Effect Sizes: Sections 41-45

How to Calculate Cohen's d

How to Calculate Cliff's delta

How to Calculate Risk, Odds and the Odds Ratio

How to Calculate Eta-Squared and Omega-Squared

How to Calculate Effect Sizes Relating to Correlation and Regression

Section 41: How to Calculate Cohen's d
Effect Size for Difference between Two Means

(Uses data files: Life Exp by Gender – nonpaired.txt, Life Exp by Gender – paired.txt)

Cohen's d is a commonly used value for measuring the effect size when you are dealing with the difference between two means.

Independent Sample Case

This example is one that deals with two independent samples. The null hypothesis is that the mean male life expectancy and the mean female life expectancy are equal. You can find the details of the initial test in Section 16: "How to Run Two Sample T-tests: Independent Samples." Only the output is repeated here for reference. The level of significance was alpha (α) = .05.

```
Welch Two Sample t-test

data:  Years by Gender
t = 9.9765, df = 35.6, p-value = 7.479e-12        ← Test statistic is 9.9765

                                                    df is calculated by a more complicated formula
                                                    than in most textbooks

                                                    p-value is 7.5 x 10^-12, which is less than α

alternative hypothesis: true difference in means is   ← Indicates a two-tailed test: one tail would
                                not equal to 0          specify "greater than" or "less than"
```

Based on the small p-value, you would reject the hypothesis of equal means. In that case, you believe the mean life expectancies are different, but you would like a measure of the effect size. That is, is the difference small, medium or large? Therefore, you probably want to calculate a measure called "Cohen's d" that is used for this purpose, and then "classify" your effect size.

To calculate Cohen's d, you should install (if necessary) and load the R package: effsize. This process is explained in Section3: "How to Find, Install and Load R Packages." Once you have loaded the package, you need the following command:

```
> cohen.d (Years, Gender, pooled = TRUE, paired = FALSE)
```

The first variable, Years, is the one whose difference in means is being measured. Gender is the grouping factor. You want to tell R to pool the standard deviations of the two groups; that is done by the subcommand "pooled = TRUE." Finally, since the data were not paired, and you specify that in the last subcommand: "paired = FALSE." The resulting output is below.

```
Cohen's d
d estimate: -3.154861 (large)
95 percent confidence interval:
    inf         sup
-4.113865   -2.195857
```

As you can see, the output is giving you an estimate for Cohen's d around -3.15. This is classified (by its magnitude, ignoring the sign) as a large effect size.

IMPORTANT NOTE: You should treat the size classification carefully. What is large, medium or small depends heavily on the context of the situation. A large effect in one situation may be relatively minor and unimportant in another. The classification is only a rule-of-thumb and should not be taken as absolute.

Dependent Sample Case (Paired Data)

This example is one that deals with paired data. In this example, the male and female life expectancies are matched pairs by county of residence. The null hypothesis is again that both genders have the same mean life expectancy, or equivalently, that the difference in their life expectancy is zero. The details of the test can be found in Section 17: "How to Run Two Sample T-Tests with Paired Data." Only the output is repeated here for reference. The level of significance used was alpha (α) = .05.

```
    One Sample t-test

data:  Difference
t = 24.9683, df = 23, p-value < 2.2e-16        ← Test statistic is 24.9683.
                                                  df=number of pairs – 1 = 23
                                                  The p-value = 2.2x10^-16, which is less than α.

alternative hypothesis: true mean is not equal to 0        ← Indicates that this is a two-tail test.
```

This p-value is less than α, leading you to reject the null hypothesis. Therefore, in this example also, you have reason to believe that the mean life expectancy differs by gender. You probably want to calculate Cohen's d. If you have not done so, install and load the package: effsize. Note that the subcommand indicating paired data is now set as TRUE.

```
> cohen.d (Male, Female, pooled = TRUE, paired = TRUE)
```

Here is the output.

```
Cohen's d
d estimate: -5.14207 (large)
95 percent confidence interval:
    inf          sup
-6.347724   -3.936416
```

As you can see, the effect size is classified (by its magnitude, ignoring the sign) as "large."

IMPORTANT NOTE: Again, you should treat the size classification carefully. What is large, medium or small depends heavily on the context of the situation. A large effect in one situation may be relatively minor and unimportant in another. The classification is only a rule-of-thumb and should not be taken as absolute.

Section 42: How to Calculate Cliff's Delta
Effect Size for Difference between Two Medians

(Uses data files: Life Exp by Gender – nonpaired.txt, Life Exp by Gender – paired.txt)

Cliff's Δ is a commonly used value for measuring the effect size when you are dealing with the difference between two medians.

Independent Sample Case

This example is one that deals with two independent samples. The null hypothesis is that the median male life expectancy and the median female life expectancy are equal. You can find the details of the initial test in Section 31: "How to Run a Mann-Whitney-Wilcoxon Test for Two Independent Samples." Only the output is repeated here for reference. The level of significance was alpha (α) = .05.

```
Wilcoxon rank sum test

data:  Years by Gender

W = 392, p-value = 2.038e-07          ← Small p-value in scientific notation; means .0000002038

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:  4.100038  6.

sample estimates:
difference in location          5.200001
```

Based on the p-value being less than α, you would reject the hypothesis of equal medians. That means that you believe the median life expectancies are different, but you would probably like a measure of the effect size. That is, is the difference small, medium or large? Therefore, you want to calculate a measure called "Cliff's Δ" that is used for this purpose, and then "classify" your effect size.

To calculate Cliff's Δ, you should install (if not already installed) and load the R package: effsize. This process is explained in Section 3: "How to Find, Install and Load R Packages." Once you have loaded the package, you need the following command:

```
> cliff.delta (Years~Gender, use.unbiased = TRUE)
```

The first variable, Years, is the one whose difference in medians is being measured. Gender is the grouping factor. The last part of the command, "use.unbiased = TRUE," refers to the method for calculating the variance of Cliff's Δ ; an extensive discussion of this is beyond the scope of this Manual.

The resulting output gives an estimate for Cliff's Δ around 0.96, which is classified as a large effect size.

```
Cliff's Delta

delta estimate: 0.96 (large)
95 percent confidence interval:
     inf       sup
0.7784777 0.9933347
```

IMPORTANT NOTE: You should treat the size classification carefully. What is large, medium or small depends heavily on the context of the situation. A large effect in one situation may be relatively minor or unimportant in another. The classification is only a rule-of-thumb and should not be taken as absolute.

Dependent Samples Case (Paired Data)

This example is one that deals with paired data. The null hypothesis was that the median male life expectancy and the median female life expectancy are equal. In this example, the male and female life expectancies are paired by county.

You can find the details of the initial test in Section 32: "How to Run a Repeated Measures Mann-Whitney-Wilcoxon Test." Only the output is repeated here for reference. The level of significance was set at alpha (α) = .05.

```
Wilcoxon signed rank test

data:  Male and Female

V = 0, p-value = 1.804e-05          ← Small p-value in scientific notation; meaning .00001804

alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(Male, Female, alternative = "two.sided",  :
cannot compute exact p-value with ties
```

Since the p-value is less than α, you conclude that you should reject the hypothesis of equal medians. Since you believe the evidence suggests that the life expectancy medians differ by gender, you would probably want to know whether the difference is large, medium or small. You can use Cliff's Δ to classify the effect size.

You need to install (if not already present) and load the package: effsize. The details of this process are explained in Section 3: "How to Find, Install and Load R Packages." Then the necessary command is modified somewhat from the version in the first example. Here is what you want; note that you have specified that the data is "paired" in this version of the command.

```
> cliff.delta (Female, Male, paired = TRUE, use.unbiased = TRUE)
```

The resulting output is as follows; it gives an estimate for Cliff's Δ around 0.93, which is classified as a large effect size. Again, treat the classification carefully. See the "Important Note" at the end of the first example.

```
Cliff's Delta

delta estimate: 0.9340278 (large)
95 percent confidence interval:
     inf       sup
0.7711221 0.9821525
```

Section 43: How to Calculate Risk, Odds and the Odds Ratio
Effect Size for Proportions
(Uses no data files)

Risk and odds are commonly used values for measuring the effect size when you are dealing with single population proportions.

EXAMPLE A (Single proportion) Refer to Example A in Section 29: "How to Test a Hypothesis about a Proportion or Comparing Two Proportions." The example dealt with a biased coin that came up heads 35 out of 50 times in a sample. The null hypothesis was that the coin produces heads 65% of the time; the alternative was the it produces heads more than 65% of the time. The example used level of significance alpha (α) = .05.

The output is reproduced below for reference.

Exact binomial test

data: 35 and 50

number of successes = 35, number of trials = 50, p-value = 0.2801 ← p-value is greater than α

alternative hypothesis: true probability of success is greater than 0.65 ← Indicates upper tail test

95 percent confidence interval: ← You can be 95% confident that the actual proportion of
0.576267 1.000000 heads produced by this coin is at least 57.5%.

sample estimates:
probability of success ← Calculated sample percentage of heads is 70%.
0.7

Based on the large p-value, you would probably not go any further with this. But suppose that you want to "quantify" the amount of bias of the coin anyhow. From the output, you can see that the coin produced a sample proportion of 70% heads; in other words, the probability of getting heads is P(heads) = 0.70.

This probability is also known as the "risk" of getting heads. The terminology sounds a bit peculiar here, but it seems more appropriate when you are dealing medical questions, such as the probability of getting a particular disease. Therefore, using the "risk" terminology, the result is that:

The "risk" of getting heads = P(heads) = 0.70.
The "risk" of getting tails = P(tails) = 0.30.

Alternatively, you may want to express the "odds" of getting heads. In general:

$$\text{Odds of an event occurring} = \frac{\text{Risk or probability of event occurring}}{\text{Risk or probability of event not occurring}}$$

So for this example, the odds of getting heads = 0.70/0.30 = 2.33. This means that, using that particular coin, the sample indicates that you are 2.33 times as likely to get heads as you are to get tails on a toss.

Now consider the situation when you are dealing with two samples. The odds ratio is frequently used when dealing with two proportions. Essentially, you compute the risk and then the odds for the sample from each population. Then calculate:

$$\text{Odds ratio of event} = \frac{\text{Odds of event in sample from first population}}{\text{Odds of event in sample from second population}}$$

EXAMPLE B (Proportions from two populations) Refer to Example B in Section 29: "How to Test a Hypothesis about a Proportion or Comparing Two Proportions." This example deals with two biased coins. Coin One produced heads on 28 out of 40 tosses. Coin Two produced heads on 26 out of 40 tosses. A proportion test was run to test the claim that Coin One is "more biased" than Coin Two. The level of significance was alpha (α)= .05.

The output is reproduced here for reference.

2-sample test for equality of proportions with continuity correction

data: heads out of samplesize

X-squared = 0.057, df = 1, p-value = 0.4057 ← X-squared is the test statistic
p-value is greater than α

alternative hypothesis: greater ← Indicates upper tail test

95 percent confidence interval: ← You can be 95% sure that the proportions of heads
-0.1470229 1.0000000 for the two coins differs by between 0 and 1 (not a particularly useful confidence interval in this case)

sample estimates:
prop 1 prop 2 ← Calculated sample proportions
0.70 0.65

Using the sample proportions, P(heads for Coin One) = 0.70 and P(heads for Coin Two) = 0.65.

The calculations then proceed as follows.

1. As shown above, the odds of heads for Coin One = 2.33.
2. Do similar calculations for Coin Two.
   "Risk" of heads for Coin Two = P(heads for Coin Two) = 0.65.
   "Risk" of tails for Coin Two = P(tails for Coin Two) = 1 – 0.65 = 0.35.
   Odds of heads for Coin Two = 0.65/0.35 = 1.86.
3. So the odds ratio of heads for these coins (in order) is: 2.33/1.86 = 1.25.

Section 44: How to Calculate Eta-Squared and Omega-Squared Effect Size for Differences among Several Means

(Uses data files: Anxiety.txt, AnxietyRepeat.txt)

Eta-squared is a commonly used value for measuring the effect size when you are dealing with the differences among multiple means. It is generally used following an ANOVA when a difference has been detected.

Note: Eta-squared applies only to the sample and SHOULD NOT BE USED for inferences about effect size in the populations. See additional comments concerning omega-squared after the two examples below.

Eta-Squared for a One-Way ANOVA

This example deals with four independent samples. It was used before in Section 21: "How to Do One-Way ANOVA," where you can look at the data set if you wish. It deals with Anxiety Scores for students from four different types of educational institutions. The null hypothesis was that the mean scores are the same for students from all four types. The alternative was that at least one type has a different mean. The test used alpha (α) = .05. The test and its output are repeated here for reference.

```
> AnxModel = lm (AnxScore ~ Type)
> anova (AnxModel)          ← Note the model has been named AnxModel. That is needed
                              below in the code for finding eta-squared.
```

The output from the analysis of variance follows.

```
Analysis of Variance Table

Response: AnxScore

           Df  Sum Sq  Mean Sq  F value  Pr(>F)     ← Test statistic is the F value.
Type        3   261.2   87.076   0.8578  0.4645     ← p-value is 0.4645.
Residuals 152 15429.9  101.513
```

Since the p-value is larger than α, you fail to reject the null hypothesis. The interpretation is that the evidence is not sufficient to conclude that any type of institution has a different mean anxiety score.

However, for purposes of illustration, suppose that the p-value had been small and the null hypothesis was rejected. Then you would believe that at least one mean was different from the rest, but you would not have a measure of how large the difference was. That is, you would need a measure of effect size. The usual one to use is eta-squared.

To calculate eta-squared, you should install (if not already installed) and load the R package: lsr. Refer to Section 3: "How to Find, Install and Load R Packages." Also note that this package is available, but as of this writing, is reportedly still under development. Its author advises caution when using it.

Once you have installed and loaded the package, you need the following command:

```
> etaSquared (AnxModel, type=2, anova=TRUE)
```

In general, the items you have to supply in the command are:

-- The first item is the name of the model that you created when you ran the original ANOVA.
-- The "type" refers to how a particular sum of squares is calculated; you can leave it as "2" as long as you are not dealing with interactions.
-- The last item tells R to display the ANOVA test as well as the effect size. You do not have to do this since you have already run an ANOVA, but it doesn't hurt to leave that option as "TRUE."

Here is the output.

	eta.sq	eta.sq.part	SS	df	MS	F	p
Type	0.01664814	0.01664814	261.2284	3	87.07615	0.8577864	0.4645137
Residuals	0.98335186	NA	15429.9190	152	101.51262	NA	NA

As you can see, the output matches that from the earlier ANOVA, except that two columns called "etq.sq" and "eta.sq.part" have been added. The second one is irrelevant, since the ANOVA was only a one-way model. Thus, the effect size for Type's contribution to the Anxiety Scores is about .0166.

The R output does not attempt to classify the effect size as small, medium or large. A set of "rules of thumb" for classifying eta-squared is the following:

0.01 = small effect 0.06 = medium effect 0.14 = large effect

(Source: http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize)

Using these as guidelines, you would probably classify the effect size here as being "small." That is what you would expect, since the original ANOVA indicated there was no significant difference in the means.

IMPORTANT NOTE: You should treat the size classification carefully. What is large, medium or small depends heavily of the context of the situation. A large effect in one situation may be relatively minor and unimportant in another. The classification is only a rule-of-thumb and should not be taken as absolute.

Eta-Squared for a Two-Way ANOVA (Repeated Measures Type)

This example also deals with Anxiety Scores, but the State variable is no longer used. Instead, each student has been tested three times in three different test sessions. Therefore, this is a repeated measures model.

This example was in Section 22: " How to Run a Repeated Measures ANOVA," where you can examine the data set if you wish. The null hypothesis was that the mean scores are the same for all test sessions. The code for the test and its output are repeated here for reference. The test used alpha (α) = .05.

```
> Repeat.Model = lm (Anx.Score ~ Test.Session + ID)  ← Creates a linear model with Anx.Score as a
                                                       function of Test.Session and ID (individual))
```

The following output results from the analysis of variance.

Analysis of Variance Table
Response: Anx.Score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Test.Session	2	383.0	191.509	2.1971	0.11874
ID	35	5235.9	149.597	1.7162	0.02781 *
Residuals	70	6101.6	87.166		

Now, as it turns out, you see that the p-value for Test.Session is greater than α, so test session shows no significant effect on the mean score. But if you had found significant results, you would probably want to measure the effect size(s).

As in the first example, if not already done, you would install and load the package: lsr. Then type the following.

```
> etaSquared (Repeat.Model, type=2, anova=TRUE)
```

The output is as follows.

	eta.sq	eta.sq.part	SS	df	MS	F	p
Test.Session	0.03267924	0.05906526	383.0185	2	191.5093	2.197054	0.11873606
ID	0.44672659	0.46181846	5235.8796	35	149.5966	1.716218	0.02780888
Residuals	0.52059418	NA	6101.6481	70	87.1664	NA	NA

Since you have two independent variables (Test.Session and ID), you want to look at the partial eta-squared column. The column labelled eta.sq.part gives these values. The partial eta-squared value for Test.Session is 0.059, which would round to a medium effect size based on the "rules of thumb" given above.

IMPORTANT NOTE: You should treat the size classification carefully. What is large, medium or small depends heavily on the context of the situation. A large effect in one situation may be relatively minor and unimportant in another. The classification is only a rule-of-thumb and should not be taken as absolute.

Comment: Eta-squared has been the most commonly used effect size associated with ANOVA. However, as noted at the start of the section, you should not use it to infer effect size for the larger population(s) involved. An alternative effect size which you may want to use is called "omega-squared." Both omega-squared and eta-squared may be found in another R package: sjstats.

The following code should work using sjstats on the data from the first example. The details, however, are not included here. Please note that the first three lines in the code are comment lines, written there instead of explanatory text. You do not actually need them to find omega-squared.

```
> # Install and load the package:  sjstats
> # The first time, when you install it, this may take a while.  Be patient.
> # Repeat first example using this package; it has both eta-squared and omega-squared calculations.

> Data = read.table ("E:/Data Files/Anxiety.txt", header = TRUE)
> attach (Data)
> Data
> AnxModel = lm (AnxScore ~ Type)
> anova (AnxModel)
> #If not already done, install and load package:  sjstats
> eta_sq (AnxModel, partial = FALSE)
> omega_sq (AnxModel)
```

Section 45: How to Calculate Effect Sizes Related to Correlation and Regression

This is largely handled by R automatically when the regression analysis is run. See the sections:

-- How to Run a Simple Linear Regression (Section 26)
-- How to Check Pairs of Values for Correlation Non-Parametrically (Section 37)
-- How to Run a Multiple Regression (Section 28)
-- How to Run a Binary Logistic Regression (Section 38)

Without repeating the details of each application, the results are summarized again here for reference.

Simple Linear Regression: The example involves finding a linear relationship between height and weight. Before starting on the regression, the Pearson's correlation coefficient was obtained.

```
> cor (Height, Weight)
```

R returned the following sample correlation coefficient.

```
0.9540717
```

This is a strong correlation. Also, a hypothesis test in Section 26 tests this value for significance. Another measure used is "R-squared," which is straightforward to calculate once you know the value of Pearson's correlation coefficient – just square it. Here the value is approximately 0.954*0.954 = 0.91. This means, roughly speaking, that 91% of the variability in Weight can be explained by its linear relationship with Height.

Multiple Regression: The example involves trying to find a linear equation the describes the severity of a cough among influenza patients, based on age and smoking status. The results are shown again here for reference; the details are in Section 28.

```
lm(formula = Cough.Severity ~ Age + Smoking)
Residuals:
   Min      1Q  Median      3Q     Max
-5.6874 -1.0199  0.2798  1.2971  2.7338

Coefficients:              ← Based on the following, the equation is:
                             Cough.Severity = 5.8 - .013 Age + .495 Smoking
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  5.813861   0.121409    47.886   < 2e-16 ***
Age         -0.013217   0.004639    -2.849   0.00457 **
Smoking      0.495108   0.245090     2.020   0.04393 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.719 on 475 degrees of freedom
Multiple R-squared:  0.0194,    Adjusted R-squared:  0.01527
F-statistic: 4.698 on 2 and 475 DF,  p-value: 0.009538
```

This time, the multiple R-squared and adjusted R-squared are the relevant measures. They are both small (only between 1 % and 2%), indicating that the model does little to explain cough severity.

Binary Logistic Regression: This example also deals with cough, but does not try to "rate" its severity. The only objective is to state the binary result: "yes, the cough is severe" (denoted 1) or "no, the cough is not severe" (denoted 0). Two models were run in Section 38 on binary regression. The first used all available input variables; the second used only age and smoking status. Part of the output is a quantity labelled AIC. This stands for Aikeke's Information Criterion.

If you look back at the output in the Section 38, you will see that:

First Model AIC = 234.65

Second Model AIC = 230.62.

Generally, the smaller value for the AIC, the better. More details about the AIC are beyond the scope of this Manual.