PSYC 6780: Lab Tutorial 6

Chris M. Fiacconi

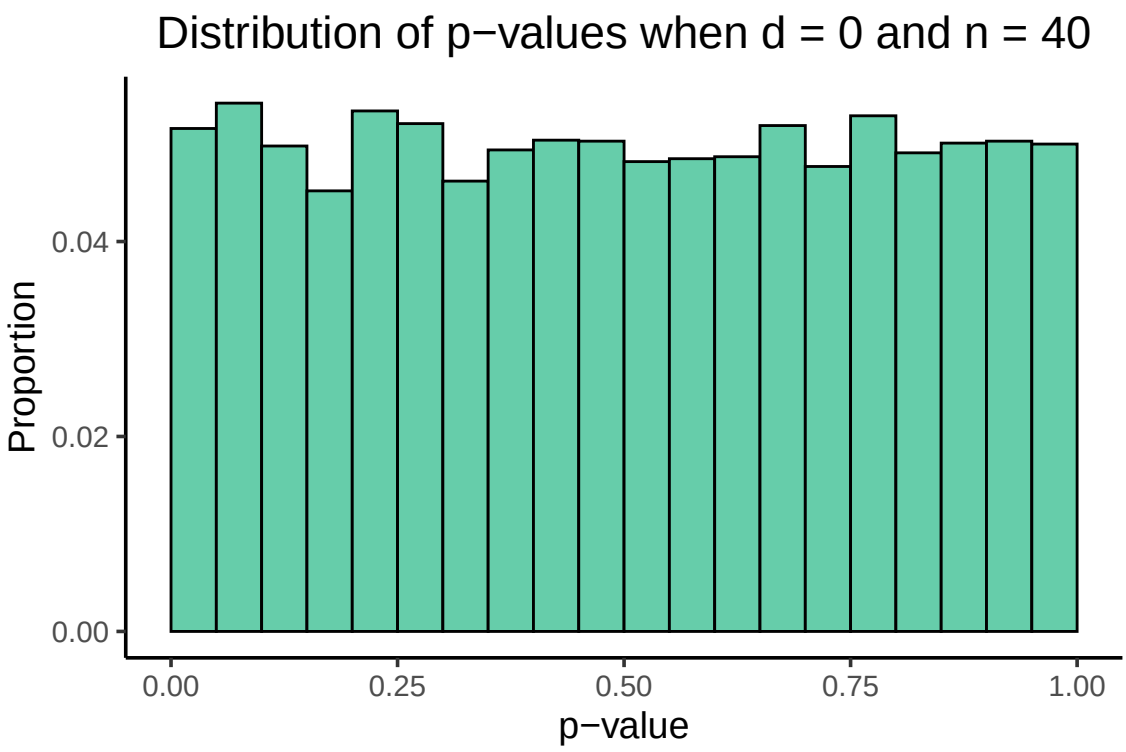
Power and the Distribution of *p*-values

You might expect that if the null hypothesis were false, you should obtain a p-value $< .05$ each time you run a replication of a given study. In fact, the proportion of times you would obtain such a p-value can be surprisingly low, and is entirely dependent on power.

Conversely, you might expect that if the null hypothesis were true, you should obtain a p-value $> .05$ each time you run a replication of a given study. In fact, you are no more likely to obtain a p-value $= .75$ than you are a p-value $= .01$.

These observations can be confirmed by running a series of simulations and varying the effect size and sample size parameters to obtain a series of p-values from an independent-groups t-test. First, let's set Cohen's $d = 0$ (null is true) and get the distribution of p-values that would be obtained from running a t-test 10000 times.

```
# Create function to return p-values after running t-test
sim_p<-function(d,nsubj){
  A<-rnorm(n=nsubj,mean=0,sd=1)
  B<-rnorm(n=nsubj,mean=d,sd=1)
  return(t.test(x=A,y=B,paired=F)$p.value)
}

d<-0 # Null hypothesis is true
nsubj<-40 # Per group sample size
n_sims<-10000

p_outcomes<-data.frame(p=replicate(n=n_sims,sim_p(d,nsubj))) # Run simulation 10000 times

hist_title<-paste("Distribution of p-values when d =",d,"and n =",nsubj) # Histogram title

library(tidyverse)

# Make histogram
ggplot(p_outcomes,aes(x=p)) +
  geom_histogram(aes(y=..count../sum(..count..)),breaks=seq(0,1,.05),color="black",
                 fill="aquamarine3") +
  theme_classic(base_size=14) +
  labs(x="p-value",y="Proportion") +
  ggtitle(hist_title) +
  theme(plot.title = element_text(hjust = 0.5))
```

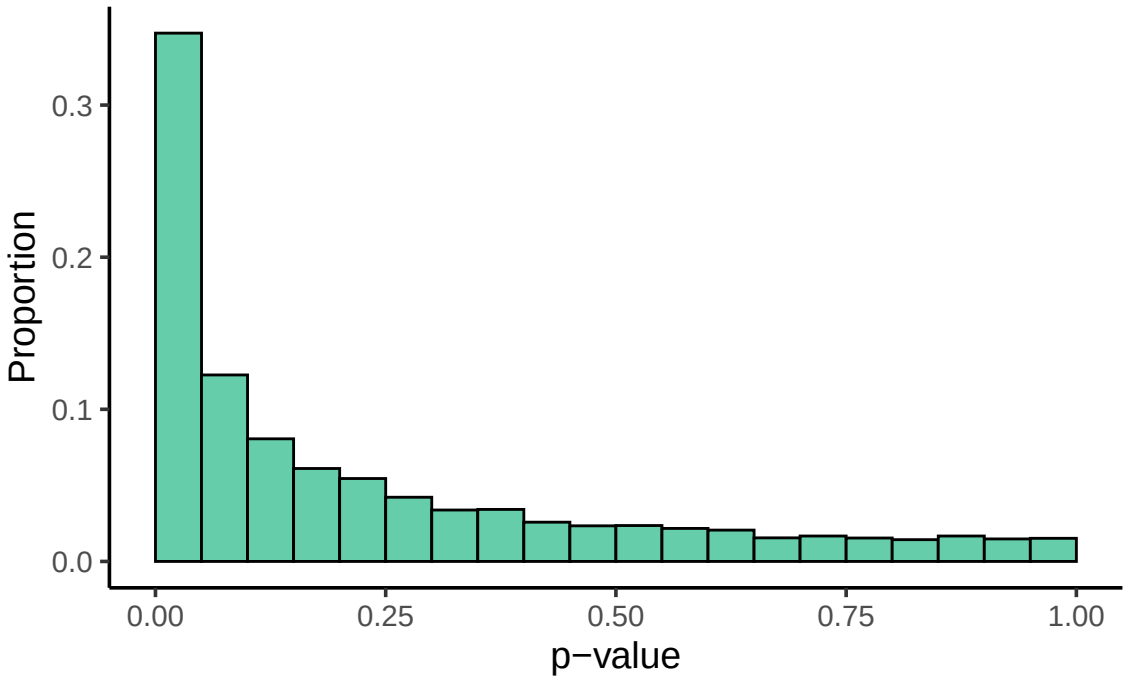



```
# Proportion of 'significant' p-values
prop_sig<-length(p_outcomes$p[p_outcomes$p<.05])/length(p_outcomes$p)
print(prop_sig)
```

```
## [1] 0.0516
```

As can be seen above, all p-values are equally likely under the null hypothesis. Now let's see what happens to the distribution of p-values when the null is false and Cohen's $d = .35$.

```
d<-.35 # Null hypothesis is false - d = .35 (small-medium effect)
nsubj<-40 # Per group sample size
n_sims<-10000

p_outcomes<-data.frame(p=replicate(n=n_sims,sim_p(d,nsubj))) # Run simulation 10000 times

hist_title<-paste("Distribution of p-values when d =",d,"and n =",nsubj) # Histogram title

# Make histogram
ggplot(p_outcomes,aes(x=p)) +
  geom_histogram(aes(y=..count../sum(..count..)),breaks=seq(0,1,.05),color="black",
                 fill="aquamarine3") +
  theme_classic(base_size=14) +
  labs(x="p-value",y="Proportion") +
  ggtitle(hist_title) +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Proportion of 'significant' p-values
prop_sig<-length(p_outcomes$p[p_outcomes$p<.05])/length(p_outcomes$p)
print(prop_sig)
```

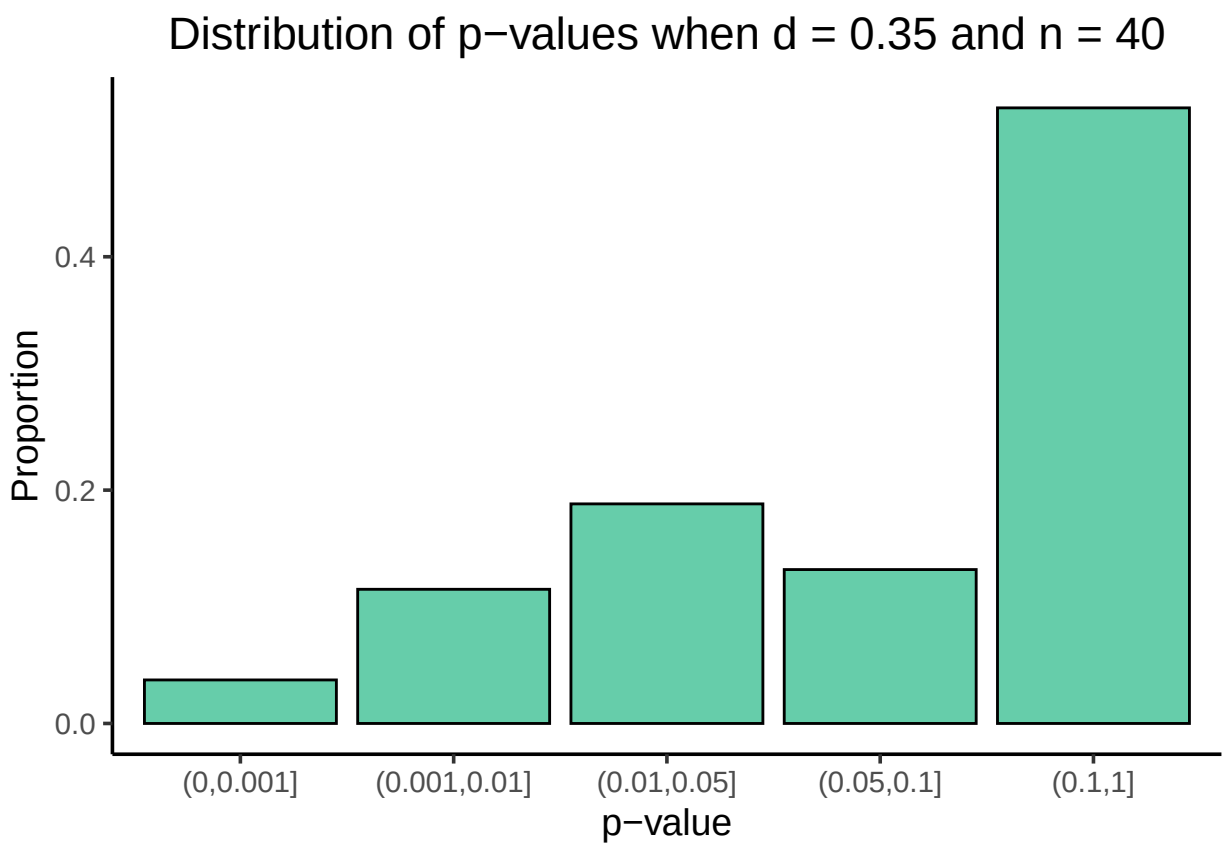
```
## [1] 0.3473
```

```
# Get actual power
library(pwr)
pwr.t.test(n=nsubj,d=d,power=NULL,type="two.sample",alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 40
##               d = 0.35
##       sig.level = 0.05
##           power = 0.3396731
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
d<-.35
nsubj<-40
n_sims<-10000

p_outcomes<-data.frame(p=replicate(n=n_sims,sim_p(d,nsubj)))

p_outcomes<-p_outcomes %>% mutate(bin=cut(x=p,breaks=c(0,.001,.01,.05,.10,1)))
```

```r
p_summary<-p_outcomes %>% group_by(bin) %>% summarize(prop=n()/n_sims)

hist_title<-paste("Distribution of p-values when d =",d,"and n =",nsubj)

ggplot(p_summary,aes(x=bin,y=prop)) +
  geom_bar(stat="identity",color="black",fill="aquamarine3") +
  theme_classic(base_size=14) +
  labs(x="p-value",y="Proportion") +
  ggtitle(hist_title) +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Proportion of 'significant' p-values
prop_sig<-length(p_outcomes$p[p_outcomes$p<.05])/length(p_outcomes$p)
print(prop_sig)
```

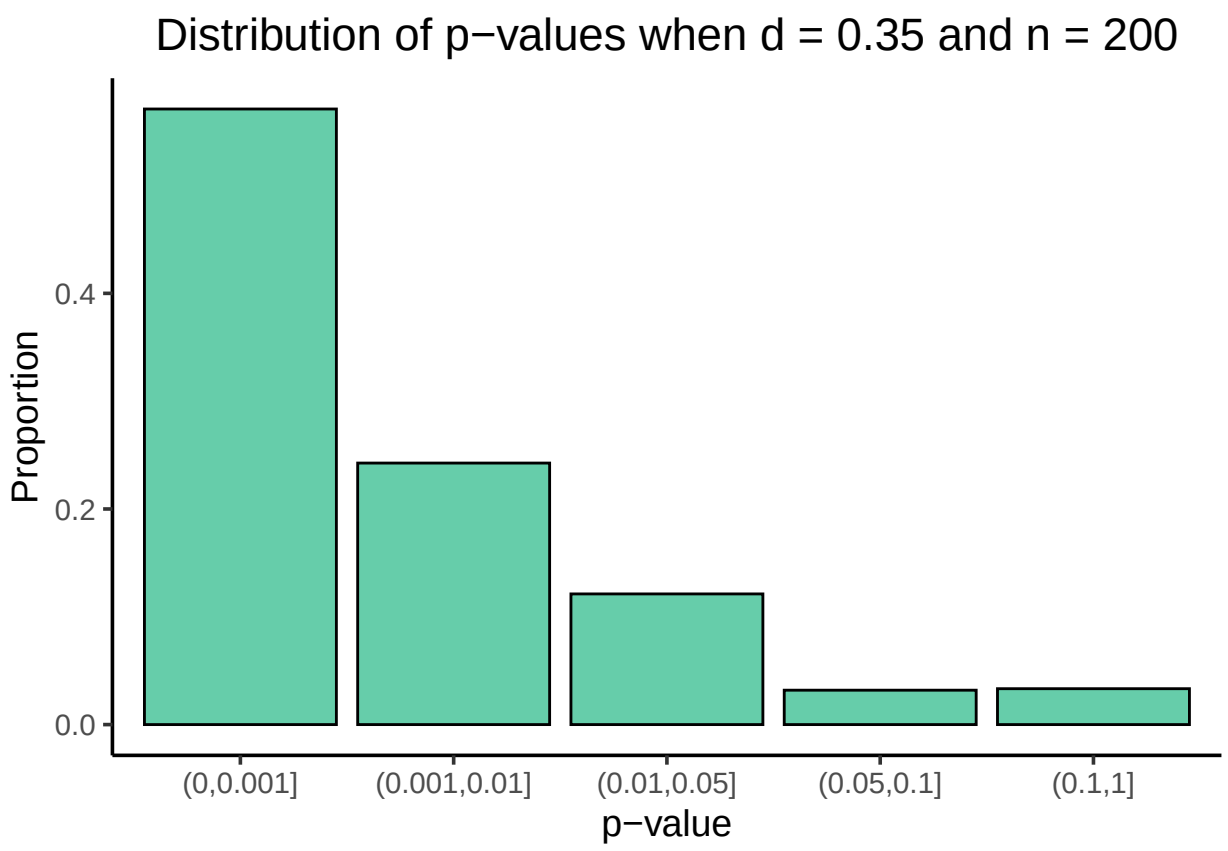
```
## [1] 0.3405
```

```r
# Get actual power
pwr.t.test(n=nsubj,d=d,power=NULL,type="two.sample",alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 40
##               d = 0.35
```

```
##       sig.level = 0.05
##           power = 0.3396731
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
d<-.35
nsubj<-200
n_sims<-10000

p_outcomes<-data.frame(p=replicate(n=n_sims,sim_p(d,nsubj)))

p_outcomes<-p_outcomes %>% mutate(bin=cut(x=p,breaks=c(0,.001,.01,.05,.10,1)))
p_summary<-p_outcomes %>% group_by(bin) %>% summarize(prop=n()/n_sims)

hist_title<-paste("Distribution of p-values when d =",d,"and n =",nsubj)

ggplot(p_summary,aes(x=bin,y=prop)) +
  geom_bar(stat="identity",color="black",fill="aquamarine3") +
  theme_classic(base_size=14) +
  labs(x="p-value",y="Proportion") +
  ggtitle(hist_title) +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Proportion of 'significant' p-values
prop_sig<-length(p_outcomes$p[p_outcomes$p<.05])/length(p_outcomes$p)
print(prop_sig)
```

```
## [1] 0.9348
```

```
# Get actual power
pwr.t.test(n=nsubj,d=d,power=NULL,type="two.sample",alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 200
##               d = 0.35
##       sig.level = 0.05
##           power = 0.9371867
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Meta-Analysis

Meta-Analysis is a method that is used to synthesize the findings from multiple studies. One of it's many purposes is to estimate the reliability of an effect by amalgamating the results across many studies investigating a particular phenomenon. In this section, we will cover the basics of meta-analysis including:

- effect size calculation and variability
- models of meta-analysis
- testing for heterogeneity among effect sizes
- summarizing the analysis with forest plots and funnel plots
- evaluating the influence of moderating variables

Effect Size Calculation and Associated Variability

The basic unit of a meta-analysis is an *effect size*. Effect sizes in their simplest form measure the difference between two conditions or groups of interest, and provide an index of *how big* the effect is. Effect sizes can be expressed in raw units, or more typically as a standardized mean difference. The latter form is critical when the goal is to compare effects across studies in which the scale of the dependent measure varies. As such, we will focus our tutorial on standardized effect sizes as these are the most commonly reported measures of effect size.

The most common standardized effect size is Cohen's d, which is also known as the *standardized mean difference*. Before we can proceed with a meta-analysis, we must compile the effect sizes from all the studies that we wish to include in our analysis. Unfortunately, Cohen's d is not always reported, and so we must calculate it based on other information.

Independent-Groups Design

For a simple between-subjects design with two groups, the formula is simply

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{within}}$$

where S_{within} is equal to:

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Recall that the formula for an independent-groups t-test (with equal n groups) is very similar to the above formula for Cohen's d:

$$t(N_1 + N_2 - 2) = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{within} \times \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

Therefore, we can easily convert a reported t-statistic into a Cohen's d effect size estimate:

$$d = t\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Now that we have calculated our measure of effect size, Cohen's d, we need to know how much uncertainty there is in our estimate of δ, the true population effect size that we're estimating with d. In other words, what we want to know is:

$$SE_d = \sqrt{v_d}$$

where v_d is equal to:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

Once we have d and SE_d, we can calculate a confidence interval capturing the likely range of values within which our true population effect size (δ) is likely to fall. I will show you how to perform these calculations using the *escalc* function in the *metafor* package for *R*.

```
library(readxl)
library(metafor)

# load in data file with relevant values for each study
JOLWF.meta<-read_xls("JOL_WFmeta.xls")
JOLWF.meta<-JOLWF.meta[,c(-12,-13)]

# Calculate effect sizes and variances assuming between-subjects design
res<-escalc(measure="SMD",m1i=m1i,m2i=m2i,sd1i=sd1i,sd2i=sd2i,n1i=ni,n2i=ni,data=JOLWF.meta)
```

The last two columns in the *res* variable, *yi* and *vi* are the effect size (Cohen's d) estimates and corresponding variances, respectively. We'll use these values when we talk about fitting models to the effect sizes.

Paired Design

The majority of studies in cognitive psychology do not utilize between-subject experimental designs. Instead, we typically use within-subject designs wherein each subject is exposed to each experimental condition of interest. With this design, the basic formula for Cohen's d holds:

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{within}}$$

but the calculation of S_{within} differs:

$$S_{within} = \frac{S_{Difference}}{\sqrt{2(1 - r)}}$$

Note that in this calculation we need to know $S_{Difference}$ which can be calculated based on the reported paired t-test for the conditions of primary interest:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\frac{S_{Difference}}{\sqrt{N}}}$$

and so therefore,

$$d = t\sqrt{\frac{2(1-r)}{n}}$$

We also need r which is the correlation between participant's scores across conditions. Unfortunately, r is almost never reported, and so we need to impute this value based on a reasonable approximation from other data (it helps if you have the raw data on hand, as this value can then be computed easily).

As was the case with the between-subject design, we also need to calculate the standard error of this effect size estimate:

$$SE_d = \sqrt{v_d}$$

where v_d is equal to:

$$v_d = (\frac{1}{n} + \frac{d^2}{2n})2(1-r)$$

You'll notice that the r value is needed again for this calculation.

Although the *escalc* function will correctly estimate the effect sizes and associated variance calculations based on group means, n's and standard deviations for between-subject designs, these values must be calculated 'by hand' for within-subject designs, as the equations used by the *escalc* function are inconsistent with the preferred approach to estimating these values (**Note:** there is controversy over how to best measure effect sizes in within-subject designs - I'm showing you how to do it based on the recommendations provided by Cumming, 2017).

Here is how you could program a loop to calculate S_{Within} for each study assuming that you know (or can estimate with precision) r, n, $S_{Difference}$, and the means (M) for each condition and that all of these values are included as their own column in your spreadsheet.

```
nStudies<-17

# Calculate S_within based on equations from Borenstein (2009)
s.within<-vector(length=nStudies) # initialize vector to store S_within for each study

for (i in 1:nStudies){
  s.within[i]<-JOLWF.meta$sddiff[i]/sqrt(2*(1-JOLWF.meta$ri[i]))
}

# Compute point estimates of Cohen's d based on S_within (Borenstein, 2009)
yi<-vector(length=nStudies) # initialize vector to store Cohen's d for each study

for (i in 1:nStudies){
  yi[i]<-(JOLWF.meta$m1i[i]-JOLWF.meta$m2i[i])/s.within[i]
}

# Compute estimates of sampling variance (vi) based on Borenstein (2009)
vi<-vector(length=nStudies) # initialize vector to store vi for each study

for (i in 1:nStudies){
  vi[i]<-(1/JOLWF.meta$ni[i]+yi[i]^2/(2*JOLWF.meta$ni[i]))*2*(1-JOLWF.meta$ri[i])
}
```

Now that we have our effect size estimates (yi) and their corresponding variances (vi) we can proceed to the modelling stage.

Models of Meta-Analysis

Meta-analyses combine effect sizes across studies using one of two models: *fixed-effect* models and *random-effect* models. We will now discuss each in turn.

Fixed-Effect Models

The basic premise of fixed-effect models is that each effect size estimate associated with each study is an estimate of a common underlying population effect size, δ. Therefore, variation among effect sizes is modeled simply as random sampling error. One consequence of such models is that the resulting model parameters apply *only* to the studies included in the fixed-effect meta-analysis. Therefore, use of fixed-effect models, while intuitively simple, limits our ability to generalize our findings to all possible studies.

Random-Effect Models

Random-effect models, on the other hand, do not assume that each effect size is estimating a common underlying population effect size. Rather, such models allow for there to be multiple *different* population effect sizes such that different effect size estimates may be estimating different population effect sizes. This aspect of random-effect models acknowledges that while some of the variation among effect sizes is due to random sampling error, there may also exist **systematic** differences between effect sizes that reflect the presence of moderator variables. The presence of this latter form of variability can be explicitly evaluated by testing for **heterogeneity** among effect sizes. The presence of significant heterogeneity would suggest that, in addition to random sampling variability, there may also be many different moderator variables are pushing the effect sizes around, and that there exist important differences between the studies in terms of methods/procedures/populations, etc.

Using a random-effect model also allows us to generalize our findings to other studies that may not have been included in the meta-analysis, as well as to future studies that fit the criteria used to select studies for the meta-analysis.

My own thinking on this issue is that one should *always* use random-effect models when conducting a meta-analysis. In the presence of significant heterogeneity among studies, random-effect models will be more conservative. When heterogeneity is not present, the results from the random- and fixed-effect models are largely the same.

Fortunately, fitting a random-effect model is easy with the *metafor* package. We will fit our model using a restricted maximum likelihood estimator (REML) using the *rma* function.

```
re.model<-rma(yi,vi,method="REML")
print(re.model)
```

```
##
## Random-Effects Model (k = 17; tau^2 estimator: REML)
##
## tau^2 (estimated amount of total heterogeneity): 0.1058 (SE = 0.0429)
## tau (square root of estimated tau^2 value):      0.3253
## I^2 (total heterogeneity / total variability):   91.94%
## H^2 (total variability / sampling variability):  12.41
##
## Test for Heterogeneity:
```

```
## Q(df = 16) = 176.3669, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.2339  0.0847  2.7634  0.0057  0.0680  0.3999  **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

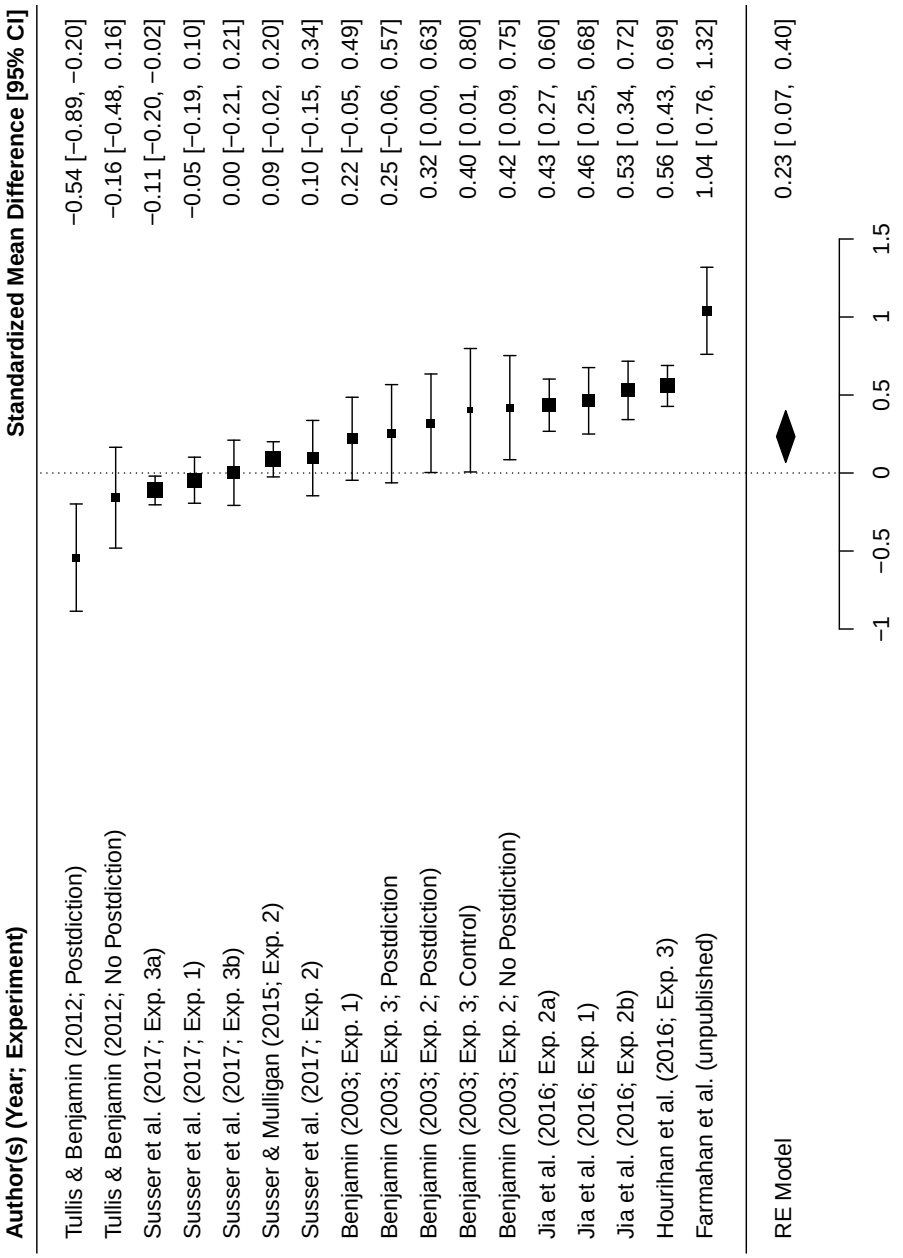
You can see from the model ouput that the overall effect of WF on JOLs was significant, $g = .23$, 95% CI [.07,.40]. Note that the overall effect size estimate here is actually *Hedges g*. That is because Cohen's d is a biased estimator of the true population effect size, and so it is corrected to an unbiased estimator, *Hedges g*.

More importantly however, the test for heterogeneity was also significant, $Q(16) = 176.37, p < .001$, meaning that the variation in effect sizes was greater than what we would expect by chance alone. In fact, approximately 92% of the total variability among effect sizes can be attributed to heterogeneity!! Therefore, it would appear as though there are systematic differences between studies that are responsible for the majority of the total variability among effect sizes, and that several moderator variables likely exist. The challenge then, is to identify what those systematic differences are, and test whether they do in fact moderate effect size magnitudes.

Before we look at how to include moderators in our meta-analysis, let's generate a forest plot to depict our results.

Visualizing the Results: Forest Plots & Funnel Plots

```
par(mar=c(2,.5,.5,.5)) # Set plot margins
par(font=1) # Change font quality

forest(re.model,xlim=c(-5,3),slab=JOLWF.meta$Study,
       order="obs",xlab="Standardized Mean Difference")

par(font=2) # bold titles

# Position labels on the figure
text(-5,nStudies+1.5,"Author(s) (Year; Experiment)",pos=4)
text(1.6,nStudies+1.5,"Standardized Mean Difference [95% CI]")
```



The forest plot depicts all the effect sizes considered in our meta-analysis, along with the 95% CIs on each effect size. Note that the average population effect size is shown as a diamond at the bottom of the plot, with the edges of the diamond representing the 95% CI. If the 95% CI does not include zero, then the overall effect size estimate is significant (i.e, a reliable difference exists). The average effect size is calculated as a weighted sum of all the individual effect sizes, where the weight assigned to each effect size is inversely proportional to its variance:

$$w_i = 1/v_i$$

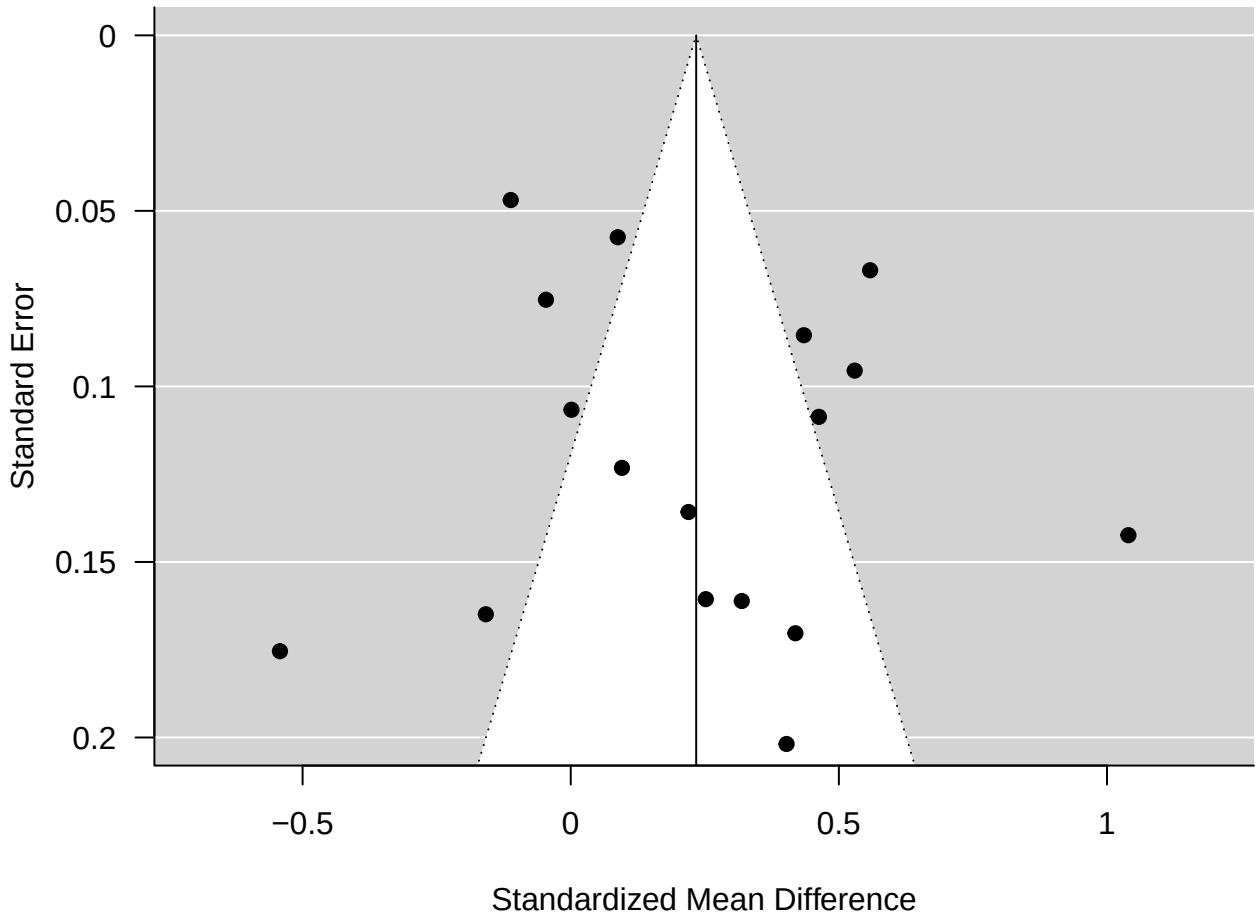
In a random-effect model, the variance of the overall effect size estimate is the sum of the heterogeneity among effect sizes σ_θ, plus random sampling variability:

$$v_\cdot = \sigma_\theta + \frac{1}{\sum_{i=1}^{k}(\frac{1}{v_i})}$$

For this reason, random-effect models tend to be more conservative than fixed-effect models *when there is significant heterogeneity among effect sizes.*

Another way to visualize the results of your meta-analysis is by using a *funnel* plot. A funnel plot depicts each effect size as a function of its associated standard error, and is useful for evaluating publication bias - the tendency to report only those findings that are statistically significant. Here is how you could make a funnel plot:

```
par(mar=c(4,4,1,2))
myTicks<-c(0,.05,.10,.15,.20)
funnel(re.model,las=1,xlab="Standardized Mean Difference",ylim=range(myTicks))
```



Testing the Influence of Moderator Variables

Given that our random-effect model detected significant heterogeneity (92%!!), there are likely many moderator variables that could explain some or all of this heterogeneity. We will evaluate whether two potential moderators, namely, whether participants produced a response to items prior to making their JOL (Response), and expected test type (Recognition vs. Recall; Test Format) can explain variation among the observed effect sizes.

```
# Fit Random-Effects Model with Response Moderator
JOLWF.meta$response<-as.factor(JOLWF.meta$response)
```

```r
JOLWF.meta$Format<-as.factor(JOLWF.meta$Format)

re.model.response<-rma.uni(yi,vi,mods=~JOLWF.meta$response,method="REML")
print(re.model.response)
```

```
##
## Mixed-Effects Model (k = 17; tau^2 estimator: REML)
##
## tau^2 (estimated amount of residual heterogeneity):     0.0794 (SE = 0.0347)
## tau (square root of estimated tau^2 value):             0.2817
## I^2 (residual heterogeneity / unaccounted variability): 88.94%
## H^2 (unaccounted variability / sampling variability):   9.04
## R^2 (amount of heterogeneity accounted for):            25.02%
##
## Test for Residual Heterogeneity:
## QE(df = 15) = 79.1213, p-val < .0001
##
## Test of Moderators (coefficient(s) 2):
## QM(df = 1) = 4.5487, p-val = 0.0329
##
## Model Results:
##
##                          estimate      se     zval    pval    ci.lb
## intrcpt                    0.3442  0.0909   3.7863  0.0002   0.1660
## JOLWF.meta$responseYes    -0.3412  0.1600  -2.1328  0.0329  -0.6547
##                           ci.ub
## intrcpt                  0.5223  ***
## JOLWF.meta$responseYes  -0.0276    *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Fit Random-Effects Model with Test Format Moderator
re.model.format<-rma.uni(yi,vi,mods=~JOLWF.meta$Format,method="REML")
print(re.model.format)
```

```
##
## Mixed-Effects Model (k = 17; tau^2 estimator: REML)
##
## tau^2 (estimated amount of residual heterogeneity):     0.1040 (SE = 0.0436)
## tau (square root of estimated tau^2 value):             0.3225
## I^2 (residual heterogeneity / unaccounted variability): 92.00%
## H^2 (unaccounted variability / sampling variability):   12.50
## R^2 (amount of heterogeneity accounted for):            1.71%
##
## Test for Residual Heterogeneity:
## QE(df = 15) = 175.0922, p-val < .0001
##
## Test of Moderators (coefficient(s) 2):
## QM(df = 1) = 1.4354, p-val = 0.2309
##
## Model Results:
```

```
##
##                               estimate      se     zval    pval    ci.lb
## intrcpt                         0.3039  0.1023   2.9702  0.0030   0.1034
## JOLWF.meta$FormatRecognition   -0.2148  0.1793  -1.1981  0.2309  -0.5663
##                                 ci.ub
## intrcpt                        0.5044  **
## JOLWF.meta$FormatRecognition   0.1366
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the moderator analysis suggest that the Response variable accounts for a significant proportion of the heterogeneity among effect sizes, but that the Test Format variable does not.