HW5 Cohen’s d报告的两分布重叠部分概率问题

Jiawen Wu

12/24/2018

参数设置

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
library(ggthemes)

# ggplot的主题参数设置
bg <- "#ECF0F2" # 背景颜色
# 把画图时用到的一些底色都调成背景颜色
theme_set(theme_economist() + theme(panel.background = element_rect(fill = bg),
                                    plot.background = element_rect(fill = bg),
                                    strip.background = element_rect(fill = bg),
                                    legend.background  = element_rect(fill = bg),
                                    legend.key = element_rect(fill = bg)))
# 给会用到的几种点上色
scale_color_rpsy <- scale_color_manual(values = c("experiment" = "#E8948E",
                                                  "control" = "#3E91BA",
                                                  "control_overlap" = "#82D9CB",
                                                  "experiment_overlap" = "#FEF3AC"
))
```

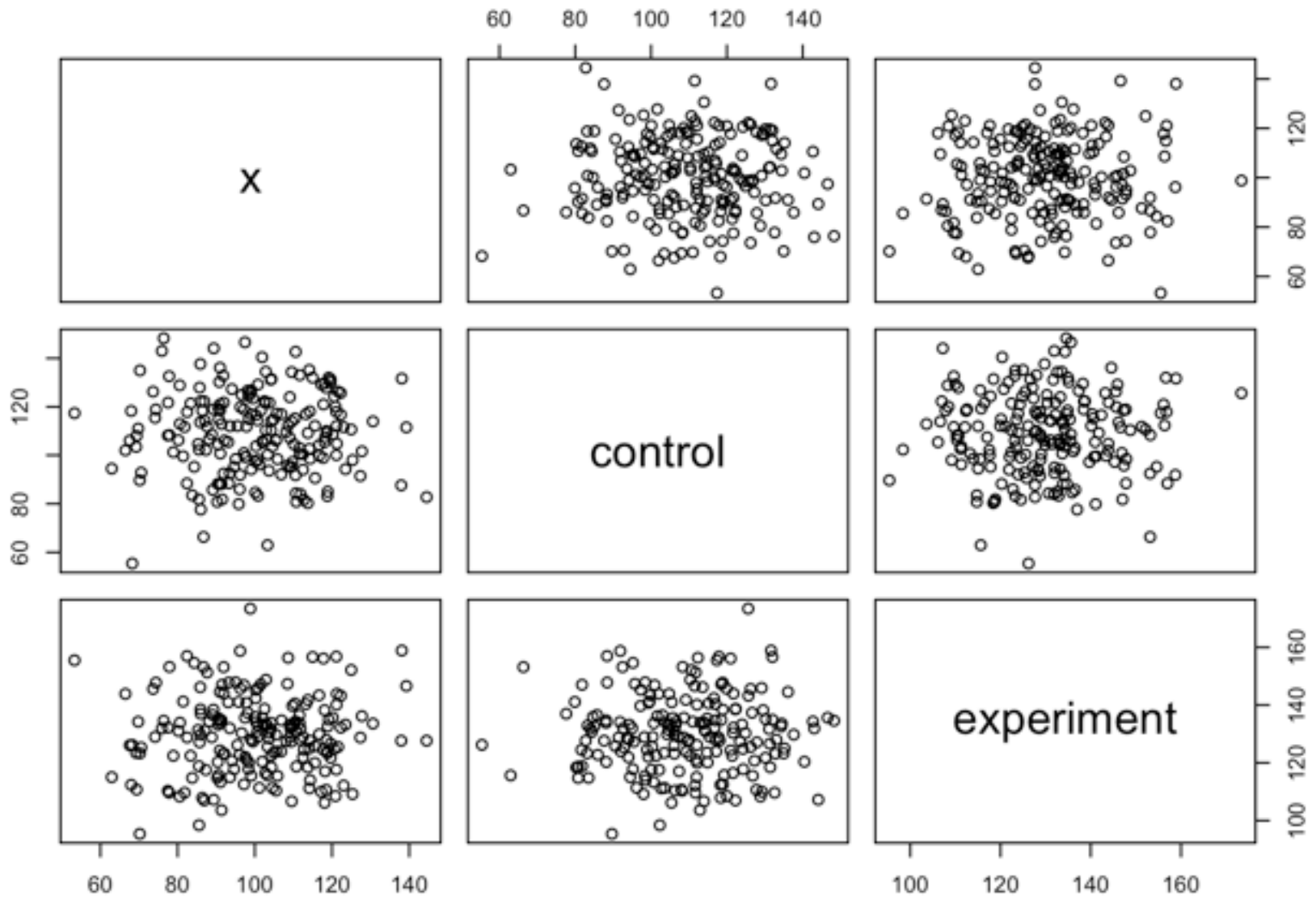
介绍一下管道函数 %>% 的用法

符号：%>%这是管道操作，其意思是将%>%左边的对象传递给右边的函数。

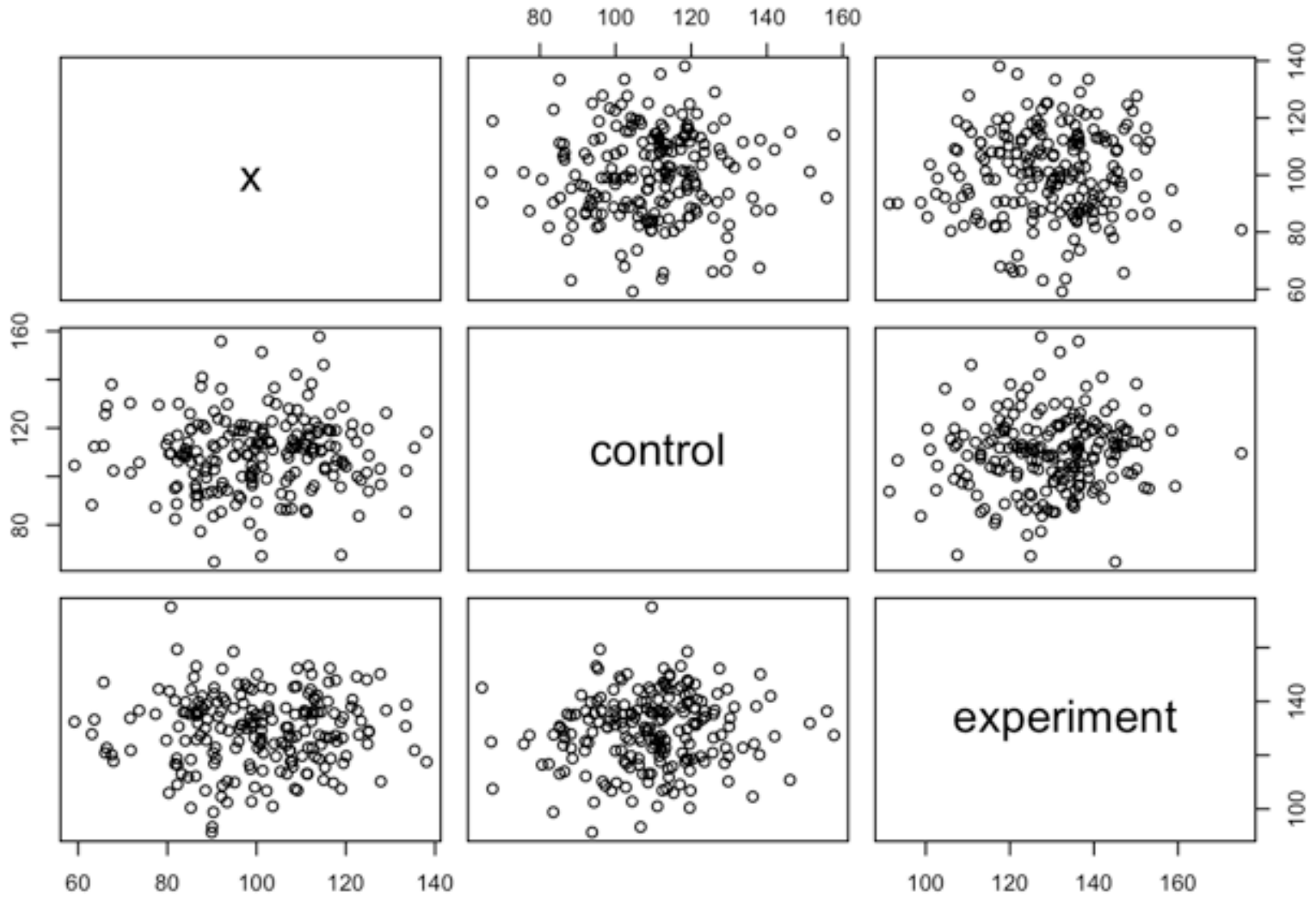
说明：%>%来自dplyr包的管道函数，其作用是将前一步的结果直接传参给下一步的函数，从而省略了中间的赋值步骤，可以大量减少内存中的对象，节省内存

- x %>% f(y) 等同于 f(x, y)
- y %>% f(x, ., z) 等同于 f(x, y, z)

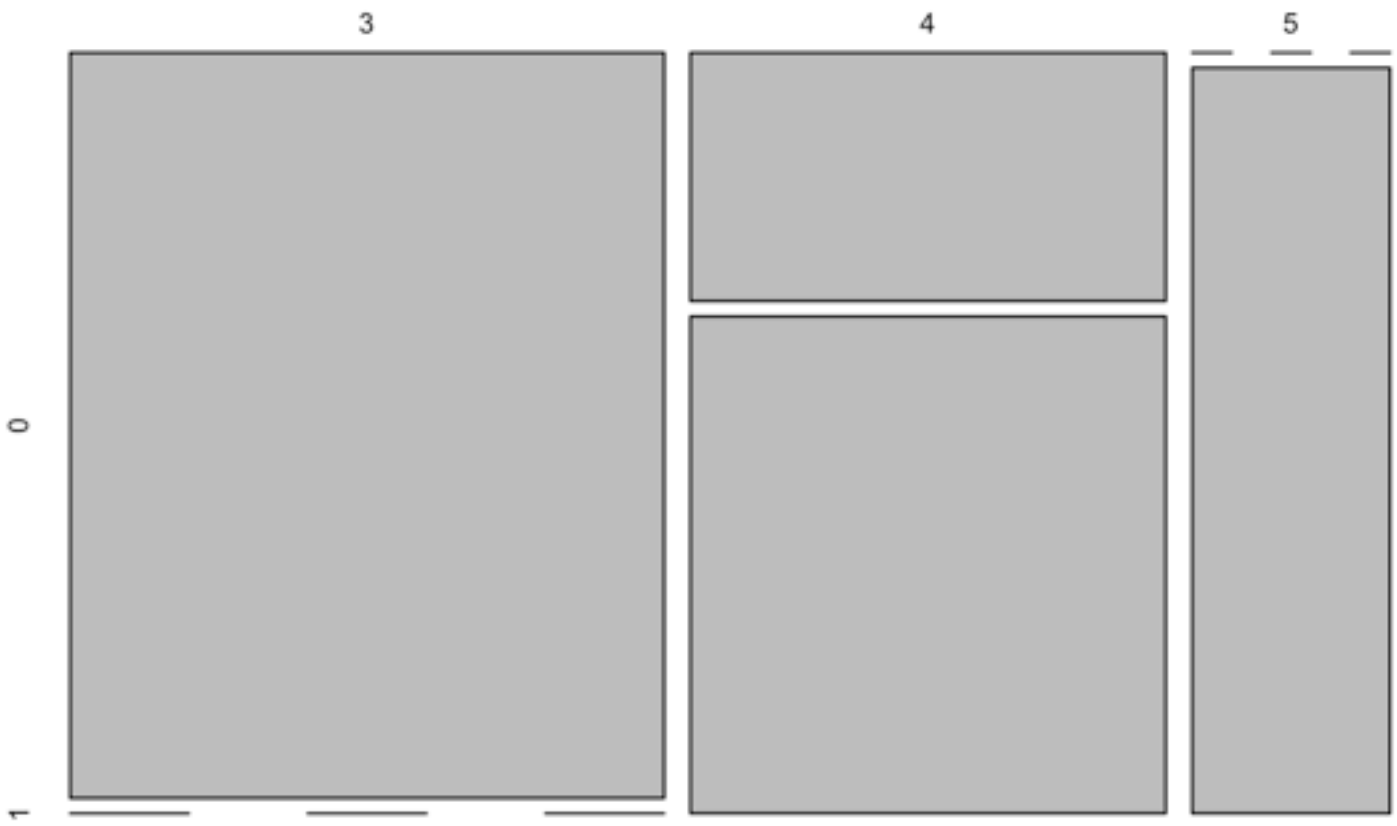
```
plot(data.frame(x = rnorm(n = 200,100,15),
                control = rnorm(n = 200,110,16),
                experiment = rnorm(n = 200,130,13)))
```



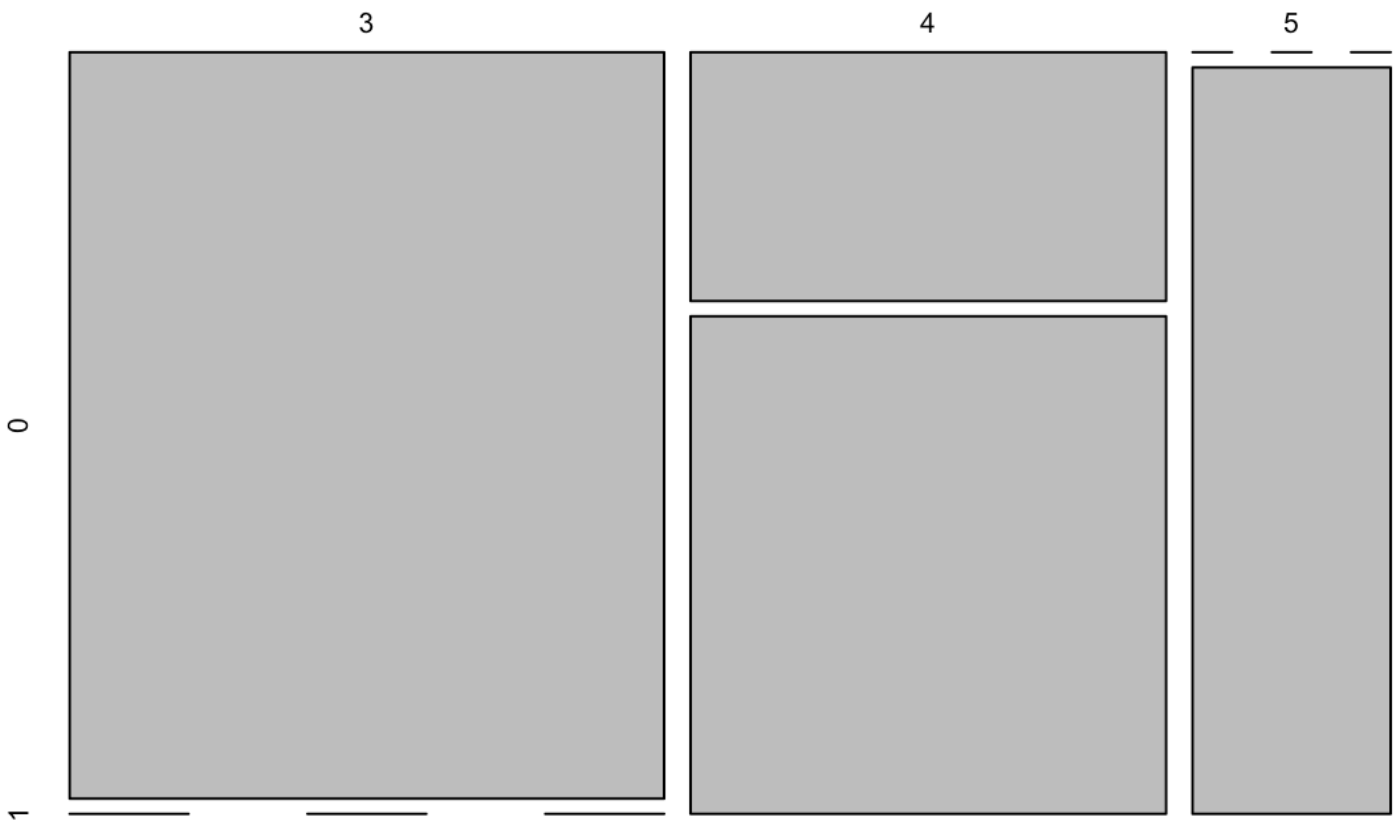

```
data.frame(x = rnorm(n = 200,100,15),
                control = rnorm(n = 200,110,16),
                experiment = rnorm(n = 200,130,13)) %>% plot()
```




```
plot(table(mtcars$gear,mtcars$am))
```




```
mtcars$gear %>%
                table(.,mtcars$am) %>%
                plot()
```



蒙特•卡罗积分

```
set.seed(4443451)
cohend <- 2
# 写一个计算overlap的函数方便后续使用
overlap <- function(x) {
    pmin(dnorm(x, 0, 1), dnorm(x, cohend, 1))
}

# 产生随机的观测值（实验组和控制组）
n <- 10000
d <- data.frame(x = runif(n, -5, 5.5),
                control = runif(n, min = 0, max = 0.4),
                experiment = runif(n, min = 0, max = 0.4)) # 从同一个分布中抽取样本
str(d)
```

```
## 'data.frame':    10000 obs. of  3 variables:
##  $ x         : num  -0.126 -2.792 -2.642 -3.352 -3.214 ...
##  $ control   : num  0.268 0.207 0.171 0.122 0.161 ...
##  $ experiment: num  0.35331 0.32134 0.32736 0.00143 0.24685 ...
```

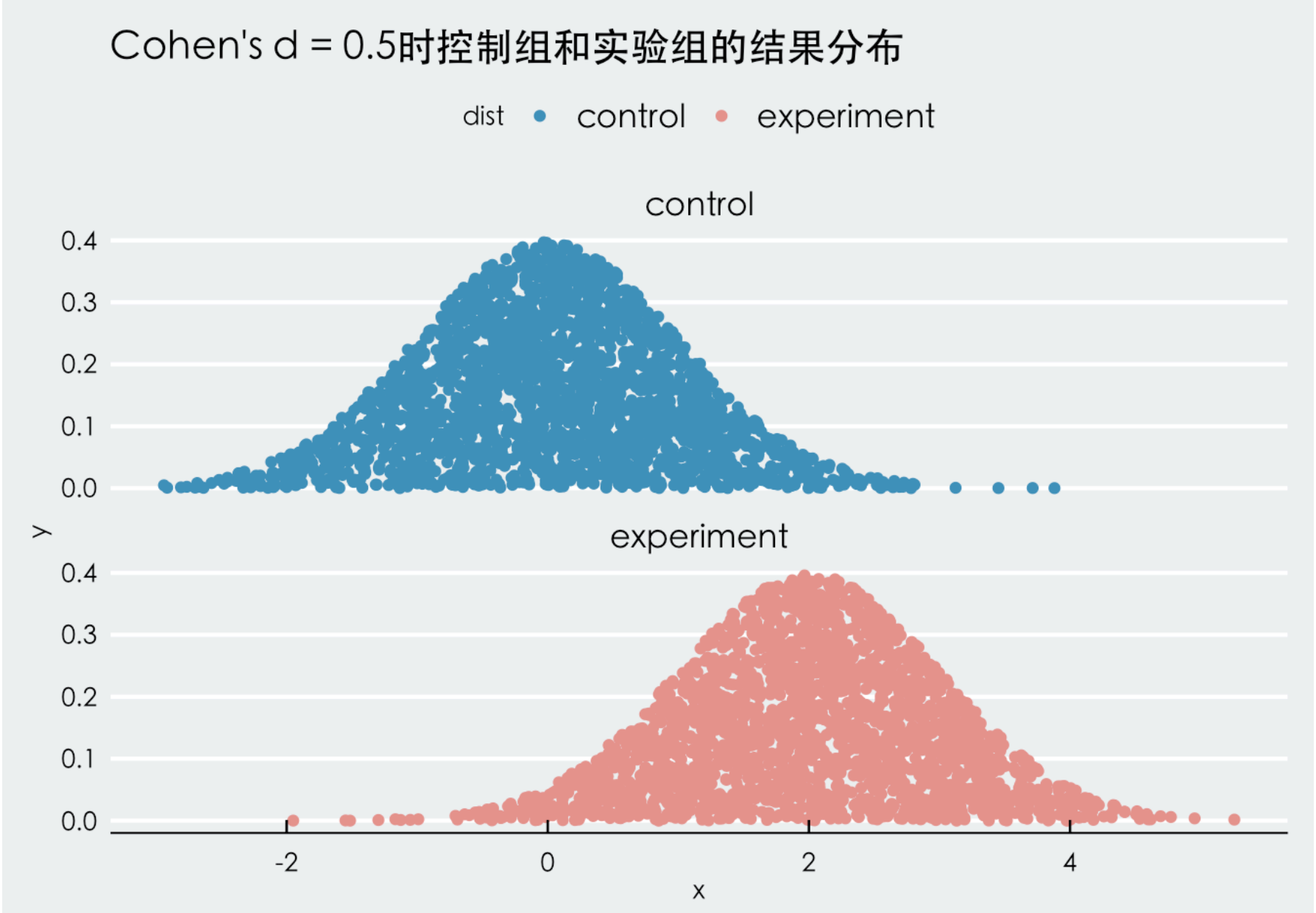
```
# 把这两个分布弄成效应量为2的两个分布
d <- d %>%
        mutate(control = ifelse(control <= dnorm(x, 0, 1), control, NA),
               experiment = ifelse(experiment <= dnorm(x, cohend, 1), experiment,
NA))

# d <- mutate(d,control = ifelse(control <= dnorm(x, 0, 1), control, NA),
#               experiment = ifelse(experiment <= dnorm(x, 0.5, 1), experiment, NA
))

str(d)
```

```
## 'data.frame':    10000 obs. of  3 variables:
##  $ x         : num  -0.126 -2.792 -2.642 -3.352 -3.214 ...
##  $ control   : num  0.268 NA NA NA NA ...
##  $ experiment: num  NA NA NA NA NA ...
```

```
d_long <- d %>%
    gather(dist, y, -x) %>%
    mutate(overlap = ifelse(y <= overlap(x), paste(dist, "_overlap", sep = ""), di
st),
           overlap = factor(overlap)) %>%
    filter(!is.na(y))
```

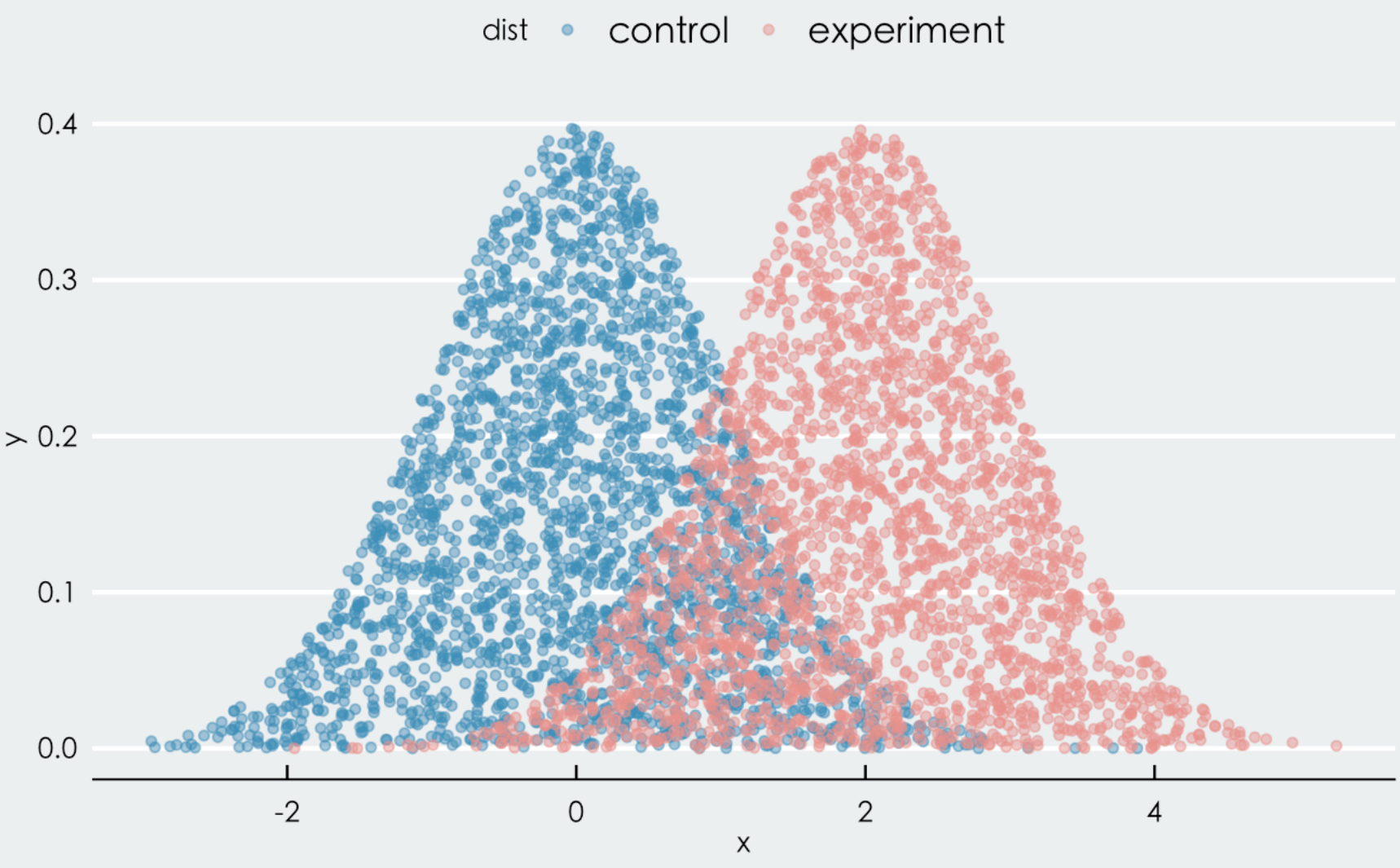
画出我们自己做出来的两个分布

```
d_long %>%
    ggplot(aes(x, y, color = dist)) +
    geom_point() +
    facet_wrap(~ dist, ncol = 1) +
    labs(title = "Cohen's d = 0.5时控制组和实验组的结果分布",
         subtile = "每个点代表了1个观测值") +
  theme(text = element_text(family = "STHeiti")) +
  scale_color_rpsy
```



画出重叠图

```
d_long %>%
    ggplot(aes(x, y, color = dist)) +
    geom_point(alpha = 0.5, size = 1.3) +
    labs(title = "两个分布的重叠部分", subtitle = "Cohen's d = 0.5") +
    theme(text = element_text(family = "STHeiti")) +
     scale_color_rpsy
```

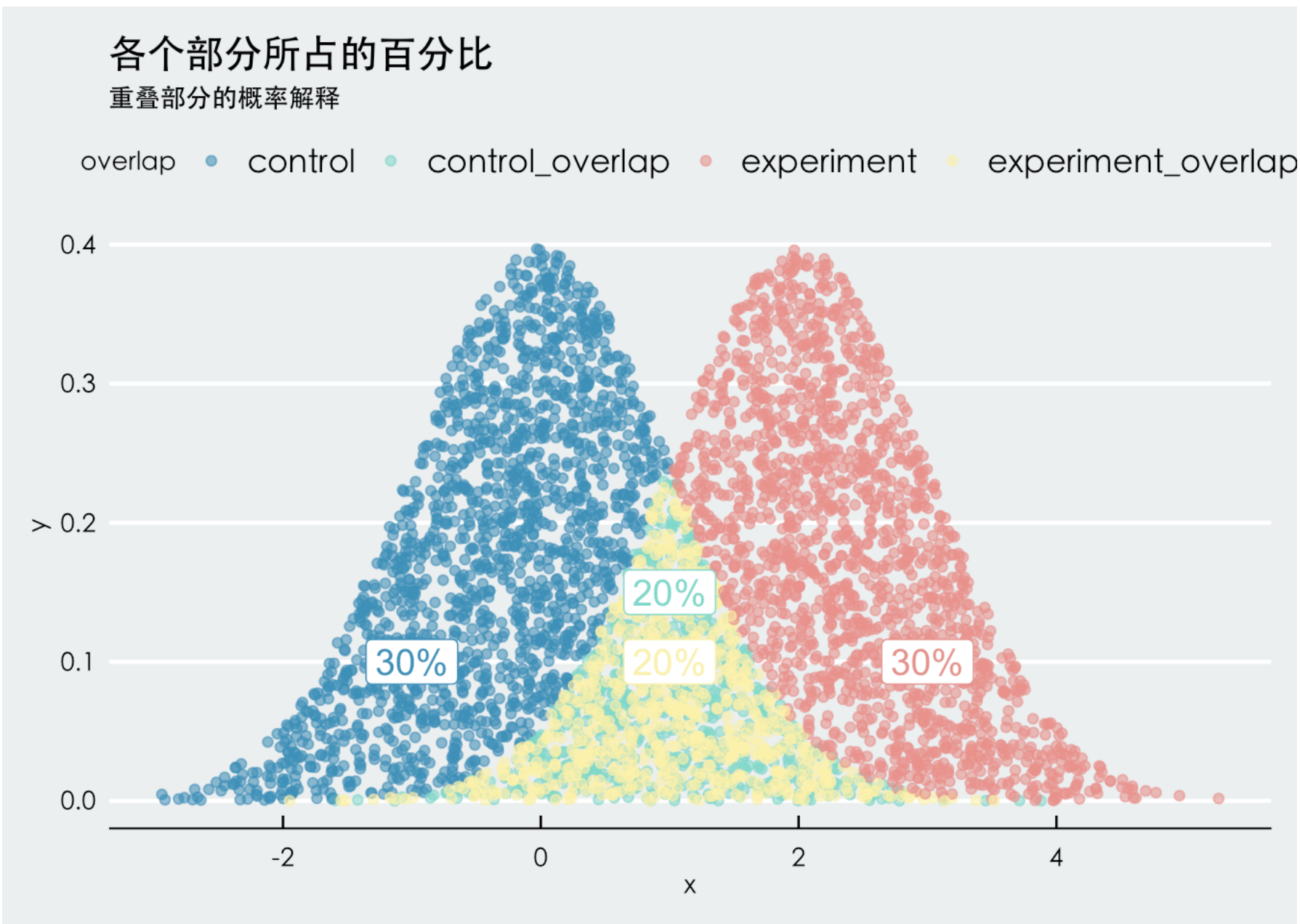


对不同的部分上不同的颜色方便查看

```
# 再次介绍一下这种代码组织方式
mutate(summarise(group_by(d_long,overlap),n = n()),
          prop = n/sum(n),
          prop = paste(round(prop, 1)*100, "%", sep = ""),
          x = c(1.5, 0.25, -1, 0.25),
          y = c(0.2, 0.25, 0.2, 0.2))
```

```
## # A tibble: 4 x 5
##   overlap                n prop      x     y
##   <fct>              <int> <chr> <dbl> <dbl>
## 1 control             1595 30%    1.5   0.2
## 2 control_overlap      780 20%    0.25  0.25
## 3 experiment          1605 30%   -1     0.2
## 4 experiment_overlap   783 20%    0.25  0.2
```

```
# 上面那条就等价于下面这个
labels <- d_long %>%
    group_by(overlap) %>%
    summarise(n = n()) %>%
    mutate(prop = n/sum(n),
           prop = paste(round(prop, 1)*100, "%", sep = ""),
           x = c(-1, 1, 3, 1),
           y = c(0.1, 0.15, 0.1, 0.1))

d_long %>%
    ggplot(aes(x, y, color = overlap)) +
    geom_point(alpha = 0.6) +
    geom_label(data = labels,
               aes(x=x, y = y, label = prop, color = overlap),
               vjust = "center", show.legend = FALSE, size = 5) +
    labs(title = "各个部分所占的百分比", subtitle = "重叠部分的概率解释") +
    theme(text = element_text(family = "STHeiti")) +
    scale_color_rpsy
```



这里计算百分比的办法：观测值的个数 / 总数 所以就是20 + 20 = 40%

Cohen计算重叠部分使用的分布

面积计算而非频率计算，所以我们需要把重叠部分的点弄掉一半

```
labels <- d_long %>%
    filter(overlap != "experiment_overlap") %>%
    group_by(overlap) %>%
    summarise(n=n()) %>%
    mutate(prop = n/sum(n),
           prop = paste(round(prop, 2) * 100, "%", sep = ""),
           x = c(-0.5, 1, 2.5))

d_long %>%
    filter(overlap != "experiment_overlap") %>%
    ggplot(aes(x, y, color = overlap)) +
    geom_point(alpha = 0.5, size = 1.3) +
    geom_label(data = labels,
               aes(x=x, y = 0.15, label = prop, color = overlap),
               vjust = "center", show.legend = FALSE, size = 5) +
    labs(title = "被两个群体分布都cover了的比例",
         subtitle = "Cohen对于overlap的比例解释") +
    theme(text = element_text(family = "STHeiti")) +
    scale_color_rpsy
```