Assignment A2

MASD 2018
Department of Computer Science
University of Copenhagen

Casper Lisager Frandsen <fsn483@alumni.ku.dk>

Version 1
Due: Dec. 11th, 23:59

Contents

1. Exercise

(a)

To find the correct value for c, we simply have to look at the definition of a uniform distribution;

$$f(x) = \frac{1}{b-a}, a \leq x \leq b, zero\ otherwise$$

We can use this to write out c:

$$c = \frac{1}{(x_{max} - x_{min}) \cdot (y_{max} - y_{min})}$$

(b)

To compute the likelihood we have to find the joint pdf: $f_{\theta}^{joint}(x_1, \ldots, x_n)$. For this specific distribution, the joint pdf can be described as thus:

$$\theta_1 = \prod_{i=1}^{4} \frac{1}{(4+1) \cdot (3+1)} = \prod_{i=1}^{4} \frac{1}{20} = \frac{1}{160000}$$

$$\theta_2 = \prod_{i=1}^{4} \frac{1}{(5+2) \cdot (6+3)} = \prod_{i=1}^{4} \frac{1}{63} = \frac{1}{15752961}$$

(c)

To find the maximum likelihood estimator, $MLE\hat{\theta}^{ML} = (\hat{x}_{max}, \hat{x}_{min}, \hat{y}_{min}, \hat{y}_{max})$, we have to look at the given values, $((0,0),(0,1),(1,1),(2,2))$. Our best estimate is then the extremes of the given values:

$$MLE\hat{\theta}^{ML} = \Big(min(x_1, \ldots, x_n), max(x_1, \ldots, x_n), min(y_1, \ldots, y_n), max(y_1, \ldots, y_n)\Big)$$

$$MLE\hat{\theta}^{ML} = (0,2,0,2)$$

We can see that this estimate has a bias for all values, since the chance that the stars in the most extreme positions are not necessarily likely to appear at the very edges of the window. We can thus say that the *min* values have a bias where estimate is likely to be larger than the actual value, and the *max* values are likely to be smaller than the actual values, but we don't know by how much and because of this can't correct for it.

2. Exercise

(a)

To find the confidence interval we have to solve the following for $\hat{\mu}$: We define the following:

$$z_a = \frac{\alpha}{2}$$
$$z_b = 1 - \frac{\alpha}{2}$$

Then, we solve the following for $\hat{\mu}$:

$$z_a \leq \sqrt{n}\frac{\hat{\mu} - \mu}{\hat{\sigma}} \leq z_b$$
$$\frac{\hat{\sigma} z_a}{\sqrt{n}} \leq \hat{\mu} - \mu \leq \frac{\hat{\sigma} z_b}{\sqrt{n}}$$
$$\frac{\hat{\sigma} z_a}{\sqrt{n}} + \mu \leq \hat{\mu} \leq \frac{\hat{\sigma} z_b}{\sqrt{n}} + \mu$$

And thus, we have found the confidence interval:

$$\left[[\frac{\hat{\sigma} z_a}{\sqrt{n}} + \mu, \frac{\hat{\sigma} z_b}{\sqrt{n}} + \mu \right]$$

(b)

326

(c)

186

3. Exercise

(a)

To complete this, we have to rewrite the given random variables into a convergence of probability:

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \text{ for } n \to \infty, \text{ and } \varepsilon > 0$$

To show that adding a constant c doesn't change this convergence, we show the following:

$$\mathbb{P}(|(X_n + c) - (X + c)| > \varepsilon) \to 0$$

In this case we can simply remove the parentheses and see that $c - c = 0$, and thus still converges to 0:

$$\mathbb{P}(|X_n - X + c - c| > \varepsilon) \to 0$$

Now we have to show that the same holds for multiplying by a constant a:

$$\mathbb{P}(|aX_n - aX| > \varepsilon) \to 0$$

We rewrite this a bit for clarity:

$$\mathbb{P}(|a(X_n - X)| > \varepsilon) \to 0$$

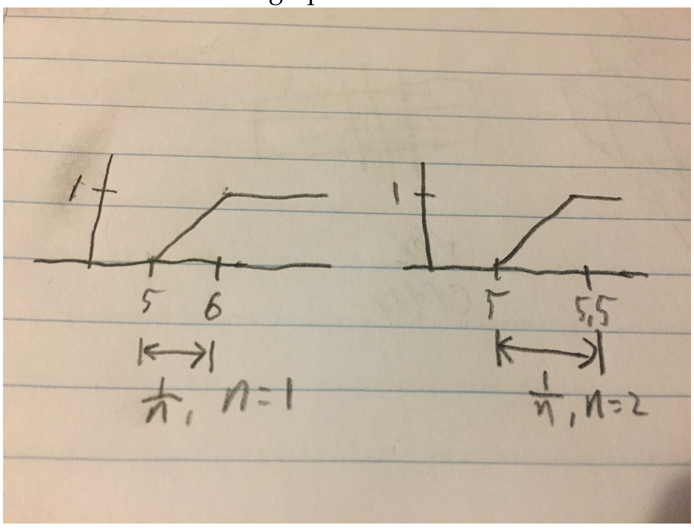
We can then divide both sides of the quality by a:

$$\mathbb{P}(|X_n - X| > \frac{\varepsilon}{a}) \to 0$$

Now, even though the ε is smaller we can see that it still converges to 0, but at $\frac{1}{a}$ the rate as when there is no constant multiplier. We can intuitively understand this, by looking at the absolute values of $|X_n - X|$, which would be a times larger at any given point in the series, up to ∞.

(b)

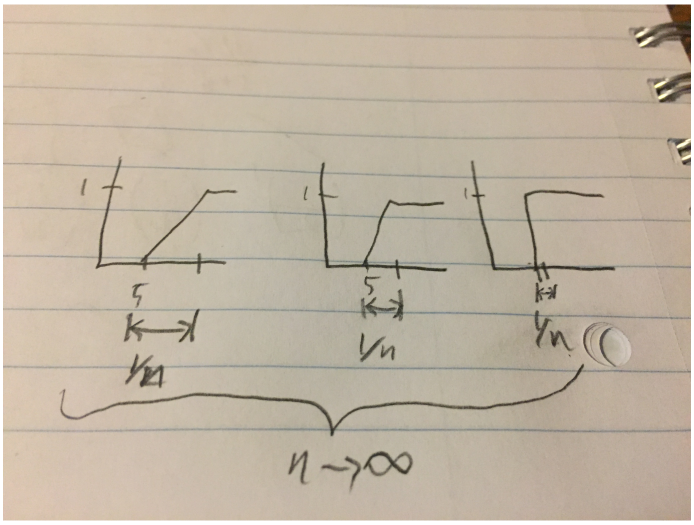
First we will draw the graph for $n = 1$ and $n = 2$:



Then, we will write down our assumptions regarding the exercise text since it is not very precise:
We have to show that the following is true:

$$F_n(X) \xrightarrow{\mathcal{L}} F(X) = 1, \textit{ for } X = 5$$

We can prove that simply by looking at the graphs, and noticing what happens. We can see that the distance between the two points on the x-axis, 5 and $5 + \frac{1}{n}$ approaches zero as n approaches infinity. This means that the value at $x = 5$ also approaches 1:



4. Exercise

(a)

The null hypothesis for this experiment would be that there is no effect on the flowering time. This means that we by default assume that the flowering time is unchanged, and we would have to prove otherwise, beyond reasonable doubt. The reason for this is that it is better to erroneously claim that you have not found a positive result than it is to erroneously claim that you have.

(b)

To perform the t-test, we have to follow the six steps. For this we will define the sequence $((X_1 - Y_1), \ldots, (X_n - Y_n))$
First, we write the model. This is given as a normal distribution:

$$T \sim \mathcal{N}(\mu, \sigma^2)$$

Then we write down the hypothesis. This hypothesis must be the case since if there is no difference between the distributions, they will have the same mean.

$$H_0 : \mu = 0, H_1 : \mu \neq 0$$

Now we write down the test:

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\sigma}$$

We are given the significance level $\alpha = 0.05$ We calculate the rejection region to be $\mathcal{R} = (-\infty, z_{0.05}] \cup [z_{0.95}, \infty) = (-\infty, -2.776] \cup [2.776, \infty)$ Finally we compute the test statistic with the data, using python. We have done this using python in the attached Jupyter Notebook:

$$T = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} = 1.809$$

This is not $\in \mathcal{R}$, which means the test passes.

(c)

Yes, he can definitely change the outcome of the test, since a "larger" dataset would look less like an outlier when computing the t-test. While the small dataset may be a random outlier, he can make it appear as if a much larger dataset, which is much less likely to have this random discrepancy.