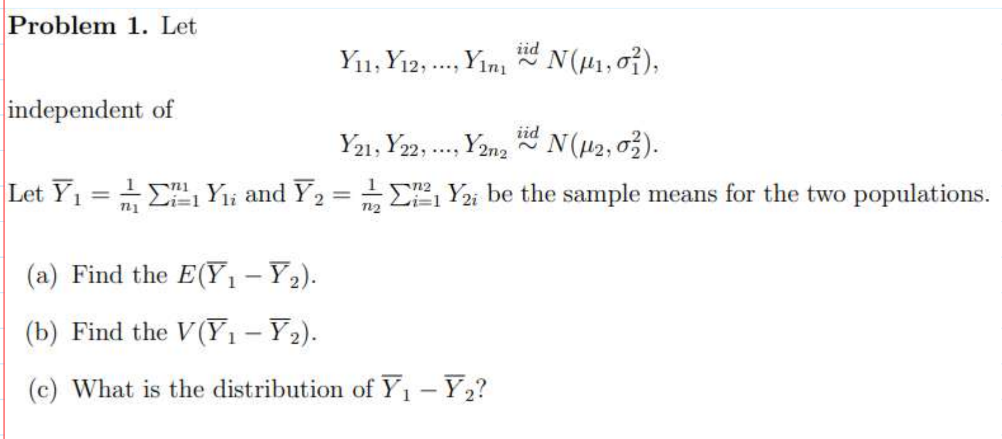
HW1

Wednesday, May 14, 2025 4:06 PM



Problem 1. Let

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2),$$

independent of

$$Y_{21}, Y_{22}, ..., Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2).$$

Let $\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i}$ and $\overline{Y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i}$ be the sample means for the two populations.

(a) Find the $E(\overline{Y}_1 - \overline{Y}_2)$.

(b) Find the $V(\overline{Y}_1 - \overline{Y}_2)$.

(c) What is the distribution of $\overline{Y}_1 - \overline{Y}_2$?

a) $E(\overline{Y}_1-\overline{Y}_2)=E(\overline{Y}_1)-E(\overline{Y}_2)$

$=E(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i})-E(\frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j})$

$=\frac{1}{n_1}\sum_{i=1}^{n_1} E(Y_{1i})-\frac{1}{n_2}\sum_{j=1}^{n_2} E(Y_{2j})$

$=\frac{n_1}{n_1}\mu_1-\frac{n_2}{n_2}\mu_2$

$=\mu_1-\mu_2$

b) $V(\overline{Y}_1-\overline{Y}_2)=E((\mu_1-\mu_2-\overline{Y}_1+\overline{Y}_2)^2)$ Note $Y_\ell \sim N(\mu_\ell, \frac{\sigma_\ell^2}{n_\ell})$ for $\ell \in \{1,2\}$

$=E(((\mu_1-\overline{Y}_1)-(\mu_2-\overline{Y}_2))^2)$

$=E((\mu_1-\overline{Y}_1)^2-2(\mu_1-\overline{Y}_1)(\mu_2-\overline{Y}_2)+(\mu_2-\overline{Y}_2)^2)$

$=V(\overline{Y}_1)+V(\overline{Y}_2)-2E(\mu_1\mu_2-\mu_1\overline{Y}_2-\mu_2\overline{Y}_1+\overline{Y}_1\overline{Y}_2)$ since $\{Y_{1i}\}$ is indep of $\{Y_{2j}\}$

$=V(\overline{Y}_1)+V(\overline{Y}_2)-\cancel{2(\mu_1\mu_2-\mu_1\mu_2-\mu_2\mu_1+\mu_1\mu_2)}$

$=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$

c) Since $\overline{Y}_1\sim N(\mu_1,\frac{\sigma_1^2}{n})$ and $\overline{Y}_2\sim N(\mu_2,\frac{\sigma_2^2}{n})$ are independent

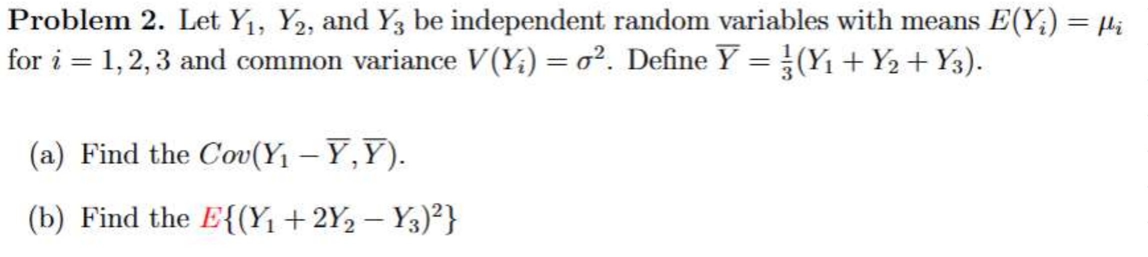
$MGF_{\overline{Y}_1-\overline{Y}_2}(t)=MGF_{\overline{Y}_1}(t)\,MGF_{\overline{Y}_2}(-t)$

$=e^{\mu_1 t+\frac{(\sigma_1 t)^2}{2n_1}}\,e^{-\mu_2 t+\frac{(\sigma_2(-t))^2}{2n_2}}$

$=e^{(\mu_1-\mu_2)t+(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})\frac{t^2}{2}}$

This is the MGF of a normal rv. Thus

$\overline{Y}_1-\overline{Y}_2\sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$



Problem 2. Let Y_1, Y_2, and Y_3 be independent random variables with means $E(Y_i) = \mu_i$ for $i = 1, 2, 3$ and common variance $V(Y_i) = \sigma^2$. Define $\overline{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$.

(a) Find the $Cov(Y_1 - \overline{Y}, \overline{Y})$.

(b) Find the $E\{(Y_1 + 2Y_2 - Y_3)^2\}$

a) $\mathrm{Cov}(Y_1-\bar{Y},\bar{Y}) = E[(Y_1-\bar{Y})\bar{Y}] - E[Y_1-\bar{Y}]E[\bar{Y}]$

$= E(Y_1\bar{Y}) - E(\bar{Y}^2) - E(Y_1)E(\bar{Y}) + E(\bar{Y})^2$

$= \mathrm{Cov}(Y_1,\bar{Y}) - V(\bar{Y})$

$\mathrm{Cov}(Y_1,\bar{Y}) = E[\frac{Y_1}{3}(Y_1+Y_2+Y_3)] - \frac{\mu_1}{3}(\mu_1+\mu_2+\mu_3)$

$= \frac{1}{3}(E[Y_1^2]+E[Y_1Y_2]+E[Y_1Y_3]) - \frac{\mu_1}{3}(\mu_1+\mu_2+\mu_3)$ From indep $E[XY]=E[X]E[Y]$

$= \frac{1}{3}E[Y_1^2] - \frac{1}{3}E[Y_1]^2$

$= \frac{1}{3}V[Y_1] = \frac{\sigma^2}{3}$

$V[\bar{Y}] = \frac{1}{9}(V(Y_1)+V(Y_2)+V(Y_3)) = \frac{3\sigma^2}{9} = \frac{\sigma^2}{3}$

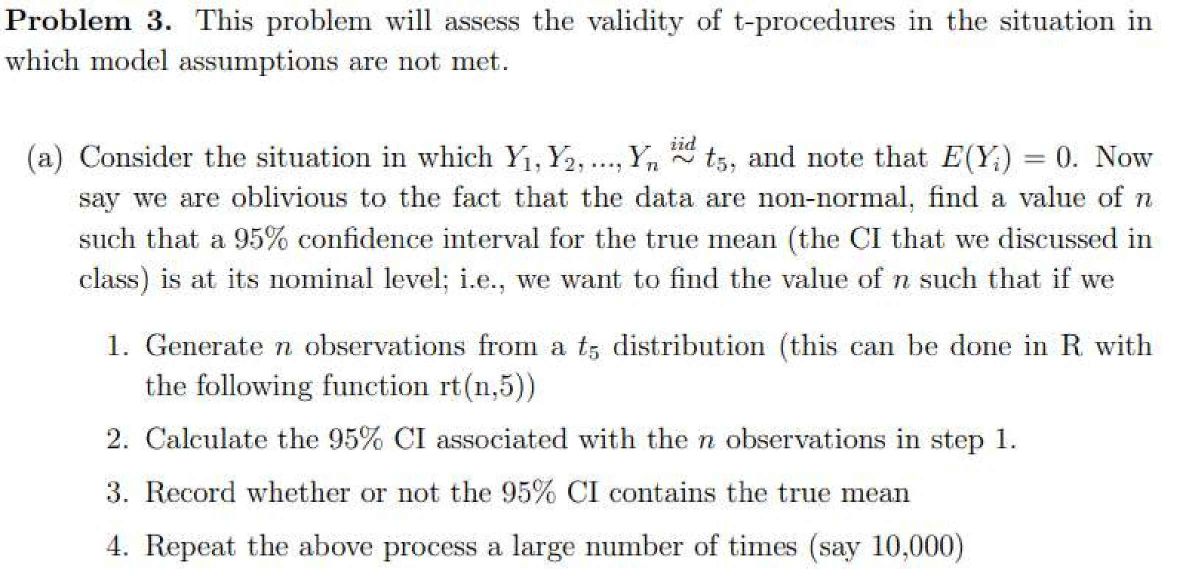
Thus $\mathrm{Cov}(Y_1-\bar{Y},\bar{Y}) = 0$

b) $E[(Y_1+2Y_2-Y_3)^2] = E[Y_1^2+4Y_2^2+Y_3^2+4Y_1Y_2-2Y_1Y_3-4Y_2Y_3]$ From indep $E[XY]=E[X]E[Y]$

$= E[Y_1^2]+4E[Y_2^2]+E[Y_3^2]+4E[Y_1Y_2]-2E[Y_1Y_3]-4E[Y_2Y_3]$

$= \sigma^2+\mu_1^2+4\sigma^2+4\mu_2^2+\sigma^2+\mu_3^2+4\mu_1\mu_2-2\mu_1\mu_3-4\mu_2\mu_3$

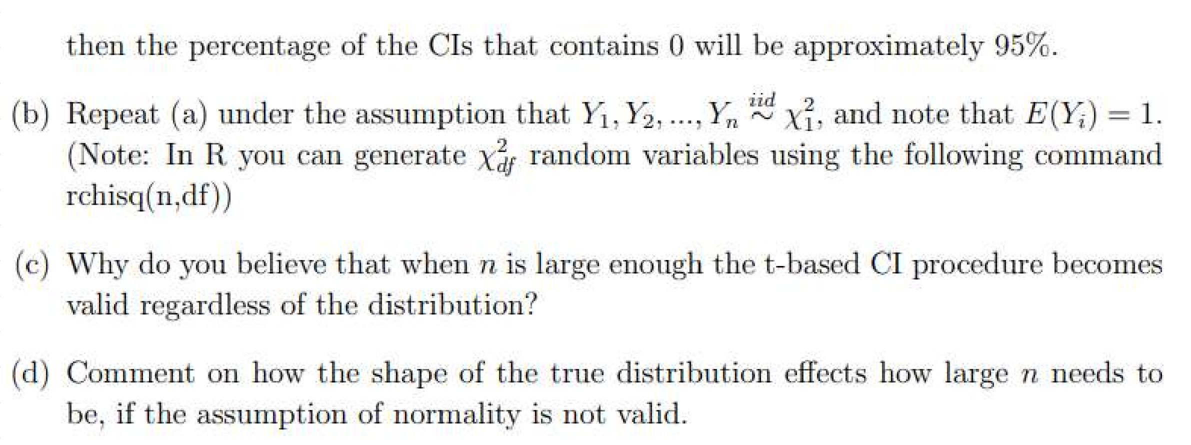
$= 6\sigma^2+\mu_1^2+4\mu_2^2+\mu_3^2+4\mu_1\mu_2-2\mu_1\mu_3-4\mu_2\mu_3$



Problem 3. This problem will assess the validity of t-procedures in the situation in which model assumptions are not met.

(a) Consider the situation in which $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} t_5$, and note that $E(Y_i) = 0$. Now say we are oblivious to the fact that the data are non-normal, find a value of n such that a 95% confidence interval for the true mean (the CI that we discussed in class) is at its nominal level; i.e., we want to find the value of n such that if we

1. Generate n observations from a t_5 distribution (this can be done in R with the following function rt(n,5))
2. Calculate the 95% CI associated with the n observations in step 1.
3. Record whether or not the 95% CI contains the true mean
4. Repeat the above process a large number of times (say 10,000)

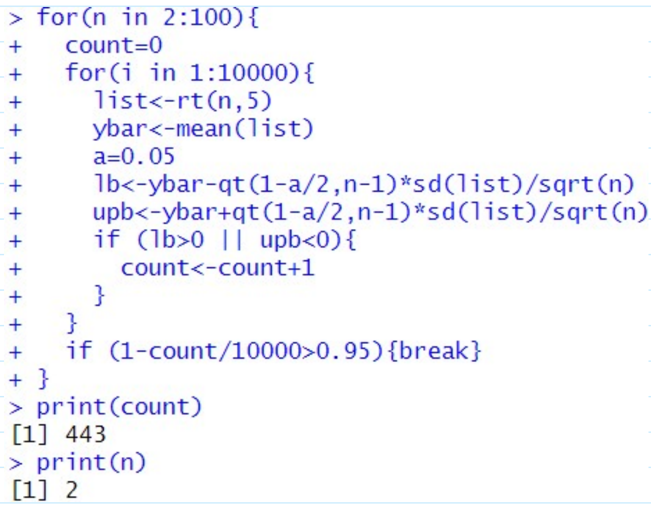

then the percentage of the CIs that contains 0 will be approximately 95%.

(b) Repeat (a) under the assumption that $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} \chi^2_1$, and note that $E(Y_i) = 1$. (Note: In R you can generate χ^2_{df} random variables using the following command rchisq(n,df))

(c) Why do you believe that when n is large enough the t-based CI procedure becomes valid regardless of the distribution?

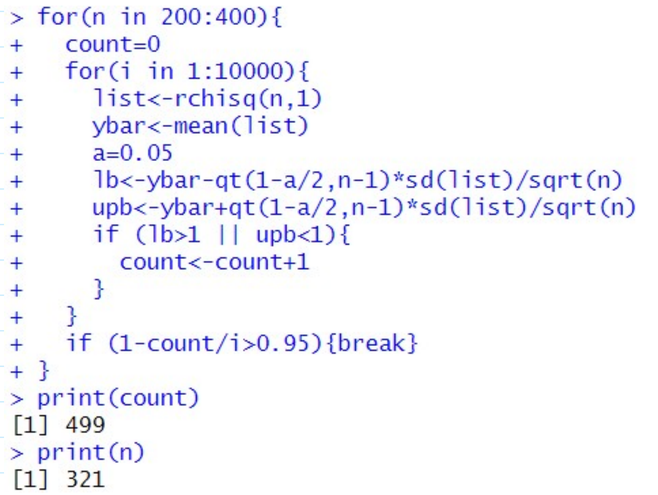
(d) Comment on how the shape of the true distribution effects how large n needs to be, if the assumption of normality is not valid.

a)



```
> for(n in 2:100){
+   count=0
+   for(i in 1:10000){
+     list<-rt(n,5)
+     ybar<-mean(list)
+     a=0.05
+     lb<-ybar-qt(1-a/2,n-1)*sd(list)/sqrt(n)
+     upb<-ybar+qt(1-a/2,n-1)*sd(list)/sqrt(n)
+     if (lb>0 || upb<0){
+       count<-count+1
+     }
+   }
+   if (1-count/10000>0.95){break}
+ }
> print(count)
[1] 443
> print(n)
[1] 2
```

b)

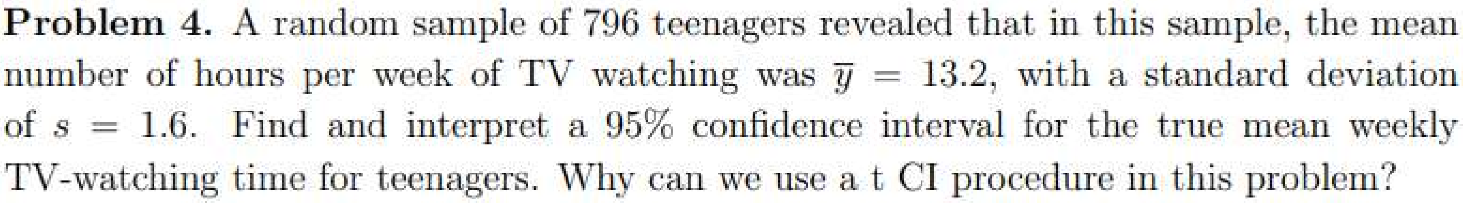


```
> for(n in 200:400){
+   count=0
+   for(i in 1:10000){
+     list<-rchisq(n,1)
+     ybar<-mean(list)
+     a=0.05
+     lb<-ybar-qt(1-a/2,n-1)*sd(list)/sqrt(n)
+     upb<-ybar+qt(1-a/2,n-1)*sd(list)/sqrt(n)
+     if (lb>1 || upb<1){
+       count<-count+1
+     }
+   }
+   if (1-count/i>0.95){break}
+ }
> print(count)
[1] 499
> print(n)
[1] 321
```

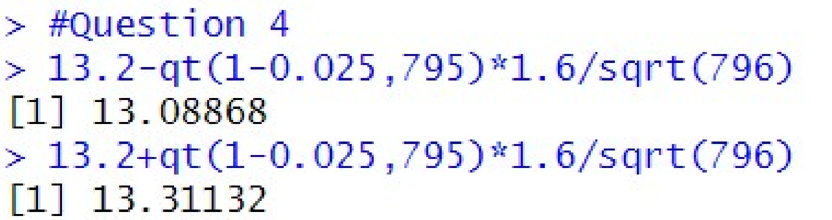
note $n \in [2, 200)$ would not terminate so I increased the lower bound for the outer loop to start at 200

c) Because the t distribution has heavy tails so eventually as n increases (so the degrees of freedom also increase) more and more will end up in the CI.

d) If the graph of the pdf of the distribution is skewed within the domain, or has a multimodal distribution, the degrees of freedom needs to be much larger. There are other more complicated reasons that would make n increase as well that the degrees of freedom needs to be much larger. There are other more complicated reasons that would make n increase as well that we have not covered in class.



Problem 4. A random sample of 796 teenagers revealed that in this sample, the mean number of hours per week of TV watching was $\overline{y} = 13.2$, with a standard deviation of $s = 1.6$. Find and interpret a 95% confidence interval for the true mean weekly TV-watching time for teenagers. Why can we use a t CI procedure in this problem?



```
> #Question 4
> 13.2-qt(1-0.025,795)*1.6/sqrt(796)
[1] 13.08868
> 13.2+qt(1-0.025,795)*1.6/sqrt(796)
[1] 13.31132
```

Thus the CI is [13.08868,13.31132]

We can use the t CI procedure because 796 is a good sample size since we have no reason to believe that the data has any issues that would require a larger sample size. We can apply the CLT which tells us in this situation

$$\frac{\bar{y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$