统计检验力（功效）分析共同笔记

Open Science China 共同写作群

2020-07-31

# 0 工作起点

digraph hierarchy {  
 nodesep=1.0 // increases the separation between nodes  
  
 node [color=Red,fontname=Courier,shape=box] //All nodes will this shape and colour  
 edge [color=Blue, style=dashed] //All the lines look like this  
  
 我要做那种power\_analysis->{posthoc prior sensitivity}  
 prior->{有预期效应量及检验力 有预期效应量及样本数}  
 sensitivity->{有预期样本数及检验力}  
 {rank=same;有预期效应量及检验力 有预期效应量及样本数 有预期效应量及检验力 有预期样本数及检验力} // Put them on the same level  
}

# 1 研究设计(依变量类型分类)

* 参考[statkat](https://statkat.com/stattest_overview.php)

依变量或第二变量

独变量或第一变量

有两组被试单一类别变量

有三组以上被试单一类别变量

第一或第二变量是名义尺度

第一或第二变量是连续尺度

所有变量都是连续尺度

无

相依样本

有两组被试单一类别变量

有三组以上被试单一类别变量

# 2 可用软件

## 2.1 G\*Power

## 2.2 jPower

## 2.3 ANOVAPower

ANOVApower 的目的是帮助研究者容易简便地模拟 ANOVA 实验设计，并且可以计算相应的统计检验力。

### 2.3.1 ANOVApower R 包的安装

在 R 中运行

# devtools::install\_github("Lakens/ANOVApower") (先注释掉了，如有需要请运行)

（如果尚未安装 devtools，请先在 Packages 中安装）

### 2.3.2 ANOVA\_design 函数

使用 ANOVApower 这个 R 包，需要在 ANOVA\_design 这个函数中设定相关的参数。目前，ANOVA\_design 这个函数可以用来构建包含**最多3个因素**的实验设计，包括被试内、被试间和混合实验设计。在构建的过程中，需要设定 design, n, mu, sd, r 和 labelnames 这几个参数，其中labelnames 是非强制设定的参数。

参数具体含义如下：

1. design：用字符串来指出你所构建的实验设计（见下）
2. n：每个被试间条件下的样本量
3. mu：每一个条件下的均值
4. sd：总体标准差，假设满足方差齐性的假设（只能设定一个标准差的值）
5. r：被试内设计中的相关（被试间为 0）
6. labelnames：非强制设定参数，用来表示因素名称和因素水平名称（见下）
7. 最后一个可选设定项为是否输出图表（plot = TRUE or FALSE)

#### 2.3.2.1 design：用字符串来指出你所构建的实验设计

1. 每一个因素用一个数字和一个字母来表示。数字代表这个因素拥有的水平的个数，字母代表这个因素是被试间因素还是被试内因素。
2. 举例：

2b 实验设计代表在这个实验中有1个被试间因素，这个因素有2个组别  
  
12w 实验设计代表一个被试内因素有12个水平  
  
2b\*3w 实验设计代表有两个因素，一个因素为被试间因素(例如药片)，有2个组别(例如，一组为药物，另外一组为安慰剂)，另外一个因素为被试内因素(例如被试健康程度)，有3个水平(例如，分别在服用药片前、服用药片后一天和服用药片后一周测量)  
  
注：ANOVApower 可最多包含3个因素，用\*号隔开

#### 2.3.2.2 mu：每一个条件下的均值

1. 在 2b\*3w 例子中，有 6 种不同的条件，所以需要设定 6 个均值。例如，

|  |  |  |  |
| --- | --- | --- | --- |
|  | 服用药片前 | 服用药片后一天 | 服用药片后一周 |
| 药物 |  |  |  |
| 安慰剂 |  |  |  |

1. 均值需要以正确的顺序输入。在 ANOVA\_design 的输出中包括了一个图，你可以用这个图来检查输入的每组的均值是否正确。
2. 一般的规则为使用程序代码来形成因素，由labelnames（见下）中输入的因素名称表示，例如condition和time。因素的水平由因素的名称和水平来共同表示，例如 drug\_time1, drug\_time2, drug\_time3, control\_time1, control\_time2, and control\_time3。
3. 如果你的实验设计中只有一个因素，你只需按照在labelnames（见下）中同样的顺序输入对应的均值。如果你的实验设计中有多个因素，通常设置如下例所示：

如果实验设计中有三个因素，而每个因素都有两个水平，顺序如下：  
1. a1b1c1  
2. a1b1c2  
3. a1b2c1  
4. a1b2c2  
5. a2b1c1  
6. a2b1c2  
7. a2b2c1  
8. a2b2c2  
如果这8个均值分别为 mu = c(1, 2, 3, 4, 5, 6, 7, 8)，则如下图所示：

均值示例

均值示例

(来自 <https://raw.githubusercontent.com/Lakens/ANOVApower/master/README_files/figure-gfm/unnamed-chunk-1-1.png>)

#### 2.3.2.3 labelnames：非强制设定参数，用来表示因素名称和因素水平名称

1. 为因素和水平命名可以提高模拟结果的图表的易读性
2. 举例，上文提到的 2\*3 实验设计的相应代码为：

c("condition", "placebo", "medicine", "time", "time1", "time2", "time3")  
# 首先写一个因素的名称，其后写该因素的不同水平；然后写另外一个因素的名称，其后写该因素的不同水平  
# 注意：因素名称或水平中不要包含空格

1. 其它例子：

c("time", "morning", "evening") # 一个被试内因素(time, 2 水平)，2w  
c("time", "morning", "evening", "group", "control", "experimental") # 两个被试间因素(time 和 group，分别包含 2 水平)， 2b\*2b  
c("time", "morning", "afternoon", "evening", "night", "group", "control", "experimental") # 两个被试间因素(time 和 group，第一个有 4 个水平，第二个有 2 个水平)，4b\*2b

#### 2.3.2.4 r：被试内设计中的相关（被试间为 0）

1. 根据实验的因素是被试内因素还是被试间因素，我们需要决定变量之间是否相关及相关程度。例如，我们可以设定被试内因素之间的相关程度，既可以假设所有的被试内因素之间的相关程度相等（例如 r = 0.7），也可以借助相关矩阵设定每一对之间的相关性
2. 假设在一个 2\*2 被试内实验设计中，有两个被试内因素 A 和 B，每一个因素有 2 个水平，所有会有 6 组相关的数值需要设定；如果实验设计为 2\*2\*4 被试内实验设计，则有120组相关的数值需要设定
3. ANOVApower 的优势：大部分的统计检验力分析的工具都只能包含一个重复测量的因素 (Potvin & Schutz, 2000)，而 ANOVApower可以包含最多 3 个；还有一些统计检验力分析工具只能将被试内因素及不同水平间的相关设置为同样的数值，而 ANOVApower 可以借助相关矩阵来设定每一组的数值
4. 相关数值的设定顺序，例如在 2\*2 的被试内实验设计中，我们需要设定6个相关数值：

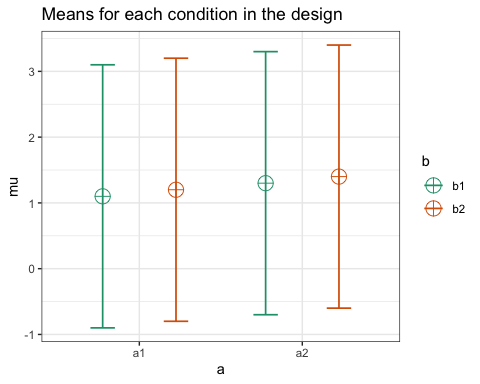
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 因素 | a1\_b1 | a1\_b2 | a2\_b1 | a2\_b2 |
| a1\_b1 | 1.00 | 0.91 | 0.92 | 0.93 |
| a1\_b2 | 0.91 | 1.00 | 0.94 | 0.95 |
| a2\_b1 | 0.92 | 0.94 | 1.00 | 0.96 |
| a2\_b2 | 0.93 | 0.95 | 0.96 | 1.00 |

然后使用自上向下、自左向右的顺序在代码中输入以上数值：

library(ANOVApower)

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

design\_result <- ANOVA\_design(design = "2w\*2w",  
 n = 80,  
 mu = c(1.1, 1.2, 1.3, 1.4),  
 sd = 2,  
 r <- c(.91, .92, .93, .94, .95, .96))



# 顺序是 a1\_b1, a1

我们可以通过检查由design\_result得到的相关矩阵来检验我们的输入是否正确：

design\_result$cor\_mat

## a1\_b1 a1\_b2 a2\_b1 a2\_b2  
## a1\_b1 1.00 0.91 0.92 0.93  
## a1\_b2 0.91 1.00 0.94 0.95  
## a2\_b1 0.92 0.94 1.00 0.96  
## a2\_b2 0.93 0.95 0.96 1.00

并且由此可以看出，我们并没有设定这些labelnames,他们是由程序自动设定的。

#### 2.3.2.5 设定样本量

你可以通过赋值给n来为每种实验条件设定样本量。此举的假设为你可以针对每种不同的实验条件来收集到相等的样本量。【在未来的ANOVApower中会加入可为不同实验条件设定不同的样本量的功能】

#### 2.3.2.6 设定标准差

你可以通过赋值给sd来设定标准差的值。当前ANOVApower只允许在满足方差齐性的假设（即在不同的实验条件下标准差相同）下来计算。需要注意的是在你计划做一个研究的时候，针对你所预测的变量的值，总会出现一些不确定性。因此，sensitivity analyses是非常有帮助的（例如在使用预期的标准差的值来做仿真模拟，但是使用更保守或者最坏情况下的值的时候）。

### 2.3.3 基于仿真模拟的统计检验力计算

基于仿真模拟，计算ANOVA设计的统计检验力有两种不同的方法。第一种方法为基于平均数、样本量、标准差和相关系数，在特定的统计假设下（例如正态分布的数据）针对每一种实验情况重复模拟数据。ANOVA\_power函数允许基于重复模拟正态分布的数据来进行统计检验力分析。第二种方法为模拟一个拥有*十分精确的*、具有所期望的特征的数据集 - 每一个设计的单元都有n个具有我们想要的均值、标准差和组间相关系数（针对被试内设计）的数据点。通过对此数据集执行ANOVA，我们可以计算用来做统计检验力分析所需要的统计量。ANOVA\_exact函数允许你基于此方法来计算统计检验力。相比而言，ANOVA\_power函数更加地灵活（例如我们可以用不同的方法来纠正多重检验，或者在以后的版本中会增加允许你进行模拟更接近现实的不为正态分布的数据集的功能），但是ANOVA\_exact函数速度更快（只需数秒，而不像其他大型的模拟需要数分钟或数小时）。

### 2.3.4 ANOVA\_power函数

ANOVA\_power函数从ANOVA\_design函数中获取结果，然后在特定的alpha水平下来模拟nsims次数据。在结果中会呈现ANOVA结果的表格和所有独立比较的结果。

需要如下输入：

1. design\_result: ANOVA\_design函数的输出结果被保存作为一个目标
2. alpha\_level: alpha水平用来设定统计显著性水平
3. p\_adjust: 使用p.ajust函数来对多重比较做出修正
4. nsims: 执行的模拟的次数
5. seed: 为可重复性结果设定seed
6. verbose: 如果不打印结果则设定为FALSE，默认值为TRUE

模拟通常会花费一段时间。更多次数的模拟会使我们得到更为精确的结果，但也需花费更长的时间。我们建议将测试设定正确与否的模拟次数设为100次，如果设定正确就将模拟次数设置为1000次（或者10000次，如果你要去喝杯咖啡；或者100000次，如果你需要小数点后更为精确的数值;))

### 2.3.5 ANOVA\_exact函数

ANOVA\_exact函数从ANOVA\_design函数中引入结果，并且模拟一个可精确匹配所需要的特征的数据集（使用MASS包中的mvrnorm函数，设定为empirical = TRUE）。这个数据集可以用来执行单个的ANOVA，并且结果可以用计算统计检验力（感谢Chris Aberson为此方法提供了灵感）。

ANOVA\_exact需要如下输入：ANOVA\_design, alpha\_level和verbose。

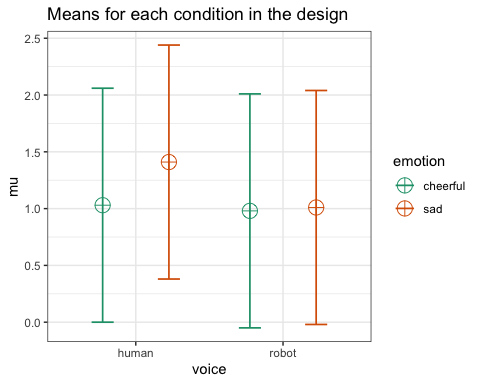
1. design\_result: 将ANOVA\_design的输出结果保存为一个目标
2. alpha\_level: alpha水平用来设定统计显著性水平
3. verbose：如果不打印结果则设定为FALSE，默认值为TRUE

与ANOVA\_power函数相比，ANOVA\_exact的方法要快很多（只需模拟一个单个的数据集）。当前唯一的不同是ANOVA\_exact不允许你来检验纠正多重比较的结果（没有p\_adjust的选项），并且ANOVA\_exact不适用与样本量小于8的情况（因为针对特别小的样本量，模拟一个拥有所需要的特征的数据集几乎不可能）。

#### 2.3.5.1 例子

假如你现在计划执行一个研究，在这个研究中，被试与人工语音助手进行互动。人工语音助手拥有两个变量 - 声音（听起来像人类 vs 听起来像机器人）和情绪（高兴 vs 沮丧）。在如下例子中，为这个2\*2的混合实验设计来执行1000次模拟（声音为被试间变量，情绪为被试内变量）。在每个被试间的实验条件下，样本量为40，所以总被试量为80。假设总体标准差为1.03，被试内因素之间的相关为0.8，平均数分别为1.03，1.21，0.98和1.01。没有针对多重比较做出修正。在模拟中使用的alpha水平为0.01。

design\_result <- ANOVA\_design(design = "2b\*2w",  
 n = 40,   
 mu = c(1.03, 1.41, 0.98, 1.01),   
 sd = 1.03,   
 r=0.8,   
 labelnames = c("voice", "human", "robot", "emotion", "cheerful", "sad"))



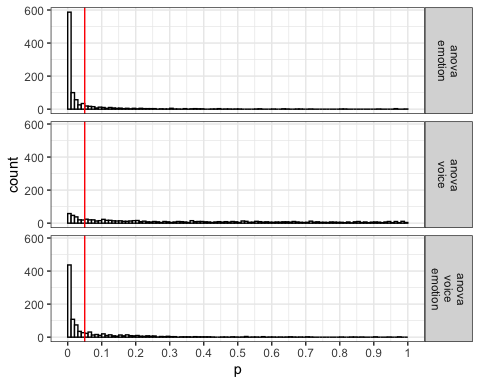
nsims <- 1000 # 注：此处根据实际需要修改，例如上文提到的先设定成100，如果正确的话再修改为1000  
power\_result\_vig\_1 <- ANOVA\_power(design\_result,   
 alpha = 0.05,   
 nsims = nsims,   
 seed = 1234)

## Power and Effect sizes for ANOVA tests  
## power effect\_size  
## anova\_voice 18.2 0.02612  
## anova\_emotion 80.4 0.10113  
## anova\_voice:emotion 68.1 0.08262  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_voice\_human\_emotion\_cheerful\_voice\_human\_emotion\_sad 95.5 0.60464  
## p\_voice\_human\_emotion\_cheerful\_voice\_robot\_emotion\_cheerful 5.1 -0.04635  
## p\_voice\_human\_emotion\_cheerful\_voice\_robot\_emotion\_sad 5.3 -0.02273  
## p\_voice\_human\_emotion\_sad\_voice\_robot\_emotion\_cheerful 47.1 -0.42257  
## p\_voice\_human\_emotion\_sad\_voice\_robot\_emotion\_sad 41.9 -0.39998  
## p\_voice\_robot\_emotion\_cheerful\_voice\_robot\_emotion\_sad 4.9 0.03820  
##   
## Within-Subject Factors Included: Check MANOVA Results

# reduce file size for github - only save what is needed  
power\_result\_vig\_1$sim\_data <- NULL  
#power\_result\_vig\_1$plot1 <- NULL  
power\_result\_vig\_1$plot2 <- NULL  
# 根据实际需要进行修改

saveRDS(power\_result\_vig\_1, file = "power\_result\_vig\_1.rds")  
power\_result\_vig\_1 <- readRDS(file = "power\_result\_vig\_1.rds")  
#可根据实际需要修改路径

power\_result\_vig\_1$plot1



统计检验力模拟的结果输出如上。结果主要分为2个部分（可以通过将verbose设定为FALSE将结果压缩显示）。第一个表格提供了ANOVA的统计检验力（从0到100%）和效应量（partial eta-squared）。我们可以看到针对声音、情绪和声音\*情绪的主效应的结果。

统计检验力模拟的结果显示情绪的主效应的统计检验力最高。因为情绪是被试内因素,并且平均数高度相关（0.8），所以针对被试内的对比，我们拥有较高的统计检验力。针对交互作用，统计检验力较低，并且声音的主效应非常低。

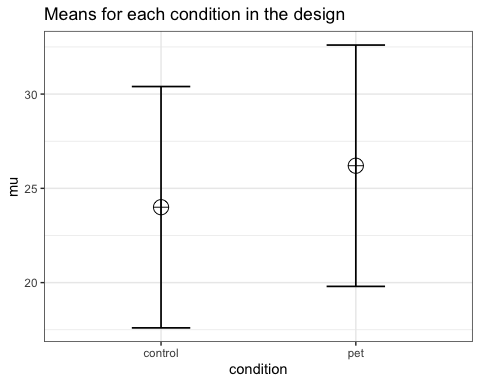
ANOVA通常的后续为事后比较。统计假设通常不只预测交互作用存在，还预测交互作用的方向。如上图所示，我们可能只对高兴和悲伤的类人声声音助理的独立效应和悲伤的类机器人和类人的声音的差异感兴趣。第二个表格显示了针对所有事后比较的*t*检验的统计检验力和效应量（针对被试间对比为Cohen’s d，针对被试内对比为Cohen’s ，针对被试内对比，见Lakens, 2013）。

通常针对被试内因素，不同实验条件下的差异的统计检验力比较高， 而对于相似的平均数（例如，1.03，0.98和1.01）之间的微小差异的统计检验力非常低。除以上两个表格之外，ANOVA\_power函数可以返回初始的模拟数据（针对每一个模拟的*p*值和效应量，见simulation\_result$sim\_data）和包含所有ANOVA检验中的*p*值分布的图表。

### 2.3.5.2 两组 One-Way ANOVA的统计检验力

假设我们计划设计一个研究来检验如下假设：照顾宠物可以提升人们的生活满意度。我们有一个控制组和一个实验组（被试有一个需要照顾的宠物），将被试随机分配到这两种组别中。我们可以模拟一个设定了alpha，样本量和效应量的One-Way ANOVA，来计算相应的统计检验力。我们预期相比于控制组，实验组中的宠物可以提升被试的生活满意度。基于Pavot和Diener（1993）的工作成果，我们相信在总体中针对生活满意度的测量均值约为24，标准差约为6.4。我们期望针对拥有宠物的被试，宠物约可以提升2.2的生活满意度分数。我们计划一共收集200个被试的数据，每一个实验条件下随机分配100名被试。我们通过如下步骤来检验这个实验设计的统计检验力：

design <- "2b"  
n <- 100  
mu <- c(24, 26.2)  
sd <- 6.4  
labelnames <- c("condition", "control", "pet") #  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,  
 mu = mu,   
 sd = sd,   
 labelnames = labelnames)



nsims <- 1000 # 注：此处根据实际需要修改，例如上文提到的先设定成100，如果正确的话再修改为1000  
  
power\_result\_vig\_2 <- ANOVA\_power(design\_result, nsims = nsims, seed = 1234)

## Power and Effect sizes for ANOVA tests  
## power effect\_size  
## anova\_condition 68.1 0.03432  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_pet 68.1 0.352

#   
# reduce file size for github - only save what is needed  
power\_result\_vig\_2$sim\_data <- NULL  
# power\_result\_vig\_2$plot1 <- NULL  
power\_result\_vig\_2$plot2 <- NULL  
saveRDS(power\_result\_vig\_2, file = "power\_result\_vig\_2.rds")  
power\_result\_vig\_2 <- readRDS(file = "power\_result\_vig\_2.rds")  
  
#Note we do not specify any correlation in the ANOVA\_design function (default r = 0), nor do we specify an alpha in the ANOVA\_power function (default is 0.05)

结果显示针对ANOVA，我们有和*t*检验一样的统计检验力结果，原因是在这个例子中只有2组被试。在这个研究中，每一组有100名被试，相应的统计检验力相对来说很低（在67.7%左右）。

只有两组被试的ANOVA和*t*检验是一样的，在上述例子中，针对控制组和实验组差异的Cohen’s d（标准均差）是d = 2.2/6.4 = 0.34375，可以很容易地通过pwr包来计算简单比较的预期的统计检验力。

library(pwr)  
pwr.t.test(d = 2.2/6.4,  
 n = 100,  
 sig.level = 0.05,  
 type="two.sample",  
 alternative="two.sided")$power

## [1] 0.6768572

我们也可以通过两个组别的Cohen’s d来直接计算Cohen’s f，正如Cohen（1988）所描述，。所以。事实上，使用pwr包中pwr.anova.test所计算的统计检验力与power.t.test结果一致。

pwr.anova.test(n = 100,  
 k = 2,  
 f = 0.171875,  
 sig.level = 0.05)$power

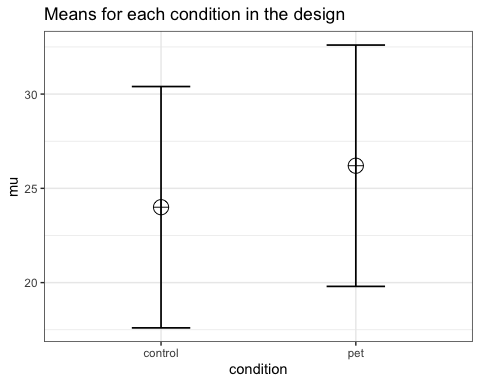
## [1] 0.6768572

以上分析表明使用每组包含100名被试的实验设定不太会导致显著的统计结果，即使我们所期望的差异是真实存在的。这显然不是最优的结果。如果我们进行一项研究，我们希望可以得到信息丰富的结果，也就是说如果效应存在，我们就应该可以探测到所期望的或者所关心的这个效应；如果效应不存在，我们就应该探测不到这个效应。

如果想达到足够的统计检验力，我们究竟需要多少被试呢？考虑到预期的差异和标准差，d = 0.34375，f = 0.171875。我们可以增大样本量进而进行模拟，直到满意为止。统计检验力与二型错误的概率密切相关。Cohen建议将所寻求的统计检验力的标准设为80%，即二型错误的概率为20%。我个人认为20%的二型错误概率太高，并且与设定地更加严格地一型错误概率（尽管你可以调整alpha，但通常设定为5%）相比并不相称。假设我们在这里想达到90%的统计检验力。

当你在增加样本量来找出究竟实验设计中需要多少被试的时候，使用ANOVA\_exact函数更为有效率。你可以看到在如下代码中我们慢慢地增加样本量，发现180个被试时可达到略高于90%的统计检验力。

design\_result <- ANOVA\_design(design = "2b",  
 n = 100,   
 mu = c(24, 26.2),   
 sd = 6.4,   
 labelnames = c("condition", "control", "pet"))



ANOVA\_exact(design\_result)$main\_results$power

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 67.69 0.029 0.1727 5.9082  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_pet 67.69 0.3437

## [1] 67.69

统计检验力为67.7%，此结果较低，因此我们将样本量从每组100人升至150人来看结果是否更接近我们的目标 - 统计检验力达到90%。

design\_result <- ANOVA\_design(design = "2b",  
 n = 150,   
 mu = c(24, 26.2),   
 sd = 6.4,   
 labelnames = c("condition", "control", "pet"),  
 plot = FALSE)  
  
ANOVA\_exact(design\_result)$main\_results$power

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 84.31 0.0289 0.1725 8.8623  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_pet 84.31 0.3437

## [1] 84.31

此次统计检验力达到84.31%，但仍然不够，我们试一下每组175人。

design\_result <- ANOVA\_design(design = "2b",  
 n = 175,   
 mu = c(24, 26.2),   
 sd = 6.4,   
 labelnames = c("condition", "control", "pet"),  
 plot = FALSE)  
  
ANOVA\_exact(design\_result)$main\_results$power

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 89.37 0.0289 0.1724 10.3394  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_pet 89.37 0.3437

## [1] 89.37

结果为89.37%，非常接近我们的目标，因此再稍添一些被试，我们来试一下每组180人的情况。

design\_result <- ANOVA\_design(design = "2b",  
 n = 180,   
 mu = c(24, 26.2),   
 sd = 6.4,   
 labelnames = c("condition", "control", "pet"),  
 plot = FALSE)  
  
ANOVA\_exact(design\_result)$main\_results$power

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 90.19 0.0288 0.1724 10.6348  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_pet 90.19 0.3437

## [1] 90.19

此次的结果为90.19%。

#### 2.3.5.3 Plot\_power函数

基于模拟的统计检验力分析需要你来增加样本量，直到统计检验力足够高来达到你想要的二型错误率。为加速这个过程，我们可以使用plot\_power函数来画出随着样本量的变化，统计检验力随之变化的曲线。

plot\_power这个函数需要如下输入：ANOVA\_design, max\_n和plot。

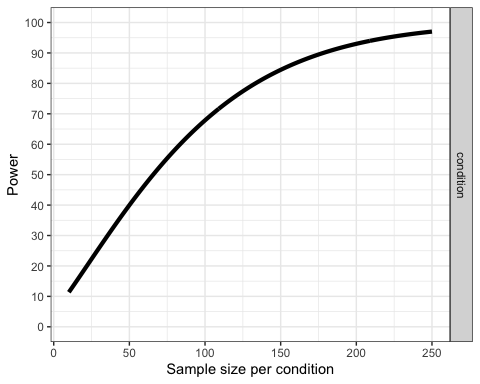
1. design\_result：由ANOVA\_design函数得到的输出（保存为一个对象）
2. min\_n：统计检验力曲线中的最小样本量
3. max\_n：统计检验力曲线中的最大样本量
4. plot: 如果不输出图像，则设定为FALSE（默认值为TRUE）

plot\_power函数可以使用ANOVA\_exact函数来模拟直至max\_n样本量的统计检验力。即使这种方法速度稍快，当样本量较大时还是需要不少时间来得出结果。如下图所示，我们可以看出，假设平均数和标准差真正的模式可代表我们所期望的平均数和标准差的模式，在每组有135名被试的情况下，我们可以得到约80%的统计检验力;在每组有180名被试的情况下，我们可以得到约90%的统计检验力；在每组有225名被试的情况下，我们可以得到约95%的统计检验力。

plot\_power(design\_result, min\_n = 10, max\_n = 250)

## Warning: Use of `plot\_data$n` is discouraged. Use `n` instead.

## Warning: Use of `plot\_data$value` is discouraged. Use `value` instead.

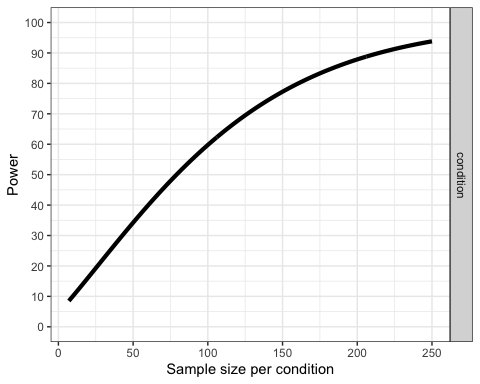


由于平均数的真正的模式是不知道的，检验一定范围内的统计检验力是非常明智的。举例来说，假如平均数之间的差别非常小（假如平均数分别为24和26，而不是24和26.2），我们可以将上图中的统计检验力曲线和此略不乐观情况下（真正的效应量稍小）的统计检验力曲线进行对比。

design\_result <- ANOVA\_design(design = "2b",  
 n = 180,   
 mu = c(24, 26),   
 sd = 6.4,   
 labelnames = c("condition", "control", "pet"),  
 plot = FALSE)  
  
plot\_power(design\_result, max\_n = 250)

## Warning: Use of `plot\_data$n` is discouraged. Use `n` instead.

## Warning: Use of `plot\_data$value` is discouraged. Use `value` instead.

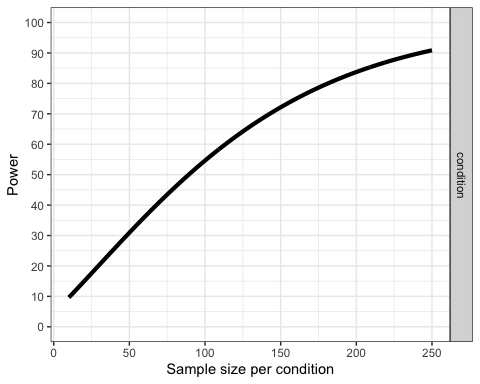


除效应量可能较小之外，标准差也可能比我们所期望的稍大。这也会进一步降低统计检验力，并且这比上面提到的情况更为不理想。我们假设真正的标准差为6.8而不是6.4。

design\_result <- ANOVA\_design(design = "2b",  
 n = 180,   
 mu = c(24, 26),   
 sd = 6.8,   
 labelnames = c("condition", "control", "pet"),  
 plot = FALSE)  
  
plot\_power(design\_result, min\_n = 10, max\_n = 250)

## Warning: Use of `plot\_data$n` is discouraged. Use `n` instead.

## Warning: Use of `plot\_data$value` is discouraged. Use `value` instead.

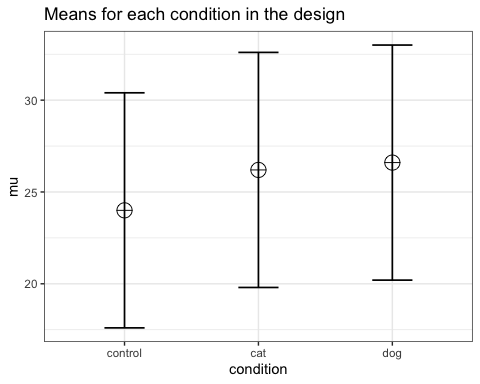


就像这些不同的图像所示，你的研究永远不可能真正地知道统计检验力。因为真正的效应量（例如，平均数和标准差的模式）未知，那么真正的统计检验力也是未知的。假设在一个特定的效应量的值下，一个研究可达到90%的统计检验力，但是如果真正的效应量与你所期望的数值不同，那么真正的统计检验力就会较高或较低。因此，当执行先验统计检验力分析时，我们应该讨论的是“期望的”统计检验力，并且应该为我们的期望提供证据（例如，被试内设计的平均数、标准差和相关的模式）。

#### 2.3.5.4 三组被试间设计的统计检验力

假如我们计划设计一项研究来检验如下假设：照顾宠物可以提升人们的生活满意度。我们有一个控制组、一个宠物为猫的实验组和一个宠物为狗的实验组。我们可以模拟一个拥有设定的alpha、样本量和效应量的One-Way ANOVA来计算ANOVA和相应的事后比较的统计检验力。我们假设和控制组相比，所有的拥有宠物组都可以提升生活满意度。并且，我们假设养狗的被试比养猫的被试的生活满意度要高。基于Pavot和Diener（1993）的研究成果，我们再一次假设在总体中针对生活满意度的测量均值约为24，标准差约为6.4。我们期望猫作为宠物可提升生活满意度约2.2个点，狗作为宠物可提升生活满意度约2.6个点。作为开始，我们收集150名被试的数据，每组50人。但是在收集数据之前，我们先计算统计检验力来看是否可以检验到我们所期望的差异。

design <- "3b"  
n <- 50  
mu <- c(24, 26.2, 26.6)  
sd <- 6.4  
labelnames <- c("condition", "control", "cat", "dog") #  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 labelnames = labelnames)



ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 47.69 0.0315 0.1804 4.7852  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_cat 39.83 0.3437  
## p\_condition\_control\_condition\_dog 52.05 0.4063  
## p\_condition\_cat\_condition\_dog 6.10 0.0625

结果显示，当每组有50个被试时，不论是总体的ANOVA（略低于50%）还是事后比较（控制组对比养猫组约为40%，控制组对比养狗组约为50%，养猫组和养狗组之间的差异的统计检验力极低，约为6%，只比一型错误率5%高一点点）统计检验力很低。我们可以很容易地使用pwr包来确认针对这些简单比较所期望的统计检验力。在我们的例子中，控制组和养猫组之间的差异的Cohen’s d（标准均差）是，控制组和养狗组之间的差异的Cohen’s d是，养猫组和养狗组之间的差异的Cohen’s d是

pwr.t.test(d = 2.2/6.4,  
 n = 50,  
 sig.level = 0.05,  
 type="two.sample",  
 alternative="two.sided")$power

## [1] 0.3983064

pwr.t.test(d = 2.6/6.4,  
 n = 50,  
 sig.level = 0.05,  
 type="two.sample",  
 alternative="two.sided")$power

## [1] 0.5205162

pwr.t.test(d = 0.4/6.4,  
 n = 50,  
 sig.level = 0.05,  
 type="two.sample",  
 alternative="two.sided")$power

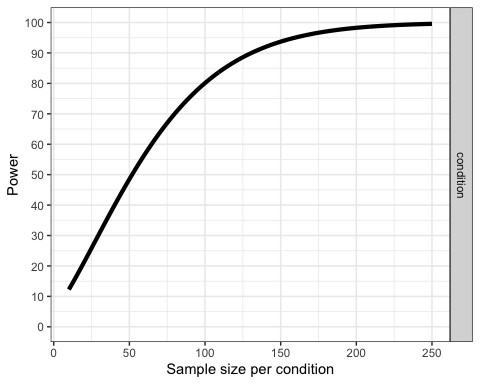
## [1] 0.06104044

很显然，我们需要更多的被试来提高统计检验力。让我们来画一个统计检验力随着样本量变化的曲线来看我们要增加多少样本。

plot\_power(design\_result, min\_n = 10, max\_n = 250)

## Warning: Use of `plot\_data$n` is discouraged. Use `n` instead.

## Warning: Use of `plot\_data$value` is discouraged. Use `value` instead.



我们可以看出在拥有134名被试时，针对ANOVA我们可以得到约90%的统计检验力。但是我们也需要检查针对控制组和养猫组的比较的统计检验力。

design <- "3b"  
n <- 134  
mu <- c(24, 26.2, 26.6)  
sd <- 6.4  
labelnames <- c("condition", "control", "cat", "dog") #  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 labelnames = labelnames,  
 plot = FALSE) #do not print the plot (same as above)  
  
ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 90.18 0.0311 0.1793 12.8242  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_cat 80.05 0.3437  
## p\_condition\_control\_condition\_dog 91.20 0.4063  
## p\_condition\_cat\_condition\_dog 8.03 0.0625

n <- 180  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 labelnames = labelnames,  
 plot = FALSE) #do not print the plot (same as above)  
  
ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 96.78 0.0311 0.1791 17.2266  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_cat 90.19 0.3437  
## p\_condition\_control\_condition\_dog 97.02 0.4063  
## p\_condition\_cat\_condition\_dog 9.09 0.0625

n <- 5380  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 labelnames = labelnames,  
 plot = FALSE) #do not print the plot (same as above)  
  
ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## condition 100 0.0309 0.1786 514.8828  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_condition\_control\_condition\_cat 100 0.3437  
## p\_condition\_control\_condition\_dog 100 0.4063  
## p\_condition\_cat\_condition\_dog 90 0.0625

当每组有134名被试时，控制组和养猫组的简单比较的统计检验力只有80%。如果我们针对事后比较都想达到90%的统计检验力，我们需要更多的被试。180名被试可以为控制组和养猫组的事后比较提供较好的统计检验力。但是养猫组和养狗组之间差异的统计检验力极低。我们需要5380名被试来使得此事后比较组达到90%的统计检验力。如果收集数据很容易，那么达到5380名被试是可能的。但是如果数据收集的代价非常高，就使得我们不可能真正得到这个问题的可靠答案。明白不可能可靠地研究一个问题是分析先验统计检验力非常重要的一个方面。

#### 2.3.5.5 重复测量ANOVA的统计检验力

在重复测量设计中，多次的测量数据是从相同的被试来收集的。最简单的情境为有两个重复测量的数据，那么重复测量ANOVA等同于非独立样本或配对样本*t*检验。让我们来看一下此中相关对于重复测量ANOVA的统计检验力的影响。

##### 2.3.5.5.1 样本内的两次测量的情况

为解释相关观测的影响，我们由非独立（或配对、或被试内）*t*检验的中等效应量开始来模拟数据。我们首先来看G\*Power。假如我们想执行一个先验统计检验力分析，我们需要先提供效应量*d\_Z*。正如Cohen（1988）所写， > 下角标Z是用来强调的初始数据的单位不再是X或者Y而是Z的事实

Z是X-Y的差异数。

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_9.png>)

因为数据之间是相关的，被试内设计比被试间设计在探测差异时拥有更高的统计检验力，并且因为每一个被试可提供多次观测数据，被试内设计需要更少的被试。独立样本*t*检验和非独立样本*t*检验的一个区别就是独立样本*t*检验的自由度为，而非独立样本*t*检验的自由度为。假设正态分布且忽视两种检验类型在自由度上的不同，两组被试内设计（NW）的样本量相对于两组被试间设计（NB）的样本量为（由Maxwell和Delaney，2004，561页，公式45）：

在公式中，除以2是因为在被试内两实验条件的设计中，每一个被试都可提供2个数据。与被试间设计相比，被试内实验设计可减少样本量的程度基于两变量的相关程度（*r*），如公式中的部分所示。如果相关为0，那么正因为每一个被试能提供两个数据点，被试内实验设计需要被试间实验设计的一半被试量（例如，需要64个而不是128个）。相关越高，被试内实验设计相对的好处越大，但是当相关为负的时候（最多至-1），这种好处就消失了。

在独立样本*t*检验中，两次观测是不相关的，而在被试内设计中，观测是相关的。这就对数据间差异的标准差产生影响。反过来，因为标准效应量是均差除以数据间差异的标准差，那么相关就会对被试内设计的标准化的均差产生影响，Cohen’s 。正如Cohen（1988，公式2.3.7）所解释的，它们之间的关系为：

因此，和d之间的关系为。正如Cohen（1988）所写：

用另一种说法来说，一个针对配对（非独立）样本的总体平均数之间的给定差异是由与其样本独立时的一样大的这个值标准化的。

如果我们将相关设定为0.5，则。当相关为0.5时，。当变量间相关变强时，例如，我们可以得到，当为1时，。反过来，，所以当时，我们可以得到。一些观点认为，当观测值都与估计效应量的“膨胀”强相关时，此相关可导致更容易地从“噪声”中区分出假设检验中的“信号”，由此又可导致数据间差异的标准差的降低，因此与d相比，的增长会导致非常明确的统计检验力上的优势。

在被试内ANOVA中，没有与Cohen’s f相对应的。针对包含两个组别的实验设计，正如Cohen（1988）所描述的，我们可以使用直接从Cohen’s d计算Cohen’s f。若，则。在G\*Power中，我们可以执行一个被试内2组实验设计ANOVA的统计检验力分析。我们将预期达到的统计检验力设为80%，通过非独立样本*t*检验来执行这个分析。之所以可行是因为当时，相关被设定为0.5，所以的转换有效。

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_1.png>)

如果我们将相关改为0.7，并且保持其他设定不变，那么重复测量的先验统计检验力可得出样本量为21。那么相关的增加导致了统计检验力的增加。

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_11.png>)

为了在G\*Power中使用非独立样本*t*检验来复现这个分析，我们需要根据来调整，得到。如果在G\*Power中的先验概率分析中输入这个值，可得到一致的结果（因为拥有两组或两次测量的重复测量ANOVA与非独立样本t检验一致）。这个例子说明变量间的相关总是会影响统计检验力分析，不论是非独立样本t检验还是重复测量ANOVA。因为非独立样本t检验使用，所以相关可能不是那么明显，但是考虑到d和之间的关系，相关一直都是在我们的考虑中，并且与被试间设计相比可以极大地为被试内设计提升统计检验力。

G\*Power示例

G\*Power示例

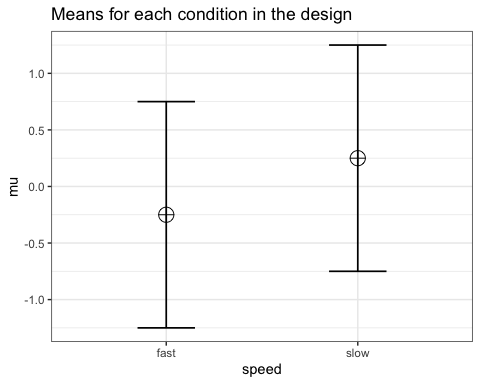
(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_10.png>)

我们也可以使用模拟来执行上述两种统计检验力分析。我们可以在模拟中将组数设置为2，（根据G\*Power可提供80.777%的统计检验力），变量重复测量的相关为0.5，alpha为0.05。在这种情况下，我们使用平均数为-0.25和0.25，标准差为1来模拟。因此，均差为0.5，Cohen’s d为。在这个例子中，我们将相关设置为0.5，结果为80.777%的统计检验力，简单效应的预估效应量为0.5。我们也可以使用计算ANOVA的偏eta方，为0.05882353。

K <- 2  
n <- 34  
sd <- 1  
r <- 0.5  
alpha = 0.05  
f <- 0.25  
f2 <- f^2  
ES <- f2/(f2+1)  
ES

## [1] 0.05882353

mu <- mu\_from\_ES(K = K, ES = ES)  
design = paste(K,"w",sep="")  
labelnames <- c("speed", "fast", "slow")  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r,   
 labelnames = labelnames)



ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## speed 80.78 0.2048 0.5075 8.5  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_speed\_fast\_speed\_slow 80.78 0.5

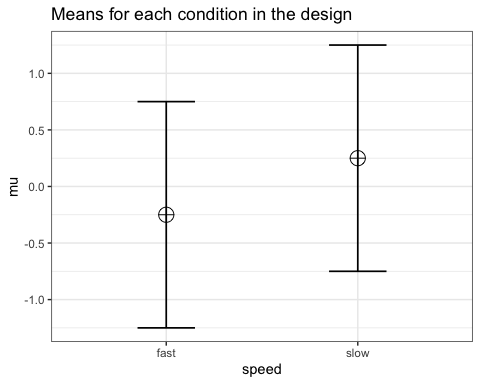
模拟的结果为80.777%。需要注意的是这个模拟为配对对比计算Cohen’s ，在此相关设定为0.5，为中等效应量。

在上述G\*Power的例子中，如果我们尝试增加相关并保持样本量一致,我们也应得到稍大的。在如下的模拟中，我们只修改了相关的值为0.7，这将使得。

K <- 2  
n <- 34  
sd <- 1  
r <- 0.7  
alpha = 0.05  
f <- 0.25  
f2 <- f^2  
ES <- f2/(f2+1)  
ES

## [1] 0.05882353

mu <- mu\_from\_ES(K = K, ES = ES)  
design = paste(K,"w",sep="")  
labelnames <- c("speed", "fast", "slow")  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r,   
 labelnames = labelnames)



ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## speed 95.47 0.3004 0.6552 14.1667  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_speed\_fast\_speed\_slow 95.47 0.6455

##### 2.3.5.5.2 多组重复测量ANOVA的统计检验力

我们来检验拥有3组被试内实验情况的重复测量实验，来解释一个重复测量的ANOVA来如何拓展非独立样本*t*检验。在两组被试内实验设计的例子中，我们提供了一个针对此类情形的特定的计算样本量的公式。假设正态分布和复合对称，并忽视两种检验中自由度的不同，相对于被试间实验设计所需要的样本量，在多于两组的被试内实验设计中的样本量为（见Maxwell和Delaney，2004，562页，公式47）：

其中a是被试内水平的个数。

##### 2.3.5.5.3 Cohen’s f和Cohen’s d之间的关系

尽管在两组重复测量ANOVA的情况下，我们可以通过我们的检验与*t*检验和Cohen’s d的对比来解释统计检验力分析的原则，但是当我们拥有多于两组的实验条件时就变得复杂了。直接解释如何计算ANOVA的统计检验力分析的Cohen’s f和效应量更为有用。根据Cohen（1988）公式8.2.1和8.2.2，Cohen’s f为：

假设我们使用拥有3组情况的被试内实验设计，自变量为我们问被试在不同条件下醒来的情绪：当闹钟叫醒时、工作日自然醒时和周末自然醒时。基于试点数据，我们预期平均数为3.8、4.2和4.3（基于7点有效的情绪量表），标准差为0.9，非独立的测量值之间的相关为0.7。我们可以计算此ANOVA的Cohen’s f和事后比较的Cohen’s 。

mu <- c(3.8, 4.2, 4.3)  
sd <- 0.9  
f <- sqrt(sum((mu-mean(mu))^2)/length(mu))/sd #Cohen, 1988, formula 8.2.1 and 8.2.2  
f

## [1] 0.2400274

r <- 0.7  
(4.2-3.8)/0.9/sqrt(2\*(1-r))

## [1] 0.5737753

(4.3-3.8)/0.9/sqrt(2\*(1-r))

## [1] 0.7172191

(4.3-4.2)/0.9/sqrt(2\*(1-r))

## [1] 0.1434438

当拥有多组时，Cohen’s d或者和Cohen’s f之间的关系变得复杂，因为此关系基于平均数的模式。Cohen（1988）讲述了三类模式的计算方法：最小变异（例如5个平均数为-0.25，0，0，0，0.25）、中等变异（例如5个平均数为-0.25，-0.25，0.25，0.25，0.25或-0.25，-0.25，-0.25，0.25，0.25）。【不知是否有最大变异】针对这三种模式可根据公式由Cohen’s d来计算Cohen’s f，其中d是由最大的均值和最小的均值的差异来计算的效应量（如果最大的均值为0.25，最小的均值为-0.25，0.25-（-0.25）=0.5.所以d是0.5除以标准差0.9）。在我们的例子中，d为（4.3-3.8）/0.9 = 0.5555556。如果我们将此值除以，我们可以得到。如果均值和标准差是已知的，可由这些值直接计算Cohen’s f。

我们可以使用模拟来计算被试内设计的统计检验力。我们将组别数设置为3，n = 20，变量间相关为0.8。如果真正的效应量f是0.25，alpha是0.05，那么统计检验力为96.9%。

在这种情况中，我们使用均值-0.3061862，0.0000000和0.3061862，并且设定标准差为1。

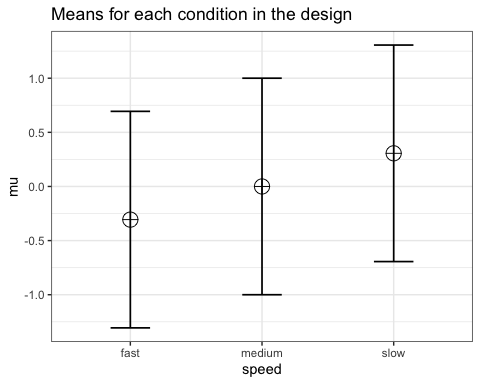
K <- 3  
n <- 20  
sd <- 1  
r <- 0.8  
alpha = 0.05  
f <- 0.25  
f2 <- f^2  
ES <- f2/(f2+1)  
ES

## [1] 0.05882353

mu <- mu\_from\_ES(K = K, ES = ES)  
sqrt(sum((mu-mean(mu))^2)/length(mu))/sd #Cohen, 1988, formula 8.2.1 and 8.2.2

## [1] 0.25

design = paste(K,"w",sep="")  
labelnames <- c("speed", "fast", "medium", "slow")  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r,   
 labelnames = labelnames)



nsims <- 1000 # 注：此处根据实际需要修改，例如上文提到的先设定成100，如果正确的话再修改为1000  
power\_result\_vig\_3 <- ANOVA\_power(design\_result, nsims = nsims)

## Power and Effect sizes for ANOVA tests  
## power effect\_size  
## anova\_speed 96.8 0.3474  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_speed\_fast\_speed\_medium 53.0 0.5012  
## p\_speed\_fast\_speed\_slow 97.9 1.0172  
## p\_speed\_medium\_speed\_slow 53.3 0.5067  
##   
## Within-Subject Factors Included: Check MANOVA Results

# reduce file size for github - only save what is needed  
power\_result\_vig\_3$sim\_data <- NULL  
# power\_result\_vig\_3$plot1 <- NULL  
power\_result\_vig\_3$plot2 <- NULL  
saveRDS(power\_result\_vig\_3, file = "power\_result\_vig\_3.rds")  
power\_result\_vig\_3 <- readRDS(file = "power\_result\_vig\_3.rds")

模拟的结果确实非常接近96.9%。我们可以看到G\*Power中一致的统计检验力预估的结果

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_12.png>)

我们甚至可以在G\*Power中使用SPSS的风格来计算Cohen’s f。我们在G\*Power中进行如上设定，然后点击’Options’，然后选择’as in SPSS’，然后点击OK，你会看到’Corr among rep measures’这个部分不见了。相关不需要单独输入，但是包含在Cohen’s f之中。Cohen’s f的原值为0.25，现为0.7024394，这个结果与SPSS的相同。目前这个值更大，并且与偏eta方相对应且包含了观测值之间的相关。

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_14.png>)

#### 2.3.5.6 被试内实验设计的2X2 ANOVA

我们可以模拟一个设定了alpha、样本量和效应量的two-way ANOVA来达到特定的统计检验力。我们可以试着在G\*Power中的*F*-test, ANOVA: Repeated measures, within-between interaction来执行统计检验力分析。

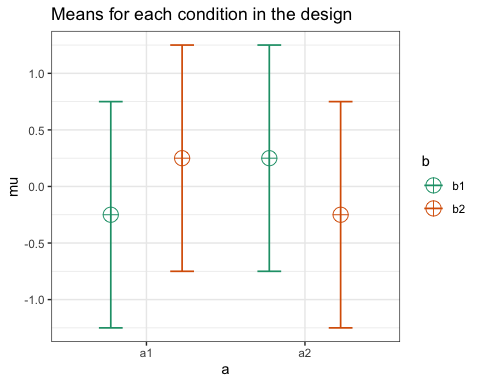
G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_5.png>)

对于2-way交互作用，如果总样本量为46，统计检验力应为91.25%。因为被试间因素有两个水平，所以我们有2组（并且每组有2次重复测量）。

mu <- c(-0.25, 0.25, 0.25, -0.25)  
n <- 23  
sd <- 1  
r <- 0.5  
design = "2w\*2b"  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r)



ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## b 5.00 0.0000 0.0000 0.0  
## a 5.00 0.0000 0.0000 0.0  
## b:a 91.25 0.2072 0.5112 11.5  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_a\_a1\_b\_b1\_a\_a1\_b\_b2 38.17 0.5  
## p\_a\_a1\_b\_b1\_a\_a2\_b\_b1 63.02 0.5  
## p\_a\_a1\_b\_b1\_a\_a2\_b\_b2 5.00 0.0  
## p\_a\_a1\_b\_b2\_a\_a2\_b\_b1 5.00 0.0  
## p\_a\_a1\_b\_b2\_a\_a2\_b\_b2 63.02 -0.5  
## p\_a\_a2\_b\_b1\_a\_a2\_b\_b2 38.17 -0.5

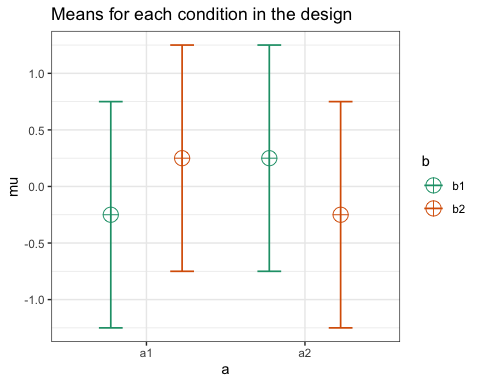
我们可以将相关调整为0.7进而模拟同样的two-way ANOVA。统计检验力应增加至98.98%。

G\*Power示例

G\*Power示例

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_6.png>)

mu <- c(-0.25, 0.25, 0.25, -0.25)  
n <- 23  
sd <- 1  
r <- 0.7  
design = "2w\*2b"  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r)



ANOVA\_exact(design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## b 5.00 0.0000 0.00 0.0000  
## a 5.00 0.0000 0.00 0.0000  
## b:a 98.98 0.3034 0.66 19.1667  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_a\_a1\_b\_b1\_a\_a1\_b\_b2 38.17 0.5000  
## p\_a\_a1\_b\_b1\_a\_a2\_b\_b1 84.09 0.6455  
## p\_a\_a1\_b\_b1\_a\_a2\_b\_b2 5.00 0.0000  
## p\_a\_a1\_b\_b2\_a\_a2\_b\_b1 5.00 0.0000  
## p\_a\_a1\_b\_b2\_a\_a2\_b\_b2 84.09 -0.6455  
## p\_a\_a2\_b\_b1\_a\_a2\_b\_b2 38.17 -0.5000

#### 2.3.5.7 被试间实验设计的two-way ANOVA

在ANOVA中的交互作用的效应量依赖于均值的模式。假如研究者计划做一项2\*2的实验：人工智能的声音是高兴或者伤心和该声音听起来更像机器人或者更像人类。我们可以期待产生不同的结果，例如不论是对于像人类的声音还是像机器人的声音，高兴的声音比伤心的声音都更让人喜欢；或者高兴的类人声比伤心的类人声更让人喜欢，而在类机器人声中高兴与伤心并无差异，或者甚至是在类机器人声中人们更喜欢伤心的声音（Marvin-the-Depressed-Robot效应）。在第一个情况中，我们将只观测声音的主效应，但是在另外两种可能的情况下，在声音的类人性和声音的情绪基调上存在交互作用。我们可以通过每组拥有80个被试的2X2的被试间实验设计的交互作用来模拟这个统计检验力（如下图所示每种实验条件的期望的均值）。

均值的模式

均值的模式

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/gpower_6.png>)

从数学上来说，交互作用是由每一个单元格的均值减去整体的均值、每一行的边缘均值减去整体的均值和每一列的边缘均值减去整体均值来计算的。例如，对于高兴的类人声的条件来说就是。由计算可得，每个单元格的值为-0.5，0.5，0.5和-0.5。Cohen’s f为。模拟的结果显示当每种实验条件下有80名被试时，统计检验力为99.379%。每个因子具有两个水平的交叉（cross-over，也称无序disordinal）交互作用与最初的两组设计拥有完全相同的统计检验力。如果针对交叉交互作用的每组样本量由80变为40，相应的统计检验力为88.315%。ANOVA中的主效应是基于一个因素相对于其它因素的均值（例如，不论是快乐还是悲伤的，类人声音的主效应与类机器人声音的主效应的对比）。这种交互作用的效应可以使用1，-1，-1，1来编码，这与基于每个单元格的分数的效应是否为非相加性的测试相似，如果每个单元格的效应可以被单纯地相加的预期偏差可以被拒绝，那么非相加效应的零假设可以被拒绝。关键的点在于不是每种情况下的样本量而是在于基于所有因素的总的样本量来决定主效应和交互作用的统计检验力。

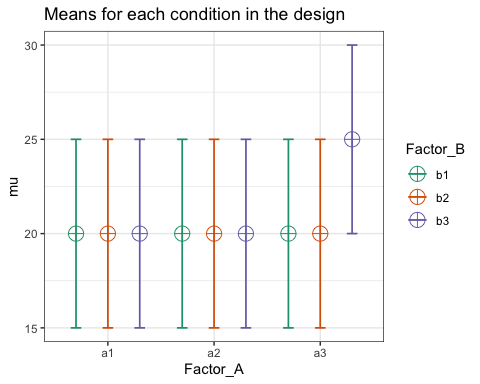
我们也可以检验表明使用高兴或者伤心的类机器人声音之间不存在差异的结果模式的统计检验力。在这种情况下，我们期望一个“序数”的交互，即类人声的均值永远不会比类机器人声的均值低，因此也就不会有交叉效应。现将均值的模式改为1，0，0，0，其中只有一个均值与其它均值不同。与两组对比相比，这种设计需要更大的样本量来拥有同样的检测到交互作用的统计检验力。原因在于效应量只有原来的一半，Cohen’s f = 0.125（在交叉交互中为0.25）。通过在模拟中稳定增加样本量，我们可以看到为了达到像两组对比的同样的统计检验力需要635名被试，几乎是两组对比的样本量（160）的四倍。

在只有一个单元格的均值与其它均值不同的2X2的序数交互的统计检验力与一个均值为其它均值的两倍大（例如2，0，0，0）的统计检验力相同。相似地，如果我们检验一个只有一个单元格与其它单元格的均值不同的2X2X2的交互，当均值为4，0，0，0，0，0，0，0时，Cohen’s f为0.25。总的来说，与无序（交叉）交互相比，在序数交互中“中等”效应量会转化为一个更为极端的均值模式,与2X2交互相比，2X2X2的交互亦是如此。因此与直接设定一个效应量相比，基于期望的均值的模式和相应计算的Cohen’s f来执行一个统计检验力分析更为直观。基于交互的种类和因素的个数，同样的效应量可代表非常不同的均值的模式。

##### 2.3.5.7.1 3X3被试间实验设计的ANOVA

如果我们将实验设计中的每一个因素拓展一个水平，我们将得到一个3X3的实验设计（2个因素，每个因素3个水平）。

design <- "3b\*3b"  
n <- 20  
mu <- c(20, 20, 20, 20, 20, 20, 20, 20, 25) #All means are equal - so there is no real difference.  
# Enter means in the order that matches the labels below.  
sd <- 5  
labelnames <- c("Factor\_A", "a1", "a2", "a3", "Factor\_B", "b1", "b2", "b3") #  
# the label names should be in the order of the means specified above.  
  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 labelnames = labelnames)



ANOVA\_exact(design\_result, alpha\_level = 0.05)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## Factor\_A 44.86 0.0253 0.1612 4.4444  
## Factor\_B 44.86 0.0253 0.1612 4.4444  
## Factor\_A:Factor\_B 64.34 0.0494 0.2280 8.8889  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a1\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a1\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a2\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a2\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a2\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a1\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a2\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a2\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a2\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a2\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a2\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a2\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a1\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a2\_Factor\_B\_b1\_Factor\_A\_a2\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b1\_Factor\_A\_a2\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a2\_Factor\_B\_b2\_Factor\_A\_a2\_Factor\_B\_b3 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a2\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b1 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a2\_Factor\_B\_b3\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a3\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b2 5.0 0  
## p\_Factor\_A\_a3\_Factor\_B\_b1\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1  
## p\_Factor\_A\_a3\_Factor\_B\_b2\_Factor\_A\_a3\_Factor\_B\_b3 86.9 1

我们可以对照解析解决方案来检查。

power\_res <- power\_twoway\_between(design\_result) #using default alpha level of .05  
  
power\_res$power\_A

## [1] 0.4486306

power\_res$power\_B

## [1] 0.4486306

power\_res$power\_AB

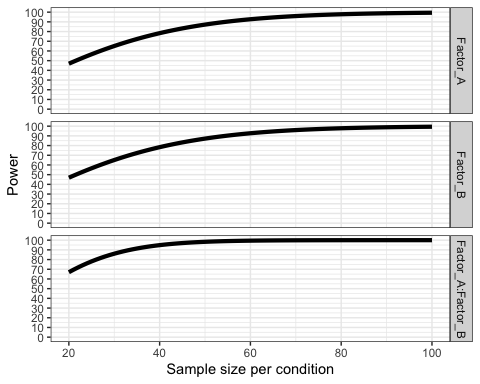
## [1] 0.6434127

我们可以看到当每个单元格有20名被试时，统计检验力很低。我们来看当每个实验条件下的样本量增加时，统计检验力如何增加。

plot\_power(design\_result, min\_n = 20, max\_n = 100)

## Warning: Use of `plot\_data$n` is discouraged. Use `n` instead.

## Warning: Use of `plot\_data$value` is discouraged. Use `value` instead.



#### 2.3.5.8 被试内实验设计的2X2 ANOVA

Potvin和Schutz（2000）模拟一个大范围的重复测量设计。给出的例子为3X3的实验设计，相关矩阵为：

相关矩阵

相关矩阵

(来自 <https://github.com/Lakens/ANOVApower/raw/master/vignettes/screenshots/PS2000.gif>)

方差设定为1（所以在模拟中所有的协方差矩阵是相同的）。在这个特定的例子中，白色区域是与A的主效应的相关相关联的（这些单元格与B拥有同样的水平，但是对A有不同的水平）。灰色的区域与B的主效应相关（这些单元格与A有相同的水平，但是与B的水平不同）。最终，黑色的区域与AXB的交互相关（与A和B的水平数均不同，对角线上的单元格拥有与A和B相同的水平。

Potvin和Schulz（2000）检验了2X2的被试内ANOVA设计的统计检验力并得出误差方差的近似值。对于一个拥有2个被试内因素（A和B）的设计来说是：

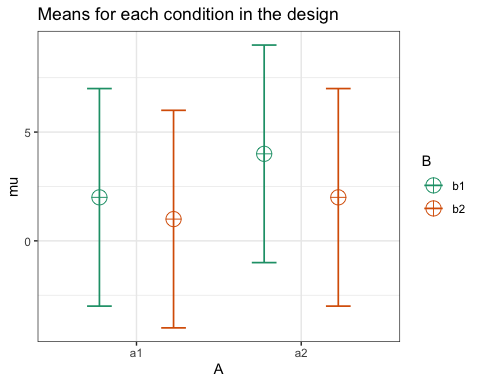
对于A的主效应：

对于B的主效应：

对于A和B的交互：

我们首先来模拟一个2X2 ANOVA被试内实验设计。

mu = c(2,1,4,2)   
n <- 20  
sd <- 5  
r <- c(  
 0.8, 0.5, 0.4,  
 0.4, 0.5,  
 0.8  
 )  
  
design = "2w\*2w"  
labelnames = c("A", "a1", "a2", "B", "b1", "b2")  
design\_result <- ANOVA\_design(design = design,  
 n = n,   
 mu = mu,   
 sd = sd,   
 r = r,   
 labelnames = labelnames)



nsims <- 1000 # 注：此处根据实际需要修改，例如上文提到的先设定成100，如果正确的话再修改为1000  
power\_result\_vig\_4 <- ANOVA\_power(design\_result, nsims = nsims)

## Power and Effect sizes for ANOVA tests  
## power effect\_size  
## anova\_A 28.4 0.1324  
## anova\_B 60.9 0.2478  
## anova\_A:B 22.9 0.1211  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_A\_a1\_B\_b1\_A\_a1\_B\_b2 25.8 -0.32199  
## p\_A\_a1\_B\_b1\_A\_a2\_B\_b1 40.5 0.41923  
## p\_A\_a1\_B\_b1\_A\_a2\_B\_b2 4.3 0.01532  
## p\_A\_a1\_B\_b2\_A\_a2\_B\_b1 64.2 0.56690  
## p\_A\_a1\_B\_b2\_A\_a2\_B\_b2 14.2 0.21776  
## p\_A\_a2\_B\_b1\_A\_a2\_B\_b2 74.8 -0.64087  
##   
## Within-Subject Factors Included: Check MANOVA Results

power\_result\_vig\_4$sim\_data <- NULL  
# power\_result\_vig\_4$plot1 <- NULL  
power\_result\_vig\_4$plot2 <- NULL  
saveRDS(power\_result\_vig\_4, file = "power\_result\_vig\_4.rds")  
power\_result\_vig\_4 <- readRDS(file = "power\_result\_vig\_4.rds")

我们也可以使用exact模拟。

ANOVA\_exact(design\_result = design\_result)

## Power and Effect sizes for ANOVA tests  
## power partial\_eta\_squared cohen\_f non\_centrality  
## A 26.92 0.0952 0.3244 2  
## B 64.23 0.2400 0.5620 6  
## A:B 26.92 0.0952 0.3244 2  
##   
## Power and Effect sizes for contrasts  
## power effect\_size  
## p\_A\_a1\_B\_b1\_A\_a1\_B\_b2 26.92 -0.3162  
## p\_A\_a1\_B\_b1\_A\_a2\_B\_b1 39.70 0.4000  
## p\_A\_a1\_B\_b1\_A\_a2\_B\_b2 5.00 0.0000  
## p\_A\_a1\_B\_b2\_A\_a2\_B\_b1 64.23 0.5477  
## p\_A\_a1\_B\_b2\_A\_a2\_B\_b2 13.60 0.2000  
## p\_A\_a2\_B\_b1\_A\_a2\_B\_b2 76.52 -0.6325

我们可以看到26.9、64.2和26.9都与exact模拟的结果一致。

## 2.4 PANGEA

## 2.5 LMM Power

## 2.6 power analysis in fMRI

## 2.7 Bayesian power analysis