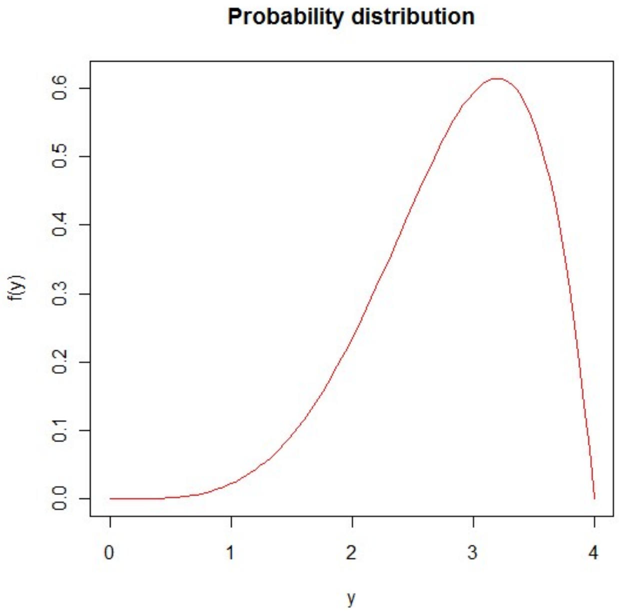
Simulation investigation of the confidence level for a confidence interval Code is available in CIconfidenceLevel.R

In class, we have learned that $T = (\bar{Y} - \mu)/(S/\sqrt{n})$ has a "t distribution" *provided* that Y_1, Y_2, …, Y_n are independent random variables with the same normal probability distribution. If the data Y_1, Y_2, …, Y_n are not normally distributed, for large samples ($n \geq 30$) the central limit theorem will "take over" for $\bar{Y}$ and the t distribution approaches a standard normal distribution. Thus, the normal distribution assumption does not matter for large enough samples. However, if the sample size is relatively small ($n < 30$) and the normal distribution assumption is violated, will the results based on the t distribution still be valid? Specifically, will the confidence interval have the claimed confidence level? We will conduct investigations via simulation in today's lab to find the answer to this question.

First, we will re-examine the GPA example in your class notes. The population is characterized by the following probability distribution

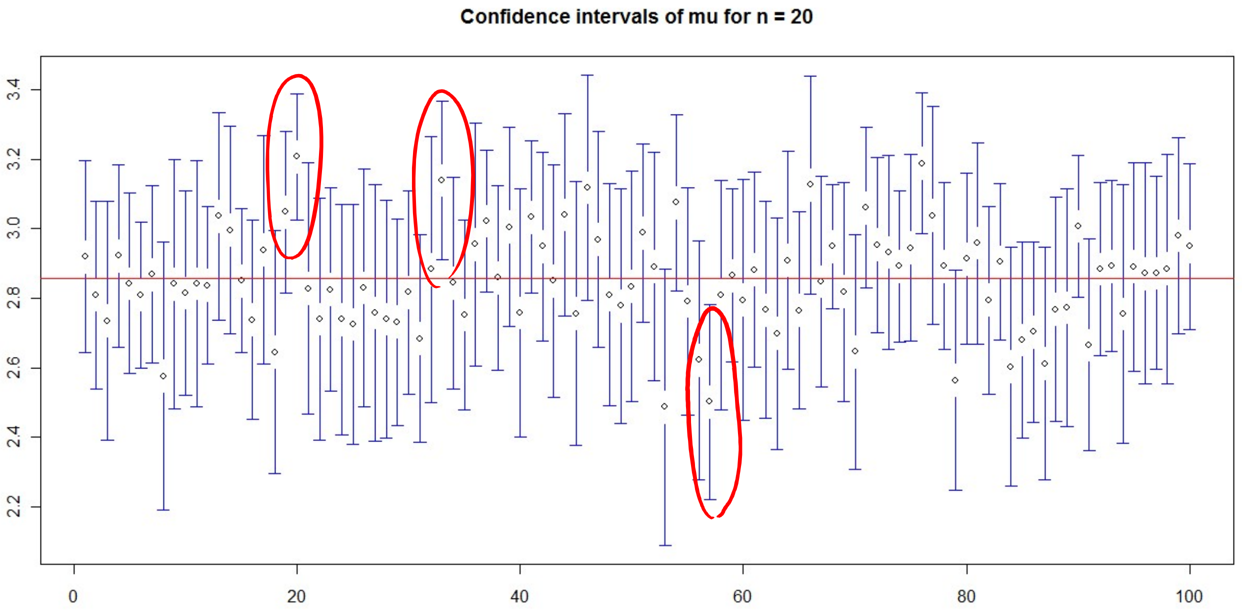


The population mean and variance are μ = 2.8571 and σ^2 = 0.4082. (This distribution is based on a beta probability distribution with parameters a = 5 and b = 2 – see the textbook "Probability and Statistics for Engineers and Scientists" for more information. You are NOT required to know this.) Remember that a 95% confidence interval means that if the sampling process was repeated 1,000 times and confidence intervals were formed *each* time, then we expect about 1,000×0.95 = 950 of the confidence intervals would contain μ. We will go through this process here with n = 20 and count the number of confidence intervals containing μ out of 1000 intervals.

```
> mu <- 2.8571
> #estimate confidence level
> n <- 20
> set.seed(1234)
> set1 <- 4*matrix(data=rbeta(n = 1000*n, shape1 = 5, shape2 = 2), nrow = 1000, ncol = n)
> means <- apply(X = set1, MARGIN = 1, FUN = mean)
> SDs <- apply(X = set1, MARGIN = 1, FUN = sd)
> lower <- means - qt(p=.975, df = 20 - 1) * SDs/sqrt(n)
> upper <- means + qt(p=.975, df = 20 - 1) * SDs/sqrt(n)
> count <- sum(lower < mu & upper > mu)
> count
[1] 946
> count/1000
[1] 0.946
```

We have already seen the use of `matrix` function in our previous lab. Now each row of `set1` contains a sample of size n = 20. Next, we use the `apply` function to find the sample mean and standard deviation for each sample. After we calculated the lower and upper limit of confidence interval for each sample, we use `lower < mu & upper > mu` (`&` in R indicates the logical AND) to find all intervals that contain the population mean. To help visualize what we are trying to do, I saved the first 100 confidence intervals in an object `CI` along with the sample means, and then used the `plotCI()` function in `gplots` package (you need to install this package first) to plot the confidence intervals:

```
> library(gplots)
> hwidth <- qt(p = 0.975, df = 20 - 1) * SDs/sqrt(n)
> plotCI(x = means, uiw = hwidth, col = "black", barcol = "blue", lwd = 0.1,
    cex = 0.8, ylab = "", xlab = "", main = "Confidence intervals of mu for n = 20")
> abline(h = mu, col = "red")
```



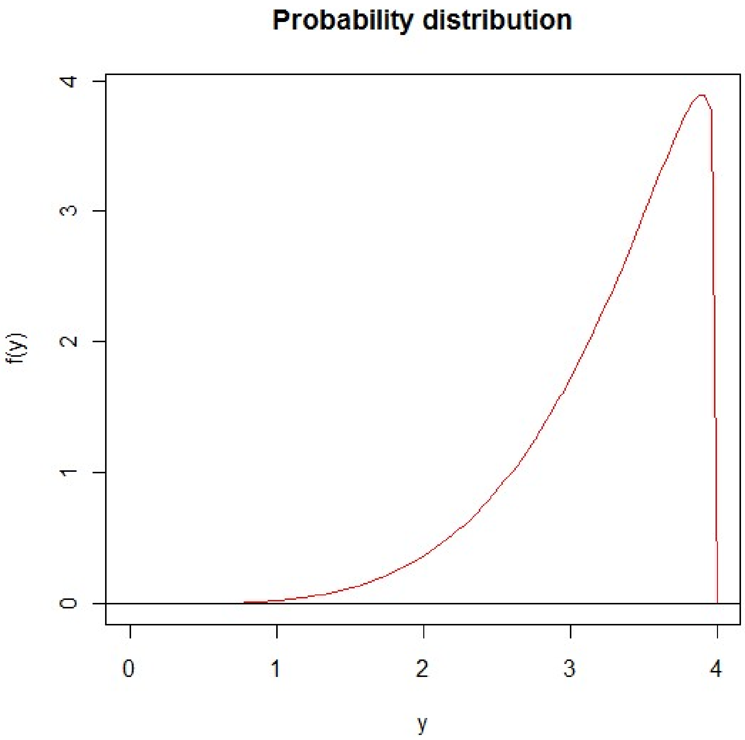
The `hwidth` is the half width of each confidence interval. The red line is the true mean of the distribution for Y. Each blue bar represents a confidence interval with the center dot denoting the sample mean of a sample. Our goal is then to count the number of blue bars that intersect with the red line. You can see in the above plot that a few intervals are completely above or below the true mean.

In the end, out of 1000 confidence intervals, 946 intervals contain $\mu = 2.8571$. Thus, the *estimated* confidence level is 0.946, very close to the claimed confidence level. It seems for this probability distribution of Y, n = 20 is large enough for the t distribution to work well.

Let us try a similar investigation for n = 10. Spend the next five minutes doing it on your own.

What sample size less than 10 results in the interval having a confidence level much less than 0.95?

Next, we would like to use a much more skewed distribution and investigate what would happen to the confidence level when the normal distribution assumption is violated for small samples. The probability distribution for the population is given as



The distribution (based on a beta probability distribution with parameters a = 5 and b = 1.1) is heavily skewed to the left and has population mean and variance: $\mu = 3.2787$ and $\sigma^2 = 0.3331$. Draw a sample of size n = 20 each time and construct a 95% confidence interval for the mean. What is the estimated confidence level for this interval?

Do you think the sample size will be smaller, the same, or larger than 20 for when the confidence level of the interval is less than that of the confidence level of the interval based on n = 20? Report the estimated confidence level and give a reason for your answer.