探索变量间的关系

数据中的变量值得去关注，是因为变量自身的变化（取常值的变量毫无价值），以及变量与变量之间的协变化。

描述统计相当于是探索单个变量自身的变化。比如，连续变量可以用均值等汇总统计量、直方图、箱线图探索其分布；离散变量可以用频率表、条形图等。

探索性数据分析另一重要内容就是探索变量间的关系，也叫作探索协变化。

协变化是两个或多个变量的值以一种相关的方式一起变化。识别出协变化的最好的方式，将两个或多个变量的关系可视化，当然也要区分变量是分类变量还是连续变量

单变量的情况：

参考：R语言描述性的数据分析 – shixiang的文章 – 知乎

https://zhuanlan.zhihu.com/p/520356428

散点图，直方图，箱线图，饼图等。

https://www.yuque.com/sdutzhou/ot65c9/kq9uc5g74dksc2da?singleDoc# 《描述性的数据分析》

https://www.yuque.com/sdutzhou/vo1gky/dgpege?singleDoc# 《箱线图》

1. 两个分类变量

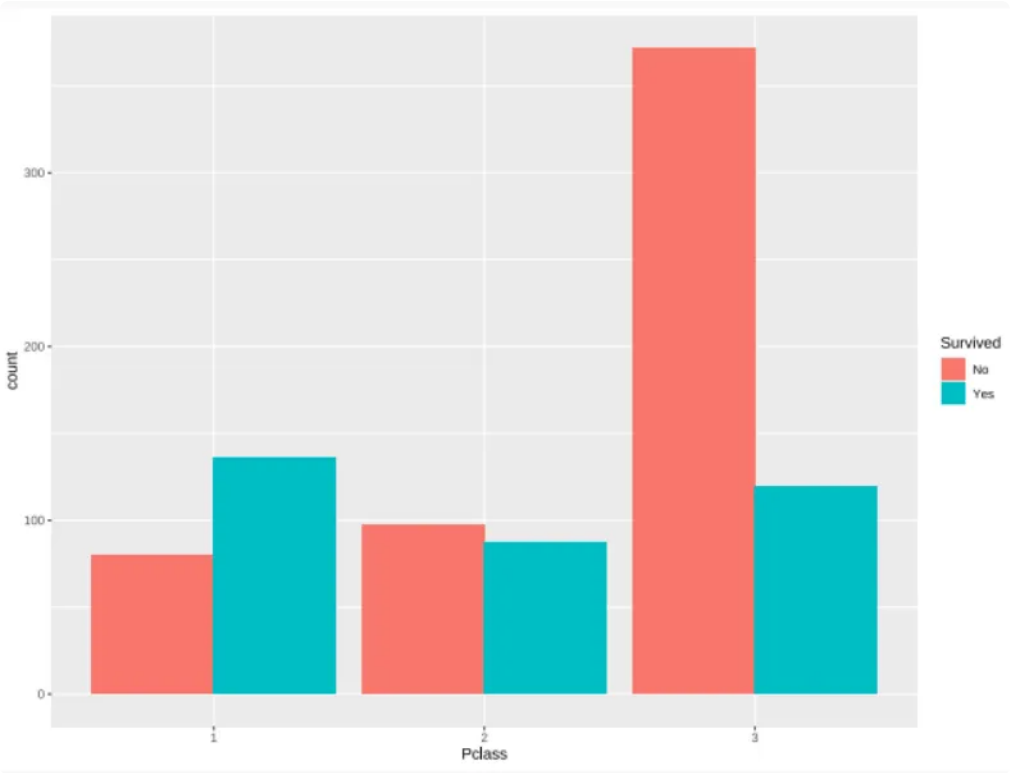
探索两个分类变量的常用方法：

- 可视化：复式条形图、堆叠条形图
- 描述统计量：交叉表
- Cramer's V 统计量：rstatix::cramer_v()
- 假设检验：检验两个比例的差、卡方独立性检验

通过百度网盘分享的文件：titanic.rds

链接：https://pan.baidu.com/s/1ahJ6MoGLmP3iP_lXyumt7Q?pwd=6868

```R
titanic = read_rds("data/titanic.rds")
titanic %>%
ggplot(aes(Pclass, fill = Survived)) +
geom_bar(position = "dodge")
```



列联表

对分类变量做描述统计，通常是计算各水平值出现的频数和占比，得到列联表（交叉表）。以上操作可以用table()函数实现，但功能很弱，也不够简洁。

janitor 包提供了更强大的 tabyl()函数，可以生成一个、两个、三个变量的列联表，再结合adorn_*()函数，可以很方便地按想要的格式添加行列的合计和占比等。

.为一维列联表添加合计行，代码如下：

```R
library(tidyverse)
library(Janitor)
# 加载mpg数据集
data("mpg")
mpg %>%
tabyl (drv) %>%
adorn_totals("row")%>%  # 添加合计行
adorn_pct_formatting()  # 设置百分比格式
```

在R语言中，mpg是一个常用的数据集，它代表了各种汽车型号的城市、高速公路等环境下的燃料经济性能数据。

A tabyl: 4 × 3

	drv	n	percent
	<chr>	<dbl>	<chr>
1	4	103	44.0%
2	f	106	45.3%
3	r	25	10.7%
4	Total	234	100.0%

```R
#为二维列联表添加列占比和频数，代码如下：
mpg %>%
tabyl (drv, cyl) %>%
adorn_percentages ("col") %>% # 添加列占比
adorn_pct_formatting(digits=2) %>%   #设置百分比格式
adorn_ns()  #添加频数


```

A tabyl: 3 × 5

	drv	4	5	6	8
	<chr>	<chr>	<chr>	<chr>	<chr>
1	4	28.40% (23)	0.00% (0)	40.51% (32)	68.57% (48)
2	f	71.60% (58)	100.00% (4)	54.43% (43)	1.43% (1)
3	r	0.00% (0)	0.00% (0)	5.06% (4)	30.00% (21)

Cramer's V 检验法探索分析的代码：

R

```r
titanic = read_rds("data/titanic.rds")
library(rstatix)
tbl = table(titanic$Pclass, titanic$Survived)
cramer_v(tbl) # Cramer'V 检验
#> [1] 0.34
prop_test(tbl) # 比例检验
#> # A tibble: 1 x 5
#> n statistic df p p.signif
#> * <dbl> <dbl> <dbl> <dbl> <chr>
#> 1 891 103. 2 4.55e-23 ****
```

卡方检验是针对无序分类变量的非参数检验，其理论依据是实际观察频数f_0与理论频数f_e(又称期望频数）之差的平方再除以理论频数所得的统计量，近似服从 χ^2 分布。

卡方检验一般用来检验无序分类变量的实际观察频数和理论频数分布之间是否存在显著差异，要求如下：

. 分类变量相互排斥，互不包容；

· 观测相互独立；

.样本容量不宜太小，理论频数大于或等于5,否则需要进行校正（合并单元格或校正卡方值）。

卡方检验常用于检验某分类变量各类的出现概率是否等于指定概率；检验两个分类变量是否相互独立；检验两组频数是否来自同一总体。

以检验Titanic船舱等级与是否生存之间是否相互独立为例

其原假设和备择假设是：H0:相互独立，H1:不相互独立

```R
titanic = read_rds("/kaggle/input/titanic-rds/titanic.rds")
tbl = titanic %>%
  janitor::tabyl(Survived,Pclass)
tbl

```

A tabyl: 2 × 4

Survived	1	2	3
<fct>	<dbl>	<dbl>	<dbl>
No	80	97	372
Yes	136	87	119

```R
rstatix::chisq_test(titanic$Survived,titanic$Pclass)
```

A rstatix_test: 1 × 6

	n	statistic	p	df	method	p.signif
	<int>	<dbl>	<dbl>	<int>	<chr>	<chr>
X-squared	891	102.889	4.55e-23	2	Chi-square test	****

P值几乎等于0,因此拒绝原假设，故结论是船舱等级与是否生存之间有关联。若要进一步比较各等级的船舱之间生存率是否有差异，可使用以下代码：

```r
library(rstatix)
pairwise_prop_test(as.matrix(tbl[,-1]))
```

A rstatix_test: 3 × 5

	group1	group2	p	p.adj	p.adj.signif
	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	1	2	2.32e-03	2.32e-03	**
2	1	3	1.21e-22	3.64e-22	****
3	2	3	1.22e-08	2.45e-08	****

```r
chisq_test(tbl) # 卡方检验
#> # A tibble: 1 x 6
#> n statistic p df method p.signi
#> * <int> <dbl> <dbl> <int> <chr> <chr>
#> 1 891 103. 4.55e-23 2 Chi-square test ****

```

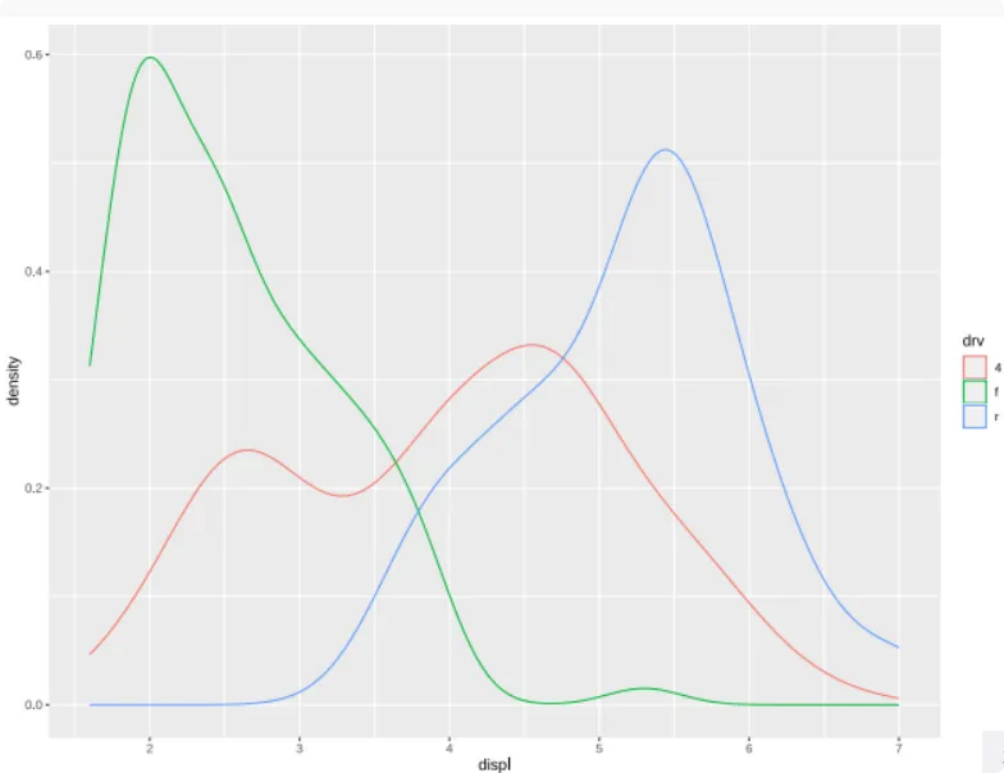
Cramer'V 统计量是修正版本的 Φ 系数，一般法则是：$|\Phi| < 0.3$, 很少或没有相关性；$0.3 \leqslant |\Phi| \leqslant 0.7$, 有弱相关性；$|\Phi| > 0.7$, 有强相关性。

2. 分类变量与连续变量

探索分类变量与连续变量的常用方法：

- 可视化：按分类变量分组的箱线图、直方图、概率密度曲线；
- 描述统计：按分类变量分组汇总；
- 比较均值的假设检验：t 检验、方差分析、Wilcoxon 秩和检验等

```R
ggplot(mpg, aes(displ, color = drv)) +
geom_density() # 概率密度曲线
```



按分类变量分组汇总

```R
mpg %>%
group_by(drv) %>%
get_summary_stats(displ, type = "five_number") # 五数汇总
#> # A tibble: 3 x 8
#> drv variable n min max q1 median q3
#> <chr> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 4 displ 103 1.8 6.5 2.9 4 4.7
#> 2 f displ 106 1.6 5.3 2 2.4 3
#> 3 r displ 25 3.8 7 4.6 5.4 5.7
```

用方差分析探索：用于分析定类数据与定量数据之间的关系情况.例如研究人员想知道三种情况下耗油量是否有显著差异

```r
mpg %>%
anova_test(displ ~ drv) # 方差分析
#> ANOVA Table (type II tests)
#>
#> Effect DFn DFd F p p<.05 ges
#> 1 drv 2 231 110 3.03e-34 * 0.487
```

A anova_test: 1 × 7

	Effect	DFn	DFd	F	p	p<.05	ges
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
1	drv	2	231	109.816	3.03e-34	*	0.487

3. 两个连续变量

探索两个连续变量的常用方法：

- 可视化：散点图（或 + 光滑曲线）、折线图，3 个连续变量可用气泡图
- 线性相关系数：协方差能反映两个变量的影响关系，定义为

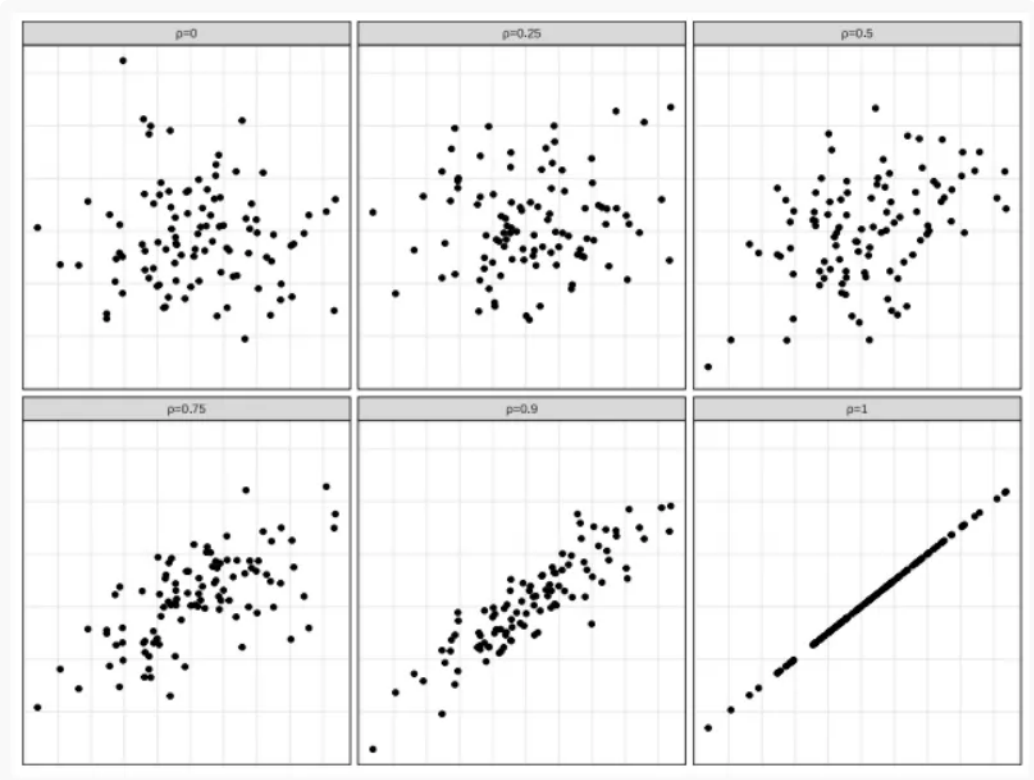
$$\mathrm{Cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x_i})(y_i-\bar{x_i})$$

但是协方差的单位是不一致的，不具有可比性，解决办法就是做标准化，

得到相关系数：

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{s_X s_Y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x_i})(y_i-\bar{x_i})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

线性相关系数介于 −1 和 1 之间，反映了线性相关程度的大小。



用 rstatix 包计算相关系数矩阵，并去掉重复，按相关系数大小排序：

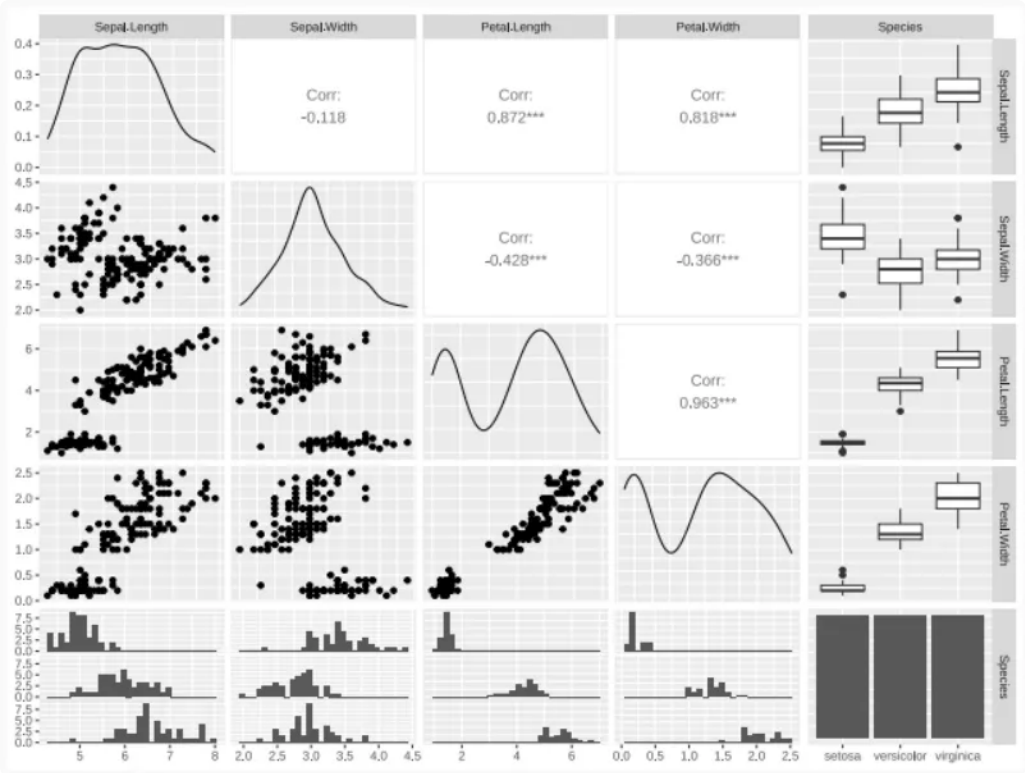
```R

iris[-5] %>%
  cor_mat() %>% # 相关系数矩阵
  replace_triangle(by = NA) %>% # 将下三角替换为 NA
  cor_gather() %>% # 宽变长
  arrange(- abs(cor)) # 按绝对值降序排列
  #> var1 var2 cor p
  #> 1 Petal.Length Petal.Width 0.96 4.68e-86
  #> 2 Sepal.Length Petal.Length 0.87 1.04e-47
  #> 3 Sepal.Length Petal.Width 0.82 2.33e-37
  #> 4 Sepal.Width Petal.Length -0.43 4.51e-08
  #> 5 Sepal.Width Petal.Width -0.37 4.07e-06
  #> 6 Sepal.Length Sepal.Width -0.12 1.52e-01
```

注意：统计相关并不代表因果相关！线性不相关也可能具有非线性关系！

GGally 包提供的 ggpair() 函数绘制散点图矩阵，非常便于可视化探索因变量与多个自变量之间的相关关系：

```
library(GGally)
ggpairs(iris, columns = names(iris))
```



实际中，经常需要从许多自变量中筛选对因变量有显著影响的，根据相关系数是方法之一，更系统的方法是机器学习中的特征选择。另外，correlationfunnel 包能够快速探索自变量，特别是大量分类变量，对因变量的相关性影响大小，并绘制”相关漏斗图”进行可视化。

最后，还可以通过构建线性回归或广义线性回归模型，查看回归系数是否显著来探索自变量（无论是连续还是分类）对因变量的影响。

参考文献

Hadley Wickham, G. G. (2017). R for Data Science. O'Reilly, 1 edition. ISBN 978-1491910399.

张敬信 (2022). R 语言编程：基于 tidyverse. 人民邮电出版社, 北京.

谢益辉 (2021). rmarkdown: Dynamic Documents for R.

锡南•厄兹代米尔, 迪夫娅•苏萨拉，. (2019). 特征工程入门与实践. 人民邮电出版社, 北京.

黄湘云 (2021). Github: R-Markdown-Template.

https://www.zhihu.com/column/c_1530113383386419200 建模专栏

https://www.zhihu.com/column/c_1509181173906006016 数学软件专栏