STAT 4352 - Mathematical Statistics Notes

JaimeGoB

March 3, 2021

1 Chapter 11 - Interval Estimation

Point Estimators
θ is a unknown parameter (feature of a population)

- Ex: population mean μ
- **Fixed.**

$\hat{\theta}$ is a point estimator of θ (it is a numerical value)

- Ex: sample mean $\bar{x}$
- **Varies from sample to sample.**
- No guarantee of accuracy
- Must be *supplemented by* Var(θ)
  Standard Error SE($\hat{\theta}$) measures how much $\hat{\theta}$ varies from sample to sample.
  small SE $\implies$ low variance thus a more reliable estimate of θ

Interval Estimators

Def: Interval Estimate
Provides a range of values that best describe the population.
Let L = L(x) be the Lower Limit
U = U(x) be the Upper Limit
Both L,U are Random Variables because they are functions of sample data.

Def: Confidence Level / Confidence Coefficient
Is the probability that the **interval estimate** will include population parameter θ.

- Sample means will follow the normal probability distribution for large sample sizes (n $\geq$ 30)
- For small sample forces us to use the t-distribution probability distribution(n $<$ 30)
- A confidence level of 95% implies that **95% of all samples would give an interval that includes θ, and only 5%of all samples would yield an erroneous interval.**
- The most frequently used confidence levels are 90%, 95%, and 99% with corresponding Z-scores 1.645, 1.96, 2.576.
- The higher the confidence level, the more strongly we believe that the value of the parameter lies within the interval.

Def: Confidence Interval
Gives plausible values for the parameter θ being estimated where degree of plausibility specified by a confidence level.

To construct an interval estimator of unknown parameter θ. We must find two statistics **L** and **U** such that:

$$P\{\mathbf{L} \leq \theta \leq \mathbf{U}\} = 1 - \alpha$$

- $P\{\mathbf{L} \leq \theta \leq \mathbf{U}\}$ **Coverage Probability**, in repeated sampling, what percent of samples or Confident Intervals capture true θ.
- 100(1- α) **Confidence Interval** - for unknown fixed parameter θ.
- L,U - **Lower and Upper Bounds** - RVs because they are functions of sample data. Vary from sample to sample.
- 1-α **Confidence Level** (Probability) estimate will include population parameter θ.
- α **Level of Significance** Percent chance Confidence Interval will not contain population parameter θ.

Def: Coverage Probability
$P\{\mathbf{L} \leq \theta \leq \mathbf{U}\}$ Gives what % of samples or Confidence Intervals capture true θ.

Ex: Coverage Probability = 95%
Will capture θ, 95% of the time.
Will NOT capture θ, 5% of the time.

Properties of Confidence Intervals

- Confidence Intervals are not unique.
- Desirable to have E[Length of CI] to be small.
- A one-sided 100(1-α) lower-confidence interval on θ: L = -∞ $\Longrightarrow$ $P\{L \leq \theta\} = 1-\alpha$
- A one-sided 100(1-α) upper-confidence interval on θ: $U = \infty$ $\Longrightarrow$ $P\{\theta \leq U\} = 1-\alpha$
- If L,U are both finite, then we have a two sided interval.

Correctly Interpreting Confidence Intervals
Not Correct
There is 90% probability that the true population mean is within the interval.
Correct
There is a 90% probability that <u>any given Confidence Interval from a random sample</u> will contain the true population mean.

How to Construct Confidence Interval Using Pivot Approach:
Suppose we have a random sample $X_1, X_2, ..., X_n$ from a population distribution and the parameter of interest is θ.

Given value $\alpha \in (0,1)$. We would like to construct a 1-α Confidence Interval using a Pivot Approach:

1. **Find a variable Y, that is function of the parameter θ and data x.**
2. **The distribution of newly created variable Y is free of θ.**

In many cases:

$$Y = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \text{ is a pivot and the distribution of Y is symmetric about 0.}$$

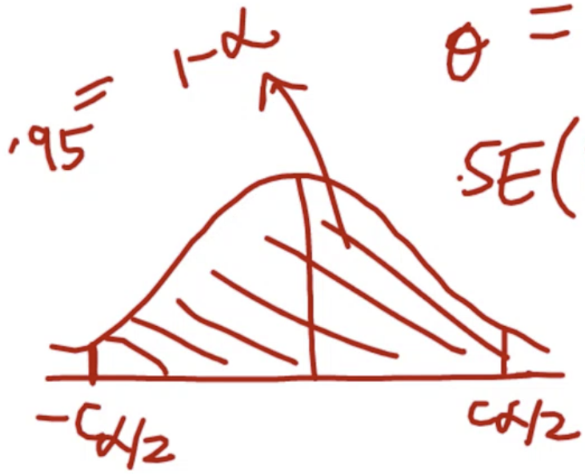
Using Pivot Approach for Two-Sided Intervals:
Find the critical points denoted $c_{\alpha/2}$ such that:

$$P\{-c_{\alpha/2} \leq Y \leq c_{\alpha/2}\} = 1 - \alpha$$

$c_{\alpha/2}$ is the upper $(\alpha\ /\ 2)$100th percentile.

Critical points- give you the area to the right of the point.

Visualizing elements from Pivot Approach:
Let μ be parameter of interest. We can construct CI using pivot approach.



$\hat{\theta} = \bar{X}$, $\theta = \mu$

$SE(\hat{\theta}) = SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

$P(-c_{\alpha/2} \leq Y \leq c_{\alpha/2}) = 1 - \alpha$

$c_{\alpha/2} = z_{.025}$

Symmetric Two-sided CI: Theorem
$\hat{\theta} \pm c_{\alpha/2}(SE(\hat{\theta}))$ is a $100(1-\alpha)\%$ confidence interval for θ

Proof:

$$1-\alpha = P\{-c_{\alpha/2} \le Y \le c_{\alpha/2}\}$$

$$= P\{-c_{\alpha/2} \le \frac{\hat{\theta}-\theta}{SE(\hat{\theta})} \le c_{\alpha/2}\}$$

$$= P\{\hat{\theta} - c_{\alpha/2}(SE(\hat{\theta})) \le Y \le \hat{\theta} + c_{\alpha/2}(SE(\hat{\theta}))\}$$

$\implies$ $\hat{\theta}$ is **within** $c_{\alpha/2}(SE(\hat{\theta}))$ of θ with **probability** 1-α

$c_{\alpha/2}(SE(\hat{\theta}))$ is known as *Margin of Error* (size of error in estimation)
Ex: In polls you might hear accurate with 0.02 (this is margin of error)

Asymmetric Two-sided CI(Non-symmetric distributions):
$[\hat{\theta} - c_{\alpha/2}(SE(\hat{\theta})), \hat{\theta} - c_{1-\alpha/2}(SE(\hat{\theta}))]$ is a $100(1-\alpha)\%$ confidence interval for θ

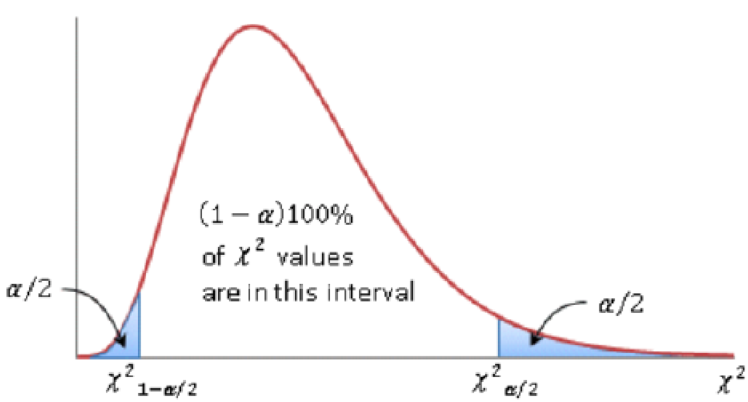
Proof:

$$1-\alpha = P\{c_{1-\alpha/2} \le Y \le c_{\alpha/2}\}$$

$$= P\{c_{1-\alpha/2} \le \frac{\hat{\theta}-\theta}{SE(\hat{\theta})} \le c_{\alpha/2}\}$$

$$= P\{\hat{\theta} - c_{\alpha/2}(SE(\hat{\theta})) \le \theta \le \hat{\theta} - c_{1-\alpha/2}(SE(\hat{\theta}))\}$$

Ex: Chi-Square distribution critical points

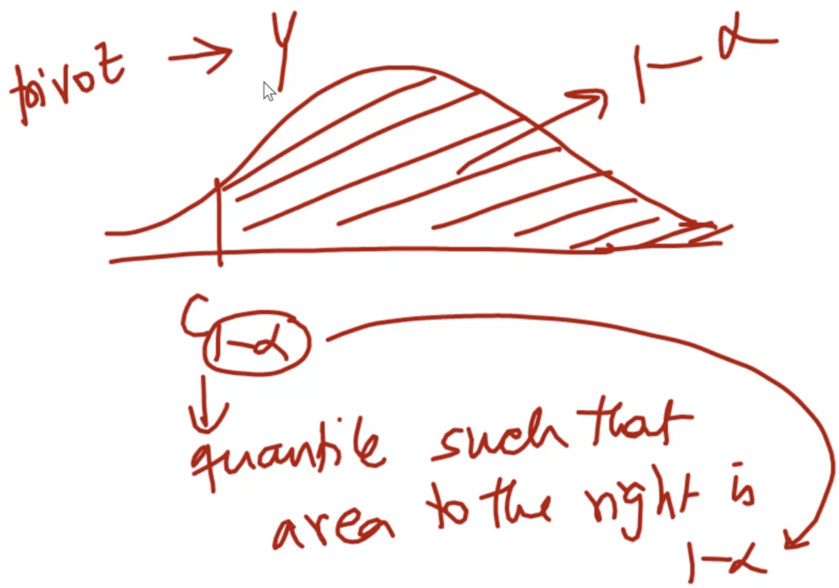


One-sided Confidence Bound:
A one-sided confidence bound defines the point where a certain percentage of the population is either higher or lower than the defined point.

Upper Bound: U = $\hat{\theta} - c_{1-\alpha}(SE(\hat{\theta}))$ when L = -∞

Lower Bound: L = $\hat{\theta} - c_{\alpha}(SE(\hat{\theta}))$ when U = ∞

Proof(Upper Bound):
Coverage probability is 1 - α.



$$1-\alpha = P\{Y \geq c_{1-\alpha}\}$$

$$= P\{\frac{\hat{\theta}-\theta}{SE(\hat{\theta})} \geq c_{1-\alpha}\}$$

$$= P\{\theta \leq \hat{\theta} - c_{1-\alpha}(SE(\hat{\theta}))\}$$

$\implies$ U= $\hat{\theta} - c_{1-\alpha}(SE(\hat{\theta}))$
The lower bound can be computed in the same manner.

How to interpret a one-sided CI?
For **Lower Bound** critical region $\in [c_\alpha, \infty]$:We are sure parameter θ is below c_α

For **Upper Bound** critical region $\in [-\infty, c_{1-\alpha}]$:We are sure parameter θ is above $c_{1-\alpha}$

Choice of Sample Size for CI for Mean when Variance Known

General form of a CI:
100(1-α)% CI for θ: $\hat{\theta} \pm c_{\alpha/2} SE(\hat{\theta})$

Mean μ of a Normal population: $\bar{x} \pm Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Margin Of Error: MOE $= Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Width of CI: $2 \times$ MOE $= 2 \cdot Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

Properties of MOE:

- **As (1-α) increases MOE increases:**
  The larger the CI, the larger critical points are needed.
  As critical points increase, Margin Of Error: MOE $= Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ increase
  $\Longrightarrow$ Increase in CI will make width of CI wider
- **As σ increases MOE increases:**

  Margin Of Error: MOE $= Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ σ is part of numerator, if it increases it will make MOE increase too.
- **As n increases MOE decreases:**

  Margin Of Error: MOE $= Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ n is part of denominator, if it increases it will make MOE decrease.

Getting n from predefined MOE:
Let E_0 be a pre-specified MOE. We can a value for n to make the following equation true.

$\frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \leq E_0$ solving for n we get: $n \geq \left(\frac{Z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right)^2$ (round up to nearest n.)

We do this when we want to know how many observations (n) will give pre-specified margin of error - E_0

Theorem 11.1: Confidence Interval on the Mean of a Normal Distribution with known Variance

Let X be normal random variable with:

Unknown mean μ

Known variance σ^2

Suppose a random sample n, $(X_1, X_2, ..., X_n)$ is taken.

A 100(1-α)% confidence interval on μ can be obtained by considering sampling distribution of the sample mean $\bar{X}$.

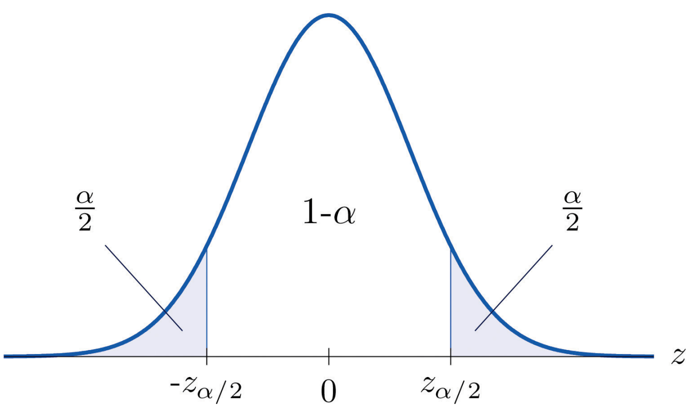
Central Limit Theorem:

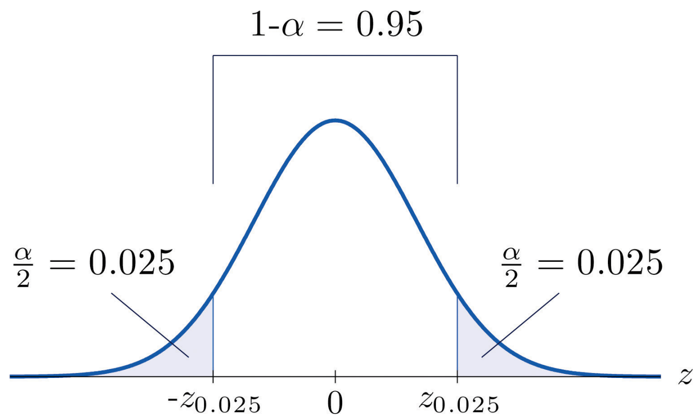
$E(\bar{X}) = \mu$ and $SD(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$, so $\bar{X} \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n})$ as n$\to \infty$

Let Z = Standardizing $\bar{X}$, Z will follow a Standard Normal Distribution

Let Z $= \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$

We can see from the image to the left: **Distribution of Z** and the image to the right: **Confidence Interval of 95%**





We can see that:

$$P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\} = 1 - \alpha$$

substituting Z into equation:

$$P\{-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}\} = 1 - \alpha$$

isolating μ:

$$P\{\bar{X} - Z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{X} + Z_{\alpha/2}(\sigma/\sqrt{n})\} = 1 - \alpha$$

Conclusion $\bar{X} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$ is a 100(1-α)% CI for μ

Confidence Interval on the Mean of a Normal Distribution
Variance Unknown and/or Small Sample:

$\bar{X} \pm t_{n-1,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ is a $100(1-\alpha)\%$ CI for μ

Proof:
We know that T $= \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

$$1-\alpha = P\{-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\}$$

$$= P\{-t_{n-1,\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2}\}$$

$$= P\{\bar{X} - t_{n-1,\alpha/2}(S\sqrt{n}) \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}(S\sqrt{n})\}$$

Confidence Interval on the Mean no Specific Distribution
Variance Known

$\bar{X} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$ is a 100(1-α)% CI for μ

Proof:
Assuming the sample size is large (n $\geq$ 30) then by CLT:

$$\text{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

The mean of any distribution **provided** n is large (n $\geq$ 30) can be approximated using a Normal Distribution.

Confidence Interval on the Mean no Specific Distribution
Variance Unknown

$\bar{X} \pm Z_{\alpha/2}(S/\sqrt{n})$ is a 100(1-α)% CI for μ

Proof:
Given the fact that S^2 is and unbiased estimator of σ^2 we can use sample variance in lieu of population variance. Also sample size is large (n $\geq$ 30) then by CLT and LLN:

$$\text{Z} = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\left(\frac{\sigma}{S}\right) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Confidence Interval on the Proportion of a Binomial Distribution

$\hat{p} \pm Z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ is a 100(1-α)% CI for p

n is random sample of size n has been taken from a large population and X($\leq n$) observations in this sample belong to a class of interest.
$\hat{p} = X/n$ is the point estimator of the proportion of the population that belongs to this class.
n and p are the parameters of a binomial distribution.

Proof:
The sampling distribution of $\hat{p}$ is approximately normal with mean p and variance p(1-p)/n, if p is not too close to 0 or 1 and n is large.

$$\text{Z} = \frac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim N(0,1)$$

To construct CI on p, note that:

$$1-\alpha \approx P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\}$$

$$\approx P\{-Z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \leq Z_{\alpha/2}\}$$

$$\approx P\{\hat{p} - Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\}$$

Since the square root is the SE of estimator $\hat{p}$ and
also contains p in lower and upper bound.
We can replace p with $\hat{p}$ and use Estimated SE instead of SE.

$$\approx P\{\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\}$$

Confidence Interval on the Variance or Standard Deviation of a Normal Distribution - Mean is Unknown

$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$ is a 100(1-α)% CI for σ^2

Proof:
According to theorem 8.11:

$$\text{Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

The critical points are: $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$

$$1 - \alpha = P\{\chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1,\alpha/2}\}$$
$$= P\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

Confidence Interval on the Variance or Standard Deviation of a Normal Distribution - Mean is Known

$\left[\frac{(n)S^2}{\chi^2_{n,\alpha/2}}, \frac{(n)S^2}{\chi^2_{n,1-\alpha/2}}\right]$ is a 100(1-α)% CI for σ^2

Proof:
Since μ is known then:
Sum of n, squared standard normal distributions
$\implies$ Sum of n Chi-Square distributions with one df
$\implies$ χ^2_n

$$\text{Y} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2} = \left(\frac{x_1 - \mu}{\sigma}\right)^2 + ... + \left(\frac{x_n - \mu}{\sigma}\right)^2 \sim \chi^2_n$$

The critical points are: $\chi^2_{n,1-\alpha/2}$ and $\chi^2_{n,\alpha/2}$

$$1 - \alpha = P\{\chi^2_{n,1-\alpha/2} \leq \frac{(n)S^2}{\sigma^2} \leq \chi^2_{n,\alpha/2}\}$$
$$= P\left[\frac{(n)S^2}{\chi^2_{n,\alpha/2}} \leq \sigma^2 \leq \frac{(n)S^2}{\chi^2_{n,1-\alpha/2}}\right]$$

Two-Sample Confidence Interval Estimation
Confidence Interval on the Difference between Means of Two Normal Distributions, Variances Known
In this case both means are unknown but variances are known.

$$\left[\bar{X}_1 - \bar{X}_2 \pm \left(Z_{\alpha/2}\right)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \text{ is a 100(1-}\alpha\text{)\% CI for } \mu_1 - \mu_2$$

Proof:
Let X_1 and X_2 be two normally distributed independent random variables.
$X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$
So, $\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \implies \text{SE}(\bar{X}_1 - \bar{X}_2) = \text{SD}(\bar{X}_1 - \bar{X}_2) = \sqrt{Var(\bar{X}_1 - \bar{X}_2)}$

Because Var(A - B) = Var(A) - Var(B) when A,B are independent

$$\text{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Independent random samples from normal populations:

$$\text{Z} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Now, to construct a CI:
$1 - \alpha = P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\}$

$$= P\{-Z_{\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq Z_{\alpha/2}\}$$

$$= P\left[(-Z_{\alpha/2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \leq (Z_{\alpha/2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

$$= P\left[\bar{X}_1 - \bar{X}_2 - \left(Z_{\alpha/2}\right)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + \left(Z_{\alpha/2}\right)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Confidence Interval on the Difference between Means of Two Normal Distributions, Variances Unknown or Small Samples

Both means and variances are unknown. However, we can assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\left[\bar{X}_1 - \bar{X}_2 \pm \left(t_{n_1+n_2-2,\alpha/2}\right) S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \text{ is a 100(1-}\alpha\text{)\% CI for } \mu_1 - \mu_2$$

Proof:

Let S_1^2 and S_2^2 be sample variances of random variables X_1 and X_2. Since both sample variances are estimates of common variance σ^2 we can obtain a "pooled estimator" of σ^2.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_2+n_1-2} \sim \chi^2_{n_1-1} + \chi^2_{n_2-1}$$

Now pivot T

$$\text{T} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Now, to construct a CI:

$1 - \alpha = P\{-t_{n_1+n_2-2,\alpha/2} \le T \le t_{n_1+n_2-2,\alpha/2}\}$

$$= P\{-t_{n_1+n_2-2,\alpha/2} \le \frac{(\bar{X}_1 - \bar{X}_)) - (\mu_1 - \mu_2)}{S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{n_1+n_2-2,\alpha/2}\}$$

$$P\left[-t_{n_1+n_2-2,\alpha/2}\left(S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) \le (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \le t_{n_1+n_2-2,\alpha/2}\left(S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)\right]$$

After solving for both population means:

$$P\left[\bar{X}_1 - \bar{X}_2 - (t_{n_1+n_2-2,\alpha/2})S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + (t_{n_1+n_2-2,\alpha/2})S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

2