W203_Supplemental_3

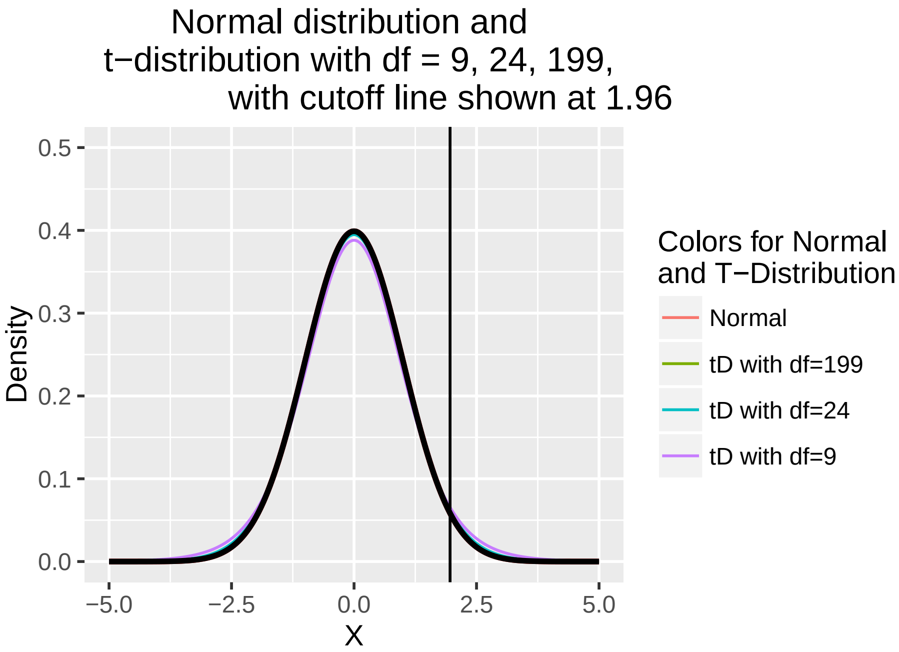
Natarajan Shankar

July 13, 2016

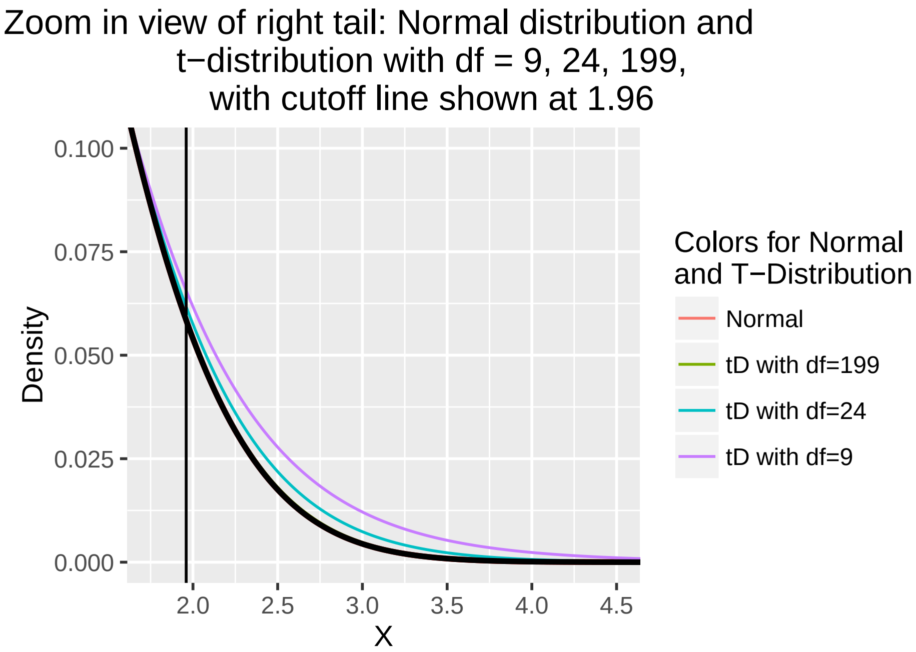
Question 1: Probability Distribution Plots

```
# Pull in all needed libraries
library("ggplot2")

# Create the base data for a Normal distribution
xNormal   <- seq(-5, 5,length=1000)
yNormal   <- dnorm(xNormal,mean=0, sd=1)

# Create  a dataframe for the Normal distribution with the created data to facilitate ggplot
dfNormal <- data.frame(xNormal,yNormal)

# Create t-distributions for various levels of degrees of freedom
yDF10 <- dt(xNormal, 9) # Sample size is 10, df is 10 - 1 = 9
yDF25 <- dt(xNormal, 24) # Sample size is 25, df is 25 - 1 = 24
yDF199 <- dt(xNormal, 199) # Sample size is 200, df is 200 - 1 = 199

# Create a data frame with all 3 distributions together
xAxis=c(xNormal, xNormal, xNormal, xNormal)
yAxis <- c(yNormal, yDF10, yDF25, yDF199)
tags <- rep(c("Normal", "tD with df=9", "tD with df=24", "tD with df=199"), each = 1000)
dfT <- data.frame(xAxis, yAxis, tags, each=1000)

# Plot just the base Normal Curve
compositePlot <- ggplot(dfNormal, aes(xNormal, yNormal))
compositePlot <- compositePlot + geom_line(size=1.0, col="Red")

# Now plot the 3 t Distributions
compositePlot <- compositePlot + geom_line(data = dfT, aes(xAxis, yAxis, col=factor(tags))) +
  geom_line(size=1.0) +
  # Set the title
  ggtitle("Normal distribution and \n t-distribution with df = 9, 24, 199,
                    with cutoff line shown at 1.96") +
  labs(x="X", y="Density") + # Set the axes
  # Themes do not seem to work in rmarkdown, comment them out
  #theme(plot.title = element_text(family = "Trebuchet MS", color="#666666",
  #face="bold", size=24, hjust=0)) +
  #theme(axis.title = element_text(family = "Trebuchet MS", color="#666666", face="bold",
  #size=22)) +  # Set theme for text
  #theme(legend.title=element_text(colour="Black", size=16, face="bold")) +  # Set the legend
  scale_colour_discrete(name = "Colors for Normal \nand T-Distribution") + # Set the legend
  #guides(col = guide_legend(reverse = TRUE)) +
  scale_y_continuous(limits = c(0.0, 0.5)) +
  scale_x_continuous(limits = c(-5, 5)) + geom_vline(xintercept = 1.96)
```

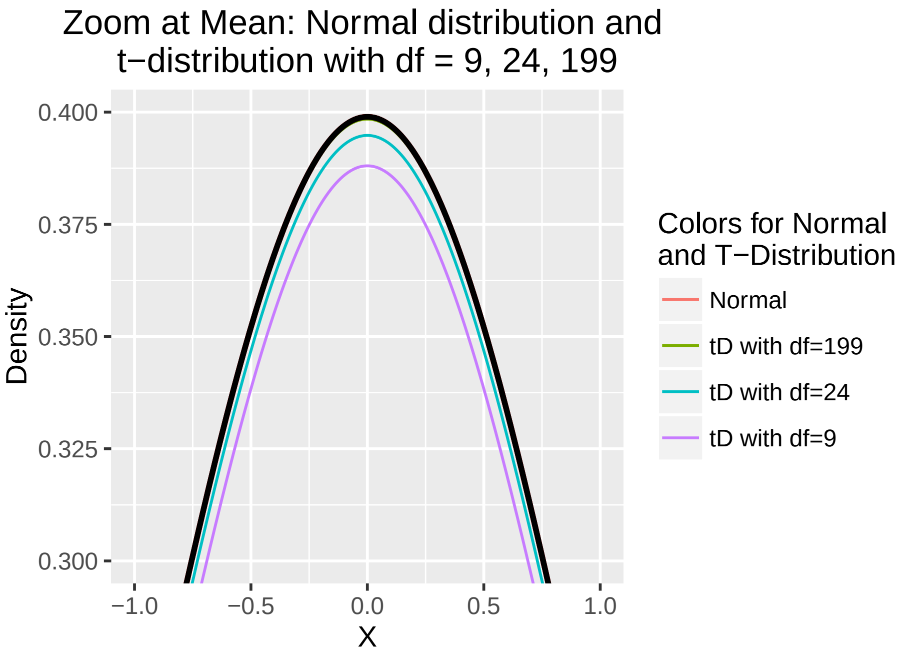
```
# Plot the 4 plots (1 Normal, 3 t-D)
compositePlot
```



```
# Zoom in on the tail to get a better view
compositePlot + coord_cartesian(xlim = c(1.75,4.5),ylim= c(0,0.1)) +
  ggtitle("Zoom in view of right tail: Normal distribution and
          t-distribution with df = 9, 24, 199,
          with cutoff line shown at 1.96")
```



```
# Zoom in on the peak (mean) to get a better view
compositePlot + coord_cartesian(xlim = c(-1.0,1.0),ylim= c(0.3,0.4)) +
  ggtitle("Zoom at Mean: Normal distribution and \nt-distribution with df = 9, 24, 199")
```



Because the t-Distribution for sample sizes less than 30 are fatter and longer tailed, for any given cutoff value they have more area underneath as compared to the Normal distribution.

Question 2 : Naive use of cutoff line at z=1.96, which corresponds to 0.025% confidence interval for 2 tailed Normal Distribution

```
# For the t Distribution, function to compute alpha levels given a cutoff value
twoTailAlpha.t <- function(cutoffValue = 1.96, sampleSize = 200) {
  # Compute and output the alpha value for the given Z value
  tDAlpha <- (1 - pt(cutoffValue, sampleSize - 1))
  return(list(tDAlpha, cutoffValue))
}
# Given a new alpha value, estimate the error
alphaResult <- function (newAlphaData) {
  # For the specific case of Z = 1.96, point out the magnification in Type 1 error area
  cat("Computed alpha value is", newAlphaData[[1]])
  if (abs(newAlphaData[[2]]) == 1.96) {
      if (abs(newAlphaData[[1]] > 0.025)) {
      cat("\nthe Type I error zone for t-Distribution is larger because of the
              incorrect use of cutoff value of 1.96")
      cat("\nEstimated error =", 100* (abs(newAlphaData[[1]] - 0.025))/0.025, "% \n")
      }
  }
}
```

Test case #1 : Z value = 1.96, Sample size = 10

```
newAlphaData <- twoTailAlpha.t(1.96, 10)
alphaResult(newAlphaData)
```

```
## Computed alpha value is 0.0408222
## the Type I error zone for t-Distribution is larger because of the
##               incorrect use of cutoff value of 1.96
## Estimated error = 63.28881 %
```

Test case #2 : Z value = 1.96, Sample size = 25

```
newAlphaData <- twoTailAlpha.t(1.96, 25)
alphaResult(newAlphaData)
```

```
## Computed alpha value is 0.03085301
## the Type I error zone for t-Distribution is larger because of the
##               incorrect use of cutoff value of 1.96
## Estimated error = 23.41205 %
```

Test case #3 : Z value = 1.96, Sample size = 200

```r
newAlphaData <- twoTailAlpha.t(1.96, 200)
alphaResult(newAlphaData)
```

```
## Computed alpha value is 0.02569592
## the Type I error zone for t-Distribution is larger because of the
##                incorrect use of cutoff value of 1.96
## Estimated error = 2.783668 %
```

What does all this mean?

With the Normal distribution, a cutoff value of 1.95 provides a confidence interval of 0.025% in a two tailed test

For the t-distribution, there is dependence of cutoff point upon sample size

1. For a sample size of 10, the cutoff value for an alpha value of 0.025 (left tail, with corresponding +ve value for the right tail) needs to be:

```
-2.2281389
```

2. For a sample size of 25, the cutoff value for an alpha value of 0.025 (left tail, with corresponding +ve value for the right tail) needs to be:

```
-2.0595386
```

3. For a sample size of 200, the cutoff value for an alpha value of 0.025 (left tail, with corresponding +ve value for the right tail) needs to be:

```
-1.9718962
```