1. **(3pts)** An employee counts and records the number of wasps on a sample of 10 clumps of vines. After calculating statistics on the data, the employee notices that he wrote down the maximum data value incorrectly. He changed the incorrectly entered 51 to 42. The value of 42 is now the highest number of wasps observed in the 10 clumps of vines. Which of the following statistics for his data does he **need to recalculate**? **Select all that apply**

   (a) Range

   (b) IQR

   (c) Standard Deviation

   (d) Average

   (e) Median

2. **(3pts)** Assume that you are rolling a fair 6-sided die, with faces 1,2,3,4,5,6 equally likely. What is the chance that in exactly 2 out of 4 rolls, you roll a 3?

   (a) 0.116

   (b) 0.019

   (c) 0.039

   (d) 0.077

   (e) 0.231

3. A summary of the number of parking tickets given out each month in one UW parking lot for the past five years is given below. Use this data for the two questions that follow.

```
> summary(monthly_tickets)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.000   6.000   7.000   8.579  10.000  17.000
```

Figure 1: Summary of number of monthly parking tickets given out in the past five years in one parking lot.

   (a) **(3pts)** The shape of this data is most likely

      i. Left Skewed

      ii. Symmetric

      iii. Right Skewed

   (b) **(3pts)** What can be said about the number of outliers that would show in a standard boxplot, as we constructed in class?

      i. 0

      ii. Exactly 1

      iii. At least 1

      iv. Two or more

4. **(3pts)** A business assigned two different employees to work on a question independently. The company thinks the chance the first employee solves the problem is 0.9 and the probability the second employee solves the problem is 0.8. Assuming the success [or failure] of the employees is independent, what is the probability at least one of the employees will solve the problem?

   (a) 0.72
   (b) 1.7
   (c) 0.9
   (d) 0.98
   (e) 0.26

5. **(3pts)** Which of the following statements are true about the standardized scores (z-scores) calculated on a normal population? **Select all that apply**

   (a) Z-scores tell us how many standard deviations above or below the mean a value falls.
   (b) Z-scores follow the standard normal distribution with mean 0 and standard deviation 1.
   (c) The area below $z = -4.5$ is slightly less than than the area above $z = 4.2$.

   all correct

6. **(3pts)** Estelle is grabbing a flashcard out of her box of eight (8). Some of the cards are white, and some are grey. Additionally, a triangle or circle are printed on some of each color, as shown in Figure 2. What is true about the events A: picking a grey card and B: picking a triangle if I were to choose one card at random?

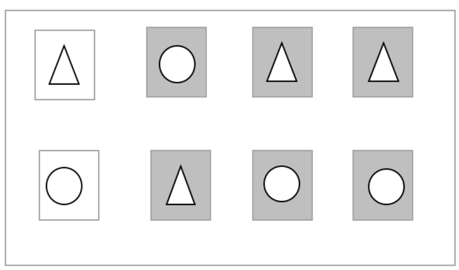


Figure 2: Color and Shape on Card in box.

   (a) The events are independent and mutually exclusive.
   (b) The events are independent and not mutually exclusive.
   (c) The events are not independent, but are mutually exclusive.
   (d) The events are neither independent nor mutually exclusive.

7. **(3pts)** Suppose you are interested in estimating the mean of a population of values μ from a random sample of size 3. The variance of the population is σ^2. Select the statement that best describes which estimator would you prefer. Assume X_1, X_2 and X_3 are iid random variables from the population. Would you prefer to use $\hat{\mu_1}$ or $\hat{\mu_2}$?

$$\hat{\mu_1} = 2X_1 - X_2 \qquad\qquad \hat{\mu_2} = X_1 + X_2 - X_3?$$

(a) I prefer $\hat{\mu_1}$ since it is more biased and has more variability.

(b) I prefer $\hat{\mu_2}$ since $\hat{\mu_1}$ is more biased and has more variability.

(c) I prefer $\hat{\mu_2}$ since $\hat{\mu_1}$ has more variability.

(d) I prefer $\hat{\mu_1}$ since $\hat{\mu_2}$ has more variability.

(e) No preference since they are both unbiased and have the same variability.

8. **(4 pts)** Identify whether the random variable X in each of the following scenerios is a binomial random variable or not. If not, in a few words explain what assumption[s] is[/are] not well met.

(I) According to a recent study, 63% of children wear their bike helmets always, 24% wear them often, and 13% hardly ever wear them. If you interview 15 randomly selected children, let X be the number that report hardly ever wearing a helmet. Works. We assume each child has 13$ chance of hardly wearing; 87% of not not hardly wearing

(II) You have a bag of 10 batteries, in which 2 are defective. Suppose you put 4 batteries into your calculator from the bag. Let X be the number of defective batteries in your calculator. Fails since probability for each battery defective isnt the same (sampling without replacement).

9. Suppose the number of days, X, that it takes for the post office to deliver a letter from City A to City B ranges from 2 to 6 days and has the following partial distribution:

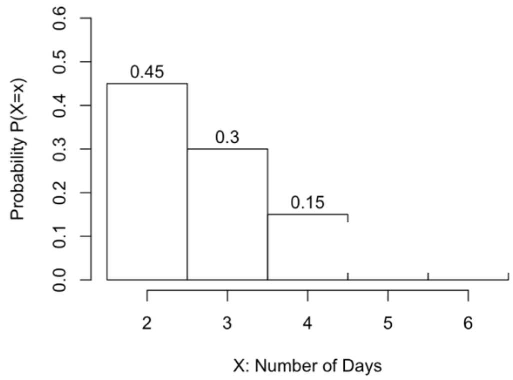


Figure 3: Number of days for letter delivery from City A to City B.

(a) **(4pts)** Assuming we know it is equally likely for the letter to take 5 or 6 days for delivery, sketch in and label bars above X=5 and X=6 to complete the probability histogram. Show the reasoning behind the values you choose.
Since 0.45+0.3+0.15=0.9, P(X=5)=P(X=6)=0.05; we know total probablity =1

(b) **(4pts)** Find the mean and standard deviation of X.
$E(X) = 2 * .45 + 3 * .3 + 4 * .15 + 5 * .05 + 6 * .05 = 2.95$
$Var(X) = (2 - 2.95)^2 * .45 + (3 - 2.95)^2 * .3 + (4 - 2.95)^2 * .15 + (5 - 2.95)^2 * .05 + (6 - 2.95)^2 * .05$ so $SD(X) = 1.1169$

(c) **(3pts)** Give a brief interpretation of the mean of X μ_X that you calculated above:
This is the long term average of the number of days that all letters will take to travel from City A to City B.

(d) **(4pts)** If three letters are to be mailed from City A all at different times of the year, what is the probability that exactly one of them will arrive in City B in four or more days?
n=3, Y=num of letters that travel from City A to City B in 4 or more days. $P(success) = 0.15 + 2 * 0.05 = 0.25$; $P(Y = 1) = 3C1(.25)^1(.75)^2 = 0.421875$

10. The population of weights X of slow pitch softball bats produced by a manufacturer is normally distributed with mean 28 ounces and standard deviation of 2.1 ounces. Consider each bat from the manufacturer as an iid random variable from the population.

(a) **(4pts)** What is the difference in bat weight between the median and first quartile weights?

Percentile: $P(Z \leq z) = 0.25$ when $z = -.675 = \frac{Obs-28}{2.1}$ so $Obs = -.675 * (2.1) + 28 = 26.5825$. Median$= 28$. Difference $= 28 - 26.5825 = 1.4175 ounces$

(b) **(4pts)** In an order of 10 bats, find the probablily that the mean weight is between 29 and 28 ounces.

$\bar{X} \sim N(28, 2.1/sqrt(10)) = N(28, 0.664)$ $P(\bar{X} < 29) = P(Z < \frac{29-28}{.664} = P(Z < 1.506) = 0.9339664 - .50 = 0.434$.

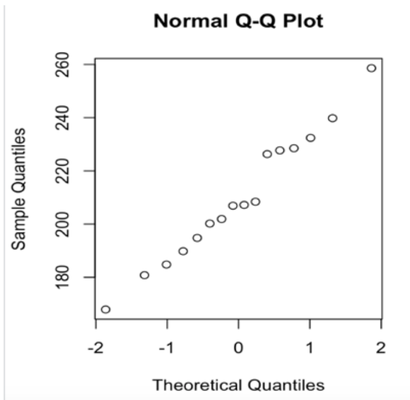
(c) **(4pts)** What total weight capacity, **t** of container would the company need to ship 30 bats in to have 0.95 probability that the total weight of the 30 bats is less than **t**?

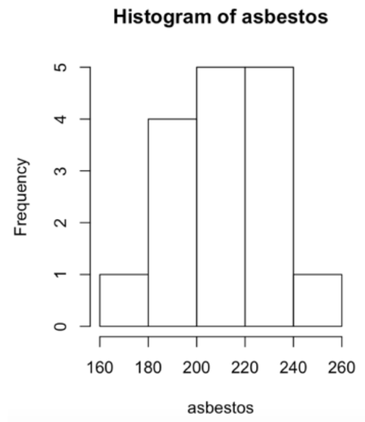
$\bar{X} \sim N(28, 2.1/sqrt(30)) = N(28, 0.383)$ $P(\bar{X} < T/30) = P(Z < \frac{t/30-28}{.383} = 1.645$ so $\frac{t}{30} = 30 * (1.645 * .383 + 28) = 858.901$.

(d) **(4pts)** One can think of the total weight of n bats shipped as the random variable $T = X_1 + X_2 + ... + X_n = n * \bar{X}$. Describe the distribution of the total weight of n bats. Does this distribution require that the population from which $X_1, X_2, ...X_n$ is drawn is normally distributed? Why or why not?

Since $\bar{X} \sim N(28, 2.1^2/(n)$, we know that $T \sim N(E(n\bar{X}) = 28n, Var(n\bar{X}) = n^2 * Var(\bar{X}) = n\sigma^2)$. If n is large enough, the distribution of T will be normal even if the population from which $X_1, X_2, ...X_n$ is drawn is not because the CLT tells us that $\bar{X}$ will be normally distributed so $T = n * \bar{X}$ would be also.

11. The effects of long term asbestos fiber on lung functionality was studied by recording pulmonary compliance (cm^3/cmH_2O) for 16 different workers that were each exposed to asbestos for one month. The sample data is shown in the graphs below.





(a) **(2pts)** Describe why the assumptions for building a confidence interval appear to be well met and what results do those give us?

It appears that the normality of population assumption is met well enough since the qqplot looks approximately linear and the histogram of sample data shows a bell curve. Also sample size of 16 reasssures me at minimum CLT makes $\bar{X}$ approximately normal. We also have to assume that these 16 workers are an independent sample

(b) **(4pts)** Statistics were computed by a calculator on the sample data and recorded in the table below. By default your calculator reports two standard deviations, both are given. Construct an 85% confidence interval for μ, the mean pulmonary compliance for all workers exposed to asbestos for one month.

n	$\bar{x}$	Median	Range	σ	s_x
16	209.75	207.05	90.7	23.39	24.16

Since we only have sample data, we need to use a t multiplier and the sample standard deviation: $209.75 \pm 1.52 * \frac{24.16}{\sqrt{16}} = (200.57, 218.93)$

(c) **(3pts)** Identify which of these statements is the most accurate way to interpret the interval (l, u) computed above.

i. We are 85% confident the interval from l to u captures the true sample mean.

ii. We are 85% confident that the next sample mean will be contained in (l, u)

iii. Of all intervals constructed in this way, about 85% of them will capture the true mean, μ.

iv. About 85% of all pulmonary compliance for workers exposed to asbestos for one month will be between l and u.

(d) **(4pts)** Suppose the researchers would like to build a 95% confidence interval, and still reduce the total width to 10 units. Using a z critical value and the standard deviation computed from the sample above, estimate the number of subjects they should recruit.

$ME = 5 = \frac{1.96*24.16}{\sqrt{n}}$ so $n = (\frac{1.96*24.16}{5})^2 = 89.69$ round up to 90.