What Confidence.. ???*

STAT 350

Project 2

As we will see in class that a $(1-\alpha)\times 100\%$ *Confidence Interval* for the unknown mean μ of a normal population (when the population standard deviation σ, is known), is given by:

$$\mu = \bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}. \tag{1}$$

To have a better understanding of the term "95% *Confidence* ", say, imagine that you can draw indefinitely many different samples of n observations each (size n), all from the same population. Then, for each such sample, you compute the sample mean $\bar{X}$, and the corresponding confidence interval, using each time the above formula. Thus now you would have *many* different confidence intervals (each computed from a different sample), of which; *about* 95% of them contain the *true* value of the population mean μ, and the other 5% fail to contain it.

In this project you will use R to simulate <u>many</u> such *95% Confidence Intervals.*

- Generate $N = 10000$ different random samples, each of size $n = 10$ observations, from *an assumed Normal population with* $\mu = 72$ *and* $\sigma = 12$.

```
N<-10000; n<-10; mu<-72; sigma=12
MySamples<-matrix(rep(0, N*n), n, N)
MySamples<-replicate(N, rnorm(n, mu, sigma))
```

- Calculate the sample average, $\bar{X}_n$ and the sample standard deviation, S for each of these $N = 10^4$ different samples.

```
Xbar<-apply(MySamples, 2, mean)
SDx<-apply(MySamples, 2, sd)
```

- Calculated values of the Lower Confidence Limit (LCL) and the Upper Confidence Limit (UCL), of the $(1-\alpha)\times 100\% = 0.95$ Confidence Interval (CI), respectively,

```
alpha<-0.05
q_alpha<-qnorm(1-alpha/2)
ME<-q_alpha*sigma/sqrt(n)  # The Margin or Error
LCL<-Xbar-ME
UCL<-Xbar+ME
Intervals<-data.frame(Xbar, ME, LCL, UCL,  Length=UCL-LCL)
head(Intervals)
```

*Spring 2021

```
##       Xbar      ME      LCL      UCL   Length
## 1 70.07804 7.43754 62.64050 77.51558 14.87508
## 2 75.90002 7.43754 68.46248 83.33756 14.87508
## 3 78.95559 7.43754 71.51805 86.39313 14.87508
## 4 70.80458 7.43754 63.36704 78.24212 14.87508
## 5 74.45777 7.43754 67.02023 81.89531 14.87508
## 6 70.95772 7.43754 63.52018 78.39526 14.87508
```

- Find out how many of the $N = 10^4$ confidence intervals do indeed contain the true value of μ. To do so, observe that a confidence interval does contain μ if and only

$$|\bar{X}_n - \mu| \leq ME \equiv Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

```
Coverage<-(abs(Xbar-mu)<=ME)
Coverage[1:10]
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

```
Confidence<-sum(Coverage)/N
Confidence
```

```
## [1] 0.951
```

- The formulation in (1) will change if the population standard deviation σ, is **unknown,** in which case, with the same assumptions as above, the $(1-\alpha)\times 100\%$ *Confidence Interval* for the unknown mean μ of a normal population is given

$$\mu = \bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}. \tag{2}$$

```
q_alpha<-qt(1-alpha/2, n-1)
ME<-q_alpha*SDx/sqrt(n)  # The Margin of Error in this Case
LCL<-Xbar-ME
UCL<-Xbar+ME
Intervals<-data.frame(Xbar, ME, LCL, UCL, Length=UCL-LCL)
head(Intervals)
```

```
##       Xbar        ME      LCL      UCL   Length
## 1 70.07804 10.256568 59.82148 80.33461 20.51314
## 2 75.90002  5.994025 69.90600 81.89405 11.98805
## 3 78.95559 11.904372 67.05122 90.85996 23.80874
## 4 70.80458  8.668159 62.13643 79.47274 17.33632
## 5 74.45777  8.930634 65.52714 83.38840 17.86127
## 6 70.95772 12.192195 58.76552 83.14991 24.38439
```

- Finding the Coverage Probability of these many intervals

```
Coverage<-(abs(Xbar-mu)<=ME)
Coverage[1:10]
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

```
Confidence<-sum(Coverage)/N
Confidence
```

```
## [1] 0.9494
```

Mispecification

Clearly, these simulations (and generation of the data) were conducted under the **assumed** $\mathcal{N}orm(\mu, \sigma^2)$ distribution. However, what would be the actual coverage probabilities, if this assumed distributional model was miss-specified, and the actual distribution generating the data was, say, the $\mathcal{G}amma(a = \mu/2, b = 2)$? How would this model miss-specification affects the actual coverage probabilities if you were to use either (1) (so-called z-Interval) or (2) (so-called t-interval) to construct a 95% Confidence Interval for μ. Recall that for if $X \sim \mathcal{G}amma(a, b)$ then

$$E(X) = ab \quad \text{and} \qquad \sigma^2 = Var(X) = ab^2$$

In particular with $a = \mu/2 = 36$ and $b = 2$ we have $E(X) = 72$ and $\sigma \equiv \sqrt{Var(X)} = 12$ too.

Your Task:

- Graph and nicely labeled (on a single figure) the density curves of the $\mathcal{N}orm(\mu, \sigma^2)$ and the $\mathcal{G}amma(a = \mu/2, b = 2)$ with $\mu = 70$. Do they appear to be similar?
- Use the previous parts to calculate the actual coverage probabilities when the data is generated from the $\mathcal{G}amma(a = \mu/2, b = 2)$ distribution, and study how the coverage probabilities are impacted by changing the sample size n as you complete the table below. Also include a copy of your code (just for one choice of n) .
- How would you explain the impact of the increased sample size on the coverage probability and the "observed Confidence Level" as compared to the "Desired Confidence Level"

Name:_____

Table 1: Observed Coverage Probabilities of a 95% CI for μ when the data is generated from the Gamma distribution $\mathcal{G}amma(a = \mu/2, b = 2)$ with $\mu = 70$.

Sample Size n	Coverage of Z-Interval	Coverage of t-Interval
10		
25		
50		
75		
100		
200		
400		
1000		