**Lecture 2. Statistical Inference:**

**Confidence Interval**

**Scenario 1: Review of Confidence Interval for One Population Mean when the Population is Normal and the Population Variance is Known**

**Motivation & simple random sample**

Eg) We wish to estimate the average height of adult US males

* Take a random sample.
* “Simple” random sample: every subject in the population has the same chance to be selected.

**Introduction to statistical inference on one population mean**

For a “**random** sample” of size n: 

**<i> Point estimatior  → sample mean ()**

Other estimators: median, mode, trimmed mean, …

**<ii> Confidence Interval (C.I.)**

Eg) 95% C.I. for μ

99.9999% C.I. (‘6-9’ in the manufacture industry)

**<iii> Hypothesis Test**

Eg) : μ 

: μ 

Point Estimator, C.I., Test  Statistical Inference

* Draw some conclusion on the population (parameters of interest) based on a random sample.

1. **The Exact Confidence Interval for** μ **when the population is normal & is known**
2. **Point estimator and confidence interval for** 

* When the population is normal and the population variance is known.
* Let  be a random sample for a normal population with mean  and variance . That is, .
* For now, we assume that  is known.

**<i> Point Estimator for  : **

 is an unbiased estimator of 

* Intuitively, this means if you take “many” samples of size n from the population, then the mean of these samples means would be equal to  if you take a large enough # of samples.
*  is also a maximum likelihood estimator (MLE) of .
*  is also a method of moment estimator (MOME) of .
* Other good properties too.

**<ii> Confidence Interval for **

- Intuitive approach (backwards derivation for the CI boundaries :

*Since we know*

We can compute the expressions for .

However, one question is that there are MANY ways to choose the C’s.

Later you will see that for pivotal quantity with symmetric pdfs, the symmetric CIs are the optimal – in that they have the shortest lengths for the given confidence level 100(1-α)%.

**Now we present a general approach to derive the CI’s.**

**General approach for deriving CI’s : the Pivotal Quantity (P.Q.) approach**

\*Definition: A pivotal quantity is a function of the sample and the parameter of interest. Furthermore, its distribution is entirely known.

1. We start by looking at the point estimator of . ****

\* Is **** a pivotal quantity for ?

→ **** is not because is unknown.

\* function of  and  : ****

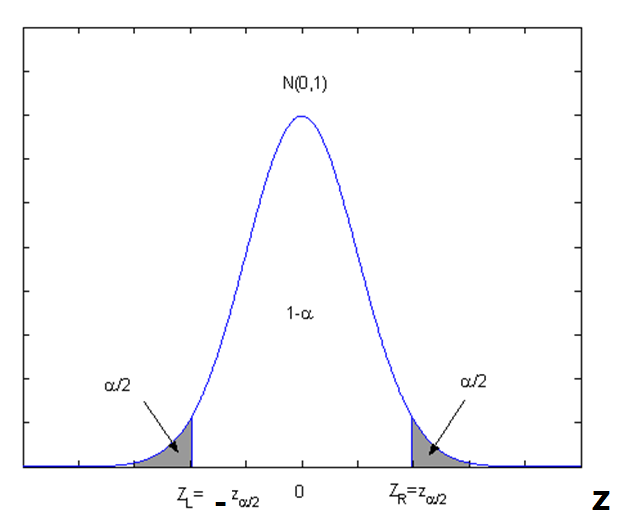
→ Yes, it is pivotal quantity.

\* Another function of  and  : ****

→ Yes, it is pivotal quantity.

So, Pivotal Quantity is not unique.

2. Now that we have found the pivotal quantity Z, we shall start the derivation for the symmetrical CI’s for µ from the PDF of the pivotal quantity Z



100(1-α)% CI for μ, 0<α<1

(e.g. α=0.05 ⇒ 95% C.I.)

****

****

****

****

****

****

∴ the 100(1-α)% C.I. for μ is ****

**\*Note, some special values for and the corresponding values are:**

**1. The 95% CI, where and the corresponding**

**2. The 90% CI, where and the corresponding**

**3. The 99% CI, where and the corresponding**

**Example 0.** A random sample of 400 adult US male was taken and the sample mean was found to be . Based on past studies, it is believed that the population distribution of all adult US male is normal and the standard deviation is 30 inches. Please construct a 95% confidence interval for the average height of all adult US male based on this sample.

**Solution:**  The 95% CI for is

That is, the estimated 95% confidence interval for the average height of all adult US male is .

… …

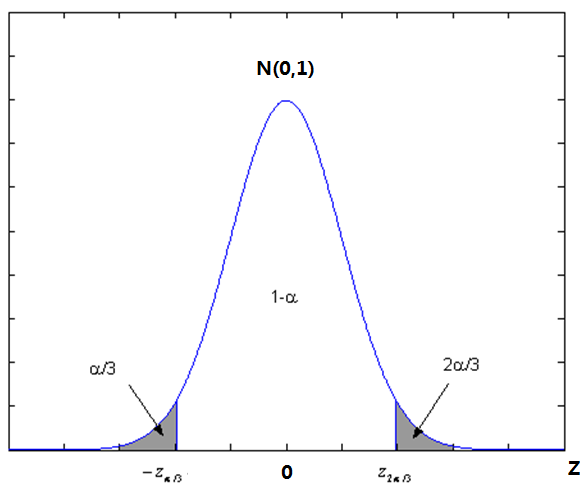
This means that we are 95% sure the population mean μ would lie between .

∴Recall the 100(1-α)% symmetric C.I. for μ is ****

**\*Please note that this CI is symmetric around**

**The length of this CI is: **

Now we derive a non-symmetrical CI:



****

100(1-α)% C.I. for μ

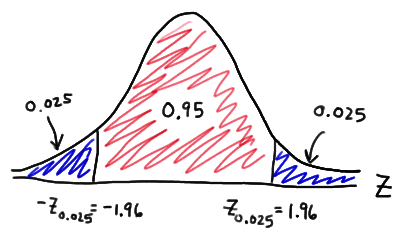
⇒ ****

Compare the lengths of the C.I.’s, one can prove theoretically that: ****

**You can try a few numerical values for α, and see for yourself. For example,**

**Example 1 (review of exact CI for mean when the population is normal and the population variance is known.).** A random sample of 126 police officers subjected to constant inhalation of automobile exhaust fumes in downtown Cairo had an average blood lead level concentration of 29.2 *μg*/*dl*. Assume *X*, the blood lead level of a randomly selected policeman, is normally distributed with a standard deviation of *σ* = 7.5 *μg*/*dl*. Historically, it is known that the average blood lead level concentration of humans with no exposure to automobile exhaust is 18.2 *μg*/*dl*. Is there convincing evidence that policemen exposed to constant auto exhaust have elevated blood lead level concentrations? (Data source: Kamal, Eldamaty, and Faris, "Blood lead level of Cairo traffic policemen," Science of the Total Environment, 105(1991): 165-170.)

**Solution.**  Let's try to answer the question by calculating a 95% confidence interval for the population mean. For a 95% confidence interval, 1−*α* = 0.95, so that *α* = 0.05 and *α*/2 = 0.025. Therefore, as the following diagram illustrates the situation, *z*0.025 = 1.96:



Now, substituting in what we know (= 29.2, *n* = 126, *σ* = 7.5, and *z*0.025 = 1.96) into the the formula for a *Z*-interval for a mean, we get:

Simplifying, we get a 95% confidence interval for the mean blood lead level concentration of all policemen exposed to constant auto exhaust: [27.89, 30.51]

That is, we can be 95% confident that the mean blood lead level concentration of all policemen exposed to constant auto exhaust is between 27.9 *μg*/*dl*and 30.5 *μg*/*dl*. Note that the interval does not contain the value 18.2, the average blood lead level concentration of humans with no exposure to automobile exhaust. In fact, all of the values in the confidence interval are much greater than 18.2. Therefore, there is convincing evidence that policemen exposed to constant auto exhaust have elevated blood lead level concentrations.

**Scenario 2 : normal population,**  **unknown**

1. Point estimation : 
2. 
3. **Theorem.** Sampling from normal population
4. 
5. 
6.  and  are independent.

**Definition.** 

------ **Derivation of CI, normal population,  is unknown** ------

 is not a pivotal quantity.

 is not a pivotal quantity.

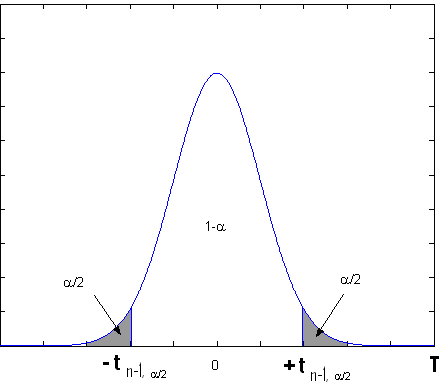
 is not a pivotal quantity.

Remove !!!

Therefore  is a pivotal quantity.

Now we will use this pivotal quantity to derive the 100(1-α)% confidence interval for μ.

We start by plotting the pdf of the t-distribution with n-1 degrees of freedom as follows:



The above pdf plot corresponds to the following probability statement:



=> 

=> 

=> 

=> 

=> 

=> Thus the  C.I. for  when  is unknown is

.

(\*Please note that )

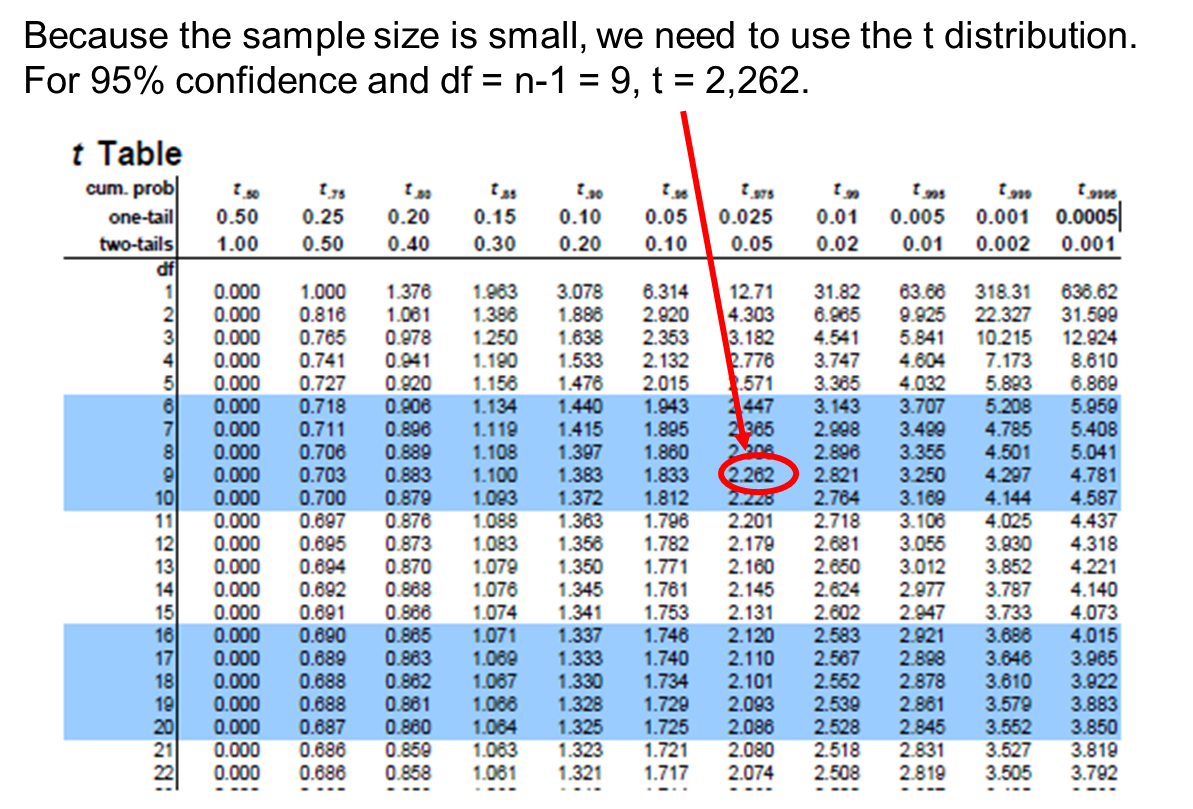
**Example 2. (Small sample CI for population mean, variance unknown, NORMAL POPULATION) (BU)**

The table below shows data on a subsample of n=10 participants in the 7th examination of the Framingham offspring Study.

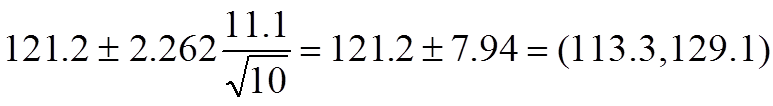
|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristic** | **n** | **Sample Mean** | **Standard Deviation (s)** |
| **Systolic Blood Pressure** | 10 | 121.2 | 11.1 |
| **Diastolic Blood Pressure** | 10 | 71.3 | 7.2 |
| **Total Serum Cholesterol** | 10 | 202.3 | 37.7 |
| **Weight** | 10 | 176.0 | 33.0 |
| **Height** | 10 | 67.175 | 4.205 |
| **Body Mass Index** | 10 | 27.26 | 3.10 |

 Suppose we compute a 95% confidence interval for the true systolic blood pressure using data in the subsample. Because the sample size is small, we must now use the confidence interval formula that involves t rather than Z.

The sample size is n=10, the degrees of freedom (df) = n-1 = 9. The t value for 95% confidence with df = 9 is = 2.262.



Substituting the sample statistics and the t value for 95% confidence, we have

.

Interpretation: Based on this sample of size n=10, our best estimate of the true mean systolic blood pressure in the population is 121.2. Based on this sample, we are 95% confident that the true systolic blood pressure in the population is between 113.3 and 129.1. Note that the margin of error is larger here primarily due to the small sample size.

**Example 3. (Small sample CI for population mean, variance unknown, NORMAL POPULATION)** In a psychological depth-perception test, a random sample of  airline pilots were asked to judge the distance between 2 markers at the other end of a laboratory.

The data (in test) are

2.7, 2.4, 1.9, 2.4, 1.9, 2.3, 2.2, 2.5, 2.3, 1.8, 2.5, 2.0, 2.2, 2.6

Please construct a 95% CI for , the average distance.

**Solution.**

*(Note: we can perform the Shapiro-Wilk test to examine whether the sample comes from a normal population or not. This test is not required in our class. Here we simply assume the population is normal. I will always give you such information in the exams.)*

CI for , small sample, normal population, population variance unknown.



95% CI for  is 



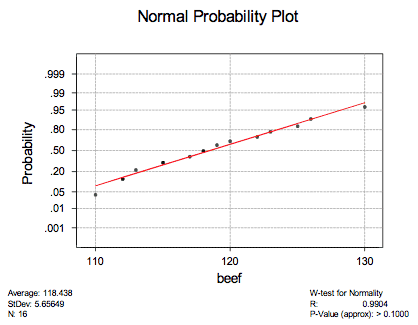
**Example 4. (Small sample CI for population mean, variance unknown, NORMAL POPULATION) (PSU)**

A random sample of 16 Americans yielded the following data on the number of pounds of beef consumed per year:

118    115    125    110    112    130    117    112  
115    120    113    118    119    122    123    126

What is the average number of pounds of beef consumed each year per person in the United States?

**Solution.** To help answer the question, we'll calculate a 95% confidence interval for the mean. As the above theorem states, in order for the *t*-interval for the mean to be appropriate, the data must follow a normal distribution. We can use a normal probability plot to provide evidence that the data are (sufficiently) normally distributed:



That is, because the data points fall at least approximately on a straight line, there's no reason to conclude that the data are not normally distributed. That's convoluted statistician talk for "we're good to go." Now, punching the *n* = 16 data points into a calculator (or statistical software), we can easily determine that the sample mean is 118.44 and the sample standard deviation is 5.66. For a 95% confidence interval with *n* = 16 data points, we need:

Now, we have all of the necessary elements to calculate the 95% confidence interval for the mean.  It is:

Simplifying, we get: 118.44±3.016  or: (115.42, 121.46)

That is, we can 95% confident that the average amount of beef consumed each year per person in the United States is between 115.42 and 121.46 pounds. Wow, that's a lot of beef!

**Scenario 3: Large Sample Confidence Interval**

**For a population mean (\*any population) and**

**For a population proportion p (\*Bernoulli population)**

**-- By the Central Limit Theorem (CLT)**

**The Central Limit Theorem (CLT):**

**By the CLT, when n is large enough, we have:**

That means Z follows approximately the normal (0,1) distribution when the sample size n is large.

**Application #1. Inference on  when the population distribution is unknown but the sample size is large.**

**(1) By the CLT, we know the following variable Z is a pivotal quantity (P.Q.) for when , for ANY population (\*not necessarily normal) as long as the sample size is large enough. In this case, the sample is considered large enough when the sample size is 30 or more (n ≥ 30).**

**P.Q.1** (when : 

Using PQ1, we can derive an **approximate** 100(1- )% large sample CI for as we have done before:

**(2) By Slutsky’s Theorem We can also obtain another pivotal quantity when is unknown by plugging in the sample standard deviation S as follows:**

**P.Q.2** (when : 

We subsequently obtain the  C.I. using the second P.Q. for :

**Application #2. Inference on one population proportion p when the population is Bernoulli(p) \*\*\***

**Setting: Let , be a random sample taken from a Bernoullis population with parameter p (\*here p is referred to as the population proportion). Please find the 100(1-α)% CI for p.**

We noticed that for the Bernoulli population, the population mean is indeed also the population proportion:

The population variance can be expressed in the parameter p as:

*Thus by the CLT we have:*



Furthermore, we notice that the sample mean is also the sample proportion:  (ex. , ). That is,

1. Therefore we obtain the following pivotal quantity Z for p:

**P.Q. 1:**

1. By Slustky’s theorem, we can replace the population proportion in the denominator with the sample proportion and obtain another pivotal quantity for p:

**\*\*\*Please note that we can construct CI for p using either PQ. However, using PQ2 will be easier.**

**# Thus the (approximate, or large sample) C.I. for p based on the second pivotal quantity**  is:



=> The large sample C.I. for p is .

# In CLT => The sample is large usually means . However, for the Bernoulli population, we have a more accurate estimation and found that large sample = the total number of successes and the total number of failures both ≥ 5:

# special case for the inference on p based on a Bernoulli population. The sample size n is large means

Let , large sample means:  (\*Here X= total # of ‘S’), and

 (\*Here n-X= total # of ‘F’)

**Example 5. (Large sample CI for population mean, variance unknown, CLT)** In a random sample of  parochial schools throughout the south, the average number of pupils per school is 379.2 with a standard deviation of 124. Use the sample to construct a 95% CI for , the mean number of pupils per school for all parochial schools in the south.

**Solution.** CI for , large sample



95% CI for  is



Example 6. (Large sample CI for population mean, variance unknown) (BU)

Descriptive statistics on variables measured in a sample of a n=3,539 participants attending the 7th examination of the offspring in the Framingham Heart Study are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristic** | **n** | **Sample Mean** | **Standard Deviation (s)** |
| **Systolic Blood Pressure** | **3,534** | **127.3** | **19.0** |
| **Diastolic Blood Pressure** | **3,532** | **74.0** | **9.9** |
| **Total Serum Cholesterol** | **3,310** | **200.3** | **36.8** |
| **Weight** | **3,506** | **174.4** | **38.7** |
| **Height** | **3,326** | **65.957** | **3.749** |
| **Body Mass Index** | **3,326** | **28.15** | **5.32** |

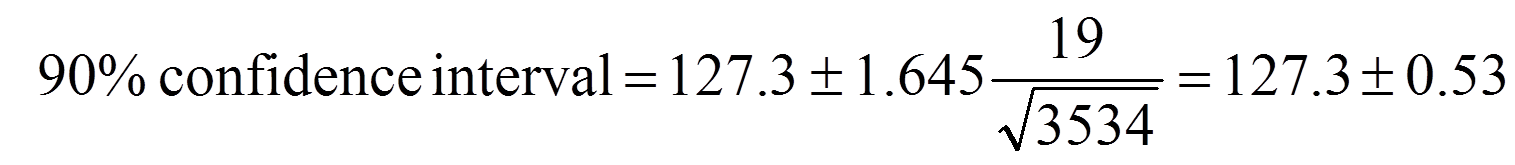
 Because the sample is large, we can generate a 95% confidence interval for systolic blood pressure using the following formula:

The Z value for 95% confidence is .

Substituting the sample statistics and the Z value for 95% confidence, we have

Therefore, the point estimate for the true mean systolic blood pressure in the population is 127.3, and we are 95% confident that the true mean is between 126.7 and 127.9. The margin of error is very small (the confidence interval is narrow), because the sample size is large.

The 90% confidence interval for mean systolic blood pressure would be:

.

That is, [126.77, 127.83].

**Example 7. (Large sample confidence interval for one population proportion)**

During one of the “beer wars” in the early 1980’s, a taste test between Schlitz and Budweiser was the focus of a TV commercial. 100 people agreed to drink 2 unmarked mugs and indicate which of the two beers they liked better. 54 chose “Bud”. Construct and interpret the corresponding 95% confidence interval for  - the proportion of beer drinkers who prefer Bud to Schlitz.

**Solution:**

**Confidence Interval for one population proportion (p) when the sample size is large**

Sample size :  

Sample proportion :  ()

\*\*\* Recall we usually denote 

***“sample is large”*** means

* For one population mean, 
* For one population proportion : 

Here we have, therefore our sample is considered large enough to use the CLT.



From 95% confidence interval, 







The 95% confidence interval for  is 

Therefore, based on the given study, the percentage of people who prefer Bud versus Schlitz could be either below 0.5 or larger than 0.5. It is inconclusive based on the given confidence interval to decide which beer is more liked.

However, if we increase our sample size to 10,000, and if still 54% of the people preferred Bud over Schlitz, we can then make a more decisive conclusion based on the CI as the following.

If ,





The 95% confidence interval for  is . Thus we are at least 95% sure is over 50% - Bud wins.