Chapter 3 - A closer look at assumptions

3.2 Robustness - When are the t-tools appropriate?

The short answer: For small sample sizes, t-tools are appropriate when the data are independent from a normal distribution. For large sample sizes, t-tools are appropriate when the data are independent from any distribution.

The technical answer: The t-tools are driven by the t-ratio $t = \frac{T-\mu_0}{SD(T)}$ for some statistic T. Technically, t-tools are only appropriate for conducting hypothesis tests and constructing CIs when:

1. the statistic T has a normal distribution;
2. the standard deviation $SD(T)$ has a (scaled) χ^2 distribution (take STAT422 if you want to know more!).

These two conditions hold when the data are independent from a normal distribution and the statistic is a sample mean. When the data are not normally distributed, the CLT assures that for *large enough* n, that the sample mean (NOT the data) is approximately normal, and hence the t-tools are still appropriate for conducting hypothesis tests and constructing CIs of the population mean. By *large enough* n, we mean:

- $n \geq 30$ for any distribution of a non-binary variable (even severely skewed ones)
- $n \geq 15$ for symmetric distributions

In other words, when working with the sample mean, the t-tools are **robust** to the strict requirement that the data are normal. We have already seen two important examples:

- For a 1-sample t-test of $H_0 : \mu = \mu_0$ based on independent data $Y_1, ..., Y_n$ with sample mean $\overline{Y}$ and sample SD S:
  - the statistic is $T = \overline{Y}$
  - $SD(T) = S/\sqrt{n}$
  - t-ratio is $t = \frac{\overline{Y}-\mu_0}{S/\sqrt{n}}$.
- For a 2-sample t-test of $H_0 : \mu_x - \mu_y = \Delta_0$ based on: independent data $X_1, ..., X_m$ and $Y_1, ..., Y_n$ with X_i's independent of Y_j's; sample means $\overline{X}$ and $\overline{Y}$; and sample SDs S_X and S_Y:
  - the statistic is $T = \overline{X} - \overline{Y}$
  - the unpooled $SD(T) = \sqrt{S_x/\sqrt{m} + S_y/\sqrt{n}}$
  - the t-ratio is $t = \frac{\overline{X}-\overline{Y}-\Delta_0}{\sqrt{S_x/\sqrt{m}+S_y/\sqrt{n}}}$.

Later in STAT411, we will also use t-tools as a follow-up to ANOVA and regression analyses.

3.2.3 Robustness of the pooled 2-sample t-tool

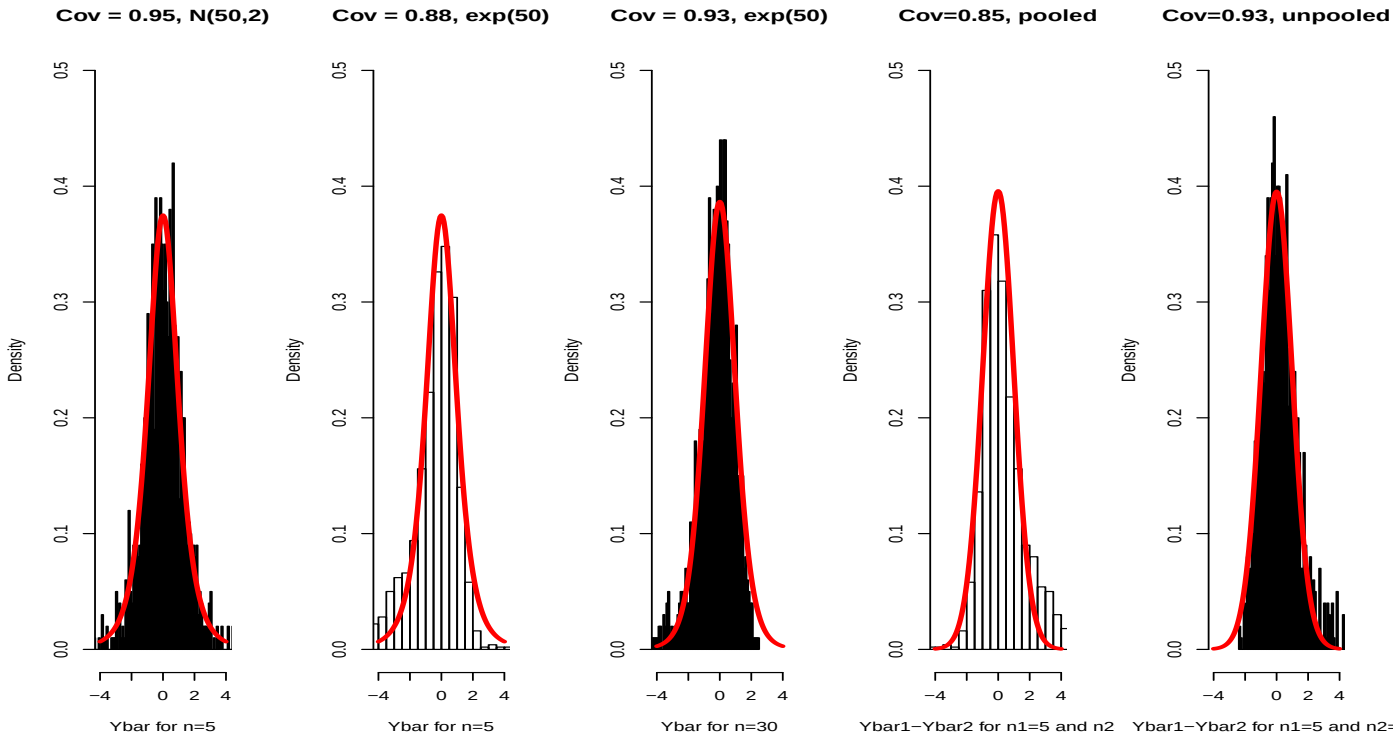
Your textbook focuses only on the robustness of the pooled 2-sample t-tools where the technical assumption is that the data are normal and that the two population SDs are equal, $\sigma_1 = \sigma_2$:

- 2-sample pooled t-tools are robust to deviations from the assumption that $\sigma_1 = \sigma_2$ when the sample sizes are roughly equal.
- 2-sample pooled t-tools are NOT robust to deviations from the assumption that $\sigma_1 = \sigma_2$ when the sample sizes are very different.

Either way, don't use a pooled 2-sample test and you will not have to worry about satisfying $\sigma_1 = \sigma_2$ for any sample sizes. **There is a scenario when the pooled t-test is useful.** In fact, the pooled test is the only possible option when you wish to compare two population means but one of your two samples only has a single datum.

This next plot illustrates some of these ideas for 10^3 simualated random samples from two different populations. The R-code is provided at the end of these notes.

- **$\overline{Y}$ is normal when the data are normal:** The first pane on the left shows a histogram of the t-ratio for a sample mean $\overline{Y}_1$ for $n_1 = 5$ data drawn from $N(\mu = 50, \sigma_1 = 2)$. A $t(4)$ distribution is overlaid on the histogram. These agree well because the data are normal. The actual confidence level of $\overline{Y}_1 \pm t_{0.975,4} S/\sqrt{5}$ is 95%; the target value is 95%.



- **$\overline{Y}$ may not be normal for small sample sizes:** The second pane from the left shows a histogram of the t-ratio for a sample mean $\overline{Y}_2$ for $n_2 = 5$ data drawn from the severely right skewed $EXP(\mu = 50, \sigma_2 = 50)$ distribution. A $t(4)$ distribution is overlaid on the histogram. The histogram and $t(4)$ DO NOT agree well because $n_2 = 5$ is small. The actual confidence level of $\overline{Y}_2 \pm t_{0.975,4} S/\sqrt{5}$ is 88%, which is a bit less than the target value of 95%.

- **$\overline{Y}$ is approximately normal for large sample sizes:** The middle pane shows a histogram of the t-ratio for a sample mean $\overline{Y}_3$ for $n_3 = 30$ data drawn from the same severely right skewed $EXP(\mu = 50, \sigma_3 = 50)$ distribution. A $t(29)$ distribution is overlaid on the histogram. These agree well by CLT because $n_3 = 30$ is sufficiently large. The actual confidence level of $\overline{Y}_3 \pm t_{0.975,4} S/\sqrt{5}$ is 93%, which is close to the target of 95%.

- **Do not use a pooled 2-sample t-test unless you have to:** The second pane from the right depicts the results from a 2-sample pooled t-test when the assumptions are violated (as your book points out). The pane shows a histogram of the t-ratio for the difference in sample means $\overline{Y}_1 - \overline{Y}_3$, where $\overline{Y}_1$ was calculated for $n_1 = 5$ data drawn from $N(\mu = 50, \sigma_1 = 2)$ and $\overline{Y}_3$ for $n_3 = 30$ data drawn from the severely right skewed $EXP(\mu = 50, \sigma_3 = 50)$. A $t(5 + 30 - 2 = 33)$ distribution, appropriate for a 2-sample pooled t-test when the assumptions hold, is overlaid on the histogram. The histogram and $t(33)$ DO NOT agree well because $\sigma_1 = 2$ is very different than $\sigma_3 = 50$

and because the samples $n_1 = 5$ and $n_3 = 30$ are very different. The actual confidence level of $\overline{Y}_1 - \overline{Y}_3 \pm t_{0.975,29} S_p^2 \sqrt{1/5 + 1/30}$ is 85%, which is NOT very close to the target of 95%. The pooled variance is $S_p^2 = \frac{4S_1^2 + 29S_3^2}{5+30-2}$.

- **Using a Welch 2-sample t-test for a large non-normal sample:** The first pane on the right depicts the results from a Welch 2-sample (unpooled) t-test that overcomes the shortcomings of the pooled approach. The pane shows a histogram of the t-ratio for the difference in sample means $\overline{Y}_1 - \overline{Y}_3$, where $\overline{Y}_1$ was calculated for $n_1 = 5$ data drawn from $N(\mu = 50, \sigma_1 = 2)$ and $\overline{Y}_3$ for $n_3 = 30$ data drawn from the severely right skewed $EXP(\mu = 50, \sigma_3 = 50)$. A $t(29)$ distribution, a bit conservative for a 2-sample unpooled t-test, is overlaid on the histogram. The histogram and $t(29)$ agree well because we do not need to assume that $\sigma_1 = \sigma_3$ and because the sample size $n_3 = 30$ is large for the non-normal sample. The actual confidence level of $\overline{Y}_1 - \overline{Y}_3 \pm t_{0.975,29} \sqrt{S_1^2/5 + S_2^2/30}$ is 93%, which is close to the target of 95%.

3.2.4 Robustness against departures from independence

Be wary of **cluster (or random) effects** and **serial effects**! Either effect means the individuals are were not chosen by a RS and hence the data are not independent! If these effects exist, a t-tool is not appropriate because the t-tool calculates the SD of the statistic assuming independence of the data. You can still analyze the data when there are cluster and serial effects, it's just that a t-tool is not appropriate!

3.3 Resistance

A statistical procedure is resistant to an outlier, or a few outliers, if statistical conclusions do not change whether you analyze either: (1) the full data set or (2) the smaller data set that does not include the outliers.

- Sample averages and sample standard deviations are not resistant to outliers.
- Any method that uses averages and/or standard deviations are not resistant to outliers.
- z-tests, t-tests, ANOVA, and regression are not resistant to outiers.

3.4 Dealing with outliers

- Do not use "automatic" outlier detection schemes (e.g., the 1.5IQR rule) or automatically exclude an outlier from a data set. Use graphical approaches to identify outliers such as boxplots, histograms, normal probability plots, individual value plots, and residual plots.
- It is imperative to investigate, vet, and report outliers.
- Only exclude or "correct" an outlier if you have a scientifically valid reason. For example, the outlier may have been entered into a computer incorrectly, or there was a severe deviation to the standard operating procedure that generated the outlying data.
- Many times an outlier is a real datum from the population, and as such is very informative about the mean and variance (and perhaps other parameters too) of the underlying population. Hence, unless there is a very good reason for exclusion, outliers ought to be left in the data set.
- Sometimes, an outlier occurs due to the *limit of detection* of the device or method used to take measurements. For example, does a 0 in the data set really mean 0 or does it mean that the device failed to detect anything? Data that occurs below (or above) a limit of detection are *censored data*. Censored data are informative about the underlying population and should not automatically be

excluded. There are many useful statistical approaches for the analysis of data sets with censored data (e.g., R's `NADA` package).

- Sometimes an outlier occurs due to a different technician being employed for a short duration during the study. Hey, everyone needs a vacation. Use a statistical method that accounts for technician-to-technician (or machine-to-machine or lab-to-lab or year-to-year) variance (R's `nlme` or `lme4` packages).
- When addressing outliers, think outside the box. Confer with others on your research team. It is not the data analysts' job to make a decision about what "to do" with an outlier. It is the study director's decision. You may very well be the study director!
- Keep a notebook when collecting data and refer back to it to see if there were any anomolies on the day when the outlying data were collected.
- *When in doubt, leave the outlier in the data set.*
- If it is determined that it is appropriate that outlying data be excluded from a data set, keep an accurate record of what data were excluded and why and whose decision it was to remove the data. In fact, you should keep a diary of all statistical analyses performed so that you can: (1) remind yourself how you analyzed the data; (2) re-analyze the data and reproduce the statistical results; and (3) more easily apply the approach to a new data set. This is the primary reason that all homework assignments require you to include all R-code in an Appendix.
- In the report or published paper that describes conclusions from data, it is appropriate to point out ALL outliers in the data set (those included and excluded).
- Display 3.6 in your textbook provides an excellent summary of a scientifically valid approach to dealing with outliers.

3.5 Evaluating and transforming to normality

For large sample sizes ($n \geq 30$), we do not need to assume that the data $\{X_i\}$ are normal to test or find a CI for μ_X. But when we have a small sample ($n < 15$) from a population which is clearly non-normal, then what do you do?

There are many statistical procedures that require that the data are a RS from a normal distribution. When your data $X_1, ..., X_n$ are not from a normal distribution, in many cases, you can "transform the data to normality," yielding transformed data $Y_1, ..., Y_n$ that are normally distributed.

One common approach is to use **BOX-COX POWER TRANSFORMATION**. Let X be the response variable before transformation and let Y be the response variable after transformation. A power transformation consists of raising X to the power λ for each case, as long as $\lambda \neq 0$. When $\lambda = 0$, the natural log transformation is appropriate. Mathematically,

$$Y_i = X_i^\lambda, \text{ if } \lambda \neq 0$$
$$Y_i = \ln(X_i), \text{ if } \lambda = 0.$$

The Box-Cox transformation is a slight modification of the power transformation: $Y_i = \frac{X_i^\lambda - 1}{\lambda}$. The reason for such a modification is that

$$\lim_{\lambda \to 0} \frac{X_i^\lambda - 1}{\lambda} = \ln(X_i).$$

Put another way, the natural log transformation is a special case of the power transformation. The $\log_{10}$ transformation can be substituted for the natural log transformation because

$$\log_{10}(X) = \log_{10}(e) \times \ln(X) \approx 0.434294 \ln(X) \text{ and}$$

$$\ln(X) = \ln(10) \times \log_{10}(X) \approx 2.3026 \log_{10}(X).$$

That is, the log in one base is just a constant multiple of the log in another base. Multiplying by a constant does not change the normality of the distribution.

The process:

1. Check for the need to transform the data $X_1, ..., X_n$. Start with the null hypothesis that **the data are normal** and then look for evidence which suggests that the data are NOT normal.
2. If the evidence suggests that the data are NOT normal, then a transformation is necessary. Find the appropriate λ to use in the transformation from a Box-Cox plot. R cannot find a Box-Cox transform, such a log-transform, if any data are not positive.
3. Check the effectiveness of your transformation: does the new data look normal?

To check for the need for a transformation in step 1 and effectiveness of the transformation in step 3, use density plots, boxplots, and normal probability plots. Another approach to help judge normality is using the *correlation coefficient* to measure the linearity on the normal probability plot.

- Boxplots and Density Plots
  - If these plots are symmetric, then the evidence fails to suggest that the data are not normal.
  - If these plots are severely skewed, then the evidence suggests that the data are not normal.
- Normal Probability Plot - A plot of the data versus values from a normal distribution.
  - If the points follow a linear pattern, then the evidence fails to suggest that the data are not normal.
  - If the points greatly deviate from a linear pattern, then the evidence suggests that the data are not normal.
- Correlation Coefficient (r)

Values to Which r Can Be Compared to Check for Normality

n	5	10	15	20	25	30	40	50	60	75
Critical r	0.832	0.880	0.911	0.929	0.941	0.949	0.960	0.966	0.971	0.976

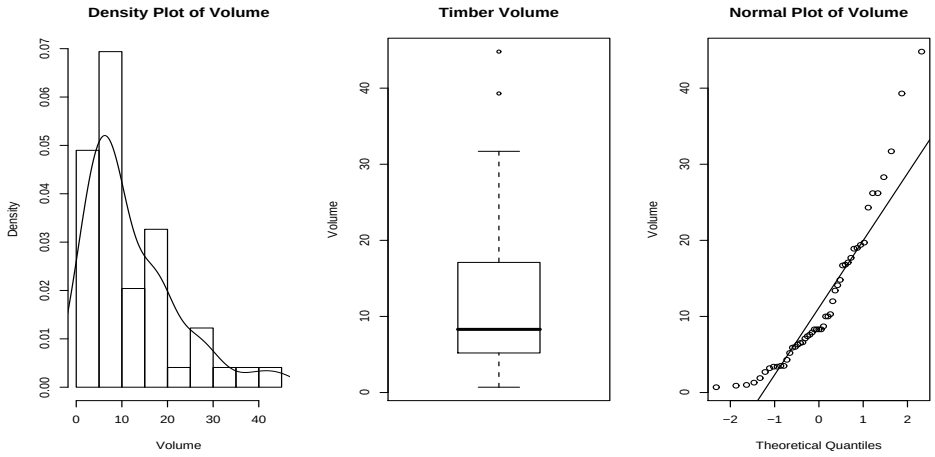
  - The correlation coefficient r quantifies the strength of the linear pattern on the normal probability plot.
    - If $r \geq$ critical r for the given sample size, then the evidence fails to suggest that the data are not normal.
    - If $r <$ critical r for the given sample size, then the evidence DOES suggest that the data are not normal.

EXAMPLE: Timber volume is estimated for $n = 49$ different trees. The R-variable name for these data is `Volume`.

1. Check the data $X_1, ..., X_n$ for normality

- Density plot, boxplot, and normal probability plot

```
d=read.table("http://www.math.montana.edu/parker/courses/STAT411/timbervolume.txt",
    header=T)
attach(d)
par(mfrow = c(1,3)) # Make three columns in the figure window
hist(Volume,freq=FALSE,main="Density Plot of Volume",xlab="Volume")
lines(density(Volume))
boxplot(Volume,main="Timber Volume",ylab="Volume")
qqnorm(Volume,main="Normal Plot of Volume",ylab="Volume")
qqline(Volume)
```



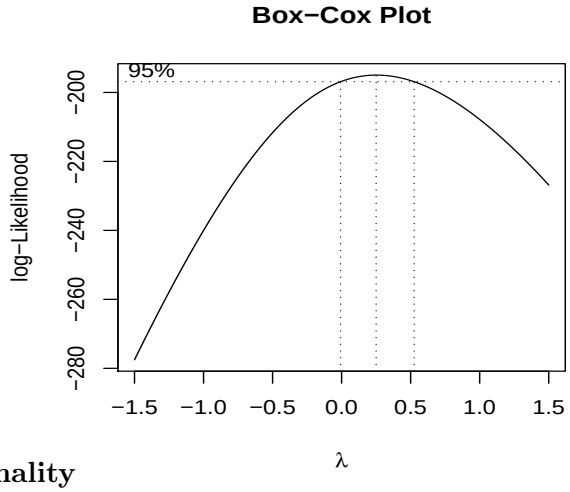
– Calculate r to check for normality:

```
xy=qqnorm(Volume)
cor(xy$y,xy$x)
[1] 0.9332381
length(Volume)
[1] 49
```

– Interpretation: From the density plot and boxplot, we see that the data are severely right-skewed. The points in the normal probability plot deviate from a linear pattern. In the Table, our sample size ($n = 49$) is not given. Instead, use the next larger n given, $n = 50$, giving the critical r value of 0.966. The correlation is $r = 0.9332 < 0.966$. Thus there is evidence to believe the data are not normal.

2. Find an appropriate λ value

```
library(MASS)
par(mfrow = c(1,1))
boxcox(Volume~1,plotit=TRUE,
    lambda=seq(-1.5,1.5,0.01))
title(main="Box-Cox Plot")
```



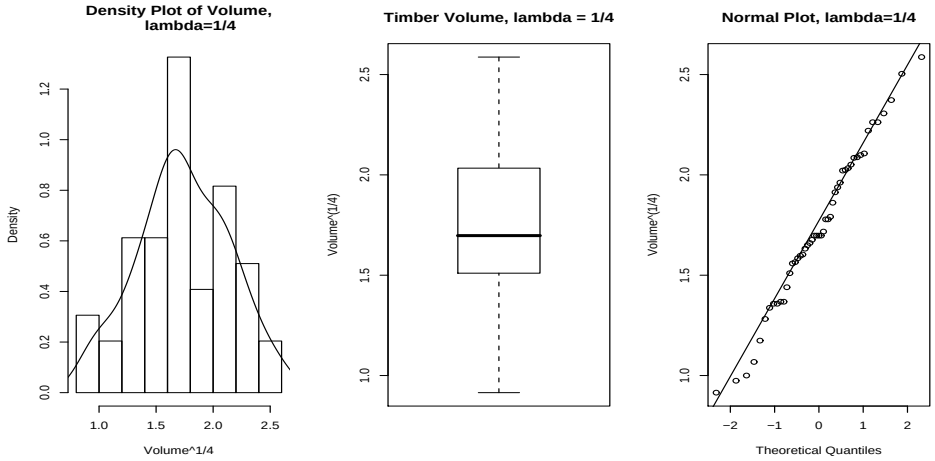
Thus, we'll transform the data $Y = X^{1/4}$.

```
Volume.new=Volume^(1/4)
```

3. Check the new data $Y_1, ..., Y_n$ for normality

– Density plot, boxplot, and normal probability plot

```
par(mfrow = c(1,3))
hist(Volume.new,freq=FALSE,main="Density Plot of Volume, lambda=1/4",xlab="Volumeˆ(1/4)")
lines(density(Volume.new))
boxplot(Volume.new,main="Timber Volume, lambda=1/4",ylab="Volumeˆ(1/4)")
qqnorm(Volume.new,main="Normal Plot, lambda=1/4",ylab="Volumeˆ(1/4)")
qqline(Volume.new)
```



– Calculate r to check for normality:

```
xy.new=qqnorm(Volume.new)
cor(xy.new$y,xy.new$x)
[1] 0.994659
```

– Interpretation: From the density plot and boxplot, we see that the transformed data symmetric. The points in the normal probability plot form a linear pattern. The correlation r calculated using the transformed data is $0.9947 > 0.966$. Thus, the evidence does not suggest that the transformed data are not normal (i.e. the transformed data looks normal).

3.5.2 Estimating μ after transforming data

After a "transform to normality", such as the Box-Cox family of transforms $Y = X^{\lambda}$, one can not directly interpret the point estimate $\overline{y}$ or a CI for μ_Y.

1. **Box-Cox transform $Y = X^{\lambda}$ with $\lambda \neq 0$**

   - The mean μ_X is always estimated by $\bar{X}$, but you may want to use medians as a measure of center due to skew or the presence of outliers.
   - $\boxed{\text{The median } \tilde{\mu}_X \text{ is estimated by } \overline{y}^{\frac{1}{\lambda}}}$ if Y is normal. This is because medians transform to medians.
   - The back-transformed interval

     $$\boxed{\left(\left(\overline{y} - t_{1-\frac{\alpha}{2},\, n-1}\frac{s_y}{\sqrt{n}}\right)^{\frac{1}{\lambda}}, \left(\overline{y} + t_{1-\frac{\alpha}{2},\, n-1}\frac{s_y}{\sqrt{n}}\right)^{\frac{1}{\lambda}}\right)}$$

     – is a $100(1-\frac{\alpha}{2})$ CI for $(\mu_Y)^{\frac{1}{\lambda}}$.
     – is a $100(1-\frac{\alpha}{2})$ CI for the median $\tilde{\mu}_X$ if Y is normal.
   - When $\lambda < 0$, you'll need to swap the endpoints of the CI

2. **Box-Cox transform $Y = \ln(X)$ $(\lambda = 0)$**

- The mean μ_X is always estimated by $\bar{X}$, but you may want to use medians as a measure of center due to skew or the presence of outliers.
- The median $\tilde{\mu}_X$ is estimated by $\exp(\overline{y})$ if Y is normal. This is because medians transform to medians. By the rules of logs, $\exp(\overline{y}) = (\Pi_i x_i)^{\frac{1}{n}}$, which is the **geometric mean** of the untransformed data.
- The back-transformed interval

$$\left(\exp\left(\overline{y} \; - \; t_{1-\frac{\alpha}{2},\; n-1}\frac{s_y}{\sqrt{n}}\right), \exp\left(\overline{y} \; + \; t_{1-\frac{\alpha}{2},\; n-1}\frac{s_y}{\sqrt{n}}\right)\right).$$

  - is a $100(1-\frac{\alpha}{2})$ CI for $\exp(\mu_Y)$, the population geometric mean of X.
  - is a $100(1-\frac{\alpha}{2})$ CI for the median $\tilde{\mu}_X$ if Y is normal.
- In the 2-sample case, when the log-transformed data have symmetric distributions,
  - $\overline{Y_1} - \overline{Y_2} = \overline{\ln X_1} - \overline{\ln X_2}$ estimates the difference in medians Median(Y_1) - Median(Y_2) = $\ln \frac{MedianX_1}{MedianX_2}$.
  - $\exp\left(\frac{\overline{Y_1}}{\overline{Y_2}}\right)$ estimates $\frac{MedianX_1}{MedianX_2}$
  - An additive effect on the log-scale, either $Y^* = Y + \delta$ or $\ln X^* = \ln X + \delta$, is a multiplicative effect on the original scale $X^* = \delta X$.

EXAMPLE continued:
For the timber volume data, we transformed volumes from $n = 49$ trees using the Box-Cox transform $Y = (Volume)^{1/4}$.

- Find an estimate for μ_Y.

- Find an estimate for the median $\tilde{\mu}_X$.

- Find a 95% CI for μ_Y.

- Find a 95% CI for the median $\tilde{\mu}_X$.

More R-code:

```
# Timber data
d=read.table("http://www.math.montana.edu/parker/courses/STAT401/data/timbervolume.txt",
               header=TRUE)
 attach(d)
 Y=Volume^(1/4)    # Box-Cox transform

# Calculate statistics on original data X
mean(Volume)
[1] 12.01837
sd(Volume)
[1] 10.03248
length(Volume)   # Sample size
[1] 49
summary(Volume)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.70    5.20    8.30   12.02   17.10   44.80

# Calculate statistics on transformed data Y
mean(Y)
[1] 1.739392
mean(Y)^4
[1] 9.153556
sd(Y)
[1] 0.3984408
summary(Y)    # 5 number summary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.9147  1.5100  1.6970  1.7390  2.0340  2.5870

# 95% CIs
# Not useful bc the Vol's are not normal
mean(Volume) + c(-1,1)*qt(.975,48)*sd(Volume)/sqrt(length(Volume))
[1]  9.136702 14.900033

# CI for mu_Y, the true mean of the Volume^(1/4)
mean(Y) + c(-1,1)*qt(.975,48)*sd(Y)/sqrt(length(Y))
[1] 1.624946 1.853838

# CI for the true median Volume
(mean(Y) + c(-1,1)*qt(.975,48)*sd(Y)/sqrt(length(Y)))^4
[1]  6.97198 11.81100


#####################################
#
# Simulation study of (1) Central Limit Theorem and (2) robustness of the t-tools

# We will take 3 samples with these sample sizes
n.1 <- 5
n.2 <- 5
n.3 <- 30   # our rule of thumb for a 'large' sample size
```

```
n.sim = 1e3 # Number of times to sample from all 3 populations
mu0=50      # The two distributions are different but have the same mean mu0
SD.norm=2   # The SD of the normal distribution

# Set up matrices for storage of simulation outputs
t.stat.vec <- matrix(NA,n.sim,5)
conf.int.mat1 <- matrix(NA, nrow=n.sim, ncol=2)
conf.int.mat2 <- matrix(NA, nrow=n.sim, ncol=2)
conf.int.mat3 <- matrix(NA, nrow=n.sim, ncol=2)
conf.int.mat4 <- matrix(NA, nrow=n.sim, ncol=2)
conf.int.mat5 <- matrix(NA, nrow=n.sim, ncol=2)

# Begin simulation
for (i in 1:n.sim) {
sample1 <- rnorm(n.1,mu0,SD.norm)   # sample from N(mu0,SD.norm)
sample2 <- rexp(n.2, 1/mu0)         # small sample from severely right-skewed exponential
sample3 <- rexp(n.3, 1/mu0)         # LARGE sample from severely right-skewed exponential

test.out1 <- t.test(sample1,mu=mu0) # 1-sample-test of small normal sample
test.out2 <- t.test(sample2,mu=mu0) # 1-sample-test of small exponential sample
test.out3 <- t.test(sample3,mu=mu0) # 1-sample-test of LARGE exponential sample
test.out12_pooled <- t.test(sample1,sample2,var.equal=TRUE) # pooled 2-sample t
test.out12_unpooled <- t.test(sample1,sample3)              # unpooled 2-sample t

# Store simulation outputs
t.stat.vec[i,1] <- test.out1$stat
t.stat.vec[i,2] <- test.out2$stat
t.stat.vec[i,3] <- test.out3$stat
t.stat.vec[i,4] <- test.out12_pooled$stat
t.stat.vec[i,5] <- test.out12_unpooled$stat
conf.int.mat1[i,] <- test.out1$conf.int[1:2]
conf.int.mat2[i,] <- test.out2$conf.int[1:2]
conf.int.mat3[i,] <- test.out3$conf.int[1:2]
conf.int.mat4[i,] <- test.out12_pooled$conf.int[1:2]
conf.int.mat5[i,] <- test.out12_unpooled$conf.int[1:2]
}  # end for simulation

# Graph results
par(mfrow=c(1,5))  # Set up 5 panes in the graph

cover.low <- ifelse(conf.int.mat1[,1] < mu0, 1, 0)
cover.up <- ifelse(conf.int.mat1[,2] > mu0, 1, 0)
coverage <- sum(cover.low*cover.up)/n.sim
hist(t.stat.vec[,1], freq=FALSE, nclass=100, ylim=c(0,.5), xlim=c(-4,4),
    xlab = sprintf("Ybar for n=%d",n.1),
    main=sprintf("Cov = %.2f, data ~ N(%d,%d)",coverage, mu0,SD.norm))
curve(dt(x, (n.1-1)), add=TRUE, col=2, lwd=3)

cover.low <- ifelse(conf.int.mat2[,1] < mu0, 1, 0)
cover.up <- ifelse(conf.int.mat2[,2] > mu0, 1, 0)
coverage <- sum(cover.low*cover.up)/n.sim
hist(t.stat.vec[,2], freq=FALSE, nclass=100, ylim=c(0,.5), xlim=c(-4,4),
```

```
    xlab = sprintf("Ybar for n=%d",n.2),
    main=sprintf("Cov = %.2f, data ~ exp(SD=%d)",coverage, mu0,SD.norm))
curve(dt(x, (n.2-1)), add=TRUE, col=2, lwd=3)

cover.low <- ifelse(conf.int.mat3[,1] < mu0, 1, 0)
cover.up <- ifelse(conf.int.mat3[,2] > mu0, 1, 0)
coverage <- sum(cover.low*cover.up)/n.sim
hist(t.stat.vec[,3], freq=FALSE, nclass=100, ylim=c(0,.5), xlim=c(-4,4),
    xlab = sprintf("Ybar for n=%d",n.3),
    main=sprintf("Cov = %.2f, data ~ exp(SD=%d)",coverage, mu0))
curve(dt(x, (n.1+n.2-2)), add=TRUE, col=2, lwd=3)

cover.low <- ifelse(conf.int.mat4[,1] < 0, 1, 0)
cover.up <- ifelse(conf.int.mat4[,2] > 0, 1, 0)
coverage <- sum(cover.low*cover.up)/n.sim
hist(t.stat.vec[,4], freq=FALSE, nclass=100, ylim=c(0,.5), xlim=c(-4,4),
    xlab = sprintf("Ybar1-Ybar2 for n1=%d and n2=%d",n.1,n.2),
    main=sprintf("Cov=%.2f, pooled 2-sample t",coverage))
curve(dt(x, (n.1+n.3-2)), add=TRUE, col=2, lwd=3)

cover.low <- ifelse(conf.int.mat5[,1] < 0, 1, 0)
cover.up <- ifelse(conf.int.mat5[,2] > 0, 1, 0)
cover <- sum(cover.low*cover.up)
coverage <- sum(cover.low*cover.up)/n.sim
hist(t.stat.vec[,5], freq=FALSE, nclass=100, ylim=c(0,.5), xlim=c(-4,4),
    xlab = sprintf("Ybar1-Ybar2 for n1=%d and n2=%d",n.1,n.3),
    main=sprintf("Cov=%.2f, unpooled 2-sample t",coverage))
curve(dt(x, (n.3-1)), add=TRUE, col=2, lwd=3)
```