Skew Normal Review

Most intro Stat courses skip analysis of skewed distributions. Unfortunately, MOST distributions in transaction environments are skewed and force fitting normal distributions on top make for severely flawed analyses.

The skew normal distribution is implemented in R with the sn package (documentation here: https://cran.r-project.org/web/packages/sn/sn.pdf) and a review of the distribution and density functions are described here: http://azzalini.stat.unipd.it/SN/ . The following is an introductory exercise:

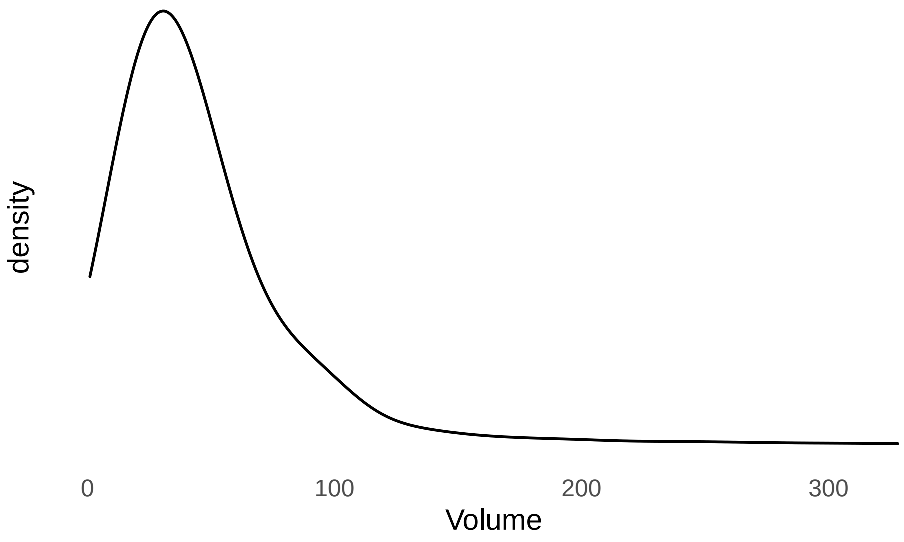
Quick Review of Skew Normal Distribution

Load the following data *(from your last class)*

```
SalesTrans =
  read_csv("C:/Users/ellen/Documents/UH/Fall 2020/Class Materials/Section II/Class 1/Data/Sales.csv")
Location =
  read_csv("C:/Users/ellen/Documents/UH/Fall 2020/Class Materials/Section II/Class 1/Data/Location.csv"
MerGroup =
  read_csv("C:/Users/ellen/Documents/UH/Fall 2020/Class Materials/Section II/Class 1/Data/MerGroup.csv"
SalesTrans = SalesTrans %>% inner_join(Location, by = "LocationID")
SalesTrans = SalesTrans %>% inner_join(MerGroup, by = "MerGroup")
LocationID = as.factor(SalesTrans$LocationID)
SalesTrans$ProductID = as.factor(SalesTrans$ProductID)
SalesTrans$Description = as.factor(SalesTrans$Description)
SalesTrans$MerGroup = as.factor(SalesTrans$MerGroup)

SalesTransSummary = SalesTrans %>%
  group_by(Description, MerGroup, MfgPromo, Wk ) %>%
  summarise(Volume = n(), TotSales = sum(Amount) )
```

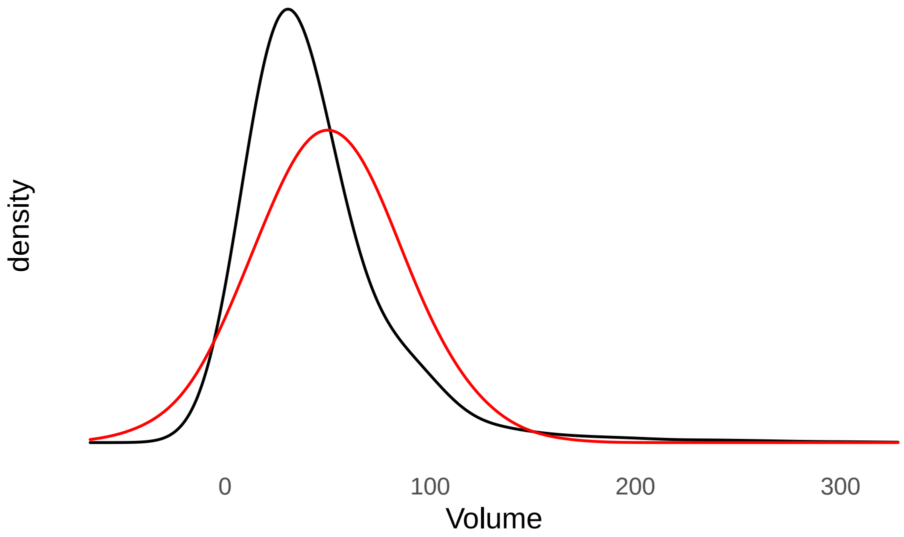
Now, using geom_density *(which uses a non-parametric kernel estimate of the density function)*, plot the estimated density:

```
p = ggplot(SalesTransSummary, aes(Volume)) + geom_density(bw = 15) +
  theme(axis.text.y=element_blank(),axis.ticks=element_blank(),
        panel.background = element_rect(fill = "white"))

p
```



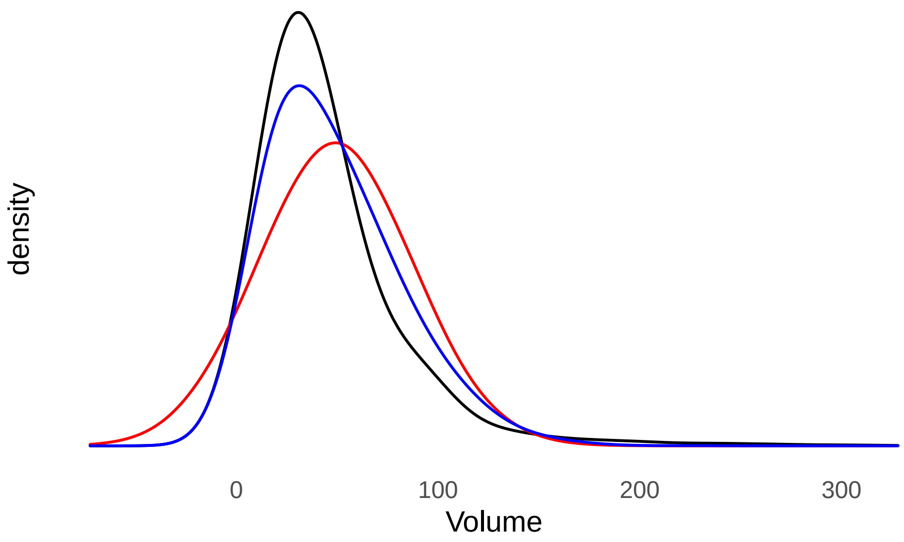
Now, let's compare the the kernel estimate with a normal distribution simulated from the estimated parameters *(mean and standard deviation)*:

```
p = p + geom_density(aes(x  = rnorm(mean = 50, sd = 35, n = nrow(SalesTransSummary))), bw = 15, color =
p
```



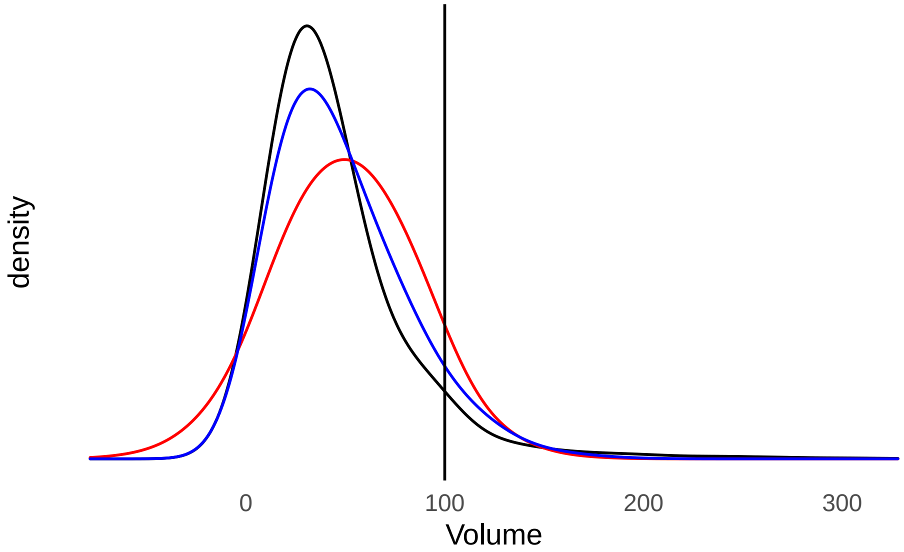
Now, let's add an skew normal distribution simulated from estimated parameters *(location - xi, scale - omega, and skew - alpha)*. Use sn to estimate as follows:

```
estMod <- sn.mple(y = SalesTransSummary$Volume, opt.method = "nlminb")$cp
estParam <- cp2dp(estMod, family = "SN")

exi <- estParam[1]
eomega <- estParam[2]
ealpha <- estParam[3]

p = p + geom_density(aes(x  = rsn(xi = exi, omega = eomega, alpha = ealpha, n = nrow(SalesTransSummary)
p
```



Now that we have these, let's estimate the probability of a locatiion / group volume exceeding 100 in any week *(say, we want to test transactions for compliance)*. Visualize the threshold:

```r
p = p + geom_vline(xintercept = 100)
p
```



Now, compare estimate based on normal distribution vs skew normal:

```r
round(1- (pnorm(100, mean = mean(SalesTransSummary$Volume), sd = sd(SalesTransSummary$Volume))),2)
```

```
## [1] 0.06
```

```r
round(1 - psn(100, xi = exi, omega = eomega, alpha = ealpha),2)
```

```
## [1] 0.07
```

So the skew normal estimates that there is a 7% probability of volume exceeding 100/wk and normal estimated a 6% probability. May not seem like much but that translates to 40 additional observations that can cause thresholds to kick in or affect assurance work.

Also, keep in mind, this is a *VERY* low volume sample - most transactions environments do this volume every seconds. AND most transaction environments have a far heavier skew *(larger values of omega and alpha)*, so this difference becomes very significant in the real world.