4. Poisson distribution



Suppose a convenience store in a bad neighbourhood is robbed at gunpoint at an average rate of 0.25 times per week.

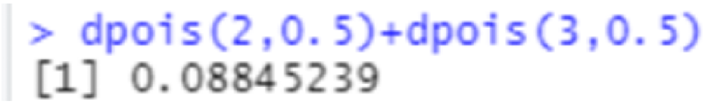
The robberies occur randomly and independently of one another, and the rate of robberies is constant through time. Poisson $\frac{\lambda^x e^{-\lambda}}{x!}$

What is the probability that in a randomly selected two week period, there are two or three robberies at gunpoint? $\lambda = 2\times 0{,}25 = 0{,}5$

$X \quad P(X=2 \text{ or } 3) = P(X=2)+P(X=3)$

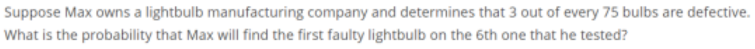
$= \frac{0.5^2 e^{-0.5}}{2!} + \frac{0.5^3 e^{-0.5}}{3!}$

$= 0.0758 + 0.0126 = 0.0885$

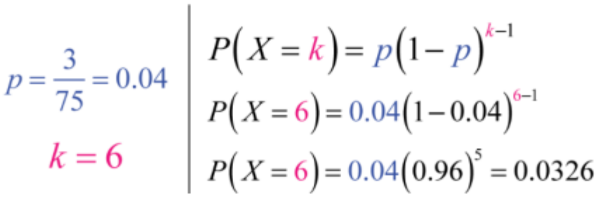


```
> dpois(2,0.5)+dpois(3,0.5)
[1] 0.08845239
```

5. Geometric distribution



Suppose Max owns a lightbulb manufacturing company and determines that 3 out of every 75 bulbs are defective. What is the probability that Max will find the first faulty lightbulb on the 6th one that he tested?

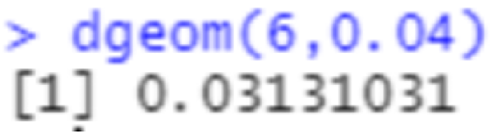


$$p = \frac{3}{75} = 0.04 \qquad k = 6$$

$$P(X=k) = p(1-p)^{k-1}$$

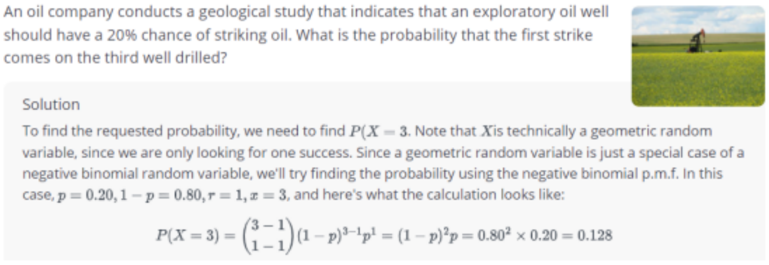
$$P(X=6) = 0.04(1-0.04)^{6-1}$$

$$P(X=6) = 0.04(0.96)^{5} = 0.0326$$



```
> dgeom(6,0.04)
[1] 0.03131031
```

6. Negative binomial distribution

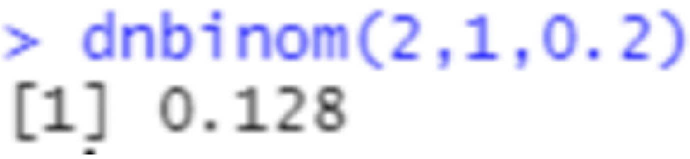


An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. What is the probability that the first strike comes on the third well drilled?

Solution

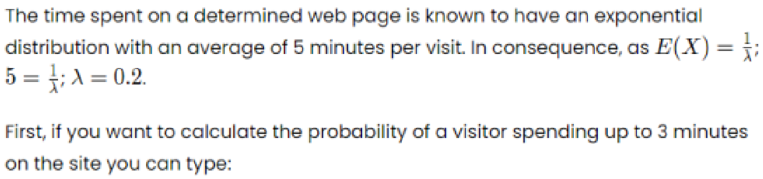
To find the requested probability, we need to find $P(X = 3$. Note that $X$is technically a geometric random variable, since we are only looking for one success. Since a geometric random variable is just a special case of a negative binomial random variable, we'll try finding the probability using the negative binomial p.m.f. In this case, $p = 0.20, 1-p = 0.80, r = 1, x = 3$, and here's what the calculation looks like:

$$P(X=3) = \binom{3-1}{1-1}(1-p)^{3-1}p^1 = (1-p)^2p = 0.80^2 \times 0.20 = 0.128$$



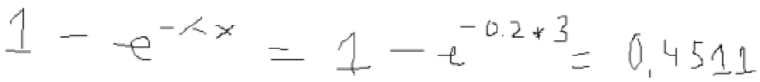
```
> dnbinom(2,1,0.2)
[1] 0.128
```

7. Exponential distribution

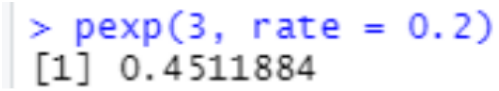


The time spent on a determined web page is known to have an exponential distribution with an average of 5 minutes per visit. In consequence, as $E(X) = \frac{1}{\lambda}$; $5 = \frac{1}{\lambda}$; $\lambda = 0.2$.

First, if you want to calculate the probability of a visitor spending up to 3 minutes on the site you can type:

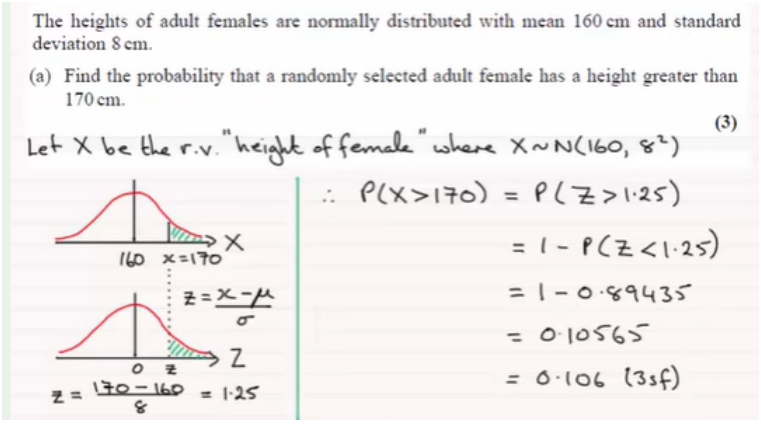


$$1 - e^{-\lambda x} = 1 - e^{-0.2*3} = 0{,}45\underline{11}$$



```
> pexp(3, rate = 0.2)
[1] 0.4511884
```

8. Normal distribution



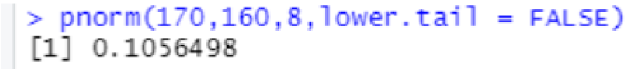
The heights of adult females are normally distributed with mean 160 cm and standard deviation 8 cm.

(a) Find the probability that a randomly selected adult female has a height greater than 170 cm.

(3)

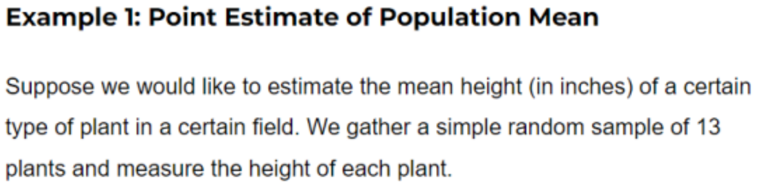
Let X be the r.v. "height of female" where $X \sim N(160, 8^2)$

$$\therefore P(X>170) = P(Z>1.25)$$
$$= 1 - P(Z<1.25)$$
$$= 1 - 0.89435$$
$$= 0.10565$$
$$= 0.106 \text{ (3sf)}$$



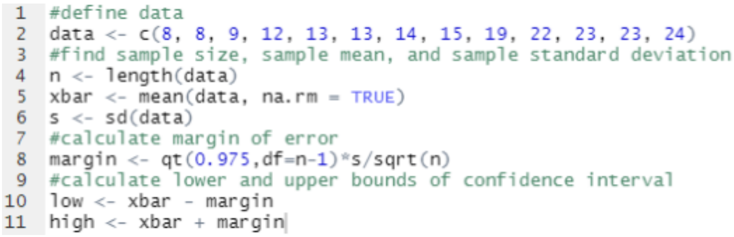
```
> pnorm(170,160,8,lower.tail = FALSE)
[1] 0.1056498
```

9. Point estimate of Population Mean by hand

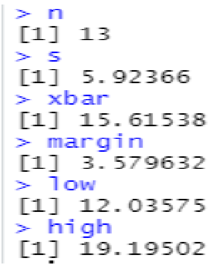


Example 1: Point Estimate of Population Mean

Suppose we would like to estimate the mean height (in inches) of a certain type of plant in a certain field. We gather a simple random sample of 13 plants and measure the height of each plant.



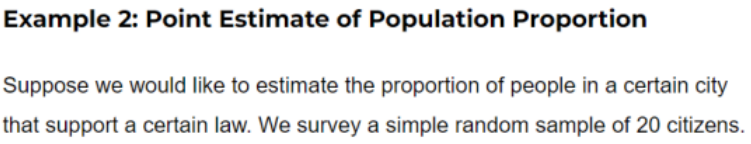
```
#define data
data <- c(8, 8, 9, 12, 13, 13, 14, 15, 19, 22, 23, 23, 24)
#find sample size, sample mean, and sample standard deviation
n <- length(data)
xbar <- mean(data, na.rm = TRUE)
s <- sd(data)
#calculate margin of error
margin <- qt(0.975,df=n-1)*s/sqrt(n)
#calculate lower and upper bounds of confidence interval
low <- xbar - margin
high <- xbar + margin
```



```
> n
[1] 13
> s
[1] 5.92366
> xbar
[1] 15.61538
> margin
[1] 3.579632
> low
[1] 12.03575
> high
[1] 19.19502
```

The 95% confidence interval for the population mean is **[12.0, 19.2]** inches.

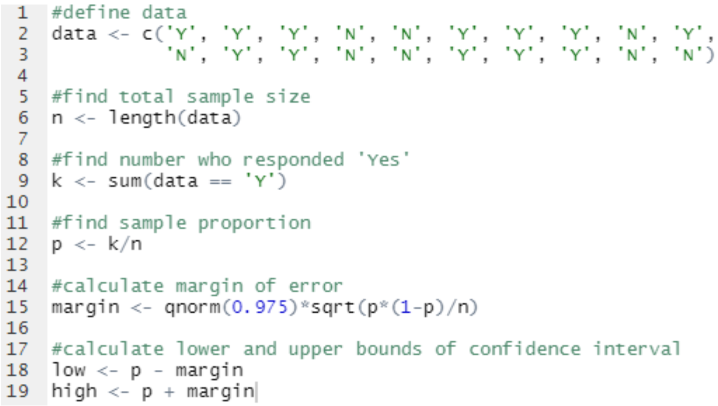
10. Point estimate of Population Proportion by hand



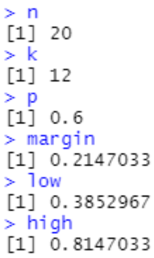
Example 2: Point Estimate of Population Proportion

Suppose we would like to estimate the proportion of people in a certain city that support a certain law. We survey a simple random sample of 20 citizens.

We can also use the following code to calculate a 95% confidence interval for the population mean:



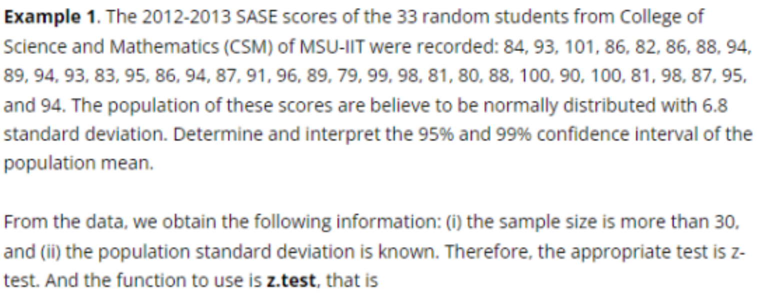
```
#define data
data <- c('Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'N', 'Y',
          'N', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'N', 'N')

#find total sample size
n <- length(data)

#find number who responded 'Yes'
k <- sum(data == 'Y')

#find sample proportion
p <- k/n

#calculate margin of error
margin <- qnorm(0.975)*sqrt(p*(1-p)/n)

#calculate lower and upper bounds of confidence interval
low <- p - margin
high <- p + margin
```



```
> n
[1] 20
> k
[1] 12
> p
[1] 0.6
> margin
[1] 0.2147033
> low
[1] 0.3852967
> high
[1] 0.8147033
```

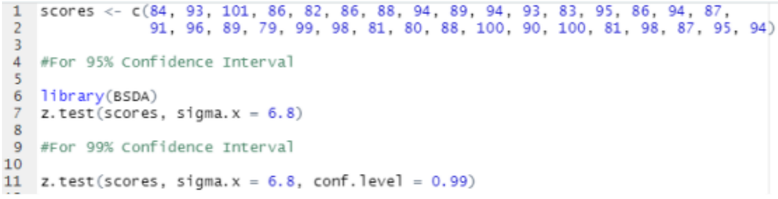
The 95% confidence interval for the population proportion is **[0.39, 0.81]**.

11. Confidence Interval on the Mean of a Normal Distribution, Variance known

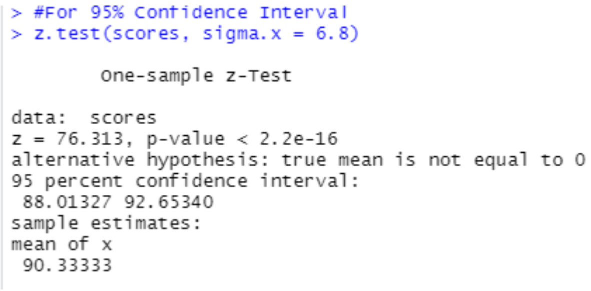


Example 1. The 2012-2013 SASE scores of the 33 random students from College of Science and Mathematics (CSM) of MSU-IIT were recorded: 84, 93, 101, 86, 82, 86, 88, 94, 89, 94, 93, 83, 95, 86, 94, 87, 91, 96, 89, 79, 99, 98, 81, 80, 88, 100, 90, 100, 81, 98, 87, 95, and 94. The population of these scores are believe to be normally distributed with 6.8 standard deviation. Determine and interpret the 95% and 99% confidence interval of the population mean.

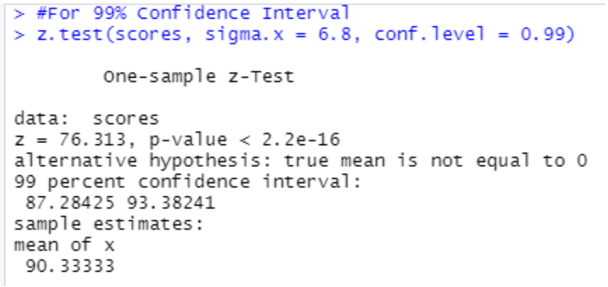
From the data, we obtain the following information: (i) the sample size is more than 30, and (ii) the population standard deviation is known. Therefore, the appropriate test is z-test. And the function to use is **z.test**, that is



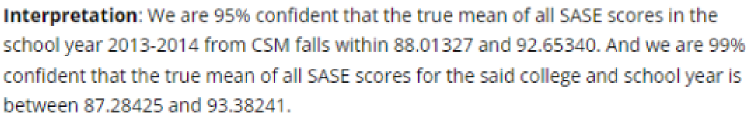
```
scores <- c(84, 93, 101, 86, 82, 86, 88, 94, 89, 94, 93, 83, 95, 86, 94, 87,
            91, 96, 89, 79, 99, 98, 81, 80, 88, 100, 90, 100, 81, 98, 87, 95, 94)

#For 95% Confidence Interval

library(BSDA)
z.test(scores, sigma.x = 6.8)

#For 99% Confidence Interval

z.test(scores, sigma.x = 6.8, conf.level = 0.99)
```



```
> #For 95% Confidence Interval
> z.test(scores, sigma.x = 6.8)

        One-sample z-Test

data:  scores
z = 76.313, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 88.01327 92.65340
sample estimates:
mean of x
 90.33333
```

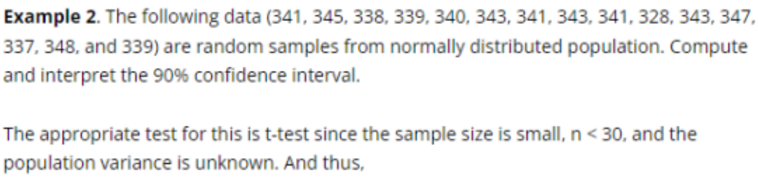


```
> #For 99% Confidence Interval
> z.test(scores, sigma.x = 6.8, conf.level = 0.99)

        One-sample z-Test

data:  scores
z = 76.313, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 87.28425 93.38241
sample estimates:
mean of x
 90.33333
```



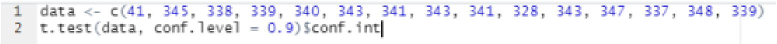
Interpretation: We are 95% confident that the true mean of all SASE scores in the school year 2013-2014 from CSM falls within 88.01327 and 92.65340. And we are 99% confident that the true mean of all SASE scores for the said college and school year is between 87.28425 and 93.38241.

12. Confidence Interval on the Mean of a Normal Distribution, Variance unknown (sample size < 30)

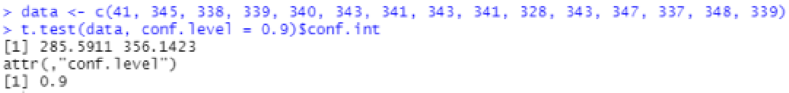


Example 2. The following data (341, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348, and 339) are random samples from normally distributed population. Compute and interpret the 90% confidence interval.

The appropriate test for this is t-test since the sample size is small, n < 30, and the population variance is unknown. And thus,



```
data <- c(41, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348, 339)
t.test(data, conf.level = 0.9)$conf.int
```



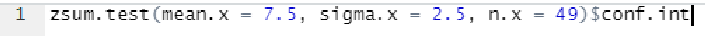
```
> data <- c(41, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348, 339)
> t.test(data, conf.level = 0.9)$conf.int
[1] 285.5911 356.1423
attr(,"conf.level")
[1] 0.9
```

Interpretation: We are 90% confident that the true mean of the population of the given data above is between 285.5911 and 356.1423.

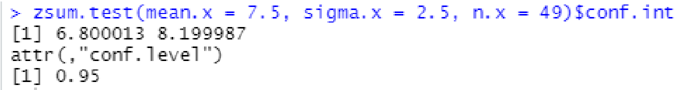
13. Confidence Interval on the Mean of a Normal Distribution, Variance unknown (sample size > 30)

Example 3. The biostatistician took a random sample of 49 patients from a list of all patients ever admitted to the hospital within a three-month period and the number of drugs prescribed per admission was determined for each. The average drug per case was found to be 7.5 with standard deviation of 2.5. Calculate and interpret the 95% confidence interval for true mean of all the patients ever admitted to the hospital.

In this example, no dataset is given, but we have the computed mean = 7.5 of this dataset, standard deviation = 2.5, and sample size = 49. Thus, to compute for the interval estimate of the population mean in R, we use the **zsum.test**



```
zsum.test(mean.x = 7.5, sigma.x = 2.5, n.x = 49)$conf.int
```

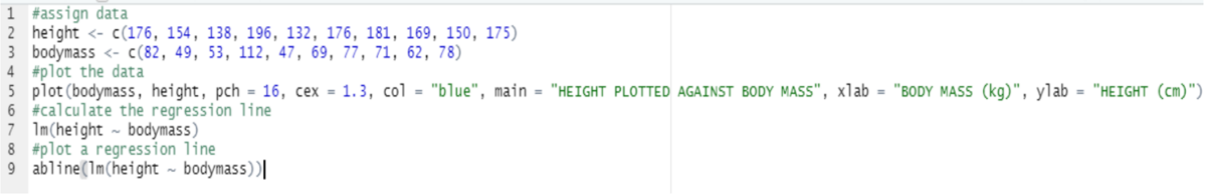


```
> zsum.test(mean.x = 7.5, sigma.x = 2.5, n.x = 49)$conf.int
[1] 6.800013 8.199987
attr(,"conf.level")
[1] 0.95
```

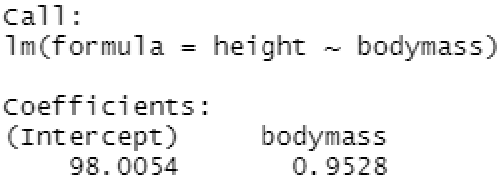
Interpretation: We are 95% confident that the true mean of all the patients ever admitted to the hospital is between 6.800013 and 8.199987.

14. Linear Regression

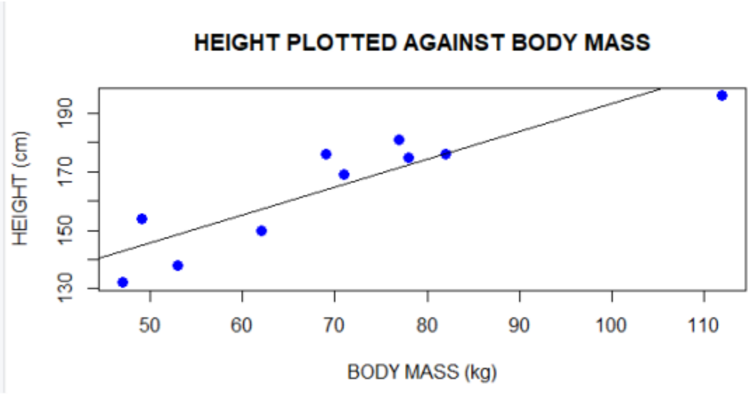
- **Today let's re-create two variables and see how to plot them and include a regression line. We take height to be a variable that describes the heights (in cm) of ten people.**



```
#assign data
height <- c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
bodymass <- c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
#plot the data
plot(bodymass, height, pch = 16, cex = 1.3, col = "blue", main = "HEIGHT PLOTTED AGAINST BODY MASS", xlab = "BODY MASS (kg)", ylab = "HEIGHT (cm)")
#calculate the regression line
lm(height ~ bodymass)
#plot a regression line
abline(lm(height ~ bodymass))
```



```
Call:
lm(formula = height ~ bodymass)

Coefficients:
(Intercept)     bodymass
    98.0054       0.9528
```


HEIGHT PLOTTED AGAINST BODY MASS
HEIGHT (cm)
130
150
170
190
50
60
70
80
90
100
110
BODY MASS (kg)