# QBS 120 - Problem Set 4

Rob Frost

***Grading: problems 1 and 6 are each worth 5 pts. See specific grading metrics below.***

1. **(Based on Rice, Chapter 6, Problem 3) Let $\bar{X}$ be the average of a sample of $n$ independent standard normal RVs.**

   (a) **Determine $c$ such that $P(|\bar{X}| < c) = 0.5$. Solve for $c$ as a function of $n$.**

   ***Grading: 2 pts total, 1 pt if the answer isn't correct but approach looks right***

   For this problem, we are told that $\mu = 0$ and that $\sigma^2 = 1$. We therefore don't need to use a t distribution and can instead model $\bar{X}$ as a normal using the known mean and variance of a sum of independent normal RVs:

   $$\bar{X} \sim \mathcal{N}(0, 1/n)$$

   Since it is often easier to work with standard normal RVs, we can divide $\bar{X}$ by the SD $(1/\sqrt{n})$ to get:

   $$\sqrt{n}\bar{X} \sim \mathcal{N}(0, 1)$$

   We're asked to find c such that. $P(|\bar{X}| < c) = 0.5$. This can be re-expressed in terms of $\sqrt{n}\bar{X}$, which is $\mathcal{N}(0, 1)$, as:

   $$\begin{aligned}
   P(|\bar{X}| < c) &= 0.5 && \\
   P(\sqrt{n}|\bar{X}| < \sqrt{n}c) &= 0.5 && \\
   P(-\sqrt{n}c < \sqrt{n}\bar{X} < \sqrt{n}c) &= 0.5 && \\
   P(\sqrt{n}\bar{X} < \sqrt{n}c) - P(\sqrt{n}\bar{X} < -\sqrt{n}c) &= 0.5 && \\
   \Phi(\sqrt{n}c) - \Phi(-\sqrt{n}c) &= 0.5 && \sqrt{n}\bar{X} \sim \mathcal{N}(0, 1) \\
   1 - \Phi(-\sqrt{n}c) - \Phi(-\sqrt{n}c) &= 0.5 && \text{symmetry of normal} \\
   1 - 2\Phi(-\sqrt{n}c) &= 0.5 && \\
   \Phi(-\sqrt{n}c) &= 0.25 && \\
   -\sqrt{n}c &= \Phi^{-1}(0.25) && \\
   c &= -1/\sqrt{n}\Phi^{-1}(0.25) &&
   \end{aligned}$$

   (b) **Using only R *norm() functions for the standard normal distribution, compute the exact value of $c$ for $n = 5, ...100$ and visualize as a plot of $c$ vs. $n$.**

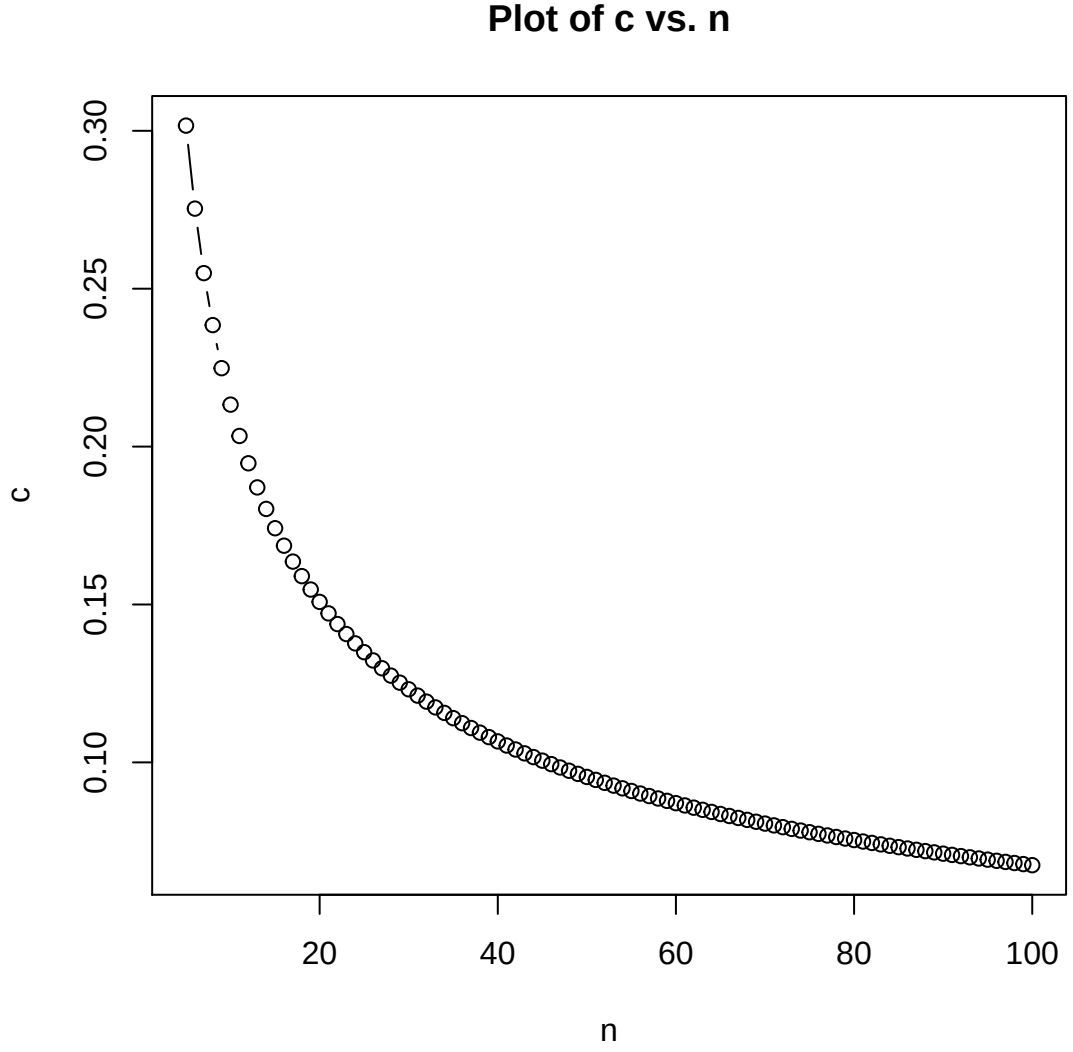
   ***Grading: 1.5 pts total, 1 pt if they generate a graph but the values look wrong***

   Solve for exact value using qnorm() (R function for $\Phi^{-1}()$) for $n = 5, ...100$:

```
> n.vals = 5:100
> c.vals = sapply(n.vals, function(x) {
+         return (-qnorm(0.25)/sqrt(x))
+         })
> names(c.vals) = n.vals
> c.vals[1:10]
        5         6         7         8         9        10        11        12
0.3016410 0.2753593 0.2549332 0.2384681 0.2248299 0.2132924 0.2033663 0.1947084
       13        14
0.1870698 0.1802650
```

Plot $c$ vs $n$

```
> plot(n.vals, c.vals, xlab="n", ylab="c", main="Plot of c vs. n", type="b")
```




(c) **If the variance was not known, how would you solve the problem and what additional piece of information would you need to get an exact answer?**

***Grading: 1 pt total, 0.5 pts to note that a t-distribution is needed, 0.5 pts to note that the sample variance is needed.***

If we were not told the variance of the $X_i$ was 1, we would need to model $\bar{X}$ using a t distribution (assuming we wanted to exact distribution; a normal approximation would also be possible). In this case, we would need to know the sample variance ($S^2$) to generate an RV with a $t_{n-1}$ distribution and get an exact answer for c.

(d) **If the $n$ RVs are independent and have the same distribution with expectation 0 and variance 1 but the exact distribution is not known, how would you approach the problem?**

***Grading: 0.5 pt total. Need to mention CLT approximation and approximate normal distribution.***

If we only know that the RVs are iid with expectation 0 and variance 1, we can rely on the central limit theorem to approximate the sample average by a normal RV. Specifically, $\sqrt{n}\bar{X} \xrightarrow{D} \mathcal{N}(0,1)$. So, our estimate of c would be identical to that found in part (a). In this case, the value is just approximate but per the CLT the approximation becomes better as $n$ grows larger.

2. **(Based on Rice, Chapter 6, Problem 6)**

(a) **Show that if $T \sim t_n$, then $T^2 \sim F_{1,n}$.**
If $T \sim t_n$:

$$T = \frac{Z}{\sqrt{X/n}}$$

Where $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_n$.
The random variable $T^2$ is therefore given by:

$$\begin{aligned} T^2 &= \frac{Z^2}{X/n} \\ &= \frac{Z^2/1}{X/n} \end{aligned}$$

Since $Z^2 \sim \chi^2_1$, this is the ratio of:

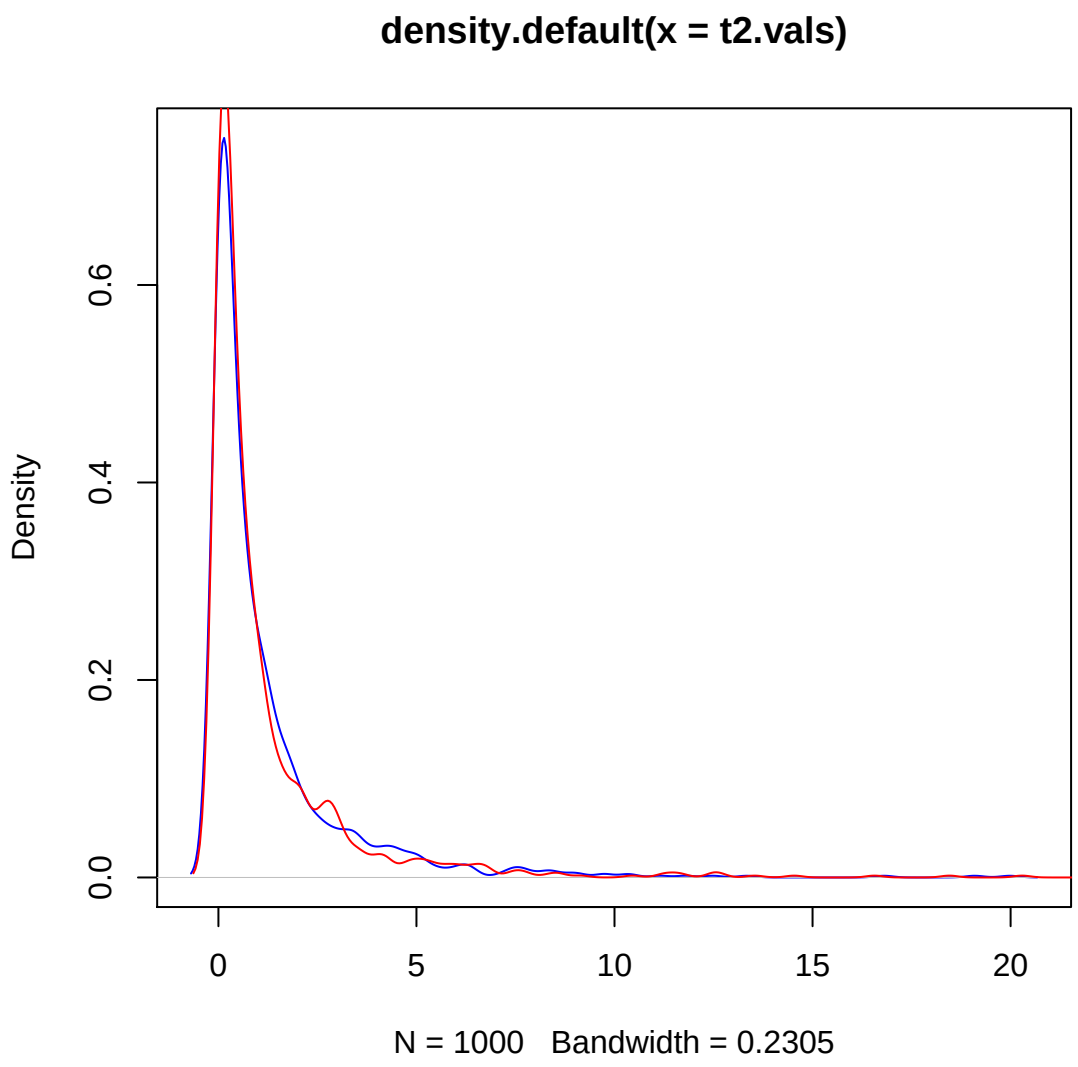
$$\frac{\chi^2_1/1}{\chi^2_n/n}$$

Therefore, by the definition of an F distribution, $T^2 \sim F_{1,n}$.

(b) **For n=10, demonstrate this equivalence numerically by plotting the kernel density estimates for 1000 randomly generated $T^2$ values and 1000 randomly generated $F_{1,n}$ values.**

```
> t.vals = rt(1000, df=10)
> t2.vals = t.vals^2
```

```
> f.vals = rf(1000, df1=1, df2=10)
> plot(density(t2.vals), type="l", col="blue")
> points(density(f.vals), type="l", col="red")
```




3. **(Optional) Rice, Chapter 6, Problem 9**

   (a) **Find the mean of $S^2$, where $S^2$ is as in Section 6.3.**

   According to section 6.3 we are dealing with iid random variables $X_1, ..., X_n$ where $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The sample variance, $S^2$, is defined as follows:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

   With $\bar{X}$, the sample mean, defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

First, the expectation of $\bar{X}$ (which is $\mu$, but let's prove that anyway):

$$\begin{aligned}
E[\bar{X}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i], \textit{ E of sum of RVs is sum of Es} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \frac{n}{n}\mu \\
&= \mu
\end{aligned}$$

Next, find the variance of $\bar{X}$ (which is $\sigma^2/n$, but let's prove that anyway):

$$\begin{aligned}
Var(\bar{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i), \textit{ var of constant is constant squared} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i], \textit{ var of sum of RVs is sum vars} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{n}{n^2}\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

The expectation of $S^2$ can be computed as follows:

$$
\begin{aligned}
E[S^2] &= E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2] \\
&= E[\frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)], \textit{ Casella and Berger, Theorem 5.2.4} \\
&= \frac{1}{n-1}(E[\sum_{i=1}^{n} X_i^2] - E[n\bar{X}^2]) \\
&= \frac{1}{n-1}(\sum_{i=1}^{n} E[X_i^2] - E[n\bar{X}^2]) \\
&= \frac{1}{n-1}(nE[X_i^2] - nE[\bar{X}^2]) \\
&= \frac{1}{n-1}(n(Var(X_i) + E[X_i]^2) - n(Var(\bar{X}) + E[\bar{X}]^2])), \textit{ from def of variance} \\
&= \frac{1}{n-1}(n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)), \textit{ plug in variance and expectations from above} \\
&= \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\
&= \frac{1}{n-1}(\sigma^2(n-1)) \\
&= \sigma^2
\end{aligned}
$$

(b) **Find the variance of $S^2$, where $S^2$ is as in Section 6.3.**
If we take of advantage of the fact that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, then the variance of $S^2$ can be found by defining a new random variable $X = n - 1S^2/\sigma^2$ and then proceeding as follows:

$$
\begin{aligned}
X &\sim \chi^2_{n-1} \textit{ , by distribution above} \\
S^2 &= (\sigma^2/n - 1)(n - 1/\sigma^2)S^2 \textit{ , algebraic trickery} \\
S^2 &= (\sigma^2/n - 1)X \\
Var(S^2) &= Var((\sigma^2/n - 1)X) \\
Var(S^2) &= \frac{\sigma^4}{(n-1)^2}Var(X) \textit{ , move constants out} \\
Var(S^2) &= \frac{\sigma^4}{(n-1)^2}2(n-1) \textit{ , variance of chi-squared} \\
Var(S^2) &= \frac{2\sigma^4}{n-1}
\end{aligned}
$$

4. **Rice, Chapter 7, Problem 3: Which of the following is a random variable?**

The way to approach this questions is to ask: does this value change between random samples? If it does, it is a RV whose distribution is induced by the random sampling scheme; if it doesn't it is a constant and not a RV. As detailed below, only cases d, f and h are RVs.

(a) **The population mean**

No, this is not a RV. For a given population, it has a fixed value that doesn't change between random samples.

(b) **The population size, N**

Again, not a RV.

(c) **The sample size, n**

Not a RV. Although not a parameter of the population, the size n is the same for all random samples of size n. The sampling process does not select n randomly.

(d) **The sample mean**

Yes, this is a RV. Since the members of the sample are random and change between samples, the sample mean will change and is a RV.

(e) **The variance of the sample mean**

No, this is not a RV but a function of the population parameters. Although the realized value of the sample mean changes between samples, they all follow the same theoretical distribution with a fixed variance. Note that we are talking about the theoretical variance of the sample mean and not an estimated variance.

(f) **The largest value in the sample**

Yes, this is an RV. If the values in the sample are random, the largest is also random.

(g) **The population variance**

Not a RV. Reasoning is similar to case a).

(h) **The estimated variance of the sample mean**

Yes, this a RV. Here we are plugging estimates of the population parameters into the theoretical variance formula. The value is therefore a function of the random sample values so is also random.

5. **(Based on Rice, Chapter 7, Problem 4) Two populations are surveyed with simple random sampling. A sample of size $n_1$ is used for population I, which has a population standard deviation of $\sigma_1$; a sample of size $n_2 = 3n_1$ is used for population II, which has a population standard deviation of $\sigma_2 = 2\sigma_1$.**

(a) **Ignoring the finite population correction, in which of the two samples would you expect the estimate of the population mean to be more accurate? Provide a mathematical justification for your answer.**

Per the LLN, the best estimate of the population mean $\mu$ is the sample average, $\bar{X}$. The variance of $\bar{X}$ (i.e, the variance of its sampling distribution) is:

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

The most accurate variance estimate will be the estimate whose sampling distribution has the smallest variance.

Ignoring the finite population correction, the variance of the estimate for population I is:

$$Var(\bar{X}_1) = \frac{\sigma_1^2}{n_1}$$

and the variance of the estimate for population II is:

$$\begin{aligned} Var(\bar{X}_2) &= \frac{(2\sigma_1)^2}{3n_1} \\ &= \frac{4\sigma_1^2}{3n_1} \\ &= 4/3Var(\bar{X}_1) \end{aligned}$$

So the estimate of $\mu$ for population I is more accurate than the estimate of $\mu$ for population II.

(b) **For what ratio of $n_2/n_1$ would the estimates have equivalent accuracy?**

- Here we want to assume that $n_2/n_1 = k$ (i.e., $n_2 = kn_1$) and then solve for k such that $Var(\bar{X}_1) = Var(\bar{X}_2)$.
- Using $n_2 = kn_1$ rather than $n_2 = 2n_1$, we have:

$$\begin{aligned} Var(\bar{X}_2) &= \frac{(2\sigma_1)^2}{kn_1} \\ &= 4/kVar(\bar{X}_1) \end{aligned}$$

- Set them equal and solve for k:

$$\begin{aligned} Var(\bar{X}_2) &= Var(\bar{X}_1) \\ 4/kVar(\bar{X}_1) &= Var(\bar{X}_1) \\ k &= 4 \end{aligned}$$

In other words, if population II has twice the variance as population I, we need a sample that is four times larger to get the same accuracy in the estimate for $\mu$.

(c) **Verify this ratio via simulation, i.e., create populations I and II by simulating 1000 normal RVs for each with $\mu = 1$ and $\sigma_1 = 1$ and generate 1000 estimates of the population mean $\mu$ using random samples with $n_1 = 100$ and $n_2$ set to give the ratio you found in b). Plot the distributions of these estimates using a kernel density estimate (the distributions should look equivalent: same mean and same variance). Why won't these empirical distributions look identical?**

Create the two populations:

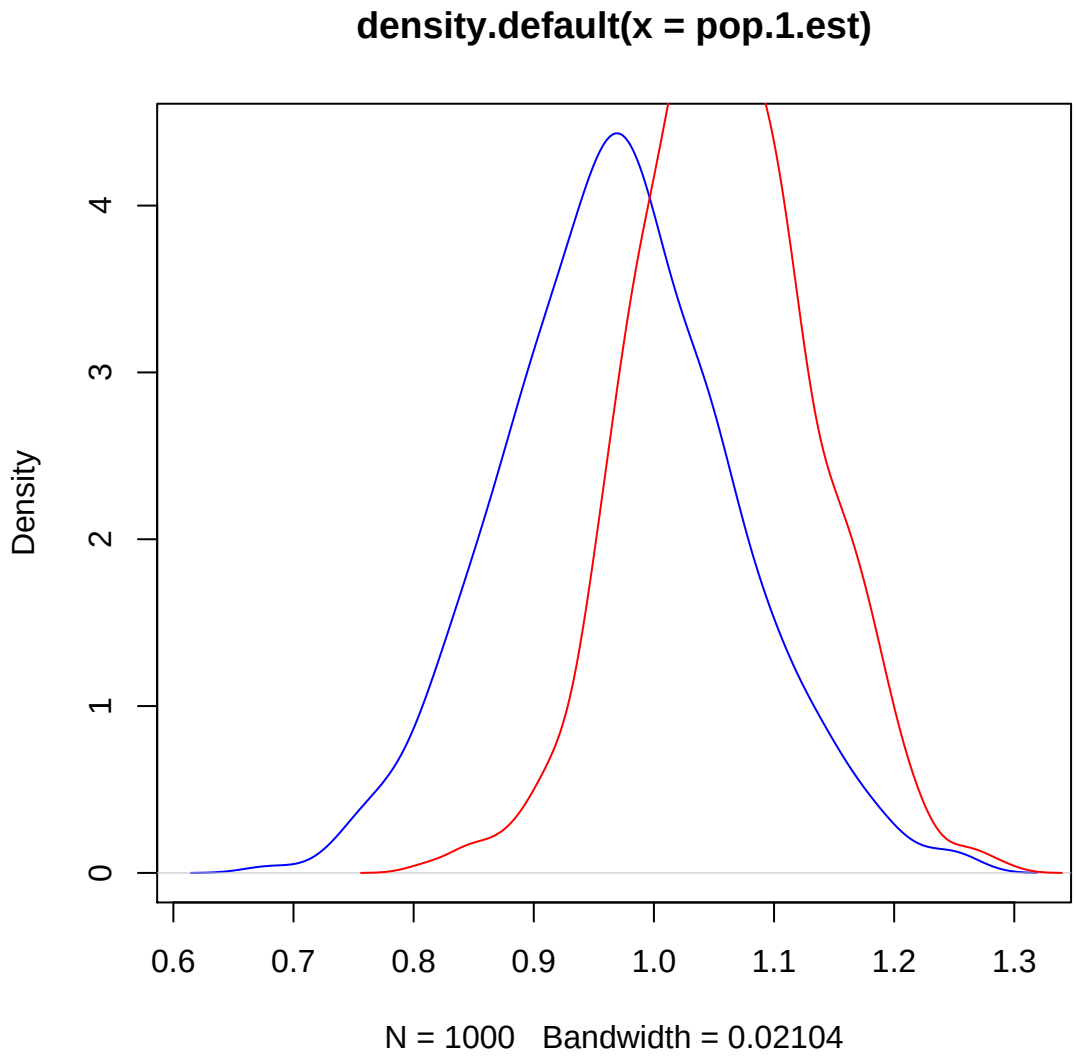
```
> pop.1 = rnorm(1000, mean=1, sd=1)
> mean(pop.1)
[1] 0.9729591
> sd(pop.1)
[1] 1.003823
> pop.2 = rnorm(1000, mean=1, sd=2)
> mean(pop.2)
[1] 1.057508
> sd(pop.2)
[1] 1.869251
```

Generate 1000 estimates of the population mean using random samples of size 100 for population 1 and size 400 for population 2:

```
> num.est = 1000
> pop.1.est = rep(NA, num.est)
> pop.2.est = rep(NA, num.est)
> for (i in 1:num.est) {
+         pop.1.est[i] = mean(sample(pop.1, 100, replace=F))
+         pop.2.est[i] = mean(sample(pop.2, 400, replace=F))
+ }
```

Plot the empirical distribution of the estimates:

```
> plot(density(pop.1.est), type="l", col="blue")
> points(density(pop.2.est), type="l", col="red")
```




These look similar. They won't be exactly equivalent since the population SDs and means are only approximate and we're using just 1000 estimates.

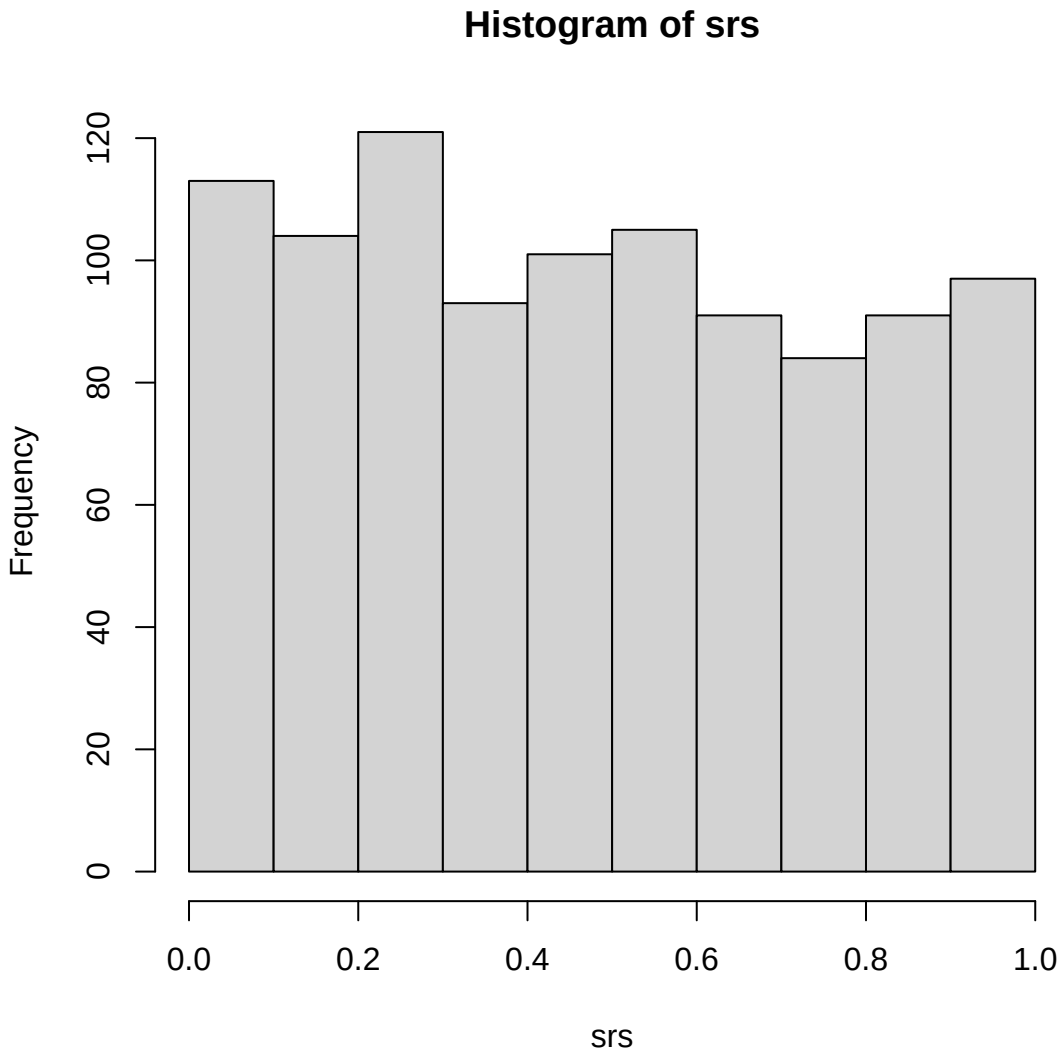
6. **Rice, Chapter 7, Problem 10: True or false (and state why): If a sample from a population is large, a histogram of the values in the sample will be appropriately normal, even if the population is not normal? Verify your answer via simulation using a population whose elements have a U(0,1) distribution.**

   The distribution of the sample values will not be approximately normal. As detailed in the chapter, each sample element $X_i$ has a discrete distribution and, if the population has distinct values, it is discrete uniform (i.e., every value has an equal probability). This is not normal. Can also think of this in terms of the CLT: are the elements of the random sample the average or sum of independent RVs? No, they are not so the CLT does not apply and the distribution will not be approximately normal.

   ***Grading: 3 pts; full points to correctly state that it is not normal and to give a plausible reason. 2 pts if they correctly state that it is not normal but don't provide a reasonable justification.***

   Let's validate using simulation (not normal!):

```
> # create a population with N=10000 by simulating from U(0,1)
> pop = runif(10000)
> # generate a simple random sample of size n=1000
> srs = sample(pop, 1000, replace=F)
> plot(hist(srs))
```




***Grading: 2 pts to generate a valid histogram for a simple random sample.***

7. **Rice, Chapter 7, Problem 16: True or false?**

   (a) **The center of a 95% confidence interval for the population mean is a random variable.**

   True. The CI is computed from the sample and the center and boundaries will therefore change with each sample.

   (b) **A 95% confidence interval for $\mu$ contains the sample mean with probability 0.95.**

   False. The CI for $\mu$ is computed based on the sample and will always contain the sample mean, the estimate for $\mu$.

   (c) **A 95% confidence interval contains 95% of the population.**

False. A x% CI is a random interval computed from each random sample; it does not contain a fixed proportion of the population.

(d) **Out of one hundred 95% confidence intervals for $\mu$, 95 will contain $\mu$.**
True. This is the exact definition of a CI.

8. **(Based on Rice, Chapter 7, Problem 19) This problem introduces the concept of a one-sided CI. Using the CLT, how should the constant k be chosen so that the interval $(\bar{X} - ks_{\bar{X}}, \infty)$ is a 90% CI for $\mu$, i.e., so that $P(\mu \geq \bar{X} - ks_{\bar{X}}) = 0.9$? This is called a one-sided CI.**

(a) **Find k such that:**

$$P(\mu \geq \bar{X} - ks_{\bar{X}}) = 0.9$$

So, we will be employing the approximate normal sampling distribution for $\bar{X}$:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

Substituting $s_{\bar{X}}$ for $\sigma/\sqrt{n}$, this becomes:

$$\bar{X} \sim \mathcal{N}(\mu, s^2_{\bar{X}})$$

Standardizing we get:

$$\frac{\bar{X} - \mu}{s_{\bar{X}}} \sim \mathcal{N}(0, 1)$$

The trick to solving for k is to rearrange terms to get a probability for $\frac{\bar{X}-\mu}{s_{\bar{X}}}$, which we know per the CLT is approximately $\mathcal{N}(0, 1)$. This can then be found as $\Phi()$ of some function of k, making k equal to the inverse CDF of some value, which we can compute using qnorm().

$$\begin{aligned} P(\mu \geq \bar{X} - ks_{\bar{X}}) &= 0.9 \\ P(\bar{X} - \mu) \leq ks_{\bar{X}}) &= 0.9 \\ P((\bar{X} - \mu)/s_{\bar{X}} \leq k) &= 0.9 \\ \Phi(k) &= 0.9 \\ k &= \Phi^{-1}(0.9) \end{aligned}$$

We can now solve for k using qnorm()

```
> qnorm(0.9)
[1] 1.281552
```

(b) **How should k be chosen so that $(-\infty, \bar{X} + ks_{\bar{X}})$ is a 95% one-sided CI?**

This is an analogous problem just starting with a different probability (note: one can use symmetry arguments to get a solution more quickly):

$$\begin{aligned} P(\mu \leq \bar{X} + ks_{\bar{X}}) &= 0.95 \\ P(\bar{X} - \mu) \geq -ks_{\bar{X}}) &= 0.95 \\ P((\bar{X} - \mu)/s_{\bar{X}} \geq -k) &= 0.95 \\ 1 - P((\bar{X} - \mu)/s_{\bar{X}} \leq -k) &= 0.95 \\ 1 - \Phi(-k) &= 0.95 \\ \Phi(-k) &= 0.05 \\ k &= -\Phi^{-1}(0.05) \end{aligned}$$

We can now solve for k using qnorm()

```
> -qnorm(0.05)
[1] 1.644854
```