Introduction to Applied Statistics
STAT 5005
Chapter 5: Inferences about Population Central Values

Jingyu Sun

Fall 2021

Introduction

Outline

Goal: Present basic ideas involved in statistical inference. This chapter focuses on inferences about population means

- Estimation of μ
- Choosing the Sample Size for Estimating μ
- A Statistical Test for μ
- The Level of Significance of a Statistical Test
- Inferences about μ for a Normal Population, σ Unknown

- Methods for making inferences about parameters fall into one of two categories.
  - Either we will estimate the value of the population parameter of interest
  - or we will test a hypothesis about the value of the parameter
- In estimating a population parameter, we are answering the question, "What is the value of the population parameter?"
- In testing a hypothesis, we are seeking an answer to the question, "Does the population parameter satisfy a specified condition?"
- Example: Consider a study in which an investigator wishes to examine the effectiveness of a drug product in reducing anxiety levels of anxious patients. What is a question related to (1) estimation? (2) statistical test?

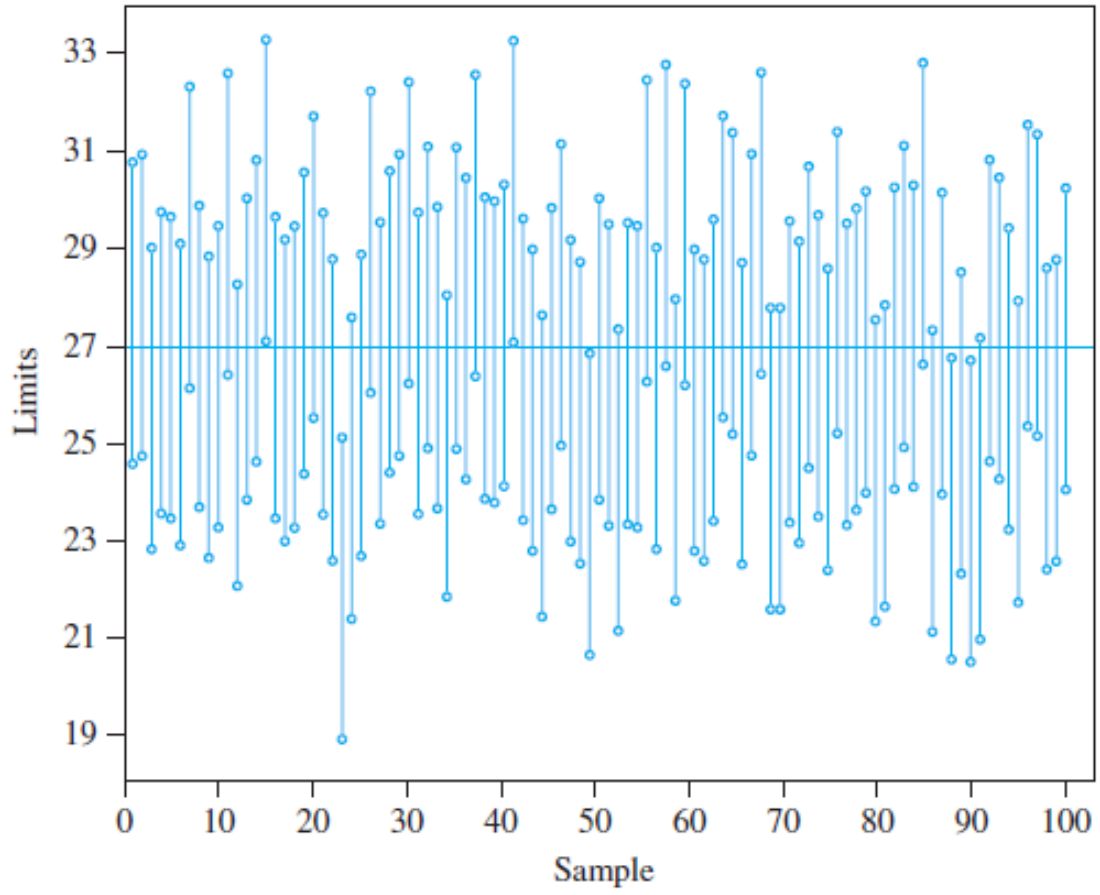
Estimation of μ

Point Estimation VS Confidence Intervals

- **Point estimation** consists in computing a single value (statistic) from the sample data to estimate a population parameter
- We may observe that the sample statistic may not be very close to the population parameter it is supposed to estimate
- We may consider an interval of possible values for the parameter, called **confidence intervals** in place of using just a single value
- In this chapter, we deal with point and interval estimation of a population mean μ

- We evaluate the goodness of an interval estimation procedure by examining the fraction of times in repeated sampling that interval estimate would encompass the parameter to be estimated
- This fraction is called the **confidence coefficient**
- For large sample size n, what is an interval estimate of μ with level of confidence 0.95?

- Consider a normal distributed population with a mean $\mu = 27$ and a standard deviation $\sigma = 10$
- We draw one hundred samples of size $n = 40$ from the population
- From each of these samples we compute the interval estimate $\bar{y} \pm 1.96(10/\sqrt{40})$



- 6 of the 100 intervals failed to capture the population mean
- Because our level of confidence is 95%, we would expect that, in a large collection of 95% confidence intervals, approximately 5% of the intervals would fail to include μ

- In most situations when the population mean is unknown, the population standard deviation σ will also be unknown
- For all practical purposes, if the sample size is relatively large (30 or more is the standard rule of thumb), we can estimate the population standard deviation σ with the sample standard deviation s in the confidence interval formula
- This estimation introduces another source of random error but the formula is still a very good approximation for large sample sizes

Confidence Interval for μ When σ is Known

- For a specified value of $(1-\alpha)$, where α is between 0 and 1, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{y} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$$

- The quantity $z_{\alpha/2}$ is a value of the standard normal distribution having a tail area of $\alpha/2$ to its right
- Here we assume that σ is known or that the sample size is large enough to replace σ with s
- For interval estimation, the width of the confidence interval and the confidence coefficient measure the goodness of the inference
- For a given value of the confidence coefficient, the smaller the width of the interval, the more precise the inference

Example

- A courier company in New York City claims that its mean delivery time to any place in the city is less than 3 hours
- The consumer protection agency decides to conduct a study to see if this claim is true
- The agency randomly selects 50 deliveries and determines the mean delivery time to be 2.8 hours with a standard deviation of 0.6 hours
- The agency wants to estimate the mean delivery time using a 95% confidence interval. Obtain this interval and then decide if the courier company's claim appears to be reasonable.
- Construct a 99% confidence interval for the mean delivery time

Choosing the Sample Size for Estimating μ

Choosing the Sample Size for Estimating μ

- Data collection costs money. If the sample is too large, time and talent are wasted
- Conversely, it is wasteful if the sample is too small, because inadequate information has been purchased for the time and effort expended
- There are two considerations in determining the appropriate sample size for estimating μ using a confidence interval
  - the tolerable error establishes the desired width of the interval
  - the level of confidence

- Suppose we want to estimate μ using a $100(1-\alpha)\%$ confidence interval having tolerable error W
- Our interval will be of the form $\bar{y} \pm E$, where $E = W/2$
- The formula for the sample size n is then

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

- The formula requires knowledge of the population variance σ^2
- We can obtain an approximate sample size by estimating σ^2, using one of these two methods:
  1. Employ information from a prior experiment to calculate a sample variance s^2
  2. Use information on the range of the observations

Example

- University officials have started to include the average amount expended on textbooks into their estimated yearly expenses for students
- In order for these estimates to be useful, the estimated cost should be within \$25 of the mean expenditure for all undergraduate students at the university
- From data collected in previous years, the university officials have determined that the annual expenditure for textbooks has a histogram that is normal in shape with costs ranging from \$250 to \$750
- How many students should the university sample in order to be 95% confident that their estimated cost of textbooks will satisfy the stated level of accuracy?

Example

- A federal agency has decided to investigate the advertised weight printed on cartons of a certain brand of cereal
- A summary of 1,500 of the weights made available to the agency indicates a mean weight of 11.80 ounces per carton and a standard deviation of .75 ounce
- Use this information to determine the number of cereal cartons the federal agency must examine to estimate the average weight of cartons being produced now, using a 99% confidence interval of width .50

A Statistical Test for μ

Example

- A consumer protection group is concerned that a soda manufacturer is filling its 591 mL bottles with less than 591 mL of soda
- The group purchases 49 bottles of this soda, measure the contents of each, and finds that the mean amount is 589 mL, and the standard deviation is equal to 2.02 mL.
- Do the data provide sufficient evidence for the consumer group to conclude that the mean fill per bottle is less than 591 mL?

- A **statistical test** is based on the concept of proof by contradiction and is composed of the five parts listed here

    1. The alternative hypothesis is proposed by the person conducting the study, denoted by H_a
    2. Null hypothesis, denoted by H_0, is the negation of H_a
    3. Test statistics
    4. Rejection region
    5. Check assumptions and draw conclusions.

- In the soda example, determine H_0 and H_a

Determining the Null and the Alternative Hypotheses

- The statement that μ equals a specific value will always be included in H_0. The particular value specified for μ is called its null value and is denoted μ_0
- The statement about μ that the researcher is attempting to support or detect with the data from the study is the alternative hypothesis, H_a
- The negation of H_a is the null hypothesis, H_0
- The null hypothesis is presumed correct unless there is overwhelming evidence in the data that H_a is supported

Test Statistic and Rejection Region

- The decision to state whether or not the data support the research hypothesis is based on a quantity computed from the sample data called the **test statistic**
- The decision will be to either reject H_0 or fail to reject H_0
- In developing our decision rule, we will assume that the value of μ is the null value μ_0
- The values of the test statistic that we are very unlikely to observe if $\mu = \mu_0$ are called the **rejection region**
- We will reject H_0 if the test statistic computed on the data is in the rejection region

- Give a test statistic and the rejection region:

The Texas A&M agricultural extension service wants to determine whether the mean yield per acre (in bushels) for a particular variety of soybeans has increased during the current year over the mean yield in the previous 2 years when μ was 520 bushels per acre.

Type I error/Type II error

- Type-I error occurs if H_0 is rejected when it is true. The probability of a Type I error is denoted by α
- Type-II error occurs if H_0 is not rejected when it is false (or H_a is true). The probability of a Type II error is denoted by β
- It is not possible to simultaneously minimize both α and β
- Usually, the experimenter specifies a tolerable probability for a Type I error of the statistical test, called **level of significance**. Popular values for α are 0.01 and 0.05
- Example: If we are willing to take the risk that 1 time in 40 we would incorrectly reject the null hypothesis, then $\alpha = 1/40 = 0.025$.

Two Types of Alternatives

Compare the two questions and give the alternative hypotheses

- Is the mean fill per bottle less than 591 mL?
- Is the mean fill per bottle different from 591 mL?

Statistical Test for μ with a Normal Population Distribution (σ Known) or Large Sample Size n

- Hypotheses:
  - Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$ (right-tailed test)
  - Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$ (left-tailed test)
  - Case 3: $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ (two-tailed test)
- Test Statistic: $z = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$
- For a fixed level of significance α, reject H_0 if
  - Case 1: $z \geq z_\alpha$. R command: `qnorm(1-alpha)`
  - Case 2: $z \leq -z_\alpha$. R command: `qnorm(alpha)`
  - Case 3: $|z| \geq z_{\alpha/2}$. R command: `qnorm(1-alpha/2)`
- Procedure is valid if the population distribution is normally distributed with σ known, or $n \geq 30$ (cf. Central Limit Theorem)

- As a part of her evaluation of municipal employees, the city manager audits the parking tickets issued by city parking officers to determine the number of tickets that were contested by the car owner and found to be improperly issued.
- In past years, the number of improperly issued tickets per officer had a normal distribution with mean $\mu = 380$ and $\sigma = 35.2$.
- Because there has recently been a change inthe city's parking regulations, the city manager suspects that the mean number of improperly issued tickets has increased.
- An audit of 50 randomly selected officers is conducted to test whether there has been an increase in improper tickets.
- Use the sample data given here and $\alpha = 0.01$ to test the research hypothesis that the mean number of improperly issued tickets is greater than 380. The audit generates the following data: $n = 50$ and $\bar{y} = 390$.

The p-value of a Statistical Test

- The observed significance or **p-value**, p is the probability, on the supposition that H_0 is true, of obtaining a result at least as contrary to H_0 and in favor of H_a as the result actually observed in the sample data
- Compare the p-value to significance level α and make a decision: reject H_0 if $p \leq \alpha$ and do not reject H_0 if $p > \alpha$

p-value for Statistical Test for μ with a Normal Population Distribution (σ Known) or Large Sample Size n

- Hypotheses:
  - Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$ (right-tailed test)
  - Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$ (left-tailed test)
  - Case 3: $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ (two-tailed test)
- The p-value can be computed as
  - Case 1: $P(z \geq$ computed $z)$. R command: `1-pnorm(z)`
  - Case 2: $P(z \leq$ computed $z)$. R command: `pnorm(z)`
  - Case 3: $2P(z \geq |$computed $z|)$. R command: `2*(1-pnorm(abs(z)))`

Example

- The total score in a professional basketball game is the sum of the scores of the two teams. An expert commentator claims that the average total score for NBA games is 202.5
- A fan suspects that this is an overstatement
- He selects a random sample of 85 games and obtains a mean total score of 199.2 with standard deviation 19.63
- Determine, at the 5% level of significance, whether there is sufficient evidence in the sample to reject the expert commentator's claim using the p-value approach

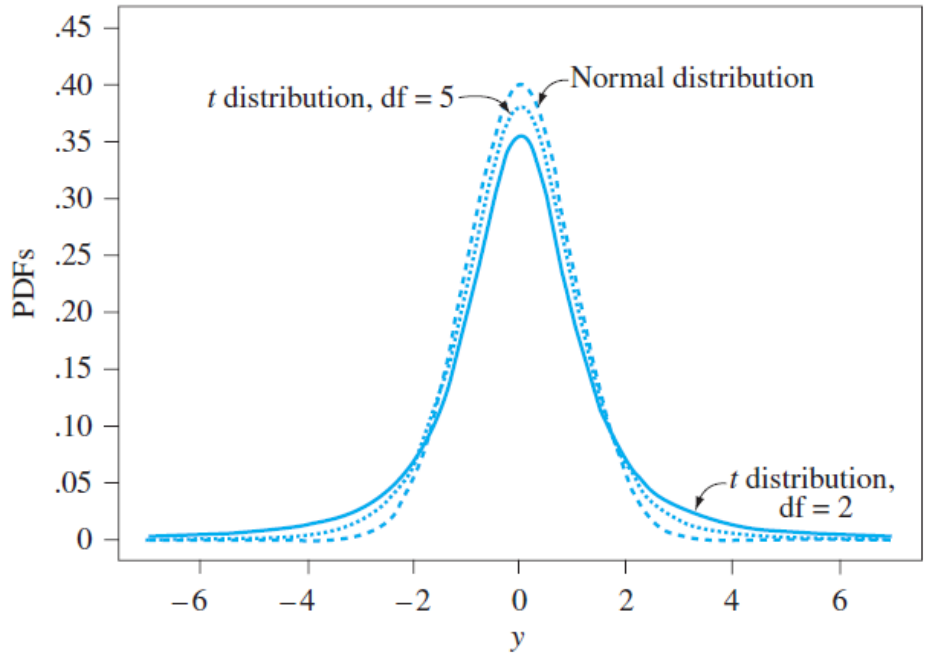
Inferences about μ for a Normal Population, σ Unknown

In this section, we present a test that can be applied when σ is unknown, no matter what the sample size, provided the population distribution is approximately normal

t-distributions

- When the population distribution is normal, the quantity $\frac{\bar{y} - \mu_0}{s/\sqrt{n}}$ is called the t statistic and its distribution is called the **Student's t distribution**
- The t distribution approximation is valid for a population with a mound-shaped distribution

- There are many different t distributions. We specify a particular one by a parameter called the degrees of freedom (df)



df

- the basic idea is that degrees of freedom are pieces of information for estimating σ using s.
- A second method of explaining degrees of freedom is to recall that σ measures the dispersion of the population values about μ, so prior to estimating s we must first estimate μ. Hence, the number of pieces of information (degrees of freedom) in the data that can be used to estimate σ is $n-1$, the number of original data values minus the number of parameters estimated prior to estimating σ.

t-distributions

- The t distribution is symmetrical about 0 and hence has mean equal to 0, the same as the z distribution
- The t distribution has variance $df/(df-2)$, and hence is more variable than the z distribution, which has variance equal to 1.
- As the df increases, the t distribution approaches the z distribution. (as df increases, the variance $df/(df-2)$ approaches 1.)
- The statistic $t = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$ has a t distribution with df$=n-1$

Statistical Test for μ with a Normal Population Distribution (σ Unknown)

- Hypotheses:
  - Case 1: $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$ (right-tailed test)
  - Case 2: $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$ (left-tailed test)
  - Case 3: $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$ (two-tailed test)
- Test Statistic: $t = \dfrac{\bar{y} - \mu_0}{s/\sqrt{n}}$
- For a fixed level of significance α and df$=n-1$, reject H_0 if
  - Case 1: $t \geq t_\alpha$. R command: `qt(1-alpha, n-1)`
  - Case 2: $t \leq -t_\alpha$. R command: `qt(alpha, n-1)`
  - Case 3: $|t| \geq t_{\alpha/2}$. R command: `qt(1-alpha/2, n-1)`
- The p-value can be computed as
  - Case 1: $P(t \geq$ computed $t)$. R command: `1-pt(t, n-1)`
  - Case 2: $P(t \leq$ computed $t)$. R command: `pt(t, n-1)`
  - Case 3: $2P(t \geq |$computed $t|)$. R command: `2*(1-pt(t, n-1))`

Confidence Interval

- In addition to being able to run a statistical test for μ when σ is unknown, we can construct a confidence interval using t
- A $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{y} \pm t_{\alpha/2}\, s/\sqrt{n}$$

Example

- A massive multistate outbreak of food-borne illness was attributed to Salmonella enteritidis. Epidemiologists determined that the source of the illness was ice cream.
- They sampled nine production runs from the company that had produced the ice cream to determine the level of Salmonella enteritidis in the ice cream.
- These levels (MPN/g) are as follows:$.593, .142, .329, .691, .231, .793, .519, .392, .418$
- Use these data to determine whether the average level of Salmonella enteritidis in the ice cream is greater than $.3$ MPN/g, a level that is considered to be very dangerous. Set $\alpha = .01$.