Foundations of Data Science Individual Project

Matthias Bartolo

21/12/2021

Task 1: Queen's Gambit

i)To find the probability that Beth wins the first game

The probability that Beth faces a beginner player is:
P(Beginner)=$\frac{1}{3}$

The probability that Beth faces an intermediate player is:
P(Intermediate)=$\frac{1}{3}$

The probability that Beth faces a master player is:
P(Master)=$\frac{1}{3}$

The probability that Beth wins against a beginner player is:
P(WinBeginner)=$\frac{4}{5}$

The probability that Beth wins against an intermediate player is:
P(WinIntermediate)=$\frac{2}{5}$

The probability that Beth wins against a master player is:
P(WinMaster)=$\frac{1}{5}$

Therefore the Probability that Beth wins the first game is:

P(Beth Wins)=P(Beginner)*P(WinBeginner)+P(Intermediate)*P(WinIntermediate)+P(Master)*P(WinMaster)

P(Beth Wins)=$(\frac{1}{3} * \frac{4}{5}) + (\frac{1}{3} * \frac{2}{5}) + (\frac{1}{3} * \frac{1}{5}) = \frac{7}{15}$

Therefore: P(Beth Wins)=$\frac{7}{15}$

ii)To find the probability that Beth wins the second game, if she won the first game

Assuming the opponent doesn't change

P(Beth Wins Second game)=
P(Beginner)*P(WinBeginner)*P(WinBeginner)+
P(Intermediate)*P(WinIntermediate)*P(WinIntermediate)+
P(Master)*P(WinMaster)*P(WinMaster)

We get:
P(Beth Wins Second game)=$(\frac{1}{3} * \frac{4}{5} * \frac{4}{5}) + (\frac{1}{3} * \frac{2}{5} * \frac{2}{5}) + (\frac{1}{3} * \frac{1}{5} * \frac{1}{5}) = \frac{7}{25}$

Therefore: P(Beth Wins Second game)=$\frac{7}{25}$

iii)Independent and Conditionally Independent Outcomes

Independent outcomes

Independent events describe situations when an event does not affect the probability of the other event occurring, as both events are disjoint events.

Conditionally Independent outcomes

Conditionally Independent events describe situations when an event which is independent from the other event, will affect the other event based on a external parameter. In general taking events A and B which can be either dependent or independent on each other, they would be independent from each other when event C occurs (C being the external parameter).

For example, if the opponent is a beginner, then the probability of winning the first game is 4/5, and the probability of winning the second game is also 4/5 since the opponent doesn't change. In case the opponent changes, the probability of winning the first game remains the same whilst the probability of winning the second game depends on the opponent's skill level.

Assumption

In this case I believe that the outcomes of the games are conditionally independent. Assuming that the opponent remains the same the probability of Beth winning each game is the same, as the probability of each game depends on the opponent's skill level.

Task 2: Conditionitis

i)To find the probability that a man has the disease given that he tested positive

Let:
1. **T** be the event that a man has a positive test
2. $\mathbf{T^C}$ be the event that a man has a negative test
3. **D** be the event that a man has actually the disease
4. $\mathbf{D^C}$ be the event that a man does not have the disease

The probability that a man test positives given he has the disease:
$\mathbf{P(T|D)}=\frac{9}{10}$

The probability that a man test negative given he has the disease:
$\mathbf{P(T^C|D)}=\frac{1}{10}$

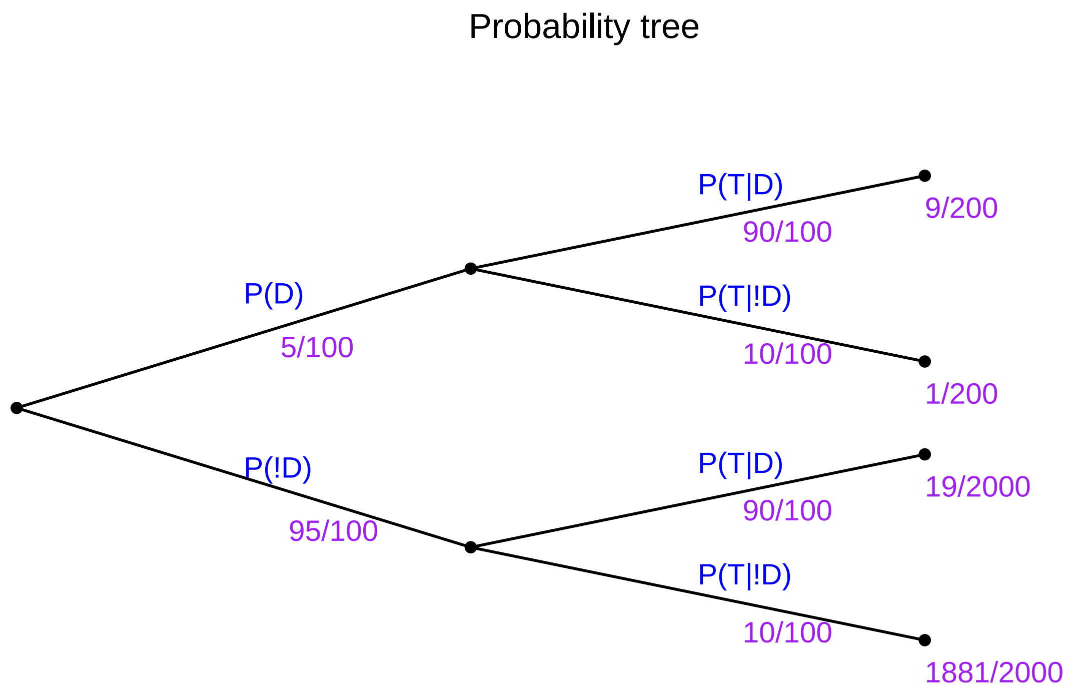
The probability that a man test positive given he doesn't have the disease:
$\mathbf{P(T|D^C)}=\frac{1}{100}$

The probability that a man test negative given he doesn't have the disease:
$\mathbf{P(T^C|D^C)}=\frac{99}{100}$

The probability that a man has the disease:
$\mathbf{P(D)}=\frac{5}{100}$

The probability that a man doesn't have the disease:
$\mathbf{P(D^C)}=\frac{95}{100}$

To obtain the following Probability Tree, a combination of the subsequent functions were used: geom_point, geom_segment as well as geom_text. Points were first plotted with the geom_point, then through the geom_segment the points were connected with a line. Furthermore, geom_text was also used to issue the respective text to the relevant lines.The following functions were also used in order to obtain a white background for the probability tree: theme_void(), theme(), xlab(NULL), ylab(NULL),scale_fill_viridis_c() and theme(). Finally, ggtitle() was used to issue a title to the plot.



Required to find: P(D|T)

Formula:

P(D|T)$=\frac{P(D)P(T|D)}{P(T)}$

where:
P(T)=P$(D \cap T)$+P$(D^C \cap T)$

P(T)=P(D)*P(T|D)+P(D$^\mathrm{C}$)*P(T|D$^\mathrm{C}$)

P(T)$=(\frac{5}{100} * \frac{90}{100}) + (\frac{95}{100} * \frac{1}{100}) = \frac{109}{2000}$

Putting into formula:

P(D|T)$=\frac{P(D)P(T|D)}{P(T)}$

We get:
$(\frac{5}{100} * \frac{90}{100})/\frac{109}{2000} = \frac{90}{109}$

Therefore: P(D|T)$=\frac{90}{109}$

ii)To find the probability that a man has the disease given that he has a negative test

Required to find: $\mathbf{P(D|T^C)}$

Formula:

$\mathbf{P(D|T^C)}$=$\frac{P(D)P(T^C|D)}{P(T^C)}$

where:
$\mathbf{P(T^C)}$=P($D \cap T^C$)+P($D^C \cap T^C$)

$\mathbf{P(T^C)}$=P(D)*P(T^C|D)+P(D^C)*P(T^C|D^C)

$\mathbf{P(T^C)}$=$(\frac{5}{100} * \frac{10}{100}) + (\frac{95}{100} * \frac{99}{100}) = \frac{1891}{2000}$

Putting into formula:

$\mathbf{P(D|T^C)}$=$\frac{P(D)P(T^C|D)}{P(T^C)}$

We get:
$(\frac{10}{100} * \frac{90}{100})/\frac{1891}{2000} = \frac{10}{1891}$

Therefore: $\mathbf{P(D|T^C)}$=$\frac{10}{1891}$

Task 3: Bottled Water

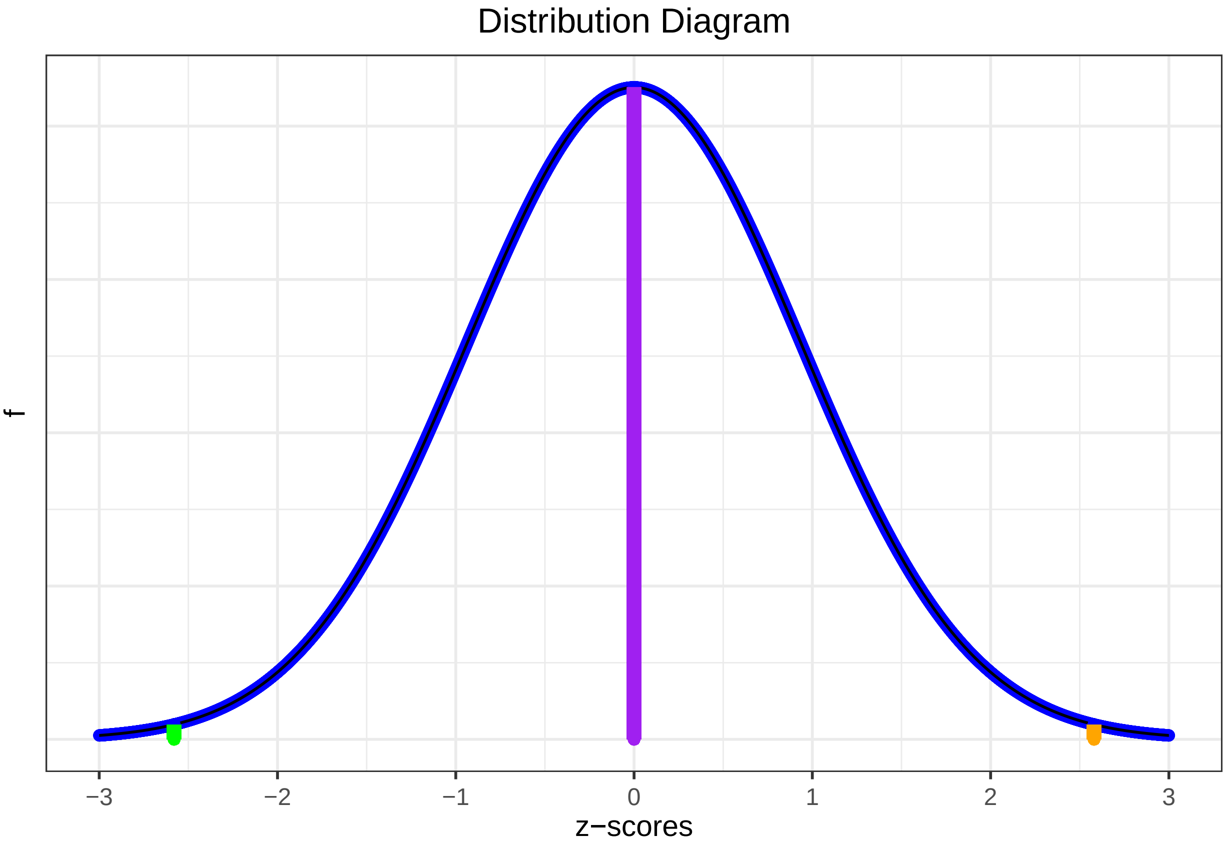
i)Null and Alternative Hypothesis

Null Hypothesis: H_0: $\theta = 240$ml
The machine continues to dispense water normally, since the amount of water dispensed is equal to 240ml, and thus the machine does not need to be stopped.

Alternate Hypothesis: H_1: $\theta \neq 240$ml
The machine should be stopped and production waits for repairs, since the amount of water is not equal to 240ml.

ii)Distribution Diagram

To obtain the following Distribution Diagram, a combination of the subsequent functions were used: geom_point, geom_segment, geom_line, ylab, xlab, theme_bw, theme and scale_x_continuous.



Given:
1. Sample Mean: $\bar{x} = 234ml$
2. Population Mean: $\mu = 240ml$
3. Sample Size: $N = 35$
4. Variance: $\sigma^2 = 0.88$
5. Significance: $\alpha = 0.99$

To obtain the Distribution diagram;

1.First the standard deviation was found by obtaining the square root of the variance (0.88). i.e $\sigma = 0.93$

2.The level of significance (0.99) was subtracted from 1 and 0.01 was obtained. The latter (0.01) was then divided by 2, since a two-tailed test was conducted and 0.005 was achieved. The value (0.005) was subtracted from 1 and the result was used in the z-table to find the critical values (green/orange values in graph).

3.The dnorm function was utilized to populate the y values to plot a normal distribution graph.

4.Furthermore, the rejection regions are the regions which are smaller than -2.58 & larger than 2.58
(-2.58 and 2.58 being the critical values).

The purple line signifies the mean which is 240ml.

Link to z-table used:
https://www.statisticshowto.com/tables/z-table/

iii)Sample parameters and test-statistics

Finding the Test-Statistic:

Formula used:
$\mathbf{z}=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$

Where:
1. z: Standard Score
2. $\bar{x}$: Observed value
3. μ: Mean of sample
4. $\sigma_{\bar{x}}$: Standard Error of means

Finding the Standard Error of means:

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} = \frac{0.93}{\sqrt{35}}$

$\sigma_{\bar{x}} = 0.159$

Substituting in the Formula:

$\mathbf{z}=\frac{234-240}{0.159}$

Therefore Test Statistic: z = -37.84

iv)Comparing the test-statistic with the critical values

When comparing the test-statistic of -37.84 with the left-most critical value of -2.58 (green value in graph), it was noted that the test-statistic was smaller than the critical value. This would imply that the test-statistic resides in the left rejection region.

v)Conclusion

In conclusion, due to the test-statistic being smaller than the critical value, the null hypothesis is rejected, since the conclusion has a significance of more than 99%. Thus, The alternate hypothesis is utilized where the machine should be stopped and production waits for repairs.