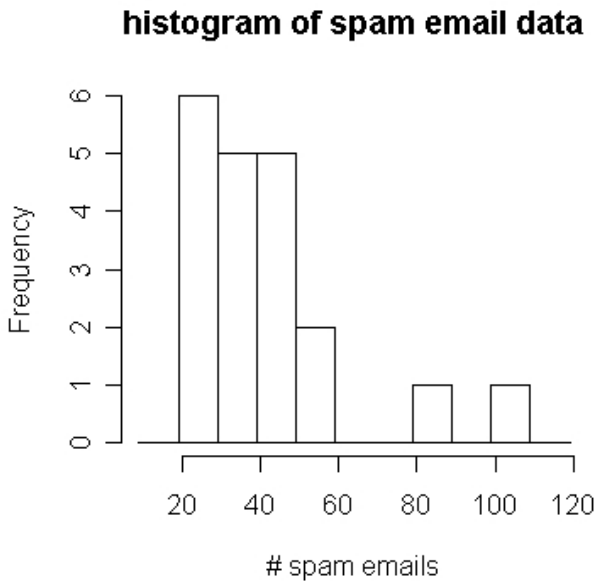
1. Harris recently installed a spam filter software, but he still saw spam emails in his inbox. He made a daily record of the number of spam emails that were delivered to his inbox over the past 20 days. The following is a frequency histogram for his data. The frequency refers to the number of days.



a) Harris also plotted a stem-and-leaf display for the data. Which of the following is a correct stem-and-leaf display for his data? Check only one answer. [2 marks]

___ **Stemplot A**
___ Stemplot B
___ Stemplot C

A.
```
 2 | 011355
 3 | 01467
 4 | 12479
 5 | 56
 6 |
 7 |
 8 | 0
 9 |
10 | 5
```

B.
```
 2 | 011355
 3 | 01467
 4 | 12479
 5 | 56
 8 | 0
10 | 5
```

C.
```
1 | 05
2 | 011355
3 | 01467
4 | 12479
5 | 56
6 |
7 |
8 | 0
```

b) Which of the following is a correct statement about the distribution of the spam email data? Check only one answer. [2 marks]

____ The distribution is roughly symmetric, and the mean is about the same as the median.
____ **The distribution is skewed, and the mean is larger than the median.**
____ The distribution is skewed, and the mean is smaller than the median.

c) What is the third quartile of the number of spam emails? Use the stem-and-leaf display you have chosen in part (a) to answer this question. Check only one answer. [2 marks]

____ 25
____ 41
____ **48**
____ 55

d) Which of the following pairs of summary statistics best describe the center and the spread of the number of spam emails received daily? Check only one answer and explain briefly. [3 marks]

____ mean and standard deviation
____ mean and IQR
____ **median and IQR**
____ median and variance

Explain: **There are possible outliers in the distribution, and the distribution is skewed, so it's best to use summary statistics that are insensitive to outliers (i.e. median and IQR).**

e) Identify any outliers in the data set using the stemplot you have chosen in part (a). It is given to you that the IQR is 23. Show your work here. [2 marks]

$Q_3 + 1.5 \times IQR = 48 + 1.5 \times 23 = 82.5$

105 > 82.5, so 105 is an outlier.

There are no outliers in the lower end (the distribution does not have a long left tail)

2. During the boxing week last year, a local bookstore offered discounts on a selection of books. The manager looks at the records of all the 2743 books sold during that week, and constructs the following contingency table:

	Discounted	Not discounted	Total
Paperback	790	389	1179
Hardcover	1276	288	1564
Total	2066	677	2743

a) What percentage of paperback books sold were discounted? [1 mark]

770/1179=67.0%

b) What percentage of hardcover books sold were discounted? [1 mark]

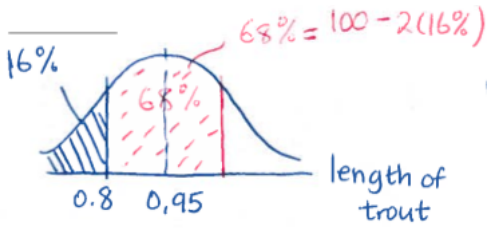
1276/1564=81.6%

c) What can you say about the two events: A = a book is paperback and B = the book is discounted? [1 mark]
$P(B) = 0.75$, $P(B|A) = 0.67$. Since $P(B) \neq P(B|A)$ they are associated / not independent.

d) Write down the marginal distribution of the book type variable. [2 marks]

Paperback	Hardcover	Total
1179	**1564**	**2743**
(42.98%)	**(57.02%)**	**(100%)**

3. The length of trout in a lake is normally distributed with mean $\mu = 0.95$ feet and an unknown standard deviation σ. If 16% of all trout are shorter than 0.8 feet, what is the value of σ? [3 marks]



By symmetry of the Normal model, we know the middle area us 68%.
Then by $68 - 95 - 99.7\%$ rule, 0.8 is 1 SD below the mean.
So, $\sigma = 0.95 - 0.8 = 0.15ft$

4. The hourly rates for highschool private tutoring follow the normal distribution with mean μ and standard deviation σ. It is given that the middle 99.7% of all the hourly rates fall between \$13 and \$43. Then (check your answers)

a) the mean μ is [1 mark]

____ **equal to \$28.**
____ greater than \$28.
____ less than \$28.

b) the standard deviation σ is roughly equal to [1 mark]

____ **\$5.**
____ \$10.
____ \$15.

c) an hourly rate of $12 has [2 marks]

____ a z-score of 0.

____ **a negative z-score.**

____ a positive z-score.

d) the IQR of the hourly rates is [2 marks]

____ equal to $30.

____ greater than $30.

____ **less than $30.**

5. Does how long children remain at the lunch table help predict how much they eat? Twenty toddlers at a nursery school were observed. On each toddler, the number of minutes he/she spent at the table when lunch was served and the number of calories that was consumed during lunch were measured. The two variables show a reasonably linear trend with a correlation coefficient $r = -0.65$.

   The least-squares regression line that predicts the amount of calories consumed from the time stayed at the table during lunch has a slope of –3.25 and an intercept of 566.5.

   a) Predict the number of calories consumed for a child who spends 25 minutes at the table during lunch. [1 mark]

      Predicted calories = 566.5 − 3.25(25) = 485.35

   b) For the following statements, check all that are correct. [3 marks]

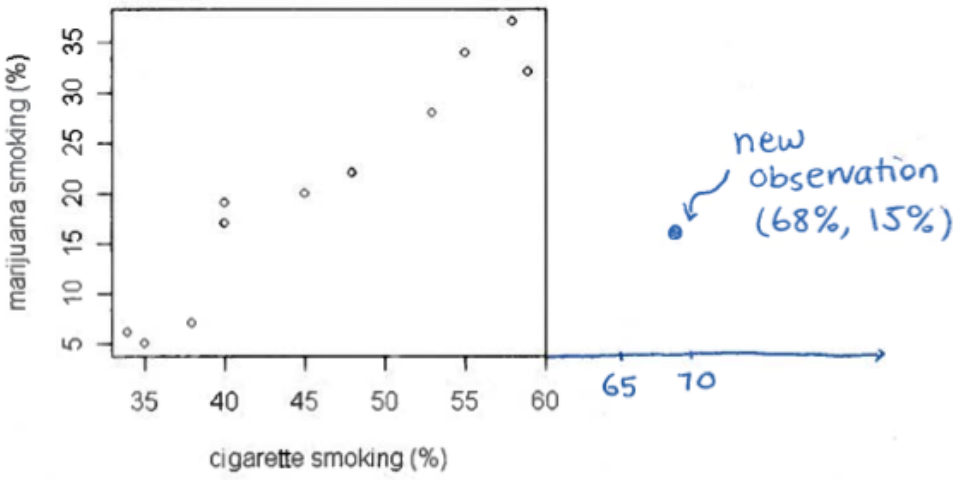
      ____ The association between the time spent at the lunch table and the amount of calories consumed is weak because r is negative.

      ____ One standard deviation (SD) increase in the number of minutes spent at the lunch table is associated with one SD decrease in the number of calories consumed.

      ____ **If one changes the unit of the amount of time spent from minutes to hours, the value of the correlation coefficient r will remain unchanged.**

      ____ None of the above.

6. A survey was conducted in 11 countries to determine the percentage of teenagers who had smoked cigarettes and used marijuana. The scatterplot for the two variables is shown below:



One more country participated in the survey, and the percentages of teenagers who have smoked cigarettes and used marijuana were found to be 68% and 15%, respectively. The correlation coefficient r is then recalculated. How do the values of r before and after the inclusion of the new observation compare? Check only one answer and explain briefly. [3 marks]

____ $r(\text{before}) < r(\text{after}) < 0$

____ **$0 < r(\text{after}) < r(\text{before})$**

____ $r(\text{before}) < r(\text{after}) < 1$

Explain: **After adding the observation of the country, the linear pattern becomes weakened. Hence the correlation coefficient will decrease, but the value will still be positive (the linear trend remains positive).**

7. You need to drive past two traffic lights on the way from your house to the nearest grocery store. The probability that you hit a red light is 0.5 at the first intersection and 0.4 at the second intersection. The probability that you run into a red light at both intersections is 0.25. On a random day you drive from home to that grocery store.

Define the following events:
E_1 = you run into a red light at the first intersection
E_2 = you run into a red light at the second intersection

Based on the information given, are E_1 and E_2 independent events? Explain briefly why or why not. [3 marks]

$P(E_1) = 0.5$, $P(E_2) = 0.4$, $P(E_1 \text{ and } E_2 \text{ happen together}) = 0.25$
If E_1, E_2 are independent,
we would have $P(E_1 \text{ and } E_2 \text{ happen together}) = P(E_1) \times P(E_2)$
but this is not the case: $0.25 \neq 0.5 \times 0.4$
$\Rightarrow E_1$ and E_2 are NOT independent.

8. A city council was planning to turn a major street in the city from a primary traffic artery to a secondary traffic artery. It sent out a questionnaire to all the 36,589 households living in the city requesting for their input concerning the plan. Thirty four percent of the 10,375 households who returned the questionnaires opposed the plan.

   True or false [2 marks]
   The percentage of the 10,375 households that opposed the plan, 34%, is a parameter.
   ___ True
   ___ **False**

9. In a university parking database with 5600 registered vehicles, records show that 43% of the registered vehicles are Asian makes, 23% are European makes and the remaining are American makes. Among all the 5600 cars, 20% once received a parking ticket.

   a) You randomly pick three vehicles with replacement from the database ("with replacement" means any drawn vehicle will be put back to the database before the next vehicle is drawn). What is the probability that at most two of the three are American makes? [3 marks]

   There are 3 makes categories, and their probabilities should add up to 1.
   So P(selecting an American make) $= 1 - 0.43 - 0.23 = 0.34$
   Let X = number of cars out of the 3 cars drawn that are American makes
   $X \sim$ Bin(n = 3, p = 0.34)

$$\begin{aligned} P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= {}_3C_0(0.34)^0(1-0.34)^3 + \ldots + {}_3C_2(0.34)^2(1-0.34)^1 \end{aligned}$$

   [or $1 - P(X = 3)$]

   b) Consider a random sample of 100 vehicles selected from the database.
   The sample proportion of the 100 selected vehicles that <u>had never received</u> a parking ticket has an approximate 95% chance of falling between
   <u>0.72</u> and **<u>0.88</u>**.
   Fill in the blanks and show your calculation below. [4 marks]

   p = population proportion of vehicles that had never received a ticket
   $= 1 - 0.2 = 0.8$
   $\hat{p}$ = sample proportion of vehicles that had never received a ticket ($n = 100$)
   Check $np = 100(0.8) = 80 > 10, n(1-p) = 100(0.2) = 20 > 10$
   So $\hat{p}$ follows approximately the Normal model with mean 0.8 and SD $= \sqrt{\frac{0.8(1-0.8)}{100}} = 0.04$
   By 68-95-99.7% rule, 95% of $\hat{p}$ values will fall between $0.8 - 2(0.04)$ and $0.8 + 2(0.04)$, i.e. (0.72, 0.88)

10. Each day the value of a particular stock goes up one unit with probability 0.3, stays the same with probability 0.5 or else goes down one unit with probability 0.2. [8 marks]

a) Consider the random variable: change in value of the stock in a day. Find the mean of this random variable.

Let X = change in value

X	**+1(up)**	**0(stays the same)**	**-1(down)**
P(X = x)	**0.3**	**0.5**	**0.2**

Mean of X = 1(0.3) + 0(0.5) + (-1)(0.2) = 0.1

b) The standard deviation of the change in the value of the stock in a day is given to be 0.70. A stockbroker reported that the average change in the stock value over 500 independent days exceeds 0.01 unit. Do you think what the stockbroker reported is unusual? Justify your answer using z-score or probability.

$\bar{X}$ = average change in the stock value over 500 independent days.
$= \frac{X(\text{day 1}) + X(\text{day 2}) + \ldots + X(\text{day 500})}{500}$
$\bar{X} \sim$ approx. N(mean = 0.1, SD = $\frac{0.70}{\sqrt{500}}$) by the CLT.
Note that the 500 days are independent and n = 500 is a sufficiently large sample size.
z-score for the value 0.01 unit is

$$\frac{0.01 - 0.1}{0.70/\sqrt{500}} = -2.87$$

P(average $\bar{X} > 0.01$) is greater than $95\% + \frac{5\%}{2} = 97.5\%$
So what the stockbroker said is highly probable (not unusual at all).