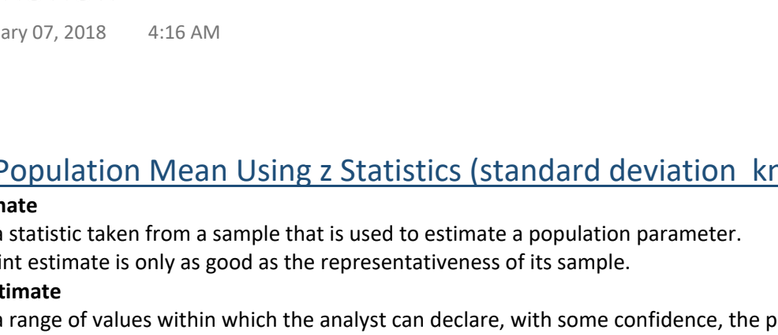
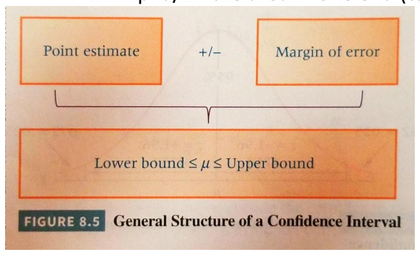
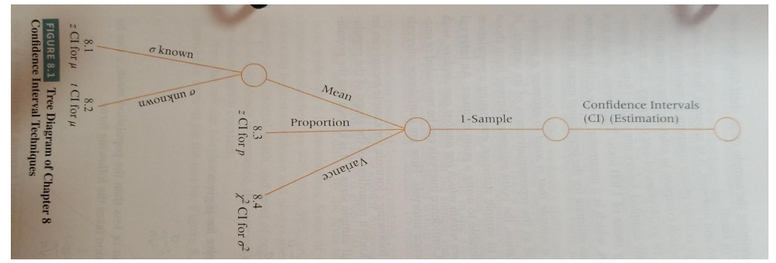
Week 5 - Review

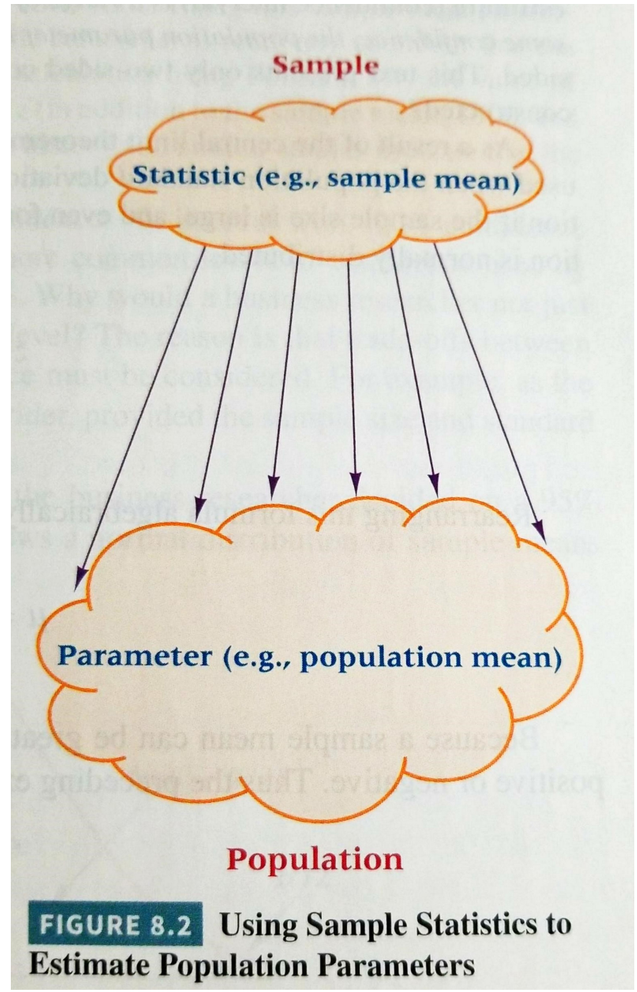
Wednesday, February 07, 2018 4:16 AM

Estimating Population Mean Using z Statistics (standard deviation known)

- **Point Estimate**
  - It is a statistic taken from a sample that is used to estimate a population parameter.
  - A point estimate is only as good as the representativeness of its sample.
- **Interval Estimate**
  - It is a range of values within which the analyst can declare, with some confidence, the population parameter lies (a.k.a. confidence interval). Confidence intervals can be two-sided or one-sided.
- **Confidence Intervals**
  - Confidence intervals provide a fundamental method for assessing how well an estimator estimates is population parameter. The probability coverage of a confidence interval refers to the probability, over multiple studies, that the resulting confidence intervals will contain the population parameter being estimated. The confidence interval captures the unknown parameter based on a certain confidence.
    - *"Every time you calculate a confidence interval or test a hypothesis, you are making an assumption regarding what the sampling distribution is for the statistic of interest. You have to make an assumption regarding that so you can make a confidence statement."*
  - ***Margin of error of the interval***
    - This is the distance between the statistics computed to estimate a parameter and the parameter. The margin of error takes into account the desired level of confidence, sample size, and standard deviation.
  - ***Upper Bound of the Confidence Interval***
    - Margin of error added to the point estimate.
  - ***Lower Bound of Confidence Interval***
    - Margin of error is subtracted from point estimate.
    - 

      - Alpha - the area under the normal curve outside of the confidence interval area
      - Alpha/2 - the area in one end (tail) of the distribution outside of the confidence interval
    - 





Estimating Population Mean Using t Statistics (standard deviation unknown)

Estimating Population Mean Using t Statistics (standard deviation unknown)

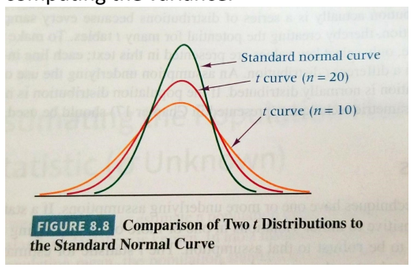
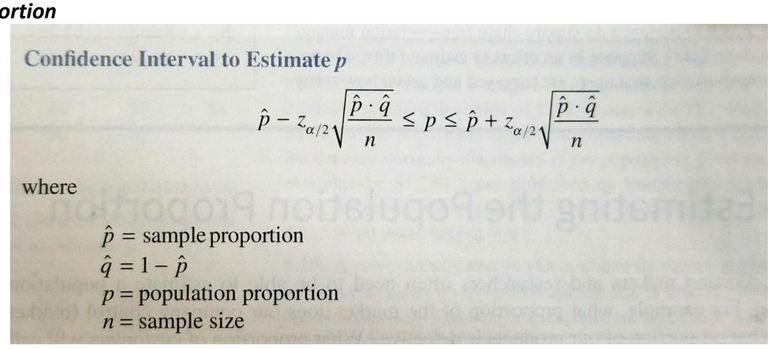
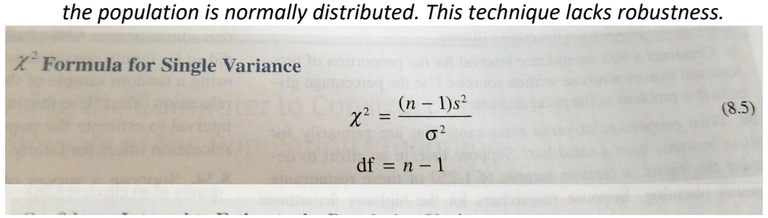
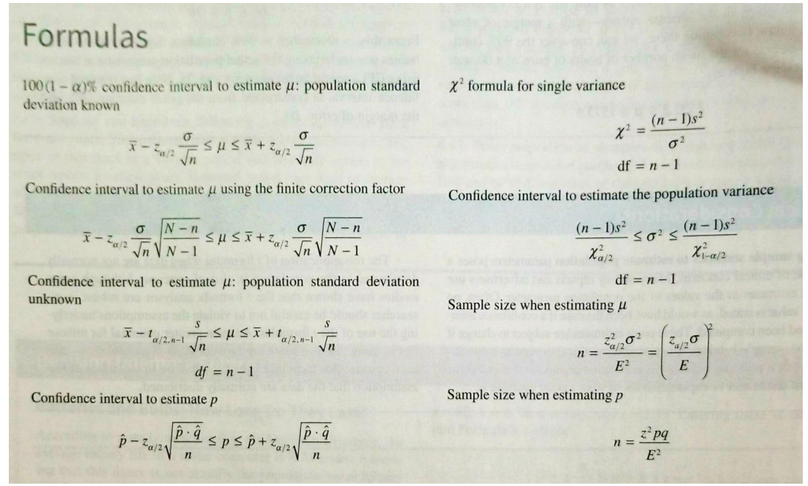
- ***t-Distribution***
  - The t distribution is a series of distribution because every sample size has a different distribution, thereby creating the potential for many t tables.
- ***Characteristics of t distributions***
  - Symmetric, unimodal, and a family of curves
  - Flatter in the middle and have more area in their tails than the standard normal distributions.
  - Values reveal that the t distribution approaches the standard normal curve as n becomes larger.
  - Are appropriate distributions to use any time the population variance and standard deviation is unknown, regardless of sample size, as long as it is known that the population of interest is normally distributed.
- ***Robustness***
  - When a statistical technique is relatively insensitive to minor violations in one or more of its underlying assumptions.
  - The t statistic for estimating a population mean is relatively robot to the assumption that the population is normally distributed.
- ***Degrees of Freedom***
  - The number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variance.
  - 

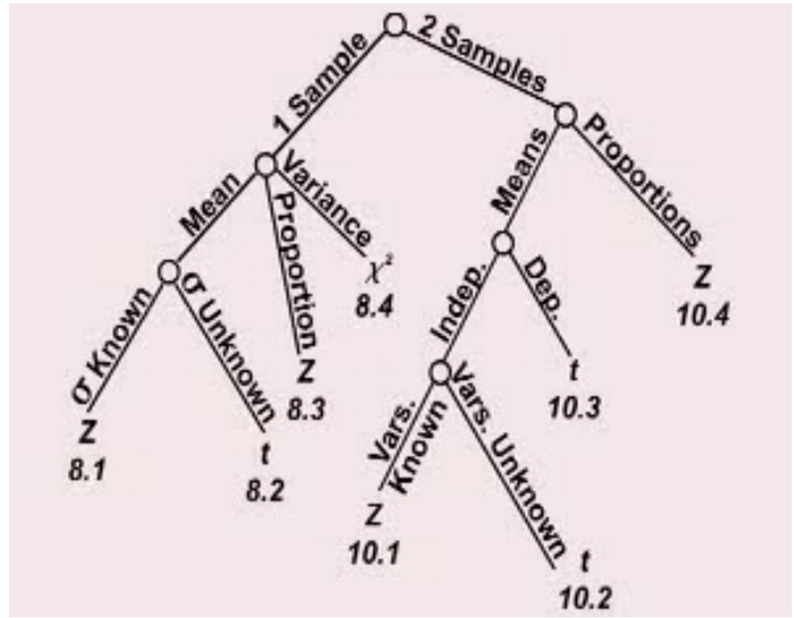
FIGURE 8.8 Comparison of Two *t* Distributions to the Standard Normal Curve

Estimating the Population Proportion and Variance

- ***Proportion***
  - 

- ***Variance***
  - Chi-square distribution
    - This is the relationship of the sample variance to the population variance
    - *Caution: the chi-squared statistic to estimate the population variance is extremely sensitive to violations of the assumption that the population is normally distributed. This technique lacks robustness.*
  - 





Type I, Type II and Sample Size

You can never be right 100% of the time.

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

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

$$\text{df} = n - 1$$

Confidence Interval to Estimate the Population Variance

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \qquad (8.6)$$

$$\text{df} = n - 1$$

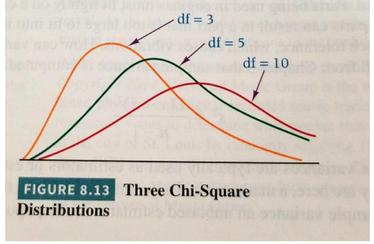


FIGURE 8.13 Three Chi-Square Distributions

- ***Sample-Size Estimation***
  - Ability to estimate the size of the sample necessary to accomplish the purposes of the study.
- ***Error of Estimation***
  - Difference between $\bar{x}$ - mu
    - Let E = ($\bar{x}$ - mu)
    - $$z = \frac{E}{\frac{\sigma}{\sqrt{n}}}$$
- **Sample Size When Estimating Mean**
  - **Sample Size When Estimating μ**

$$n = \frac{z^2_{\alpha/2}\sigma^2}{E^2} = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

- **Sample Size when Estimating Proportion**
  - **Sample Size When Estimating *P***

$$n = \frac{z^2 pq}{E^2}$$

where

p = population proportion
$q = 1 - p$
E = error of estimation
n = sample size

Student's *t*

- ***Symmetric distribution***
  - *Outliers tend to be rare*
    - Student's t performs well in terms of both probability coverage and the length of the resulting confidence intervals relative to other methods.
  - *Outliers tend to be common*

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

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

What is a suitable tradeoff of the error rates?

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.**

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?

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

Student's t performs well in terms of both probability coverage and the length of the resulting confidence intervals relative to other methods.

- *Outliers tend to be common*
  - The length of the confidence interval based on Student's t might be very high relative to other methods that might be used.

- ***Asymmetric distribution***
  - *Outliers tend to be rare*
    - A sample size of 200 or more might be needed to get a reasonably accurate confidence interval.
    - It is possible that Student's t is providing inaccurate confidence intervals unless the sample size is fairly large.
  - *Outliers tend to be common*
    - A sample size of 300 or more might be needed to get accurate probability coverage
    - The length of the confidence interval will tend to be large than that obtained by alterative techniques

Bootstrap *t* Method

- This type of method can provide more accurate confidence intervals.
- With a large enough sample size, bootstrap methods are not needed.
  - Except when dealing with the median, it is unclear how large a sample size must be for this to be the case.

Statistical Hypothesis Testing

- ***Null hypothesis***
  - Researcher tries to disprove, reject, or nullify the hypothesis
    - Ex: drug has no effect
- ***Alternative hypothesis***
  - Alternative view of the true state of nature
    - Ex: drug has a beneficial effect

Type 1, Type II, and Sample Size

- ***Type 1 Error***
  - Example: Fire alarm when there is no fire
  - Error Rate - alpha
- ***Type II Error***
  - No fire alarm when there is a fire
  - Error Rate: beta
- *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.*

P-Value to Test Hypothesis

- For 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.

$$n-1$$

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

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.