HOMEWORK 3

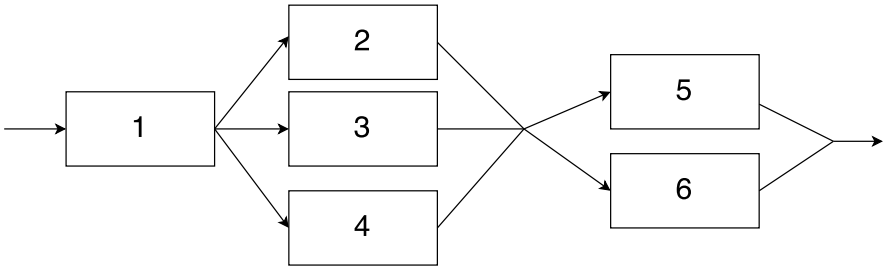
Due Friday, Oct 21, start of class

This homework covers Chapters 4.3, 6.4, 6.5, and 8.1-8.3. You should be working on your homework throughout these two weeks. If you can't solve some of the problems, please come to office hours. Email is fine only for very short questions.

THEORETICAL PORTION

The theoretical problems should be **neatly** numbered, written out, and solved. Do not turn in messy work.

1. Consider a system of 6 components pictured in the following diagram:



For the system to work, **all** of the following have to be satisfied:

- Component 1 has to work.
- At least one of components 2, 3 or 4 has to work.
- At least one of components 5 or 6 has to work.

Now, suppose component 1 has an exponentially distributed lifetime with a mean of 1/2 year. Components 2, 3 and 4 each have an exponentially distributed lifetime with a mean lifetime of 1 year and components 5 and 6 have exponentially distributed lifetimes with mean lifetime of 1.5 year. Assume all components function independently.

What is the probability that the system will function uninterrupted for at least 2 years?

Solution:

For the system to work for at least 2 years, the following things need to happen:

(a) Component 1 must work for at least two years.

(b) *At least one* of components 2, 3, or 4 must work for at least two years.

(c) *At least one* of components 5 or 6 must work for at least two years.

Let X be a r.v. that denotes whether or not component 1 works, let Y be a r.v. that denotes how many of components 2, 3, and 4 work, and let Z be a r.v. that denotes how many of components 5 and 6 work. Since each component is independent, and working for two or more years is a "success", then we can state the following:

- $X \sim B(1, p_x)$
- $Y \sim B(3, p_y)$
- $Z \sim B(2, p_z)$

where p_x, p_y, and p_z are the probability the components last for more than 2 years. We can then write:

$$\begin{aligned} P(\text{system lasts} \geq 2 \text{ years}) &= P(X=1) \times P(Y \geq 1) \times P(Z \geq 1) \\ &= P(X=1) \times (1 - P(Y=0)) \times (1 - P(Z=0)) \\ &= \binom{1}{1} p_x^1 (1-p_x)^{1-1} \times \left(1 - \binom{3}{0} p_y^0 (1-p_y)^{3-0}\right) \times \left(1 - \binom{2}{0} p_z^0 (1-p_z)^{2-0}\right) \\ &= p_x \times \left(1 - (1-p_y)^3\right) \times \left(1 - (1-p_z)^2\right). \end{aligned}$$

All the components are exponentially distributed (i.e. $f(x) = \lambda \exp(-\lambda x), F(x) = \int_0^x f(t)dt = 1 - \exp(-\lambda x)$), but have different average lifetimes, so the parameter λ is different for both.

- Component 1 has a mean lifetime of 1/2 a year, so is distributed exponential with parameter $\lambda = 2$. This means that
$$p_x = 1 - F(2, \lambda = 2) = \exp(-2 * 2) = \exp(-4).$$
- Components 2, 3, and 4 have mean lifetimes equal to 1 year, so are distributed exponential with parameter $\lambda = 1$. So
$$p_y = 1 - F(2, \lambda = 1) = \exp(-2).$$
- Components 5 and 6 have mean lifetimes equal to 1.5 years, so are distributed exponential with parameter $\lambda = 2/3$. So
$$p_z = 1 - F(2, \lambda = 2/3) = \exp(-4/3).$$

We can then actually calculate the probability the system lasts more than two years:

$$\begin{aligned} P(\text{system lasts} \geq 2 \text{ years}) &= p_x \times \left(1 - (1 - p_y)^3\right) \times \left(1 - (1 - p_z)^2\right) \\ &= \exp(-4)\left(1 - (1 - \exp(-2))^3\right)\left(1 - (1 - \exp(-4/3))^2\right) \\ &= 0.002963801. \end{aligned}$$

This can also be done in R:

```
px = 1-pexp(2, rate = 2)
py = 1-pexp(2, rate = 1)
pz = 1-pexp(2, rate = 2/3)

p_last2plusyrs = dbinom(1, 1, px)*(1-dbinom(0, 3, py))*(1-dbinom(0, 2, pz))
```

2. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.
   (a) Show (without integration), that $Var(\bar{X}) = \sigma^2/n$.
   (b) What is $E(\bar{X}^2)$?

   Solution:

   (a)
$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\ &= \frac{1}{n^2}(Var(X_1) + Var(X_2) + \cdots + Var(X_n)) \text{ Since they are i.i.d., no need to worry about covariance terms} \\ &= \frac{1}{n^2}\left(\sigma^2 + \cdots + \sigma^2\right) \\ &= \frac{1}{n^2} \times n\sigma^2 \\ &= \sigma^2/n \end{aligned}$$

   (b) We know that $Var(\bar{X}) = E(\bar{X}^2) - (E\bar{X})^2$, therefore $E(\bar{X}^2) = Var(\bar{X}) + (E\bar{X})^2$. We know that $E\bar{X} = \mu$, so
$$E(\bar{X}^2) = \sigma^2/n + \mu^2.$$

3. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and $Y \sim \mathcal{N}(\theta, \gamma)$. Assume X and Y are independent. Show all steps in solving the problems below:
   (a) What is E(X+Y)? Var(X+Y)?
   (b) What is the distribution of X+Y and how do you know?
   (c) What is the distribution of 3X+20Y+8? Include not only the type of distribution but also the values of any parameters of that distribution.

   Solution:

(a) $E(X+Y) = EX + EY = \mu + \theta$, and $Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y) = \sigma^2 + \gamma^2$.

(b) $X + Y \sim \mathcal{N}\left(\mu + \theta, \sigma^2 + \gamma^2\right)$, because a linear function of normally distributed random variables is normally distributed.

(c) $3X + 20Y + 8 \sim \mathcal{N}\left(3\mu + 20\theta + 8, 9\sigma^2 + 400\gamma^2\right)$.

4. Let $X \sim U(a,b)$. What is $E(X^3)$? Show your steps. *Major hint:* $E(X^3) \neq (E(X))^3$*!*

Solution:

$$\begin{aligned}
E(X^3) &= \int_a^b x^3 \frac{1}{b-a} dx \\
&= \frac{1}{b-a} \int_a^b x^3 dx \\
&= \frac{1}{b-a} \left(\frac{1}{4} x^4 \mid_a^b \right) \\
&= \frac{1}{4(b-a)} \left(b^4 - a^4\right) \\
&= \frac{\left(b^2 - a^2\right)\left(b^2 + a^2\right)}{4(b-a)} \\
&= \frac{(b-a)(b+a)\left(b^2 + a^2\right)}{4(b-a)} \\
&= \frac{1}{4}(b+a)\left(b^2 + a^2\right) \\
&= \frac{1}{4}\left(b^3 + a^2 b + a b^2 + a^3\right)
\end{aligned}$$

5. You are given a data set of time to recovery after knee surgery for a sample of 22 elderly women.

(a) Write an equation for the 98% confidence interval you would use for this data, assuming it is normally distributed.

(b) You make a histogram of the data, and realize it is not normally distributed, but in fact looks like it is exponentially distributed. What could you do to improve on your original confidence interval?

Solution:

(a) If the data is normally distributed, then the equation for our confidence interval is:

$$\bar{x} \pm t_{.01,21} \times s/\sqrt{n},$$

where s is the sample standard deviation, calculated as $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

(b) In the instance where the data is not normally distributed, we have to get a little clever. There are a number of sufficient ways to do this, but a more robust way would be to assume the data are exponentially distributed. We then could calculate the 2.5% and 97.5% of the exponential distribution, replacing λ with $1/\bar{x}$ (the MLE). We would then anticipate that the true λ would fall in this interval approximately 95% of the time. More advanced methods would include techniques such as bootstrapping or jackknifing.

6. Under specific conditions, the drying time of a certain type of paint is normally distributed with a mean value of 55 minutes, and a standard deviation of 7 minutes. Use only the values given to you to solve this problem:

$$\Phi(0.385) = 0.65, \Phi(0.428) = 0.67, \Phi(0.71) = 0.762, \Phi(1.43) = 0.924, \Phi(1.92) = 0.973.$$

$$\Phi^{-1}(0.972) = 1.917, \Phi^{-1}(0.740) = 0.643, \Phi^{-1}(0.95) = 1.645$$

$$z_{0.0017} = 2.923, z_{0.016} = 2.14, z_{0.027} = 1.93.$$

(a) What is the probability that the paint will dry in 45 minutes or less?

(b) What is the probability that the paint will dry between 50 and 70 minutes?

(c) How many minutes represents the 35^{th} percentile in paint drying?

Solution:

(a)

$$P(X < 45) = P\left(\frac{X-55}{7} < \frac{45-55}{7}\right) = P(Z < -1.43)$$
$$= P(Z > 1.43) = 1 - P(Z < 1.43) = 1 - \Phi(1.43) = 1 - 0.924 = 0.076.$$

(b)

$$P(50 < X < 70) = P\left(\frac{50-55}{7} < \frac{X-55}{7} < \frac{70-55}{7}\right) = P(-0.71 < Z < 2.14)$$
$$= \Phi(2.14) - \Phi(-0.71) = \Phi(2.14) - (1 - \Phi(0.71))$$
$$= 1 - 0.016 - (1 - 0.762) = 0.746. \text{ where the first number is taken from } z_{0.016} = 2.14.$$

(c) First, represent the 35^{th} percentile: $P(X < x_{0.35}) = 0.35 \longrightarrow P(\frac{X-55}{7} < z_{0.35}) = 0.35$. This can be re-written as: $\Phi\left(\frac{X-55}{7}\right) = 0.35$. This value is not available above, but we do know that $P(\frac{X-55}{7} < z_{0.35}) = 0.35 \rightarrow P(\frac{X-55}{7} > z_{0.35}) = 0.65$, which is available to us. So $z_{0.35} = -0.385$.
Then solve for X:

$$\frac{X-55}{7} = -0.385 \longrightarrow X = -0.385 * 7 + 55 = 52.305.$$

So, 52.305 minutes represents the 35^{th} percentile of the distribution of the time it takes the paint to dry.

7. A rock specimen is randomly selected and weightd two different times. Let w denote the true weight (a number) of the rock, and let X_1 ad X_2 be the two measured weights. Then, $X_1 = w + E_1$ and $X_2 = w + E_2$, where E_1 and E_2 are the two measurement errors. Suppose that E_1 and E_2 are independent and distributed normally with mean 0 and variance 0.1 (*i.e.*, E_1, $E_2 \sim \mathrm{N}(0, 0.1)$).

(a) What is the mean of X_1? What is the mean of X_2?

(b) What is $V(X_1)$? What is $V(X_2)$?

(c) What is $\mathrm{Corr}(X_1, X_2)$?

Solution:

(a) $EX_1 = E(w + E_1) = w + E(E_1) = w + 0 = w$. Similarly, $EX_2 = w$.

(b) $Var(X_1) = Var(w + E_1) = Var(E_1) = 0.1$. Similarly, $Var(X_2) = 0.1$.

(c) The correlation between the two variables is 0, since they are independent (because E_1 and E_2 are independent), since:

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}$$
$$= \frac{E(X_1X_2) - EX_1 \cdot EX_2}{0.1}$$
$$= \frac{E\left((w + E_1)(w + E_2)\right) - w^2}{0.1}$$
$$= \frac{E\left(w^2 + wE_1 + wE_2 + E_1 \cdot E_2\right) - w^2}{0.1}$$
$$= \frac{w^2 + 0 + 0 + E(E_1 \cdot E_2) - w^2}{0.1}$$
$$= \frac{w^2 + E(E_1)E(E_2) - w^2}{0.1}$$

(If two r.v.s X and Y are independent, then $E(XY) = EX \cdot EY$) $= \frac{w^2 + 0 \cdot 0 - w^2}{0.1} = 0$

8. Let

$$f(x) = \begin{cases} \frac{x^3}{4} & \text{if } 0 < x < c \\ 0 & \text{otherwise.} \end{cases}$$

(a) What value of c is needed so that $f(x)$ is a pdf?

(b) Calculate $F(x)$.

(c) In class we discussed that for any cdf, $F(x) \sim U(0,1)$. For this problem, calculate the inverse of $F(x)$ using this property. i.e., if $F(x) = U$, solve for x in terms of U.

Solution:

(a)

$$\begin{aligned} \int_0^c f(x)dx = \int_0^c \frac{x^3}{4}dx &= 1 \\ \rightarrow &= \frac{x^4}{16}\Big|_0^c = 1 \\ \rightarrow &= c^4 = 16 \rightarrow c = \pm 2 \end{aligned}$$

We know that c must equal 2 (not negative 2), since 2 is greater than 0. So, $f(x) = \frac{x^3}{4}, 0 < x < 2$.

(b)

$$F(x) = \int_0^x \frac{t^3}{4}dt = \frac{t^4}{16}\Big|_0^x = \frac{x^4}{16}.$$

(c)

$$\begin{aligned} F(x) = \frac{x^4}{16} &= U \\ \rightarrow x^4 &= 16U \\ \rightarrow x &= \pm 2U^{1/4} \\ \rightarrow x &= +2U^{1/4} \text{ because } U \text{ must be between 0 and 1.} \end{aligned}$$

9. **APPM 5570 students only:** Let $X_1, X_2, \ldots, X_{15} \overset{iid}{\sim} f(x)$.

(a) Write down the equation for a confidence interval for the mean if $f(x)$ is the normal distribution with known parameter σ.

(b) Write down the equation for a confidence interval for the mean if $f(x)$ is the normal distribution with unknown σ.

(c) If $f(x)$ is a Poisson distribution with parameter λ, derive a reasonable equation you could use to calculate an α-level confidence interval for the mean.

(d) If $f(x)$ is a Uniform distribution with parameters a and b, derive a reasonable equation you could use to calculate an α-level confidence interval for the mean.

Solution:

(a) $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$

(b) $\bar{x} \pm t_{n-1,\alpha/2}s/\sqrt{n}$

(c) A general equation for a confidence interval can be thought of as:

$$\hat{\theta} \pm t_{n-1,\alpha/2}\sqrt{\widehat{Var}\left(\hat{\theta}\right)}. \tag{1}$$

In the Poisson distribution, $\theta = \lambda = E(X)$. We can approximate the mean, $E(X)$, with $\bar{x}$. This means that we can then say $\hat{\theta} = \hat{\lambda} = \bar{X}$. Then,

$$Var\left(\bar{X}\right) = \lambda/n \rightarrow \widehat{Var\left(\bar{X}\right)} = \hat{\lambda}/n = \bar{x}/n.$$

We can then re-write Equation (**??**) for the Poisson distribution as:

$$\bar{x} \pm t_{n-1,\alpha/2}\sqrt{\bar{x}/n}.$$

(d) If $f(x)$ is the $Unif(a,b)$ distribution, then $\theta = E(X) = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}$. Since a is the minimum any value can take on, and b is the maximum any value can take on, then we could let the smallest observation estimate a and the largest observation estimate b. Let the smallest observation be denoted as $X_{(1)}$ and the largest observation be denoted as $X_{(15)}$. Then,

$$\hat{\theta} = \frac{\widehat{a+b}}{2} = \frac{X_{(1)} + X_{(15)}}{2}.$$

The variance of $X_{(1)}$ and $X_{(15)}$ are very complicated to calculate theoretically (although it can be done), so we don't do that here. Instead, we *approximate* as follows:

$$Var\left(\hat{\theta}\right) = Var\left(\frac{\widehat{a+b}}{2}\right) = \frac{1}{4}\left(Var(X) * 2\right) = \frac{Var(X)}{2} = \frac{\frac{(b-a)^2}{12}}{2} = \frac{(b-a)^2}{6}.$$

$$\rightarrow \widehat{Var\left(\hat{\theta}\right)} = \frac{(\hat{b}-\hat{a})^2}{6} = \frac{(X_{(15)} - X_{(1)})^2}{6}.$$

In this way, we re-write Equation (**??**) as:

$$\frac{X_{(1)} + X_{(15)}}{2} \pm t_{n-1,\alpha/2}\sqrt{\frac{(X_{(15)} - X_{(1)})^2}{6}}.$$

COMPUTATIONAL PORTION

The computational portion of your homework should be neatly done and include all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do *not* put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade. **LABELS ARE YOUR FRIEND, USE THEM.** If you turn in something that is messy or out of order, it will be returned to you with a zero. All computations should be done using R, which can be downloaded for free at https://cran.r-project.org/.

1. Aphid infestation of fruit rees is usually controlled via pesticides or via ladybug inundation. In a particular area, 2 different (and well isolated) groves, with 15 fruit trees each, are selected for an experiment. The trees in both groves are of the same age, roughly the same size and can be assumed to be independent. One grove is sprayed with pesticides, and one is infested with ladybugs. The fruit yield (in pounds) for each tree is given below:

   Treatment #1, Grove with pesticide:

   55.57109, 36.50319, 47.80090, 33.34822, 36.16251,
   35.28337, 41.50154, 44.18931, 40.81439, 33.88648,
   44.90427, 49.97089, 22.85414, 27.84301, 38.49843

   Treatment #2, Grove with ladybugs:

   45.44505, 35.52320, 46.97865, 45.76921, 41.66216,
   54.69599, 58.77678, 49.08538, 48.53812, 70.17137,
   51.86253, 39.59365, 42.10194, 47.39945, 39.04648

   You can read in this data using the "HW3TreeData.txt" data set.

   (a) Plot a relative frequency histogram of the yields in the two groves. Make sure both histograms have the same range on the x-axis.

   (b) Comment on the histogram shapes. Which densities do they resemble? In particular do they appear normal?

   (c) Find the sample mean of yields for the two groves.

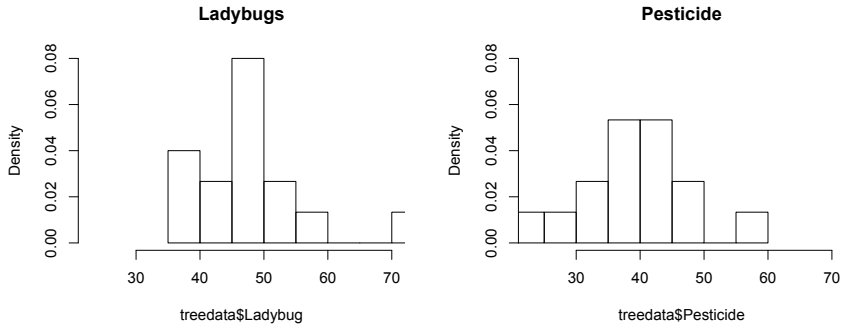
   (d) Assuming both samples come from a normal population and using your answer from part (c), provide the two 95% confidence intervals for the true mean yields for trees under the two treatments.

   (e) Interpret the confidence interval you constructed for the grove treated with pesticides (*i.e.* what does a confidence interval really signify?)

   (f) Find(or approximate) the sample mean variance, *i.e.* variance of the sample mean, $\overline{X}$, for the yields in the two groves.

   (g) Using a chi-square distribution, construct the 95% confidence interval for the true variance of yield for both groves.

   Solution:



   (a)

(b) We can see from the histograms that the trees with pesticide seem to have a lower yield than the ladybug trees, and also a larger variance (the data is spread further apart). The distribution of the yield for the pesticide trees seems to be normally distributed. The yield from the lady bug trees also appears to be symmetric, although the distribution is not as symmetric, and there appears to be an outlier with a high yield.

(c) The mean yield for the lady bug trees is 47.8, and the mean yield for the pesticide trees is 39.3

(d) Since $n = 15$ for both groups, we cannot use the CLT. However, it appears the data is normally distributed (based on the histograms), and we can estimate σ^2 with s^2, so we can use the t distribution. The equation we want is:

$$\bar{x} \pm t_{n-1,\alpha/2} * s/\sqrt{n}.$$

When we calculate these confidence intervals, we get $(43.0, 52.6)$ for the mean yield of the ladybug trees, and we get $(34.5, 44.0)$ for the mean yield of the pesticide trees.

(e) The 95% confidence interval for the yield from the pesticide trees is between 34.56 and 43.99 pounds. This means that 95% of the time, the true mean yield for the pesticide trees will be in these bounds. In other words, if we do this same experiment 100 times, and create the 95% confidence interval around the sample mean, we expect that 5 of those interval will not contain the true mean yield for pesticide trees.

(f) Note that by a proposition in 5.4 on p. 223 we have that $V(\overline{X}) = \dfrac{\sigma^2}{n}$ but since we do not know σ^2 we use $\sigma^2 \approx s^2$, thus we approximate the "sample mean variance", i.e. variance of $\overline{X}$ as $V(\overline{X}) = \dfrac{\sigma^2}{n} \approx \dfrac{s^2}{n}$ thus, the variance of the sample mean yield of the ladybug trees is 5.06, and the variance of the sample mean yield of the pesticide trees is 4.83.

(g) The equation we want to use to create a confidence interval for the variance is:

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}.$$

The resulting intervals are:

$$(2.71 < \sigma_L^2 < 12.6), (2.6 < \sigma_P^2 < 12.0).$$

Code for this problem is below:

```
##############################################################################################
#
# Problem 1
#
##############################################################################################
# Read in the data set:
treedata = read.table('HW3TreeData.txt', header = TRUE)
# The first column is not numeric, so remove it

# (a) Plot both relative frequency hists on the same x-axis and same figure.
par(mfrow = c(1,2))
hist(treedata$Ladybug, xlim = range(treedata), main = 'Ladybugs', freq = FALSE, ylim = c(0,0.08))
hist(treedata$Pesticide, xlim = range(treedata), main = 'Pesticide', freq = FALSE, ylim = c(0,0.08))

# (c)
ladybug.mean = mean(treedata$Ladybug)
pesticide.mean = mean(treedata$Pesticide)

# (d)
# Calculate the t value used for these CIs
t.CI = qt(.975, df = 14)
# Ladybug lower and upper bounds:
ladybug.mean-t.CI*sd(treedata$Ladybug)/sqrt(15)
```

```
ladybug.mean+t.CI*sd(treedata$Ladybug)/sqrt(15)
# Pesticide lower and upper bounds:
pesticide.mean-t.CI*sd(treedata$Pesticide)/sqrt(15)
pesticide.mean+t.CI*sd(treedata$Pesticide)/sqrt(15)

# (f)
ladybug.var = var(treedata$Ladybug)/15
pesticide.var = var(treedata$Pesticide)/15

# (g)
# Calculate the chi-square value needed for these CIs
# Lower and upper bound chi-square values are different!
chisq.CI.LB = qchisq(.975, 14)
chisq.CI.UB = qchisq(.025, 14)
# Ladybug lower and upper bounds:
14*ladybug.var/chisq.CI.LB
14*ladybug.var/chisq.CI.UB
# Pesticide lower and upper bounds:
14*pesticide.var/chisq.CI.LB
14*pesticide.var/chisq.CI.UB
```

2. Generate 500 samples (each with $n = 300$) from an exponential distribution with a mean equal to $1/2$. Retain the entire first sample and store it in "expSample", and then in "means.of.expSamples" retain only the means of the remaining 499 samples.

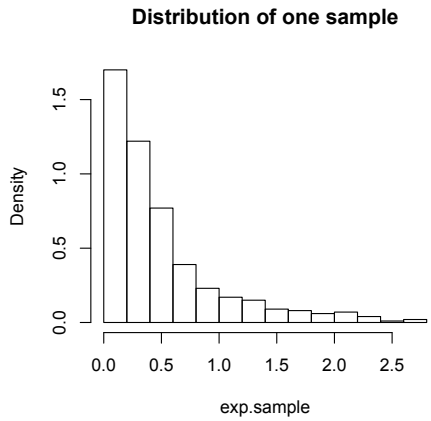
   (a) Report the mean and standard deviation for "expSample" and "means.of.expSamples".

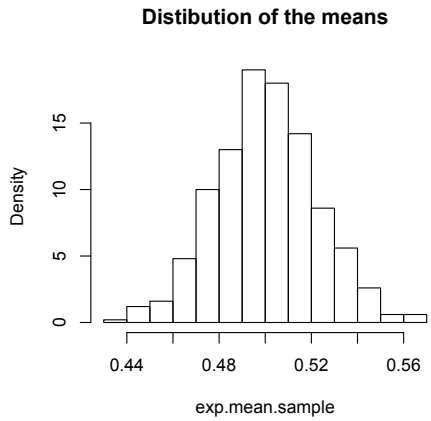
   (b) Create a histogram for "expSample" and for "means.of.expSamples".

   Both of these data sets originate from the same distribution. Why are they so different? What do you notice about the the means and the standard deviations of the samples?

   Solution:

   (a) In the first data set, my mean is 0.49 and my standard deviation is 0.51, which is about what I expected since I specified a random sample from an exponential distribution with $\lambda = 2$, so the mean and standard deviation are each equal to 0.5. In my second data set of stored means, my mean is 0.50, and the standard deviation of is 0.022. This is also what I expected, since $E(\bar{X}) = 1/\lambda = 0.5$, and $sd(\bar{X}) = (1/\lambda)/\sqrt{n} = 0.5/\sqrt{500} = 0.022$. The histogram shows that the distribution of my first sample looks exactly like the exponential distribution, but the distribution of my second sample looks like the normal distribution. *This is why the Central Limit Theorem works!* Even though our original distribution is not normal, the distribution of our sample means is. Therefore, we can use the standard normal distribution to perform hypothesis tests, calculate confidence intervals, and perform other types of inference on data sets. It is because we know the distribution of the mean (when n is large enough) is a normal distribution.





Code is below:

```
##############################################################################################
#
# Problem 2
#
##############################################################################################
# (a)
normal.sample = rnorm(500, mean = 5, sd = 10)
hist(normal.sample, main = 'Normal Sample')
mean(normal.sample)
sd(normal.sample)
# mean.sample will keep track of the means for all of the 100 samples
mean.sample = NULL
for(i in 1:500){
    mean.sample = c(mean.sample, mean(rnorm(500, mean = 5, sd = 10)))
}
mean(mean.sample)
sd(mean.sample)
# Put the two histograms (of one sample and of the means) side by side
par(mfrow = c(1,2))
hist(normal.sample, main = 'Distribution of one sample', freq = FALSE)
hist(mean.sample, main = 'Distribution of the means', freq = FALSE)
# exp.sample will hold the one sample of size 500
exp.sample = rexp(500, rate = 2)
# exp.mean.sample will hold the 500 means from the 500 samples (each of size 500)
exp.mean.sample = NULL
for(i in 1:500){
    exp.mean.sample = c(exp.mean.sample, mean(rexp(500, rate = 2)))
}
mean(exp.sample)
sd(exp.sample)
mean(exp.mean.sample)
sd(exp.mean.sample)
par(mfrow = c(1,2))
hist(exp.sample, main = 'Distribution of one sample', freq = FALSE)
hist(exp.mean.sample, main = 'Distibution of the means', freq = FALSE)
```

3. Let X be a normally distributed random variable with mean 3 and variance 4.

(a) Let $Y = 5X + 2$, what is the distribution of Y? What are it's mean and variance? (You don't need the computer for this first one.)

(b) Find $P(Y < 10)$ and $P(X < 10)$.

(c) What value of Y marks the 67^{th} percentile?

(d) The 25^{th} percentile of the standard normal distribution is -0.674. How can you use this information to find the value of X that marks the 25^{th} percentile?

(e) Use a normal table to calculate the answers for parts (b)-(c). Make sure your answers match. I have no way of knowing if you actually do this or not, but this is a good time to practice this skill.

Solution:

(a) A linear function of normally distributed random variables is also normally distributed. So,

$$EY = 5EX + 2 = 5(3) + 2 = 17,$$

and

$$Var(Y) = 5^2 Var(X) = 25(4) = 100.$$

(b) $P(Y < 10) = 0.2419637, P(X < 10) = 0.9997674.$

(c) We need to find y such that $P(Y < y) = 0.67$. We can use R to determine that this value is 21.399.

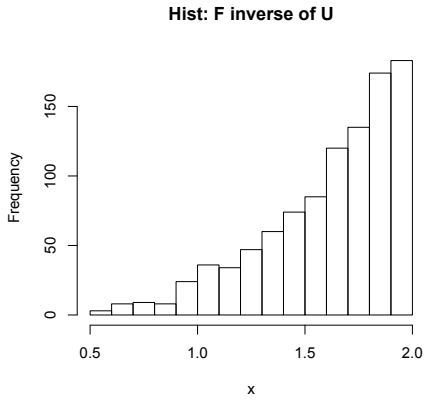
(d) We need to find an x such that $P(X < x) = 0.25$. Since we are given the value of z at the 25^{th} percentile of the standard normal distribution, we adjust:

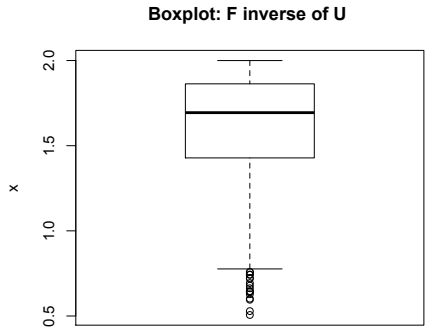
$$\begin{aligned} &P(Z < -0.674) = 0.25 \\ &P\left(\frac{X-3}{\sqrt{4}} < -0.674\right) = 0.25 \\ &\frac{X-3}{\sqrt{4}} = -0.67 \\ &X = 1.65102 \end{aligned}$$

Code for this problem:

```
##############################################################################################
#
# Problem 3
#
##############################################################################################
# (b)
pnorm(10, mean = 17, sd = sqrt(100)) # Same as pnorm((10-17)/sqrt(100), mean = 0, sd = 1)
pnorm(10, mean = 3, sd = sqrt(4)) # Same as pnorm((10-3)/sqrt(4), mean = 0, sd = 1)
# (c)
qnorm(.67, mean = 17, sd = sqrt(100)) # Same as qnorm(.67, 0, 1)*sqrt(100)+17
```

4. Refer back to theoretical problem number 8. Generate a random sample of size 1000 from your distribution and make a histogram and a boxplot. What features do you notice?





Solution:

Code:

```
##############################################################################################
#
# Problem 4
#
##############################################################################################
set.seed(511379)

Usample = runif(1000)
Finverse = 2*Usample^.25

par(mfrow = c(1,2))
hist(Finverse, main = 'Hist: F inverse of U', xlab = 'x')
boxplot(Finverse, main = 'Boxplot: F inverse of U ', ylab = 'x', xlab = '')
```

5. **APPM 5570 students only:** Refer back to problem 9 in the theoretical section. Create 3 data sets, where each row of the data set is a sample of 15 i.i.d random variables, and each sample is taken 1000 times (i.e., your dataset is a 1000 by 15 matrix). The random variables are distributed as follows:

   - Dataset 1: Normally distributed with mean 5 and standard deviation equal to 2.3.
   - Dataset 2: Poisson distributed with λ equal to 5.
   - Dataset 3: Uniform distributed with $a = 1, b = 9$.

   For each row of each data set, you should calculate the confidence interval using each of your four methods, and count how many times the true mean (which is equal to 5 for all data sets) is in the interval. Keep track of this for each data set for each type of interval and report these numbers in a table.

   What do you notice about your results? Did anything surprise you?

   Solution: Code for this problem is below:

R.V. Distribution	Normal CI	t CI	Poisson CI	Uniform CI
Normal	0.061	0.043	0.048	0
Poisson	0.054	0.043	0.027	0
Uniform	0.049	0.047	0.041	0

Table 1: Table of the percentage of times the mean, 5, was not in the bounds of the 95% confidence interval.

```
##############################################################################################
#
# Problem 5
```

```
#
#############################################################################################
################## Make functions for each type of CI: #################
z.025 = qnorm(.975)
t.025 = qt(.975, 14)
# When sigma is known
CI.z = function(x){
    mean.x = mean(x)
    n = length(x)
    CIl = mean.x - z.025*sqrt(5/n)
    CIu = mean.x + z.025*sqrt(5/n)
    return(c(CIl, CIu))
}
# When sigma is unknown
CI.t = function(x){
    mean.x = mean(x)
    sd.x = sd(x)
    n = length(x)
    CIl = mean.x - t.025*sd.x/sqrt(n)
    CIu = mean.x + t.025*sd.x/sqrt(n)
    return(c(CIl, CIu))
}
# For Poisson
CI.po = function(x){
    mean.x = mean(x)
    n = length(x)
    CIl = mean.x - t.025*sqrt(mean.x/n)
    CIu = mean.x + t.025*sqrt(mean.x/n)
    return(c(CIl, CIu))
}
# For uniform
CI.un = function(x){
    mean.x = (min(x)+max(x))/2
    var.x = (max(x)-min(x))^2/6
    CIl = mean.x - t.025*sqrt(var.x)
    CIu = mean.x + t.025*sqrt(var.x)
    return(c(CIl, CIu))
}
################## Generate data #################
set.seed(100234)
data.norm = matrix(rnorm(15*1000, mean = 5, sd = 2.3), nrow = 1000)
data.pois = matrix(rpois(15*1000, lambda = 5), nrow = 1000)
data.unif = matrix(runif(15*1000, min = 1, max = 9), nrow = 1000)

################## Calculate the CIs for each data set #################
# Normal
cis.z = cis.t = cis.po = cis.un = 0
test = NULL
for(ctr in 1:1000){
        ciz = CI.z(data.norm[ctr,])
        cit = CI.t(data.norm[ctr,])
        cipo = CI.po(data.norm[ctr,])
        ciun = CI.un(data.norm[ctr,])

        if(5 < ciz[1] | ciz[2] < 5){    cis.z = cis.z + 1   }
```

```
        if(5 < cit[1] | cit[2] < 5){    cis.t = cis.t + 1   }
        if(5 < cipo[1] | cipo[2] < 5){    cis.po = cis.po + 1   }
        if(5 < ciun[1] | ciun[2] < 5){  cis.un = cis.un + 1     }
}
out.table = c(cis.z, cis.t, cis.po, cis.un)

# Poisson
cis.z = cis.t = cis.po = cis.un = 0
for(ctr in 1:1000){
    ciz = CI.z(data.pois[ctr,])
    cit = CI.t(data.pois[ctr,])
    cipo = CI.po(data.pois[ctr,])
    ciun = CI.un(data.pois[ctr,])

    if(5 < ciz[1] | ciz[2] < 5){    cis.z = cis.z + 1   }
    if(5 < cit[1] | cit[2] < 5){    cis.t = cis.t + 1   }
    if(5 < cipo[1] | cipo[2] < 5){    cis.po = cis.po + 1   }
    if(5 < ciun[1] | ciun[2] < 5){    cis.un = cis.un + 1   }
}
out.table = rbind(out.table, c(cis.z, cis.t, cis.po, cis.un))

# Uniform
cis.z = cis.t = cis.po = cis.un = 0
for(ctr in 1:1000){
    ciz = CI.z(data.unif[ctr,])
    cit = CI.t(data.unif[ctr,])
    cipo = CI.po(data.unif[ctr,])
    ciun = CI.un(data.unif[ctr,])

    if(5 < ciz[1] | ciz[2] < 5){    cis.z = cis.z + 1   }
    if(5 < cit[1] | cit[2] < 5){    cis.t = cis.t + 1   }
    if(5 < cipo[1] | cipo[2] < 5){    cis.po = cis.po + 1   }
    if(5 < ciun[1] | ciun[2] < 5){    cis.un = cis.un + 1   }
}
out.table = as.data.frame(rbind(out.table, c(cis.z, cis.t, cis.po, cis.un))/ctr)
names(out.table) = c('norm','t','pois','unif')
row.names(out.table) = c('norm', 'pois','unif')
```

DATA FOR FINAL PROJECT

One of the components of our final project will involve a technique called "microsimulations". Microsimulations are a type of simulations that are often used to determine how a new policy or intervention will impact a community 5, 10, or even 50 years in the future. A common way in which microsimulations are often used are in assessing traffic patterns. For example, we might do microsimulations to assess the impact of a new lane on a highway will impact traffic flow. Other examples of ways in which microsimulations may be used include examining the impact or effects of:

- A certain economic policy on job rates.
- Addition of a non-native plant species on the habitat of a certain animal.
- Increasing Medicare coverage on the rate of cancer diagnoses in the elderly.

- Increased education spending on poverty rates.

Your job is to find information on an intervention that can be used for this part of your final project. The information must satisfy these requirements:

1. You must have specific numbers/information about how the current scheme works.
2. You must have specific numbers/information about how the proposed scheme may work or is supposed to work.
3. The information must come from a reliable source. This can be a journal article or reputable news source. Be scientific here.
4. **APPM 5570:** You must have specific numbers/information about the margin of error for the proposed intervention. This means you need either standard errors or confidence intervals for the possible outcomes(s).

You need to email me your proposed project idea by **October 28th**. You should include a brief summary of why this intervention could be important (or perhaps useless if you don't think it is a good intervention), the numbers you have, and your sources. It should be about 1 paragraph long (longer is okay). I will read your proposal and make sure you have enough information for the project to work.