taken together, can provide strong support. Using a method for combining probabilities, it can be determined that combining the probability values of 0.11 and 0.07 results in a probability value of 0.045. Therefore, these two non-significant findings taken together result in a significant finding.

Although there is never a statistical basis for concluding that an effect is exactly zero, a statistical analysis can demonstrate that an effect is most likely small. This is done by computing a confidence interval. If all effect sizes in the interval are small, then it can be concluded that the effect is small. For example, suppose an experiment tested the effectiveness of a treatment for insomnia. Assume that the mean time to fall asleep was 2 minutes shorter for those receiving the treatment than for those in the control group and that this difference was not significant. If the 95% confidence interval ranged from -4 to 8 minutes, then the researcher would be justified in concluding that the benefit is eight minutes or less. However, the researcher would not be justified in concluding the null hypothesis is true, or even that it was supported.

Steps in Hypothesis Testing

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

Learning Objectives

1. Be able to state the null hypothesis for both one-tailed and two-tailed tests
2. Differentiate between a significance level and a probability level
3. State the four steps involved in significance testing

1. The first step is to specify the null hypothesis. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu_1 - \mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is $\mu_2 - \mu_1 \geq 0$. This is equivalent to $\mu_1 \leq \mu_2$.
2. The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.
3. The third step is to compute the probability value (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.
4. Finally, compare the probability value with the α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

Significance Testing and Confidence Intervals

by David M. Lane

Prerequisites

- Chapter 10: Confidence Intervals Introduction
- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Significance Testing

Learning Objectives

1. Determine from a confidence interval whether a test is significant
2. Explain why a confidence interval makes clear that one should not accept the null hypothesis

There is a close relationship between confidence intervals and significance tests. Specifically, if a statistic is significantly different from 0 at the 0.05 level then the 95% confidence interval will not contain 0. All values in the confidence interval are plausible values for the parameter whereas values outside the interval are rejected as plausible values for the parameter. In the Physicians' Reactions case study, the 95% confidence interval for the difference between means extends from 2.00 to 11.26. Therefore, any value lower than 2.00 or higher than 11.26 is rejected as a plausible value for the population difference between means. Since zero is lower than 2.00, it is rejected as a plausible value and a test of the null hypothesis that there is no difference between means is significant. It turns out that the p value is 0.0057. There is a similar relationship between the 99% confidence interval and significance at the 0.01 level.

Whenever an effect is significant, all values in the confidence interval will be on the same side of zero (either all positive or all negative). Therefore, a significant finding allows the researcher to specify the direction of the effect. There are many situations in which it is very unlikely two conditions will have exactly the same population means. For example, it is practically impossible that aspirin and acetaminophen provide exactly the same degree of pain relief. Therefore, even before an experiment comparing their effectiveness is conducted, the researcher knows that the null hypothesis of exactly no difference is false. However, the researcher does not know which drug offers more relief. If a test of the difference is significant, then the direction of the difference is established because the values in the confidence interval are either all positive or all negative.

If the 95% confidence interval contains zero (more precisely, the parameter value specified in the null hypothesis), then the effect will not be significant at the 0.05 level. Looking at non-significant effects in terms of confidence intervals makes clear why the null hypothesis should not be accepted when it is not rejected: Every value in the confidence interval is a plausible value of the parameter. Since zero is in the interval, it cannot be rejected. However, there is an infinite number of other values in the interval (assuming continuous measurement), and none of them can be rejected either.

Misconceptions

by David M. Lane

Prerequisites

- Chapter 11: Introduction to Hypothesis Testing
- Chapter 11: Statistical Significance
- Chapter 11: Type I and II Errors

Learning Objectives

1. State why the probability value is not the probability the null hypothesis is false
2. Explain why a low probability value does not necessarily mean there is a large effect
3. Explain why a non-significant outcome does not mean the null hypothesis is probably true

Misconceptions about significance testing are common. This section lists three important ones.

1. Misconception: The probability value is the probability that the null hypothesis is false.
Proper interpretation: The probability value is the probability of a result as extreme or more extreme given that the null hypothesis is true. It is the probability of the data given the null hypothesis. It is not the probability that the null hypothesis is false.

2. Misconception: A low probability value indicates a large effect.
Proper interpretation: A low probability value indicates that the sample outcome (or one more extreme) would be very unlikely if the null hypothesis were true. A low probability value can occur with small effect sizes, particularly if the sample size is large.

3. Misconception: A non-significant outcome means that the null hypothesis is probably true.
Proper interpretation: A non-significant outcome means that the data do not conclusively demonstrate that the null hypothesis is false.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 11: Interpreting Non-Significant Results

Research in March, 2012 reported here found evidence for the existence of the Higgs Boson particle. However, the evidence for the existence of the particle was not statistically significant.

What do you think?

Did the researchers conclude that their investigation had been a failure or did they conclude they have evidence of the particle, just not strong enough evidence to draw a confident conclusion?

> One of the investigators stated, "We see some tantalizing evidence but not significant enough to make a stronger statement." Therefore, they were encouraged by the result. In a subsequent study, the evidence was significant.

References

Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, 67, 160-167

Exercises

Prerequisites

- All material presented in the Logic of Hypothesis Testing chapter

1. An experiment is conducted to test the claim that James Bond can taste the difference between a Martini that is shaken and one that is stirred. What is the null hypothesis?

2. The following explanation is incorrect. What three words should be added to make it correct?

   The probability value is the probability of obtaining a statistic as different (add three words here) from the parameter specified in the null hypothesis as the statistic obtained in the experiment. The probability value is computed assuming that the null hypothesis is true.

3. Why do experimenters test hypotheses they think are false?

4. State the null hypothesis for:

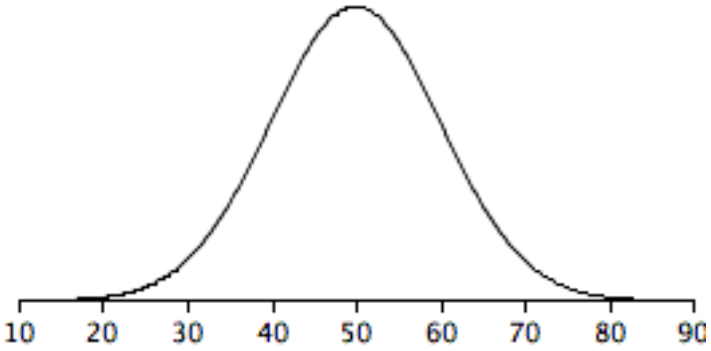
   a. An experiment testing whether echinacea decreases the length of colds.

   b. A correlational study on the relationship between brain size and intelligence.

   c. An investigation of whether a self-proclaimed psychic can predict the outcome of a coin flip.

   d. A study comparing a drug with a placebo on the amount of pain relief. (A one-tailed test was used.)

5. Assume the null hypothesis is that $\mu = 50$ and that the graph shown below is the sampling distribution of the mean (M). Would a sample value of M= 60 be significant in a two-tailed test at the .05 level? Roughly what value of M would be needed to be significant?



6. A researcher develops a new theory that predicts that vegetarians will have more of a particular vitamin in their blood than non-vegetarians. An experiment is conducted and vegetarians do have more of the vitamin, but the difference is not significant. The probability value is 0.13. Should the experimenter's confidence in the theory increase, decrease, or stay the same?

7. A researcher hypothesizes that the lowering in cholesterol associated with weight loss is really due to exercise. To test this, the researcher carefully controls for exercise while com- paring the cholesterol levels of a group of subjects who lose weight by dieting with a control group that does not diet. The difference between groups in cholesterol is not significant. Can the researcher claim that weight loss has no effect?

8. A significance test is performed and $p = .20$. Why can't the experimenter claim that the probability that the null hypothesis is true is .20?

9. For a drug to be approved by the FDA, the drug must be shown to be safe and effective. If the drug is significantly more effective than a placebo, then the drug is deemed effective. What do you know about the effectiveness of a drug once it has been approved by the FDA (assuming that there has not been a Type I error)?

10. When is it valid to use a one-tailed test? What is the advantage of a one-tailed test? Give an example of a null hypothesis that would be tested by a one-tailed test.

11. Distinguish between probability value and significance level.

12. Suppose a study was conducted on the effectiveness of a class on "How to take tests." The SAT scores of an experimental group and a control group were

compared. (There were 100 subjects in each group.) The mean score of the experimental group was 503 and the mean score of the control group was 499. The difference between means was found to be significant, p = .037. What do you conclude about the effectiveness of the class?

13. Is it more conservative to use an alpha level of .01 or an alpha level of .05? Would beta be higher for an alpha of .05 or for an alpha of .01?

14. Why is "Ho: "M1 = M2" not a proper null hypothesis?

15. An experimenter expects an effect to come out in a certain direction. Is this sufficient basis for using a one-tailed test? Why or why not?

16. How do the Type I and Type II error rates of one-tailed and two-tailed tests differ?

17. A two-tailed probability is .03. What is the one-tailed probability if the effect were in the specified direction? What would it be if the effect were in the other direction?

18. You choose an alpha level of .01 and then analyze your data.

    a. What is the probability that you will make a Type I error given that the null hypothesis is true?

    b. What is the probability that you will make a Type I error given that the null hypothesis is false?

19. Why doesn't it make sense to test the hypothesis that the sample mean is 42?

20. True/false: It is easier to reject the null hypothesis if the researcher uses a smaller alpha (α) level.

21. True/false: You are more likely to make a Type I error when using a small sample than when using a large sample.

22. True/false: You accept the alternative hypothesis when you reject the null hypothesis.

23. True/false: You do not accept the null hypothesis when you fail to reject it.

24. True/false: A researcher risks making a Type I error any time the null hypothesis is rejected.

12. Testing Means

A. Single Mean
B. Difference between Two Means (Independent Groups)
C. All Pairwise Comparisons Among Means
D. Specific Comparisons
E. Difference between Two Means (Correlated Pairs)
F. Specific Comparisons (Correlated Observations)
G. Pairwise Comparisons (Correlated Observations)
H. Exercises

Many, if not most experiments are designed to compare means. The experiment may involve only one sample mean that is to be compared to a specific value. Or the experiment could be testing differences among many different experimental conditions, and the experimenter could be interested in comparing each mean with each of the other means. This chapter covers methods of comparing means in many different experimental situations.

The topics covered here in sections C, D, F, and G are typically covered in other texts in a chapter on Analysis of Variance. We prefer to cover them here since they bear no necessary relationship to analysis of variance. As discussed by Wilkinson (1999), it is not logical to consider the procedures in this chapter as tests to be performed subsequent to an analysis of variance. Nor is it logical to call them post-hoc tests as some computer programs do.

Testing a Single Mean

by David M. Lane

Prerequisites

- Chapter 7: Normal Distributions
- Chapter 7: Areas Under Normal Distributions
- Chapter 9: Sampling Distribution of the Mean
- Chapter 9: Introduction to Sampling Distributions
- Chapter 10: t Distribution
- Chapter 11: Logic of Hypothesis Testing

Learning Objectives

1. Compute the probability of a sample mean being at least as high as a specified value when σ is known
2. Compute a two-tailed probability
3. Compute the probability of a sample mean being at least as high as a specified value when σ is estimated
4. State the assumptions required for item 3 above

This section shows how to test the null hypothesis that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50. In other words, the null hypothesis is that the population mean (μ) is 50. The (hypothetical) data are shown in Table 1. The data in Table 1 have a sample mean (M) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table 1. Distribution of scores.

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from μ by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As shown in Chapter 9, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_M = \mu$$

and

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

respectively. It is clear that $\mu_M = 50$. In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation (σ).

The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know σ and therefore you would use s, the sample estimate of σ. However, it is

instructive to see how the probability is computed if σ is known before proceeding to see how it is calculated when σ is estimated.

For the current example, if the null hypothesis is true, then based on the binomial distribution, one can compute that variance of the number correct is

$$\sigma^2 = N\pi(1-\pi)$$

$$= 100(0.5)(1-0.5)$$

$$= 25.$$

Therefore, σ = 5. For a σ of 5 and an N of 9, the standard deviation of the sampling distribution of the mean is 5/3 = 1.667. Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or more when the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute this probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use the normal distribution calculator (external link) as shown in Figure 1.

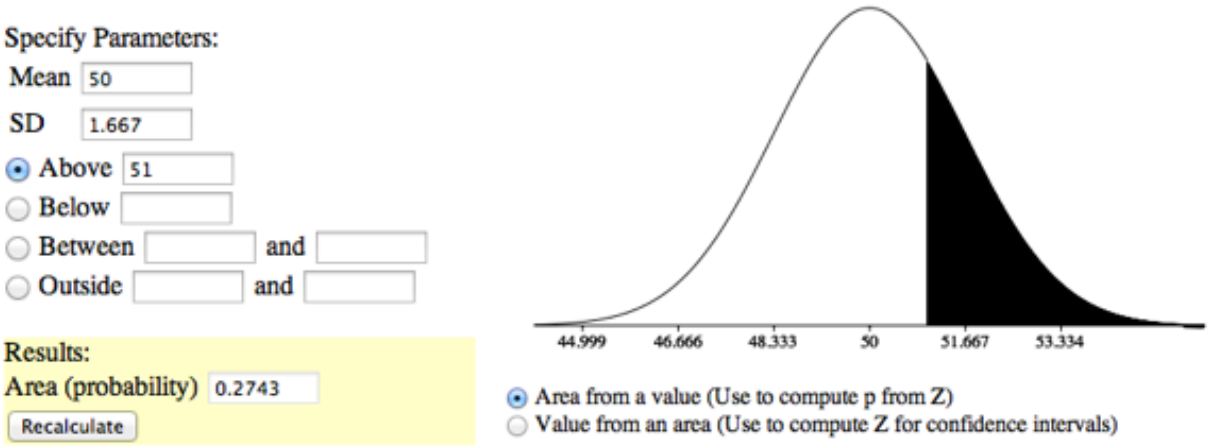


Figure 1. Probability of a sample mean being 51 or greater.

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area **above** 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results of the normal distribution calculator are shown in Figure 2.

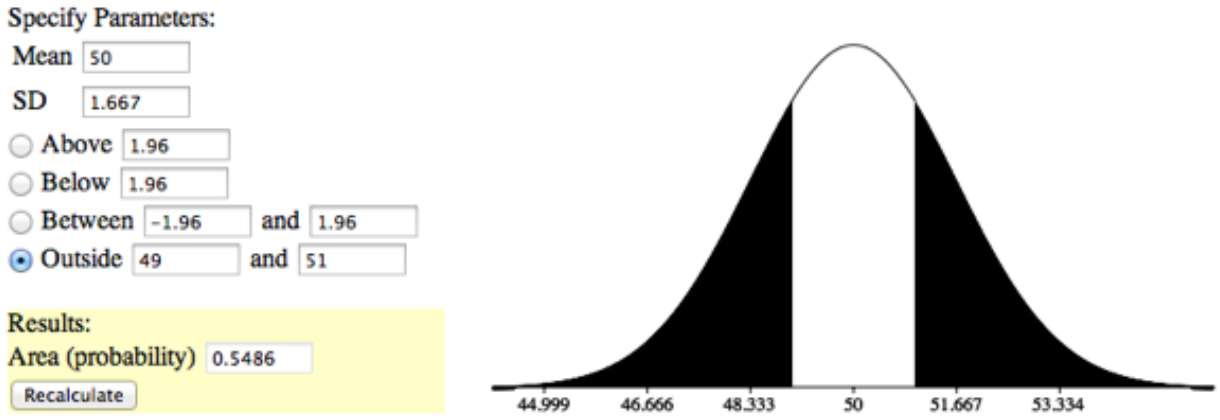


Figure 2. Probability of a sample mean being less than or equal to 49 or greater than or equal to 51.

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure 1.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by computing Z based on the formula

$$Z = \frac{M - \mu}{\sigma_M}$$

where Z is the value on the standard normal distribution, M is the sample mean, μ is the hypothesized value of the mean, and σ_M is the standard error of the mean. For this example, Z = (51-50)/1.667 = 0.60. The normal calculator with a mean of 0 and a standard deviation of 1 is shown in Figure 3.

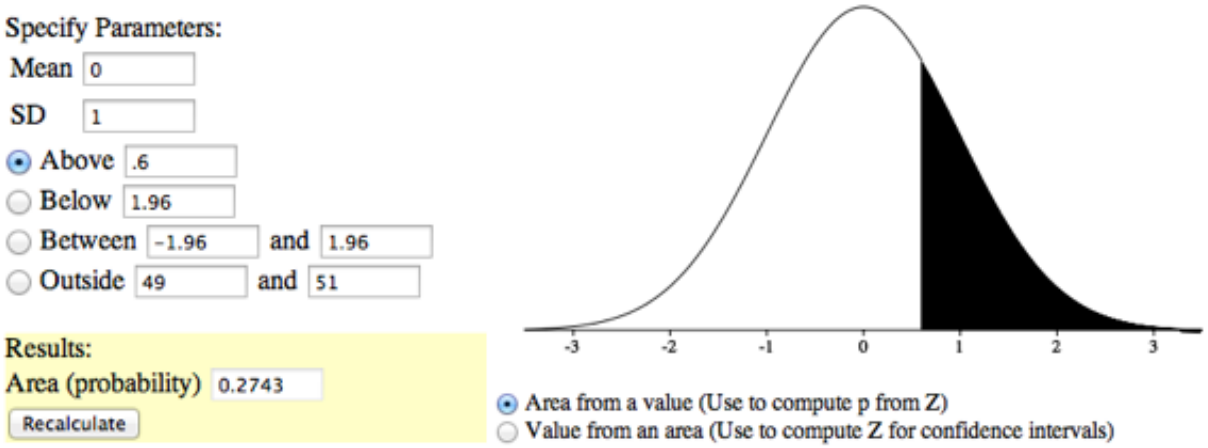


Figure 3. Calculation using the standardized normal distribution.

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analyses it is very rare that you would know σ and wish to estimate μ. Typically σ is not known and is estimated in a sample by s, and σ_M is estimated by s_M. For our next example, we will consider the data in the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table 2 shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled "Diff" that shows the difference in performance between the 0.6 mg (D60) and the 0 mg (D0) conditions. These difference scores are positive for children who performed better in the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

Table 2. DOG scores as a function of dosage.

D0	D60	Diff
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

To test this null hypothesis, we compute t using a special case of the following formula:

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\textit{s}\text{tandard error of the statistic}}$$

The special case of this formula applicable to testing a single mean is

$$t = \frac{M - \mu}{S_M}$$

where t is the value we compute for the significance test, M is the sample mean, μ is the hypothesized value of the population mean, and s_M is the estimated standard error of the mean. Notice the similarity of this formula to the formula for Z.

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean (M) of the N = 24 difference scores is 4.958, the hypothesized value of μ is 0, and the standard deviation (s) is 7.538. The estimate of the standard error of the mean is computed as:

$$s_m = \frac{s}{\sqrt{N}} = \frac{7.5382}{\sqrt{24}} = 1.54$$

Therefore, t = 4.96/1.54 = 3.22. The probability value for t depends on the degrees of freedom. The number of degrees of freedom is equal to N - 1 = 23. A t distribution calculator shows that a t less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.

Review of Assumptions

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

Differences between Two Means (Independent Groups)

by David M. Lane

Prerequisites

- Chapter 9: Sampling Distribution of Difference between Means
- Chapter 10: Confidence Intervals
- Chapter 10: Confidence Interval on the Difference between Means
- Chapter 11: Logic of Hypothesis Testing
- Chapter 12: Testing a Single Mean

Learning Objectives

1. State the assumptions for testing the difference between two means
2. Estimate the population variance assuming homogeneity of variance
3. Compute the standard error of the difference between means
4. Compute t and p for the difference between means
5. Format data for computer analysis

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. This section covers how to test for differences between means from two separate groups of subjects. A later section describes how to test for differences between the means of two conditions in designs where only one group of subjects is used and each subject is tested in each condition.

We take as an example the data from the "Animal Research" case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 1.

Table 1. Means and Variances in Animal Research study.

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very

important. What is important is whether there is a difference in the population means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent. The analysis of data with two scores per subject is shown in the section on the correlated t test later in this chapter.

Small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

We saw the following general formula for significance testing in the section on testing a single mean:

$$t = \frac{\text{statistic} - \text{hypothesized value}}{\textit{s}\text{tandard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0. The hypothesized value is the null hypothesis that the difference between population means is 0.

We continue to use the data from the "Animal Research" case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males. For this calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$$M_1 - M_2 = 5.3529 - 3.8824 = 1.4705.$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means so the estimated standard error of the statistic is ($s_{M_1-M_2}$). Recall from the relevant section in the chapter on

sampling distributions that the formula for the standard error of the difference between means is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate σ^2 and use that estimate in place of σ^2. Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2. In this example,

```
MSE = (2.743 + 2.985)/2 = 2.864.
```

Since n (the number of scores **in each group**) is 17,

$$S_{M_1-M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

The next step is to compute t by plugging these values into the formula:

```
t = 1.4705/.5805 = 2.533.
```

Finally, we compute the probability of getting a t as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to $(n_1 - 1) + (n_2 - 1)$, where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees for freedom is 16 + 16 = 32.

Once we have the degrees of freedom, we can use a t distribution calculator to find that the probability value for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. This is the probability of a $t < -2.533$ or a $t > 2.533$. A one-tailed test would result in a probability of 0.0082, which is half the two-tailed probability.

Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 2.

Table 2. Example Data.

Group 1	Group 2
3	2
4	6
5	8

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 2 is shown in Table 3.

Table 3. Reformatted Data

G	Y
1	3
1	4
1	5
2	2
2	6
2	8

Computations for Unequal Sample Sizes (optional)

The calculations are somewhat more complicated when the sample sizes are not equal. One consideration is that MSE, the estimate of variance, counts the group

with the larger sample size more than the group with the smaller sample size. Computationally, this is done by computing the sum of squares error (SSE) as follows:

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2$$

where M_1 is the mean for group 1 and M_2 is the mean for group 2. Consider the following small example:

Table 4. Unequal n

Group 1	Group 2
3	2
4	4
5	

$M_1 = 4$ and $M_2 = 3$.

$$SSE = (3-4)^2 + (4-4)^2 + (5-4)^2 + (2-3)^2 + (4-3)^2 = 4$$

Then, MSE is computed by:

$$MSE = \frac{SSE}{df}$$

The formula

$$S_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}}$$

is replaced by

$$S_{M_1-M_2} = \sqrt{\frac{2MSE}{n_h}}$$

where n_h is the harmonic mean of the sample sizes and is computed as follows:

$$n_h = \frac{2}{\frac{1}{n_1} + \frac{1}{n_1}} = \frac{2}{\frac{1}{3} + \frac{1}{2}} = 2.4$$

and

$$S_{M_1-M_2} = \sqrt{\frac{(2)(1.333)}{2.4}} = 1.054$$

Therefore,

```
t = (4-3)/1.054 = 0.949
```

and the two-tailed p = 0.413.

All Pairwise Comparisons Among Means

by David M. Lane

Prerequisites

• Chapter 12: Difference Between Two Means (Independent Groups)

Learning Objectives

1. Define pairwise comparison
2. Describe the problem with doing t tests among all pairs of means
3. Calculate the Tukey HSD test
4. Explain why Tukey test should not necessarily be considered a follow-up test

Many experiments are designed to compare more than two conditions. We will take as an example the case study "Smiles and Leniency." In this study, the effect of different types of smiles on the leniency showed to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table 1.

The problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the Type I error rate is inflated. Figure 1 shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

Table 1. Six Comparisons among Means.

false vs felt		
false vs miserable		
false vs neutral		
felt vs miserable		
felt vs neutral		
miserable vs neutral		

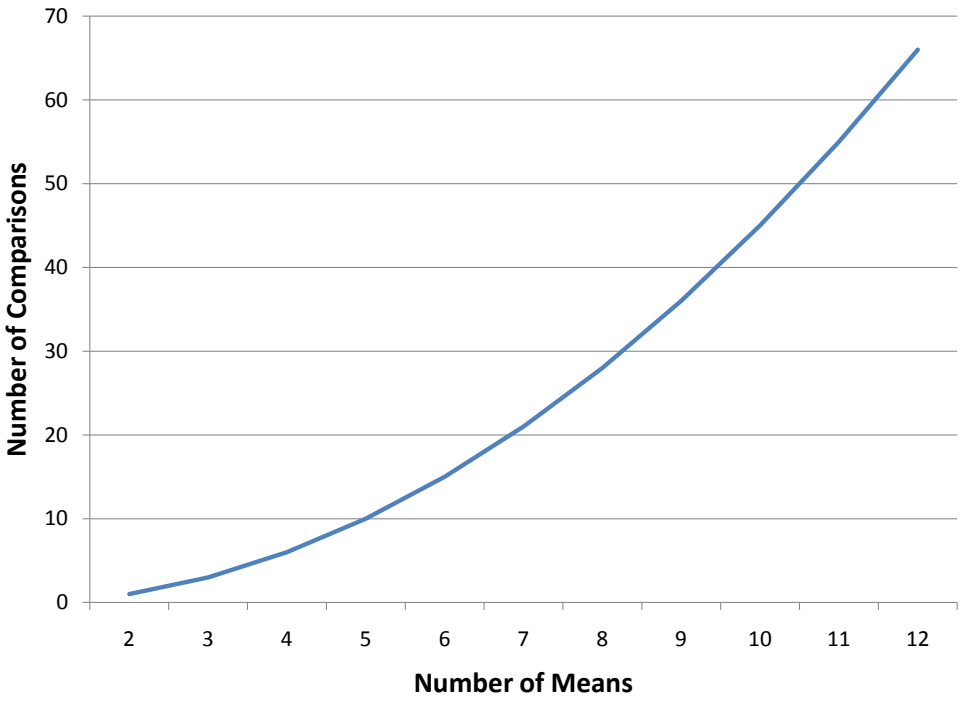


Figure 1. Number of pairwise comparisons as a function of the number of means.

Figure 2 shows the probability of a Type I error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

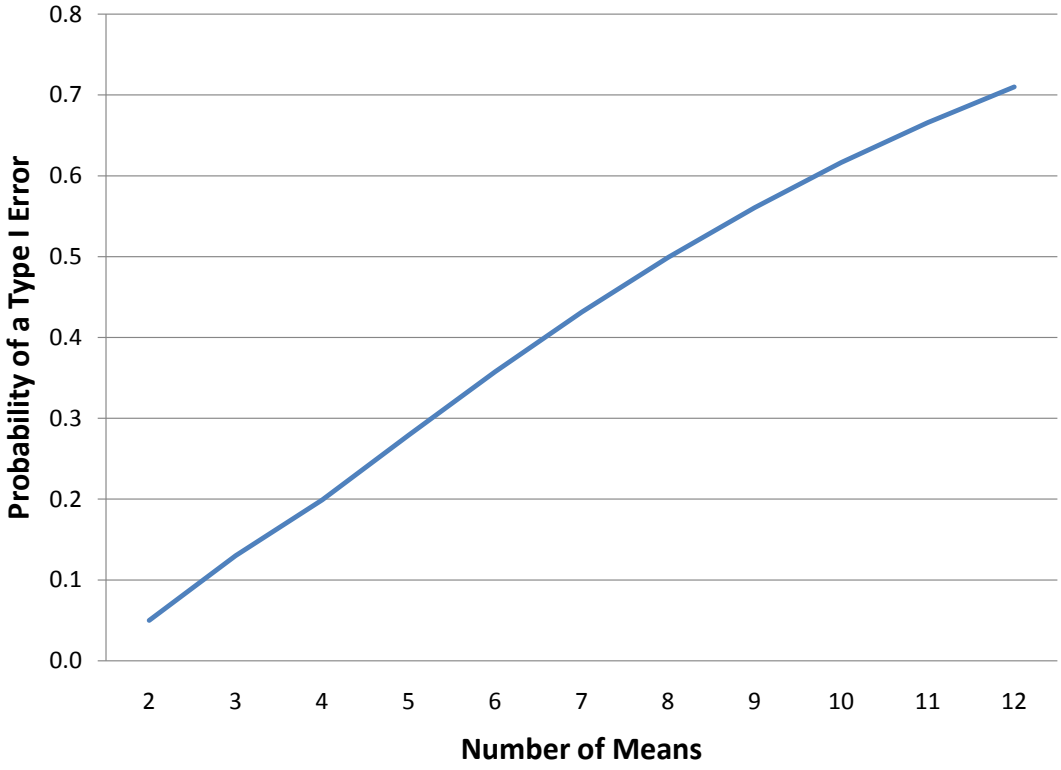


Figure 2. Probability of a Type I Error as a Function of the Number of Means.

The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or Tukey HSD for short. The Tukey HSD is based on a variation of the t *distribution* that takes into account the number of means being compared. This distribution is called the *studentized range distribution*.

Let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

Condition	Mean	Variance
FALSE	5.37	3.34
Felt	4.91	2.83

Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.
3. Compute

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

for each pair of means, where M_i is one mean, M_j is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute p for each comparison using the Studentized Range Calculator (external link; requires Java). The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, df = 136 - 4 = 132.

The tests for these data are shown in Table 2. The only significant comparison is between the false smile and the neutral smile.

Table 2. Six Pairwise Comparisons.

Comparison	Mi-Mj	Q	p
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject

it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in-between.

Assumptions

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

Computer Analysis

For most computer programs, you should format your data the same way you do for independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

Although full-featured statistics programs such as SAS, SPSS, R, and others can compute Tukey's test, smaller programs (including Analysis Lab) may not. However, these programs are generally able to compute a procedure known as Analysis of Variance (ANOVA). This procedure will be described in detail in a later chapter. Its relevance here is that an ANOVA computes the MSE that is used in the calculation of Tukey's test. For example, the following shows the ANOVA summary table for the "Smiles and Leniency" data.

Source	df	SSQ	MS	F	p
Condition	3	27.5349	9.1783	3.4650	0.0182
Error	132	349.6544	2.6489		
Total	135	377.1893			

The column labeled MS stands for "Mean Square" and therefore the value 2.6489 in the "Error" row and the MS column is the "Mean Squared Error" or MSE. Recall that this is the same value computed here (2.65) when rounded off.

Tukey's Test Need Not Be A Follow-Up to ANOVA

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Wilkinson and the Task Force on Statistical Inference (1999).

Computations for Unequal Sample Sizes (optional)

The calculation of MSE for unequal sample sizes is similar to its calculation in an independent-groups t test. Here are the steps:

1. Compute a Sum of Squares Error (SSE) using the following formula

$$SSE = \sum (X - M_1)^2 + \sum (X - M_2)^2 + \cdots + \sum (X - M_k)^2$$

where Mi is the mean of the i^{th} group and k is the number of groups.

2. Compute the degrees of freedom error (df_e) by subtracting the number of groups (k) from the total number of observations (N). Therefore,

df_e N - k.

Compute MSE by dividing SSE by df_e:

MSE = SSE/df_e.

For each comparison of means, use the harmonic mean of the n's for the two means (n_h).

All other aspects of the calculations are the same as when you have equal sample sizes.

Specific Comparisons (Independent Groups)

by David M. Lane

Prerequisites

- Chapter 12: Difference Between Two Means (Independent Groups)

Learning Objectives

1. Define linear combination
2. Specify a linear combination in terms of coefficients
3. Do a significance test for a specific comparison

There are many situations in which the comparisons among means are more complicated than simply comparing one mean with another. This section shows how to test these more complex comparisons. The methods in this section assume that the comparison among means was decided on before looking at the data. Therefore these comparisons are called *planned comparisons*. A different procedure is necessary for *unplanned comparisons*.

Let's begin with the made-up data from a hypothetical experiment shown in Table 1. Twelve subjects were selected from a population of high-self-esteem subjects (esteem = 1) and an additional 12 subjects were selected from a population of low-self-esteem subjects (esteem = 2). Subjects then performed on a task and (independent of how well they really did) half in each esteem category were told they succeeded (outcome = 1) and the other half were told they failed (outcome = 2). Therefore, there were six subjects in each of the four esteem/outcome combinations and 24 subjects all together.

After the task, subjects were asked to rate (on a 10-point scale) how much of their outcome (success or failure) they attributed to themselves as opposed to being due to the nature of the task.

Table 1. Data from Hypothetical Experiment.

outcome	esteem	attrib
1	1	7
1	1	8
1	1	7
1	1	8
1	1	9
1	1	5
1	2	6
1	2	5
1	2	7
1	2	4
1	2	5
1	2	6
2	1	4
2	1	6
2	1	5
2	1	4
2	1	7
2	1	3
2	2	9
2	2	8
2	2	9
2	2	8
2	2	7
2	2	6

The means of the four conditions are shown in Table 2.

Table 2. Mean ratings of self-attributions of success or failure.

Outcome	Esteem	Mean
Success	High Self-Esteem	7.333
	Low Self-Esteem	5.500
Failure	High Self-Esteem	4.833
	Low Self-Esteem	7.833

There are several questions we can ask about the data. We begin by asking whether, on average, subjects who were told they succeeded differed significantly from subjects who were told they failed. The means for subjects in the success condition are 7.333 for the high-self-esteem subjects and 5.500 for the low-self-esteem subjects. Therefore, the mean for all subjects in the success condition is (7.3333 + 5.5000)/2 = 6.4167. Similarly, the mean for all subjects in the failure condition is (4.8333 + 7.8333)/2 = 6.3333. The question is: How do we do a significance test for this difference of 6.4167-6.3333 = 0.083?

The first step is to express this difference in terms of a linear combination using a set of coefficients and the means. This may sound complex, but it is really pretty easy. We can compute the mean of the success conditions by multiplying each success mean by 0.5 and then adding the result. In other words, we compute

```
(.5)(7.333) + (.5)(5.500)
  = 3.67 + 2.75
  = 6.42
```

Similarly we can compute the mean of the failure conditions by multiplying each failure mean by 0.5 and then adding the result:

```
(.5)(4.833) + (.5)(7.833)
  = 2.417 + 3.917
  = 6.33
```

The difference between the two means can be expressed as

```
.5 x 7.333 + .5 x 5.500 -(.5 x 4.833 + .5 x 7.833)=
.5 x 7.333 + .5 x 5.500 - .5 x 4.833 - .5 x 7.8333
```

We therefore can compute the difference between the "success" mean and the "failure" mean by multiplying each "success" mean by 0.5, each "failure" mean by -0.5 and adding the results. In Table 3, the coefficient column is the multiplier and the product column in the result of the multiplication. If we add up the four values in the product column we get:

```
L = 3.667 + 2.750 - 2.417 - 3.917 = 0.083
```

This is the same value we got when we computed the difference between means previously (within rounding error). We call the value "L" for "linear combination."

Table 3. Coefficients for comparing low and high self-esteem.

Outcome	Esteem	Mean	Coeff	Product
Success	High Self-Esteem	7.333	0.5	3.667
	Low Self-Esteem	5.500	0.5	2.750
Failure	High Self-Esteem	4.833	-0.5	-2.417
	Low Self-Esteem	7.833	-0.5	-3.917

Now, the question is whether our value of L is significantly different from 0. The general formula for L is

$$L = \sum c_i M_i$$

where c_i is the i^{th} coefficient and M_i is the i^{th} mean. As shown above, L = 0.083. The formula for testing L for significance is shown below:

$$t = \frac{L}{\sqrt{\frac{\sum c_i^2 MSE}{n}}}$$

In this example,

$$\sum c_i^2 = .5^2 + .5^2 + (-.5)^2 + (-.5)^2 = 1$$

MSE is the mean of the variances. The four variances are shown in Table 4. Their mean is 1.625. Therefore MSE = 1.625.

Table 4. Variances of attributions of success or failure to oneself.

Outcome	Esteem	Variance
Success	High Self-Esteem	1.867
	Low Self-Esteem	1.100
Failure	High Self-Esteem	2.167
	Low Self-Esteem	1.367

The value of n is the number of subjects in each group. Here n = 6.

Putting it all together,

$$t = \frac{0.083}{\sqrt{\frac{(1)(1.625)}{6}}} = 0.16.$$

We need to know the degrees for freedom in order to compute the probability value. The degrees of freedom is

```
df = N - k
```

where N is the total number of subjects (24) and k is the number of groups (4). Therefore, df = 20. Using the Online Calculator, we find that the two-tailed probability value is 0.874. Therefore, the difference between the "success" condition and the "failure" condition is not significant.

A more interesting question about the results is whether the effect of outcome (success or failure) differs depending on the self-esteem of the subject. For example, success may make high-self-esteem subjects **more** likely to attribute the outcome to themselves, whereas success may make low-self-esteem subjects **less** likely to attribute the outcome to themselves.

To test this, we have to test a difference between differences. Specifically, is the difference between success and failure outcomes for the high-self-esteem subjects different from the difference between success and failure outcomes for the low-self-esteem subjects? The means in Table 5 suggest that this is the case. For the high-self-esteem subjects, the difference between the success and failure

attribution scores is 7.333 - 4.833 = 2.500. For low-self-esteem subjects, the difference is 5.500 - 7.833 = -2.333. The difference between differences is 2.500 - (-2.333) = 4.833.

The coefficients to test this difference between differences are shown in Table 5.

Table 5. Coefficients for testing differences between differences.

Self-Esteem	Outcome	Mean	Coefficient	Product
High	Success	7.333	1	7.333
	Failure	4.833	-1	-4.833
Low	Success	5.500	-1	-5.500
	Failure	7.833	1	7.833

If it is hard to see where these coefficients came from, consider that our difference between differences was computed this way:

```
(7.33 - 4.83) - (5.5 - 7.83)

= 7.3 - 4.83 - 5.5 + 7.83

= (1)7.3 + (-1)4.83 + (-1)5.5 + (1)7.83
```

The values in parentheses are the coefficients.

To continue the calculations,

$$L = 4.83$$

$$\sum c_i^2 = 1^2 + (-1)^2 + (-1)^2 + (1)^2 = 4$$

$$t = \frac{4.83}{\sqrt{\frac{(4)(1.625)}{6}}} = 4.64$$

The two-tailed p value is 0.0002. Therefore, the difference between differences is highly significant.

In a later chapter on Analysis of Variance, you will see that comparisons such as this are testing what is called an *interaction*. In general, there is an interaction when the effect of one variable differs as a function of the level of another variable. . In this example, the effect of the outcome variable is different depending on the subject's self-esteem. For the high-self-esteem subjects, success led to more self-attribution than did failure; for the low-self-esteem subjects, success led to less self-attribution than did failure.

Multiple Comparisons

The more comparisons you make, the greater your chance of a Type I error. It is useful to distinguish between two error rates: (1) the *per-comparison error rate* and (2) the *familywise error rate*. The per-comparison error rate is the probability of a Type I error for a particular comparison. The *familywise error rate* is the probability of making one or more Type I errors in a family or set of comparisons. In the attribution experiment discussed previously, we computed two comparisons. If we use the 0.05 level for each comparison, then the per-comparison rate is simply 0.05. The familywise rate can be complex. Fortunately, there is a simple approximation that is fairly accurate when the number of comparisons is small. Defining α as the per-comparison error rate and c as the number of comparisons, the following inequality always holds true for the familywise error rate (FW):

$$FW \leq c\alpha$$

This inequality is called the *Bonferroni inequality*. In practice, FW can be approximated by $c\alpha$. This is a conservative approximation since FW can never be greater than $c\alpha$ and is generally less than $c\alpha$.

The Bonferroni inequality can be used to control the familywise error rate as follows: If you want the familywise error rate to be α, you use α/c as the per-comparison error rate. This correction, called the *Bonferroni correction*, will generally result in a familywise error rate less than α. Alternatively, you could multiply the by c and use the original α level.

Should the familywise error rate be controlled? Unfortunately, there is no clear-cut answer to this question. The disadvantage of controlling the familywise error rate is that it makes it more difficult to obtain a significant result for any given comparison: The more comparisons you do, the lower the per-comparison rate must be and therefore the harder it is to reach significance. That is, the power

is lower when you control the familywise error rate. The advantage is that you have a lower chance of making a Type I error.

One consideration is the definition of a family of comparisons. Let's say you conducted a study in which you were interested in whether there was a difference between male and female babies in the age at which they started crawling. After you finished analyzing the data, a colleague of yours had a totally different research question: Do babies who are born in the winter differ from those born in the summer in the age they start crawling? Should the familywise rate be controlled or should it be allowed to be greater than 0.05? Our view is that there is no reason you should be penalized (by lower power) just because your colleague used the same data to address a different research question. Therefore, the familywise error rate need not be controlled. Consider the two comparisons done on the attribution example at the beginning of this section: These comparisons are testing completely different hypotheses. Therefore, controlling the familywise rate is not necessary.

Now consider a study designed to investigate the relationship between various variables and the ability of subjects to predict the outcome of a coin flip. One comparison is between males and females; a second comparison is between those over 40 and those under 40; a third is between vegetarians and non-vegetarians; and a fourth is between firstborns and others. The question of whether these four comparisons are testing different hypotheses depends on your point of view. On the one hand, there is nothing about whether age makes a difference that is related to whether diet makes a difference. In that sense, the comparisons are addressing different hypotheses. On the other hand, the whole series of comparisons could be seen as addressing the general question of whether anything affects the ability to predict the outcome of a coin flip. If nothing does, then allowing the familywise rate to be high means that there is a high probability of reaching the wrong conclusion.

Orthogonal Comparisons

In the preceding sections, we talked about comparisons being independent. Independent comparisons are often called orthogonal comparisons. There is a simple test to determine whether two comparisons are orthogonal: If the sum of the products of the coefficients is 0, then the comparisons are orthogonal. Consider again the experiment on the attribution of success or failure. Table 6 shows the

coefficients previously presented in Table 3 and in Table 5. The column “C1” contains the coefficients from the comparison shown in Table 3; the column “C2” contains the coefficients from the comparison shown in Table 5. The column labeled “Product” is the product of these two columns. Note that the sum of the numbers in this column is 0. Therefore, the two comparisons are orthogonal.

Table 6. Coefficients for two orthogonal comparisons.

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	1	0.5
	Low Self-Esteem	0.5	-1	-0.5
Failure	High Self-Esteem	-0.5	-1	0.5
	Low Self-Esteem	-0.5	1	-0.5

Table 7 shows two comparisons that are not orthogonal. The first compares the high-self-esteem subjects to the low-self-esteem subjects; the second considers only those in the success group and compares high-self-esteem subjects to low-self-esteem subjects. The failure group is ignored by using 0's as coefficients. Clearly the comparison of high-self-esteem subjects to low-self-esteem subjects for the whole sample is not independent of the comparison for the success group only. You can see that the sum of the products of the coefficients is 0.5 and not 0.

Table 7. Coefficients for two non-orthogonal comparisons.

Outcome	Esteem	C1	C2	Product
Success	High Self-Esteem	0.5	0.5	0.25
	Low Self-Esteem	-0.5	-0.5	0.25
Failure	High Self-Esteem	0.5	0.0	0.0
	Low Self-Esteem	-0.5	0.0	0.0

Difference Between Two Means (Correlated Pairs)

by David M. Lane

Prerequisites

- Chapter 4: Values of the Pearson Correlation
- Chapter 10: t Distribution
- Chapter 11: Hypothesis Testing
- Chapter 12: Testing a Single Mean
- Chapter 12: Difference Between Two Means (Independent Groups)

Learning Objectives

1. Determine whether you have correlated pairs or independent groups
2. Compute a t test for correlated pairs

Let's consider how to analyze the data from the "ADHD Treatment" case study. These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. In this section, we will be concerned only with testing the difference between the mean of the placebo (D0) condition and the mean of the highest dosage condition (D60). The first question is why the difference between means should not be tested using the procedure described in the section Difference Between Two Means (Independent Groups). The answer lies in the fact that in this experiment we do not have independent groups. The scores in the D0 condition are from the same subjects as the scores in the D60 condition. There is only one group of subjects, each subject being tested in both the D0 and D60 conditions.

Figure 1 shows a scatter plot of the 60-mg scores (D60) as a function of the 0-mg scores (D0). It is clear that children who get more correct in the D0 condition tend to get more correct in the D60 condition. The correlation between the two conditions is high: $r = 0.80$. Clearly these two variables are not independent.

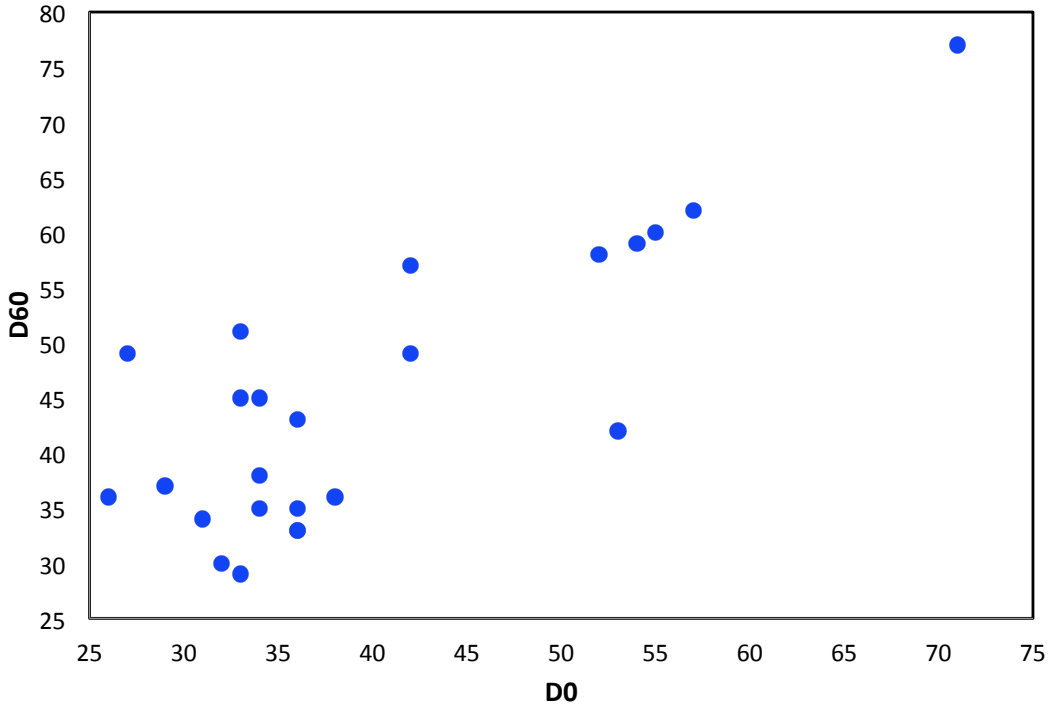


Figure 1. Number of correct responses made in the 60-mg condition as a function of the number of correct responses in the 0-mg condition.

Computations

You may recall that the method to test the difference between these means was presented in the section on "Testing a Single Mean." The computational procedure is to compute the difference between the D60 and the D0 conditions for each child and test whether the mean difference is significantly different from 0. The difference scores are shown in Table 1. As shown in the section on testing a single mean, the mean difference score is 4.96 which is significantly different from 0: $t = 3.22$, $df = 23$, $p = 0.0038$. This t test has various names including "*correlated t test*" and "*related-pairs t test*."

In general, the correlated t test is computed by first computing the differences between the two scores for each subject. Then, a test of a single mean is computed on the mean of these difference scores.

Table 1. DOG scores as a function of dosage.

D0	D60	D60-D0
57	62	5
27	49	22

32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

If you had mistakenly used the method for an independent-groups t test with these data, you would have found that $t = 1.42$, $df = 46$, and $p = 0.15$. That is, the difference between means would not have been found to be statistically significant. This is a typical result: correlated t tests almost always have greater power than independent-groups t tests. This is because in correlated t tests, each difference score is a comparison of performance in one condition with the performance of that same subject in another condition. This makes each subject "their own control" and

keeps differences between subjects from entering into the analysis. The result is that the standard error of the difference between means is smaller in the correlated t test and, since this term is in the denominator of the formula for t, results in a larger t.

Details about the Standard Error of the Difference between Means (Optional)

To see why the standard error of the difference between means is smaller in a correlated t test, consider the variance of difference scores. As shown in the section on the Variance Sum Law, the variance of the sum or difference of the two variables X and Y is:

$$s^2_{X \pm Y} = s^2_X + s^2_Y \pm 2rs_X s_Y$$

Therefore, the variance of difference scores is the variance in the first condition (X) plus the variance in the second condition (Y) minus twice the product of (1) the correlation, (2) the standard deviation of X, and (3) the standard deviation of Y. For the current example, r = 0.80 and the variances and standard deviations are shown in Table 2.

Table 2. Variances and Standard Deviations

	D0	D60	D60 - D0
Variance	128.02	151.78	56.82
Sd	11.31	12.32	7.54

The variance of the difference scores of 56.82 can be computed as:

```
128.02 + 151.78 - (2)(0.80)(11.31)(12.32)
```

which is equal to 56.82 except for rounding error. Notice that the higher the correlation, the lower the standard error of the mean.

Specific Comparisons (Correlated Observations)

by David M. Lane

Prerequisites

- Chapter 10: t Distribution
- Chapter 12: Hypothesis Testing, Testing a Single Mean
- Chapter 12: Specific Comparisons
- Chapter 12: Difference Between Two Means (Correlated Pairs)

Learning Objectives

1. Determine whether to use the formula for correlated comparisons or independent-groups comparisons
2. Compute t for a comparison for repeated-measures data

In the "Weapons and Aggression" case study, subjects were asked to read words presented on a computer screen as quickly as they could. Some of the words were aggressive words such as injure or shatter. Others were control words such as relocate or consider. These two types of words were preceded by words that were either the names of weapons, such as shotgun or grenade, or non-weapon words, such as rabbit or fish. For each subject, the mean reading time across words was computed for these four conditions. The four conditions are labeled as shown in Table 1. Table 2 shows the data from five subjects.

Table 1. Description of Conditions.

Variable	Description
aw	The time in milliseconds (msec) to name an aggressive word following a weapon word prime.
an	The time in milliseconds (msec) to name an aggressive word following a non-weapon word prime.
cw	The time in milliseconds (msec) to name a control word following a weapon word prime.
cn	The time in milliseconds (msec) to name a control word following a non-weapon word prime.

Table 2. Data from Five Subjects

Subject	aw	an	cw	cn
1	447	440	432	452
2	427	437	469	451
3	417	418	445	434
4	348	371	353	344
5	471	443	462	463

One question was whether reading times would be shorter when the preceding word was a weapon word (aw and cw conditions) than when it was a non-weapon word (an and cn conditions). In other words, is

$$L_1 = (an + cn) - (aw + cw)$$

greater than 0? This is tested for significance by computing L_1 for each subject and then testing whether the mean value of L_1 is significantly different from 0. Table 3 shows L_1 for the first five subjects. L_1 for Subject 1 was computed by

$$L_1 = (440 + 452) - (447 + 432) = 892 - 879 = 13$$

Table 3. L1 for Five Subjects

Subject	aw	an	cw	cn	L1
1	447	440	432	452	13
2	427	437	469	451	-8
3	417	418	445	434	-10
4	348	371	353	344	14
5	471	443	462	463	-27

Once L_1 is computed for each subject, the significance test described in the section "Testing a Single Mean" can be used. First we compute the mean and the standard error of the mean for L_1. There were 32 subjects in the experiment. Computing L_1 for the 32 subjects, we find that the mean and standard error of the mean are 5.875 and 4.2646, respectively. We then compute

$$t = \frac{M - \mu}{s_M}$$

where M is the sample mean, μ is the hypothesized value of the population mean (0 in this case), and s_M is the estimated standard error of the mean. The calculations show that t = 1.378. Since there were 32 subjects, the degrees of freedom is 32 - 1 = 31. The t distribution calculator shows that the two-tailed probability is 0.178.

A more interesting question is whether the priming effect (the difference between words preceded by a non-weapon word and words preceded by a weapon word) is different for aggressive words than it is for non-aggressive words. That is, do weapon words prime aggressive words more than they prime non-aggressive words? The priming of aggressive words is (an - aw). The priming of non-aggressive words is (cn - cw). The comparison is the difference:

L_2 = (an - aw) - (cn - cw).

Table 4 shows L_2 for five of the 32 subjects.

Table 4. L2 for Five Subjects

Subject	aw	an	cw	cn	L2
1	447	440	432	452	-27
2	427	437	469	451	28
3	417	418	445	434	12
4	348	371	353	344	32
5	471	443	462	463	-29

The mean and standard error of the mean for all 32 subjects are 8.4375 and 3.9128, respectively. Therefore, t = 2.156 and p = 0.039.

Multiple Comparisons

Issues associated with doing multiple comparisons are the same for related observations as they are for multiple comparisons among independent groups.

Orthogonal Comparisons

The most straightforward way to assess the degree of dependence between two comparisons is to correlate them directly. For the weapons and aggression data, the comparisons L_1 and L_2 are correlated 0.24. Of course, this is a sample correlation and only estimates what the correlation would be if L_1 and L_2 were correlated in the population. Although mathematically possible, orthogonal comparisons with correlated observations are very rare.

Pairwise Comparisons (Correlated Observations)

by David M. Lane

Prerequisites

- Chapter 12: Difference between Two Means (Independent Groups)
- Chapter 12: All Pairwise Comparisons Among Means
- Chapter 12: Difference Between Two Means
- Chapter 12: Difference Between Two Means (Correlated Pairs)
- Chapter 12: Specific Comparisons (Independent Groups)
- Chapter 12: Specific Comparisons (Correlated Observations)

Learning Objectives

1. Compute the Bonferroni correction
2. Calculate pairwise comparisons using the Bonferroni correction

In the section on all pairwise comparisons among independent groups, the *Tukey HSD test* was the recommended procedure. However, when you have one group with several scores from the same subjects, the Tukey test makes an assumption that is unlikely to hold: The variance of difference scores is the same for all pairwise differences between means.

The standard practice for pairwise comparisons with *correlated observations* is to compare each pair of means using the method outlined in the section "Difference Between Two Means (Correlated Pairs)" with the addition of the *Bonferroni correction* described in the section "Specific Comparisons." For example, suppose you were going to do all pairwise comparisons among four means and hold the familywise error rate at 0.05. Since there are six possible pairwise comparisons among four means, you would use 0.05/6 = 0.0083 for the per-comparison error rate.

As an example, consider the case study "Stroop Interference." There were three tasks each performed by 47 subjects. In the "words" task, subjects read the names of 60 color words written in black ink; in the "color" task, subjects named the colors of 60 rectangles; in the "interference" task, subjects named the ink color of 60 conflicting color words. The times to read the stimuli were recorded. In order to compute all pairwise comparisons, the difference in times for each pair of conditions for each subject is calculated. Table 1 shows these scores for five of the 47 subjects.

Table 1. Pairwise Differences

W-C	W-I	C-I
-3	-24	-21
2	-41	-43
-1	-18	-17
-4	-23	-19
-2	-17	-15

The means, standard deviations (Sd), and *standard error of the mean* (Sem), t, and p for all 47 subjects are shown in Table 2. The t's are computed by dividing the means by the standard errors of the mean. Since there are 47 subjects, the degrees of freedom is 46. Notice how different the standard deviations are. For the Tukey test to be valid, all population values of the standard deviation would have to be the same.

Table 2. Pairwise Comparisons.

Comparison	Mean	Sd	Sem	t	p
W-C	-4.15	2.99	0.44	-9.53	<0.001
W-I	-20.51	7.84	1.14	-17.93	<0.001
C-I	-16.36	7.47	1.09	-15.02	<0.001

Using the Bonferroni correction for three comparisons, the p value has to be below 0.05/3 = 0.0167 for an effect to be significant at the 0.05 level. For these data, all p values are far below that, and therefore all pairwise differences are significant.

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 12: Single Mean

Research on the effectiveness of surgery for weight loss reported here found that "The surgery was associated with significantly greater weight loss [than the control group who dieted] through 2 years (61.3 versus 11.2 pounds, $p<0.001$)."

What do you think?

What test could have been used and how would it have been computed?

> For each subject a difference score between their initial weight and final weight could be computed. A t test of whether the mean difference score differs significantly from 0 could then be computed. The mean difference score will equal the difference between the mean weight losses of the two groups (61.3 - 11.2 = 50.1).

References

Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Exercises

Prerequisites

- All material presented in the Testing Means chapter

1. The scores of a random sample of 8 students on a physics test are as follows: 60, 62, 67, 69, 70, 72, 75, and 78.

   a. Test to see if the sample mean is significantly different from 65 at the .05 level. Report the t and p values.

   b. The researcher realizes that she accidentally recorded the score that should have been 76 as 67. Are these corrected scores significantly different from 65 at the .05 level?

2. A (hypothetical) experiment is conducted on the effect of alcohol on perceptual motor ability. Ten subjects are each tested twice, once after having two drinks and once after having two glasses of water. The two tests were on two different days to give the alcohol a chance to wear off. Half of the subjects were given alcohol first and half were given water first. The scores of the 10 subjects are shown below. The first number for each subject is their per- formance in the "water" condition. Higher scores reflect better performance. Test to see if alcohol had a significant effect. Report the t and p values.

water	alcohol
16	13
15	13
11	10
20	18
19	17
14	11
13	10
15	15
14	11
16	16

3. The scores on a (hypothetical) vocabulary test of a group of 20 year olds and a group of 60 year olds are shown below.

20 yr olds	60 yr olds
27	26

26	29
21	29
24	29
15	27
18	16
17	20
12	27
13	

a. Test the mean difference for significance using the .05 level.

b. List the assumptions made in computing your answer.

4. The sampling distribution of a statistic is normally distributed with an estimated standard error of 12 (df = 20). (a) What is the probability that you would have gotten a mean of 107 (or more extreme) if the population parameter were 100? Is this probability significant at the .05 level (two-tailed)? (b) What is the probability that you would have gotten a mean of 95 or less (one-tailed)? Is this probability significant at the .05 level? You may want to use the t Distribution calculator for this problem.

5. How do you decide whether to use an independent groups t test or a correlated t test (test of dependent means)?

6. An experiment compared the ability of three groups of subjects to remember briefly-presented chess positions. The data are shown below.

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

a. Using the Tukey HSD procedure, determine which groups are significantly different from each other at the .05 level.

b. Now compare each pair of groups using t-tests. Make sure to control for the familywise error rate (at 0.05) by using the Bonferroni correction. Specify the alpha level you used.

7. Below are data showing the results of six subjects on a memory test. The three scores per subject are their scores on three trials (a, b, and c) of a memory task. Are the subjects get- ting better each trial? Test the linear effect of trial for the data.

a	b	c
4	6	7
3	7	8
2	8	5
1	4	7
4	6	9
2	4	2

a. Compute L for each subject using the contrast weights -1, 0, and 1. That is, compute (-1)(a) + (0)(b) + (1)(c) for each subject.

b. Compute a one-sample t-test on this column (with the L values for each subject) you created.

8. Participants threw darts at a target. In one condition, they used their preferred hand; in the other condition, they used their other hand. All subjects performed in both conditions (the order of conditions was counterbalanced). Their scores are shown below.

Preferred	Non-preferred
12	7
7	9
11	8
13	10
10	9

a. Which kind of t-test should be used?

b. Calculate the two-tailed t and p values using this t test.

c. Calculate the one-tailed t and p values using this t test.

9. Assume the data in the previous problem were collected using two different groups of subjects: One group used their preferred hand and the other group used

their non-preferred hand. Analyze the data and compare the results to those for the previous problem.

10. You have 4 means, and you want to compare each mean to every other mean. (a) How many tests total are you going to compute? (b) What would be the chance of making at least one Type I error if the Type I error for each test was .05 and the tests were independent? (c) Are the tests independent and how does independence/non-independence affect the probability in (b).

11. In an experiment, participants were divided into 4 groups. There were 20 participants in each group, so the degrees of freedom (error) for this study was 80 - 4 = 76. Tukey's HSD test was performed on the data. (a) Calculate the p value for each pair based on the Q value given below. You will want to use the Studentized Range Calculator. (b) Which differences are significant at the .05 level?

Comparison of Groups	Q
A - B	3.4
A - C	3.8
A - D	4.3
B - C	1.7
B - D	3.9
C - D	3.7

12. If you have 5 groups in your study, why shouldn't you just compute a t test of each group mean with each other group mean?

13. You are conducting a study to see if students do better when they study all at once or in intervals. One group of 12 participants took a test after studying for one hour continuously. The other group of 12 participants took a test after studying for three twenty minute sessions. The first group had a mean score of 75 and a variance of 120. The second group had a mean score of 86 and a variance of 100.

    a. What is the calculated t value? Are the mean test scores of these two groups significantly different at the .05 level?

    b. What would the t value be if there were only 6 participants in each group? Would the scores be significant at the .05 level?

14. A new test was designed to have a mean of 80 and a standard deviation of 10. A random sample of 20 students at your school take the test, and the mean score turns out to be 85. Does this score differ significantly from 80?

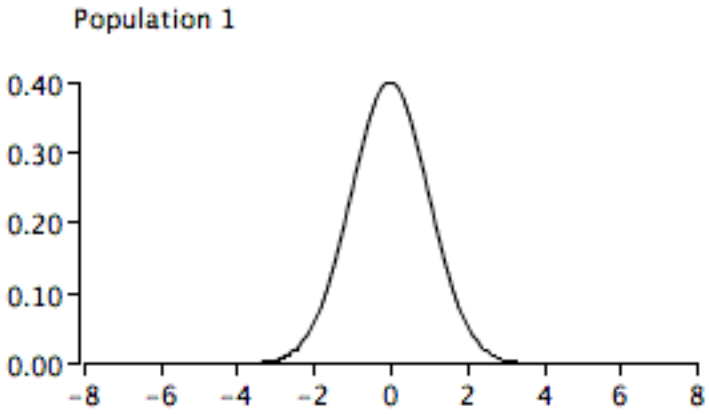
15. You perform a one-sample t test and calculate a t statistic of 3.0. The mean of your sample was 1.3 and the standard deviation was 2.6. How many participants were used in this study?

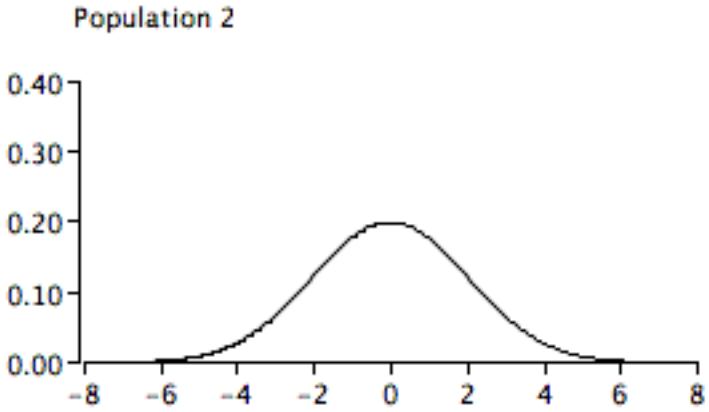
16. True/false: The contrasts (-3, 1 1 1) and (0, 0 , -1, 1) are orthogonal.

17. True/false: If you are making 4 comparisons between means, then based on the Bonferroni correction, you should use an alpha level of .01 for each test.

18. True/false: Correlated t tests almost always have greater power than independent t tests.

19. True/false: The graph below represents a violation of the homogeneity of variance assumption.





20. True/false: When you are conducting a one-sample t test and you know the population standard deviation, you look up the critical t value in the table based on the degrees of freedom.

Questions from Case Studies

Angry Moods (AM) case study

21. (AM) Do athletes or non-athletes calm down more when angry? Conduct a t test to see if the difference between groups in Control-In scores is statistically significant.

22. (AM) Do people in general have a higher Anger-Out or Anger-In score? Conduct a t test on the difference between means of these two scores. Are these two means independent or dependent?

Smiles and Leniency (SL) case study

23. (SL) Compare each mean to the neutral mean. Be sure to control for the familywise error rate.

24. (SL) Does a "felt smile" lead to more leniency than other types of smiles? (a) Calculate L (the linear combination) using the following contrast weights false: -1, felt: 2, miserable: -1, neutral: 0. (b) Perform a significance test on this value of L.

Animal Research (AR) case study

25. (AR) Conduct an independent samples t test comparing males to females on the belief that animal research is necessary.

26. (AR) Based on the t test you conducted in the previous problem, are you able to reject the null hypothesis if alpha = 0.05? What about if alpha = 0.1?

27. (AR) Is there any evidence that the t test assumption of homogeneity of variance is violated in the t test you computed in #25?

ADHD Treatment (AT) case study

28. (AT) Compare each dosage with the dosage below it (compare d0 and d15, d15 and d30, and d30 and d60). Remember that the patients completed the task after every dosage. (a) If the familywise error rate is .05, what is the alpha level you will use for each comparison when doing the Bonferroni correction? (b) Which differences are significant at this level?

29. (AT) Does performance increase linearly with dosage?

    a. Plot a line graph of this data.

    b. Compute L for each patient. To do this, create a new variable where you multiply the following coefficients by their corresponding dosages and then sum up the total: (-3)d0 + (-1)d15 + (1)d30 + (3)d60 (see #7). What is the mean of L?

    c. Perform a significance test on L. Compute the 95% confidence interval for L.