1930026143 薛劭杰

EBIS 3103 Introduction to Business Data Analytics - Individual Assignment 2

Instructor: Dr. Jingzhi Zhang

*For all the questions below, please use R to help you answer the questions if necessary.

1. A company has at most 6 days of delivery of a good. The fastest delivery time is 1 day, and the probabilities for the different delivery times are shown in the table below. Let *X* be the number of days of delivery. [15 marks]

Delivery time in days	1	2	3	4	5	6
Probability in %	55	20	10	5	5	5

(a) Find the cumulative distribution *F(x)* of *X*. How can we interpret *F(x)* in this case?
(b) Find the expected days of delivery *E[X]=μ.*

Ans:
(a) Initially, we know the pmf of this data, then we can calculate the cdf of it.

$F(1) = f(X \le 1) = 0.55$
$F(2) = f(X \le 2) = 55\% + 20\% = 0.75$
$F(3) = f(X \le 3) = 55\% + 20\% + 10\% = 0.85$
$F(4) = (X \le 4) = 55\% + 20\% + 10\% + 5\% = 0.9$
$F(5) = f(X \le 5) = 55\% + 20\% + 10\% + 5\% + 5\% = 0.95$
$F(6) = f(X \le 6) = 55\% + 20\% + 10\% + 5\% + 5\% + 5\% = 1$

(b) $E(x) = \sum_{i=1}^{6} x_i f(x_i) = 1 * 55\% + 2 * 20\% + 3 * 10\% + 4 * 5\% + 5 * 5\% + 6 * 5\% = 2$

2. Let X denote the income (in USD) of a randomly selected person. We have made 25 independent observations and found

$$\overline{X} = 35{,}600, \qquad S_X^2 = 441{,}000{,}000.$$

(a) Assume that X is approximately normal and find a 95% confidence interval for *E[X]=μ*. [15 marks]

Ans:
(a) We know that the number of independent observations $n = 25$, then we can use the $t-$

$distribution$ because we the σ is unknown and X is approximately normal.

$$T = \frac{\overline{X} - \mu}{S[\overline{X}]} = \frac{\overline{X} - \mu}{S_x/\sqrt{n}}$$

And the degree of freedom $df = 25 - 1 = 24$, the confidence interval is 95%. According to the $T - test\ table$, we can get the t of $P(T_{24} \geq z) = \frac{1-95\%}{2} = 2.5\%$.

n'	P(1):单側	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
1		1	3.078	6.314	12.706	31.821	63.657	127.321	318.309
2		0.816	1.886	2.92	4.303	6.965	9.925	14.089	22.327
3		0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215
4		0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5		0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6		0.718	1.44	1.943	2.447	3.143	3.707	4.317	5.208
7		0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8		0.706	1.397	1.86	2.306	2.896	3.355	3.833	4.501
9		0.703	1.383	1.833	2.262	2.821	3.25	3.69	4.297
10		0.7	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11		0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12		0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.93
13		0.694	1.35	1.771	2.16	2.65	3.012	3.372	3.852
14		0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15		0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16		0.69	1.337	1.746	2.12	2.583	2.921	3.252	3.686
17		0.689	1.333	1.74	2.11	2.567	2.898	3.222	3.646
18		0.688	1.33	1.734	2.101	2.552	2.878	3.197	3.61
19		0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20		0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21		0.686	1.323	1.721	2.08	2.518	2.831	3.135	3.527
22		0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23		0.685	1.319	1.714	2.069	2.5	2.807	3.104	3.485
24		0.685	1.318	1.711	$\boxed{2.064}$	2.492	2.797	3.091	3.467
25		0.684	1.316	1.708	2.06	2.485	2.787	3.078	3.45
26		0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27		0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421

We can get $z = 2.064$, so the interval of 95% confidence is:

$$\overline{X} \pm t * \left(\frac{S_x}{\sqrt{n}}\right) = 35600 \pm 2.064 * \left(\frac{21000}{5}\right) = [26931.2,\ \ 44268.8]$$

3. A tele-communication company's past records indicate that individual customers pay on average $220 per month for local data usage. A random sample of 15 customers' local data usage bills during a particular month produced the following amounts:

260	180	290	170	300	210	320	240	280	250	150	270	350	230	200

(a) Comment on the distribution of the above sample data. Does it appear to follow normal distribution? [15 marks]

(b) At 5% level of significance, is there enough evidence to reveal that the average amount of bills for local data usage per month is high than $250 using both the critical value approach and the p-value approach to test the hypotheses. What assumption has to be made? [15 marks]

Ans:

(a) Firstly, we can use the Shapiro-Wilk test to check whether it follow the normal distribution by using r language:

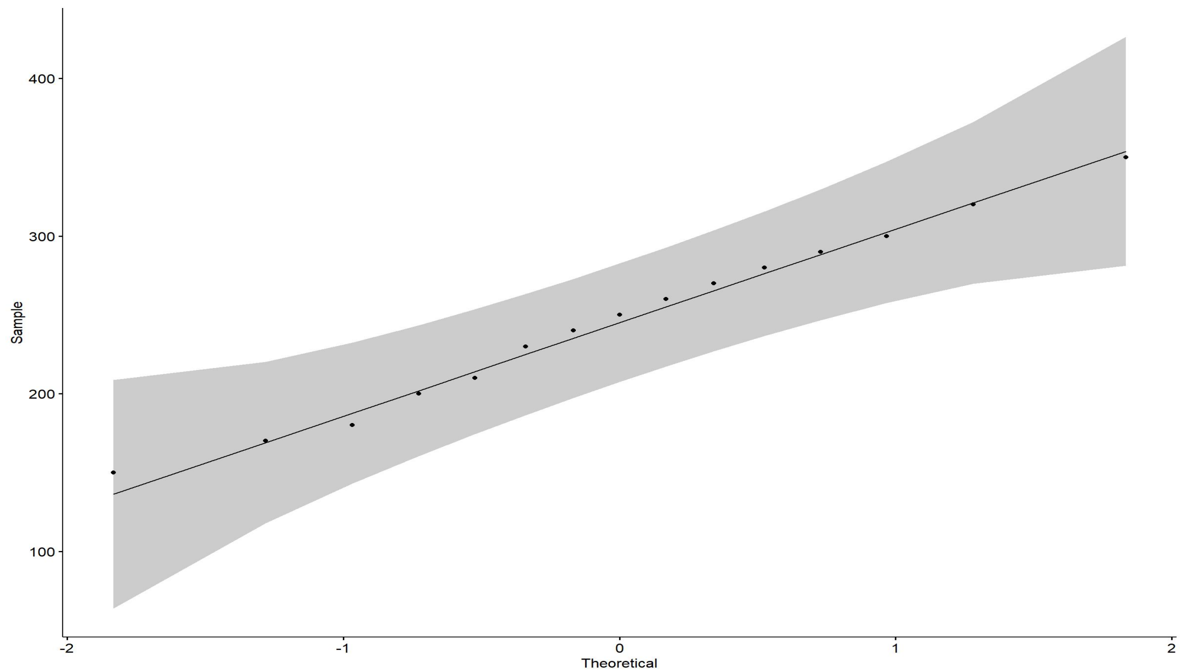
```
> shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.98617, p-value = 0.9954
```

We can see that the p-value is close to 1 which means that it is hard to reject the null hypothesis and we have strong evidence that it is normal distribution.

In addition, we can use visual method to judge the distribution by $qq-plot$, we can find that all of points roughly on that reference line which means that it have strong evident that it follows the normal distribution.



(b) 1. Critical Value: From the data, we can get the mean and variance of it.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{260 + 284 + \dots + 230 + 200}{15} \approx 246.6667$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} \approx 57.4042$$

So we can set : $H_0 : \mu \leq 250, H_1 : \mu > 250$ and calculate the T distribution:

$$T = \frac{\overline{X} - \mu}{S[\overline{X}]} = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{246.6667 - 250}{57.4042/\sqrt{15}} \approx -0.2249$$

At 5% level of significance ($\alpha = 0.05$), we can find the value of $t_{0.05,\ 14} = 1.761$.
Since $t_{0.025,\ 14} > T$, we cannot reject the null hypothesis (H_0).

2. $p - Value$: we calculate the p-value by the function $pnorm()$:

```
> pnorm(250, mean(X), sd(X) / sqrt(15))
[1] 0.5889698
```

We can find the result is larger than 0.05 which means that we cannot reject the H_0. We have no evidence to support that the average amount of bills is larger than 250.

4. The retail manager of a supermarket chain wants to determine whether product location has an effect on the sale of pet toys. Three different aisle locations are considered: front, middle, and rear. A random sample of 18 stores is selected with 6 stores randomly assigned to each aisle location. The size of the display area and price of the product are constant for all stores. At the end of a 1-month trial period, the sales volumes (in thousands of dollars) of the product in each store were reported as follows:

	Aisle Location	
Front	Middle	Rear
8.6	3.4	4.6
7.2	2.4	6.0
5.4	2.0	4.0
6.2	1.4	2.8
5.0	2.0	2.2
4.0	1.6	2.8

(a) At the 0.05 level of significance, is there evidence of a significant difference in mean sales among the various aisle locations? [20 marks]

(c) What should the retail manager conclude? Fully describe the retail manager's options with respect to aisle locations. [20 marks]

Ans:
(a) We set the $H_0: \mu_F = \mu_M = \mu_R,\ H_1: not\ all\ of\ the\ \mu\ are\ the\ same$
Initially, we can set the type of location number be c, and the number of sample in each group be n. Then we can calculate the mean and variance of each group:
Front: $\bar{x}_F = \frac{1}{n}\sum_{i=1}^{n} x_{Fi} \approx 6.0667$, $S_F^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{F_i} - \overline{X_F}\right)^2 \approx 2.7146$.
Middle: $\bar{x}_M = \frac{1}{n}\sum_{i=1}^{n} x_{Mi} \approx 2.1333$, $S_M^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{M_i} - \overline{X_R}\right)^2 \approx 0.5067$.
Rear: $\bar{x}_R = \frac{1}{n}\sum_{i=1}^{n} x_{Ri} \approx 3.7333$, $S_R^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{R_i} - \overline{X_R}\right)^2 \approx 2.0107$.
And the mean of the groups: $\overline{\overline{X}} = \frac{\bar{x}_F + \bar{x}_M + \bar{x}_R}{c} = 3.9778.$

Then we can get the SSA, MSA and the SSW, MSW:

$$SSA := \sum_{j=1}^{c} n_j\left(\bar{x}_j - \bar{\bar{X}}\right)^2 \approx 46.9529 \qquad MSA := \frac{SSA}{c-1} \approx 23.4764$$

$$SSW := (n-1)\sum_{j=1}^{c} S_j^2 = 26.16 \qquad MSW := \frac{SSW}{c(n-1)} = 1.744$$

And then we can get the F test:

$$F = \frac{MSA}{MSW} \approx 13.4585$$

The degree of freedom of MSA and MSW are $c-1$ and $n-c$ respectively, then we can check the $F - test\ table$:

/	df_1=1	2	3	4	5	6	7	8	9	10	12
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	243.9060
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4125
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753

$F_{0.05}(2,3) = 9.5521 < F$ which means that we can reject the H_0 and not all of the μ are same.

(b) From the above analysis, we can infer that not all of the μ are same. And we can get the mean of sale for each group, it is clearly that the sale in the front location is the best, which means that we can put the most important thing to sell here such as products whose shelf life is nearing its end, products with short shelf life like foods as well as popular products. By contrast, the sale in the rear and the middle location are relatively lower, we can put some items that can be kept for a long time there.