Homework 3 Soln: Random Variables: Discrete, Continuous, and Combining

Accuracy Points: 35 + 10 Completion points= Total of 45 possible points

1. A chemical supply company ships a certain solvent in 10-gallon drums. Let X represent the number of drums ordered by a randomly chosen customer. Assume X has the following probability mass function (pmf). The mean and variance of X is : $\mu_X = 2.3$ and $\sigma_X^2 = 1.81$:

X	1	2	3	4	5
p(X=x)	0.4	0.2	0.2	0.1	0.1

   (a) Find $P(X \leq 2)$ and describe what it means in the context of the problem.

   Answer: $P(X \leq 2) = 0.6$ *This is the probability a randomly chosen custom orders 2 or fewer drums*

   (b) Let Y be the number of gallons ordered, so $Y = 10X$. Find the probability mass function of Y.

   Answer: **2 points**

Y	10	20	30	40	50
p(X=x)	0.4	0.2	0.2	0.1	0.1

   (c) Find the mean number of gallons ordered μ_Y using the pmf from part (b) and a second time using the expectectation of linear combination formula shown in class.

   Answer: ***4 points; 2 for each way*** $E(Y) = 10 * .4 + 20 * .2 + 30 * .2 + 40 * .1 + 50 * .1 = 23.$ $E(Y) = E(10X) = 10 * E(X) = 10 * 2.3 = 23$

   (d) Find the standard deviation of the number of gallons ordered σ_Y using the pmf from part (b) and a second time using the standard deviation of linear combinationformulas shown in class.

   Answer: ***4 points; 2 for each way*** $Var(Y) = (10-23)^2 * .4 + (20-23)^2 * .2 + (30-23)^2 * .2 + (40-23)^2 * .1 + (50-23)^2 * .1 = 181$. $Var(Y) = Var(10X) = 100 * Var(X) = 100 * 1.81 = 181$; $SD(Y) = \sqrt{181} = 13.45362 = 10 * 1.345362$

   (e) What possible values of Y are within two standard deviations of the mean value (that is, in the interval from $(\mu - 2\sigma, \mu + 2\sigma)$)? What is the probability that the observed value of Y is within two standard deviations of the mean value?

   Answer: ***2 points; 1 point for values; 1 point for percent*** $(\mu - 2\sigma = 23 - 2 * 13.45 = -3.9, \mu + 2\sigma = 49.9)$ *contains all values except y=50)* 90% *of customers ordered gallons within 2 sd of the mean amount*

2. Consider a large population which has true mean μ and true standard deviation σ. We take a sample of size 3 from this population, thinking of the sample as the RVs X_1, X_2, X_3 where X_i can be considered iid. We are interested in estimating μ.

   (a) Consider the estimator $\hat{\mu_1} = X_1 + X_2 - X_3$. Is this estimator unbiased?

   Answer: ***2 points; +1show expectation = mu; +1 says unbiased.*** *Since* $E(\hat{\mu_1}) = \mu$*, yes, this is unbiased for* μ.

(b) Find the variance of $\hat{\mu_1}$.

Answer: ***2 points*** $Var(\hat{\mu_1}) = 3\sigma^2$

(c) When estimating μ, explain why someone would prefer the estimator $\bar{X} = \frac{X_1+X_2+X_3}{3}$ over $\hat{\mu_1}$.

Answer: *$\bar{X}$ is unbiased (so mean of $\bar{X}$ is μ). but $\bar{X}$ has lower variability. $Var(\bar{X}) = \frac{\sigma^2}{n}$ compared to $3\sigma^2$*

(d) Now consider the estimator $\hat{\mu_2} = \frac{X_1+X_2+X_3}{2}$. Is this estimator unbiased? Calculate the bias.

Answer: *$E(\hat{\mu_2}) = \frac{3\mu}{2}$, so $\hat{\mu_2}$ is biased. $bias(\hat{\mu_2}) = E(\hat{\mu_2}) - \mu = \frac{3\mu}{2} - \mu = \frac{\mu}{2}$*

(e) Compute the MSE for $\hat{\mu_2}$.

Answer: ***4 points; +1 bias, +1 variance +2 combine correctly.*** $MSE(\hat{\mu_2}) = Var(\hat{\mu_2}) + bias(\hat{\mu_2})^2 = 3/4\sigma^2 + 1/4\mu^2$

3. Let F be an RV that represents the operating temperature in Fahrenheit of one instance of a manufacturing process, and assume $F \sim N(100, Var(F) = 5^2)$. Let C be an RV that represents the same process, but measured in Celsius. Fahrenheit can be converted to Celsius using $C = \frac{5}{9}(F - 32)$. Using Normal table in Canvas, solve for the following (You can also check your answers using R):

(a) Find the probability that one randomly selected instance of the process will have operating temperature greater than 98.6 Fahrenheit.

2 points *Standardizing and using the normal table, $P(F \geq 98.6) = P(Z \geq \frac{(98.6-100)}{5}) = P(Z \geq -0.28) = P(Z \leq 0.28) = 0.61026$.*

(b) Find the distribution of C. (Hint: $C \sim ?(?, ?)$)

3 points, +1 N, +1 E(C), +1 Var(C) *Using our rules of E and VAR, $E(C) = (5/9)(E(F) - 32) = 37.78$, and $VAR(C) = (5/9)^2 VAR(F) = 7.72$. Since any linear function of normals is normal, the distribution of C is normal, thus $C \sim N(37.78, Var(C) = 7.72)$.*

(c) Find the probability that one randomly selected instance of the process will have operating temperature below 32 Celsius.
Standardizing and using the normal table, $P(C \leq 32) = P(Z \leq \frac{(32-37.78)}{\sqrt{7.72}}) = P(Z \leq -2.08) = 0.01876$.

(d) Above what temperature (in Celsius) is the top 10% of operating temperatures?
2 points *$P(C \leq c) = .90$ equivalent $P(Z \leq z) = .90$ when $z = 1.28 = \frac{c-37.78}{2.778}$ so $c = 1.28 * 2.778 + 37.78 = 41.34$*

(e) Find the probability in a sample of 6 instances, more than 4 instances have operating temperature above 32 Celsius.(Assuming observations in the sample are independent)
3 points*Let X be the number of instances in 6 that have operating temperature above 32 Celsius so $X \sim Bin(6, 0.9812)$. $P(X > 4) = P(X = 5) + P(X = 6)$=$\binom{6}{5} * 0.9812^5 * (0.0188)^1 + \binom{6}{6} * 0.9812^6 * (0.0188)^0$= $0.1025881 + 0.8923706 = 0.9949587$*

(f) Find the distribution of $\bar{C}$ for n=6, then find the probability that the average operating temperature in a sample of 6 instances is above 32 Celsius.
5 points; +3 for correct dist of $\bar{C}$; +2 for solving *Since $C \sim N(37.78, \sigma^2 = 7.72)$, so $\bar{C} \sim N(37.78, \sigma^2 = \frac{7.72}{6})$ so $\bar{C} \sim N(37.78, \sigma = \sqrt{\frac{7.72}{6}} = 1.1343)$ Since C is Normal, the sampling distribution of C-bar is also normally distributed since its a linear combination of Normal. $P(\bar{C} > 32) = P(Z > \frac{32-37.78}{1.1343} = -5.095654)$ over 99.9%.*

4. Retail stores experience their heaviest returns on December 26th and December 27th each year. The distribution for the Number of Items Returned (X) for Hilldale Macy's on those days last year is given in the table below. It has mean: $\mu = 2.61$ and variance $\sigma^2 = 1.80$ Assume the probability distribution also holds for this year.

Number of Items	Probability
1	0.25
2	0.28
3	0.20
4	0.17
5	0.08
6	0.02

(a) In this year, a random sample of size 45 returns is selected for review. Describe the sampling distribution of the sample mean (shape, center, and spread).

Answer: *CLT applies so X-bar is approximately normal with mean:* $\mu_{\bar{X}} = 2.61$ *and var:* $\sigma^2_{\bar{X}} = \frac{1.80}{45} = 0.04$. $\bar{X} \sim N(2.61, .2^2)$

(b) What is the probability that the sample mean will be greater than 2.9 items?

Answer: $P(\bar{X} > 2.9) = P(Z \geq \frac{2.9-2.61}{\sqrt{1.80/45}}) = P(Z \geq \frac{2.9-2.61}{.2}) = P(Z \geq 1.45) = 1 - P(Z < 1.45) = 1 - 0.9264707 = 0.07352926$

(c) Find an upper bound b such that the total number of items returned by 45 customers will be less than b with probability 0.95.

Answer: $P(X_1 + X_2 + ... + X_{45} < b) = P(\bar{X} < b/45) = P(Z < \frac{b/45-2.61}{\sqrt{1.8/45}}) = .95$. *So we need* $P(Z \leq z) = .95$, $z = 1.645 = \frac{b/45-2.61}{0.2}$, *so* $b = (1.645 * .2 + 2.61) * 45 = 132.3$ *Round up to 133 as the bound.*

5. We will be exploring the difference between using the standard deviation formula: $s_1 = \sqrt{\frac{(X-\bar{X})^2}{n-1}}$ and the population formula: $s_2 = \sqrt{\frac{(X-\bar{X})^2}{n}}$ through a simulation.

In the code below, I have defined a population of values, named pop1. I have also written a function sample.sd to compute the sample standard deviation on a set of numbers passed in.

(a) Copy and paste the entire chuck of code from set.seed(1) (so that we're ll using the same data) through par(mfrow=c(1,1)) (so we reset the graphics pane). Then copy and paste the three lines of code within the samp.sd function into the pop.sd function. Update the three lines as necessary so the pop.sd function will calculate the population standard deviation formula for a set of values.

(b) Run the entire chunk of code (from set.seed(1) through par(mfrow=c(1,1))). (i) What do you notice about the average of the standard deviations produced using the samp.sd function compared to the pop.sd function compared to the true population standard deviation? (ii) Why might we prefer to use the sample.sd formulation when we have a sample of data and are interested in estimating the population standard deviation? (You can compare the resulting histograms to help you answer the question.)

I notice that the mean of the histogram from the sample (n-1) formulation is closer to the true population sd (red). Using the sd.pop formula resulted in an estimate of the true popultion sd that was biased down furter, since mean of the sd.pop values is further below the true population sd)

```r
set.seed(1)
pop1<-rnorm(10000, 4, 2)
```

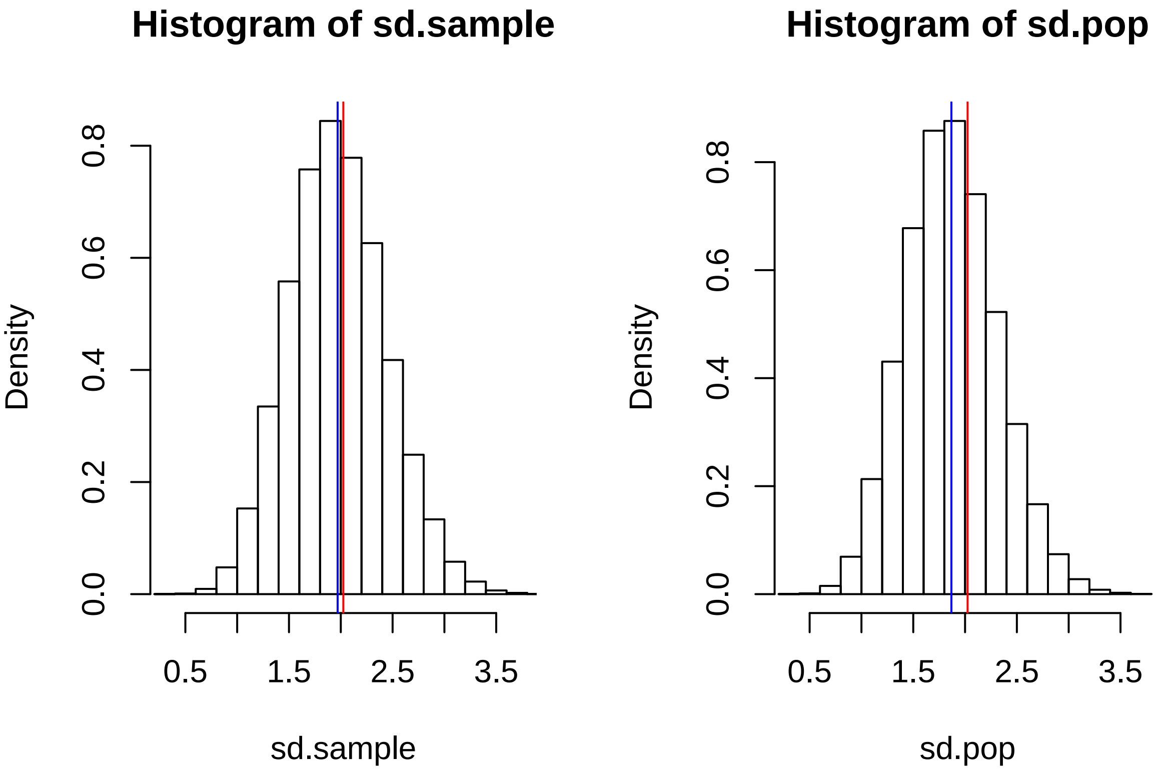
```r
samp.sd<-function(data){
  n<-length(data)
  sum.sq.devs<-sum((data-mean(data))^2)
  av.dev<-sqrt(sum.sq.devs/(n-1))
  return(av.dev)
}

pop.sd<-function(data){
  n=length(data)  #remove for nonsoln
  sum.sq.devs<-sum((data-mean(data))^2)  #remove for nonsoln
  av.dev<-sqrt(sum.sq.devs/n)   #remove for nonsoln
  return(av.dev)
}

#pop.sd<-function(data){
  #fill function based on samp.sd
  #fill function based on samp.sd
  #fill function based on samp.sd
 # return(av.dev)
#}


#Simulation Section
#Building sampling distribution of pop standard deviation estimators
#estimator 1 is sd.sample and estimator 2 is sd.pop formulation
nsamples<-100000
sd.sample<-rep(0, nsamples)
sd.pop<-rep(0, nsamples)
for (i in 1:nsamples){
  samp<-sample(pop1, 10, replace=TRUE)  #taking a new sample of size 10 from population
  sd.sample[i]<-samp.sd(samp)         #calculating and storing sd using sample formula on new sample
  sd.pop[i]<-pop.sd(samp)             #calculating and storing sd using pop formula on new sample
}

#Displaying histograms of the 100000 standard deviations we calculated used
#the population and sample equations;
#adding true population standard deviation in red and
#mean of the simulated standard deviations in blue.
par(mfrow=c(1,2))
hist(sd.sample, freq=FALSE, xlim=c(0.3, 3.75));   #histogram of generated sample sds
abline(v=pop.sd(pop1), col="red"); abline(v=mean(sd.sample), col="blue")
hist(sd.pop, freq=FALSE, xlim=c(0.3, 3.75));      #histogram of generated pop sds
abline(v=pop.sd(pop1), col="red"); abline(v=mean(sd.pop), col="blue")
```



```r
par(mfrow=c(1,1))
```