Inferential Statistics Project - Aviv Gelfand & Maor Moshe

Recall that in the previews project we used WHO table of Average BMI and life expectancy for 177 countries in the world. Source.

Predicted / dependent variable Y: Life expectancy - refers to average number of years a person can expect to live in a country.

Undependable variable X: BMI - Body Mass Index units. A calculation using a person's height and weight.

We hypothesized that BMI is an explanatory variable for a person's life expectancy.

Question 1) Point Estimators

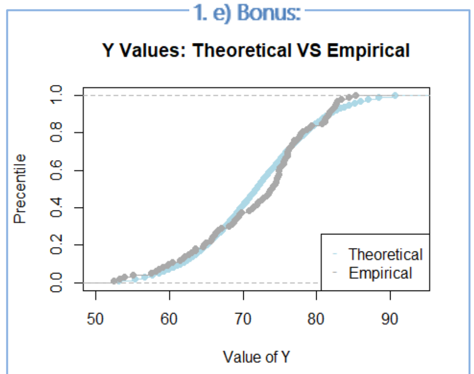
1. a) Estimating E[Y] and Var(Y)

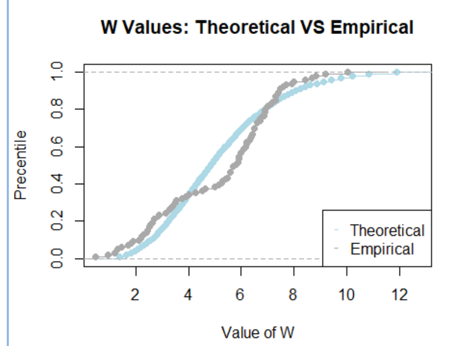
Let's assume that the predicted value, the average life expectancy in the country, is normally distributed. We define an estimator for Y according to the method of MLE or the moment's method (the same, a result in a normal distribution): $\widehat{\mu_y} := \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ = 71.82. Moreover, we will define a variance estimator according to the MLE method: $\widehat{\sigma_y^2} := \overline{Y^2} - \overline{Y}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2$ = 64.89

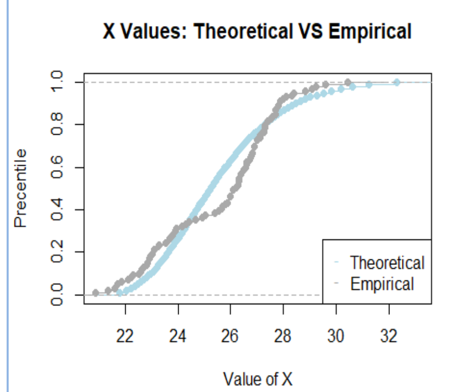
In addition, we learned in class about a conventional unbiased estimator for the variance of a normal distribution: $\widehat{\sigma_y^2} := S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2$ = 65.26

Both of the estimators are unbiased. We can see that moment estimator has smaller variety and hence smaller mse rate. therefore we will proceed the assignment with mom astimator.

1. e) Bonus:







1. b) Estimating Expectency and Variance of X using Gamma

We denote the independent variable, BMI as X, then let $m = min(X)$ and Let $W := X - m$ and assume $W \sim Gamma(\alpha, \lambda)$. Then, by the formulas we prooved in class, the point estimator of α would be:

$\hat{\alpha} = \frac{\overline{W}^2}{\overline{W}^2 - \overline{W^2}}$ = 5.19 . The point estimator of λ would be: $\hat{\lambda} = \frac{\overline{W}}{\overline{W}^2 - \overline{W^2}}$ = 1

Thus, $W := X - m \sim Gamma(\alpha = 5.19\ , \lambda = 1\ )$

1. c-d) Theoretical and Empirical Quantiles [0.1,0.5,0.75,0.9] of X, Y and W:

	0.1	0.5	0.75	0.9
Y_Theoretical	61.471228	71.823729	77.272324	82.176229
Y_Empirical	59.900000	74.000000	77.000000	81.840000
X_Theoretical	22.967133	25.260743	26.893732	28.639136
X_Empirical	22.500000	26.200000	27.200000	27.900000
W_Theoretical	2.567133	4.860743	6.493732	8.239136
W_Empirical	2.100000	5.800000	6.800000	7.500000

Question 2 - Confidence Intervals

CI of E[Y] where confidence level is 97%

We don't know What σ_Y is. Hence, we calculate the confidence interval of an expected life expectancy in a country (our Y) as follows:

$$CI = \left[\overline{Y_n} \pm t_{n-1,1-\frac{a}{2}} \frac{S_n}{\sqrt{n}}\right] = \left[71.82 \pm t_{n=177,\, df=0.985} \frac{8.07}{\sqrt{177}}\right] = [70.49,\ 73.15]$$

CI of variance(Y) where confidence level is 92%

We calculate the confidence interval of the variance of life expectancy in a country (our Y) as follows:

$$CI = \left[\frac{(n-1)S_n^2}{\chi^2_{(n-1),1-\frac{a}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{(n-1),\frac{a}{2}}}\right] = \left[\frac{(176) * 65.25}{\chi^2_{(176),\,0.96}}, \frac{(176) * 65.25}{\chi^2_{(176),\,0.04}}\right] = [54.95, 78.94]$$

Question 3 - Hypotheses Testing

U denotes the Higher BMI world $:=$ countries where the average person has a higher BMI than the median BMI in the world.

L denotes the Lower BMI world $:=$ countries where the average person has a lower BMI than the median BMI in the world.

3. a) Verbal formulation of research hypotheses

Our null hypothesis (H_0) assumes no correlation between Life expectancy (Y) and Average BMI (X). In other words, H_0 assumes both of our examined groups are identical, such that the difference of their means is equal to zero. Our alternative hypothesis ($\boldsymbol{H_1}$) is that the higher BMI world would have a higher Life Expectancy then the lower BMI world (higher BMI where countries are richer).

3. b) Statistical formulation of our hypotheses

Let us denote $D = \overline{Y_U} - \overline{Y_L}$ as our test statistic. Then our hypotheses will be:

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D > 0$$

3. c) Building a statistical test.

Our test-statistic is $\overline{D}$. We don't know D's real variance then $D \sim t_{88}$ under H_0.

Since we don't know the real variance of D, we estimate it as $\sigma_D^2 = S_D^2 = 96$ with the above-mentioned formula.

Then: $P_{val} := p_{H_0}(Reject\ H_0) \underset{\overline{D}>0}{=} 1 - pt_{88}\left(\frac{\overline{D} - 0}{\sqrt{S_n^2/n}}\right) =$ 1.849632e-13 ≈ 0

Hence, let our test be: Reject H_0 if $P_{val} > \alpha$

This is a very small p-value, which suggests that the difference in means between the two groups is statistically significant for every α. In other words, the data provides strong evidence against H_0.

3. d) Applying our statistical test

We set our significance level as $\alpha = 0.03$ and we achieve: $P_{val} \approx 0 < \alpha = 0.03 \ \rightarrow$ We Reject H_0 with the significance level of 0.03.

Based on the p-value, we can conclude that the mean life expectancy of the "High BMI" group is significantly greater than the mean life expectancy of the "Low BMI" group.