## ads

https://www.cnblogs.com/chen-heybro/p/18212750
https://www.zhihu.com/people/djj-3-36?utm_id=0

## 改pdf引擎

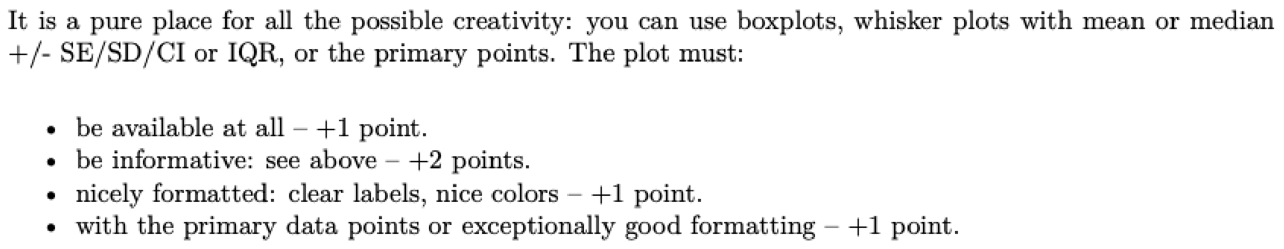
output:
pdf_document:
latex_engine: xelatex

## knit

```
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
knitr::opts_chunk$set(tidy = TRUE, tidy.opts = list(width.cutoff = 60))
library(tidyverse)
library(knitr)
```

## ggplot

老师要求



It is a pure place for all the possible creativity: you can use boxplots, whisker plots with mean or median +/- SE/SD/CI or IQR, or the primary points. The plot must:

- be available at all – +1 point.
- be informative: see above – +2 points.
- nicely formatted: clear labels, nice colors – +1 point.
- with the primary data points or exceptionally good formatting – +1 point.

标题肯定要的，xy轴也必须重命名成清楚的，带上单位

箱线图案例

```
# Load the ggplot2 package
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)

# Create sample data
set.seed(123) # Set a random seed for reproducibility
data <- data.frame(
  Category = rep(c("A", "B", "C"), each = 100),
  Value = c(rnorm(100, mean = 10, sd = 2),
            rnorm(100, mean = 12, sd = 3),
            rnorm(100, mean = 8, sd = 1))
)

# Create a boxplot using ggplot2
ggplot(data, aes(x = Category, y = Value)) +
  geom_boxplot() + # Add boxplot layer
  geom_point(position = position_jitter(width = 0.2), color = "blue") + # Add jittered points layer to show raw
data
  labs(title = "Example Boxplot with ggplot2", # Set the chart title
       x = "Categories", # Set the x-axis label
       y = "Values") + # Set the y-axis label
  theme_minimal() + # Use a minimalist theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1), # Rotate x-axis labels for better readability
        plot.title = element_text(hjust = 0.5)) + # Center the title
  scale_y_continuous(breaks = seq(0, 15, by = 2)) + # Set y-axis tick marks
  scale_color_brewer(palette = "Pastel1") + # Use a color brewer palette for points
  theme(plot.background = element_rect(fill = "gray95"), # Set the plot background color
        axis.line = element_line(color = "black"), # Set axis line color
        panel.grid.major = element_line(color = "gray80"), # Set major grid line color
        panel.grid.minor = element_line(color = "gray90")) # Set minor grid line color
```

### 一。 数据清洗

```
# 加载数据集
data(airquality)

# 查看数据集
str(airquality)
head(airquality)
```

### 查看NA

```
colSums(is.na(airquality))
```

### 去除NA

```
airquality_no_na <- na.omit(airquality)
```

### 再次check NA

```
colSums(is.na(airquality_no_na))
```

### 查看duplicates

```
duplicated_rows <- duplicated(airquality_no_na)
sum(duplicated_rows)
```

### 去除duplicates

```
airquality_no_dup <- unique(airquality_no_na)
```

### 查看typo（在这个例子中，数据集没有文本列，所以跳过这一步）用另外一个数据集举例子

```
# 加载数据集
data(mtcars)
mtcars
```

```
# 假设在"carb"列中，"4"被错误地输入为"4x"
mtcars$carb[c(1, 10)] <- "4x"

# 查看"carb"列的唯一值
unique(mtcars$carb)
```

```
# 查找包含"x"的行（假设"x"是一个typo标志）
grep("x", mtcars$carb)
```

```
# 将"4x"替换为"4"
mtcars$carb <- gsub("4x", "4", mtcars$carb)
```

```
# 再次查看"carb"列的唯一值，确认typo已被去除
unique(mtcars$carb)
```

### 检查数据类型和数据位置是否合理

有的时候，数据类型并不合适，比如分类变量我们通常采用factor的数据格式；而且有的时候，因变量在自变量的列前面，这个也不合适

```
teeth <- teeth %>%
mutate(dose = factor(dose, levels = c(0.5, 1, 2), ordered = T),
supp = as.factor(supp)) %>% relocate(supp, dose)
str(teeth)
```

### 进行长宽数据转换（如果有必要）

#### 长数据转换

```
library(tidyr)
airquality_long <- gather(airquality_no_dup, key = "variable", value = "value", -Month, -Day)
```

- airquality_no_dup：这是输入数据集，它应该是一个数据框（data frame）或类似的结构，其中包含多个列，这些列将被转换成长格式。
- key：这是新数据集中用于存储原始列名的列名。在这个例子中，它被设置为 "variable"。
- value：这是新数据集中用于存储原始列值的列名。在这个例子中，它被设置为 "value""。

  -Month 和 -Day：这些是选择器，表示在转换过程中，Month 和 Day 这两列将保持不变，不被转换为长格式。- 前缀表示排除这些列。

#### 宽数据转换

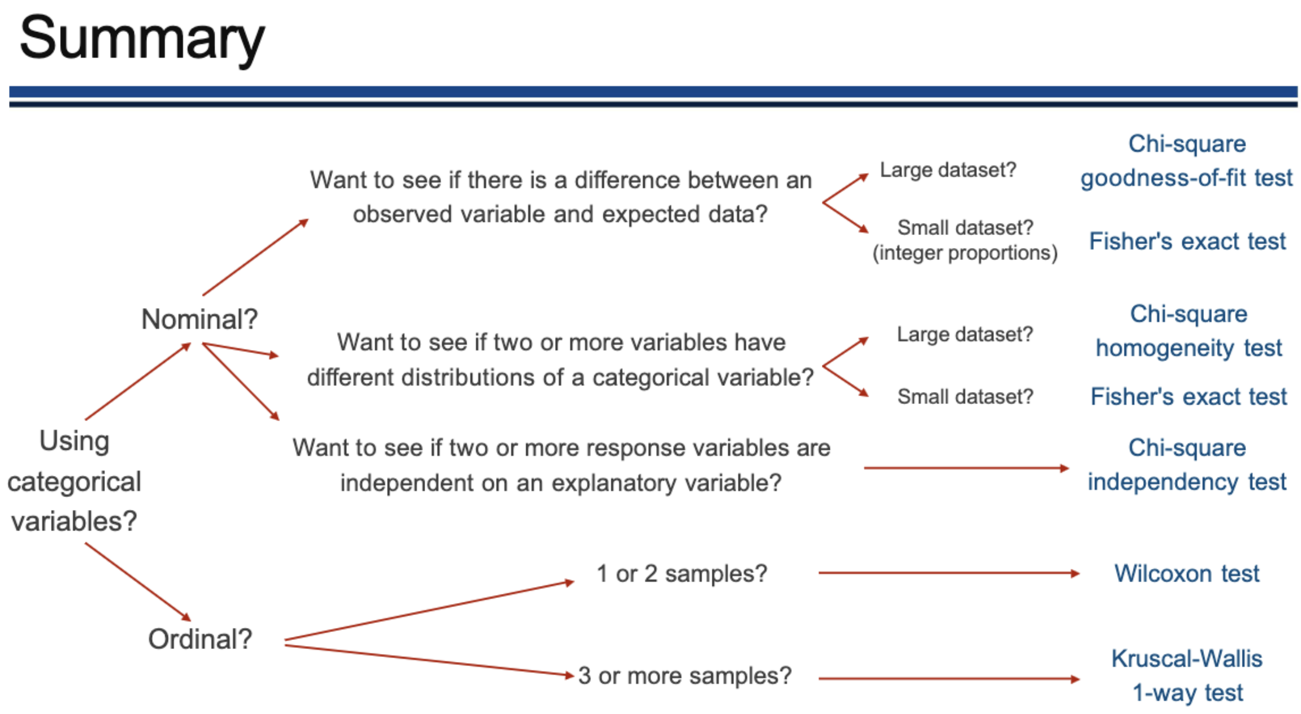
```
# 假设df是长格式的数据框
wide_data <- df %>%
  pivot_wider(names_from = column_to_spread, values_from = column_to_copy)
```

```
data_wide <- airquality_long %>%
  pivot_wider(names_from = variable,values_from = value)
```

### 再次查看数据集确认满足要求，清洗完成

```
str(airquality_long)
head(airquality_long)
head(data_wide)
```

## 二。 chisq



Summary

Using categorical variables?

Nominal?

Want to see if there is a difference between an observed variable and expected data?

Large dataset? → Chi-square goodness-of-fit test

Small dataset? (integer proportions) → Fisher's exact test

Want to see if two or more variables have different distributions of a categorical variable?

Large dataset? → Chi-square homogeneity test

Small dataset? → Fisher's exact test

Want to see if two or more response variables are independent on an explanatory variable? → Chi-square independency test

Ordinal?

1 or 2 samples? → Wilcoxon test

3 or more samples? → Kruscal-Wallis 1-way test

### 卡方分布

，是为了解决衡量categorical data,尤其是nominal variables而创造出来的方法

> 如果你的研究目的是比较实际观测值与某个理论分布或期望值，使用适合性检验。
> 如果你在比较两个或多个独立样本的分类变量分布是否有差异，使用同质性检验。
> 如果你在分析两个分类变量之间是否存在关联，使用独立性检验。

这三个检验类型——适合性检验、同质性检验和独立性检验——都是统计分析中用于检验不同假设的方法，但它们检验的目的和应用场景各不相同。下面我会详细解释它们的区别：

1. **适合性检验（Goodness-of-fit test）**：
   - **目的**：检验实际观测值是否与某个理论分布或期望值相符合。
   - **应用场景**：当你有一个理论模型或分布，想要检验实际数据是否遵循这个模型或分布时使用。例如，检验抛硬币实验的结果是否符合理论上的50%正面和50%反面的概率分布。
   - **常用方法**：卡方适合性检验（Chi-squared goodness-of-fit test）是常用的适合性检验方法。
2. **同质性检验（Homogeneity test）**：
   - **目的**：检验两个或多个独立样本的分类变量分布是否存在差异。
   - **应用场景**：当你想要比较不同样本或不同群体在某个分类变量上的分布是否一致时使用。例如，比较不同地区人群的吸烟率是否有显著差异。
   - **常用方法**：卡方同质性检验（Chi-squared test for homogeneity）是常用的同质性检验方法。
3. **独立性检验（Independence test）**：
   - **目的**：检验两个分类变量之间是否存在关联或依赖关系。
   - **应用场景**：当你想要分析两个分类变量之间是否存在某种联系时使用。例如，检验性别（男性/女性）和职业选择（工程师/非工程师）之间是否存在关联。
   - **常用方法**：卡方独立性检验（Chi-squared test for independence）是常用的独立性检验方法。

**区别总结**：

- **适合性检验**关注的是实际观测值与理论模型或期望值之间的拟合程度。
- **同质性检验**关注的是不同样本或群体在某个分类变量上的分布是否一致。
- **独立性检验**关注的是两个分类变量之间是否存在关联。

在实际应用中，选择哪种检验方法取决于你的研究问题和数据类型。例如，如果你的数据是分类变量，并且你想要检验它们之间的关联，那么独立性检验是合适的；如果你想要检验实际观测值是否符合某个理论分布，那么适合性检验是正确的选择；如果你想要比较不同样本在某个分类变量上的分布，那么同质性检验是合适的。

### Assumptions 使用的前提 (large dataset)

- The variables must be categorical.
- Observations must be independent
  - Can assume from the task
- Cells in the contingency table are mutually exclusive.
  - Fits
- The expected value of cells should be 5 or greater in at least 80% of cells.

### test of goodness-of-fit

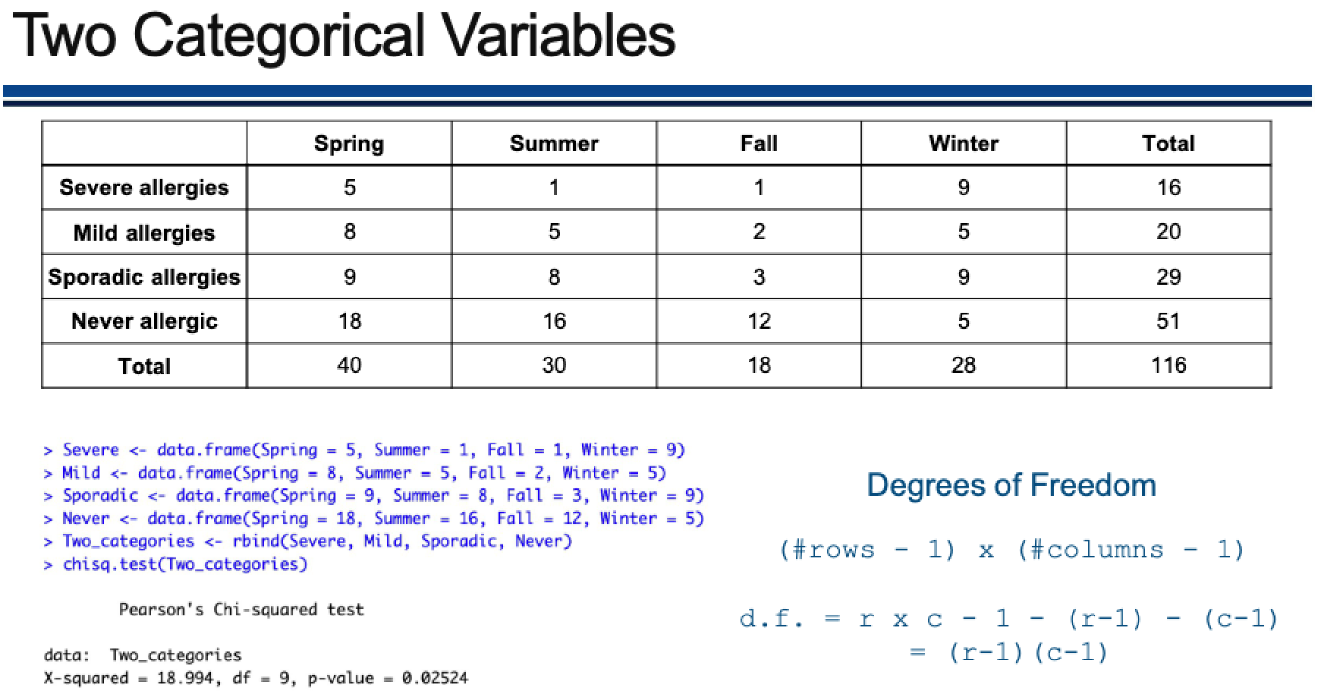
例子： Is there a difference between the season preferences?
H0: There is no difference between the observed and expected season preferences
H1: There is a difference between the observed and expected season preferences

```
chisq.test(Poll_seasons, correct = FALSE, p = rep(1/4, 4))
```

### test for homogeneity(需要判断两个分类变量是否是一样的或者不一样的)

Question: Is there a difference between the distribution of allergic reactions in the different seasons?
H0: The distribution of allergic reactions is the same for the people who preferred different seasons
H1: The distribution of allergic reactions is not the same for the people who preferred different seasons

问：不同季节过敏反应的分布有区别吗？H0：喜欢不同季节的人过敏反应分布相同 H1：喜欢不同季节的人过敏反应分布不一样



## Two Categorical Variables

| | Spring | Summer | Fall | Winter | Total |
|---|---|---|---|---|---|
| **Severe allergies** | 5 | 1 | 1 | 9 | 16 |
| **Mild allergies** | 8 | 5 | 2 | 5 | 20 |
| **Sporadic allergies** | 9 | 8 | 3 | 9 | 29 |
| **Never allergic** | 18 | 16 | 12 | 5 | 51 |
| **Total** | 40 | 30 | 18 | 28 | 116 |

```
> Severe <- data.frame(Spring = 5, Summer = 1, Fall = 1, Winter = 9)
> Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5)
> Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9)
> Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5)
> Two_categories <- rbind(Severe, Mild, Sporadic, Never)
> chisq.test(Two_categories)

	Pearson's Chi-squared test

data:  Two_categories
X-squared = 18.994, df = 9, p-value = 0.02524
```

Degrees of Freedom

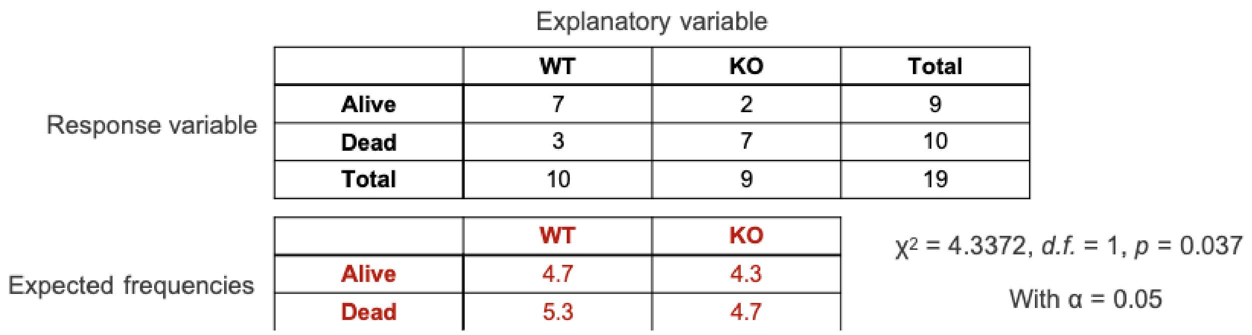
```
(#rows - 1) x (#columns - 1)

d.f. = r x c - 1 - (r-1) - (c-1)
     = (r-1)(c-1)
```

### test of independency

Question: We need to analyse the survival data of a geneX knockout mice at 1 year. Does geneX affect lifespan of mice?
H0: The survival of mice is independent on geneX
H1: The survival of mice is dependent on geneX

The lifespan is dependent on geneX



Explanatory variable

Response variable

| | WT | KO | Total |
|---|---|---|---|
| **Alive** | 7 | 2 | 9 |
| **Dead** | 3 | 7 | 10 |
| **Total** | 10 | 9 | 19 |

Expected frequencies

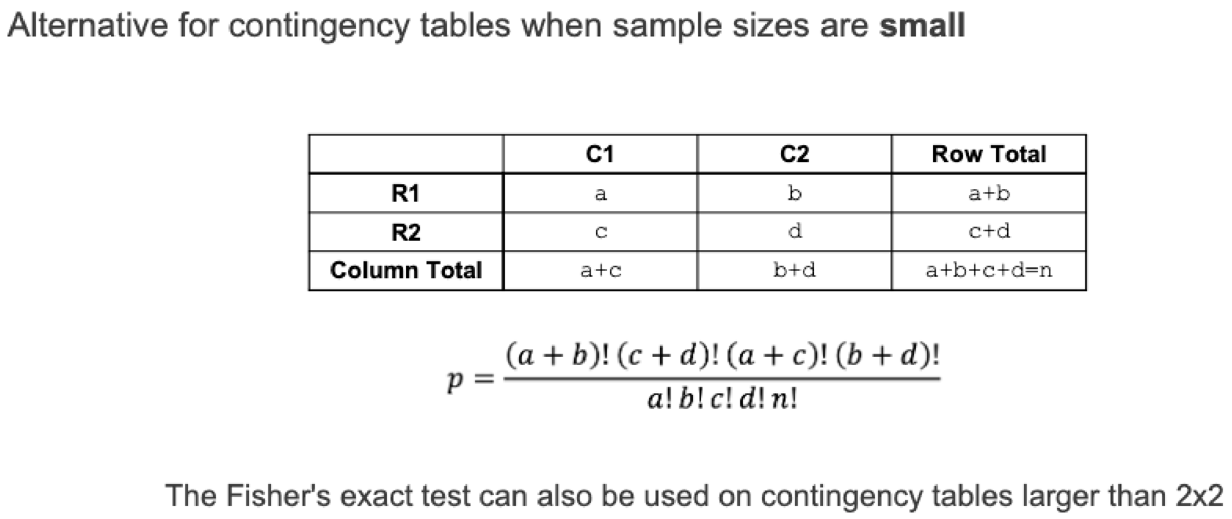
| | WT | KO |
|---|---|---|
| **Alive** | 4.7 | 4.3 |
| **Dead** | 5.3 | 4.7 |

$\chi^2 = 4.3372$, *d.f.* = 1, $p = 0.037$

With α = 0.05

## Fisher's exact test (在卡方检验不适合的时候就可以用)

当样本量很小的时候就可以用这个



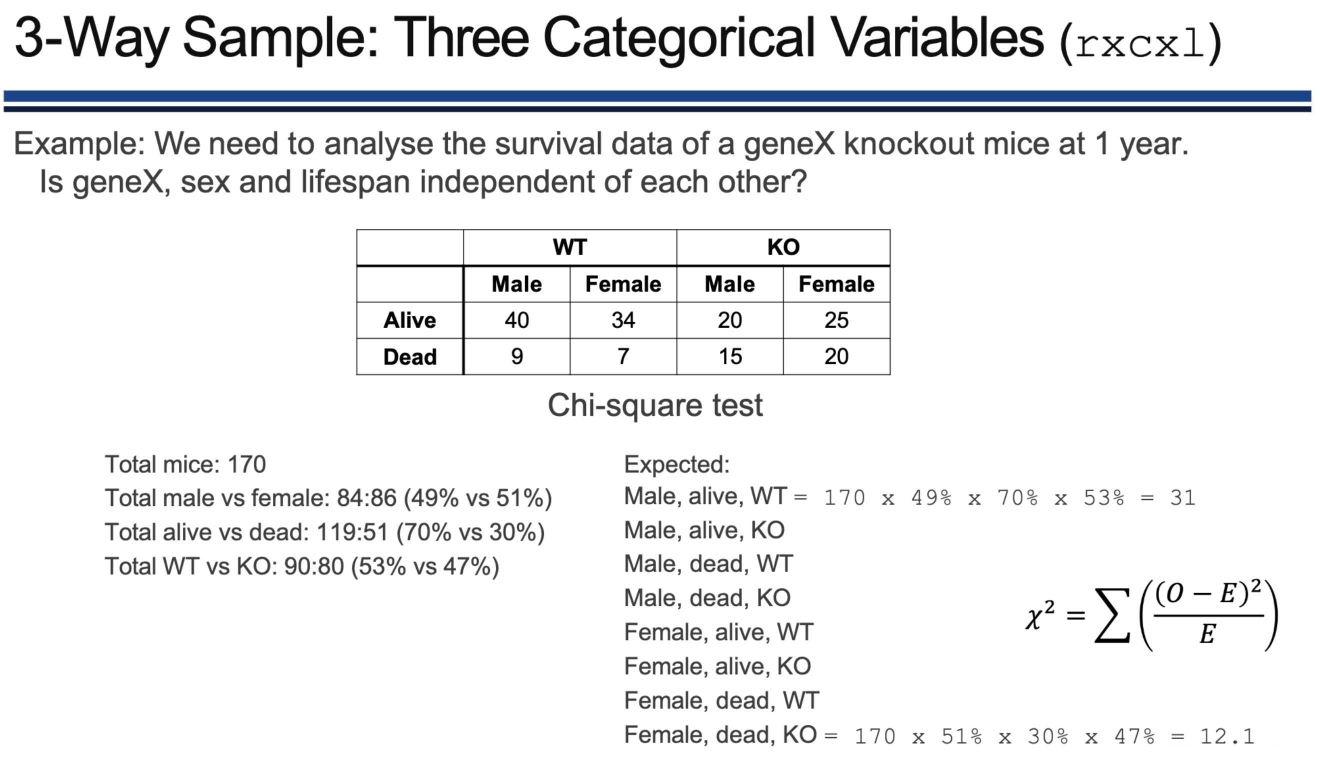
Alternative for contingency tables when sample sizes are **small**

| | C1 | C2 | Row Total |
|---|---|---|---|
| **R1** | a | b | a+b |
| **R2** | c | d | c+d |
| **Column Total** | a+c | b+d | a+b+c+d=n |

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

The Fisher's exact test can also be used on contingency tables larger than 2x2

## 3-way ANNOVA



# 3-Way Sample: Three Categorical Variables (rxcxl)

Example: We need to analyse the survival data of a geneX knockout mice at 1 year. Is geneX, sex and lifespan independent of each other?

| | WT | | KO | |
|---|---|---|---|---|
| | **Male** | **Female** | **Male** | **Female** |
| **Alive** | 40 | 34 | 20 | 25 |
| **Dead** | 9 | 7 | 15 | 20 |

Chi-square test

Total mice: 170
Total male vs female: 84:86 (49% vs 51%)
Total alive vs dead: 119:51 (70% vs 30%)
Total WT vs KO: 90:80 (53% vs 47%)

Expected:
Male, alive, WT = 170 x 49% x 70% x 53% = 31
Male, alive, KO
Male, dead, WT
Male, dead, KO
Female, alive, WT
Female, alive, KO
Female, dead, WT
Female, dead, KO = 170 x 51% x 30% x 47% = 12.1

$$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$$

## Kruskal-Wallis H test

Kruskal-Wallis H test（也称为一元方差分析的非参数替代方法）用于比较三个或更多个独立样本的中位数是否存在显著差异。当数据不满足ANOVA的正态分布假设时，Kruskal-Wallis检验是一个有用的非参数选择。

以下是使用R语言执行Kruskal-Wallis检验的步骤：

1. **确定假设**:
   - 零假设 (H0): 所有组的中位数相同。
   - 备择假设 (H1): 至少有一个组的中位数与其他组不同。
2. **准备数据**:
   - 确保数据是分成三个或更多组的独立样本。
3. **使用R执行Kruskal-Wallis检验**:
   - 在R中，使用`kruskal.test()`函数来进行Kruskal-Wallis检验。

假设你有一个向量`group`表示样本所属的组，以及一个向量`value`表示对应的观测值，以下是如何使用`kruskal.test()`函数的示例：

```r
# 假设有以下数据
group <- factor(c("A", "A", "B", "B", "C", "C"))
value <- c(10, 12, 15, 18, 20, 22)

# 执行Kruskal-Wallis检验
kruskal_result <- kruskal.test(value ~ group)

# 打印结果
print(kruskal_result)
```

在这个示例中，`group`是一个因子类型变量，用于指示每个观测值所属的组，而`value`是对应的数值型数据。`kruskal.test(value ~ group)`函数将执行检验，并返回一个包含检验统计量、自由度、P值等信息的对象。

**结果解释**:

- 如果P值小于常用的显著性水平（例如0.05），则拒绝零假设，认为至少有两个组之间存在显著差异。
- 如果P值大于显著性水平，则不能拒绝零假设，即没有足够证据表明组间中位数存在差异。

Kruskal-Wallis检验的结果只能告诉你至少有两个组之间存在差异，但它不会告诉你具体哪些组之间存在差异。如果检验结果显著，通常需要进一步的事后比较（如Mann-Whitney U检验）来确定具体哪些组之间存在差异。

Mann-Whitney U检验（也称为Wilcoxon秩和检验）是一种非参数检验，用于比较两个独立样本的中位数是否存在显著差异。当数据不满足正态分布假设或样本量较小时，此检验是一个合适的选择。

以下是使用R语言执行Mann-Whitney U检验的步骤：

1. **确定假设**:
   - 零假设 (H0): 两个独立样本的中位数相同。
   - 备择假设 (H1): 两个独立样本的中位数不同。
2. **准备数据**:
   - 确保你有两组独立的数据。
3. **使用R执行Mann-Whitney U检验**:
   - 在R中，使用`wilcox.test()`函数来进行Mann-Whitney U检验。

假设你有两个向量`sample1`和`sample2`，分别代表两组独立样本的数据，以下是如何使用`wilcox.test()`函数的示例：

```r
# 假设有以下两组独立样本数据
sample1 <- c(10, 12, 15, 18)
sample2 <- c(8, 14, 11, 17)

# 执行Mann-Whitney U检验
u_result <- wilcox.test(sample1, sample2)

# 打印结果
print(u_result)
```

在这个示例中，`sample1`和`sample2`是两组独立的数值型数据。`wilcox.test(sample1, sample2)`函数将执行Mann-Whitney U检验，并返回一个包含检验统计量、P值等信息的对象。

**结果解释**:

- **检验统计量**: `wilcox.test()`返回的检验统计量是U值，它是根据秩次计算的。U值越小，表示两个样本之间的差异越大。
- **P值**: 如果P值小于常用的显著性水平（例如0.05），则拒绝零假设，认为两个独立样本的中位数存在显著差异。
- 如果P值大于显著性水平，则不能拒绝零假设，即没有足够证据表明两个样本的中位数存在差异。

请注意，`wilcox.test()`函数默认执行的是Wilcoxon秩和检验，它是Mann-Whitney U检验的一个变体，两者在解释上是相同的。此外，如果你的数据是配对样本，应该使用Wilcoxon符号秩检验（也称为Wilcoxon符号检验），这是通过在`wilcox.test()`函数中设置`paired=TRUE`参数来实现的。

```
# 假设有以下两组配对样本数据
paired_sample1 <- c(10, 12, 15, 18)
paired_sample2 <- c(8, 14, 11, 17)

# 执行Wilcoxon符号秩检验（配对样本）
signed_result <- wilcox.test(paired_sample1, paired_sample2, paired=TRUE)

# 打印结果
print(signed_result)
```

在配对样本的情况下，检验将考虑观测值之间的差异，而不是像独立样本那样分别对两组数据进行秩次排序。

## CrossTable函数

P.S. 如何快速提取出列联表（甚至直接计算出chi-sq，但是直接计算有点不保险，最好还是提取出来再说）

```
library(gmodels)
head(mydata)
cross_table <- CrossTable(mydata$x, mydata$y) #应该要设置chisq=flase
chisq_result<- chisq.test(cross_table$t)
```

## 原始画表格

```
new_genotype <- data.frame(
  WT = c(nrow(subset(genotype, sex == "female" & genotype == "WT")),
         nrow(subset(genotype, sex == "male" & genotype == "WT"))),
  het = c(nrow(subset(genotype, sex == "female" & genotype == "het")),
          nrow(subset(genotype, sex == "male" & genotype == "het"))),
  mut = c(nrow(subset(genotype, sex == "female" & genotype == "mut")),
          nrow(subset(genotype, sex == "male" & genotype == "mut"))),
  row.names = c("female", "male")
)
...
```

## 三。提建议

- 缺乏生物学意义
- 需要后续研究
- 机制不明确，等等

increase sample size:但是必须给出统计学上的理由

- Power estimation

1. Chi-sq
   X是我要计算的数据集，可以是向量也可以是matrix

```
model <- chisq.test(x)
chisq_test <- model$statistic
N <- sum(X)
R <- row_numbers
C <- col_numbers
V <- sqrt(chisq_stat / (N * min(R-1, C-1)))  # 计算Cramer's V


current <-pwr.chisq.test(w=V,df=(R-1)*(C-1),sig.level = 0.05,N = N)
current
expected <- pwr.chisq.test(w=V, df=(R-1)*(C-1), sig.level=0.05, power=0.8)
expected
```

情况一：满足了power
Based on the power analysis for the chi-square test, at a significance level of 0.05, the current power is current$power. This power value has reached the commonly used standard of 0.8 in statistical analysis, indicating that with the current sample size, we have sufficient ability to detect an actual effect if it exists.

Since the power has reached the desired level, there is no need to increase the sample size. We can proceed with the actual chi-square test and subsequent statistical analyses based on the current sample size.

情况二：不满足power
Based on the power analysis for the chi-square test, at a significance level of 0.05, the current power is current$power, which is lower than the commonly used standard of 0.8 in statistical analysis. This indicates that with the current sample size, our ability to detect an actual effect, if it exists, is relatively low.

To increase the power to the desired level of 0.8, it is recommended to increase the sample size. Based on the expected power of 0.8, the minimum required sample size is expected$n.Therefore,Irecommendincreasingthesamplesizetoatleastexpected$n to ensure sufficient statistical power to detect an actual effect.

2. Anova （只能做one-way的，两变量不考虑这个）

```
library(pwr)

# 计算eta squared效应量
x <- c(10, 12, 14, 16, 18, 20, 22, 24, 8, 10, 12, 14, 16, 18, 20, 22, 6, 8, 10, 12, 14, 16, 18, 20)  # 输入所有数据
group <- c(rep("A", 8), rep("B", 8), rep("C", 8))  #


anova_model <- aov(x ~ group)
anova_summary <- anova(anova_model)
ss_total <- sum(anova_summary$"Sum Sq")
ss_group <- anova_summary$"Sum Sq"[1]
ss_residual <- anova_summary$"Sum Sq"[2]
eta_squared <- ss_group / ss_total


# 计算所需样本量
k <- length(unique(group))  # 组数
current <- pwr.anova.test(f = eta_squared / (1 - eta_squared),
                          k = k, n= sum(x)/k,
                          sig.level = 0.05)
expected <- pwr.anova.test(f = eta_squared / (1 - eta_squared),
                           sig.level = 0.05, k = k,
                           power = 0.8)
current
expected
```

3. t test

```
library(pwr)
```

```
# 计算Cohen's d效应量
x <- c(12,212,3232,323,31,31,43)  # 输入第一组数据
y <- c(12,33,22,4,42,111,3,222)  # 输入第二组数据
n1 <- length(x)
n2 <- length(y)
d <- (mean(x) - mean(y)) / sqrt(((n1-1)*var(x) + (n2-1)*var(y)) / (n1 + n2 - 2))

# 计算所需样本量
current <- pwr.t.test(n = n1, d = d,
                      sig.level = 0.05,
                      type = "two.sample")
expected <- pwr.t.test(d = d,
                       sig.level = 0.05,
                       power = 0.8,
                       type = "two.sample")

cat("Current power is:", current$power, "\n")
cat("With an expected power of 0.8, the minimum sample size required is:", expected$n, "\n")
```

Based on your request, here is a possible report for the power analysis of a two-sample t-test, discussing two scenarios:

Scenario 1: Power meets the desired level

Report:
The power analysis for the two-sample t-test was conducted to assess the ability to detect a significant difference in means between two independent groups. The effect size, Cohen's d, was calculated based on the provided data.

At a significance level of 0.05, the current power is current$power. This power value meets the commonly accepted standard of 0.8 in statistical analysis, indicating that with the current sample size, we have sufficient ability to detect an actual effect if it exists.

Recommendation:
Since the power has reached the desired level of 0.8, there is no need to increase the sample size. We can proceed with the actual two-sample t-test and subsequent statistical analyses based on the current sample size of n1 = [current sample size for group 1] and n2 = [current sample size for group 2].

Scenario 2: Power does not meet the desired level

Report:
The power analysis for the two-sample t-test was conducted to assess the ability to detect a significant difference in means between two independent groups. The effect size, Cohen's d, was calculated based on the provided data.

At a significance level of 0.05, the current power is current$power, which is lower than the commonly accepted standard of 0.8 in statistical analysis. This indicates that with the current sample size, our ability to detect an actual effect, if it exists, is relatively low.

Recommendation:
To increase the power to the desired level of 0.8, it is recommended to increase the sample size. Based on the expected power of 0.8, the minimum required sample size for each group is expected$n. Therefore, I recommend increasing the sample size to at least n1 = expected$n and n2 = expected$n to ensure sufficient statistical power to detect an actual effect.

It is important to note that the power analysis results depend on the estimation of the effect size, Cohen's d. If the actual effect size deviates from the estimated value, further adjustment of the sample size may be necessary.

Please note that in this report, you should replace "current$power'' and '' expected$n" with the actual values obtained from your power analysis. Additionally, you should replace "[current sample size for group 1]" and "[current sample size for group 2]" with the respective sample sizes for each group in the current data.

- 旧版本，不删除了怕还是有点用

```
library(pwr)

# t检验
power_t_test <- function(m1, m2, sd1, sd2, n1, n2, alpha = 0.05) {
  sd_pooled <- sqrt(((n1 - 1) * sd1^2 + (n2 - 1) * sd2^2) / (n1 + n2 - 2))
  d <- (m1 - m2) / sd_pooled
  power <- pwr.t.test(n = n1 + n2, d = d, sig.level = alpha, type = "two.sample", alternative = "two.sided")$power
  return(list(d = d, power = power))
}

# Wilcoxon秩和检验
power_wilcox_test <- function(z, n, alpha = 0.05) {
  r <- z / sqrt(n)
  power <- pwr.r.test(n = n, r = r, sig.level = alpha)$power
  return(list(r = r, power = power))
}

# 单因素方差分析
power_anova_test <- function(ss_between, ss_total, n, k, alpha = 0.05) {
  eta_sq <- ss_between / ss_total
  f <- sqrt(eta_sq / (1 - eta_sq))
  power <- pwr.anova.test(k = k, n = n, f = f, sig.level = alpha)$power
  return(list(eta_sq = eta_sq, f = f, power = power))
}

# Fisher's精确检验和卡方检验
power_fisher_chisq_test <- function(chi_sq, n, k, alpha = 0.05) {
  if (k == 2) {
    effect_size <- sqrt(chi_sq / n)
  } else {
    effect_size <- sqrt(chi_sq / (n * (k - 1)))
  }
  power <- pwr.chisq.test(w = effect_size, N = n, df = (k - 1) * (k - 1), sig.level = alpha)$power
  return(list(effect_size = effect_size, power = power))
}

# 示例用法
t_test_result <- power_t_test(m1 = 10, m2 = 12, sd1 = 2, sd2 = 2.5, n1 = 30, n2 = 30)
print(t_test_result)

wilcox_test_result <- power_wilcox_test(z = 2.5, n = 50)
print(wilcox_test_result)

anova_test_result <- power_anova_test(ss_between = 100, ss_total = 500, n = 10, k = 3)
print(anova_test_result)

fisher_chisq_test_result <- power_fisher_chisq_test(chi_sq = 8, n = 100, k = 2)
print(fisher_chisq_test_result)
```

好的,我来解释一下这些函数中使用的统计量:

1. `power_t_test`函数:
    - `d`: 这是Cohen's d效应量,用于衡量两组平均值之间的差异大小。它是两组平均值的差值除以pooled标准差。
    - `power`: 这是在给定效应量和样本量的情况下,t检验能够检测到真实效应的概率。
2. `power_wilcox_test`函数:
    - `r`: 这是Wilcoxon秩和检验的效应量,计算方式为标准化的Wilcoxon统计量z除以sqrt(n)。
    - `power`: 这是在给定效应量和样本量的情况下,Wilcoxon秩和检验能够检测到真实效应的概率。
3. `power_anova_test`函数:
    - `eta_sq`: 这是η^2效应量,用于衡量在单因素方差分析中,因子对总变异的解释程度。它等于因子平方和除以总平方和。
    - `f`: 这是f效应量,与η^2相关,计算方式为sqrt(eta_sq / (1 - eta_sq))。
    - `power`: 这是在给定效应量和样本量的情况下,单因素方差分析能够检测到真实效应的概率。
4. `power_fisher_chisq_test`函数:
    - `effect_size`: 这是效应量,对于2x2列联表,计算方式为sqrt(chi_sq / n);对于其他情况,计算方式为sqrt(chi_sq / (n * (k - 1)))。
    - `power`: 这是在给定效应量和样本量的情况下,Fisher's精确检验或卡方检验能够检测到真实效应的概率。

总的来说,这些函数中使用的统计量包括Cohen's d、Wilcoxon秩和检验的r、η^2、f和卡方检验的效应量,它们都是衡量效应大小的指标。通过这些指标和样本量,可以计算出相应检验的统计力。

怎么阐述这些理由：
The current study has not enough power - it can detect differences between groups with the probability of xx. That may cause high-level of type-II error. If we want to achieve at least 0.8 power, we need to have that many animals in each group, and the groups must be balanced.

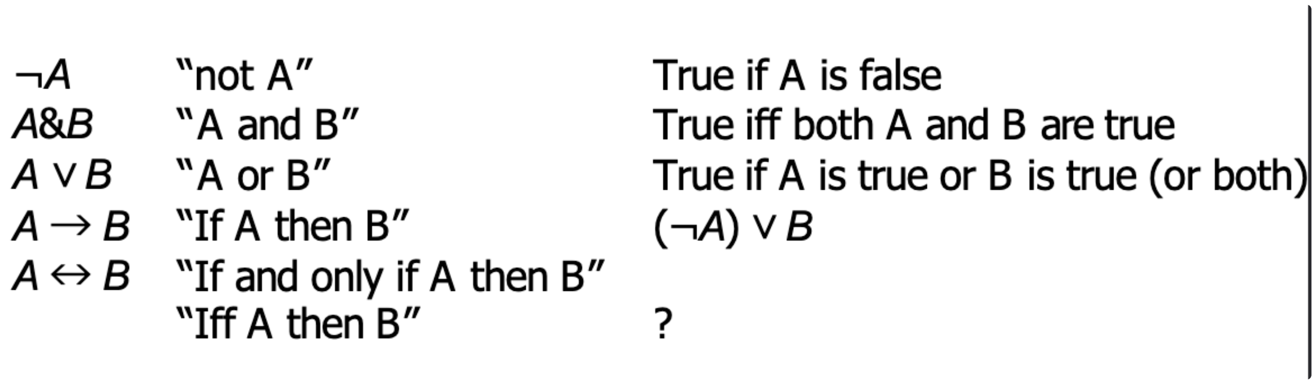
然后下面用power = 0.8作为阈值，看看需要的n大小是多少

```
pwr.anova.test(k = k, f = f, sig.level = alpha, power=0.8)
```

## 四。概率类

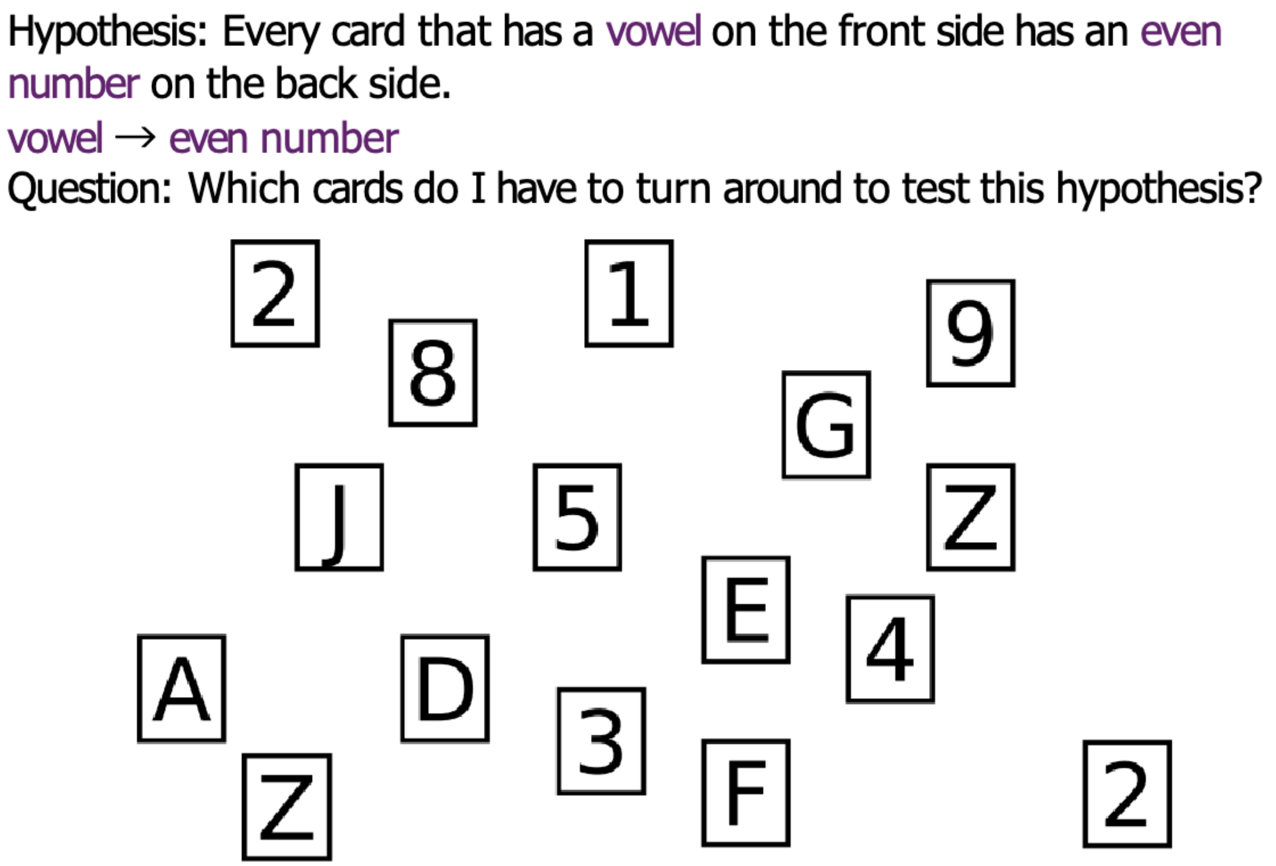
贝叶斯定理

### Logical foundations



| | | |
|---|---|---|
| $\neg A$ | “not A” | True if A is false |
| $A\&B$ | “A and B” | True iff both A and B are true |
| $A \vee B$ | “A or B” | True if A is true or B is true (or both) |
| $A \rightarrow B$ | “If A then B” | $(\neg A) \vee B$ |
| $A \leftrightarrow B$ | “If and only if A then B”<br>“Iff A then B” | ? |

### Conditional probabilities



Hypothesis: Every card that has a vowel on the front side has an even number on the back side.

vowel → even number

Question: Which cards do I have to turn around to test this hypothesis?

- we need to turn around odd number

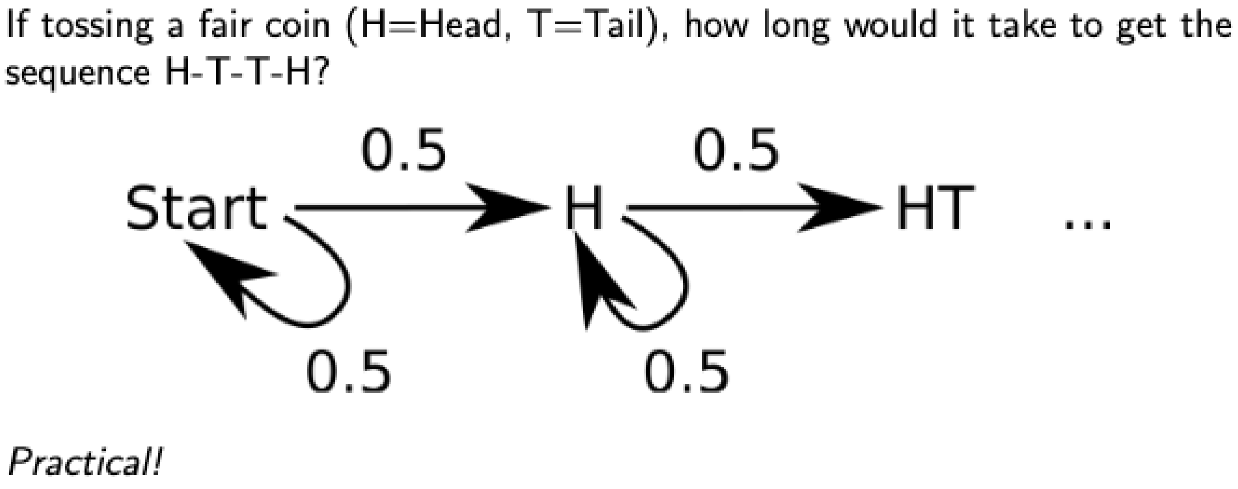
- Because we already know only when we get the: condition True, but conclusion false situation can we fully reject the hypothesis. Other situation can't strongly prove our hypothesis

### Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

### Markov chain

- Stochastic model
- 我的理解，本质上就是计算出每一个状态的转变概率矩阵，链图是可视化的过程
- Probabilities of the state transitions depend on the state the system is currently in, not its history



If tossing a fair coin (H=Head, T=Tail), how long would it take to get the sequence H-T-T-H?

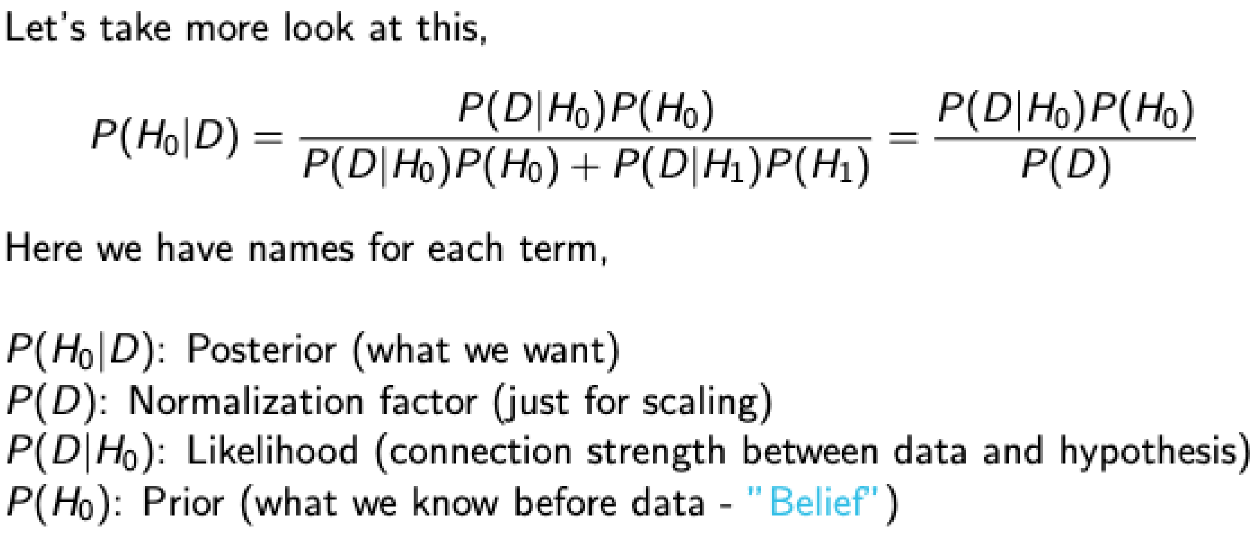
*Practical!*

### Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

### Beyes factor

We have two hypothesis H0 and H1, and some data D



Let's take more look at this,

$$P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0) + P(D|H_1)P(H_1)} = \frac{P(D|H_0)P(H_0)}{P(D)}$$

Here we have names for each term,

$P(H_0|D)$: Posterior (what we want)
$P(D)$: Normalization factor (just for scaling)
$P(D|H_0)$: Likelihood (connection strength between data and hypothesis)
$P(H_0)$: Prior (what we know before data - "Belief")

- Prior: 在本次计算收集数据之前，基于经验和之前的计算得到的概率
  My prior is just a guess. What if it's a bad guess?
- That's OK. The priors are just a starting point. The whole point is that we are updating our beliefs when we get new information.

How do I get P(D|H)?

- Often this can be done using simulation. You have done this before!

How do I know P(D)?

- If you are comparing two hypotheses, the nice thing is you don’t need to know it

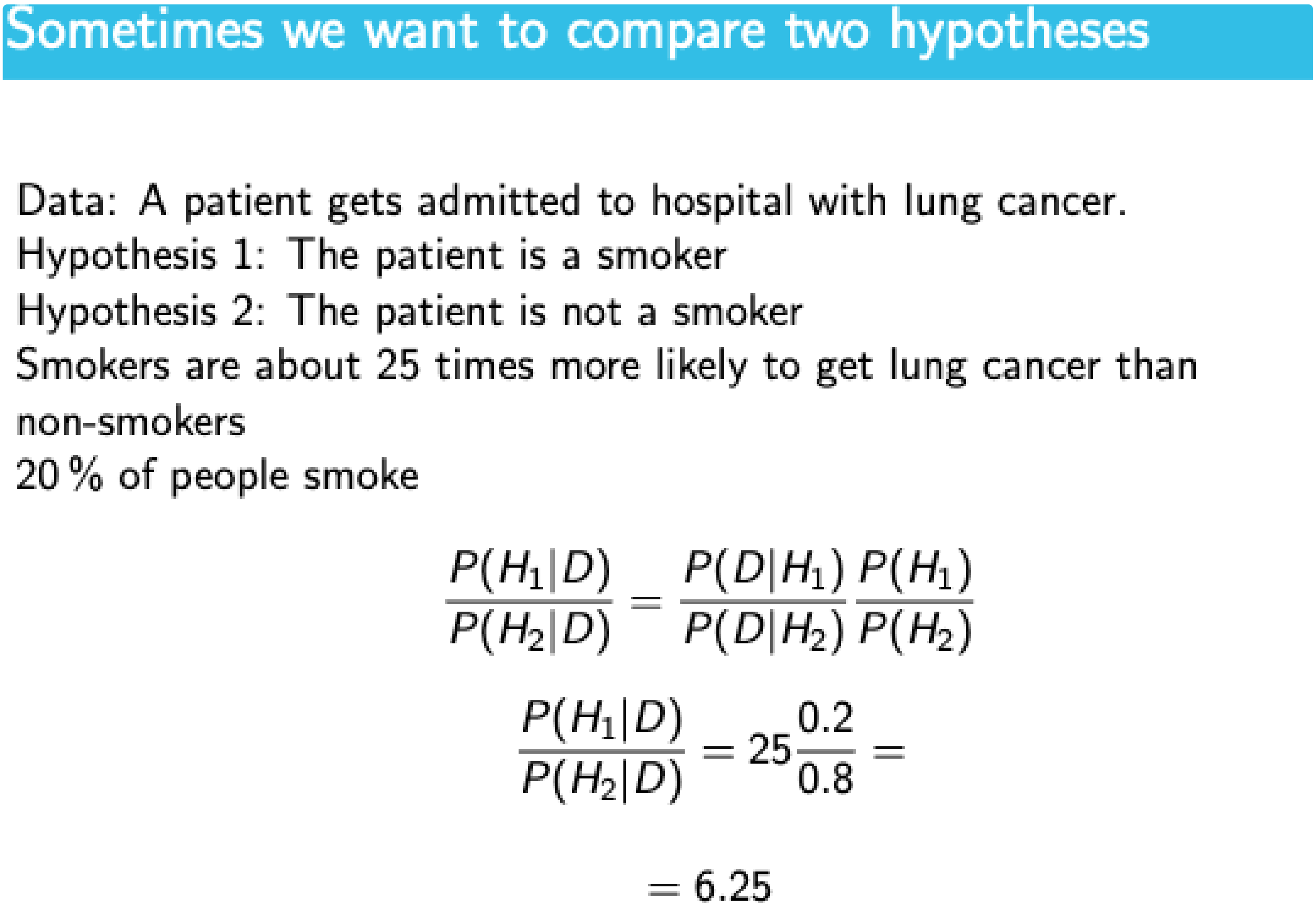
some times we want to compare two hypotheses

If we consider two hypothesis: $H_2, H_1$
Then

$$\frac{P(D|H_1)}{P(D|H_2)}$$

we called Bayes factor ($\alpha$): to judge whether the two hypothesis are equal or not

$$\frac{P(H_1|D)}{P(H_2|D)} = \alpha \frac{P(H_1)}{P(H_2)}$$



## Sometimes we want to compare two hypotheses

Data: A patient gets admitted to hospital with lung cancer.
Hypothesis 1: The patient is a smoker
Hypothesis 2: The patient is not a smoker
Smokers are about 25 times more likely to get lung cancer than non-smokers
20 % of people smoke

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \frac{P(H_1)}{P(H_2)}$$

$$\frac{P(H_1|D)}{P(H_2|D)} = 25 \frac{0.2}{0.8} =$$

$$= 6.25$$

## 五。Kmeans

### 1. turtles

### Introduction

In this report, we will analyze a problem related to turtle populations on a small island with two beaches: West Beach and East Beach. The goal is to determine the probability of being on East Beach given that a Loggerhead Turtle is found. We will use Bayes' theorem and R programming to solve this problem.

### Problem Statement

An ecologist studying turtles on a small island with two beaches knows the following information about the turtle population:

- On West Beach, 90% of turtles are Green Sea Turtles, and the remaining 10% are Loggerhead Sea Turtles.
- On East Beach, 60% of turtles are Green Sea Turtles, while 40% are Loggerhead Turtles.

On a foggy day, the ecologist gets lost on the island. After hours of walking, they reach a beach but cannot determine which one it is due to the dense fog. The ecologist finds a turtle and examines it, discovering that it is a Loggerhead Turtle.

The question is: What is the probability that the ecologist is on East Beach? Additionally, we need to state the assumptions made to arrive at this probability.

### Assumptions

To solve this problem, we make the following assumptions:

1. The ecologist is either on West Beach or East Beach; there are no other possibilities.
2. In the foggy weather, the probability of reaching West Beach or East Beach is equal, i.e., 50% each.

### Solution

We will use Bayes' theorem to calculate the probability of being on East Beach given that a Loggerhead Turtle is found.

```
# Define known conditions
p_west <- 0.5  # Probability of reaching West Beach
p_east <- 0.5  # Probability of reaching East Beach
p_loggerhead_given_west <- 0.1  # Probability of finding a Loggerhead Turtle on West Beach
p_loggerhead_given_east <- 0.4  # Probability of finding a Loggerhead Turtle on East Beach

# Apply Bayes' theorem to calculate the probability of being on East Beach given a Loggerhead
Turtle is found
p_east_given_loggerhead <- (p_loggerhead_given_east * p_east) /
  (p_loggerhead_given_west * p_west + p_loggerhead_given_east * p_east)

# Print the result
cat("The probability of being on East Beach given that a Loggerhead Turtle is found is:",
p_east_given_loggerhead, "\n")
```

### Conclusion

Based on the given information and assumptions, the probability of being on East Beach given that a Loggerhead Turtle is found is `r p_east_given_loggerhead`, or 80%.

This result relies on the assumptions that the ecologist is either on West Beach or East Beach and that the probability of reaching each beach in the foggy weather is equal. If these assumptions do not hold, the calculated probability may differ. For example, if the probability of reaching West Beach in the foggy weather is higher, the ecologist might still be more likely to be on West Beach even after finding a Loggerhead Turtle.

## 2. Classifying neuron types from electrophysiological recordings

```
library(tidyverse)
vmndata <- read.csv("/Users/chen_yiru/Desktop/Desk/Projects/incourse/大二下/ADS_files/vmndata.csv")
head(vmndata)
```

```
colSums(is.na(vmndata))
```

No NA value here.

```
vmndata_duplicate <- duplicated(vmndata)

sum(vmndata_duplicate)
```

No duplicates.

```
ggplot(data = vmndata, aes(x = hap1, y = hap2, color= type)) + geom_point() + ggtitle("Original
classification")
```

#### 寻找最佳聚类数

虽然这里因为已经知道一共有五类，但是如果是不知道的情况下还是要这一步

```
dots <- vmndata[,c(2,3)]
library(factoextra)
set.seed(123)
fviz_nbclust(dots, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)
```

```
model <- kmeans(dots, 5)
model
vmndata$cluster <- as.factor(model$cluster)
ggplot(data = vmndata, aes(x = hap1, y = hap2, color= cluster)) + geom_point() + ggtitle("Cluter
classification")
```

### Test the clustering using different subsets of the fit parameters

```
sub_hap1 <- vmndata[,2]
hap1_result <- kmeans(sub_hap1,5)
vmndata$hap1_cluster <- as.factor(hap1_result$cluster)
ggplot(data = vmndata, aes(x = hap1, y = hap2, color= hap1_cluster)) + geom_point() +
ggtitle("Hap1 classification")
```

```
sub_hap2 <- vmndata[,3]
hap2_result <- kmeans(sub_hap2,5)
vmndata$hap2_cluster <- as.factor(hap2_result$cluster)
ggplot(data = vmndata, aes(x = hap1, y = hap2, color= hap2_cluster)) + geom_point() +
ggtitle("Hap2 classification")
```

### Comparison between the origional classification and cluster classification，评估结果

Adjusted Rand Index (ARI) 是一种用于评估两个数据分配（例如，真实标签和由聚类算法得到的标签）一致性的统计量。它的取值范围是 [-1, 1]，其中：

```
1 表示两个分配完全一致。
0 表示随机一致性，即两个分配的一致性与随机标签的一致性相同。
-1 表示完全不一致，即两个分配的一致性比随机标签的一致性还要差。
```

虽然 ARI 没有绝对的“好”或“坏”的阈值，但通常认为：

```
接近 1：非常好
0.7 到 0.9：好
0.4 到 0.69：一般
0.2 到 0.39：较差
接近 0：随机
负值：比随机还差
```

```
# Example of calculating ARI
library(CommKern)
ari_score <- adj_RI(vmndata$type, vmndata$cluster)
print(ari_score)
```

## Report

Certainly! Below is a brief report written in English based on the R code you provided:

---

**Report on the Analysis of Neuronal Electrical Activity Classification**

**Introduction:**
This report presents an analysis of electrical activity recordings from 25 neurons, classified into 5 types by a colleague based on their activity patterns. We aim to independently classify these recordings using a model with two fit parameters, `hap1` (half-life or time constant) and `hap2` (magnitude), and compare the results with the original classifications.

**Data Import and Preliminary Checks:**
We began by importing the original data from the file `vmndata.csv` using the `tidyverse` package in R. Preliminary checks for missing values and duplicates were performed, confirming that the data set contains no NA values and no duplicated entries.

**Original Classification Visualization:**
To visualize the original classifications, a scatter plot was created with `hap1` on the x-axis, `hap2` on the y-axis, and colors representing the `type` of neuron classification. This plot provides a clear visual representation of how the neurons were originally classified based on their electrical activity patterns.

**Clustering Analysis:**
Using the `kmeans` function in R, we performed clustering on the model fit data to create our own classification of the neuron recordings. The choice of 5 clusters was informed by the original analysis. The resulting clusters were added to the data set for further visualization and comparison.

**Clustering Visualization:**
A scatter plot similar to the original classification plot was created, but this time with colors representing the clusters identified by the `kmeans` algorithm. This visual comparison allows us to assess the differences and similarities between the original classifications and the clusters derived from our model.

**Subset Clustering Analysis:**
To test the robustness of our clustering, we performed separate `kmeans` analyses using only `hap1` and `hap2` as the sole variables. This allowed us to see how sensitive the clustering results are to each parameter individually.

**Comparison and Evaluation:**
To quantitatively compare the original classifications with our cluster classifications, we employed the Adjusted Rand Index (ARI), a statistical measure that assesses the agreement between two data partitions. An ARI score close to 1 indicates perfect agreement, while a score around 0 suggests random agreement.

**Conclusion:**
The ARI score obtained from our analysis provides a quantitative measure of how well our clustering aligns with the original classifications. A high ARI score would suggest that our model-based classification is consistent with the colleague's analysis, while a lower score would indicate discrepancies.

**Recommendations:**
Based on the ARI score and visual comparisons, we can make recommendations for further refinement of the model or suggest areas where additional data collection or analysis might be beneficial.

---

This report provides a concise summary of the steps taken in the analysis, the methods used, and the implications of the findings. It is written in clear, non-technical language suitable for a general audience.

## 六。线性回归

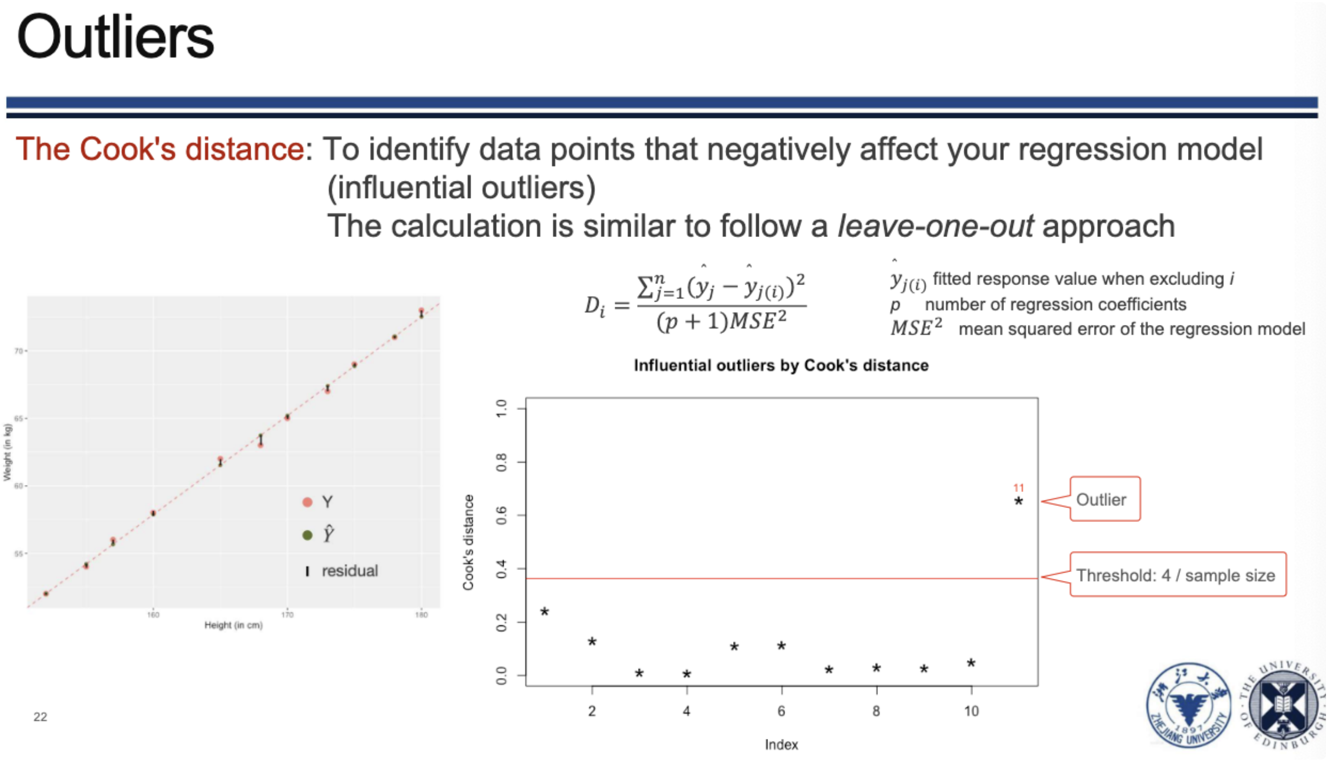
### Correlation

sample correlation coefficient: r, from -1 to 1

### Linear regression

Assumptions:
- The residuals are normally distributed and homeostatic
- The errors are independent
- The relationships are linear

## Outliers



# Outliers

The Cook's distance: To identify data points that negatively affect your regression model (influential outliers)
The calculation is similar to follow a *leave-one-out* approach

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)MSE^2}$$

$\hat{y}_{j(i)}$ fitted response value when excluding *i*
$p$ number of regression coefficients
$MSE^2$ mean squared error of the regression model

Influential outliers by Cook's distance

Outlier

Threshold: 4 / sample size

22

### 具体代码

#### 首先构建模型。再进行下一步的假设检验

```
model <- lm(y ~ x, data = df)
summary(model)
```

Then we check the assumptions of linear regression for each of the fitted models in the models_late object. It iterates over the list of models and generates four diagnostic plots for each model: residuals vs. fitted values, normal Q-Q plot, scale-location plot, and residuals vs. leverage plot. These plots help assess the assumptions of linearity, normality, homoscedasticity, and the absence of influential observations.

```
par(mfrow = c(2,2))
plot(model)
```

Although not all subsets of real-world data can perfectly meet the assumptions, but we consider that most models fit these assumption and no model strangely break these rules, therefore these linear regression models are robust and acceptable.

### 七。anova非参数

Kruskal-Wallis 测试是一种非参数方法，用于比较三个或更多个独立样本的中位数是否存在显著差异。在R语言中，你可以使用 `kruskal.test()` 函数来执行Kruskal-Wallis 测试。

以下是使用 `kruskal.test()` 函数的基本步骤：

1. **准备数据**：确保你的数据是向量或因子形式，并且每个向量代表一个组。
2. **使用 `kruskal.test()` 函数**：将数据作为参数传递给 `kruskal.test()` 函数。
3. **查看结果**：函数将返回一个列表，其中包含测试结果，包括p值。

下面是一个简单的示例，演示如何使用R语言进行Kruskal-Wallis 测试：

```
# 假设我们有三组数据，分别代表三个不同的处理组
group1 <- c(10, 12, 15, 18)
group2 <- c(8, 9, 11, 13)
group3 <- c(7, 8, 9, 10)
```

```
# 将数据合并为一个向量，并为每个值指定对应的组
data <- c(group1, group2, group3)
groups <- factor(rep(c("Group1", "Group2", "Group3"), each = length(group1)))

# 执行Kruskal-Wallis 测试
kruskal.test(data ~ groups)
```

在这个示例中，`data`是一个包含所有观测值的向量，`groups`是一个因子，指示每个观测值属于哪个组。`kruskal.test(data ~ groups)`会进行Kruskal-Wallis 测试，比较三个组的中位数是否存在显著差异。

输出结果将包括以下几个部分：

- **statistic**：Kruskal-Wallis 统计量。
- **parameter**：自由度。
- **p.value**：测试的p值，用于判断组间差异是否显著。

如果p值小于常用的显著性水平（例如0.05），则可以认为至少有两个组之间存在显著差异。然而，Kruskal-Wallis 测试本身并不告诉我们哪些组之间存在差异。为了确定具体哪些组之间有显著差异，你需要进行后续的多重比较测试，如Dunn的测试。在R中，可以使用`dunn.test`包来进行这种多重比较。

结果阐述：
Kruskal-Wallis rank sum test（也称为Kruskal-Wallis H test）是一种非参数统计检验，用于检验两个或多个独立样本的分布是否存在显著差异。在R语言中，`kruskal.test()`函数用于执行此测试。

根据你提供的测试结果：

- **Kruskal-Wallis chi-squared**: 这是Kruskal-Wallis检验的统计量，值为5.6263。
- **df**: 表示自由度（degrees of freedom），在这个测试中，自由度等于组的数量减1，所以这里是2，意味着有两个比较组。
- **p-value**: 检验的p值为0.06001。

**如何解释这些结果：**

1. **Kruskal-Wallis chi-squared**: 这个值表示了组间排名的总体差异。数值越大，组间差异越大。
2. **df**: 这里的自由度是2，意味着除了一个比较组之外，其他所有组都被考虑在内。
3. **p-value**: p值是判断统计显著性的关键。如果p值小于常用的显著性水平（通常是0.05），则结果被认为是统计显著的，这意味着组间至少存在一个显著差异。在你的例子中，p值为0.06001，这大于0.05，因此我们不能拒绝零假设，即没有足够的证据表明组间存在显著差异。

**结论**：根据这个Kruskal-Wallis检验的结果，我们可以得出结论，没有足够的证据表明所比较的三个组的中位数存在显著差异。

**后续步骤**：尽管Kruskal-Wallis检验表明没有显著差异，但如果你仍然对组间可能存在的差异感兴趣，可以考虑进行后续的多重比较测试，比如使用Nemenyi后续测试，来确定哪些组之间的差异是显著的。在R中，可以使用`pairwise.wilcox.test()`函数或`dunn.test()`包来进行这种多重比较。

## 八。bootstrapping

在这段R Markdown代码中，代表bootstrapping思想的代码片段是以下几段：

1. 这部分代码使用bootstrapping方法来估计活性（Active）和抑制（Repressed）状态下`ave`列的中位数：

```
active_med <- c()
repress_med <- c()
for (rep in 1:100) {
  active_sample <- sample(active_rep$ave, size = length(active_rep), replace = T)
  repress_sample <- sample(repress_rep$ave,size = length(repress_rep),replace = T)
  active_med <- c(c(active_med),median(active_sample))
  repress_med <- c(c(repress_med),median(repress_sample))
}
```

2. 这部分代码通过多次随机抽样来估计`result`中`1`的数量的分布，并计算其均值和标准差：

```
num_count <- c()
for (rep in 1:1000) {
  sample_num <- sample(result,276,replace = T)
  num_count <- c(length(sample_num[sample_num==1]),c(num_count))
}
```

3. 这部分代码使用一个双层循环来为`movie`数据集中的每个电影计算95%置信区间的上下界，这也是bootstrapping方法的应用：

```
min_list <- c()
max_list <- c()
for (i in 1:length(movie$students)){
  size0 <- 267 - movie$students[i]
  size1 <- movie$students[i]
  sample0 <- rep(0, size0)
  sample1 <- rep(1, size1)
  result <- c(sample0, sample1)
  num_count <- c()
  for (rep in 1:1000) {
    sample_num <- sample(result,276,replace = T)
    num_count <- c(length(sample_num[sample_num==1]),c(num_count))
  }
  quan <- quantile(num_count,probs = c(0.025,0.975))
  result <- as.matrix(quan)
  min_list <- c(c(min_list),result[1])
  max_list <- c(c(max_list),result[2])
}
```

Bootstrapping是一种统计方法，它通过从数据集中进行多次随机抽样（有放回），来估计统计量的分布。在上述代码中，这种方法被用来估计中位数、数量的分布以及构建置信区间。

### 使用bootsrapping进行代替chisq的测试，当chisq不满足

- Null hypothesis: there is no difference between the proportion of students who are satisfied with an early or late opening time
- Alternative hypothesis: there is a difference between the proportion of students or students prefer a late opening time (this would be the equivalent of a one-tailed test)

可以构建verctors, 然后用bootstrapping绘制出分布和置信区间
bootstrapping 的结果比较的时候，尽量用概率（正则化，两组的总数可能不一样）

```
for (a in 1:100) { first_sample <-
mean(sample(first_results, length(first_results), replace = T)) second_sample <-
mean(sample(second_results, length(second_results), replace = T))
first_bootstraps <- c(first_bootstraps, first_sample)
second_bootstraps <- c(second_bootstraps, second_sample)
}
first_upper <- quantile(first_bootstraps, probs = c(0.975))
second_lower <- quantile(second_bootstraps, probs = c(0.025))

boxplot(
first_bootstraps,
second_bootstraps,
notch = T,
names = c('early', 'late'),
ylab = 'Prop. of satisfied button presses'
)
```

## 九。 anova

### 对于ANNOVA的理解

什么情况下可以使用annova：

1. More than 2 populations
   对于多种不同药物对于某种疾病的效果的研究；比较不同国家指标的研究
2. More than 1 predictive variable (factor)
   锻炼和饮食对于健康的影响； effect of genetic background and drugs on stress levels
3. 如果是多way test
   2-way test: 2 factors, 比如effect of age and sex on salary

### One-way annova

Null hypothesis: means of different <> groups are the same (老师标准写法)
或者说
Null hypothesis: There is no effects of class attendance or previous grades on course performance
Alternative hypothesis: means of different <> groups are not the same

核心思想：组内差异对比组间差异，如果这两者差异大就说明组内之间确实有差异；否则可以认为没有什么组间差异

统计前假设条件：

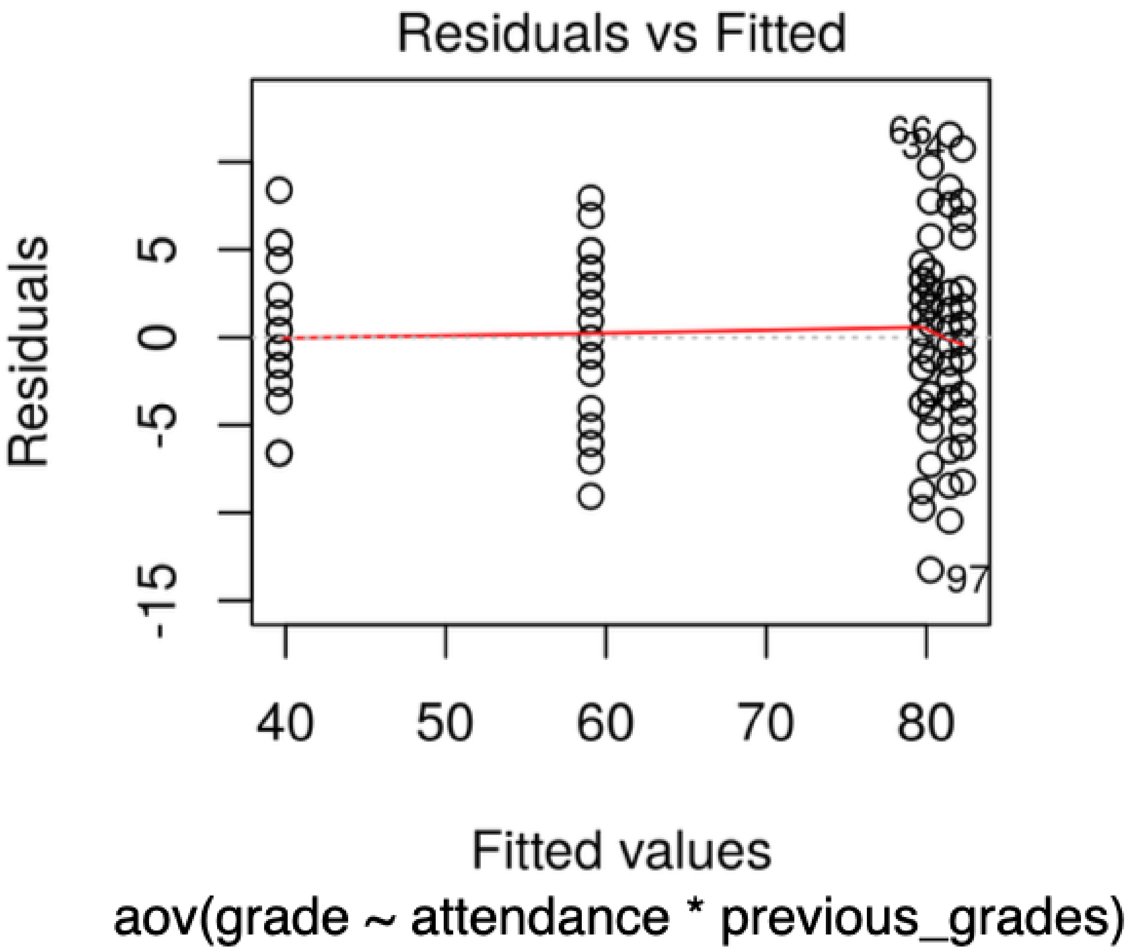
1. Independent random sampling
   We believe that this is true given the description of the experiment itself
2. normality of residuals (distance from group mean)

```
model <- aov(grade ~ attendance * previous_grades) # 这个地方，老师似乎认为没有理由能不使用interaction
hist((resid(model), main = "residuals")# 选一个，方法一
shapiro.test(resid(model)) #方法二
```

3. Equality of variances
   通过作图，“residuals vs fitted” plot进行查看

```
plot(model,1)
```

好的情况：



接下来查看统计量

```
summary(model)
```

实例结果：
Df Sum Sq Mean Sq F value Pr(>F)
Treatment 2 4.46 2.228 1.064 0.353
Measurement 1 2.05 2.049 0.979 0.328
Residuals 47 98.39 2.093

进行完ANNOVA 测试后，如果还想要知道具体是哪一组不同于另外几组，可以采用post-hoc tests。比如Tukey's HSD test

```
TukeyHSD(model)
```

如果想要探索，也可以思考两个factor之间是否有interaction， hypotheses变化：
• H0: There is no interaction between class attendance and previous grades
• HA: There is an interaction between class attendance and previous grades