Solutions to the Review 1

(1) $\pi = 0.6$ calls resolved within 5 minutes $= S$

(a) $\underbrace{SS\ldots\ldots S}_{20} = 0.6^{20}$ or $P(x=20) = \binom{20}{20} 0.6^{20} 0.4^{0}$

$= \boxed{0.6^{20}}$

(b) $\underbrace{FF\ldots\ldots F}_{20} = 0.4^{20}$ or $P(x=0) = \binom{20}{0} 0.6^{0} 0.4^{20}$

$= \boxed{0.4^{20}}$

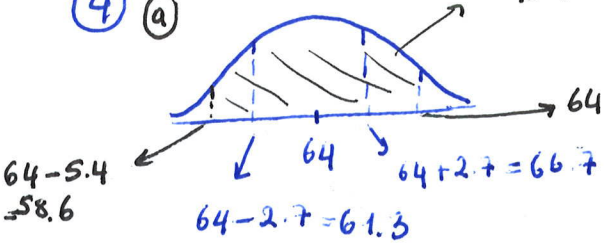
(c) $P(x \geq 1) = 1 - P(x=0) = \boxed{1 - 0.4^{20}}$

(2) observational, because ...

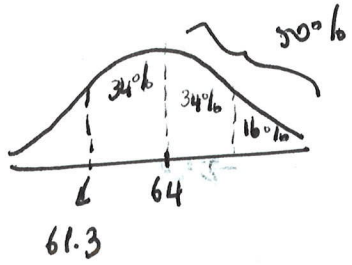
(3) (a) $4316 + 8986 + 1182 = 14,484$

(b) No, this poll clearly consists of volunteers who decided to participate. Note that we have no guarantee that these responses came from 14,484 distinct people; some may have voted more than once.

(4) (a)



(b)



$34\% + 34\% + 16\% = 84\%$ of the women are taller than 61.3.

(5) $\dfrac{3.1+2.4+1.5+x+1.7}{5} = 2.3$

$8.7 + x = 11.5$

$\boxed{x = 2.8}$

(b) New data is (with $\bar{x}$ ↓)

3.1 2.4 1.5 2.8 5 $\qquad \bar{x} = \dfrac{3.1+2.4+1.5+2.8+5}{5} = 2.96$

Mean is a sensitive measure of the center.

(6) x = depth of flooding
y = flood damage

a) $\hat{y} = 13.96 + 3.170x$

b) The plot with the line drawn in suggests that perhaps a simple linear regression model may not be appropriat. It suggests that there might be a curvilinear relationship may exist between depth and damage.

c) when $x = 6.5$ $\hat{y} = 13.96 + 3.170(6.5) = 34.565$

d) Since $x = 18$ is outside of the range of x-values for which data has been collected, we have no information concerning the relationship in the vicinity of $x = 18$. So one would not want to use the least squares line to predict flood damage when depth of flooding is 18 feet.

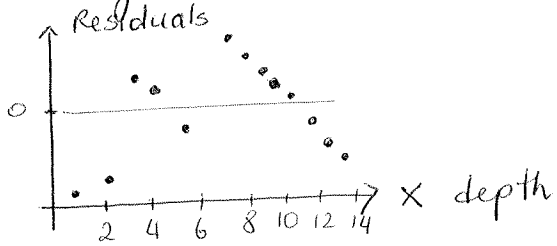
e) We want to find first the mean and the standard deviation of the flood damage data.

$\bar{x}_y = 36.15$
$s_y = 13.31$

z score for 26 $= \frac{26 - 36.15}{13.31} = -0.763$.

So when flood damage is 26 you know that you are within one standard deviation of the mean of the flood damage data set.

f)



Finding the residuals on minitab
Stat → Regression → Fitted Line Plot
→ storage ✓ Residuals
✓ Fits.
This will create a new column with residuals and fits.
Then graph a scatterplot with
Y → residuals
x → flood depth.

g) Stat → Display Descriptive Statistics.

Variables: y

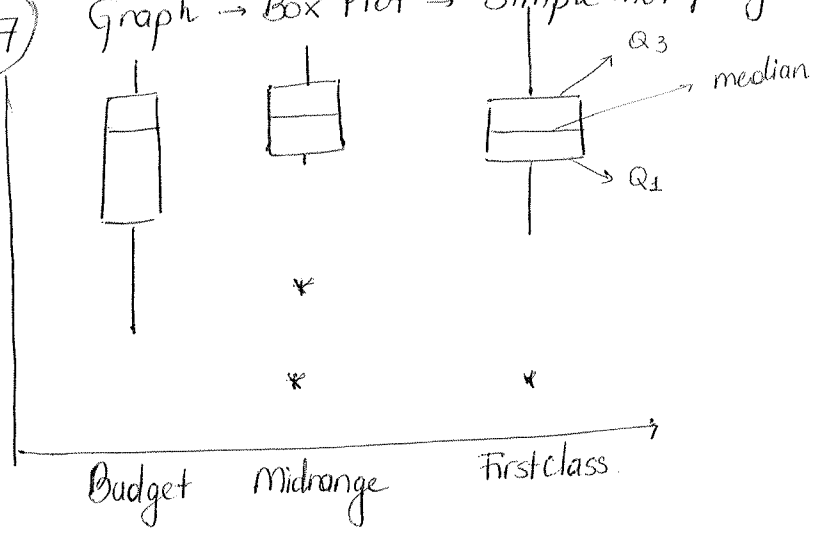
Mean	St Dev	Min	Q_1	Median	Q_3	Max
36.15	13.31	10	27	43	46.5	49

Five # summary consists of
min Q_1 median Q_3 max.

h) Mean and median are very off from each other. This would be a data set that's skewed negatively that's why the mean is pulled

down towards the left tail. For center and spread I would use five number sommary and also IQR.

(7) Graph → Box Plot → Simple multiple y's.



The median franchise cost is about the same for each type of the hotel. The variability of franchice cost differs substancially for the three types. With the exception of outliers first class hotels very little in the cost of franchises while budget hotels vary a lot. Ignoring the outliers the distribution of franchise costs for first class hotels is quite symmetric, midrange hotels slightly skewed to the right (positive) while budges hotels distribution is skewed to the left (negative)

(8) The ordered data is

33 41 41 42 44 44 46 46 46 46 48 50 50 51 53 53 54 55 60

Q1 = 44, median = 46, Q3 = 53

a) $IQR = 53 - 44 = 9$

$1.5 \cdot 9 = 13.5$

$Q_3 + 13.5 = 53 + 13.5 = 66.5$ mild no outliers on the positive side

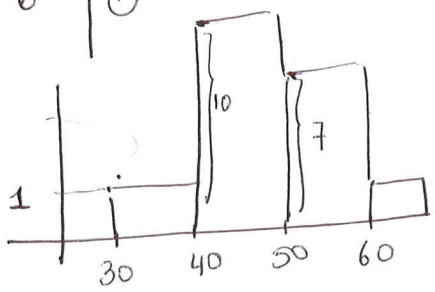
$Q_1 - 13.5 = 44 - 13.5 = 30.5$ // negative side.

Hence there arent any extreme outliers.

(b)

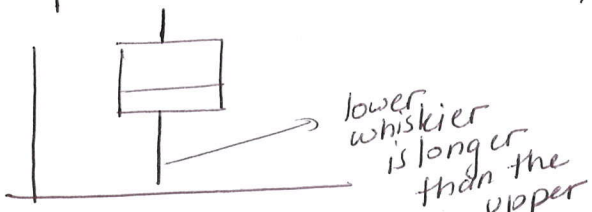
3	3
4	1 1 2 4 4 6 6 6 6 8
5	0 0 1 3 3 4 5
6	0

(c)



(d) The mean of this data is 47.53 and the median is 46. this suggests a little bit of positive skewerless in the middle half of the data And looking at the boxplot we see negative skeweness in the extremes of the data.

Boxplot



(10) (a) Correlation coefficient is between -1 and 1 and doesn't have a unit and is a numerical value between 2 numerical variables.

(b) Correlation coefficient is a numerical value between two numerical variables, however gender is a categorical variable

(11) a)

	Mean	Median	Q_1	Q_3
Pure tone	106.2	72	38	155.5
Monkey call	176.6	141	91	205.5

These describe very well that all the numerical measurements are higher for monkey call than of pure tones.

b) Call = 93.92 + 0.7783 Tone

c) r squared = .408 or 40.8%

40.8% of the variability in monkey calls can be explained by the linear relationship between pure tones and monkey calls.

(d) The 3rd point (241, 485) has the largest residual (203.5045)

e) The first point (474, 500) is an outlier in the x-direction. The value of $r = .639$ with this point but r drops to .479 when this point is removed. This point has a major impact on r.

f) No, the tones range from 19 to 474 (the outlier) in the data set so this would go well beyond the range of variable data. The regression line should not be used for extrapolation.

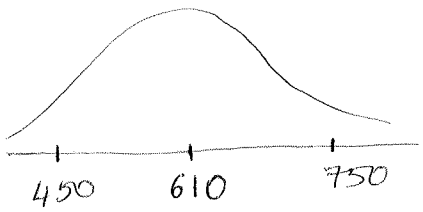
(12) • The different varieties have different lengths. Yellow is the shortest, red is next and bihai is the longest. The means, medians or any other measures of location clearly shows this fact.

- The variability as measured by IQR (height of each box) or by standard deviation is highest for red. The yellow has the lowest/smallest variability and bihai is in between.
- The distribution of lengths for the yellow variety is symmetric, the other two distributions are skewed to the right.

(13) (Box-plots would be a good one to derive the above)
Restricting the range of GPAs to perfect 4.0 will not allow us to explore the relationship with the explanatory variables.

b) Many possible answers. For example a simple random sample of students could be selected or a stratified sample to obtain equal numbers of students for each year could be obtained. Randomization should be used and association between GPA and other explanatory variables should be examined.

(14) mean = 610



30 schools

2 schools is 6.6% of 30 schools.

We do know for normal distributions.

95% of the observations fall within 2 standard deviation away from the mean

As a guess we can take $\begin{array}{r} 750 \\ 610 \\ \hline 140 \end{array}$ as 2 st deviation

and then standard deviation $= \frac{140}{2} = 70$.

(b) z-score $= \frac{594 - 610}{70} = -0.228$

(15) (a) weight versus diameter $r^2 = 94.4\%$

weight versus length $r^2 = 72.3\%$

Diameter seems to give a better linear association then length.

(b) when we look at weight versus squared diameter we see that we get even a better linear association with $r^2 = 99.7\%$