Chapter 6
Compare Two Groups

Section 6.1:
The Mean Difference

David W. Gerbing

The School of Business
Portland State University

- The Mean Difference
  - The Model
  - Confidence Interval
  - Hypothesis Test
  - Application
    - Data and Descriptive Statistics
    - Assumptions of Inference
    - Inferential Analysis
    - Conclusion

6.1a
The Model

Managerial Application: Introduction

Which of two suppliers has the smaller average delivery time?

- Managers at one firm decided to choose one of two current suppliers of a component needed in the manufacture of their primary product
- One aspect of the evaluation of supplier performance is Ship Time, the elapsed time from the receipt of the purchase order to the delivery of the part
- Is one supplier faster, on average? What is the difference between the mean ship time for each of the two companies?
- **Problem**: The past, the source of the data for the analysis of the descriptive statistics, represents just one *arbitrary* difference between the two sample mean ship times, replete with sampling error
- **Solution**: Use inferential statistics to estimate the true population mean difference

Compare Means

Question of interest

- **Key Concept**: Compare means of the same variable across different groups
- Example: Does one supplier have a faster Ship Time, on average, than the other?
- To answer this question, compare the mean Ship Time for each of the two suppliers
  - Collect the ship times on a group of shipments for each of the two suppliers
  - From *each* group of data values separately, calculate the summary statistics such as sample mean, m_1 and m_2
- The result of this data collection is a data table with two variables that defines two groups of data

Two Variables in Analysis of the Mean Difference

Compare the responses across the two groups

- **Grouping Variable**: A variable, generically called X, with two values, called levels or categories, that define the two groups
  - The grouping variable, X, has exactly two values
  - Example: Variable X is Supplier, with values: S1 and S2
- **Response Variable**: Continuous variable, the response of interest, measured for all observations in each group, generically called Y
  - The response variable, Y, is compared across the groups
  - Example: Ship Time
- Each row of data in the data table contains one of the two values of X and the corresponding value of Y

The Mean Difference

Compare the Group Means by Their Difference

- For the two groups defined by Grouping Variable X, compare the sample means of Response Variable Y, m_1 and m_2
- **Key Concept**: Compare m_1 to m_2 by their difference, $m_1 - m_2$
  - Example: Y is Ship Time for Suppliers S1 and S2
  - First calculate m_1 and m_2, the respective sample means of each group of recorded Ship Times, and then their difference
- The analysis includes descriptive statistics, plus the confidence interval and hypothesis test of the mean difference

> Analyze with `lessR` function `ttest()`, here with response variable Y and grouping variable X
>
> `> tt_brief(Y ~ X)` or `tt(Y ~ X)`
>
> Read the tilde, "~" to mean "is related to" or "depends on", so to what extent does the value of Y relate to the value of X?
>
> Can also run `tt_brief()` to obtain briefer output

More Examples: Analysis of the Mean Difference

Values of Categorical Variable X	Continuous Response Variable Y	
Groups to Compare (levels)	*Variable Compared*	*Question of Interest*
men vs. women	salaries	Do average Salaries differ according to Gender?
red vs. blue packages	sales	Do average Sales differ according to Package Color?
two different training methods	job performance	Does Training Method affect average Performance?
two investment portfolios	rate of return	Does one Portfolio have a higher average Rate of Return?
our product vs. competitor's	satisfaction	Does average Satisfaction differ among the two Products?

Fluctuations of Sample Mean Difference

The presence of sampling error requires statistical inference

- Suppose the population average Ship Times for the two suppliers are exactly the same
- Yet samples from both suppliers with the *same* population mean Ship Time yield *different* sample means
- Even with the same average Ship Time, two different samples, such as from two different companies, yield two different sample means, m_1 and m_2
- Why? Each sample mean is contaminated with sampling error
- **Key Concept**: The sample mean difference only approximates the corresponding population mean difference

$$m_1 - m_2 \approx \mu_1 - \mu_2$$

- Hence the need for statistical inference to infer the true underlying, stable reality, $\mu_1 - \mu_2$

Focus on a Zero Population Mean Difference

Equality of Population Means

- The analysis of the mean difference is the analysis of $\mu_1 - \mu_2$
- μ_1 and μ_2, though with unknown values, are each a constant for any one application, a number such as 5 or 5.42 or -16.1
- The difference between population means is also a constant, the subtraction of two constants
- To compare two values to each other by subtraction leads to the value of zero as the natural point of comparison
- Three possibilities for the relation of μ_1 and μ_2 with each other
  - $\mu_1 - \mu_2 = 0$, average of both groups is the same
  - $\mu_1 - \mu_2 > 0$, average of the first group is larger
  - $\mu_1 - \mu_2 < 0$, average of the second group is larger
- **Purpose of inference** of the mean difference: Identify which of these three conditions is true, and if > or <, by how much

Distribution of the Sample Mean Difference

What happens over, usually hypothetical, multiple samples?

- Statistical inference, as applied to confidence intervals and hypothesis tests, is based on the distribution of the sample mean difference, $m_1 - m_2$, over hypothetical multiple samples
- As with the distribution of the sample mean, the sample mean difference over multiple samples must be normal before inference can be based on normal or t-distribution probabilities
- **Normality of the Mean Difference**: If m_1 and m_2 are normal, the sample mean difference $m_1 - m_2$ is also normal
- Evaluate the normality of m for each group separately, as shown previously, with the Central Limit Theorem:
- m is at least approximately normal if . . .
  - $n \geq 30$ for any distribution of the population from which the data are sampled, or,
  - for small n when Y is normal or at least not skewed

Underlying Reality

The data values in each group should reflect a stable process

- Underlying the data is the focus of the analysis, true reality, μ, free of sampling error, but not directly observed
- As before, the stable process model is of Y_i, the ith data value:
  - underlying constant value, μ, shared with other data values from that specific population
  - a random error component, ϵ_i, unique to that data value
  - the individual data values, the Y_i's, randomly vary about μ, but the overall level of variation, σ, remains constant
- **Stable process model** of a data value from one of two groups:
  - Group #1: $Y_i = \mu_1 + \epsilon_{1i}$ with constant σ_1
  - Group #2: $Y_i = \mu_2 + \epsilon_{2i}$ with constant σ_2
- **Key Concept**: The interpretation of $\mu_1 - \mu_2$, the population mean difference, is meaningful only if each data value, Y_i, is sampled from a stable process consistent with μ_1 or μ_2

Comment on the Standard Error

Two different versions

- For inference of the mean, a crucial aspect of the analysis is the estimated standard error of the mean, s_m
- For confidence intervals and hypothesis tests of the mean difference, the corresponding standard error is the *estimated* standard error of the mean difference, $s_{m_1-m_2}$
- The concept of the standard error is the same for both the mean and the mean difference, the hypothetical variation of the underlying statistic over repeated samples
- In terms of implementation, however, there are two different versions of the estimated standard error of the mean difference
- To focus here on the meaning of the analysis, the technical details of the standard error are deferred until Section 6.2
- Here continue with the estimated standard error of the mean difference as a given, such as from a computer output

6.1b
The Confidence Interval

Purpose of Confidence Interval of a Mean Difference

Focus on Zero

- **Key Concept**: The confidence interval of the mean difference estimates the value of the population mean difference, $\mu_1 - \mu_2$
- **Confidence Interval**: Range of plausible values for $\mu_1 - \mu_2$, the difference between population means of the same variable across two different groups, at the specified level of confidence
- The one true value of the population mean difference, $\mu_1 - \mu_2$, is likely somewhere within the confidence interval
- The confidence interval answers the following question:

  Is one population mean plausibly larger than the other population mean and, if so, by how much?
- Of particular interest is the value that specifies the equality of population means:

  Is the value of zero in the confidence interval for the population mean difference?

Form of the Confidence Interval of a Mean Difference

Statistical inference requires the standard error

- Confidence interval of the mean and mean difference have the same form
- The t-cutoff, which defines the 95% range of sampling variability, is the same as before, $t_{.025} \approx 2$
- **Margin of Error**: About 2 standard errors on either side of the sample *mean difference*
- 95% confidence interval of the *mean difference* is the margin of error added and subtracted from the sample mean difference
- Notation for the estimated standard error of the mean difference: $s_{m_1 - m_2}$
- **Confidence interval for the mean difference**:

$$\mu_1 - \mu_2 \text{ likely within } m_1 - m_2 \pm (t_{.025})\, s_{m_1 - m_2}$$

Illustration: Zero in Confidence Interval of Mean Difference

- Suppose that a given data analysis yields a
  - Sample mean difference of $m_1 - m_2 = 1.25$
  - t-cutoff and standard error, of which the specific values are not given here, but from this information . . .
  - Confidence interval, which extends from -1.2 to 0 to 3.7

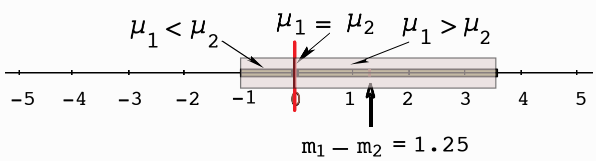


Figure: Confidence interval about sample mean difference of 1.25

Illustration: Zero in Confidence Interval of Mean Difference

Focus on 0

- Conf interval includes negative values, where $\mu_2 > \mu_1$
- Conf interval includes zero, the value at which $\mu_1 = \mu_2$
- Conf interval includes positive values, where $\mu_1 > \mu_2$
- **Key Concept**: When zero is in the confidence interval of the mean difference, the two population means could plausibly equal each other, or either one could be larger than the other
- In this example, the population mean of first group is plausibly up to less than 1.2 units *smaller*, or equal to, or up to 3.7 units larger than population mean of second group
- **Interpretation**: No difference detected between the population means of the two groups

Illustration: Zero *not* in Conf Int of the Mean Difference

Consider a 95% confidence interval, which centers around $m_1 - m_2 = 2.67$ and extends from 0.2 to 4.9, so does *not* include 0

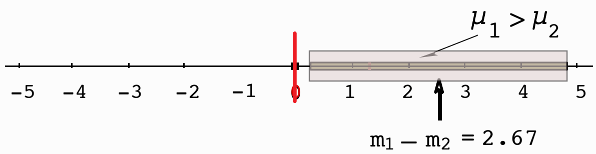


Figure: Confidence interval about the sample mean difference of 2.67

Illustration: Zero *not* in Conf Int of the Mean Difference

- This confidence interval includes *only* positive values, where $\mu_1 > \mu_2$
- **Key Concept**: When zero is *not* in the confidence interval of the mean difference, then one of the population means is plausibly larger than the other
- In this example, all plausible values of the mean difference are positive, so the population mean of the first group is plausibly larger than the population mean of the second group
- **Interpretation**: A difference between population means detected, the first population mean is likely somewhere between .2 to 4.9 units larger than the second population mean

6.1c
The Hypothesis Test

Null and Alternative Hypotheses

The hypothesis that is evaluated

- The usual hypothesis test here is called a non-directional or two-tailed test because the interest is to interpret a deviation for *either* a positive or a negative mean difference
- In general, the null hypothesis may specify any constant difference in population means, but the vast majority of applications focus on the value of zero
- Null hypothesis for the two-tailed test of the equality of population means, no difference between population means

$$H_0: \ \mu_1 - \mu_2 = 0$$

- With the focus on zero, the alternative hypothesis for this two-tailed test is that of unequal population means

$$H_1: \ \mu_1 - \mu_2 \neq 0$$

Test Statistic for Hypothesis Test of a Mean Difference

A test statistic, here t, assesses distance from the null value

$$\text{test statistic} = \frac{\text{sample value} - \text{hypothesized value}}{\text{standard error}}$$

- As already shown, for a single mean: $t_m = \frac{m - \mu_0}{s_m}$
- What is the t-statistic for evaluating a mean difference, $m_1 - m_2$, against its standard error?
- The mean difference specified by the null hypothesis, $(\mu_1 - \mu_2)_0$, is usually 0
- For a mean difference:

$$t_{m_1 - m_2} = \frac{(m_1 - m_2) - (\mu_1 - \mu_2)_0}{s_{m_1 - m_2}} \quad \text{usually} \quad t_{m_1 - m_2} = \frac{m_1 - m_2}{s_{m_1 - m_2}}$$

p-value Criterion for the Mean Difference

Decide to reject or not reject the null hypothesis

- To perform the test, see how far the sample result, here the mean difference, is from the hypothesized value, usually 0
- ***p*-value** for the mean difference: Probability of obtaining a difference between sample means as large or larger than the observed difference, *assuming* the null hypothesis
- Compare *obtained* p-value to the *set* alpha, usually $\alpha = 0.05$
- *If* the t-statistic, distance from zero, is large, then the p-value is small, indicating an unlikely event
  - **Statistical decision**: Reject null hypothesis of no difference
  - **Interpretation**: Population group means differ
- *If* the t-statistic, the distance from zero, is small, then the p-value is large, indicating a likely event
  - **Statistical decision**: Do *not* reject null hypothesis
  - **Interpretation**: No difference between pop means detected

Agreement of Confidence Interval and Hypothesis Test

Same general statistical result if done correctly

- The analyses of the mean, mean difference or any other statistic, in terms of the confidence interval and the hypothesis test, are always consistent with each other
- If a difference is detected between population means
  - Zero is outside of the 95% confidence interval
  - p-value $< .05$
  - *Follow up*: Assess the potential extent of the difference, such as with the confidence interval
- If a difference is *not* detected between population means
  - Zero is inside the 95% confidence interval
  - p-value $> .05$
  - *Follow up*: Use power analysis (discussed later) to assess if a meaningful difference may not have been detected

6.1d Application

Managerial Question

Which of two suppliers has the smaller average delivery time?

- Choose the supplier with the smaller average delivery time of the two suppliers
- Last year's sample mean difference is $m_1 - m_2 = 1.48$ days, so Suppler 1 was almost one and a half days, on average, slower than Supplier 2
- **Problem**: The past, assessed by descriptive statistics, represents just one *arbitrary* sample mean difference replete with sampling error
- **Response**: Use inferential statistics to estimate $\mu_1 - \mu_2$, the true or population mean difference
- Assessment of the population difference in average delivery time provides a forecast about what will happen *next* year
- The analysis of the average difference begins with the descriptive statistics of last year's performance

Data and Descriptive Statistics

Data Format

Stacked vs unstacked data representation

- The generally preferred data format is the standard data table, one column per variable, which allows for many variables
- Analyze two variables, Supplier with Time, the ship time

Stacked data: Standard table

Supplier	Time
S1	8.7
S2	8.1
S1	7.3
S1	9.1
S2	6.5
S2	9.8
...	...

Unstacked data: One column of Y values for each group (e.g., Excel)

S1	S2
8.7	8.1
7.3	6.5
9.1	9.8
...	...

Data

Two variables

- Supplier: Grouping variable with two values, S1 and S2
- Time: Response variable, Ship time
- csv data file: `http://lessRstats.com/data/twogroup.csv`
- One row of data for each shipment, stacked format
- Two variables so two data values per row
- The data are ordered by time, listed in the order in which the shipments occurred

```
Supplier,Time
S1,8.7
S2,8.1
⋮
S2,9.7
S1,10.4
```

Analysis of the Mean Difference with lessR ttest

Descriptive statistics

The lessR ttest() function, provides the analysis of the mean difference, here rely on the brief version

```
> tt_brief(Time ~ Supplier)
```

- The first part of the output provides the descriptive statistics

```
Time for Supplier S1:  n = 19,  mean = 9.57,  sd = 1.45
Time for Supplier S2:  n = 15,  mean = 8.09,  sd = 1.59
```

- In these data samples, Supplier S1 had a longer average delivery time, $m_{S1} = 9.57$ versus $m_{S2} = 8.09$
- **Question of interest**: Does this observed difference in sample means generalize to a difference in the population means

Assumptions of Inference

Three Assumptions to Evaluate

When possible, evaluate each assumption

- Before beginning the inferential analysis, evaluate the underlying assumptions
  - Stability of each response variable Y over time
  - Normality of the sample mean difference of response variable Y over (hypothetical) repeated samples
  - The classic implementation of the estimated standard error of the mean difference assumes what is called homogeneity of variance, evaluated in Section 6.2

Evaluate Stability of Process Assumption

Stable process for Y of each group

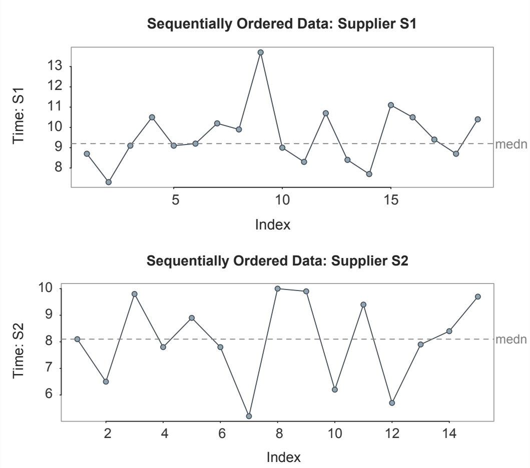
- Regarding stability, the goal is the analysis of the current process as it relates to future performance
- For either supplier, data from different shipping procedures in the same sample would not assist in forecasting the future of the *current* procedure
- The data for the suppliers are listed in the order they were collected over time, so evaluate each sample with a run chart
- Verify if random variation about a stable mean
- Search for any invalidating patterns such as a steady increase or decrease in ship times, or a sudden shift in ship time

Use the `ttest()` function option `line.chart=TRUE` to separate the data values by group and obtain the respective line charts

```
> tt_brief(Time ~ Supplier, line_chart=TRUE)
```


Random Variation about the Mean?

Generate the run chart for each group



- Run charts show no pronounced pattern, so assume stability of each μ and σ for the sampled values within the same group

Evaluate Normality of Mean Difference Assumption

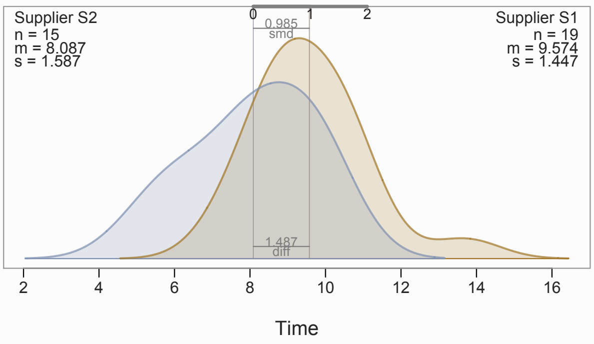
Normality of the sample mean difference

- Normality of the sample mean difference follows from normality of each sample mean
- The sample size of Ship Time for each supplier is below 30, so the Central Limit Theorem cannot guarantee normality of the sample mean for either sample
- If Ship Time for each group is normal in the population, or at least not too skewed, then each mean also follows a normal distribution even for low sample sizes
- Sample data that roughly follow a normal distribution provide evidence that the population distribution from which the data are sampled is normal
- Evaluate the normality of the data for *each* sample

The `ttest()` function automatically provides the density curves for the response variable for each *group* on the same graph

Density Curves Approximately Normal?

A density curved is a "smoothed" histogram



- Each distribution of Ship Time appears reasonably normal

Inferential Analysis

Analysis of the Mean Difference with `lessR` `ttest()`

Test of the Null Hypothesis of Equal Population Means

> The `lessR` `ttest()` function, provides the analysis of the mean difference, here as the brief version, to compare the mean ship Time across the two Suppliers
>
> ```
> > tt_brief(Time ~ Supplier)
> ```

- The first inferential analysis is of the hypothesis test

```
Hypothesis Test of 0 Mean Diff:
    t = 2.85,   df = 32,   p-value = 0.008
```

- First examine the *p*-value
- **Statistical Decision**: $p\text{-value} = 0.008 < \alpha = 0.05$, so reject the null hypothesis of no difference
- The sample mean difference is far from zero, the hypothesized population mean difference

Hypothesis Test: Conclusion

Inferential Statistics: Interpretation and Forecast

- The statistical decision is *not* the interpretation or managerial decision, which follow, in part, from the statistical decision
- The true average delivery times for Suppliers S1 and S2 apparently differ
- Informally conclude that the sample result in which Supplier S1 is slower than S2, on average, generalizes to the population

Managerial Decision: Resulting Action

- The ultimate purpose of the analysis is guide policy and decision making
- On the basis of this analysis, choose Supplier S2

Further Analysis

Estimate the extent of the difference in average Ship Times

- Hypothesis test shows that Supplier S2 is, on average, likely faster
- The natural follow-up question is, How much faster?
- The confidence interval provides the best available estimate of how much faster, on average, Supplier S2 is than Supplier S1

Analysis of the Mean Difference with `lessR` `ttest()`

The confidence interval

> The `lessR` `ttest()` function, provides the analysis of the mean difference, here as the brief version, to compare the mean ship Time across the two Suppliers
>
> ```
> > tt_brief(Time ~ Supplier)
> ```

- The second inferential analysis is the confidence interval

```
95% Confidence Interval for Mean Diff:  0.42 to 2.55
```

- The confidence interval estimates the true mean difference of how much slower is the first supplier
- All values in the confidence interval are positive, the range of plausible values for the population mean difference

Analysis with Traditional Notation

The confidence interval

- **Standard Error** $s_{m_1-m_2} = 0.522$ Discussed in Section 6.2
- **$t_{.025}$-cutoff** $t_{.025} = 2.037$ for $df = 32$
- ***t*-value** $t_{m_1-m_2} = \dfrac{(m_1 - m_2) - (\mu_1 - \mu_2)}{s_{m_1-m_2}}$

$$= \frac{(9.574 - 8.087) - 0}{0.522} = 2.851$$

- **Margin of Error** $E = (t_{.025})(s_m) = 2.037(0.522) = 1.062$
- **Confidence interval**

  $(m_1 - m_2) - E = 1.487 - 1.062 = 0.42$ days

  $(m_1 - m_2) + E = 1.487 + 1.062 = 2.55$ days

Conclusion

Confidence Interval: Conclusion

Descriptive Statistics: Summary of the Past

- Average of 15 Supplier S2 delivery times is 1.48 days smaller than the average for 19 deliveries by Supplier S1

Inferential Statistics: Interpretation and Forecast

- With 95% confidence, the true average delivery time, the forecast of future average delivery times, for Supplier S2 is between .42 and 2.54 days faster than for Supplier S1

Managerial Decision: Resulting Action

- On the basis of this criterion, choose Supplier S2, which apparently provides faster deliveries than Supplier S1, on average, from about one-half to two and one-half days

Index

Subtract 2 from each listed value to get the Slide #

- The End