Sync #3 Session Agenda

- Hypothesis Testing
- Type I and Type II Errors
- Sampling Distributions
- z-statistic
- Eight Step Hypothesis Testing Process
- Student’s t-statistic
- Central Limit Theorem Convergence
- Exponential Sampling Distribution
- Asymmetric Distributions
- Transformations
- Bootstrapping
- Contaminated Distributions
- Robust Estimation
- Test #3
- Final Exam

Statistical Hypothesis Testing

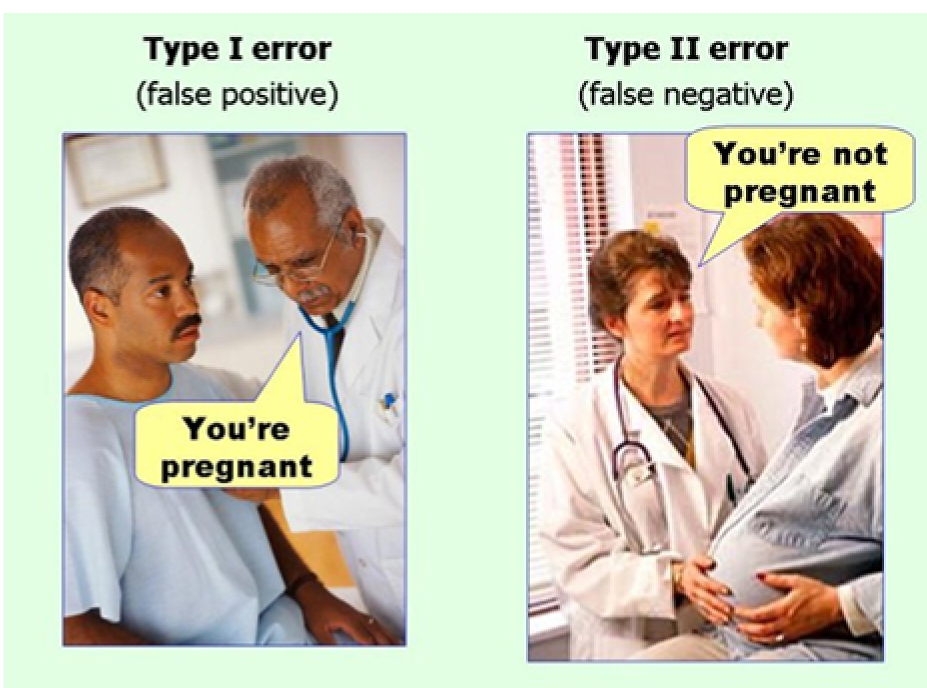
Statistical hypothesis testing involves two parts:

1) Null hypothesis
2) Alternative hypothesis

Used for decision making. The Neyman-Pearson approach involves making a choice between alternatives.

True State of Nature

		H_o True	H_A True
Decision	Do Not Reject	No Error	Type 2 Error
Matrix	Reject	Type 1 Error	No Error



Type I, Type II and Sample Size

Type I error: Fire alarm when there is no fire.

Type II error: No fire alarm when there is a fire.

You can never be right 100% of the time.

What is a suitable tradeoff of the error rates?

- **Efficacious or not in a drug trial?**

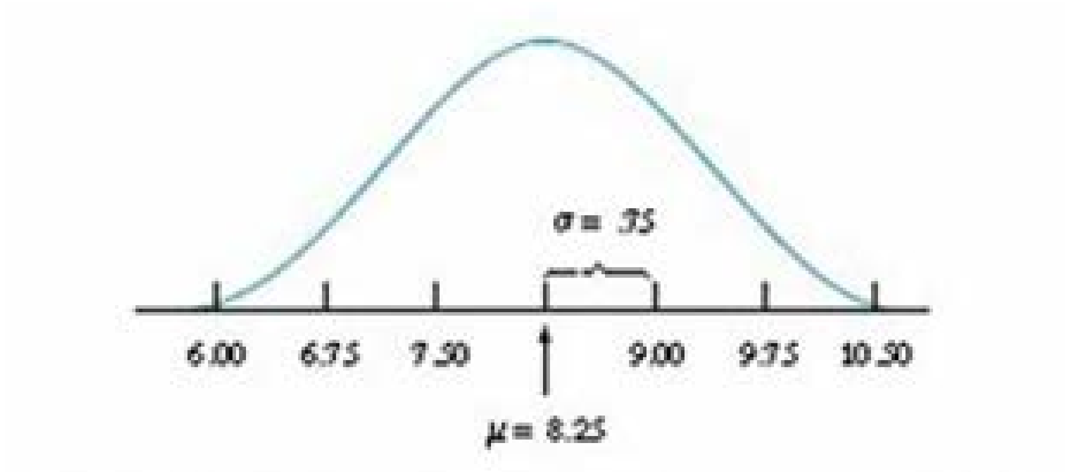
The Type 1 error rate (alpha), the Type 2 error rate (beta) and the sample size (n) are connected.

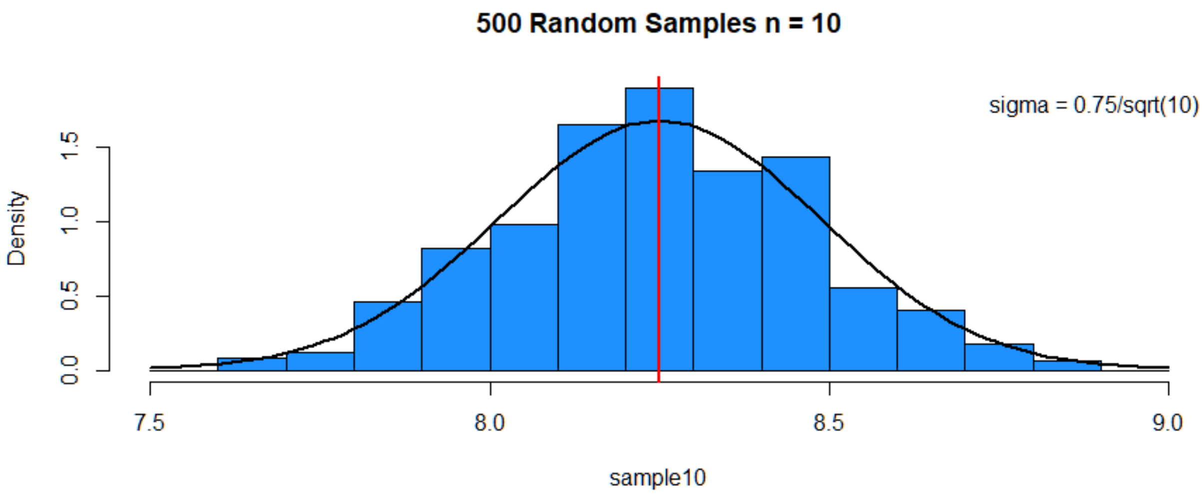
- **If the sample size is held constant, and the Type 1 error rate is increased, the Type 2 error rate is decreased and conversely.**

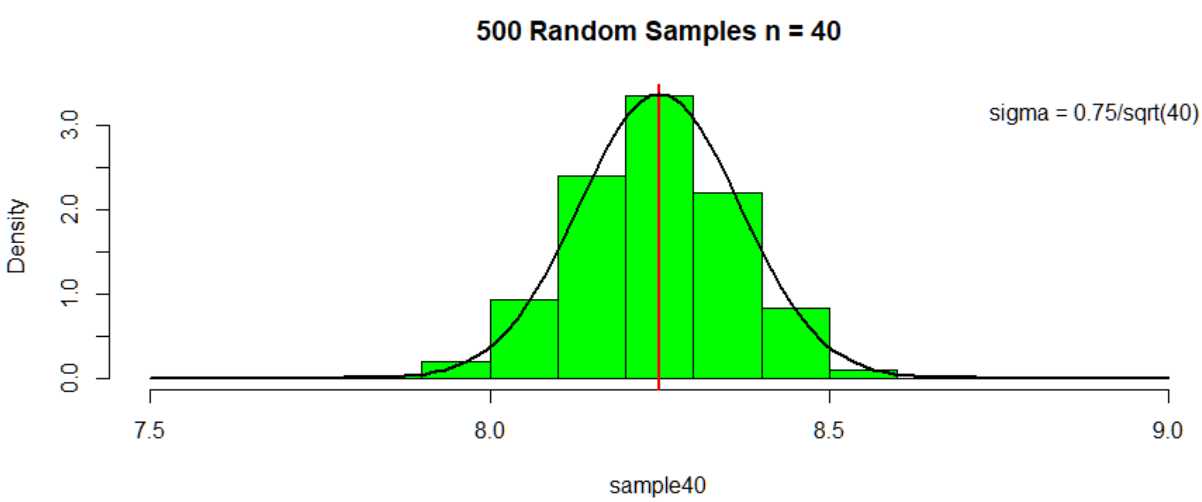
- **If the Type 1 error rate is held constant, and the sample size is increased the Type 2 error rate is decreased and conversely.**

- **If the Type 2 error rate is held constant, and the sample size is increased the Type 1 error rate is decreased and conversely.**

Sampling Distributions from Normal Population







Normal Distribution Point and Interval Estimation

Normal Distribution

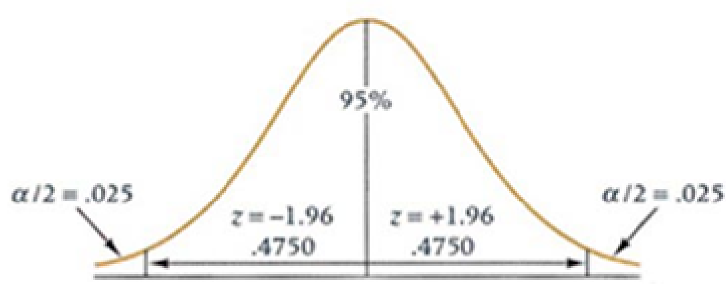
A random sample of independent observations from a normal distribution with mean μ and variance σ^2 yields a sample mean $\bar{x}$ distributed normally with mean μ and variance σ^2 / n .

Implications

The point estimate of a normal population mean is the sample mean; and, the z-statistic can be used to construct a confidence interval.

$$P\left[z_{\alpha/2} \le \left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right) \le z_{1-\alpha/2} \right] = 1 - \alpha$$

$$P\left[\bar{x} - z_{1-\alpha/2}\sigma / \sqrt{n} \le \mu \le \bar{x} - z_{\alpha/2}\sigma / \sqrt{n} \right] = 1 - \alpha$$



The standard normal distribution is used as the sampling distribution for the z-statistic. This allows quantiles (percentiles) to be selected corresponding to the desired overall confidence level. These quantiles are used for hypothesis testing and confidence interval construction.

Refer to Wilcox Section 4.5 pages 64-65 dealing with historical remarks.

z-test Examples

Decade old data indicates the national average net income for sole-proprietor CPAs was \$98,500 with a standard deviation of \$14,530. A random sample of 112 CPAs has an average of \$102,220. Has the average net income changed? Test with alpha = 0.05. Assume the income data are normally distributed.

$$H_o: \mu = \$98{,}500$$
$$H_a: \mu \neq \$98{,}500$$

Hypothesis testing:

$$P\left[z_{\alpha/2} \le \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right) \le z_{1-\alpha/2}\right] = 1-\alpha$$

```
> z <- (102220 - 98500)/(14530/sqrt(112))
> z
[1] 2.709482
> qnorm(c(0.025,0.975), mean = 0, sd = 1, lower.tail = TRUE)
[1] -1.959964  1.959964
> pnorm(z, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.003369414
```

Confidence interval:

$$P\left[\bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} - z_{\alpha/2}\sigma/\sqrt{n}\right] = 1-\alpha$$

```
> z.alpha <- qnorm(c(0.975, 0.025), mean = 0, sd = 1, lower.tail = TRUE)
> round(102220 - z.alpha*14530/sqrt(112), digits = 0)
[1]  99529 104911
```

Review Black Section 9.2 page 278 "Using the p-Value to Test Hypotheses". For a two-sided test, divide the alpha value in half and compare to the p-value. For a one-sided test, do not perform this division.

Eight-Step Process for Testing Hypotheses

Problem: Decade old data indicates the national average net income for sole-proprietor CPAs was \$98,500 with a standard deviation of \$14,530. Has this average net income changed? A random sample of 112 CPAs is planned.

Steps	z-statistic example
1) Establish a null and an alternative hypothesis.	**H_o:** $\mu = \$98{,}500$ **H_a:** $\mu \neq \$98{,}500$
2) Determine the appropriate statistical test.	$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$
3) Set the value of the Type I error rate (alpha).	**This is up to the investigator.** **How conservative does this test need to be?** $\alpha = 0.05$
4) Establish the decision rule.	**For a two-sided test, the critical z value is** ± 1.96 **since half of 0.05 is used for each tail.** $P[z_{\alpha/2} \leq (\bar{x} - \mu) / (\sigma / \sqrt{n}) \leq z_{1-\alpha/2}] = 1 - \alpha$
5) Gather sample data.	**The study should be planned in advance of collecting data. This is sound practice.**
6) Analyze the data.	**Do an EDA first. Check on outliers.** **Verify assumptions.** $z = \dfrac{102,220 - 98,500}{\dfrac{14,530}{\sqrt{112}}} = 2.71$
7) Reach a statistical conclusion.	**Reject the null hypothesis since** **2.71 is greater than 1.96.** **Alternatively, a p-value could also be reported.** **p = 0.00336 < 0.025**
8) Make a business decision.	**What does this mean?** **Is this result of practical importance?**

Connection to confidence intervals:

$$P[z_{\alpha/2} \leq (\bar{x} - \mu) / (\sigma / \sqrt{n}) \leq z_{1-\alpha/2}] = 1 - \alpha$$

$$\bar{x} - (\sigma / \sqrt{n}) z_{1-\alpha/2} \leq \mu \leq \bar{x} - (\sigma / \sqrt{n}) z_{\alpha/2}$$

The t Statistic

Point and Interval Estimation

With unknown variance, and a normal distribution, the Student's t distribution can be used. For these purposes,

$$P\left[t_{\alpha/2} \le (\bar{x}-\mu)/(s/\sqrt{n}) \le t_{1-\alpha/2}\right] = (1-\alpha)$$

$$\bar{x}-(s/\sqrt{n})t_{1-\alpha/2} \le \mu \le \bar{x}-(s/\sqrt{n})t_{\alpha/2}$$

Degrees of freedom?

$$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

$$\sum_{i=1}^{n}(X_i-\bar{X}) \ = 0$$

The degrees of freedom are (n-1) because the sum of the differences always equals zero.

If the original population is <u>not</u> normally distributed but "well behaved" with unknown variance, the Student t distribution may be used with sample standard deviation s provided n >= 40 (n >= 30 minimum).

A very large sample size may be needed with symmetric and asymmetric distributions that have outliers.

t test Example

A simple random sample of 112 sole-proprietor CPAs results in a sample mean of $102,220 with a <u>sample</u> standard deviation of $14,530. Test the null hypothesis that the population mean is $98,500 with alpha = 0.05. Determine a 95% two-sided confidence interval for the population mean.

Hypothesis testing:

$$P\left[t_{\alpha/2} \le (\bar{x}-\mu)/(s/\sqrt{n}) \le t_{1-\alpha/2}\right] = (1-\alpha)$$

```
> t <- (102220 - 98500)/(14530/sqrt(112))
> t
[1] 2.709482
> qt(c(0.025, 0.975), df = 111, lower.tail = TRUE)
[1] -1.981567  1.981567
> pt(t, df = 111, lower.tail = FALSE)
[1] 0.003904094
```

Confidence interval:

$$\bar{x}-(s/\sqrt{n})t_{1-\alpha/2} \le \mu \le \bar{x}-(s/\sqrt{n})t_{\alpha/2}$$

```
> t.alpha <- qt(c(0.975, 0.025), df = 111, lower.tail = TRUE)
> round(102220 - t.alpha*14530/sqrt(112), digits = 0)
[1]  99499 104941
```

For a one-sided lower 95% confidence interval, use the upper 0.95 quantile.

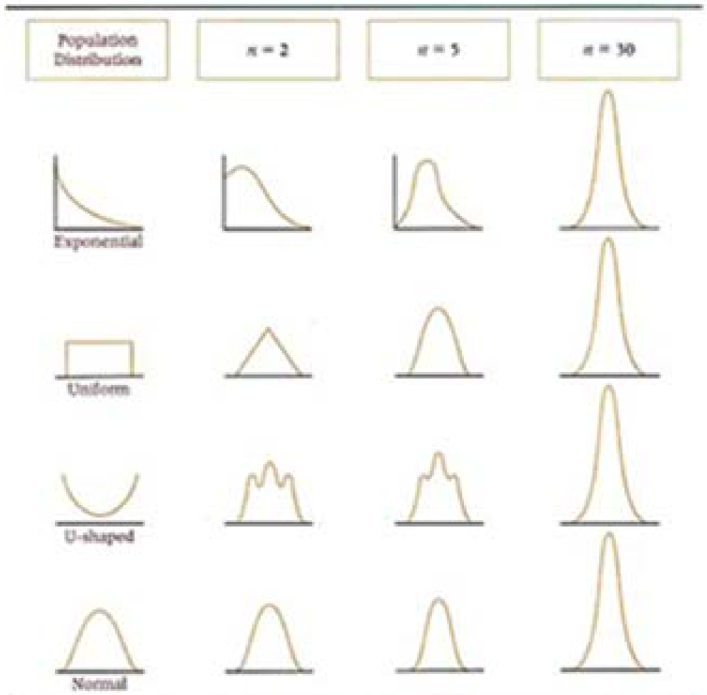
```
> t.alpha <- qt(0.95, df = 111, lower.tail = TRUE)
> round(102220 - t.alpha*14530/sqrt(112), digits = 0)
[1] 99943
```

For a one-sided upper 95% confidence interval, use the lower 0.05 quantile.

```
> t.alpha <- qt(0.05, df = 111, lower.tail = TRUE)
> round(102220 - t.alpha*14530/sqrt(112), digits = 0)
[1] 104497
```

R provides t.test() which can be used in these instances if you have a sample of data points.

Central Limit Theorem Convergence



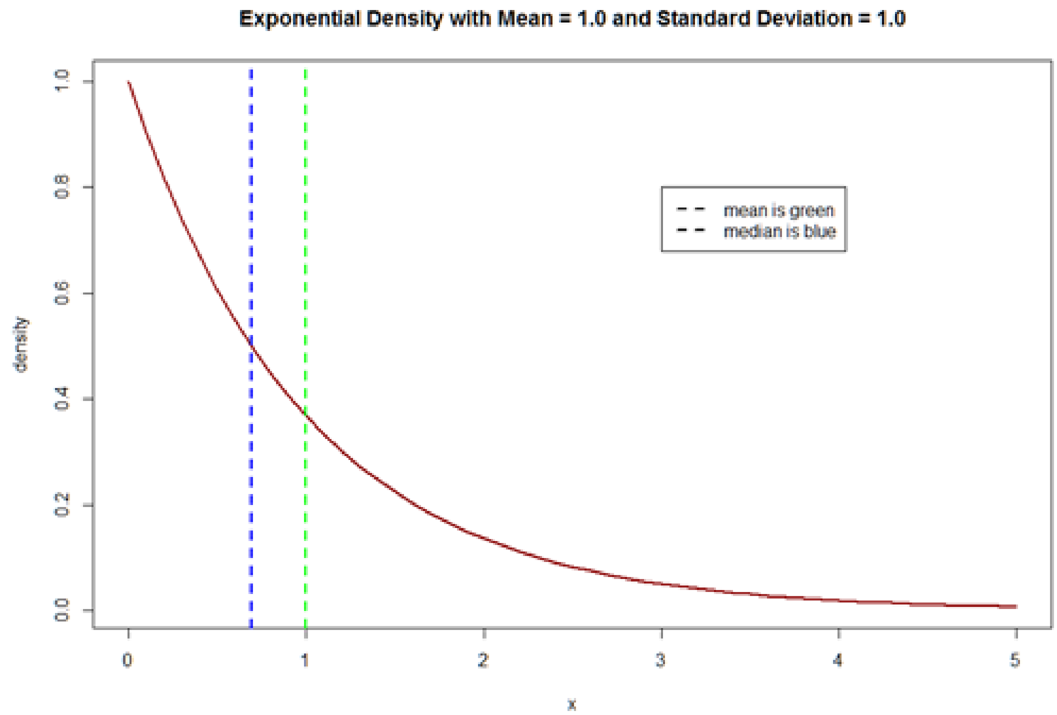
The mean $\bar{x}$ of a random sample drawn from a population with mean μ and standard deviation σ can be assumed to have approximately a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$ if n is large enough.

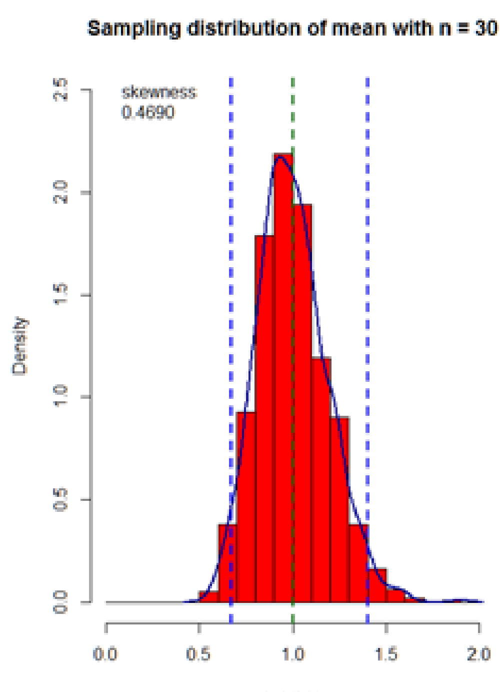
How large must be the sample size n?

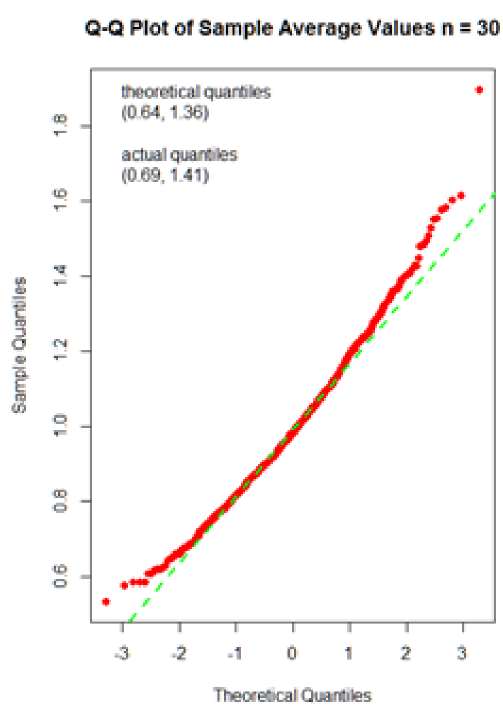
Black (page 241) "…in this text (as in many others), a sample of size 30 or larger will suffice…." Wilcox states (page 90) in general n >= 40 will suffice.

If the original population is <u>not</u> normally distributed but "well behaved" and the mean and variance are known, for n >= 40 (n >= 30 minimum) the z score has an approximate normal distribution with mean = 0 and variance = 1.

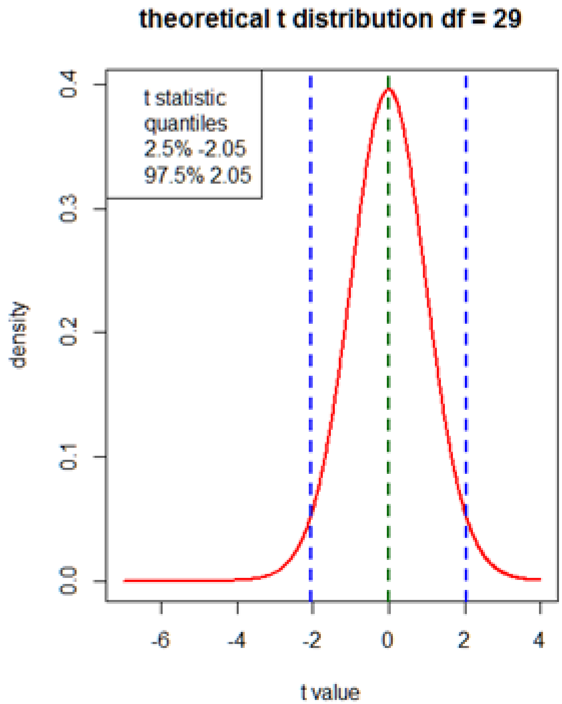
Exponential Sampling Distributions

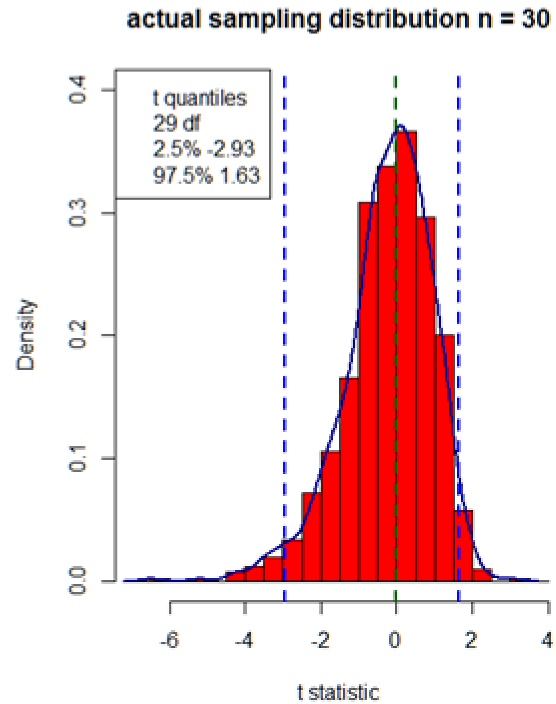


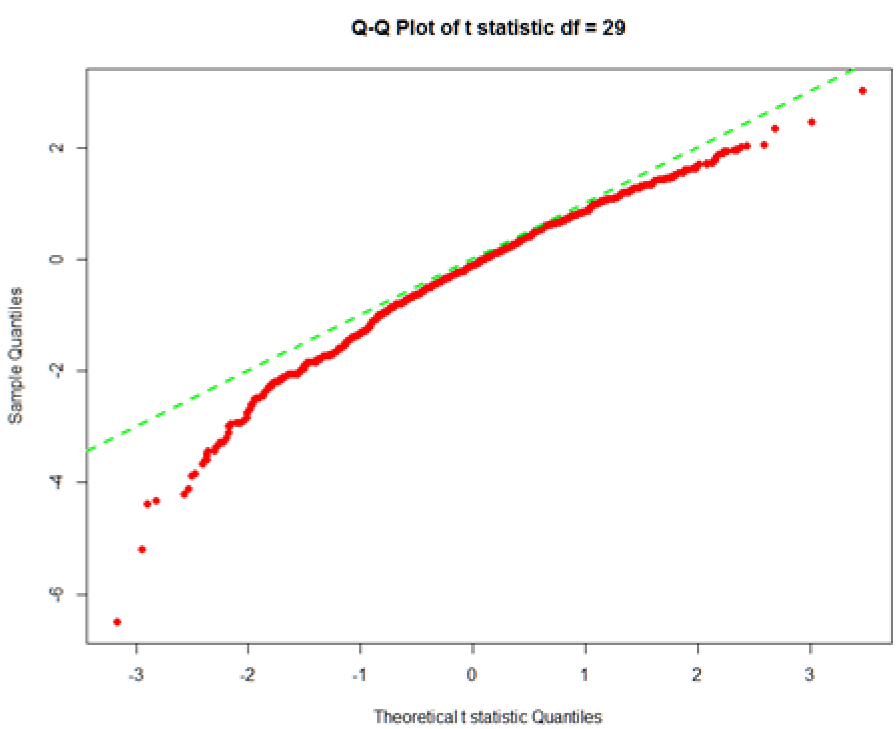




Student's t statistic Performance







Four Types of Distributions

Symmetric with few outliers (example: normal distribution)	**Asymmetric with few outliers (example: exponential distribution)**
Symmetric with outliers commonly occurring	**Asymmetric with outliers commonly occurring**

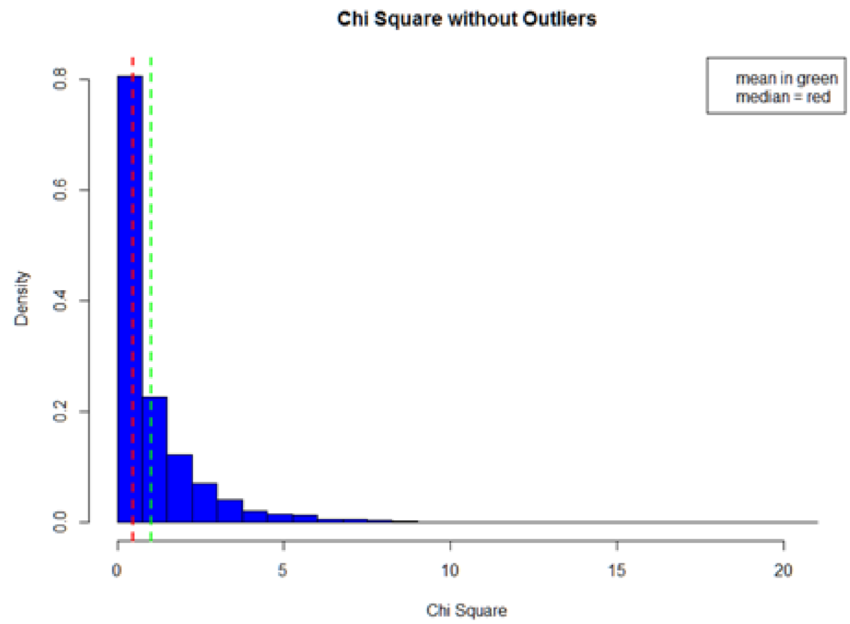
- **If the data are collected from a distribution close to normal, well established statistical inference procedures exist for this situation.**

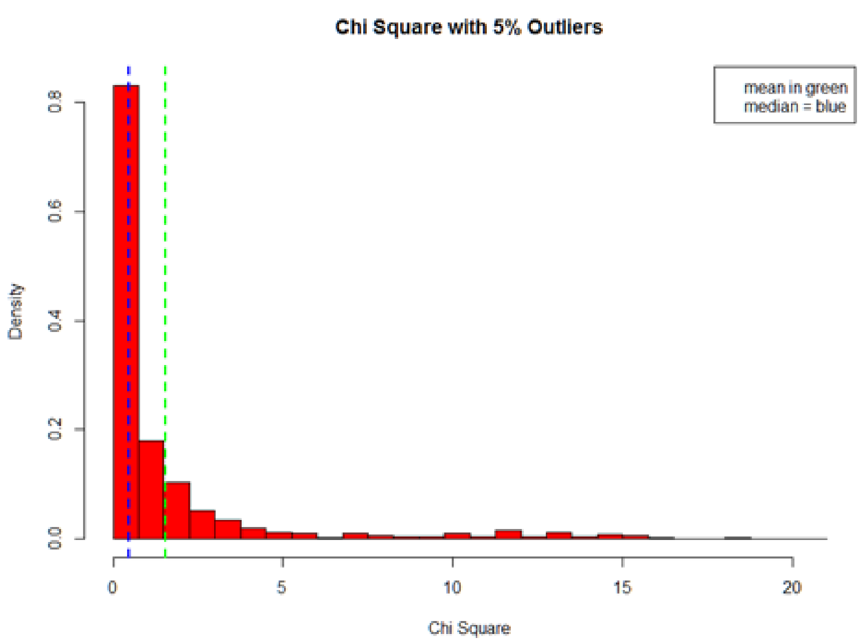
- **For asymmetric distributions with few outliers, the Central Limit Theorem may pertain with a moderate sample size.**

- **For distributions with outliers commonly occurring, a large sample size will be required for the Central Limit Theorem to be effective.**
  - **The sample mean is sensitive to outliers.**
  - **Outliers inflate the sample variance.**

Depending on the population and the statistic, the sampling distribution may be complicated. Formulas may not exist in closed-form. Estimated quantiles may be necessary for hypothesis testing and confidence intervals. Bootstrapping!

Asymmetric Distributions





Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen, "The Importance of the Normality Assumption in Large Public Health Data Sets", Annu. Rev. Public Health 2002. 23:151–69.

Transformations to Normality

If the population is substantially skewed and the sample size is at most moderate, the approximation provided by the central limit theorem can be poor, and the resulting confidence interval for the population mean will likely have the wrong coverage probability. Various methods have been used to address this problem. They include data transformations and bootstrapping.

Examples:

- **Arcsine Transform: The arcsine transform equals the inverse sine of the square root of the proportion or Y = arcsine(sqrt(p)) = sin−1(sqrt(p)) where p is the proportion and Y is the result of the transformation.**

- **Square Root Transform: The square root transformation is simply Y = sqrt(X). It is often used for counts and for other measures where group means are correlated with within group variances. The square root is used with counts that follow a Poisson distribution.**

- **Log Transform (base 10 or base e): Take the logarithm, giving Y = log(X). Useful for dealing with ratios and multiplicative variables.**

- **Power Transform: A power transform is also called a Box-Cox transform. The simplest equation for the transform for positive variables is Y = (X** λ −1)/ λ, λ ≠ 0, Y = log(X), λ = 0. The value of λ must be found using computer algorithms.**

Caveat: There is often a broader question of whether the mean and standard deviation are appropriate summary measures of central tendency and spread. In highly skewed distributions, the median might be a better reflection of what is typical of the population.

http://www.biostathandbook.com/transformation.html
https://statswithcats.wordpress.com/2010/11/21/fifty-ways-to-fix-your-data/
Kabacoff Section 8.5.2 pages 199-200.

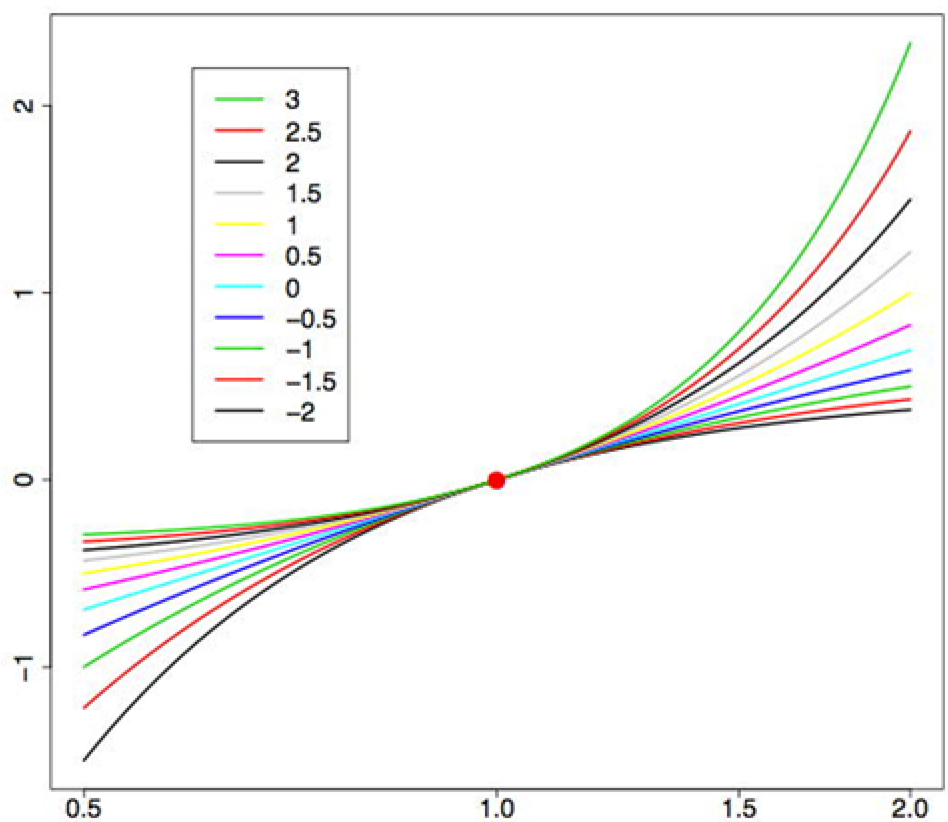
Box-Cox Power Transformations

What are the Box-Cox power transformations?

► The original form of the Box-Cox transformation, as appeared in their 1964 paper, takes the following form:

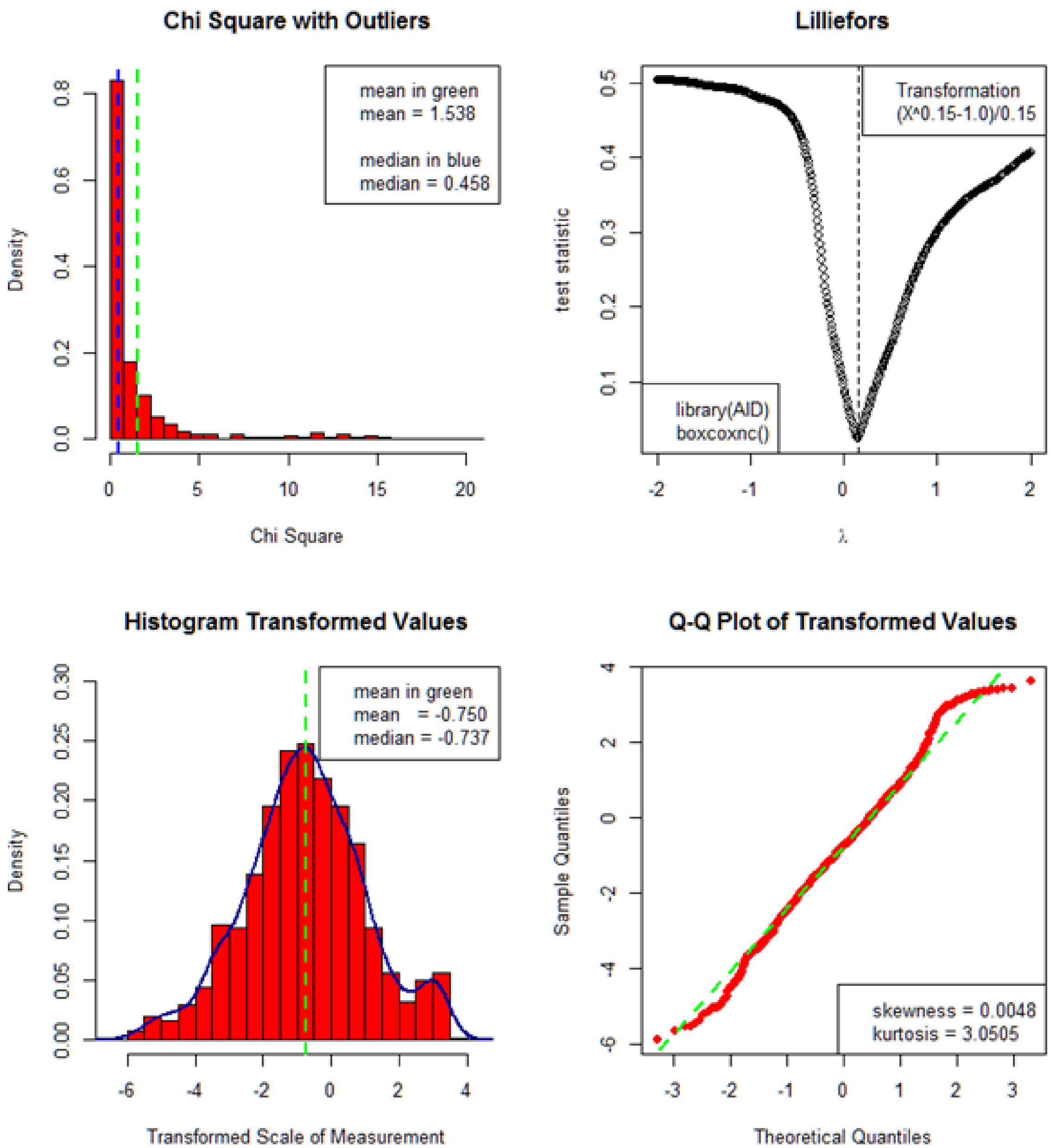
$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

What does it do to the data?



Note that the value y = 1.0 is always mapped to 0.

Power Transformation Results



Original scale	Reverse Transformation
Median = 0.458	0.15*(-0.75+1)^(1/0.15)= 0.451
Mean = 1.538	
20% trimmed mean = 0.597	

Osborne, J. W. "Improving Your Data Transformations, Applying the Box-Cox Transformation". http://pareonline.net/pdf/v15n12.pdf

Bootstrapping

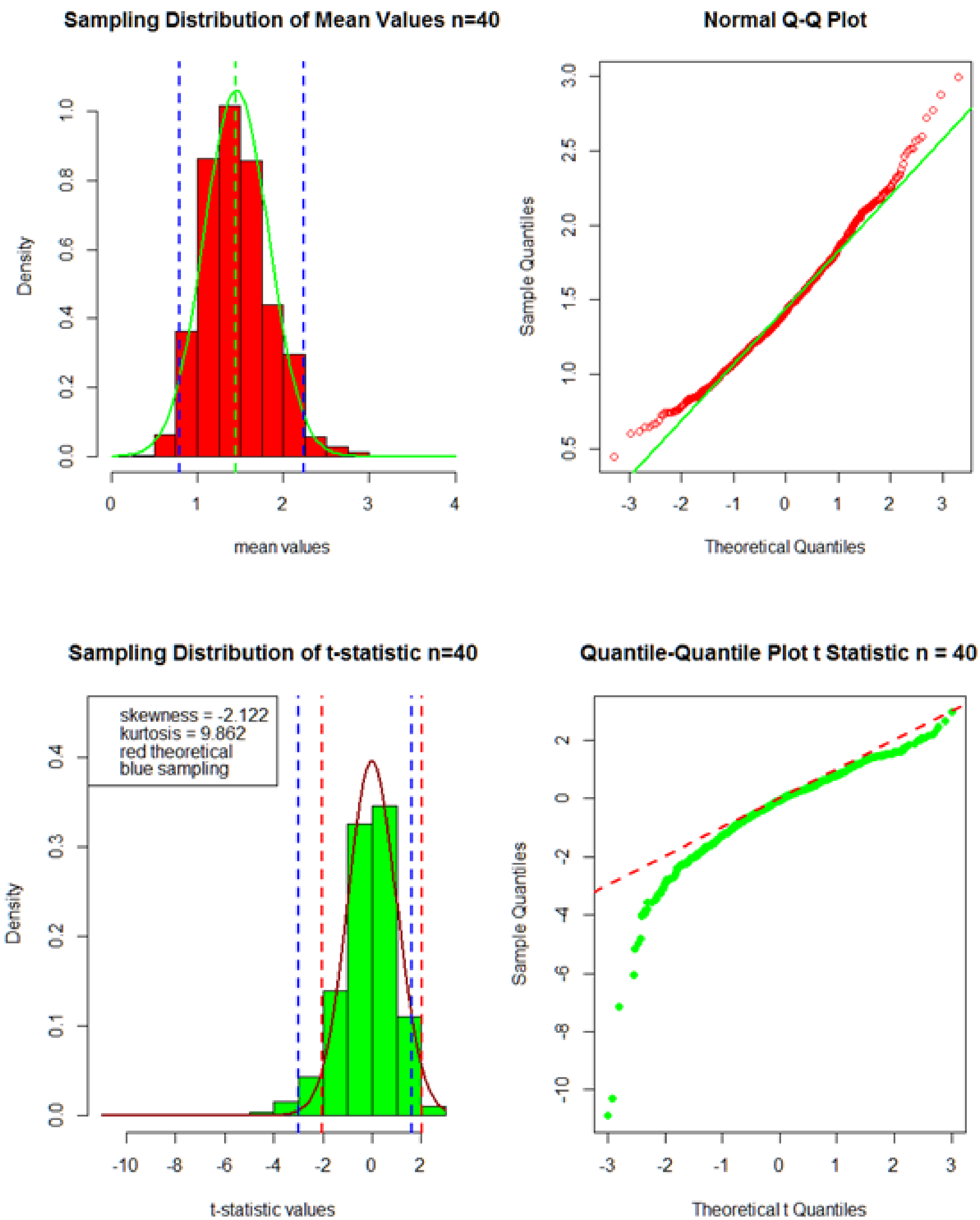
What do you do with a non-normal distribution if your sample size is small and outliers are present?

- **Student's t-statistic is unsatisfactory with small sample sizes when the distribution is non-normal particularly when outliers are present. The sample variance can inflate. The location and width of confidence intervals will be affected, which in turn affects the confidence level.**
- **Bradley Efron has been the inspiration behind the bootstrap. His book with Robert J. Tibshirani, *An Introduction to the Bootstrap,* is a classic. Over 1000 articles have been published.**
- **Bootstrap methods have been refined and are applied to:**
  - **Estimate location,**
  - **Measure association,**
  - **Perform regression.**
- **Two bootstrap methods for estimating quantiles are:**
  - **the percentile method,**
  - **the bootstrap t method.**

The idea behind all bootstrap methods is to use the data obtained from a study to approximate the sampling distribution of the test statistic so that confidence intervals may be determined and hypotheses tested.

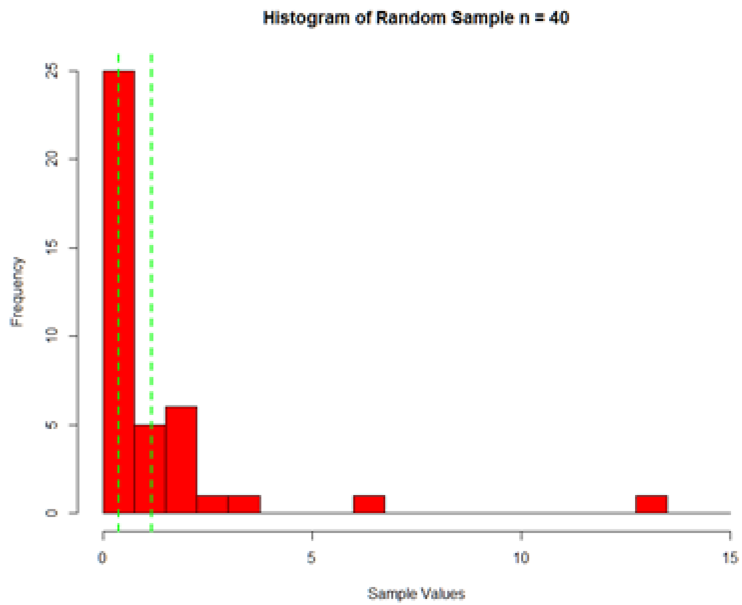
See Kabacoff Section 12.6 pages 292-298 for information on bootstrapping with the boot package.

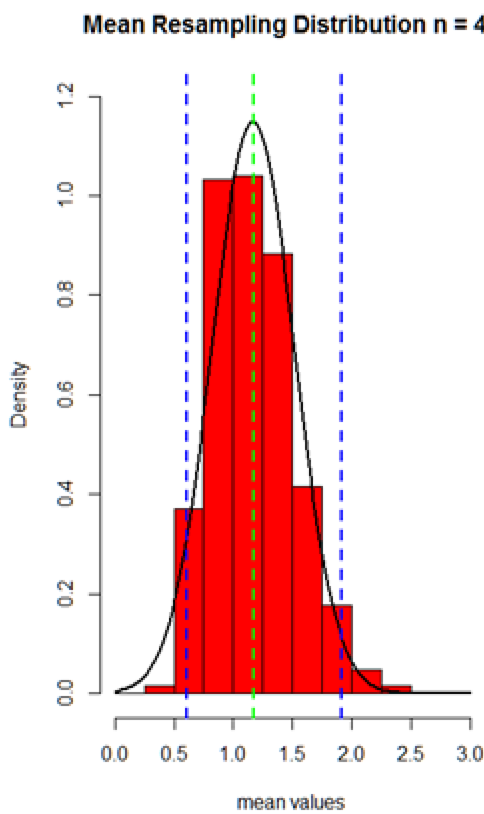
Sampling Distribution n = 40

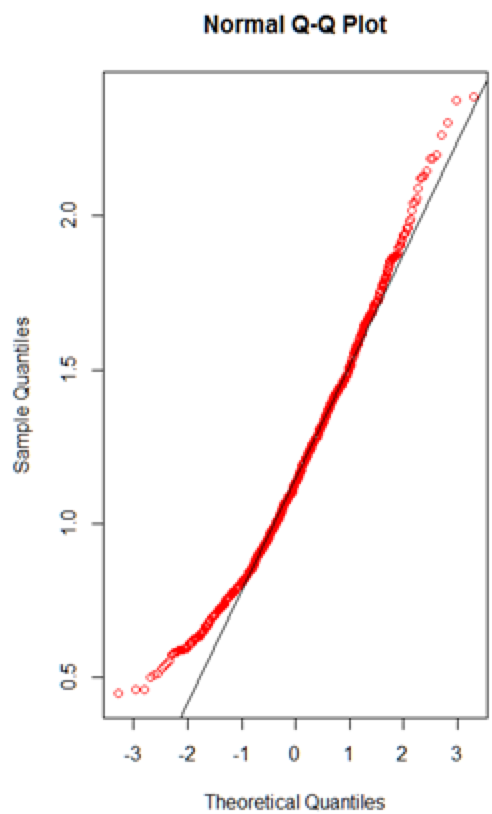


The vertical blue dotted lines show the location of the 2.5% and 97.5% quantiles.

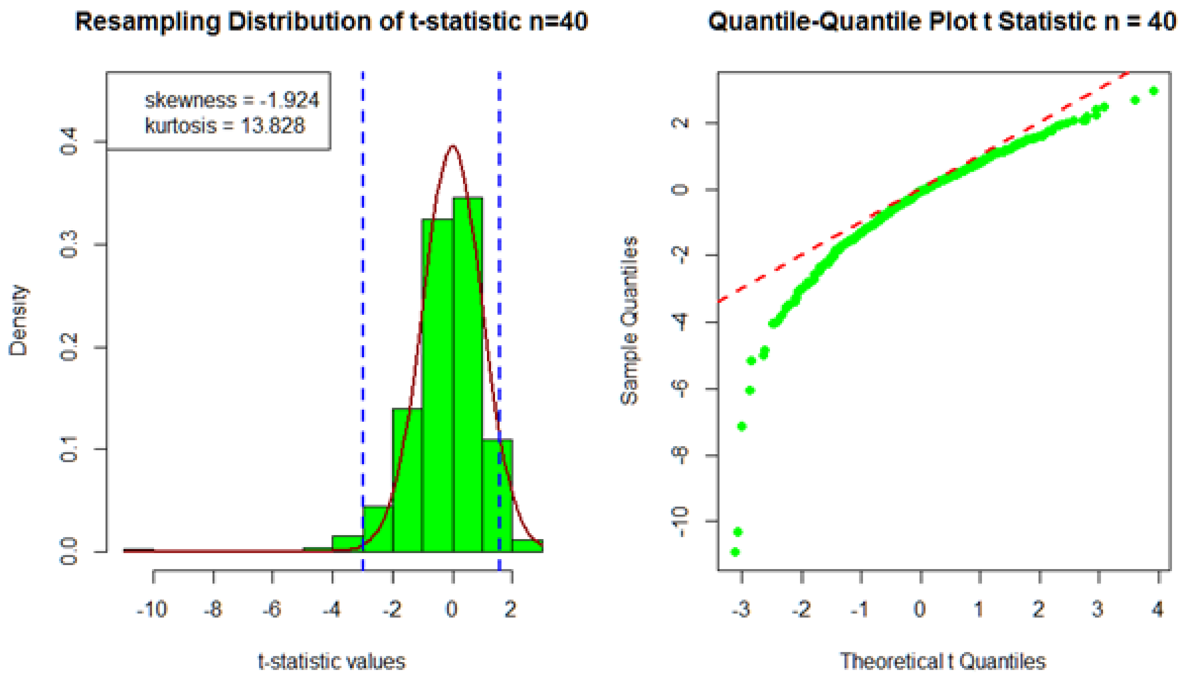
What to Do With a Single Sample?







Bootstrap t Method n = 40



$$P\left[t_{\alpha/2} \leq (\bar{x}-\mu)/(s/\sqrt{n}) \leq t_{1-\alpha/2}\right] = (1-\alpha)$$

$$\bar{x}-(s/\sqrt{n})t_{1-\alpha/2} \leq \mu \leq \bar{x}-(s/\sqrt{n})t_{\alpha/2}$$

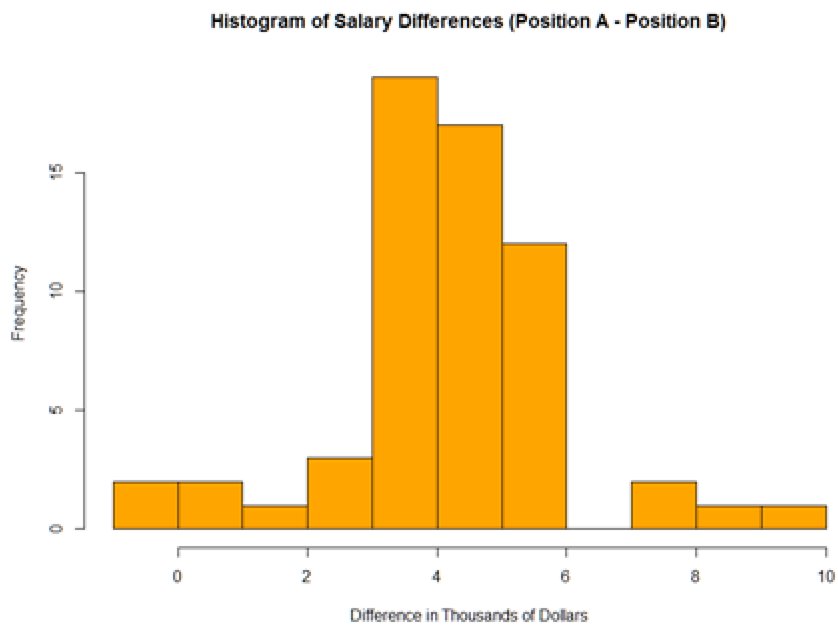
Use the sample mean and standard deviation. Substitute the empirical 2.5% and 97.5% quantiles from the resampled t distribution to obtain the 95% bootstrap t confidence interval: (0.643, 2.880).

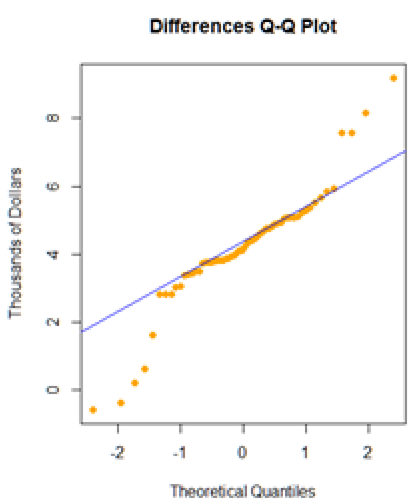
Percentile bootstrap confidence interval: (0.611, 1.912).

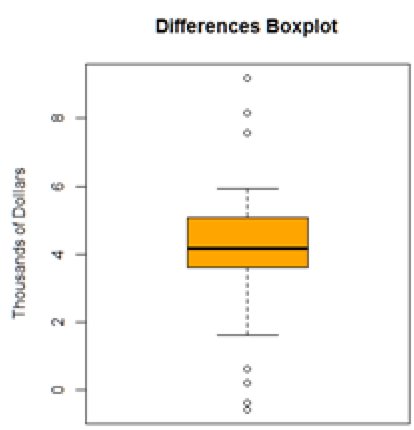
The traditional 95% confidence interval using the theoretical t distribution is: (0.432, 1.913). It is symmetric about the sample mean and does not reflect the skewed nature of the sampling distribution involved.

Heavy-Tailed (Contaminated) Distribution

A random sample of 60 colleges is taken to evaluate salary levels for different positions. Two different job titles have been selected for analysis. Position parameters for each job have been reviewed and found to be comparable across the colleges for each of the positions. The average salary difference in compensation is of interest.







Robust Measure of Location

<u>20% trimmed mean example</u> (Wilcox procedure):

4, 10, 11, 12, 15, 19, 24, 29, 31, 33, 38, 46, 69
0.2(13) = 2.6 rounded down to 2

Drop 4, 10, 46, 69 and calculate the mean: 23.56

In R use the function mean(x, trim = 0.2) on the original vector of data.

<u>20% winsorized variance example</u> (Wilcox procedure):

Replace 4, 10 with 11, and 46, 69 with 38.
11, 11, 11, 12, 15, 19, 24, 29, 31, 33, 38, 38, 38

The average of the winsorized values is used to calculate the variance.
The variance is 11.17919.

In R load the package asbio from CRAN. Use the function win() on the original vector of data. var(win(x, lambda = 0.2)) or sd(win(x,lambda = 0.2).

The variance for the 20% trimmed mean is: $s_w^2/(0.36n)$.

s_w^2 **is the sample winsorized variance and n the sample size. This results in the trimmed t-statistic for bootstrapping:**

$$T_t = \frac{\bar{X}_t^* - \bar{X}_t}{s_w^*/(0.6\sqrt{n})}$$

Here $\bar{X}_t^*$ **is a trimmed resample mean and** $\bar{X}_t$ **is the overall trimmed mean.** T_t **will have approximately a Student's t distribution with degrees of freedom equal to (n-2g-1) with g equal to the greatest integer less than or equal to 0.2n.**

Comparison of Methods

Traditional t procedure ignoring the heavy tails:

t.test(diff, alternative = c("two.sided"), mu =0, conf.level = 0.95)
95 percent confidence interval: (3.76, 4.65) mean of x = 4.21

Using a trimmed mean three different ways:

Using 20% Trimmed Mean and Winsorized Variance		
Procedure	**95% Confidence Intervals**	**20% trimmed mean**
Student's t	(3.97, 4.55)	4.26
percentile bootstrap	(3.96, 4.54)	4.26
bootstrap t	(3.94, 4.58)	4.26

Trimming reduces the confidence interval widths by approximately a third. This is the equivalent of increasing the sample size from n = 60 to n = 135, a factor of 2.25.

Rand R. Wilcox *Fundamentals of Modern Statistical Methods* 2nd ed 2009 Springer. (Available for free to NWU students from the Springer Library.)

Some Sync Session Learning Points

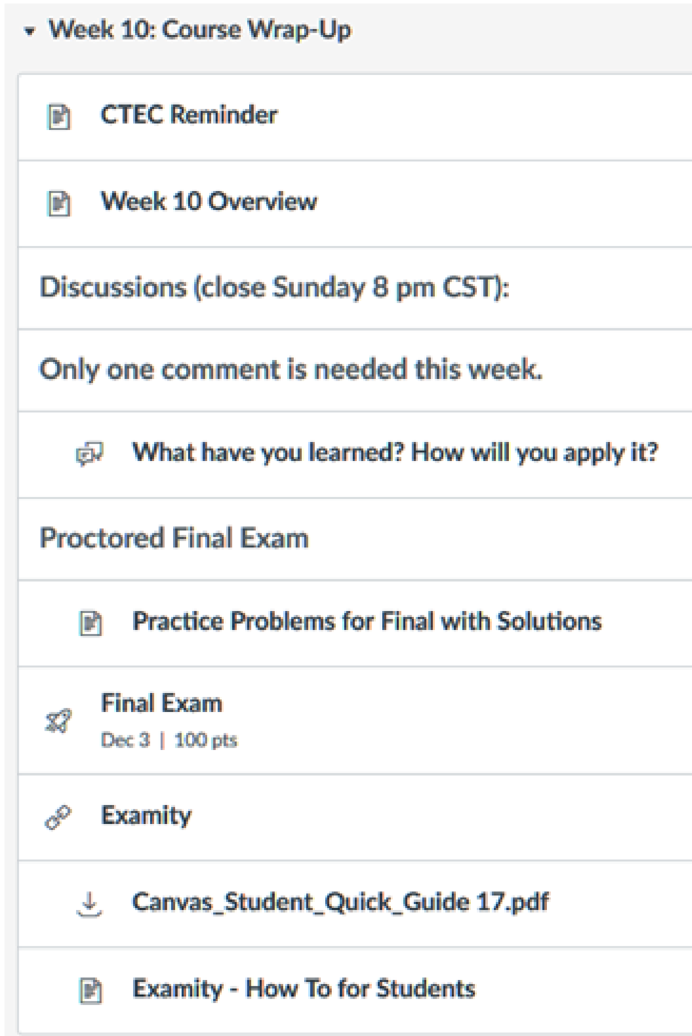
- The following statements are true for sampling distributions:
  - Confidence intervals are constructed using quantiles determined from a sampling distribution.
  - It may not be possible to express a sampling distribution in closed form mathematically.
  - A sampling distribution depends on the nature of the population being sampled.
  - The sampling distribution for a statistic is the probability distribution for that statistic based on all possible random samples from a population.
- The choice of statistical method depends on the nature of the population being studied. Symmetric and asymmetric distributions with outliers common generally require large sample sizes for application of traditional methods.
- The z-statistic may be used when the sampling distribution is normal and the population standard deviation is known.
- Student's t-statistic may be used for a random sample from a population for which the standard deviation is not known.
- Hypotheses may be tested using confidence intervals.
- Bootstrapping uses resampling with replacement. It provides an approximate sampling distribution for a statistic.
- Transformations are used to improve the symmetry of skewed distributions prior to statistical analysis.
- A symmetric heavy-tailed distribution may be detected using a box plot and QQ chart.
- The field of statistics is more a part of science than a branch of mathematics. It is not algorithmic. As more or better data come available a statistical model may be revised and conclusions may change.

Comments on Test #3

- **Confidence intervals**
  - **Means**
  - **Proportions**
  - **Standard Deviation**
- **Sample sizes**
  - **Means**
  - **Proportions**
- **Hypothesis tests**

This test will require computation for most questions. Review and practice with the solved problems on the course site could be helpful.

Final Exam



- You are responsible for scheduling your final exam.
- SPS will cover exam costs to students. You will be responsible for any late scheduling, cancellation, or no-show fees.
- Arrange a "dry run" with Examity in advance to test your equipment and get any questions answered.
- This is an open-book exam, however only one screen is allowed.
- The two-hour exam consists of ten multiple choice questions. No questions ask about R, however R may be used for calculations.