BAN 5733

Individual Exercise 1

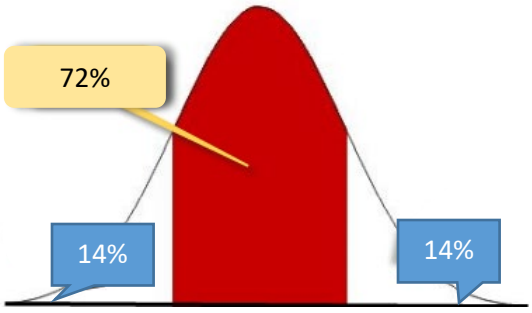
10 Points

1. Assume that average time spent by customers in a company's web site is normally distributed with mean of 41 secs and a standard deviation of 1.8 seconds. The most typical 72% of customers spend between what two time periods on the company's web site? **(1 point)**

   Solution: This is an application of Normal distribution problem. You need to find the interval where the middle 72% of customers are using the Normal distribution of mean of 41 and standard deviation of 1.8 seconds. This means you need to find the inverse of cumulative probability of 15% and 85% of that normal distribution. So, using Excel formula,

   Lower number = NORMINV(0.14,41,1.8) = 39.06
   Upper number = NORMINV(0.86,41,1.8) = 42.94

   Diagrammatically it looks like this:



2. If two balanced dice are rolled, what is the probability that the difference between the two numbers that appear will be less than 4? Explain how you arrived at your answer (hint: write out all possibilities of numbers in the throw of the dice 1 and dice 2). **(1 point)**

Solution:

First Dice	Second Dice	Difference
6	6	0
6	5	1
6	4	2
6	3	3
6	2	4
6	1	5
5	6	-1
5	5	0
5	4	1
5	3	2

First Dice	Second Dice	Difference
5	2	3
5	1	4
4	6	-2
4	5	-1
4	4	0
4	3	1
4	2	2
4	1	3
3	6	-3
3	5	-2

First Dice	Second Dice	Difference
3	4	-1
3	3	0
3	2	1
3	1	2
2	6	-4
2	5	-3
2	4	-2
2	3	-1

First Dice	Second Dice	Difference
2	2	0
2	1	1
1	6	-5
1	5	-4
1	4	-3
1	3	-2
1	2	-1
1	1	0

For the difference of the two dice to be less than 4, here are the possibilities:

- **First dice could be 6 with probability of 1/6. Then second dice has to be either 6 or 5 or 4 or 3. Probability of 6 or 5 or 4 or 3 is 4/6. Since the throw of dice are independent, the joint probability is: (1/6)*(4/6) = 4/36**
- **First dice could be 5 with probability of 1/6. Then second dice has to be either 6 or 5 or 4 or 3 or 2. Probability of 6 or 5 or 4 or 3 or 2 is 5/6. Since the throw of dice are independent, the joint probability is: (1/6)*(5/6) = 5/36**
- **First dice could be 4 with probability of 1/6. Then second dice has to be either 6 or 5 or 4 or 3 or 2 or 1. Probability of 6 or 5 or 4 or 3 or 2 or 1 is 6/6. Since the throw of dice are independent, the joint probability is: (1/6)*(6/6) = 6/36**
- **First dice could be 3 with probability of 1/6. Then second dice has to be either 6 or 5 or 4 or 3 or 2 or 1. Probability of 6 or 5 or 4 or 3 or 2 or 1 is 6/6. Since the throw of dice are independent, the joint probability is: (1/6)*(6/6) = 6/36**
- **First dice could be 2 with probability of 1/6. Then second dice has to be either 5 or 4 or 3 or 2 or 1. Probability of 5 or 4 or 3 or 2 or 1 is 5/6. Since the throw of dice are independent, the joint probability is: (1/6)*(5/6) = 5/36**
- **First dice could be 1 with probability of 1/6. Then second dice has to be either 4 or 3 or 2 or 1. Probability of 4 or 3 or 2 or 1 is 4/6. Since the throw of dice are independent, the joint probability is: (1/6)*(4/6) = 4/36**

Adding up all possibilities, the overall probability is 30/36 or 5/6.

3. Your friend, Jack, comes to you with the following situation and asks for your comments. Jack has recently inherited $5,000 and he wants to invest it. He is looking at several options.
    a. Option A: Invest all of it in a broad-based mutual fund that has generated about an average of 8% return (after tax and fees) over the years with a standard deviation of 5%.
    b. Option B: Invest all of it in a mixture of long/medium/short-term bonds that has generated an average of 4% return (after tax and fees) over the years with a standard deviation of 2%.

    What would you tell Jack and why? Assume past performance is a good indicator of what may happen in the future. **(1 point)**

    Solution: Option A has higher expected return but more risk (standard deviation). Option B has lower expected return but less risk (standard deviation). Coefficient of variation for A is

(5/8)*100% or, 62.5%. Coefficient of variation for B is (2/4)*100% or, 50%. So, risk-adjusted basis, Option B is better than Option A.

Data Set Information:

You will work with Forest Fires data set that contains 13 variables and over 500 observations. The variables in the data set are shown below with their appropriate description.

Attribute Information:

Name	Description
X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
MONTH	Month of the year: "jan" to "dec"
DAY	Day of the week: "mon" to "sun"
FFMC	FFMC index from the FWI system: 18.7 to 96.20
DMC	DMC index from the FWI system: 1.1 to 291.3
DC	DC index from the FWI system: 7.9 to 860.6
ISI	ISI index from the FWI system: 0.0 to 56.10
TEMP	Temperature in Celsius degrees: 2.2 to 33.30
RH	relative humidity in %: 15.0 to 100
WIND	wind speed in km/h: 0.40 to 9.40
RAIN	outside rain in mm/m2 : 0.0 to 6.4
AREA	the burned area of the forest (in ha): 0.00 to 1090.84

1. Explore the distribution of the **RH** via histogram and moments.
    a. Overlay a Normal curve on the histogram.
    b. Report the following information for the **RH** variable: mean, median, interquartile range, kurtosis and skewness values (**1 point**)
    c. Include a screenshot or copy of the histogram with the superimposed Normal curve and provide comments about the distribution of the variable**. (1 point**)

d. Create a Normal Probability plot for the variable **RH** and then interpret this plot. (**1 point**)

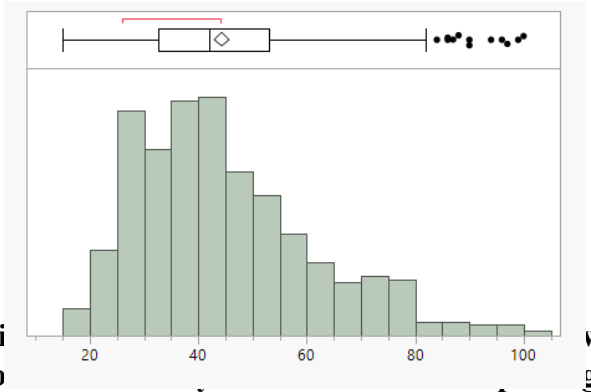
Solution:

As shown here, the histogram indicates that the RH variable is slightly right skewed. The skewness and kurtosis values indicate that the distribution is not exactly normal as does the overlay of the normal line.

Table 1: Summary Statistics for Relative Humidity

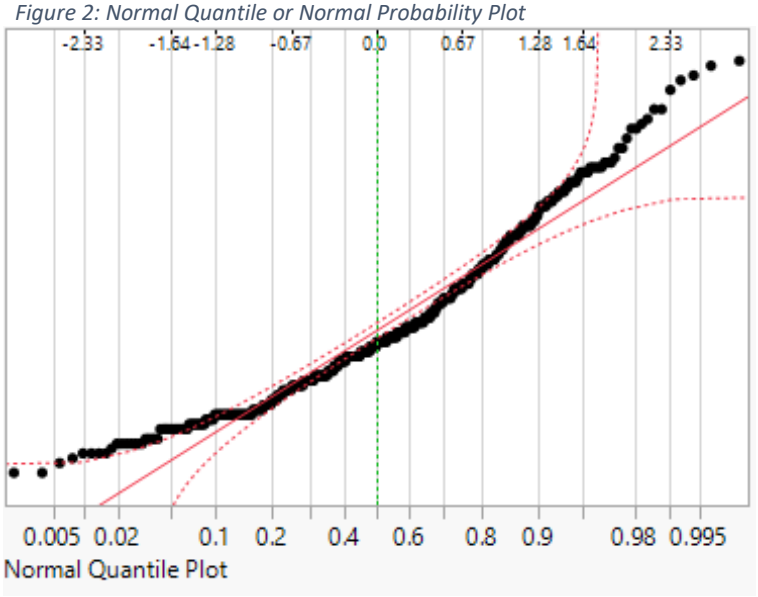
Mean	44.288201
Std Dev	16.317469
Std Err Mean	0.7176415
Upper 95% Mean	45.69806
Lower 95% Mean	42.878343
N	517
Skewness	0.862904
Kurtosis	0.4381829
Range	85
Interquartile Range	20.5

Figure 1: Histogram and Box-Whisker Plot of Relative Humidity



Accordi w, it appears that there are several violatio g the solid line and they exceed the red dotted lines indicating the normal probability ranges are violated.

Figure 2: Normal Quantile or Normal Probability Plot



2. Relationship between **2 VARIABLES**.
   a. Explore the relationship between **AREA** (Y) and **TEMP** (X) with a Fit Y by X graph or graph builder. Does there appear to be a relationship between these two variables? Write a few lines about what you observe in this data. **(2 points)**

Solution: When graphing the Area by Temp, you can quickly see that there are outliers within the data that may affect any analysis. In general, as the temperature increased the more land area that was burned. This does not indicate a causation but does show a potential correlation. Removing the outliers shows a slightly more distributed area. ***Note: The Y axis scale has changed.***

Figure 4: Scatter Plot with Outliers

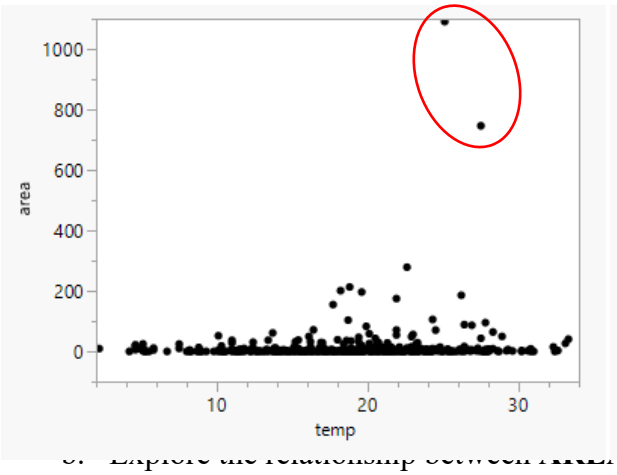
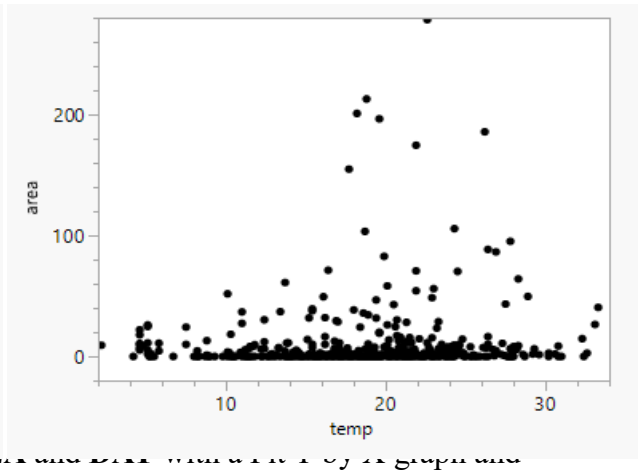


Figure 3: Scatter Plot without Two Outliers



the grouped means. Does there appear to be a relationship between these two variables write a few lines about what you observe in this data. **(2 points)**

Solution: While the graph itself is hard to see a difference in mean values across the days in which the fires started, examining the table of means by day (grouped means) does show some variation within the mean values of area burned by day of the week. Interestingly, Fridays show the lowest area burned (5.3 ha) while Saturdays show the largest average area burned (25.5 ha).

Figure 5: Area Burned by Day of the Week the Fire Started

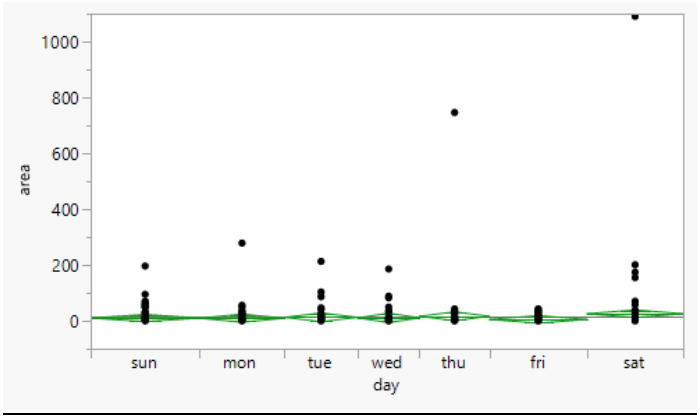


Table 2: Mean Area Burned by Start Day of the Week

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
sun	95	10.1045	6.5363	-2.74	22.946
mon	74	9.5477	7.4059	-5.00	24.098
tue	64	12.6217	7.9635	-3.02	28.267
wed	54	10.7148	8.6696	-6.32	27.747
thu	61	16.3459	8.1570	0.32	32.371
fri	85	5.2616	6.9101	-8.31	18.837
sat	84	25.5340	6.9511	11.88	39.190

Figure 6: Area Burned by Start Day of the Week without Two Outliers