

1. (10 points) Random Number Generation

- (a) Fill in the appropriate blanks to generate 30 values from a normal distribution of mean 2 and variance 3 starting at the cell A1 in your spreadsheet.

Answers:

- Number of Variables: 1
  - Number of Random Numbers: 50
  - Mean = 2
  - Standard Deviation = 1.73
  - Output Range: \$A\$1
- (b) What is the probability that any generated number is greater than 3 or smaller than 2.  
Answer: Let  $X$  be the value of the generated number, then  $X \sim N(2, \sqrt{3})$ . Let A and B be the events:
- A: the generated number is greater than 3:  $\{X > 3\}$
  - B: The generated number is smaller than 2:  $\{X < 2\}$

The probability that we are looking for is  $P(A \text{ or } B)$ .

Since A and B are disjoint, we can apply the additive rule for disjoint events and we have

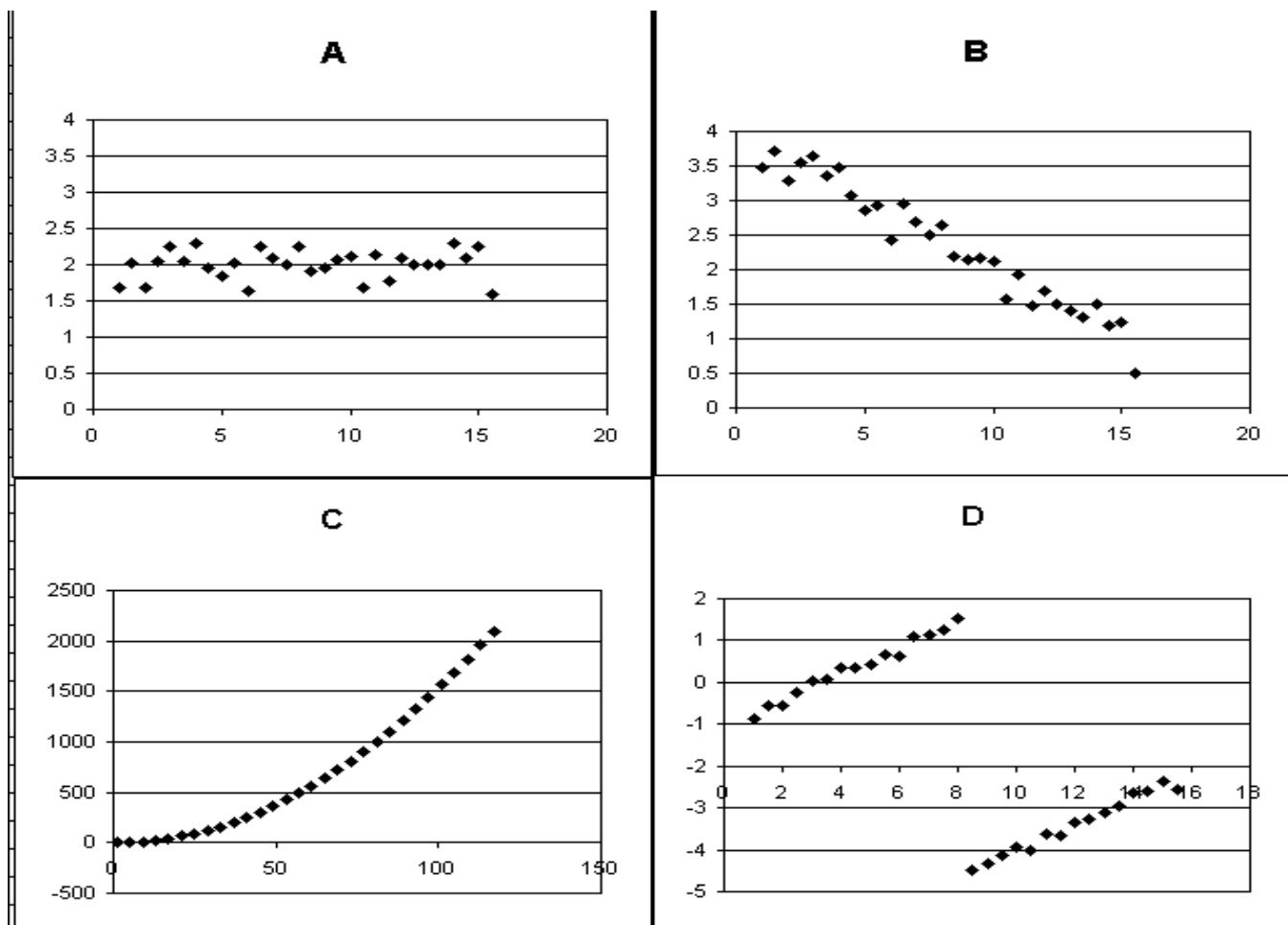
that  $P(A \text{ or } B) = P(A) + P(B)$ .

$$P(A) = P(X > 3) = P(Z > 0.577) = 0.28$$

$$P(B) = P(X < 2) = P(Z > 0) = 0.5$$

$$\text{So, } P(A \text{ or } B) = 0.78$$

2. (30 points) Answer the following questions for the four scatterplots A, B, C, and D.



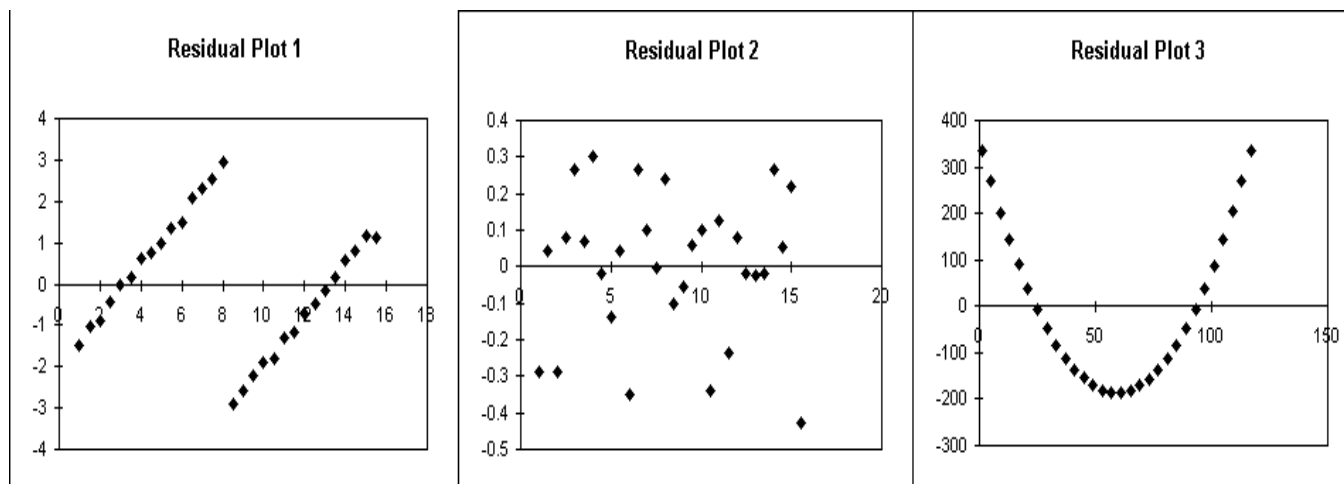
(a) Match each correlation to the appropriate scatterplot

- $r \simeq 0.96$  Answer: C
- $r \simeq 0.08$  Answer: A
- $r \simeq -0.97$  Answer: B
- $r \simeq -0.65$  Answer: D

(b) There is a strong relationship between the two variables in which scatterplot(s)? Answer: From the scatterplot we see that: in B the relationship is strong, linear, negative. In C the relationship is strong curved. In D, we have two clusters and the relationship is strong within each cluster

Continued...

- (c) There is no apparent relationship between the two variables in which scatterplot(s)? **There is no apparent relationship between the variables in A (we can see that from the scatterplot and the low value of the correlation)**
- (d) Match each scatterplot with its corresponding least squares regression residual plot. Note that one of the following residual plot corresponds to two scatterplots.



Answer:

- Residual 1, Scatterplot D.
  - Residual 2, Scatterplot A and B
  - Residual 3, Scatterplot C
- (e) How much of the variation in the response variable is explained by the least squares regression of the response on the explanatory in Scatterplot A.  
 $R^2 = (0.08)^2 = 0.0064 = .64\%$ , so only 0.64% of the variation in the response is explained by the least squares regression of the response on the explanatory.
- (f) The least squares regression fit would be a good fit in which scatterplot(s) (*Don't forget to justify your answers*)?

Answer: by looking at the scatterplot, the value of  $R^2$  and the residuals, the least squares regression would be a good fit for scatterplot B. It will not be a good fit for A because there is no relationship between the variables. It will not be a good fit in C because the relationship is curves and there is a pattern in the residuals. It will not be a good fit in B because there are clusters and the least squares regression would show a negative slope whereas the relationship between each cluster is positive, we can see this pattern also in the residuals.

Continued...

3. (30 points) Alice runs the "Wonderland" grocery store and she would like to know if customers who spend a lot during the week before Thanksgiving spend a lot during the week before Christmas of the same year. 890 regular customers have a "Queen(or King) of Heart customer" or (HeC) discount card, and the store keeps a database of amount of purchases for each "HeC" holder. Alice asks Lewis, a stat 31 student, for a quick preliminary analysis on a sample of customers. Lewis randomly picks 30 "HeC" holders from the database.

- (a) What is the target population?

Answer: the target population is all "HeC" holders or all regular customers in the store

- (b) Using the following line from table B 90597 93600 54973 86278 88737 74351, what are the first 3 "HeC" in the sample?

The three first numbers in the sample are: 360, 54, and 862

- (c) Lewis finds that in his sample, the equation of the regression line of spending in the week before Christmas (dollars) on spending in the week before Thanksgiving(dollars) is  $y = 1.2*x + \$20$ . Mr. Rabbit spends \$10 above the sample average spending in Thanksgiving, how many dollars above the sample average spending in Christmas do you predict that he will spend?

We know that the least squares regression line goes through the point  $(\bar{x}, \bar{y})$ . Let  $x_R$  be the amount Mr. Rabbit spends for Thanksgiving, and let  $y_R$  be the predicted amount he will spend for Christmas. Then, since (1.2) is the slope we have that

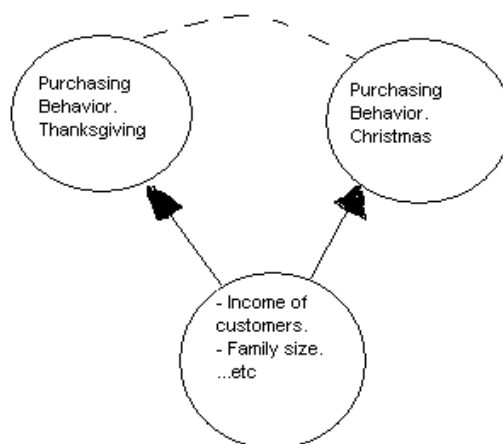
$$\frac{y_R - \bar{y}}{x_R - \bar{x}} = 1.2$$

Since  $x_R - \bar{x} = \$10$  then

$$y_R - \bar{y} = \$12$$

So the predicted amount Mr. Rabbit would spend in Christmas above the average in Christmas is \$12.

- (d) Does the purchasing behavior of a customer in Thanksgiving causes the purchasing behavior of a customer in Christmas? Use a diagram like those we saw in Chapter 2 to present your answer. No, even if we find a relationship between purchasing behavior in Thanksgiving and purchasing behavior in Christmas, this doesn't mean causation. The two purchasing behaviors are probably common responses of several variables such as: income and family size



Continued...

- 
- (e) If 712 people who own a "HeC" discount card have children in the household. What is the probability that none of the people in Lewis' sample have children in the household? If we pick a person at random from the population, the probability that this person doesn't have kids is  $1 - \frac{712}{890} = 0.2$ . If we pick 30 people at random, then the probability of each person having a kid is **independent** of the choice of the other people in the sample. The event: none in the sample have a child could be written as: the 1st person doesn't have a child and the 2nd person doesn't have a child ... and the 30th person doesn't have a child. So, by applying the multiplication rule for independent events we have that the probability that none in the sample have a child is  $(0.2)^{30} = 1.07 * 10^{-2}$
- (f) Alice thinks now that Lewis should take into account in his sampling scheme whether the customers holding a "HeC" card have children in the household or not, give the name of the sampling scheme you think Lewis should use and briefly explain how to choose such a sample.

Answer: Lewis should use a stratified random sample. To obtain such a sample, we can divide the population into two strata (customers with children in the household and customers without children in the household). Then, we should take a SRS from each strata and combine them together to get the stratified sample.

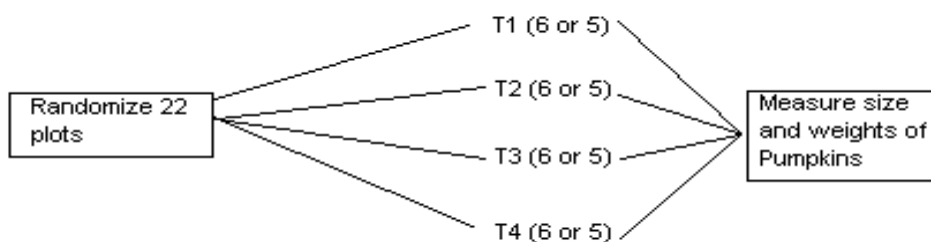
4. (10 points) The size and weight of a pumpkin depends on the seed variety of the pumpkin and also on the quality of the soil. Before starting to grow pumpkin in his farm, Jack decided to experiment in 22 small and equal size plots which seed: Atlantic Giant pumpkin seed or Spooktacular pumpkin seed grows to a larger pumpkin and if the "fish emulsion" soil mineral additive makes any difference in the size and weight of the pumpkin. In each plot, Jack plants seeds of one of the two varieties and he either adds compost alone or compost with the "fish emulsion" to the soil of the plot. In the fall, Jack measures the average weight and size of pumpkins in every plot.

(a) What are the experimental units, factors, treatments, and response?

Answer:

- Experimental units: plots
- Factors: 1. Seed variety (levels: Atlantic giant, or Spooktacular). 2. Quality of soil (levels: compost, or compost with fish emulsion).
- Treatments:
  1. Atlantic Giant in compost
  2. Atlantic Giant in compost with fish emulsion.
  3. Spooktacular in compost.
  4. Spooktacular in compost with fish emulsion
- Response: weight and size of pumpkins in each plot.

(b) Outline the design of this experiment.



Randomize to see which treatment group will have 5 plots or 6 plots.

Continued...

5. (25 points) Ella is considering buying a stock of a hot company for \$1000. Every day, the stock either gains 30% or loses 25% of its value from the previous day, each with probability 0.5. Its returns on consecutive days are independent of each other. If Ella buys the stock, she plans to sell it after two days. Let  $X$  be the value in dollars of the stock after two days.

- (a) Fill in the dots the three missing values in the probability table of the random variable  $X$  and justify your answers

*Note that if you don't know how to answer this question, you can fill the probability table with your "guessed" values to be able to answer the three following questions in this Problem.*

X	\$562.5	\$975	\$1690
Probability	0.25	0.50	0.25

Answer: there are four possibilities (loss both days, gain and then loss, loss and then gain, and gain both days). Since the two days are independent, and the probability of gain or loss is (0.5) then each possibility has a probability of  $(0.25 = (0.5) * (0.5))$ .

- Value of  $X$  if loss both days:  $(1000 * (0.75)) * (0.75) = \$562.5$
- Value of  $X$  if gain and then loss:  $(1000 * (1.3)) * (0.75) = \$975$ .
- Value of  $X$  if loss and then gain:  $(1000 * (0.75)) * (1.3) = \$975$ .
- Value of  $X$  if gain both days:  $(1000 * (1.3) * (1.3)) = \$1690$ .

- (b) What is the probability that the stock is worth more after two days than the \$1000 Ella would pay for it?

Answer: from the table

$$P(X \geq 1000) = P(X = 1690) = 0.25$$

- (c) What is the expected value of the stock after 2 days?

$$\mu_X = \sum x_i * p_i = 0.25 * 562.5 + 0.5 * 975 + 0.25 * 1690 = \$1050.62$$

- (d) What is the standard deviation of the stock after 2 days?

$$\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 * p_i} = 405.74$$

- (e) Which one of your answers in b) or c) would you use to help Ella make an informed decision. Justify your choice.

Answer: Since Ella is buying only one stock, I would use answer (b) because this probability reflects better her chances of winning. If Ella was buying multiple independent stocks with the same behavior as  $X$  then I would use (c) because then according to the Law of Large Numbers she would make an average profit of \$50.62.

6. (1 points) Bonus question: Find in the Problems 1 to Problem 6, the value of two parameters and the value of one statistic. Write those values, specify in each case in which problem you found them and if it is a statistic or a parameter.

Answer: There are multiple correct answers. A parameter is a number describing the population and a statistic is a number describing a sample. In problem 3, the slope (1.3) of the linear regression is a statistic (obtained from the sample). In problem 3, the probability of not having children in the household for the population of customers is (0.2) and is a parameter. In problem 5, the expected value of  $X$  is a parameter.

Happy Halloween!