

PRIME COLLEGE

Khusibun, Kathmandu



LAB REPORT OF STATISTICS - I

Submitted by:

Name: Aavash Devkota

College roll no: 01

Faculty: BSc. CSIT

Semester: 2nd

Submitted to:

Subita Baidya

TABLE OF CONTENTS

Unit	Title	Date	Signature
2	Descriptive Statistics		
6	Probability Distributions		
7	Correlation and Linear Regression		

1. One of the major measures of the quality of service provided by an organization is the speed with which it responds to customer complaints. An internet service provider had undergone major improvements by recruiting well trained installation crews, supervisors and office staffs. The business objective of the company was to reduce the time between when the complaint is received and when it is resolved. During a recent month, the company received 50 complaints concerning internet installation. The data from the 50 complaints, collected by ISP, represent number of hours between the receipt and the solution of the complaint:

27, 4, 52, 30, 22, 36, 26, 20, 23, 33, 68, 165, 32, 29, 28, 29, 26, 25, 1, 14, 13, 13, 10, 5, 19, 126, 110, 110, 29, 61, 35, 94, 31, 26, 5, 12, 4, 54, 5, 35, 137, 31, 27, 152, 2, 123, 81, 74, 27, 11

- Compute the mean, median, first quartile and third quartile.
- Compute the range, interquartile range, variance, standard deviation and coefficient of variation.
- Construct a boxplot. Are the data skewed? If so, how?
- On the basis of the results of (a) through (c), if you had to tell the president of the company how long a customer should expect to wait to have a complaint resolved, what would you say? Explain

Solution:

Syntax:

DATASET ACTIVATE DataSet1.

DESCRIPTIVES VARIABLES=Time

/STATISTICS=MEAN STDDEV MIN MAX.

EXAMINE VARIABLES=Time

/PLOT BOXPLOT

/PERCENTILES (5, 10, 25, 50, 75, 90, 95) HAVERAGE

/STATISTICS DESCRIPTIVES

/NOTOTAL

OUTPUT:

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Time between complaint and resolution of problem	50	1	165	43.04	41.926
Valid N (list-wise)	50				

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Time between complaint and resolution of problem	50	100.0%	0	0.0%	50	100.0%

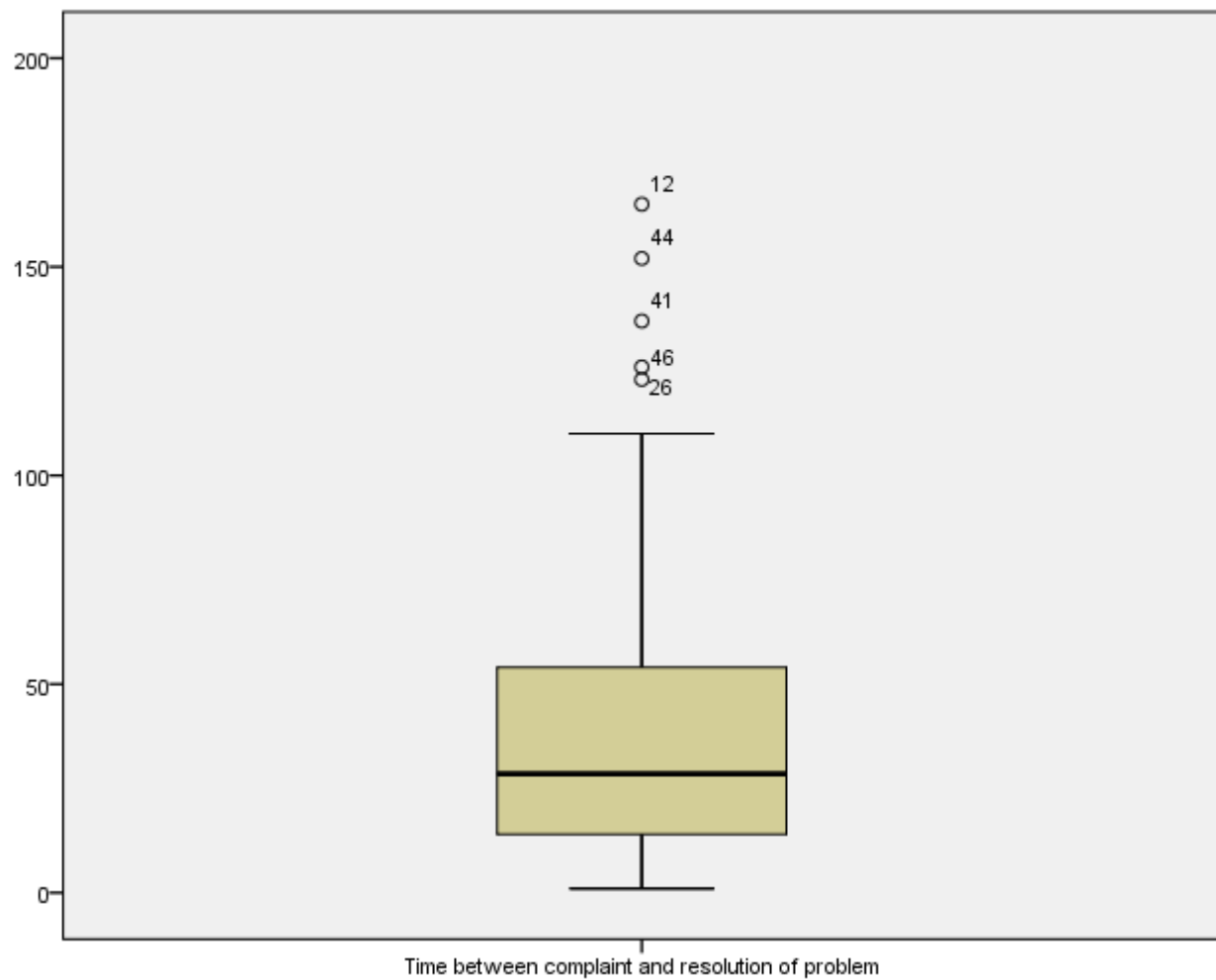
Descriptives

			Statistic	Std. Error
Time between complaint and resolution of problem	Mean		43.04	5.929
	95% Confidence Interval for Mean	Lower Bound	31.12	
		Upper Bound	54.96	
	5% Trimmed Mean		39.14	
	Median		28.50	
	Variance		1757.794	
	Std. Deviation		41.926	
	Minimum		1	
	Maximum		165	
	Range		164	
	Interquartile Range		42	
	Skewness		1.488	.337
	Kurtosis		1.309	.662

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Time between complaint and resolution of problem	3.10	5.00	13.75	28.50	55.75	121.70	143.75
Tukey's Hinges	Time between complaint and resolution of problem			14.00	28.50	54.00		

Time between complaint and resolution of problem



Conclusion:

Therefore,

a. Mean=43.04

Median= 28.5

First Quartile=13.75

Third Quartile=55.75

b. Range=164

Interquartile range=42

Variance=1757.794

Standard Deviation=41.926

Coefficient of variation=97.411%

c. Coefficient of skewness=1.488>0, so the data is positively skewed.

d. The five point summary of the data is as follows:

(1, 13.75, 28.5, 55.75, 165)

Here the minimum time required to solve the complaint is 1 hour and the maximum time required to solve the complaint is 165 hours.

So the customer should expect between 1 to 165 hours for the resolution of their complaint.

2. Fit Binomial distribution.

x	0	1	2	3	4	5	6	7	8	Total
f	5	25	35	48	65	41	28	9	4	260

Solution:

SYNTAX:

DATASET ACTIVATE DataSet1.

COMPUTE ex=PDF.BINOM(x,8,3.72/8).

EXECUTE.

COMPUTE ef=260*ex.

EXECUTE.

COMPUTE rndef=rnd(ef).

EXECUTE.

OUTPUT:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help									
10 : mdef									
	x	f	ex	ef	mdef	var	var	var	var
1	0	5	.01	1.75	2.00				
2	1	25	.05	12.13	12.00				
3	2	35	.14	36.91	37.00				
4	3	48	.25	64.16	64.00				
5	4	65	.27	69.71	70.00				
6	5	41	.19	48.47	48.00				
7	6	28	.08	21.06	21.00				
8	7	9	.02	5.23	5.00				
9	8	4	.00	.57	1.00				
10									
11									

Conclusion:

Therefore, the fitted binomial distribution is:

X	0	1	2	3	4	5	6	7	8
F	2	12	37	64	70	48	21	5	1

3. Fit the Poisson distribution.

Mistakes per page	0	1	2	3	4	5
No. of pages	142	156	69	27	5	1

Solution:

SYNTAX:

COMPUTE $p_x = \text{PDF.POISSON}(x, 1.34)$.

EXECUTE.


COMPUTE $\text{expfx} = 400 * p_x$.

EXECUTE.

COMPUTE $\text{rndef} = \text{rnd}(\text{expfx})$.

EXECUTE

OUTPUT:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help							
							
1 : rndef		105.00					
	x	f	px	expfx	rndef	var	var
1	0	142	.26	104.74	105.00		
2	1	156	.35	140.35	140.00		
3	2	69	.24	94.03	94.00		
4	3	27	.11	42.00	42.00		
5	4	5	.04	14.07	14.00		
6	5	1	.01	3.77	4.00		
7							

Conclusion:

Therefore, the fitted Poisson distribution is:

Mistakes per page	0	1	2	3	4	5
No. of pages	105	140	94	42	14	4

4. Calculate Karl Pearson's correlation coefficient and Spearman's Rank correlation coefficient test its significance. Find coefficient of determination.

Nutrition	Child Mortality
12.1	9.5
9.1	9.2
26	11.8
6.4	6.4
9.5	7.3
18.5	20.3
22.8	24.4
17.4	21.1
13.9	10.7
3.2	3.5
30.2	11.8
15.7	12.3
8.7	11.8
5.6	9.4
11.2	8.3
9.8	9
8.4	4.7

Solution:

SYNTAX:

CORRELATIONS

/VARIABLES=Nutrition ChildMortality

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE.

NONPAR CORR

```
/VARIABLES=Nutrition ChildMortality  
/PRINT=SPEARMAN TWOTAIL NOSIG  
/MISSING=PAIRWISE.
```

REGRESSION

```
/MISSING LISTWISE  
/STATISTICS R ANOVA  
/CRITERIA=PIN(.05) POUT(.10)  
/NOORIGIN  
/DEPENDENT ChildMortality  
/METHOD=ENTER Nutrition.
```

OUTPUT:

Karl Pearson's Correlations

Correlations

		Nutrition	ChildMortality
Nutrition	Pearson Correlation	1	.626**
	Sig. (2-tailed)		.007
	N	17	17
ChildMortality	Pearson Correlation	.626**	1
	Sig. (2-tailed)	.007	
	N	17	17

** . Correlation is significant at the 0.01 level (2-tailed).

Spearman's Rank Correlation

Correlations

		Nutrition	ChildMortality
Spearman's rho	Correlation Coefficient	1.000	.779**
	Nutrition	17	17
	Sig. (2-tailed)		
	N		
	Correlation Coefficient	.779**	1.000
	ChildMortality	17	17
	Sig. (2-tailed)		
	N		

** . Correlation is significant at the 0.01 level (2-tailed).

Coefficient of determination

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.626 ^a	.391	.351	4.5963

Conclusion:

Therefore,

- Karl Pearson's correlation coefficient= 0.626 and it is significant
- Spearman's Rank correlation coefficient= 0.779 and it is significant
- Coefficient of determination= 0.391

5. Omprakash Sharma, owner of the Kathmandu Precast Company, has hired you as a part-time analyst. He was extremely pleased when you uncovered a positive relationship between the number of building permits issued and the amount of work available to his company. Now he wonders if it's possible to use knowledge of interest rates on first mortgages to predict the number of building permits that will be issued each month. You collect a sample of data covering nine months.

Month	Building Permits(Y)	Interest rate (X)
1	786	10.2
2	494	12.6
3	289	13.5
4	892	9.7
5	343	10.8
6	888	9.5
7	509	10.9
8	987	9.2
9	187	14.2

- Calculate the correlation coefficient between building permits and interest rate and test its significance at 1%.
- Estimate the best fitting regression line and compute residual for month 9.
- Compute the coefficient of determination and interpret its meaning.
- Predict building permits when the interest rate increases by 9.7%.

Solution:

SYNTAX:

CORRELATIONS

/VARIABLES=Y X

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA CHANGE

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT Y

/METHOD=ENTER X

/SAVE PRED RESID.

OUTPUT:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help								
22 :								
	month	Y	X	PRE_1	RES_1	var	var	var
1	1	786	10.2	738.94861	47.05139			
2	2	494	12.6	391.07475	102.92525			
3	3	289	13.5	260.62206	28.37794			
4	4	892	9.7	811.42233	80.57767			
5	5	343	10.8	651.98014	-308.98014			
6	6	888	9.5	840.41181	47.58819			
7	7	509	10.9	637.48540	-128.48540			
8	8	987	9.2	883.89605	103.10395			
9	9	187	14.2	159.15885	27.84115			
10								
11								
12								
13								
14								
15								

Correlations

Correlations

		building permits	interest rate
building permits	Pearson Correlation	1	-.891 **
	Sig. (2-tailed)		.001
	N	9	9
interest rate	Pearson Correlation	-.891 **	1
	Sig. (2-tailed)	.001	
	N	9	9

**. Correlation is significant at the 0.01 level (2-tailed).

Regression

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.891 ^a	.793	.764	144.298	.793	26.876	1	7	.001

a. Predictors: (Constant), interest rate

b. Dependent Variable: building permits

ANOVA^a

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	559606.629	1	559606.629	26.876	.001 ^b
Residual	145752.926	7	20821.847		
Total	705359.556	8			

a. Dependent Variable: building permits

b. Predictors: (Constant), interest rate

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	2217.412	316.204		7.013	.000
interest rate	-144.947	27.959	-.891	-5.184	.001

a. Dependent Variable: building permits

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	159.16	883.90	597.22	264.482	9
Residual	-308.980	103.104	.000	134.978	9
Std. Predicted Value	-1.656	1.084	.000	1.000	9
Std. Residual	-2.141	.715	.000	.935	9

a. Dependent Variable: building permits

Conclusion:

Therefore,

i. The correlation coefficient = -0.891 and it is significant at 1%.

ii. We know the regression equation is:

$$y = a + bx$$

Where, y is dependent variable (Building permits)

a is intercept

b is correlation coefficient

x is independent variable (Interest Rate)

From calculation we get,

$$a = 2217.412$$

$$b = -144.947$$

So, the best fitting regression line is:

$$y = 2217.412 - 144.947x$$

And,

$$\text{Residual for month 9} = 27.84115$$

iii. Coefficient of Determination = $0.793 = 79.3\%$

This shows 79.3% is explained by independent variable i.e. interest rate and remaining 20.7% is unexplained.

iv. When interest rate increases by 9.7%, building permits = 811.42233

6. Management of a soft-drink bottling company wants to develop a method for allocating delivery costs to customers. Although one cost clearly relates to travel time within a particular route, another variable cost reflects the time required to unload the cases of soft drink at the delivery point. A sample of 10 deliveries within a territory was selected. The delivery times and the number of cases delivered were recorded as follows:

Customer	Number of cases	Delivery times (minutes)
1	52	32.1
2	64	34.8
3	95	37.8
4	116	38.5
5	143	44.2
6	161	43.0
7	184	49.4
8	218	56.8
9	254	61.2
10	267	58.2

- i. Find the correlation coefficient between delivery times and the number of cases delivered.
- ii. Develop a regression model to predict delivery time, based on the number of cases delivered.
- iii. Interpret the meaning of slope in this problem.
- iv. Predict the delivery time for 150 cases of soft drink.
- v. Compute the standard error of the estimate and interpret its meaning.
- vi. Determine the coefficient of determination and explain its meaning in this problem
- vii. Compute residual for customer 7.

Solution:

SYNTAX:

DATASET ACTIVATE DataSet1.

CORRELATIONS

/VARIABLES=X Y

```
/PRINT=TWOTAIL NOSIG
```

/MISSING=PAIRWISE.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA CHANGE

```
/CRITERIA=PIN(.05) POUT(.10)
```

/NOORIGIN

/DEPENDENT Y

~~/METHOD=ENTER X~~

/SAVE PRED RESID.

OUTPUT:

[illegible]

Correlations

		Number of cases	Delivery time(minutes)
Number of cases	Pearson Correlation	1	.981**
	Sig. (2-tailed)		.000
	N	10	10
Delivery time(minutes)	Pearson Correlation	.981**	1
	Sig. (2-tailed)	.000	
	N	10	10

** . Correlation is significant at the 0.01 level (2-tailed).

Regression

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.981 ^a	.962	.958	2.1232	.962	205.149	1	8	.000

a. Predictors: (Constant), Number of cases

b. Dependent Variable: Delivery time(minutes)

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	924.797	1	924.797	205.149	.000 ^b
	Residual	36.063	8	4.508		
	Total	960.860	9			

a. Dependent Variable: Delivery time(minutes)

b. Predictors: (Constant), Number of cases

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	24.744	1.603		15.432	.000
Number of cases	.134	.009	.981	14.323	.000

a. Dependent Variable: Delivery time(minutes)

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	31.723	60.578	45.600	10.1368	10
Residual	-3.3516	2.7986	.0000	2.0018	10
Std. Predicted Value	-1.369	1.478	.000	1.000	10
Std. Residual	-1.579	1.318	.000	.943	10

a. Dependent Variable: Delivery time(minutes)

Conclusion:

Therefore,

- i. The correlation coefficient = 0.981
- ii. We know the regression equation is:

$$y = a + bx$$

Where, y is dependent variable (Deliver times)

a is intercept

b is correlation coefficient

x is independent variable (No of Cases)

From calculation we get,

$$a = 24.744$$

$$b = 0.134$$

So, the regression equation is:

$$y = 24.744 + 0.134x$$

iii. Here,

Slope = b i.e. regression coefficient

b= 0.134 means that y changes by 0.134 per unit change in x.

iv. When no of cases (x) = 150, then

$$y = 24.744 + 0.134 * 150$$

$$y = 44.844.$$

The delivery time is 44.84 minutes when number of cases is 150.

v. The standard error of the estimate is 2.1232 which means that variation around the data point is less and hence it is reliable.

vi. Coefficient of Determination = 0.962 = 96.2%

This shows 96.2% is explained by independent variable i.e. number of cases and remaining 3.8% is unexplained.

vii. Residual for customer 7 = -.03833