Meta-Análise

José Luiz Padilha da Silva

06/08/2019

A meta-análise é uma técnica estatística que objetiva combinar resultados de estudos independentes, sobre uma mesma questão de pesquisa, combinando, em uma medida resumo, os resultados de tais estudos.

A meta-análise tem um papel de grande importância na síntese de evidência em muitas áreas de pesquisa, como nas Ciências Sociais, Medicina e Economia.

Uma Simulação Simples de Um Estudo Multicêntrico

Para introduzirmos alguns aspectos de uma meta-análise vamos simular um estudo multicêntrico simples com o objetivo de comparar uma nova droga anti-hipertensiva (denotada por `Drug`) com uma droga convencional controle (denotada por `CTRL`) na redução da pressão sanguínea diastólica em homens adultos hipertensivos.

Vamos assumir que uma *análise apropriada de poder* indicou que o tamanho amostral necessário para detectar uma diferença especificada entre os tratamentos seja $n = 1.000$. Já que é difícil recrutar 1.000 participantes em um lugar durante uma janela definida de tempo, a equipe de pesquisadores decidiu conduzir um estudo multicêntrico para recrutar os participantes em cinco centros distintos.

Para estes n participantes, nós registramos a idade e medimos a pressão diastólica *baseline* (inicial) logo antes da aleatorização já que a idade é um importante fator de risco relacionado com a pressão sanguínea.

As drogas foram então administradas e a pressão sanguínea foi medida e registrada periodicamente, incluindo uma medição ao fim do estudo. A variável resposta é a mudança na pressão sanguínea entre o *endpoint* (final) e *baseline* (linha de base) foi calculado e usado para avaliar a eficácia anti-hipertensiva da nova droga.

Dados

Vamos então simular os dados por centro. Por exemplo, no centro 1 assumimos que a pressão diastólica baseline para os 200 participantes recrutados ($n = 100$ para cada tratamento) é normalmente distribuída com média $\mu = 100$ (mmHg) e desvio padrão $sd = 20$ (mmHg). A idade é normalmente distribuída com média $age.mu = 50$ (anos) e desvio padrão $age.sd = 10$ (anos). Além disso, assumimos que a nova droga diminuirá a pressão sanguínea diastólica em $mu.d = 10$ (mmHg).

No `R` fazemos:

```
# Number of participants each arm
n   = 100
#  Mean  blood  pressure  at  baseline
mu  = 100
# Standard deviations for blood pressure
sd  = 20
# Mean changes for blood pressure
mu.d    = 10
#  Mean  age  for participants
age.mu = 50
#  sd  of  age  for  participants
age.sd = 10
```

Primeiramente simulados os dados para os n participantes do grupo `CTRL` com idade (denotada por `age`), pressão sanguínea baseline (denotada por `bp.base`), pressão sanguínea final (denotada por `bp.end`) e mudança na pressão sanguínea (denotada por `bp.diff=bp.end-bp.base`):

```r
# Fix the seed for random number generation
set.seed(123)
#  Use  "rnorm"  to  generate  random  normal
age =  rnorm(n,  age.mu,  age.sd)
bp.base =  rnorm(n,mu,sd)
bp.end  =  rnorm(n,mu,sd)
# Take the difference between endpoint and baseline
bp.diff = bp.end-bp.base
# put the data together using "cbind" to column-bind
dat4CTRL    =  round(cbind(age,bp.base,bp.end,bp.diff))
#
head(dat4CTRL)
```

```
##      age bp.base bp.end bp.diff
## [1,]  44      86    144      58
## [2,]  48     105    126      21
## [3,]  66      95     95       0
## [4,]  51      93    111      18
## [5,]  51      81     92      11
## [6,]  67      99     90      -9
```

Similarmente, simulamos os dados para a nova droga.

```r
#  Simulate  `age'
age =  rnorm(n,  age.mu,  age.sd)
#  Simulate  `baseline' blood  pressure
bp.base =  rnorm(n,mu,sd)
# Simulate `endpoint' blood pressure
bp.end  =   rnorm(n,mu-mu.d,sd)
#  The  changes  in  blood  pressure
bp.diff = bp.end-bp.base
#  Make  the  data  matrix
dat4drug   =   round(cbind(age,bp.base,bp.end,bp.diff))
```

Para facilitar a manipulação, nós empilhamos os conjuntos de dados `CTRL` e `Drug`, criamos duas colunas indicadora de tratamento e centro.

```r
# Make a dataframe to hold all data
dat1    =    data.frame(rbind(dat4CTRL,dat4drug))
# Make "TRT" as a factor for treatment.
dat1$TRT    =   as.factor(rep(c("CTRL",   "Drug"),   each=n))
#  Make  a  "Center"  to  represent  the  center  number
dat1$Center  =  1
```

O conjunto de dados `dat1` tem 200 observações sendo 100 `CTRL` e 100 `Drug`.

```r
# check the data dimension
dim(dat1)
```

```
## [1] 200   6
```

```r
# print the first 6 obervations to see the variable names
head(dat1)
```

```
##   age bp.base bp.end bp.diff  TRT Center
## 1  44      86    144      58 CTRL      1
## 2  48     105    126      21 CTRL      1
## 3  66      95     95       0 CTRL      1
```

```
## 4  51      93    111      18 CTRL      1
## 5  51      81     92      11 CTRL      1
## 6  67      99     90      -9 CTRL      1
```

A seguir, usamos uma função para executar o processo de geração de dados para os outros centros.

```
data.generator = function(n,age.mu,age.sd,mu,mu.d,sd, center){
# Data from CTRL
age = rnorm(n, age.mu, age.sd)
bp.base = rnorm(n,mu,sd)
bp.end = rnorm(n,mu,sd)
bp.diff = bp.end-bp.base
dat4CTRL = round(cbind(age,bp.base,bp.end,bp.diff))
# Data from Drug
age = rnorm(n, age.mu, age.sd)
bp.base = rnorm(n,mu,sd)
bp.end = rnorm(n,mu-mu.d,sd)
bp.diff = bp.end-bp.base
dat4drug = round(cbind(age,bp.base,bp.end,bp.diff))
# Put both data matrices together
dat = data.frame(rbind(dat4CTRL,dat4drug))
# Make "TRT" as a factor for treatment.
dat$TRT = as.factor(rep(c("CTRL", "Drug"), each=n))
# Make a "Center" to represent the center number
dat$Center = center
# Return the simulated data
dat
} # end of function
```

Para o centro 1, podemos obter novamente os dados fazendo:

```
d1  =  data.generator(n,age.mu,age.sd,mu,mu.d,sd,  1)
```

Para gerar os dados para os outros centros, vamos assumir média e desvio padrão para pressão sanguínea baseline similares em todos os centros, mas que a nova droga tem diferentes efetividades para cada centro com $mu.d2 = 13$ no centro 2, $mu.d3 = 15$ no centro 3, $mu.d4 = 8$ no centro 4 e $mu.d5 = 10$ no centro 5. Geramos os dados para cada centro da seguinte forma:

```
# Data from Center 2
mu.d2 = 13
d2  =         data.generator(n,age.mu,age.sd,mu,mu.d2,sd,2)
# Data from Center 3
mu.d3 = 15
d3  =         data.generator(n,age.mu,age.sd,mu,mu.d3,sd,3)
# Data from Center 4
mu.d4 = 8
d4  =         data.generator(n,age.mu,age.sd,mu,mu.d4,sd,4)
# Data from Center 5
mu.d5 = 10
d5  =         data.generator(n,age.mu,age.sd,mu,mu.d5,sd,5)
```

Combinando os dados dos 5 centros, criamos um conjunto de dados chamado `dat`, como segue:

```
dat=data.frame(rbind(d1,d2,d3,d4,d5))
# Change `Center' from numeric to factor
dat$Center  =  as.factor(dat$Center)
dim(dat)
```

```
## [1] 1000    6
```

```
head(dat)
```

```
##   age bp.base bp.end bp.diff  TRT Center
## 1  61      85    107      22 CTRL      1
## 2  50      69     87      18 CTRL      1
## 3  50      86    117      31 CTRL      1
## 4  35     102    123      21 CTRL      1
## 5  58      73    106      33 CTRL      1
## 6  48     112    103      -9 CTRL      1
```

Análise Exploratória

Medidas Resumo

Agregamos os estudos individuais por tamanho amostral, médias e desvio padrão:

```
# Get the study sample size
ndat  =  aggregate(dat$bp.diff, list(Center=dat$Center,TRT = dat$TRT), length)
# Print the study specific sample size
ndat
```

```
##    Center  TRT   x
## 1       1 CTRL 100
## 2       2 CTRL 100
## 3       3 CTRL 100
## 4       4 CTRL 100
## 5       5 CTRL 100
## 6       1 Drug 100
## 7       2 Drug 100
## 8       3 Drug 100
## 9       4 Drug 100
## 10      5 Drug 100
```

```
# Calcuate  the means by  study
mdat  =  aggregate(dat$bp.diff, list(Center=dat$Center,TRT  =  dat$TRT),  mean)
# Print the means
mdat
```

```
##    Center  TRT      x
## 1       1 CTRL  -0.24
## 2       2 CTRL   3.05
## 3       3 CTRL  -3.84
## 4       4 CTRL  -4.43
## 5       5 CTRL   3.84
## 6       1 Drug -12.86
## 7       2 Drug -14.14
## 8       3 Drug -11.43
## 9       4 Drug  -7.04
## 10      5 Drug  -8.77
```

```
#  Calculate  the  standard  deviations
sddat   =   aggregate(dat$bp.diff, list(Center=dat$Center,TRT  =  dat$TRT),  sd)
# Print the SDs
sddat
```

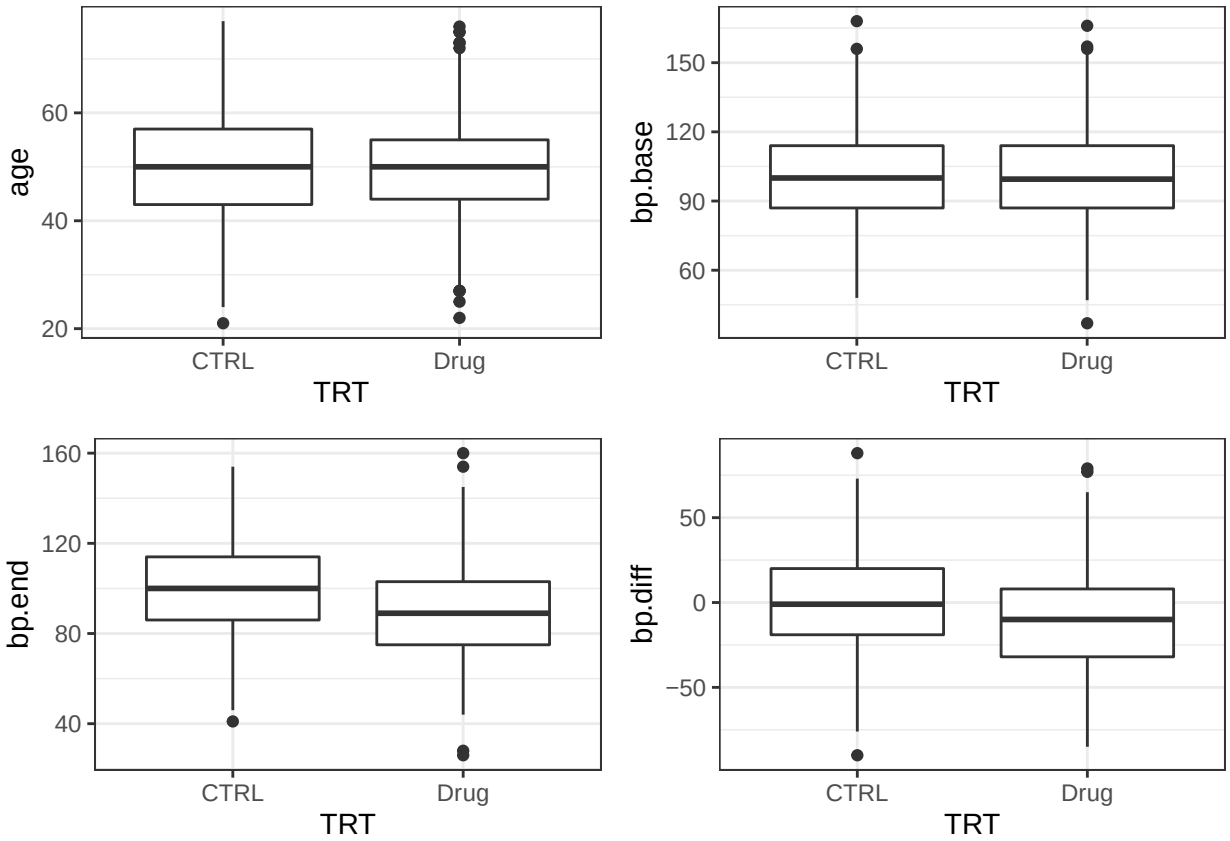
```
##    Center  TRT        x
```

```
## 1        1 CTRL 29.17558
## 2        2 CTRL 27.42497
## 3        3 CTRL 30.04600
## 4        4 CTRL 28.33360
## 5        5 CTRL 28.30131
## 6        1 Drug 28.54874
## 7        2 Drug 30.30919
## 8        3 Drug 27.45879
## 9        4 Drug 28.37196
## 10       5 Drug 30.64059
```

Gráficos

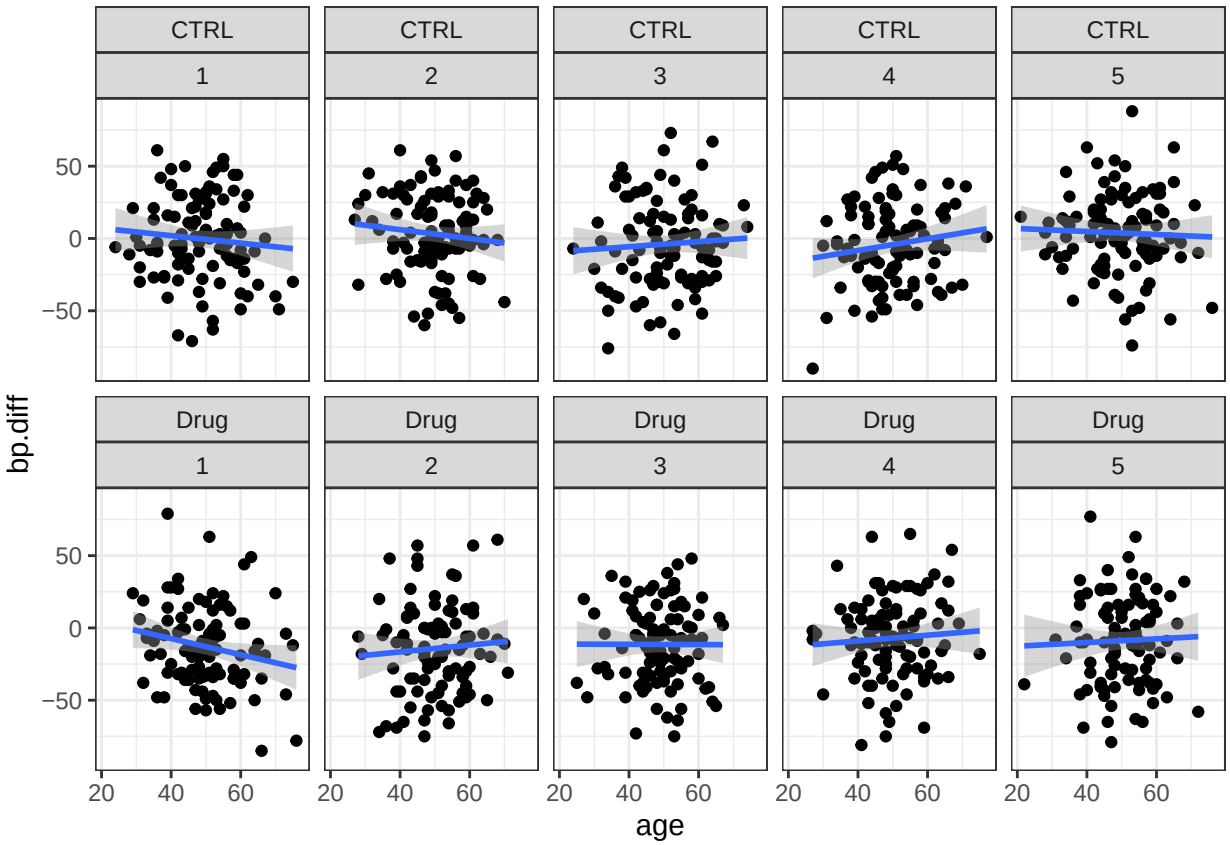
Avaliamos a distribuição dos dados por meio de boxplots:

```
library(ggplot2); library(gridExtra)
p1 = ggplot(dat, aes(x=TRT, y=age)) + geom_boxplot() + theme_bw()
p2 = ggplot(dat, aes(x=TRT, y=bp.base)) + geom_boxplot() + theme_bw()
p3 = ggplot(dat, aes(x=TRT, y=bp.end)) + geom_boxplot() + theme_bw()
p4 = ggplot(dat, aes(x=TRT, y=bp.diff)) + geom_boxplot() + theme_bw()
grid.arrange(p1, p2, p3, p4)
```



Na sequência mostramos a relação entre a diferença da pressão sanguínea como uma função da idade para cada tratamento com o objetivo de avaliarmos se existe uma relação *estatisticamente significativa* além de uma diferença de tratamento.

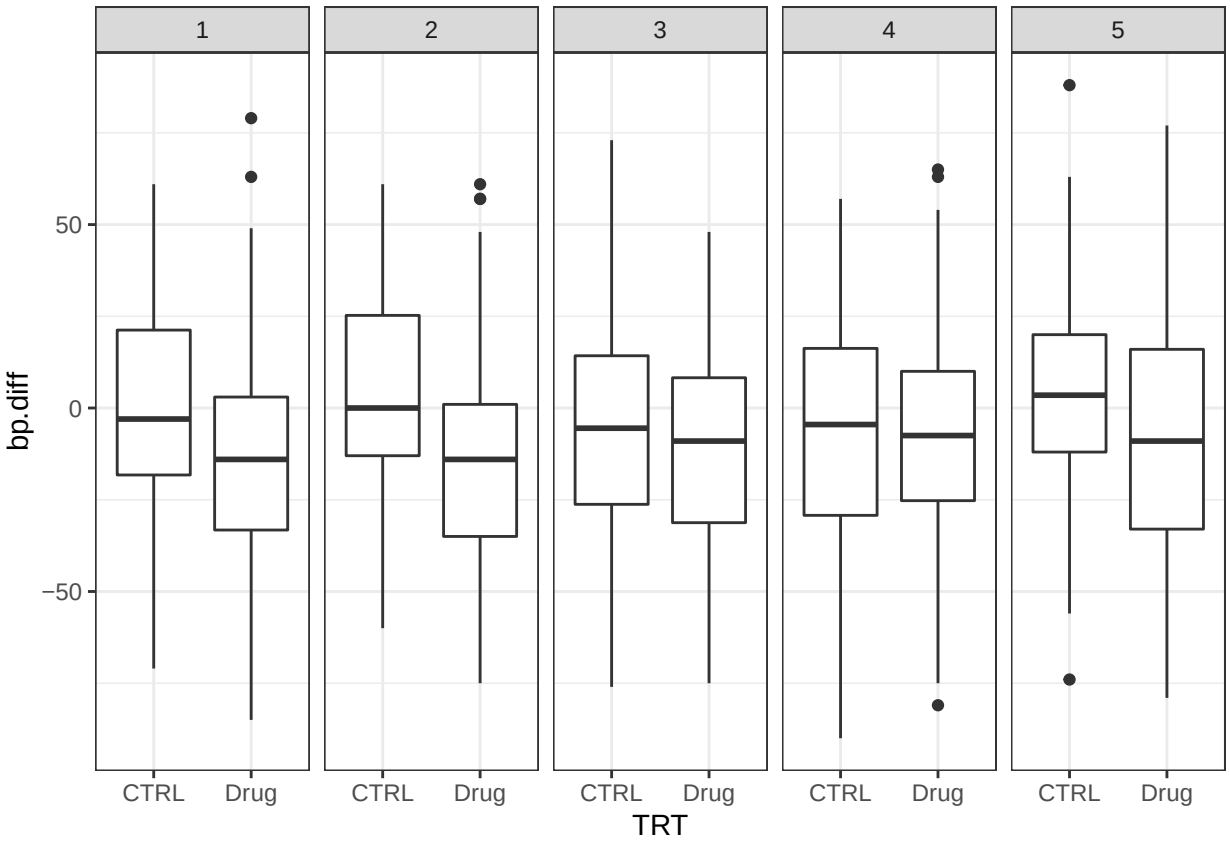
```
ggplot(dat, aes(x=age, y=bp.diff)) + geom_point() + geom_smooth(method = "lm") +
  facet_wrap(~TRT+Center, ncol = 5) + theme_bw()
```



- A relação entre a pressão sanguínea e idade parece ser significativa?
- A nova droga parece reduzir a pressão sanguínea?
- Como podemos modelar estes dados?

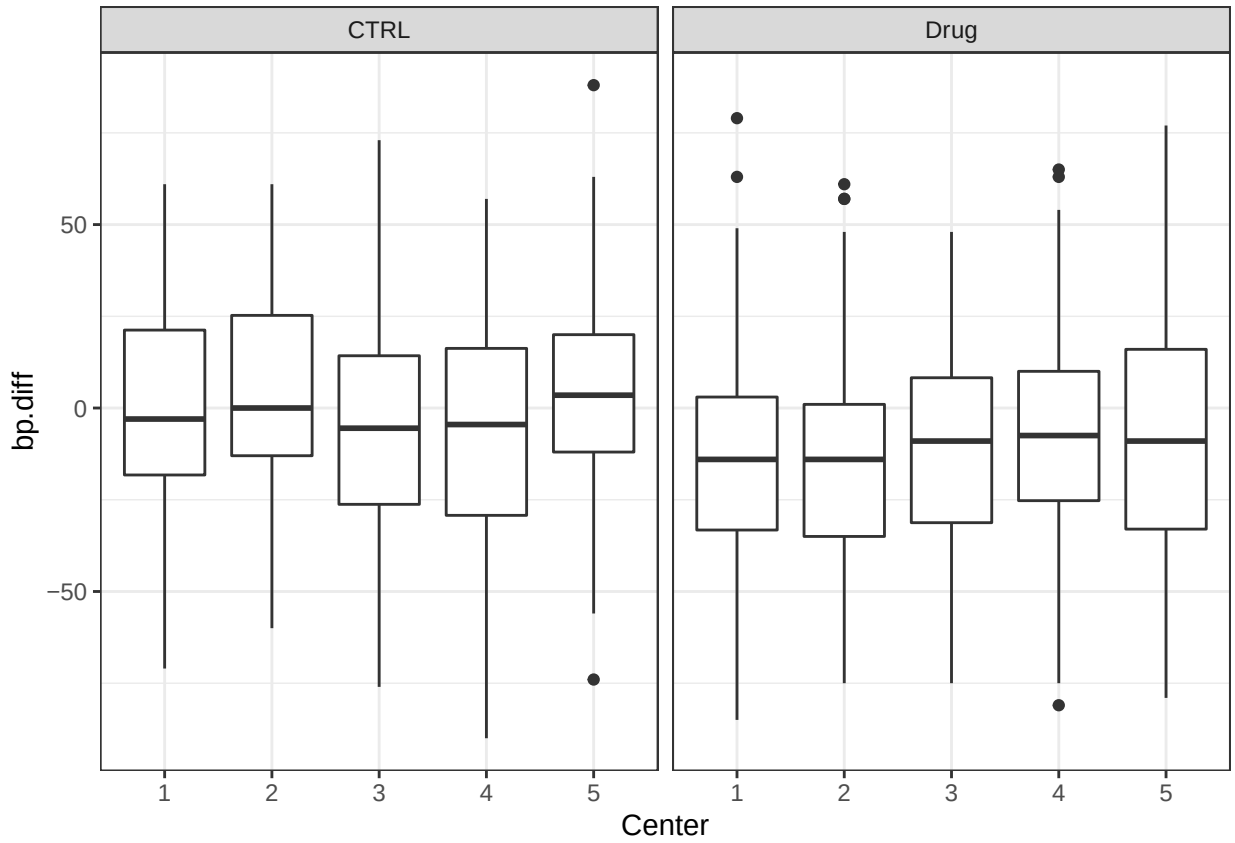
Verifiquemos agora o efeito do tratamento por centro.

```
ggplot(dat, aes(x=TRT, y=bp.diff)) + geom_boxplot() + facet_wrap(~Center, nrow=1) + theme_bw()
```



Alternativamente, ilustramos o efeito do tratamento pelo gráfico a seguir.

```
ggplot(dat, aes(x=Center, y=bp.diff)) + geom_boxplot() + facet_wrap(~TRT, nrow=1) + theme_bw()
```



Vemos algumas variações dentro dos centros, consequência de simularmos dados com diferentes valores de *mu.d*.

Análise com Dados de Cada Centro

Para os centros individuais nós esperamos observar resultados estatisticamente menos significativos

```
# Model for Center 1
summary(lm(bp.diff~TRT,  data=dat[dat$Center==1,]))
```

```
##
## Call:
## lm(formula = bp.diff ~ TRT, data = dat[dat$Center == 1, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72.14 -19.30  -2.14  19.11  91.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.240      2.886  -0.083  0.93382
## TRTDrug      -12.620      4.082  -3.092  0.00228 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.86 on 198 degrees of freedom
## Multiple R-squared:  0.04605,    Adjusted R-squared:  0.04123
```

```
## F-statistic: 9.558 on 1 and 198 DF,  p-value: 0.002277
```

```
# Model for Center 2
summary(lm(bp.diff~TRT,  data=dat[dat$Center==2,]))
```

```
##
## Call:
## lm(formula = bp.diff ~ TRT, data = dat[dat$Center == 2, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.05 -18.25  -0.86  19.45  75.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.050      2.890   1.055    0.293
## TRTDrug      -17.190      4.088  -4.205 3.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.9 on 198 degrees of freedom
## Multiple R-squared:  0.082,  Adjusted R-squared:  0.07736
## F-statistic: 17.69 on 1 and 198 DF,  p-value: 3.944e-05
```

```
# Model for Center 3
summary(lm(bp.diff~TRT,  data=dat[dat$Center==3,]))
```

```
##
## Call:
## lm(formula = bp.diff ~ TRT, data = dat[dat$Center == 3, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.160 -21.718   0.635  18.988  76.840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.840      2.878  -1.334   0.1837
## TRTDrug       -7.590      4.070  -1.865   0.0637 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.78 on 198 degrees of freedom
## Multiple R-squared:  0.01726,    Adjusted R-squared:  0.0123
## F-statistic: 3.477 on 1 and 198 DF,  p-value: 0.0637
```

```
# Model for Center 4
summary(lm(bp.diff~TRT,  data=dat[dat$Center==4,]))
```

```
##
## Call:
## lm(formula = bp.diff ~ TRT, data = dat[dat$Center == 4, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.570 -20.960  -0.265  19.138  72.040
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.430      2.835  -1.562    0.120
## TRTDrug       -2.610      4.010  -0.651    0.516
##
## Residual standard error: 28.35 on 198 degrees of freedom
## Multiple R-squared:  0.002135,   Adjusted R-squared:  -0.002904
## F-statistic: 0.4237 on 1 and 198 DF,  p-value: 0.5159
```

```
# Model for Center 5
summary(lm(bp.diff~TRT,  data=dat[dat$Center==5,]))
```

```
##
## Call:
## lm(formula = bp.diff ~ TRT, data = dat[dat$Center == 5, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77.84 -18.19  -0.23  20.12  85.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.840      2.949   1.302  0.19444
## TRTDrug      -12.610      4.171  -3.023  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.49 on 198 degrees of freedom
## Multiple R-squared:  0.04412,    Adjusted R-squared:  0.0393
## F-statistic:  9.14 on 1 and 198 DF,  p-value: 0.002832
```

Dos ajustes acima, vemos que o efeito de tratamento é estatisticamente significativo ao nível de 5% para os centros 1, 2 e 5, mas não para os centros 3 e 4.

Como podemos combinar estes resultados em uma única inferência? Será que um modelo linear com centro como fator fixo é adequado?

Meta-análise

Para análise de dados provenientes de diferentes estudos vamos utilizar métodos de meta-análise. Em resumo:

- Precisamos de uma medida de efeito e variabilidade para cada estudo e então calculamos uma média ponderada dessas medidas de efeito.
- Geralmente atribuímos mais peso aos estudos mais precisos, mas as regras para atribuição dos pesos dependem das nossas suposições distribucionais dos verdadeiros efeitos.

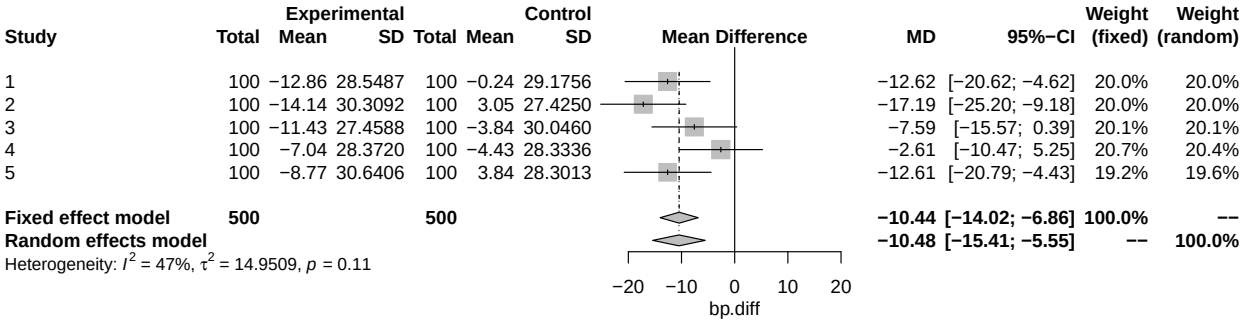
O produto final de uma meta-análise pode ser representado em um gráfico do tipo *forest plot*, como exemplo a seguir:

```
meta.dat = data.frame(Study=ndat$Center,TRT=ndat$TRT,n=ndat$x,m=mdat$x,s=sddat$x)
meta.dat = cbind(subset(meta.dat, TRT == "Drug"), subset(meta.dat, TRT == "CTRL"))[-c(2,6,7)]

library(meta)
```

```
## Loading 'meta' package (version 4.9-5).
## Type 'help(meta)' for a brief overview.
```

```r
# Do meta-analysis
m <- metacont(n, m, s, n.1, m.1, s.1, studlab=Study, data=meta.dat)
# Produce forest plot
forest(m, xlab="bp.diff")
```



Algumas Questões

- Como podemos planejar e conduzir uma meta-análise? Quais protocolos devem ser seguidos na revisão sistemática para obtenção dos dados de diferentes estudos.
- Qual método de análise é mais adequado para o tipo de resposta considerada? Modelos de efeitos fixos ou efeitos aleatórios?
- Como estimar a variabilidade entre estudos? Quais medidas de heterogeneidade são adequadas?
- É necessário realizar análise de subgrupos? Uma meta-regressão é indicada?
- Existe *efeito de pequenos estudos*? Viés de publicação?
- Como lidar com dados ausentes?