Cooper Golemme

Question 1: Pearson Test

1.1 A

We will do a Pearson χ^2 test for independence. Under the null hypothesis, the data are independent; graduating high school has no effect on making more or less than $25,000$.

	No Graduate	Graduate
$<$ 25,000	210	79
$>$ 25,000	153	558

The test statistic U looks like

$$U = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}$$

	No Grad	Grad	Marginals
<25,000	210	79	289
>25,000	153	558	711
Marginals	363	637	1000

Plugging these values in we get

$$U = \frac{[210 - \frac{289\cdot 363}{1000}]^2}{\frac{289\cdot 363}{1000}} + \frac{[79 - \frac{637\cdot 289}{1000}]^2}{\frac{637\cdot 289}{1000}} + \frac{[153 - \frac{711\cdot 363}{1000}]^2}{\frac{711\cdot 363}{1000}} + \frac{[558 - \frac{637\cdot 711}{1000}]^2}{\frac{637\cdot 711}{1000}}$$

We can compute this U statistic and compare it to a χ^2_1 with 1 degree of freedom, evaluated at a $\alpha = 0.05$ level. If the test statistic exceeds the value for the chi-squared at 0.05, then we can reject H_0 and we conclude that high school graduation does not effect making more than $25,000$.

1.2 B

The age of the people would affect the results of the study. If some of the 1000 people interviewed were still in high school, they obviously would not have been a high school graduate and very likely would be making less than 25,000.

Location is also a very important factor. The cost of living is different in different places in the country and around the world. $25,000$ looks very different depending on where the people interviewed are from.

Question 2: Derivation of Linear Regression

2.1 Derivation

$$\text{argmin}_{\beta_1,\beta_0} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2 \tag{1}$$

$$\frac{d}{d\beta_0} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2 \tag{2}$$

$$= \sum_{i=1}^{n} \frac{d}{d\beta_0} (\beta_1 x_i + \beta_0 - y_i)^2 \tag{3}$$

$$= \sum_{i=1}^{n} 2(\beta_1 x_i + \beta_0 - y_i) \tag{4}$$

$$0 = \sum_{i=1}^{n} 2(\beta_1 x_i + \beta_0 - y_i) \qquad \text{to minimize} \tag{5}$$

$$0 = \sum_{i=1}^{n} 2\beta_1 x_i + 2\beta_0 - 2y_i) \tag{6}$$

$$0 = \sum_{i=1}^{n} 2\beta_1 x_i + \sum_{i=1}^{n} 2\beta_0 - \sum_{i=1}^{n} 2y_i \tag{7}$$

$$0 = \sum_{i=1}^{n} \beta_1 x_i + \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} y_i \tag{8}$$

$$-\sum_{i=1}^{n} \beta_0 = \sum_{i=1}^{n} \beta_1 x_i - \sum_{i=1}^{n} y_i \tag{9}$$

$$-n\beta_0 = \sum_{i=1}^{n} \beta_1 x_i - \sum_{i=1}^{n} y_i \tag{10}$$

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} \beta_1 x_i \tag{11}$$

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i \tag{12}$$

$$\beta_0 = \bar{Y}_n - \beta_1 \bar{X}_n \tag{13}$$

$$\tag{14}$$

We can use this to plug in for β_0 to minimize with respect to β_1:

$$\text{argmin}_{\beta_1,\beta_0} \sum_{i=1}^{n} (\beta_1 x_i + \bar{Y}_n - \beta_1 \bar{X}_n - y_i)^2 \tag{15}$$

$$\frac{d}{d\beta_1} \sum_{i=1}^{n} (\beta_1 x_i + \bar{Y}_n - \beta_1 \bar{X}_n - y_i)^2 \tag{16}$$

$$\sum_{i=1}^{n} \frac{d}{d\beta_1} (\beta_1 x_i + \bar{Y}_n - \beta_1 \bar{X}_n - y_i)^2 \tag{17}$$

$$\sum_{i=1}^{n} \frac{d}{d\beta_1} (\beta_1 x_i - \beta_1 \bar{X}_n + \bar{Y}_n - y_i)^2 \tag{18}$$

$$0 = \sum_{i=1}^{n} 2(\beta_1 x_i - \beta_1 \bar{X}_n + \bar{Y}_n - y_i)(x_i - \bar{X}_n) \quad \text{to minimize} \tag{19}$$

$$0 = \sum_{i=1}^{n} 2(\beta_1 (x_i - \bar{X}_n) + \bar{Y}_n - y_i)(x_i - \bar{X}_n) \tag{20}$$

$$0 = \sum_{i=1}^{n} 2(\beta_1 (x_i - \bar{X}_n)^2 + \bar{Y}_n (x_i - \bar{X}_n) - y_i (x_i - \bar{X}_n)) \tag{21}$$

$$0 = \sum_{i=1}^{n} \beta_1 (x_i - \bar{X}_n)^2 + \sum_{i=1}^{n} \bar{Y}_n (x_i - \bar{X}_n) - \sum_{i=1}^{n} y_i (x_i - \bar{X}_n) \tag{22}$$

$$-\beta_1 \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 = \sum_{i=1}^{n} \bar{Y}_n (x_i - \bar{X}_n) - \sum_{i=1}^{n} y_i (x_i - \bar{X}_n) \tag{23}$$

$$-\beta_1 \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 = \sum_{i=1}^{n} \bar{Y}_n (x_i - \bar{X}_n) - y_i (x_i - \bar{X}_n) \tag{24}$$

$$-\beta_1 \sum_{i=1}^{n} (x_i - \bar{X}_n)^2 = \sum_{i=1}^{n} (\bar{Y}_n - y_i)(x_i - \bar{X}_n) \tag{25}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{Y}_n)(x_i - \bar{X}_n)}{\sum_{i=1}^{n} (x_i - \bar{X}_n)^2} \tag{26}$$

Thus,

$$\beta_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{Y}_n)(x_i - \bar{X}_n)}{\sum_{i=1}^{n} (x_i - \bar{X}_n)^2} \quad \textit{and} \quad \beta_0 = \bar{Y}_n - \beta_1 \bar{X}_n$$

2.2 Data changes

Suppose the data point (9, 5) is replaced with (9, 50). At a descriptive level, how will this change the solution in (a)? Explain.

Answer:

β_1 would increase i.e. the slope of the line of best fit would increase. This is because the numerator of β_1 term would increase while the denominator would stay the same as the x values are not changing. The only things that would change would be the final y_i value and the average $\bar{Y}_n$.

b_0 would likely change as well as it is dependent on both β_1 and the averages of the y values. This is the intercept of the line.

Overall, the line will become much steeper and the intercept will change. This will result in generally a worse fitting line for the majority of the data as the least squares linear regression is very sensitive to outliers.

Question 3: Histogram Estimator

3.1 A

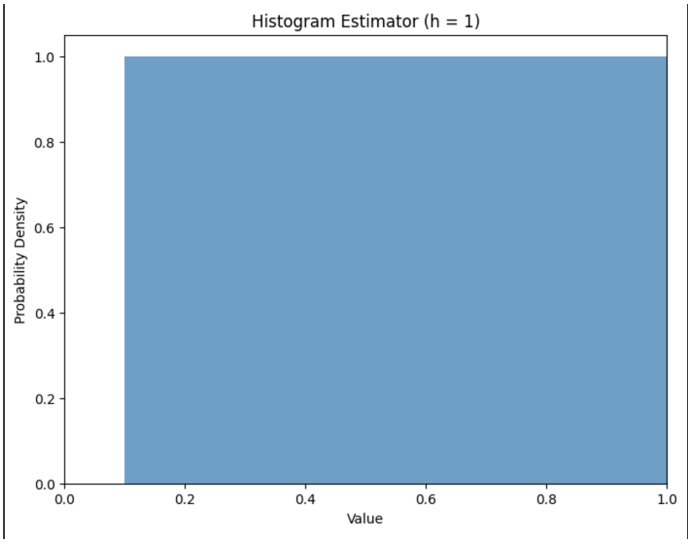

Figure 1: Histogram estimator with 1 Bin Size

3.2 B

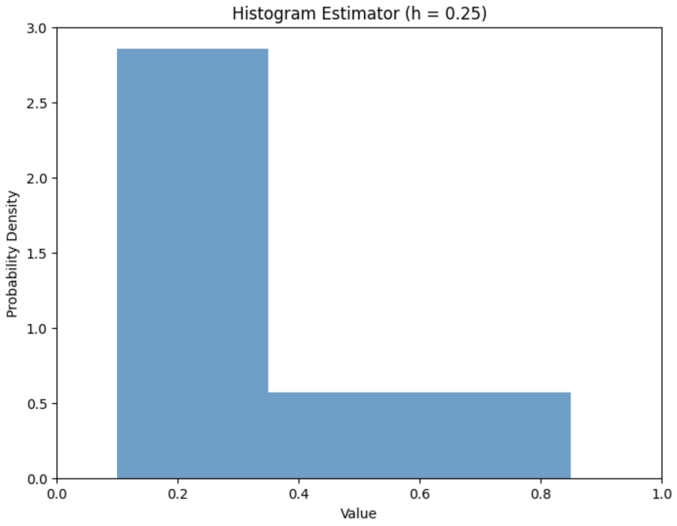

Figure 2: Histogram estimator with h=0.25

3.3 C

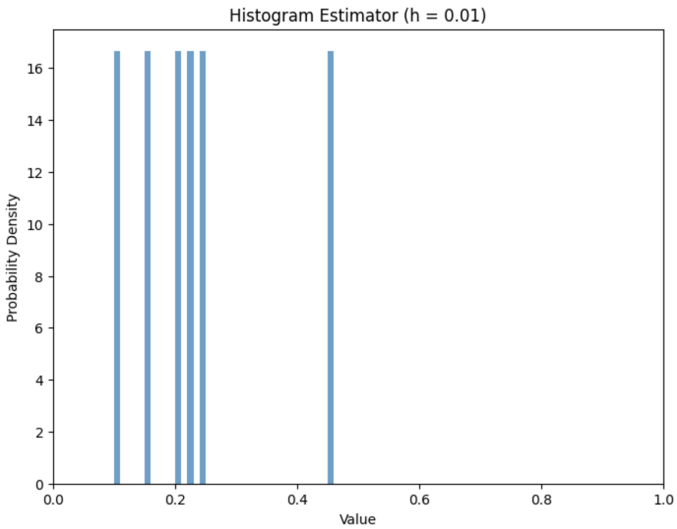

Figure 3: Histogram with 0.01 Bin Size

3.4 Conclusions:

In light of (a)-(c), discuss the role of bin size in histogram estimation. Which of the above makes the most sense to you? Why?

Answer:

Bin sizes play an important role in histogram estimation. Firstly, the bin size determines the granularity of the estimator. A large bin size flattens out the variations in the underlying density of the data and makes the data appear more uniform. Large bin sizes remove the complexities of the underlying data. Small bin sizes amplify small deviations based on the random nature of data. You can see this in the graph as with the smaller bin sizes there are higher peaks and sparse regions between peaks.

In the above example, I think that the 0.25 bin size is the most convincing. It shows that most of the data falls withing 0.1 -0.4 roughly speaking and the rest is kind of evenly distributed over the rest of the range of the data. (a) shows that the data is exactly uniform; all of the positions that the data fell in were equally likely to occur, which does not fit with the provided data being skewed toward the lower range. In (c), the bin size is too small and the small variations of the data are amplified. The resulting density estimation is unconvincing; it is highly unlikely that the underlying distribution makes it impossible to generate data from 0.25 to 0.4 for example.

Overall, the 0.25 bin size yields the best estimate that makes sense with the data sampled and the our assumptions about the underlying density of the random process.