Assignment 2 - Data Analysis and Statistical Methods

Frans Simanjuntak - S3038971

September 20, 2017

1 Question 1

Given:
P(A) = 0.6 (Probability that a student passes a course)
P(B) = 0.2 (Probability of the students who study > 30 hours)
P(A|B) = 0.9 (Probability that students pass the exam given that they studied > 30 hours)

Answers to the given questions:

(A) Probability that a randomly selected student studies > 30 hours and passes the exam:
P(B ∩ A) = P(B) * P(A|B)
= 0.2 * 0.9
= 0.18

(B) Probability that a student studied more than 30 hours, given that the student passed the exam:
P(B|A) = P(B ∩ A) / P(A)
= 0.18 / 0.6
= 0.3

(C) Events are called independent **if P(B|A)=P(B)**. Since the probability of P(B|A) is greater than P(B) therefore they are not independent events.

2 Question 2

Given:
W = The event when the warning is on
H = The event when the pump is overheated
P(H) = 0.05 (Probability that a pump overheats)
$P(H^C)$ = 0.95 (Probability that a pump does not overheat)

Answer to the given questions:

(A) The formulation of the last two statements as conditional probabilities as follows:
The probability that the warning light is switched on correctly, when the pump is overheated is formulated as **P(W|H) = 0.99**.
The probability that the the light is switched on when nothing is wrong is formulated as **$P(W|H^C)$ = 0.02**.

(B) Probability that the warning light is on during high loads is:
W = (W∩H) ∪ (W∩H^C)
= P(W∩H) + P(W∩H^C)
= (P(H)* P(W|H)) + (P(H^C)* P(W|H^C))
= (0.05*0.99) + (0.95*0.02)
= 0.0495 + 0.19
= 0.0685

(C) The probability that the pump is overheated, given the warning light is on.
$P(H|W) = P(H \cap W)/P(W)$
$= (P(H)*P(W|H))/P(W)$
$= (0.05*0.99)/(0.0685)$
$= (0.0495)/(0.0685)$
$= 0.72262$

3 Question 3

Given:
p = 0.6 remote controller for PS-2 function longer than 1 year.
(1-p) = 0.4

Answer to the given questions:

(A)
- The distribution that can describe the probability of x out of 10 tested controllers function longer than one year is **Binomial distribution**. This statement can be formulated as: $f_B(x;10,0.6) = \binom{10}{x} (0.6)^x (1\text{-}0.6)^{10-x}$
- The expected value is $E(x;n,p) = n*p = 10*0.6 = 6$

(B)
- The probability that all 10 controllers fail the quality control:
$f_B(10;10,0.6) = \binom{10}{10} (0.6)^{10} (0.4)^0$
$= \frac{10!}{10!(10-10)!} 0.6^{10}$
$= 0.6^{10}$
$= 0.0060$
- The probability that 8 or more pass the quality control:
$x \geq 8 = 1 - f_B(10;7,0.6)$
$= 1 - \binom{10}{7} (0.6)^7 (0.4)^3$
$= 1 - \frac{10!}{7!(10-7)!} 0.6^7 0.4^3$
$= 1 - \frac{10!}{7!3!} (0.6)^7 (0.4)^3$
$= 1 - \frac{10.9.8}{3.2.1} (0.6)^7 (0.4)^3$
$= 1 - (120*0.0279936*0.064)$
$= 1 - 0.214990848$
$= 0.785009152$

4 Question 4

Given:
$\mu = 100$
$\sigma = 35$

Answers to the given questions:

(A) The z values that corresponding to:

- $50 = (50\text{-}100)/35 = -1.428$
- $95 = (95\text{-}100)/35 = -0.1428$
- $135 = (135\text{-}100)/35 = 1$

(B) Below is the plotting of the original distribution and the corresponding standard normal distribution.

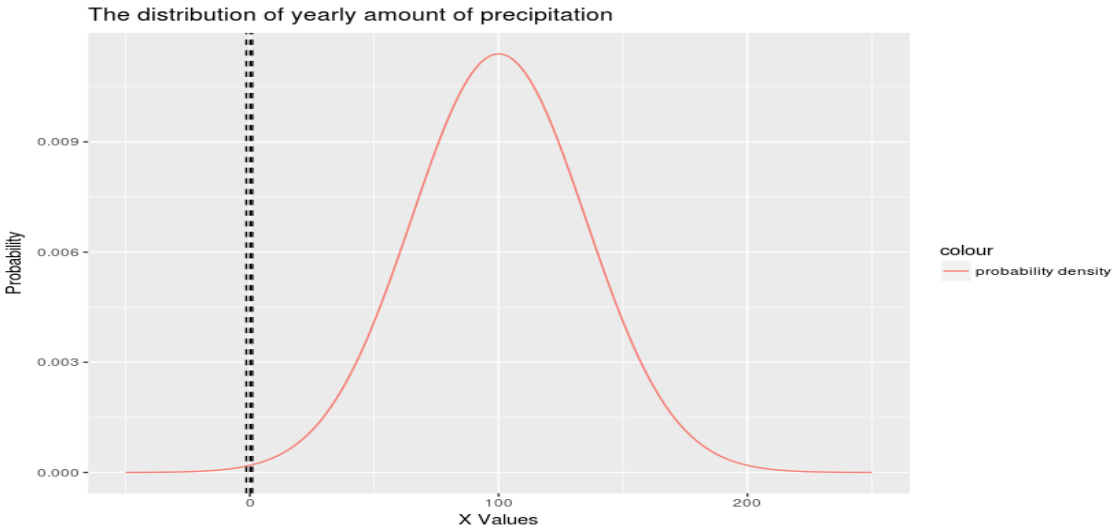

Figure 1: The Standard Normal Distribution.

Source code

```
library(tidyverse)
library(ggplot2)

# define the mean, standard deviation, and z values
x_avg <- 100
sd <- 35
z1 <- -1.428
z2 <- 0.1428
z3 <- 1

# create a series of x values
xvals = seq(-50, 250, 1)
dist_data = data.frame(x=xvals,dist = dnorm(xvals, mean=x_avg, sd=sd))
# plot standard normal distribution
ggplot() +
geom_line(data=dist_data, aes(x=x, y=dist, color="probability density")) +
ylab("Probability") +
xlab("X Values")+
geom_vline(xintercept = z1, linetype="dashed") +
geom_vline(xintercept = z2, linetype="dashed") +
geom_vline(xintercept = z3, linetype="dashed") +
ggtitle("The distribution of yearly amount of precipitation")
```

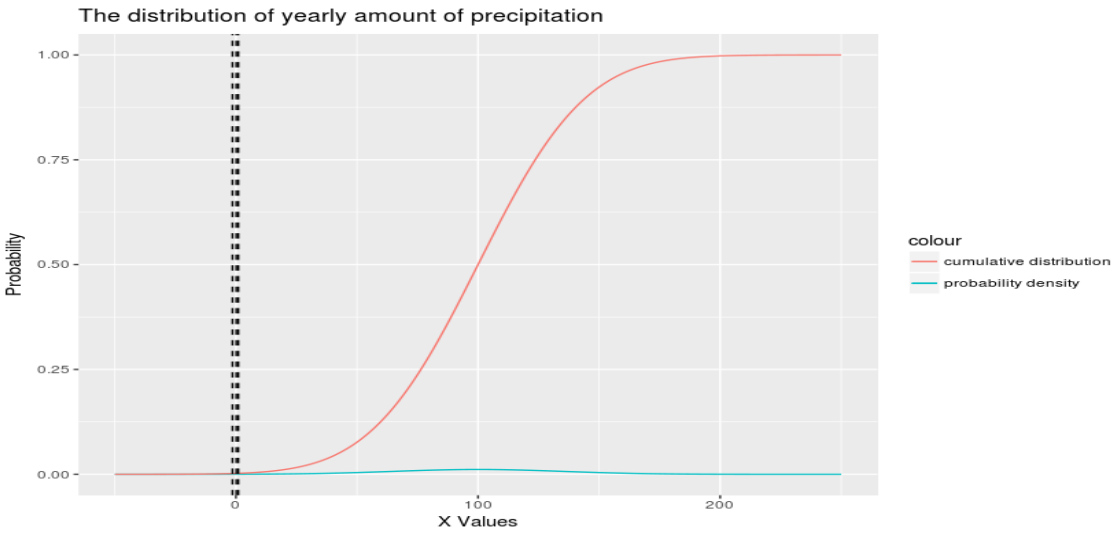

Figure 2: Cumulative Distribution and Probability Density.

Source code

```
library(tidyverse)
library(ggplot2)

# define the mean, standard deviation, and z values
x_avg <- 100
sd <- 35
z1 <- -1.428
z2 <- 0.1428
z3 <- 1

# create a series of x values
xvals = seq(-50, 250, 1)
dist_data = data.frame(x=xvals,
                       dist = dnorm(xvals, mean=x_avg, sd=sd),
                       cumulative_dist = pnorm(xvals, mean=x_avg, sd=sd))
# plot cumulative distribution and probability density
ggplot() +
geom_line(data=dist_data, aes(x=x, y=dist, color="probability density")) +
geom_line(data=dist_data, aes(x=x, y=cumulative_dist,
color="cumulative distribution")) +
ylab("Probability") +
xlab("X Values")+
geom_vline(xintercept = z1, linetype="dashed") +
geom_vline(xintercept = z2, linetype="dashed") +
geom_vline(xintercept = z3, linetype="dashed") +
ggtitle("The distribution of yearly amount of precipitation")
```

(C) The probability that in a given year, the precipitation exceeds 150mm
z_1= x-μ/σ
= 150-100/35
= 1.43

Fn($z>z_1$) = 1 - Fn($z<z_1$)
= 1 - 0.9236 = **0.0764**

(D) The amount of rain corresponds to a record wet year with a probability of 1 in 100.

The probability of the wet year:
Fn(z1) = $\frac{1}{100}$ = 0.01
Fn(z<z_1) = 1 - Fn(z)
= 1 - 0.01
= 0.99

Based on the z table, the z value for the probability of 0.99 is **2.326348**.
The amount of rain (x) = (2.326348*35)+100
x= 181.42218
Therefore the rounded amount of rain is **181**.

5 Question 5

Given:
n = 15
$\overline{x}$ = 30
s = 5

Answers to the given questions:

(A)
- The 95% confidence interval of the mean:
  α = 0.05
  $\frac{s}{\sqrt{n}} = \frac{5}{\sqrt{15}}$ = 1.290
  df=(n-1)= 14
  $t_{\alpha/2}$=$t_{0.025}$=-2.145
  $t_{1-\alpha/2}$=$t_{0.975}$=2.145
  Confidence Interval = $\overline{x}$+$t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x}$+$t_{1-\alpha/2}\frac{s}{\sqrt{n}}$
  = [30-(2.145*1.290), 30+(2.145*1.290)]
  = [27.230 , 32.769]
- The 99% confidence interval of the mean:
  α = 0.01
  $\frac{s}{\sqrt{n}} = \frac{5}{\sqrt{15}}$ = 1.290
  df=(n-1)= 14
  $t_{\alpha/2}$=$t_{0.005}$=-2.977
  $t_{1-\alpha/2}$=$t_{0.995}$=2.977
  Confidence Interval = $\overline{x}$+$t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x}$+$t_{1-\alpha/2}\frac{s}{\sqrt{n}}$
  = [30-(2.977*1.290), 30+(2.977*1.290)]
  = [26.156 , 33.843]

(B) In my opinion, setting the confidence interval to 95% is commonly used in statistic and highly encouraged. The 95% confidence interval can be translated as if I take 15 samples, then 95% of my sample will cover the real mean value of the entire sample or population which makes the interval much larger. The higher the confidence, the larger the interval.

If we set the confidence interval to 99%, the interval will be much larger than 95%. Conversely, if we set the interval to 80%, the interval will be smaller because it will cover only 80% of the mean of sample or population.

(C) The 95% confidence interval for the standard deviation:
$\sigma^2 = \left(\frac{s^2(n-1)}{X^2_{1-\alpha/2}} \quad , \quad \frac{s^2(n-1)}{X^2_{\alpha/2}}\right)$
σ^2 = [(25*14)/qchisq(0.025, 14), (25*14)/qchisq(0.975, 14)]
σ^2 = [62.181 , 13.400]
σ = [7.88 , 3.660]

References