Heterogeneity

Introduction to variation

The goal of this document is to get intimate with variation and variance. This is because we will then move on to understanding heterogeneity and to do that we need to be on very solid grounds regarding variance, since they are literally synonymous in meta-analysis. Let's start with the very basics and then move on to more complex calculations.

Let's start by creating an entire population.

```
population <- c(2,5,1,5,7,3,7,1)
```

This is our population. Variance is the average squared distances from the mean of the population, that is, the sum of the squared residuals divided by the population size. Let's calculate the variance of our population using the formula:

$$\sigma^2 = \frac{\sum x - \mu}{N}$$

```
pop_var <- (sum((mean(population)-population)^2))/(length(population))
pop_var
```

```
[1] 5.359375
```

Nice, now consider that we take a sample from our population, the sample variance will be different than the variance within the population.

```
set.seed(6539)
our_sample <- sample(population, 5, replace = F)
```

We use the same formula but include Bessel's correction by subtracting 1 from the size of our sample.

$$sd^2 = \frac{\sum x - \mu}{n - 1}$$

```
(sum((mean(our_sample)-our_sample)^2))/(length(our_sample)-1)
```

```
[1] 2.8
```

Notice how the current estimated variance is much lower than the true variance in the population. This is because the population is small and somewhat diverse. The variation/preposition with which we estimate the mean is known as the standard error, or the standard deviation of the sampling distribution. Our accuracy will increase as our sample size increases. But when we only have a small population, our sample can only be so big. Let's do the same thing we did previously but on a larger scale.

```
new_population <- rnorm(mean = mean(population),
                        sd   = sqrt(pop_var),
                        n    = 1e5)
mean(population)
```

```
[1] 3.875
```

Let's take a random sample of 30 observations from the population. And compare the sampled mean from the true mean.

```
new_sample <- sample(new_population, size = 30, replace = F)
mean(new_sample)
```

```
[1] 3.855494
```

Pretty close, we can also estimate the standard error to get an approximation of the precision of our mean.

```
sd(new_sample)/sqrt(length(new_sample))
```

```
[1] 0.4542812
```

If we take our estimate =/- the standard error we can get a pretty good picture of the population mean. Now let's estimate the variation again.

```
new_pop_var <- (sum((mean(new_population)-new_population)^2))/(length(new_population))
new_pop_var
```

```
[1] 5.359762
```

Let's compare it to our estimation of the variance from our sample, let's use the var() function for some brevity.

```
var(new_sample)
```

```
[1] 6.191143
```

Our estimated variation(average squared distance from the mean) is smaller than that of the population. Larger samples will provide more accurate estimations. Let's finish of by taking a larger sample, say n = 100.

```
set.seed(675)
big_sample <- sample(new_population, size = 100, replace = F)

mean(big_sample)
```

```
[1] 3.777773
```

```
mean(new_population)
```

```
[1] 3.877644
```

Quite close, let's compare variation as well.

```
var(big_sample)
```

```
[1] 5.52892
```

```
new_pop_var
```

```
[1] 5.359762
```

```r
sd(big_sample)
```

```
[1] 2.351366
```

```r
sqrt(new_pop_var)
```

```
[1] 2.315116
```

Yup, very close. Let's move on to a more complex problem - meta analysis.

Meta-analysis

A fair assumption in most meta-analyses is that the true population effect of some stimuli actually consists of a subgroup of true effects that together constitute an overall true effect. That is, the population effect has variation, or heterogeneity. This variability can stem from many sources but for the moment we will not concern ourselves with *why* heterogeneity is present. Note also that we move from having a complete population, to having a population that is studies, meaning that the population is generated from an unobservable hyper-parameter. This makes it so that we never can have a definite answer - however, that is almost never the case anyways. Our data points are now a number of studies containing an effect size and a variation. Let's generate them ourselves.

```r
set.seed(673)
k_studies <- 10

es <- rnorm(k_studies, 1, 1)
sigma <- runif(k_studies, .3, 1)
```

The next step is to calculate the inverse variance weights. We do this to give more prominence/weight to studies with less variation.

```r
weights <- 1/sigma
```

Next we calculate a pooled effect size using our observed effects and our variance dependent weights. we do this by dividing the sum of the weighted effects with the sum of the weights - making it ''unbiased".

```r
pooled_effect <- sum(weights*es)/sum(weights)
```

Next we can calculate the heterogeneity, the believe the simplest way is with DerSimonian and Laird's functions. We do this by computing a Q statistic by taking the sum of the weights times the squared difference between our observed effects and our pooled effect. Then we subtract Q by our number of studies to then divide it by the sum of our weights to get tau^2, for tau we simply take the square root of our calculation.

$$\tau^2_{DL} = max(0, \frac{Q - (n-1)}{S_1 - S_2/S_1})$$

I believe that the denominator here is simply the sum of the weights, but i could be wrong. That is, the S denominator is the gained using the following formula:

$$S_r = \sum_{i=1}^{n} w_i^r$$

Higgins and Thompson gives the following formula for $\hat{\tau}^2_{DL}$:

$$\hat{\tau}^2_{DL} = \frac{Q - (k-1)}{\sum w_i - \sum w_i^2 / \sum w_i}$$

```
Q <- sum(weights * (es - pooled_effect)^2)
tau <- (Q - (k_studies-1))/sum(weights)

tau <- sqrt((Q - (k_studies-1))/sum(weights))
tau_denom <-sum(weights) - ((sum(weights^2)) / sum(weights))
tau_higgins <- (Q - (k_studies-1))/tau_denom
```

Great, we now have an estimation of the variation in the effect. We have large variation here, though given that the sigma we specified ranged from 1 to 3 with a true effect size of N(1, 1). Now we only need to check if we are correct, that can easily be done using the metafor package.

```
library(metafor)

tau_2 <- rma.uni(yi = es, vi = sigma, method = 'DL',
                 weights = weights)

sqrt(tau_2$tau2)
```

```
[1] 1.261183
```

hmm, neither are the same. I was pretty sure it would match up. It could be due to rounding, but that seems like a stretch. My calculation is probably wrong - must have missed something. It was pretty hard to get tau from the original paper but the silver lining is that the Q is the same in both estimates. Let's do a little simulation so see how the hand calculation and the metafor calculation relates to each other, it could shed some light about what is potentially wrong in my formula.

```
tau_data <- data.frame(hand_calc = NA, metafor_calc = NA,
                       hand_q = NA, meta_q = NA )

set.seed(679)
k_studies <- 10
true_es <- .5
sigma <- runif(k_studies, .3, 1)
nsim <- 1000

for (i in 1:nsim) {
  es <- rnorm(k_studies, true_es, 1)
  sigma <- runif(k_studies, .3, 1)
  weights <- 1/sigma
  pooled_effect <- sum(weights*es)/sum(weights)
  hand_q <- sum(weights * (es - pooled_effect)^2)

  hand_calc <- sqrt((hand_q - (k_studies-1))/sum(weights))

  tau_2 <- rma.uni(yi = es, vi = sigma, method = 'DL',
                   weights = weights)
  metafor_calc <- sqrt(tau_2$tau2)
  meta_q <- tau_2$QE

  tau_data[i,] <- c(hand_calc, metafor_calc, hand_q, meta_q)

}
```

Let's plot this out and replace NaN with 0.

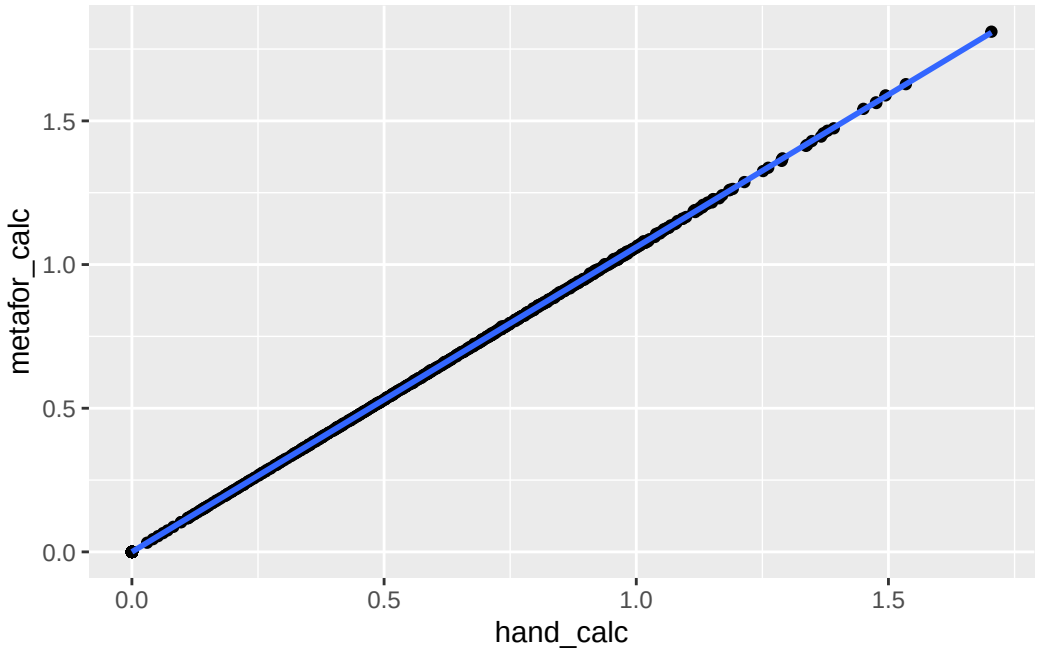
```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
```

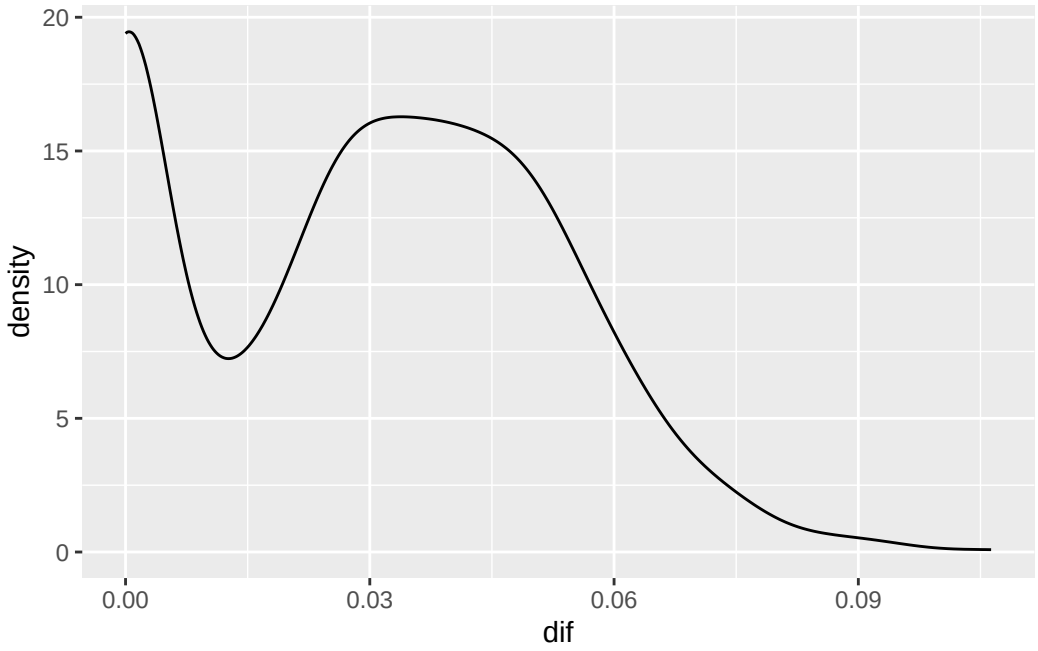
```
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x tidyr::expand() masks Matrix::expand()
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x tidyr::pack()   masks Matrix::pack()
x tidyr::unpack() masks Matrix::unpack()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
tau_data$hand_calc[is.nan(tau_data$hand_calc)] <- 0

tau_data %>%
  ggplot(aes(x = hand_calc, y = metafor_calc))+
  geom_point()+
  geom_smooth()
```

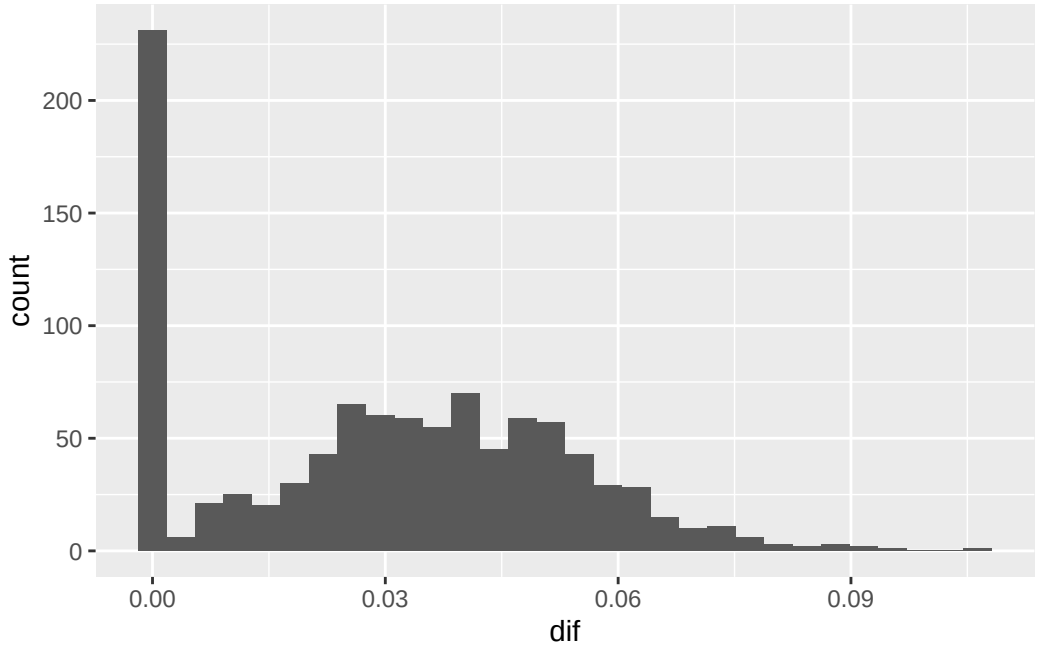
```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```r
dif <- tau_data$metafor_calc- tau_data$hand_calc

ggplot()+
  geom_density(aes(dif))
```




```r
ggplot()+
  geom_histogram(aes(dif), bins = 30)
```



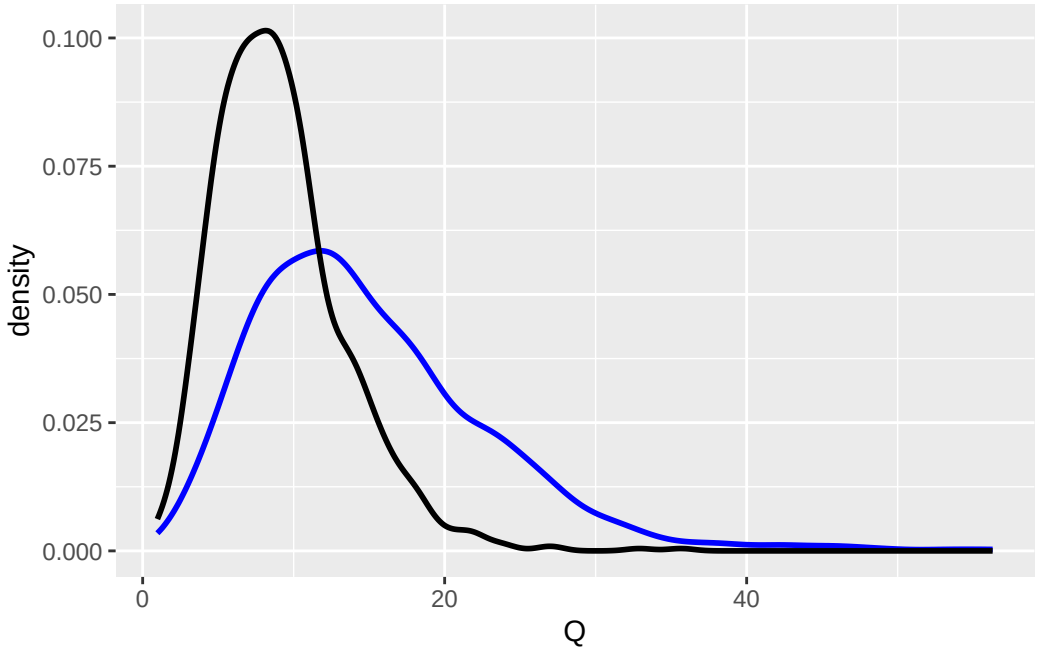

```
mean(dif)
```

```
[1] 0.02999343
```

I am not sure exactly what is going on here, the difference varies but the Q statistic is the same. I feel as if it was only due to me calculating it wrongly the difference should be constant given that the data used is the same. Before looking at if this changes depending on the sample size, let's look at the Q statistic. It should follow a χ^2 distribution with df = k-1. Deviation from this distribution reflects sampling variation beyond what would be expected randomly.

```
set.seed(92895)
chi2 <- rchisq(1000, 9)
ggplot()+
  geom_density(aes(tau_data$meta_q), col = 'blue', size = 1)+
  geom_density(aes(chi2), size = 1)+
  xlab(label = 'Q')
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



The best way to illustrate this is not with random values, but it is nice to see that we get the expected relationship between the expected values and the observed values on just 1000 random Q and χ^2 statistics. Since our observed Qs (blue density) is more right skewed than the random χ^2 statistics(black density) we have evidence that we observe more variation in our effects than to be expected if no heterogeneity is in the sample was present.

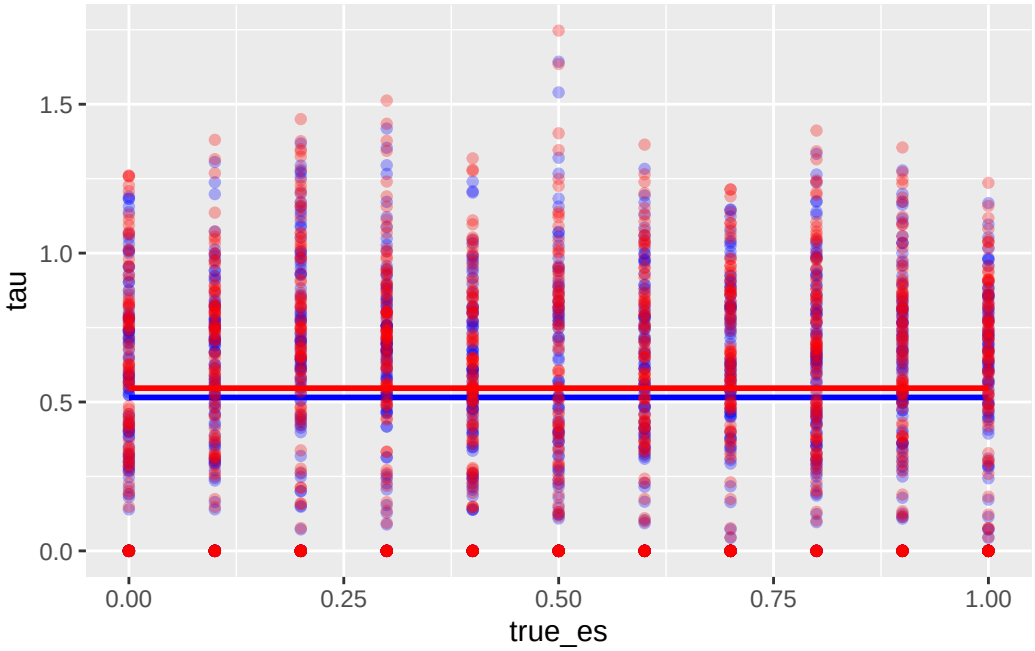
```
tau_data_2 <- data.frame(hand_calc = NA, metafor_calc = NA,
                         true_es=NA)


set.seed(666)
es_range <- seq(0, 1, 0.1)
nsim <- 100
k_studies <- 10

for (i in 1:length(es_range)) {
  for (j in 1:nsim) {

    true_es <- es_range[i]
    es <- rnorm(k_studies, es_range[i], 1)
    sigma <- runif(k_studies, .3, 1)
    weights <- 1/sigma
```

```
    pooled_effect <- sum(weights*es)/sum(weights)
    Q <- sum(weights * (es - pooled_effect)^2)

    hand_calc <- sqrt((Q - (k_studies-1))/sum(weights))

    tau_2 <- rma.uni(yi = es, vi = sigma, method = 'DL',
                     weights = weights)
    metafor_calc <- sqrt(tau_2$tau2)


    index <- (i - 1) * nsim + j
    tau_data_2[index, ] <- c(hand_calc, metafor_calc, true_es)
  }

}
```

Cool, lets plot it out and set the NaN to 0.

```
tau_data_2$hand_calc[is.nan(tau_data_2$hand_calc)] <- 0

tau_data_2 %>%
  ggplot()+
  geom_smooth(aes(x = true_es, y = hand_calc),
              col = 'blue', se = F)+
  geom_point(aes(x = true_es, y = hand_calc),
             col = 'blue', alpha = .3)+
  geom_smooth(aes(x = true_es, y = metafor_calc),
              col = 'red', se = F)+
  geom_point(aes(x = true_es, y = metafor_calc),
             col = 'red', alpha = .3)+
  ylab(label = 'tau')
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



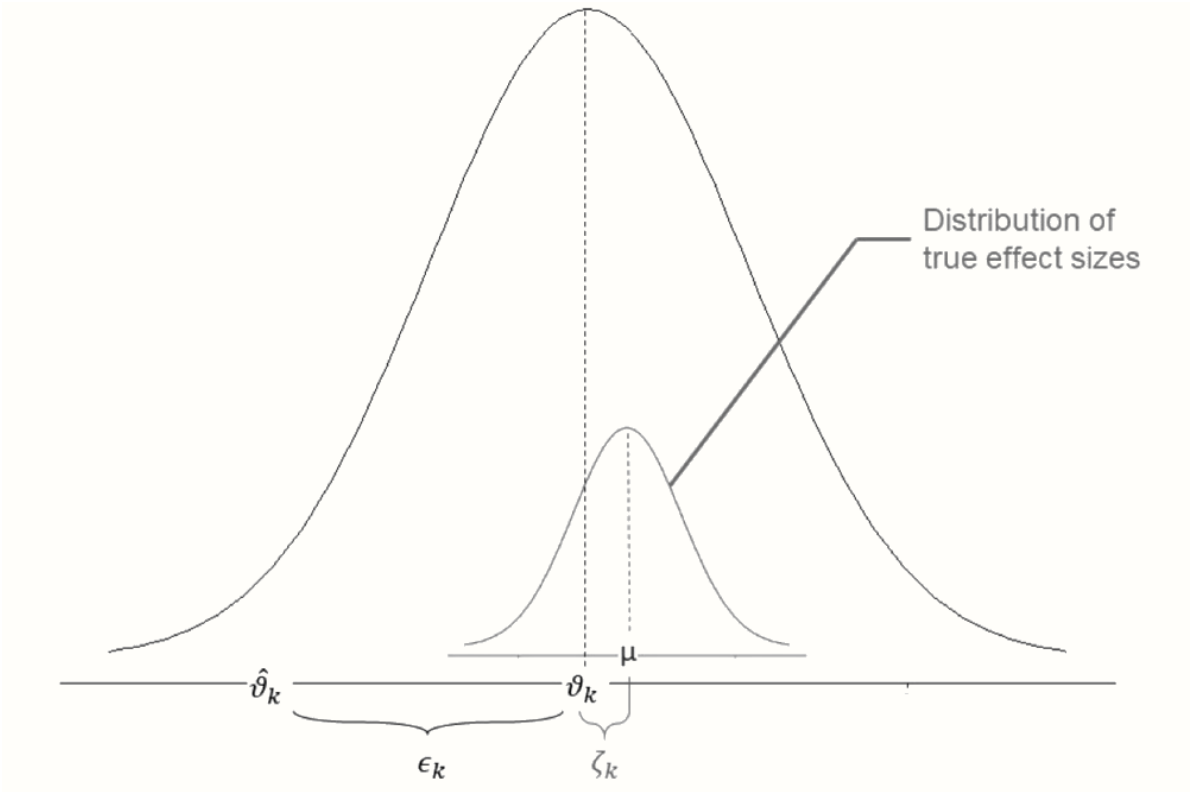
I consistently make an underestimation around roughly .03 compared to metafor regardless of the size of the effect, but there is variation in the difference meaning that it is not a constant miss calculation.

Now, it appears that we have quite a bit of variation in our effect since our tau is consistently around .5. This means that if we have a true effect of 1, that effect has a standard deviation of .5, meaning that the average effect should be between .5 and 1.5 in ca 66% of observed effects. Since we have random effects, we assume that we have multiple true effects in the data and that the random estimate is the average of these true effects. The width of the distribution is what we call heterogeneity and estimate with τ. My understanding is that τ and standard errors are similar but targets different levels of the analysis, standard error being the variation in means and τ being variation in true effects. τ should not be confused with the standard deviation of the observed effects, that is the variation in the sample. For example, let's generate some effects to illustrate.

```
set.seed(834)
k_studies <- 10

es <- rnorm(k_studies, 1, 1)
sigma <- runif(k_studies, .3, 1)
```

We have effects and variances, the standard deviation of the effects is NOT τ, it is simply the sample standard deviation. Again, τ is the standard deviation of the distribution of true effects,

and this distribution of true effects is what have generated the observed sample with moment 1 being the random effect estimate and moment 2 being the heterogeneity. The following figure illustrates the point well.



The distribution of the true effect μ is the heterogeneity. The distribution of θ_k is the variation in the observed true effects. Tau is the standard deviation of the distribution of μ, the standard error is the standard deviation of mean θ_k. $\hat{\theta}_k$ is the estimated true effect dependent on the error between μ and θ_k, and θ_k and $\hat{\theta}_k$.

Higgins and Thompson does a good job explaining it through the following example: They distinguish between key concepts very eloquently, stating that: we have k studies with true treatments effects of θ_i such that $E[\theta_i\ ] = \mu$ and $\text{var}(\theta_i\ ) = \tau^2$. From each study an estimate y_i of θ_i is available such that $E[y_i|\theta_i] = \theta_i$ and $E[y_i|\theta_i] = \sigma_i$. The parameters underlying the scenario are μ, τ^2, σ^2 and k, where μ and τ^2 are unknown and k and σ^2 is assumed known.

A Framework for Approaching Heterogeneity

A key concepts within statistics is to analyse how variation in a value can be reduced by introducing explanatory variables which can lower the average squared distances from the mean. For example, using ANOVA we can try to reduce the variation in a dependent value

by accounting for the variation in the effect introduced by the different treatments and within participant variation if we have repeated measures. A similar perspective can be employed within meta-analysis, but instead of using the term variance we use the word heterogeneity.

A common typology of heterogeneity is that from Cochrane institute which distinguished between clinical, methodological and statistical heterogeneity. Clinical heterogeneity is the variation in effects that is introduced by the participants within a study, the methodological heterogeneity is variation introduced by the study design, and the statistical heterogeneity is the statistically quantified variation in effects\results.

Similarly, Weiss 2014 provides a similar but slightly different view of how to evaluate heterogeneity in the effect of programs. They, like Cochrane argue that heterogeneity has tree sources: Treatment contrast, within participants moderators, and contextual moderators.

Treatment contrast refer to the strength of the measured effect. Consider that we have a completely valid concept X, but when we study X we can get either large X or small X dependent on the stimuli we use to elicit effects. Say we want to examine pain and we agree that this can be measured by putting your hand in cold water, the degree of coldness can be a source of difference in effects that then contributes to unexpected heterogeneity observed across a set of k studies examining this treatment. This is an interesting source of heterogeneity since the concept might be the same, but the true effect is not.

Within participants moderators refer to variation coming from individual participants. Say that we do our cold water experiment and have a single stimuli of water with a certain coldness, but our sample consists of two unaccounted for subgroups – one with a high pain tolerance and one with low pain tolerance. This difference within participants affects the strength of the effect, and thus is a systematic source of variation which we cannot attribute to normal sampling error – thus providing heterogeneity.

Contextual moderating factors are factors external to the participants and separate from the stimuli provided to elicit the effect. They cover contextual aspects of the ''study environment" that can cause potential differences in effects. For example, say that the coldness of water could be moderated by the size of the cup you place your hand into, and that our experiment uses multiple cups which vary in size. This could artificially introduce variation in the effect due to the context of the measurement and again provide heterogeneity.

It is important to note that the contexts of these perspectives on heterogeneity differs, one is from medicin and the other from political science and policy. However, they both have striking similarities. The main difference is that Cochrane places less weight on stimuli strength, which presumably would fall into methodological heterogeneity.

Cook, Campbell and Shadish makes the point that experiments, which are often used in psychology and medicine but cannot always be used in policy research, are highly local but have very general aspirations. That is, an experiment answers a very specific question about a specific group of people at a specific point in time. As such, there is a conflict between the aim of a scientific inquiry and that which is answerable through an experiment. However, if

the construct we examine through our experiment is valid, we can make these generalizations more confidently.

Validity

Exercises for Your Understanding

1. **Is there such a thing as a ‘’True” effect? If so, what is it? Is it an exact parameter or a distribution of values?**

   I believe a true effect does exist in reality, but whether it is a fixed point/magnitude is hard to say. A true effect is a relation between two or more things. The effect can sometimes be generative of a concept. I feel as if a safe place to start thinking about true effects is through John Locke, who states: ’ ’That which produces any simple or complex idea, we denote by the name *cause*, and that which is produces, *effect’ ’*. Through this lens, a true effect is simply the constant thing caused by something else. That is, a true effect is *the* thing that can be derived from some other thing. Put in context of an experimental paradigm, a true effect can be described as the true counterfactual difference between the worlds where something occurred and where that something didn’t occur. This is obviously an impossible thing to observe empirically, but we can with the help of experimental design artificially construct a couterfactual.

   In light of this definition of effect, I think we need to make the concession that it is a point estimate. The true differences between states is not variable. If that argument were to be made – that the true effect is distributed across some other meta/latent effect – we can argued that to infinity, there is always another underlying effect causing the distribution of a true effect. We need to bite the bullet with true effects and assume that they are fixed values that do not change.

2. **When are true effects causal and when are the not causal? How does this impact our research question?**

   All effects are products of a cause, but not all effects are causes of some other effect. They can be correlated with other effects but they do not have to be related in any other manner. For example, Y can cause the effects X and Z, but X does not have to cause Z. X does not have to be unrelated to Z either, there common cause – Y is what cements their dependency. For example, say we have two effects, a sound, and a physical sensation in our hands. The cause of both of these effects is me clapping my hands. This does not mean that the sound caused the sensation or that the sensation caused the sound, but they occur together and are therefore related. For our research into heterogeneity, we need to divide types of heterogeneity since they have different causes. What we might be interested in how the different heterogeneity effects are related to each other. For this we need to think about the laws of causality – temporal presidency, co-occurance,

and logical consistency. Let's come back to this after we have a more consistent view of heterogeneity topology.

3. **Where does a true effect live ontological and epistemicaly?**

   I believe it resides within critical realism. We have to assume a consistent realism for effects to have a meaning. That is, it needs to exist in reality for us to systematically and reliably observe it. However, we must acknowledge that observation is flawed, and that there will be variation in interpretation of the value and meaning of empirical observation.

4. **What is the difference between an ''effect" and a ''true effect"?**

   For our purposes, there is no difference between the general term effect and ''true effect" since we already have recognized an ''effect" to mean the fixed counterfactual difference. One might make the distinction that the ''true effect" is the contextualized version of *the* effect, meaning that the ''true effect" has clear mathematical properties depending on the data generative of the "observed effect" – which is separate from both the general effect and the ''true effect". Another way of illustrating this might be trough looking at the observed effect $\widehat{\theta}_k$. The observed effect differs only from the true effect through random/sampling error. However, this might not be true for the general counterfactual effect, though i might be wrong on this. In the end, they are effectively the same for our intents. The Cochrane handbook provides a pretty nice description of what they refer to as **effect measures**. What they mean by this is statistical constructs that compare outcome data between two interventions groups, that is, an effect size or an observed effect. I think this definition of an observed effect is good and to the point. It also shares language with validity theory (construct validity).

5. **Some articles claim that methodological heterogeneity is crucial for theory development/testing/generalization and the progress of science. And some others (try to find examples) found that methodological heterogeneity led to increased uncertainty. How can these viewpoints be reconciled?**

   I think the meaning of uncertainty in this case is important to define. If uncertainty just IS heterogeneity then increased heterogeneity is a natural thing to expect when making a conceptual replication. That is the amount of unexplainable variation introduced through the replication. One aspect of Linden and Hönekopp covers this, they find that conceptual replications have larger standard deviations than close replications, meaning that the variation in effect size is larger than expected just from random error. On of the main ideas behind conceptual replication is to widen the space of potential explainable values of a parameter, meaning that through our acquisition of knowledge we gain a better ability to predict an outcome. However, if this generalization only causes us to make worse predictions then we have a serious issue. The more general a research question become, the more dependencies it requires for a accurate prediction, so when we generalize without trying to account for the generality, we will be doing a big mistake.

I think the concept of incorporating our knowledge of the phenomenon in our predictions marries the idea that generality is good for theory but also bad for precision.

6. **Does methodological heterogeneity cause statistical heterogeneity? When?**

   I think it always will be some miniscule increase in statistical heterogeneity that will be caused by methodological heterogeneity unless the true effects observed are distinct but identical. Say the outcome of method A is identical to method B, can we be sure that they truly are distinct in their methodological effect? This might be logically true, but it is also a theoretical question. In other words, how can we be sure of their internal validity? As methodological heterogeneity is explored it introduces statistical heterogeneity by virtue of adding complexity to the model if not mathematically adjusted for. However, we can have interesting edge cases. Consider a situation where we observe effect Y from 5 procedurally identical studies, we then replicate effect Y from another 5 studies who are different in measurement from the previous experiment but procedurally identical to each other. In the fist synthesis we get a tau of 5, in the second we get a tau of 4, but when we pool the results and adjust for the different methodological approach, we get a tau of 3. One possible explanation of such a scenario is that we have a poor understanding of the effect to the extent that instead of capturing effect Y, we capture effect Z. I think it might be best illustrated through a Q distribution. Image the same scenario of studying effect Y, but both our methods of measuring are skewed in relation to the chi^2(k-1) distribution, but when we pool our results, the variation of the data is captured by the new chi^2(k-1) distribution. I should test this through simulation.

7. **Can it be that methodological heterogeneous studies have low statistical heterogeneity?**

   I don't see a reason for why there might not be cases like this. It appears to be uncommon but I do not doubt that some examples exist.

8. **When can we distinguish between methodological and clinical heterogeneity?**

9. **Do meta-analyses try to restrict the methodological heterogeneity of the included study in order to potentially reduce statistical heterogeneity? Is this a good approach?**

   It depends on what ''restrict" means. If restrict is post-hoc changing the inclusion criteria to reduce statistical heterogeneity i do not think that is very smart. However, exclusion of some possibly influential studies as a sensitivity analysis is good. I do not think there is any harm in exploring the garden of forking paths so long as you can keep track of where you have been and how much you have explored.

10. **How much does null-hypothesis testing have to do with the problem of heterogeneity and failure to replicate?**

11. **Can informative hypothesis testing as an alternative to null-hypothesis testing be a remedy? E.g. by having more power?**

12. **Is it fair to think about heterogeneity as unaccounted for variation? Or perhaps unexpected variance?**

13. **Are the Cochrane categories(clinical, methodological, statistical) of heterogeneity correct? How does it relate to Weiss?**