Statistics for Data Analytics

Continuous Assessment

Lecturer: Dr Shahram Azizi Sazi

Student Name: Noel Linnane
Student Number: 10389479

Contents

Question 1

In a financial network, an agent works properly with p=0.8. Let us assume that 5 agents work in this network.

a) Define X to be the possible numbers of agents who properly work, compute the probability table for X.
b) What is P(X>4)?
c) Find the expectation and variance X.

Solution – Part A

n = 5
p = 0.8
q = 1-p = 0.2
Possible values for X: {0,1,2,3,4,5}

Probability for binomial random variables:

$$P(x) = \frac{n!}{x!\,(n-x)!}p^x q^{n-x}$$

Find the probabilities for each possible value of X:

$$P(0) = \frac{5!}{0!\,5!}(0.8)^0(0.2)^5 = (1)(1)(0.0032) = 0.00032$$

$$P(1) = \frac{5!}{1!\,4!}(0.8)^1(0.2)^4 = (5)(0.8)(0.0016) = 0.0064$$

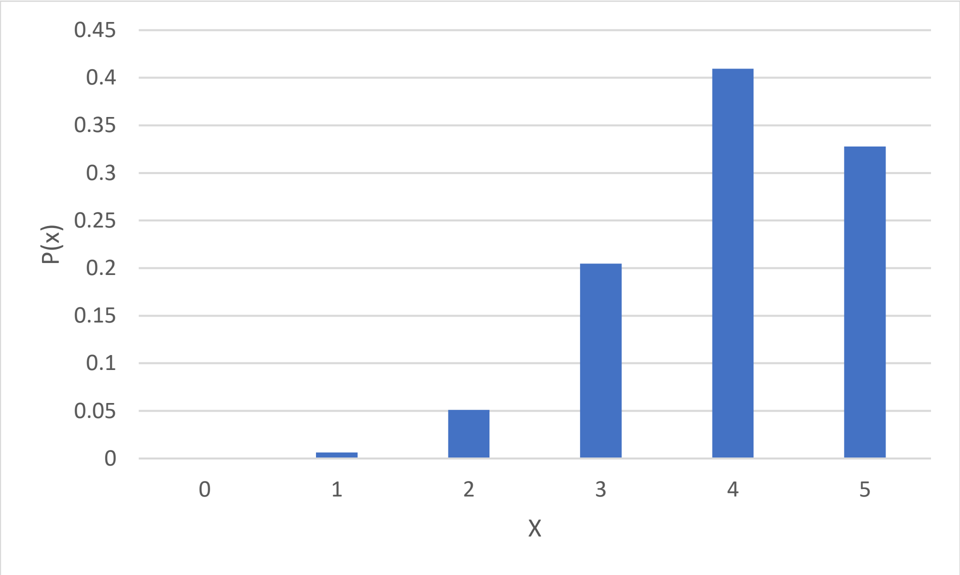
$$P(2) = \frac{5!}{2!\,3!}(0.8)^2(0.2)^3 = (10)(0.64)(0.008) = 0.0512$$

$$P(3) = \frac{5!}{3!\,2!}(0.8)^3(0.2)^2 = (10)(0.512)(0.04) = 0.2048$$

$$P(4) = \frac{5!}{4!\,1!}(0.8)^4(0.2)^1 = (5)(0.4096)(0.2) = 0.4096$$

$$P(5) = \frac{5!}{5!\,0!}(0.8)^5(0.2)^5 = (1)(0.32768)(1) = 0.32768$$

Probability Table for X:



x	0	1	2	3	4	5
P(x)	0.00032	0.0064	0.0512	0.2048	0.4096	0.32768

Solution - Part B

P(X>4) = P(5) = 0.32768

Solution – Part C

Find the expectation of X.

Expectation of X (or Mean):

$$\mu = \boldsymbol{E}(\boldsymbol{x}) = \sum \boldsymbol{x}\boldsymbol{P}(\boldsymbol{x})$$

(0x0.00032) + (1x0.0064) + (2x0.0512) + (3x0.2048) + (4x0.4096) + (5x0.32768) = 4

For binomial random variables this equation can be shortened to:

$$\mu = \boldsymbol{np}$$

(5)(0.8) = 4

The expectation of X is 4

The Variance of X

$$\boldsymbol{\sigma}^2 = \sum (\boldsymbol{x} - \mu)^2 \boldsymbol{P}(\boldsymbol{x})$$

For binomial random variable this can be shortened to:

$$\boldsymbol{\sigma}^2 = \boldsymbol{npq}$$

(5)(0.8)(0.2) = 0.8

Question 2

A manufacturing process produces ball bearings with diameters that have a normal distribution with known standard deviation of .04 centimetres. Ball bearings with diameters that are too small or too large are undesirable. In order to test the claim that μ=0.50 centimetres, perform a two-tailed hypothesis test at the 5% level of significance. Assume that a random sample of 25 gave a mean diameter of 0.51 centimetres. Perform a hypothesis test (step procedure outlined in class) and state your decision.

Solution

Step 1: State the hypotheses

H_0: μ = 0.50
H_1: μ ≠ 0.50

Step 2: State the level of significance

$\alpha = 0.05$

Step 3: Compute the test value

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{0.51 - 0.50}{0.04/\sqrt{25}} = 1.25$$

Step 4: Find the critical value
n < 30, but σ is known, use standard normal distribution.

$\alpha = 0.05$
$\alpha/2 = 0.025$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Critical value = 1.96

Step 5: Decision
If |test value| > critical value then reject H_0.
1.25 < 1.96, do not reject H_0.

Conclusion
As the test value is less than the critical value, there is not sufficient information to reject H_0.

Question 3

A specific price dataset is analysed and, the summary of ANOVA table is given as follows:

Oneway

Descriptives

PRICE

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
1.00	5	81.2000	29.8781	13.3619	44.1015	118.2985	40.00	120.00
2.00	5	77.0000	20.5061	9.1706	51.5383	102.4617	59.00	110.00
3.00	5	55.4000	13.5019	6.0382	38.6352	72.1648	40.00	73.00
Total	15	71.2000	23.7523	6.1328	58.0464	84.3536	40.00	120.00

ANOVA

PRICE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1916.400	2	958.200		.189
Within Groups	5982.000	12	498.500		
Total	7898.400	14			

Find the F-statistic and express your decision.

Solution

Step 1: State the hypotheses.

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: Not all population means are equal.

Step 2: State level of significance.

$\alpha = 0.05$

Step 3: Test Statistic

$$F = \frac{MST}{MSE}$$

$$MST = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{K-1}$$

$$MST = \frac{5(81.2 - 71.2)^2 + 5(77 - 71.2)^2 + 5(55.4 - 71.2)^2}{2} = 958.2$$

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_2 - 1)s_2^2}{n - K}$$

$$MSE = \frac{(4)(892.7) + (4)(420.5) + (4)(182.3)}{12} = 498.5$$

$$F = \frac{958.2}{498.5} = 1.9222$$

Step 4: Critical Value

Degrees of freedom

$df_1 = K - 1 = 2$

$df_2 = n - K = 12$

$\alpha = 0.05$

Lookup F-Distribution Table for $\alpha = 0.05$ with $df_1 = 2$ and $df_2 = 12$

/	df_1=1	2	3	4	5	6	7	8
df_2=1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452
⋮								
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669

Critical Value = 3.8853

Step 5: Decision

As the F-statistic is less than the critical value we accept the null hypothesis H_0 and can conclude at a 5% level of significance that all population means are equal.

Question 4

An opinion poll surveyed a simple random sample of 1000 students. Respondents were classified by gender (male or female) and by opinion (Reservation for women, No Reservation, or No Opinion). Results are shown in the observed contingency table below.

	Opinion on Women's Reservation			
	Yes	**No**	**Can't Say**	**Row Total**
Male	200	150	50	400
Female	250	300	50	600
Column Total	450	450	100	1000

Are gender and opinion on women's reservation independent? Use a 0.05 level of significance. To do so,

a. State the hypotheses.
b. Find the statistic value.
c. Find the critical value.
d. Explain your decision and Interpret results.

Solution

Step 1: State the Hypotheses

H_0: Gender and Opinion are independent.
H_1: Gender and Opinion are not independent.

Step 2: State the level of significance

$\alpha = 0.05$
We will do a chi-squared test for independence

Step 3: Expected Contingency Values

E = (R x C)/n

Where:

R: The row total
C: The column total
n: The sample size

1st row in 1st column	(400x450)/1000 = 180
1st row in 2nd column	(400x450)/1000 = 180
1st row in 3rd column	(400x100)/1000 = 40
2nd row in 1st column	(600x450)/1000 = 270
2nd row in 2nd column	(600x450)/1000 = 270
2nd row in 3rd column	(600x100)/1000 = 60

This gives us the below expected contingency table:

*Expected Contingency Table				
	Opinion on Women's Reservation			
	Yes	**No**	**Can't Say**	**Row Total**
Male	180	180	40	400
Female	270	270	60	600
Column Total	450	450	100	1000

Step 4: Test Statistic

$$\chi^2 = \boldsymbol{\Sigma}\frac{(\boldsymbol{O}-\boldsymbol{E})^2}{\boldsymbol{E}}$$

Where:

O = Observed Value
E = Expected Value

$$\chi^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} + \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60} = \mathbf{16.2}$$

The test statistic is 16.2

Step 5: Critical Value

Degrees of Freedom:

DF = (I-1) (J-1)

Where:

I = Number of levels in the first factor, gender.
J = Number of levels in the second factor, opinion.

DF = (2-1) (3-1) = 2

The test statistic follows a chi-squared distribution with 2 degrees of freedom.

Critical Value:

Table of the chi square distribution – Appendix J, p. 915

	Level of Significance α								
df	0.200	0.100	0.075	0.050	0.025	0.010	0.005	0.001	0.0005
1	1.642	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	3.219	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	4.642	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.731
4	5.989	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.998

The critical value is 5.991.

Step 6: Decision

If Test Value > Critical Value, then reject H_0

16.2 > 5.991. This means the test value is within the rejection region and we can reject the null hypothesis H_0.

Conclusion

There is sufficient evidence, at 5% level of significance, to conclude that gender and opinion on women's reservation are not independent.

Question 5

The delivery dataset is analysed in R and the output of Regression analysis is as follows

```
> fit <- lm(Time ~ Cases + Distance , data = delivery)
> summary(fit)

Call:
lm(formula = Time ~ Cases + Distance, data = delivery)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7880 -0.6629  0.4364  1.1566  7.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.341231   1.096730   2.135 0.044170 *  
Cases       1.615907   0.170735   9.464 3.25e-09 ***
Distance    0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared: 0.9596, Adjusted R-squared: 0.9559 
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

(a) List the assumptions for the linear regression.
(b) Using the above output, specify the response and independent variables. Find the coefficients' estimates for independent variables.
(c) Identify the significant independent variables at level α = 0.05.
(d) Provide the predictive model and find the predicted value of time where cases are two and Distance is three.

Solution

5a. List the assumptions for the linear regression

- The mean of the probability distribution of ε is 0. $E(\varepsilon_i) = 0$.
- The variance of ε is constant.
- ε has a normal distribution.
- The values of ε associated with any observed values of y are independent.

5b(i). Using the above output, specify the response and independent variables.

Independent Variables	Response Variables
Cases	Time
Distance	

5b(ii). Find the coefficients estimates for independent variables.

Cases: 1.615907
Distance: 0.014385

5c. Identify the significant independent variables at level α = 0.05

The stars on the end of coefficient estimates give an indication to their level of significance.

Intercept	One star = 95% level of significance

Cases	Three stars = 100% level of significance
Distance	Three start = 100% level of significance

Therefore these 3 are the significant independent variables.

5d. Provide the predictive model and find the predicted value of time where cases are two and Distance is three.

Predictive model:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

β_0 = 2.341231
β_1 = 1.615907
β_2 = 0.014385
x_1 = 2
x_2 = 3

$y = 2.341231 + (1.615907)(2) + (0.014385)(3) = 5.6162$

Bibliography

McClave, J. and Sincich, T. (2013). *Statistics*. Boston: Pearson.

Saylordotorg.github.io. (2018). *Introductory Statistics*. [online] Available at: https://saylordotorg.github.io/text_introductory-statistics/index.html [Accessed 1 Jul. 2018].

www.SOCR.ucla.edu, I. (2018). *F-Distribution Tables*. [online] Socr.ucla.edu. Available at: http://www.socr.ucla.edu/Applets.dir/F_Table.html#FTable0.05 [Accessed 1 Jul. 2018].

Users.stat.ufl.edu. (2018). [online] Available at: http://users.stat.ufl.edu/~athienit/Tables/Ztable.pdf [Accessed 1 Jul. 2018].

R, S. (2018). *Simple Linear Regression in R - Articles - STHDA*. [online] Sthda.com. Available at: http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/ [Accessed 1 Jul. 2018].