CmpE 343 - Assignment II

Karahan Sarıtaş - 2018400174

December 29, 2021

1 Answers

1. (a) $\sigma = 0.1$ is given and we want to construct 95% confidence interval for the μ for the daily number of customers. We can apply Central Limit Theorem.

$$\begin{aligned}
1 - \alpha &= 0.95 \\
\alpha/2 &= 0.025 \\
z_{\alpha/2} &= 1.96 \\
\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} &< \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \\
\bar{x} - 1.96 \cdot \frac{0.1}{\sqrt{n}} &< \mu < \bar{x} + 1.96 \cdot \frac{0.1}{\sqrt{n}}
\end{aligned}$$

Mean of the sample is calculated using the **benan.npy** file.

$$\bar{x} = 54.9993 \text{ and } n = 32$$

$$54.9993 - 1.96 \cdot \frac{0.1}{\sqrt{32}} < \mu < 54.9993 + 1.96 \cdot \frac{0.1}{\sqrt{32}}$$

$$54.9646 < \mu < 55.0339$$

We are 95% confident that mean of the population lies within the given range: $54.9646 < \mu < 55.0339$.

(b) Null hypothesis : $\mu = 55$
Alternative hypothesis : $\mu \neq 55$

$$\begin{aligned}
\alpha/2 &= 0.025 \\
z_{\alpha/2} &= 1.96 \\
z &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
z &= \frac{54.9993 - 55}{0.1/\sqrt{32}} \\
z &= -0.0396
\end{aligned}$$

Critical region: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

$$-1.96 < -0.0396 < 1.96$$

Computed test statistic is not in the critical region. Therefore it **fails to reject** the hypothesis $\mu = 55$.

(c) The P-value corresponding to z = -0.0396 given by:
$P = P(|Z| > 0.0396) = 2P(Z < -0.0396) \approx 2 \cdot 0.4840 = 0.9680$
P-value is significantly greater than suggested level of significance α, therefore it again fails to reject the hypothesis $\mu = 55$.

(d) Testing of null hypothesis $\mu = \mu_0$ against alternative hypothesis $\mu \neq \mu_0$ at a significance level $\alpha = 0.05$ is equivalent to computing a $100(1-\alpha)\%$ confidence interval on μ, and rejecting the null hypothesis if μ_0 is outside of the confidence interval. But in our case, there was no sufficient evidence in favor of the alternative hypothesis, therefore we concluded by saying "**fail to reject**". One should not consider confidence interval estimation and hypothesis testing as separate forms of statistical inference. However, there are differences between the classic hypothesis testing with a fixed α and the P-value approach. In P-value approach, we do not have a fixed α, at the end one has to use judgment based on the P-value and knowledge of the scientific system. α refers to a pre-determined probability and P-value is calculated after the experiment.

(e) Since σ is not given, and we know that the population has Gaussian distribution, we should use t-distribution.

$$\begin{aligned} 1 - \alpha &= 0.95 \\ \alpha/2 &= 0.025 \\ v &= 31 \\ t_{\alpha/2} &\approx 2.04 \\ \bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu &< \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \end{aligned}$$

Variance of the sample is calculated using the **benan.npy** file.

$$\bar{x} = 54.9993 \text{ and } s^2 = 0.01085$$
$$54.9993 - 2.04 \cdot \frac{0.1042}{\sqrt{32}} < \mu < 54.9993 + 2.04 \cdot \frac{0.1042}{\sqrt{32}}$$
$$54.9624 < \mu < 55.0369$$

New confidence interval is slightly wider than the former interval. That's because of two reasons: Firstly, variance of the sample is slightly greater than the variance of the population given in the first question. Greater variance reduces the certainty and therefore increases the width of the interval (about which we can be $100(1-\alpha)\%$ sure). Secondly, t-value above which there is an area of α is greater than the corresponding z-value for the same α. It is because probability of getting values far from the mean is larger with a T-distribution than a normal distribution since T-values depend on the fluctuations of two quantities, $\bar{X}$ and S^2.

2.

$$H_0 : \beta = 3$$
$$H_1 : \beta \neq 3$$

Type I error: H_0 is true, however result lies in the critical region, therefore we reject it.
Type II error: H_0 is false, however result doesn't lie in the critical region, therefore we do not reject the H_0.
Uniform distribution:

$$f(x; 0, B) = \begin{cases} \frac{1}{B}, & 0 \leq x \leq B \\ 0, & \text{elsewhere} \end{cases}$$

(a) Probability of type I error:

$$\begin{aligned} &P(k \leq 0.1 \mid \beta = 3) + P(k \geq 2.9 \mid \beta = 3) \\ &= \int_0^{0.1} \frac{1}{3}\, dx \; + \int_{2.9}^{3} \frac{1}{3}\, dx \\ &= \frac{0.1}{3} + \frac{3-2.9}{3} \\ &= \frac{1}{15} = 0.067 \end{aligned}$$

(b) Probability of type II error:

$$\begin{aligned} &P(0.1 \leq k \leq 2.9 \mid \beta = 3.5) \\ &= \int_{0.1}^{2.9} \frac{1}{3.5}\, dx \\ &= \frac{2.9-0.1}{3.5} \\ &= \frac{4}{5} = 0.8 \end{aligned}$$

(c) To check the consistency, let's calculate the other probabilities as well.

$P(0.1 \leq k \leq 2.9 \mid \beta = 3) = 0.93$ (correct hypothesis - correct decision)

$\frac{14}{15} + \frac{1}{15} = 1$ (consistent)

$P(k \leq 0.1 \mid \beta = 3.5) + P(k \geq 2.9 \mid \beta = 3.5) = 0.2$ (wrong hypothesis - correct decision)

$0.8 + 0.2 = 1$ (consistent)

(d) Let's first explain what's going on in the meme, then talk about the bottom line. Usually when doing statistical testing, null hypothesis refers to the hypothesis of "no difference" or "no link", such as in this example where H_0 refers to "no link between jelly beans and acne". And of course, alternative hypothesis is "there is a link between jelly beans and acne". If the calculated P-value at the end of the experiment is less than the chosen significance level (which is equal to 0.05 in this meme), then we reject the null hypothesis, because a small P-value indicates a significant result. In other words, probability that the hypothesis is true given the evidence, is very low.

At first, scientists find out that $p > 0.05$, which means null hypothesis is true and apparently the hypothesis they tested is "no link between jelly beans and acne". After revealing the results, claimer tries to change the hypothesis a little bit by adding the color effect. Essentially, there is no difference between the colors, therefore "no link between jelly beans and acne" argument should stay valid. However, when the scientists tested the hypotheses multiple times, probability of observing at least one significant result just due to chance has increased and unsurprisingly they've encountered with a significant result while experimenting on green jelly beans.

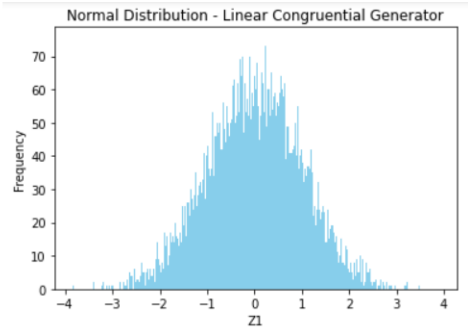
For a mathematical analysis, think about this way: There were 20 hypothesis tests carried out after the "specific color" claim. Probability of not having a significant result is 0.95. It means that we expect 1 incorrect rejection in 20 tests. This rejection is caused by Type I error, in which hypothesis is true but result lies in the critical region. Upon moving the mouse over the meme, a text shows up saying that "So, uh, we did the green study again and got no link. It was probably a–'RESEARCH CONFLICTED ON GREEN JELLY BEAN/ACNE LINK; MORE STUDY RECOMMENDED!'. It shows that the previous result of the green study was due to the chance.

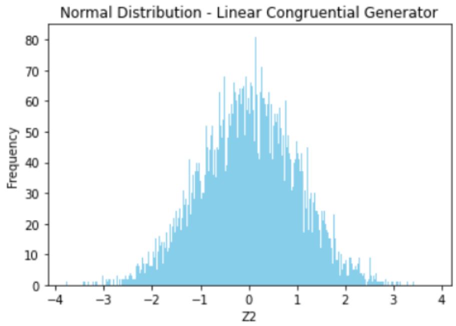
That was a good one, I'd also like to share another xkcd meme from here, which involves both Bayesian statistics and hypothesis testing (and p-value).

3. (a) I've implemented a linear congruential generator based on the values used by Donald Knuth for the development of MMIX. Source: A Search for Good Pseudo-random Number Generators: Survey and Empirical Studies. The values of the multipler, increment and module are defined as:

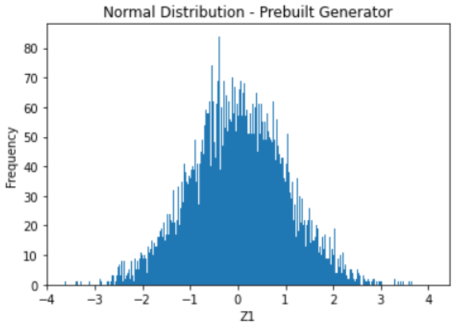
$$
\begin{aligned}
a &= 6364136223846793005 \\
c &= 1442695040888963407 \\
m &= 2^{64}
\end{aligned}
$$

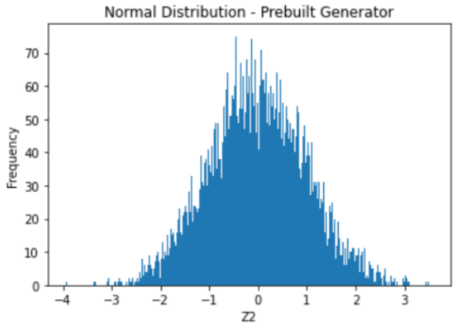
(b) Distributions of Z_1 and Z_2 observed with the linear congruential generator can be found below:



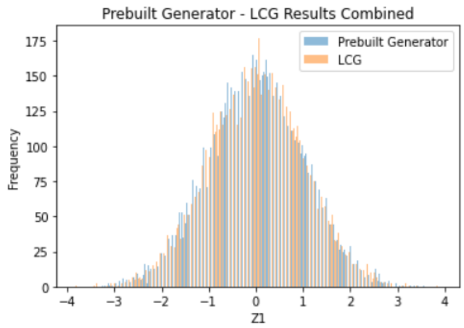


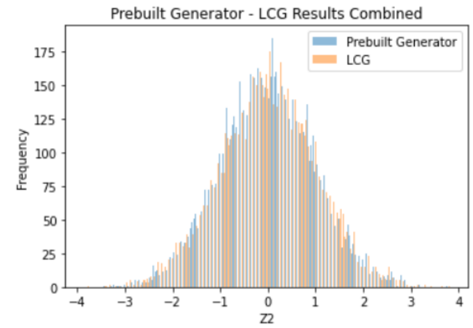
Distributions of Z_1 and Z_2 observed with the prebuilt generator can be found below:





For the comparison of the result, you can find the superimposed histograms for Z_1 and Z_2 generated by linear congruential generator and prebuilt generator. As can be seen, distributions are very similar to each other with $\mu = 0$ and $\sigma^2 = 1$:





(c) First let's leave the random variable U_1 and U_2 alone:

$$\begin{aligned} Z_1 &= \sqrt{-2ln(U_1)} \cdot \cos(2\pi U_2) \\ Z_2 &= \sqrt{-2ln(U_1)} \cdot \sin(2\pi U_2) \\ Z_1^2 &= -2ln(U_1) \cdot \cos^2(2\pi U_2) \\ Z_2^2 &= -2ln(U_1) \cdot \sin^2(2\pi U_2) \\ Z_1^2 + Z_2^2 &= -2ln(U_1) \cdot (\cos^2(2\pi U_2) + \sin^2(2\pi U_2)) \\ ln(U_1) &= -\frac{(Z_1^2 + Z_2^2)}{2} \\ U_1 &= e^{-\frac{(Z_1^2+Z_2^2)}{2}} \end{aligned}$$

$$\begin{aligned} Z_1 &= \sqrt{-2ln(U_1)} \cdot \cos(2\pi U_2) \\ Z_2 &= \sqrt{-2ln(U_1)} \cdot \sin(2\pi U_2) \\ \frac{Z_1}{Z_2} &= \frac{1}{\tan(2\pi U_2)} \\ U_2 &= \frac{1}{2\pi} \cdot \arctan(\frac{Z_2}{Z_1}) \end{aligned}$$

Now, we have U_1 and U_2, which are **continuous** random variables with joint probability distribution $f(u_1, u_2)$. $Z_1 = m_1(U_1, U_2)$ and $Z_2 = m_2(U_1, U_2)$ define a one-to-one transformation between the points (u_1, u_2) and (z_1, z_2) so that the equations $z_1 = m_1(u_1, u_2)$ and $z_2 = m_2(u_1, u_2)$ may be uniquely solved for u_1 and u_2 in terms of z_1 and z_2, say $u_1 = w_1(z_1, z_2)$ and $u_2 = w_2(z_1, z_2)$. Then the joint probability distribution of Z_1 and Z_2 is

$$\begin{aligned} g(Z_1, Z_2) &= f[w_1(Z_1, Z_2), w_2(Z_1, Z_2)]|J|, \\ J &= \begin{vmatrix} \frac{\partial u_1}{\partial z_1} & \frac{\partial u_1}{\partial z_2} \\ \frac{\partial u_2}{\partial z_1} & \frac{\partial u_2}{\partial z_2} \end{vmatrix} \end{aligned}$$

This equation comes from the Theorem 7.4, source: Probability & Statistics for Engineers & Scientists, 9th Edition (pg. 214). For the derivative of $\arctan(f(x))$:

$$\begin{aligned} \arctan(f(x)) &= y \\ \tan(y) &= f(x) \\ (1 + \tan^2 y) \cdot \frac{dy}{dx} &= \frac{d(f(x))}{dx} \\ \frac{dy}{dx} &= \frac{\frac{d(f(x))}{dx}}{1 + f^2(x)} \end{aligned}$$

Continue with calculating the Jacobian:

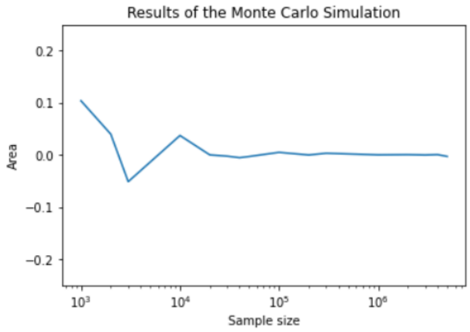
$$\frac{\partial u_1}{\partial z_1} = -Z_1 \cdot e^{-\frac{(Z_1^2+Z_2^2)}{2}}$$
$$\frac{\partial u_1}{\partial z_2} = -Z_2 \cdot e^{-\frac{(Z_1^2+Z_2^2)}{2}}$$
$$\frac{\partial u_2}{\partial z_1} = \frac{1}{2\pi} \cdot \frac{-\frac{Z_2}{Z_1^2}}{1+\frac{Z_2^2}{Z_1^2}} = \frac{1}{2\pi} \cdot \frac{-Z_2}{Z_1^2+Z_2^2}$$
$$\frac{\partial u_2}{\partial z_2} = \frac{1}{2\pi} \cdot \frac{\frac{1}{Z_1}}{1+\frac{Z_2^2}{Z_1^2}} = \frac{1}{2\pi} \cdot \frac{Z_1}{Z_1^2+Z_2^2}$$

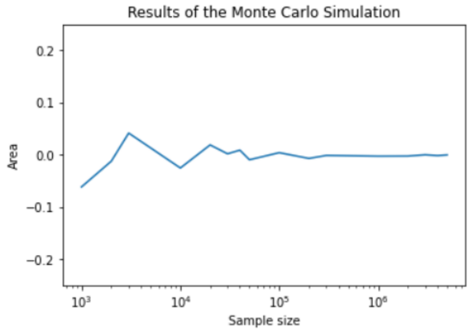
$$J = -Z_1 e^{-\frac{(Z_1^2+Z_2^2)}{2}} \cdot \frac{1}{2\pi} \cdot \frac{Z_1}{Z_1^2+Z_2^2} - (-Z_2 e^{-\frac{(Z_1^2+Z_2^2)}{2}}) \cdot \frac{1}{2\pi} \cdot \frac{-Z_2}{Z_1^2+Z_2^2}$$
$$J = -\frac{1}{2\pi} e^{-\frac{(Z_1^2+Z_2^2)}{2}} \frac{(Z_1 Z_1 + Z_2 Z_2)}{Z_1^2+Z_2^2}$$
$$|J| = \frac{1}{2\pi} e^{-\frac{(Z_1^2+Z_2^2)}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_2^2}{2}}$$

$$f[w_1(Z_1,Z_2), w_2(Z_1,Z_2)] = 1 \quad \text{(uniform distribution between 0 and 1)}$$
$$g(Z_1,Z_2) = 1 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_1^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{Z_2^2}{2}}$$

According to the Theorem 7.4, final result is equal to the joint probability distribution of Z_1 and Z_2. Since, Z_1 and Z_2 are independent random variables, $g(Z_1, Z_2)$ is equal to the multiplication of probability distribution functions of two standard normal random variable. We mathematically proved that Z_1 and Z_2 in Box-Muller's transformation are normally distributed.

4. (a) We have to integrate cos(x) from 0 to π. Expected result is zero. I've implemented Monte Carlo Simulation for different sample sizes. As it can be seen from the graph below, accuracy of the results gets better and better as sample size increases. Exact results with 4 decimal point precision are printed. Two output examples can be found below:





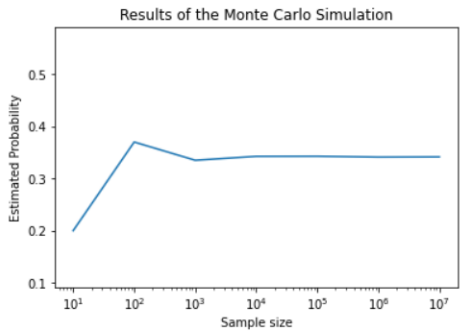
(b) Expected result of the probability $P(0 \leq Z \leq 1)$ is calculated below.

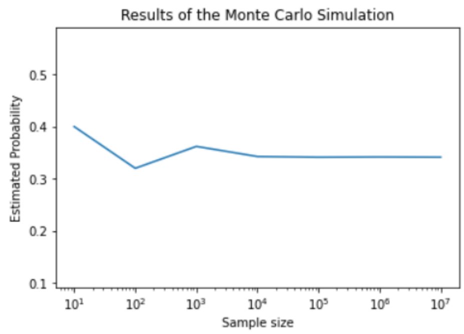
$$P(0 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq 0) = 0.8413 - 0.5000 = 0.3413$$

Monte Carlo Simulation is implemented for different sample sizes. Two output examples can be found below. As input size increases, accuracy of the results get better and better.

```
Estimated probability: 0.2000 with size 10
Estimated probability: 0.3700 with size 100
Estimated probability: 0.3350 with size 1000
Estimated probability: 0.3425 with size 10000
Estimated probability: 0.3427 with size 100000
Estimated probability: 0.3411 with size 1000000
Estimated probability: 0.3414 with size 10000000
```

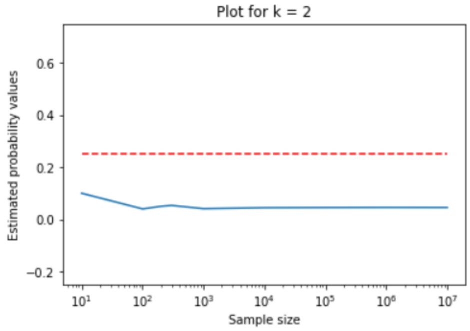
```
Estimated probability: 0.4000 with size 10
Estimated probability: 0.3200 with size 100
Estimated probability: 0.3620 with size 1000
Estimated probability: 0.3426 with size 10000
Estimated probability: 0.3412 with size 100000
Estimated probability: 0.3417 with size 1000000
Estimated probability: 0.3412 with size 10000000
```



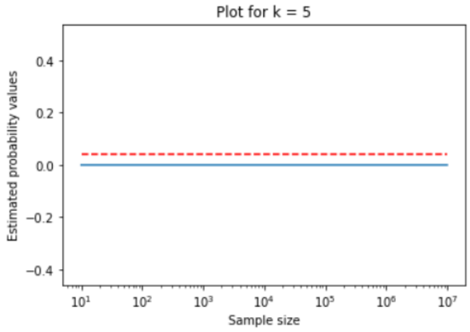


(c) According to the Chebyshev's Inequality, for values of $k = 2, 5, 10$, we expect the right hand side to be 0.25, 0.04 and 0.01 respectively. For a random variable $X \sim \mathcal{N}(0, 1)$, inequality reduces to: $P(|X| \geq k) \leq \frac{1}{k^2}$. Results of the Monte Carlo Simulation for different input sizes given below:

```
Right hand side of the inequality = 0.25
Estimated probability : 0.1000 with size 10
Estimated probability : 0.0400 with size 100
Estimated probability : 0.0500 with size 200
Estimated probability : 0.0533 with size 300
Estimated probability : 0.0410 with size 1000
Estimated probability : 0.0446 with size 10000
Estimated probability : 0.0453 with size 100000
Estimated probability : 0.0457 with size 1000000
Estimated probability : 0.0455 with size 10000000
```




```
Right hand side of the inequality = 0.04
Estimated probability : 0.0000 with size 10
Estimated probability : 0.0000 with size 100
Estimated probability : 0.0000 with size 200
Estimated probability : 0.0000 with size 300
Estimated probability : 0.0000 with size 1000
Estimated probability : 0.0000 with size 10000
Estimated probability : 0.0000 with size 100000
Estimated probability : 0.0000 with size 1000000
Estimated probability : 0.0000 with size 10000000
```




```
Right hand side of the inequality = 0.01
Estimated probability : 0.0000 with size 10
Estimated probability : 0.0000 with size 100
Estimated probability : 0.0000 with size 200
Estimated probability : 0.0000 with size 300
Estimated probability : 0.0000 with size 1000
Estimated probability : 0.0000 with size 10000
Estimated probability : 0.0000 with size 100000
Estimated probability : 0.0000 with size 1000000
Estimated probability : 0.0000 with size 10000000
```