Data Analysis Foundation

Descriptive Statistics and Probability Bacis Thory

Ma Ming
Big Data Platform Architect and Data Scientist.

Course Introduction

- Prelimitary
- 总体和样本(Population and Sample)
- 描述性统计(Descriptive Statistics)
- 可视化(Visualization)
- 概率分布(Probability distribution)
  - Normal Distribution
  - Binomial Distribution
  - Possion Distribution

Prelimitary

- Data Basics
  - Observations,Variables and data matrics or data frame
  - Type of Variables
- Random Variables
- Relationships between Variables
- PMF,PDF,CDF,SF

mtcars dataset

	MPG	CYL	DISP	HP	DRAT	WT	QSEC	VS	AM	GEAR	CARB
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Type of Variables

- Numerical (Quantitative)
  - Continous
  - Discrete
- Categorical (Qualitative)
  - Nominal
  - Ordinal

Airquality Dataset

OZONE	SOLAR.R	WIND	TEMP	MONTH	DAY
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

Random variables（随机变量）

- 一个 random variable 是一个实验或观察结果的数字形式输出.
- Random Variable有两种形式,
  discrete or continuous.
- 离散型random variable ： 有限的可能性.
  $P(X = k)$
- 连续型random variable接受一个实数范围.
  $P(X \in A)$

可以被考虑为是随机变量的例子

- 扔硬币的输出数字化$(0-1)$
- 扔骰子的输出
- 某天某网站的访问量
- 点击广告的人数
- Question? : How to describe the distribution of a random variable?

Example 1 : Titanic dataset

X	NAME	PCLASS	AGE	SEX	SURVIVED	SEXCODE
1	Allen, Miss Elisabeth Walton	1st	29.00	female	1	1
2	Allison, Miss Helen Loraine	1st	2.00	female	0	1
3	Allison, Mr Hudson Joshua Creighton	1st	30.00	male	0	0
4	Allison, Mrs Hudson JC (Bessie Waldo Daniels)	1st	25.00	female	0	1
5	Allison, Master Hudson Trevor	1st	0.92	male	1	0
6	Anderson, Mr Harry	1st	47.00	male	1	0

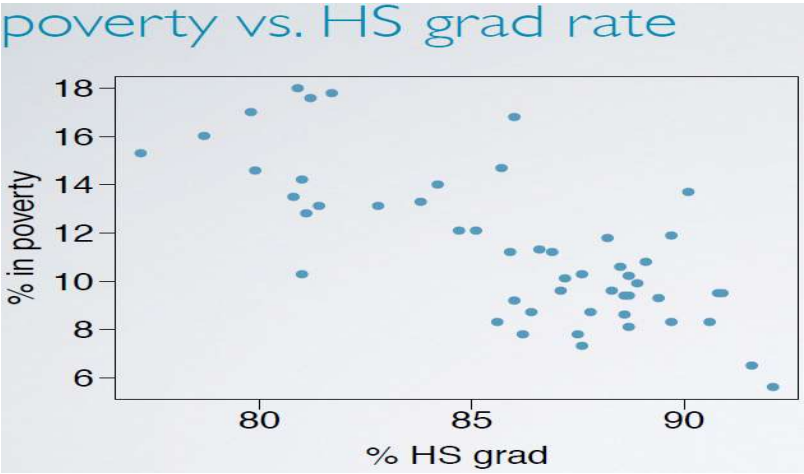
Example 2 : iris dataset

SEPAL.LENGTH	SEPAL.WIDTH	PETAL.LENGTH	PETAL.WIDTH	SPECIES
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

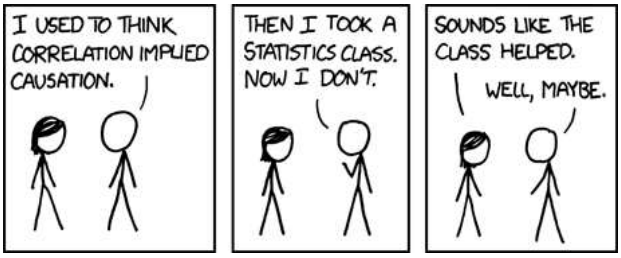
Example 3 : mtcars dataset

	MPG	CYL	DISP	HP	DRAT	WT	QSEC	VS	AM	GEAR	CARB
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Relation between variables

- Correlation
  - Explanatory Variable and Response Variable
  - Positive and Negative
  - Strong and Weak
- 

Example



- Correlation DONOT imply Causation!!!

Example again

- 某调查机构调查2379个9~19岁女生显示大部分吃早餐的女生都比较苗条，如果我们由此得出结论： 吃早餐可以使女士苗条。
  你认为正确吗？
- Correlation DONOT imply Causation!!!

PMF (Probability Mass Function)

概率密度函数的值是随机变量取该值的概率，p

1. $p(x) \geq 0$ for all x
2. $\sum_x p(x) = 1$

Example for PMF

令 X 代表抛硬币的结果，这里 $X = 0$ 代表反面而 $X = 1$ 代表正面。

$$p(x) = (1/2)^x (1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

假如我们不知道硬币是否均匀; θ 表示正面出现的概率(在 0 和 1 之间).

$$p(x) = \theta^x (1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

PDF (Probability Density Function)

概率密度函数描述连续型随机变量的概率特征。pdf下的面积代表了该随机变量在相应范围出现的概率

pdf = f 必须满足

1. $f(x) \geq 0$ for all x
2. The area under $f(x)$ is one.

Example of PDF

思考从求助热线打入的求救电话，其中得到解决的比例

$$f(x) = \begin{cases} 2x & \text{for } 1 > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

这是一个有效的概率密度函数吗？

CDF and Survival function

- 一个随机变量的累积分布函数cumulative distribution function (CDF) X 定义为

$$F(x) = P(X \leq x)$$

- 适用于连续型和离散型随机变量
- 留存函数survival function 定义为

$$S(x) = P(X > x)$$

- 注意 $S(x) = 1 - F(x)$
- 对连续型随机变量来说，PDF 是 CDF 的导数

Example

回想前面热线电话的例子，其CDF和SF:

For $1 \geq x \geq 0$

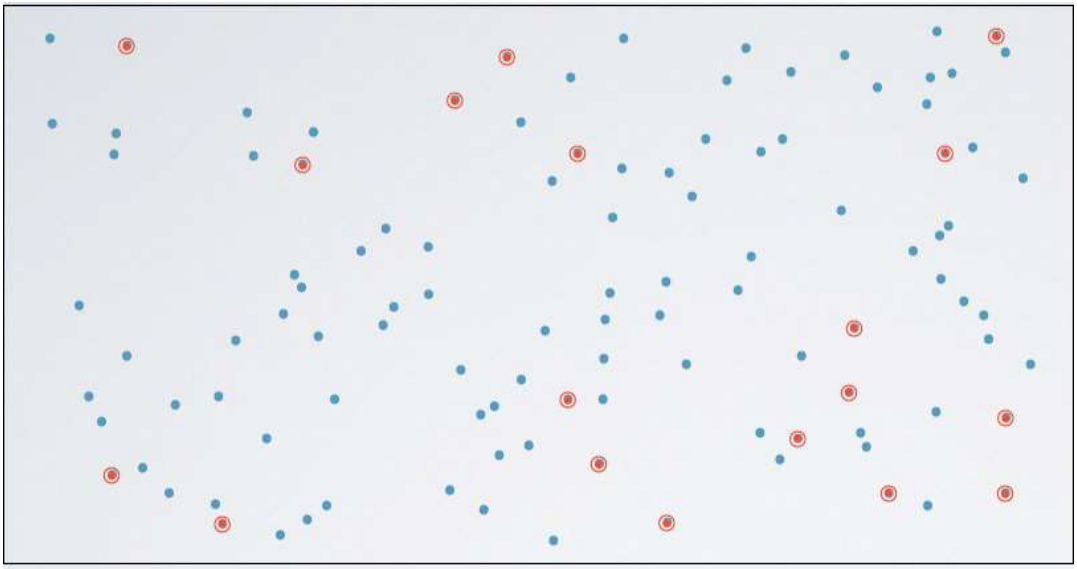
$$F(x) = P(X \leq x) = \frac{1}{2} Base \times Height = \frac{1}{2}(x) \times (2x) = x^2$$

$$S(x) = 1 - x^2$$

Sampling & Sources of bias

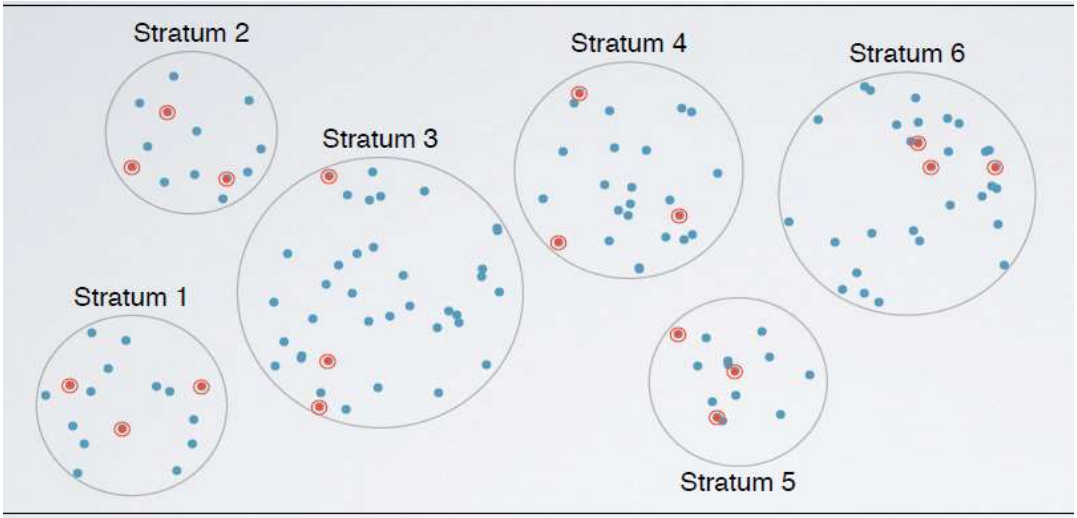
- Population vs Sample
- Sources of Bias
  - Convenience Sample
  - No Response
  - Voluntary Respone
- Sampling methods
  - Simple random sampling (SRS)
  - Stratified sample
  - Cluster sample
  - Multistage sample

Simple Random Sampling



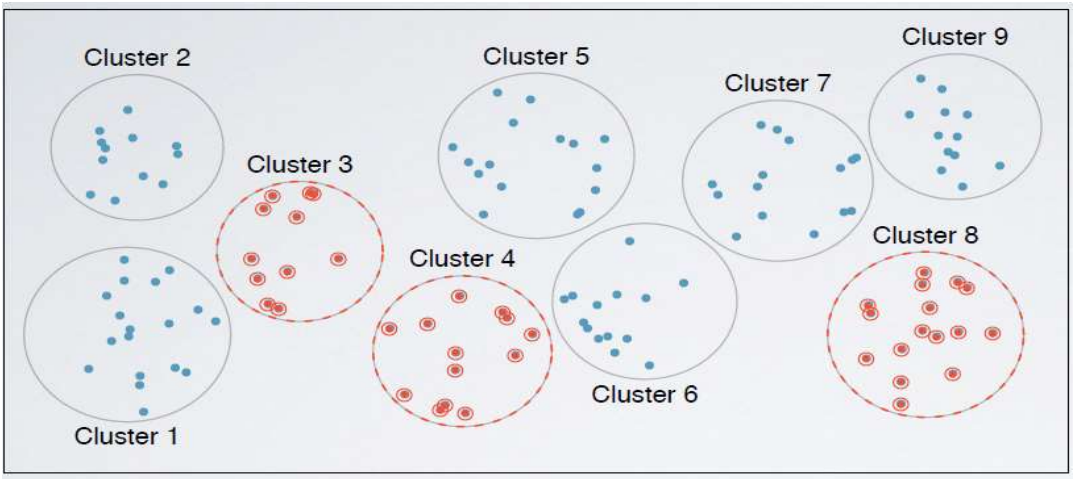
each case is equally likely to be selected

Stratified Sampling



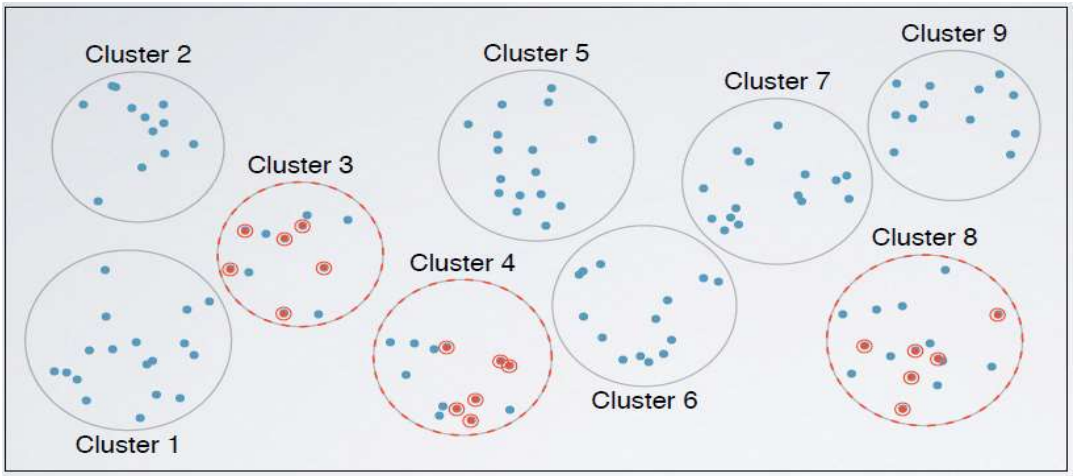
divide the population into homogenous strata, then randomly sample from within each stratum

Cluster Sampling



divide the population clusters,
randomly sample a few clusters,
then sample all observations within these clusters

Multistage Sampling

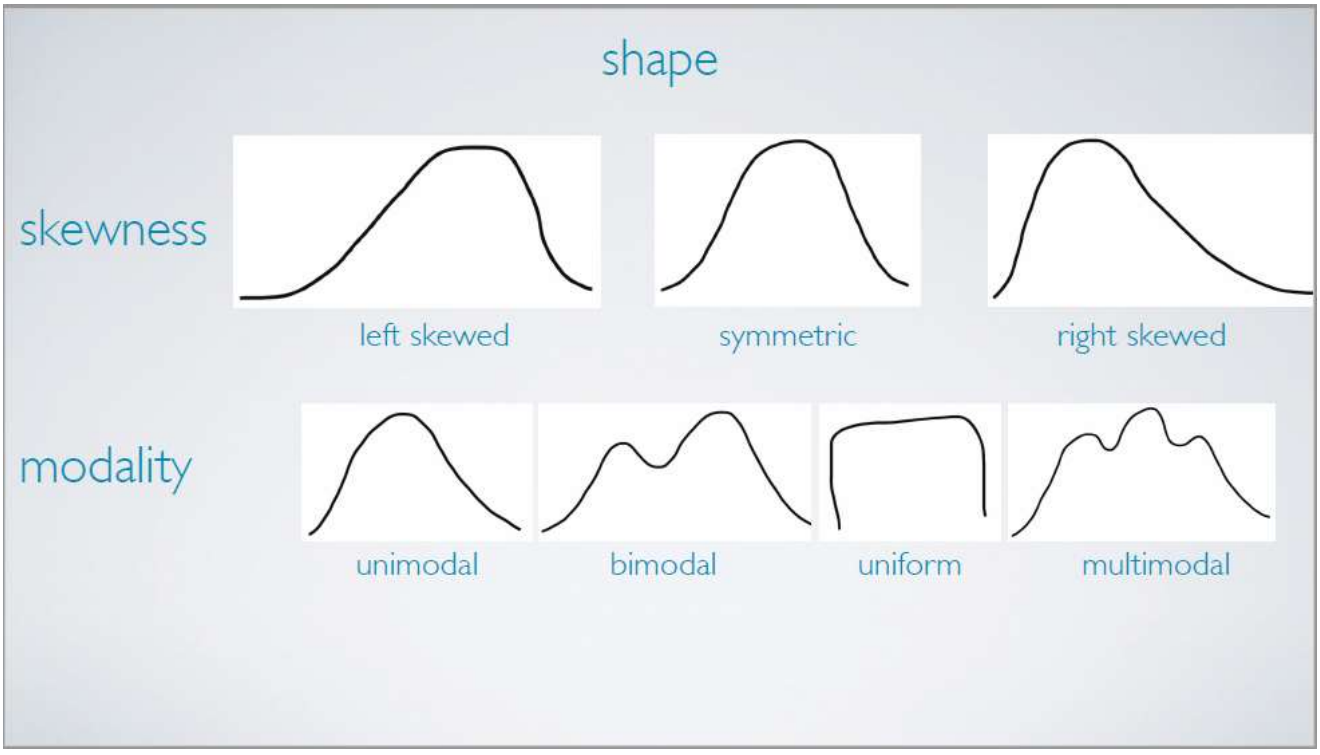


divide the population clusters,
randomly sample a few clusters,
then randomly sample within these clusters

Descriptive Statistics

- Shape
- Center
  - Mean
  - Median
  - Mode
- Spread
  - Range
  - Variance
  - IQR (Inter Quarter Range)
- Fivenum : $(Min, Q_1, Median, Q_3, Max)$
- Visualization

Measure of Shape



Measure of Center

mean

arithmetic average

$\bar{x}$ sample mean

μ population mean

median

midpoint of the distribution (50th percentile)

mode

most frequent observation

sample statistic

point estimate

population parameter

Measure of Spread

- Range : (Max - Min)
- Variance : $S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$
- Standart Deviation : $\sqrt{Var(x)}$
- IQR (Inter Quartile Range) : $Q_3 - Q_1$
- Robust Statistics : Mean , IQR

The population mean(总体均值)

- 随机变量期望值 (expected value) 或者 均值 被认为是随机变量的分布的中心点
- 对于离散型随机变量 X ，其 PMF 为 $p(x)$ 定义为

$$E[X] = \sum_x xp(x).$$

  这里的和是包括所有可能的 x
- $E[X]$ 表示位置和权重的集合中心, $\{x, p(x)\}$

The sample mean(样本均值)

- 样本均值用来估算总体均值
- 经验均值 (empirical mean)

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$$

where $p(x_i) = 1/n$

The population variance(总体方差)

- 方差描述了一个随机变量的 spread(散布程度)
- X 是一个随机变量， 均值为 μ, 其方差定义为：

$$Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

标准差(The standart deviation)： $\sigma = \sqrt{Var}$

- 到均值的期望距离
- 高方差意味着较大的散步程度或强波动性。
- 方差的平方根成为标准差 standard deviation
- 标准差与 X 有同样的单位
- Question: 为什么我们要对到均值距离平方来计算方差?

样本方差(The sample variance)

- The sample variance is

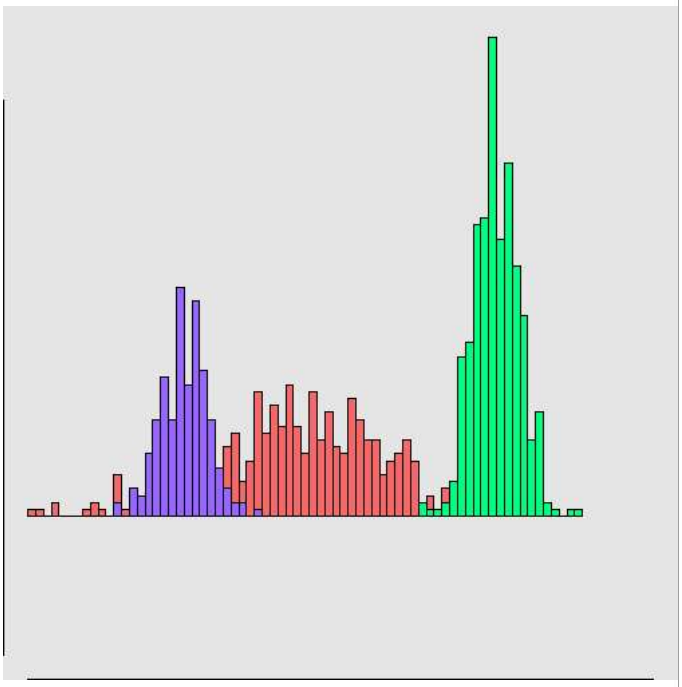
$$S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$$

- 也是一个随机变量
- 其平方根是样本标准差(sample standard deviation)

Histogram (直方图)

直方图是对数值型数据分布的一种图形化表示。

A	B	C
-3.290495	17.64978	-12.13383
-13.176570	22.03377	-23.51500
-16.852019	22.26957	-18.29376
-9.154496	21.23076	-16.71893
-11.472137	19.91917	-19.96881
-6.012846	15.63001	-20.67114
-3.373150	18.12184	-21.80643
-12.350203	20.88682	-20.47307
-10.078285	26.20302	-20.13695
-9.394277	21.68058	-17.90722
2.450722	19.13475	-17.74624
-3.808834	18.22566	-20.19697



Parameters of Histogram

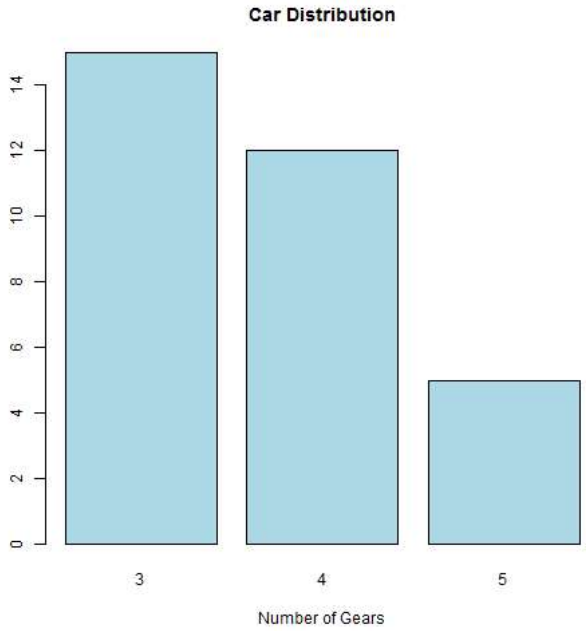
- Number of bins : k
- Or bin width : h
- How to choose k
  - Square-root choice : $k = \sqrt{n}$
  - Sturges' formula : $k = \log_2 n + 1$

BarPlot (条形图)

- 条形图是用矩形条来展现分组数据，矩形条的长度和分组数据的值是成比例的。

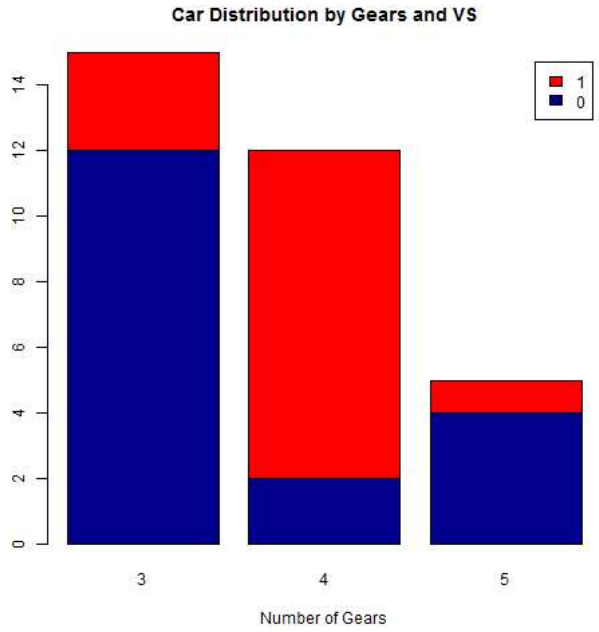
mtcars counting for gear

```
##
##  3  4  5
## 15 12  5
```



Stacked Barplot

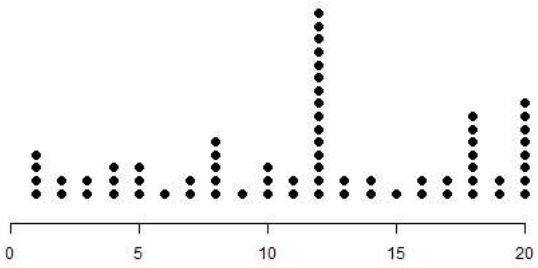
```
##
##     3  4  5
##  0 12  2  4
##  1  3 10  1
```



DotPlot

- The dot plot as a representation of a distribution consists of group of data points plotted on a simple scale. Dot plots are used for continuous, quantitative, univariate data.

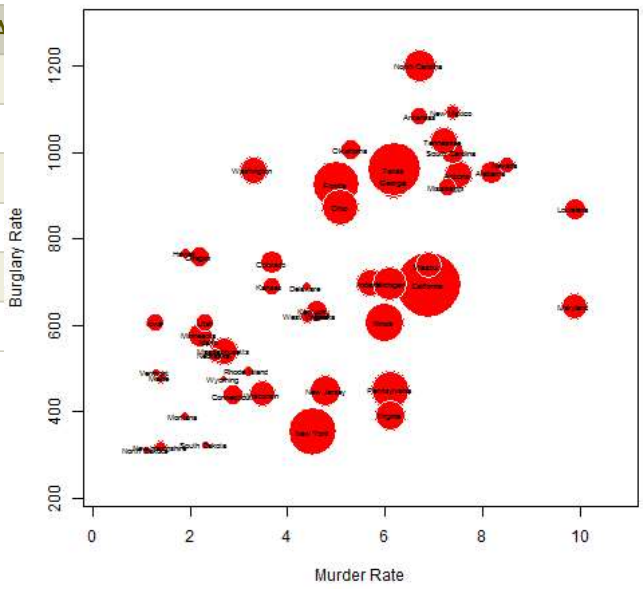
(2,2,3,5,13,20,1,4,3,5,7,8,8,8,10,10,8,8,4,10,15,5,12,
12,14,14,16,16,17,19,19,4,6,9,7,12,13,11,11,17,18,1
18,18,18,18,18,12,12,12,12,12,12,12,12,12,12,12,
12,20,20,20,20,20,20,1,1,1,20)



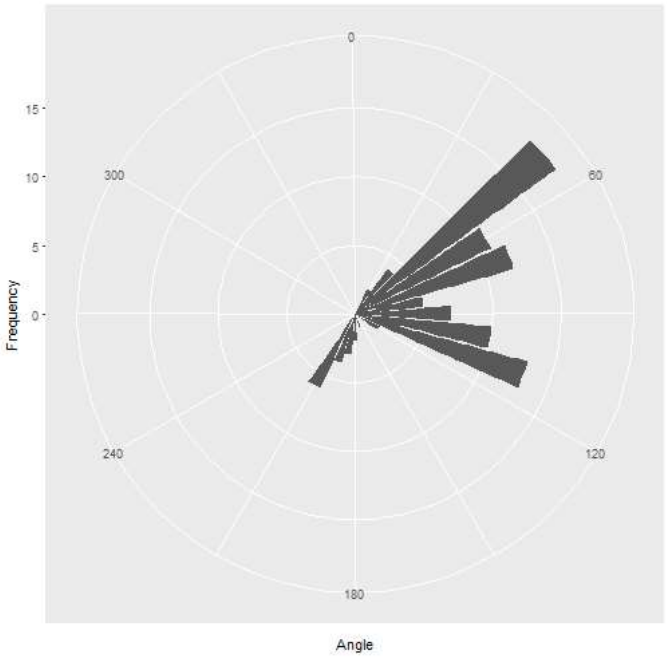
Bubble Plot

The Crime Dataset

STATE	MURDER	FORCIBLE_RATE	ROBBERY	AGGRAVA
Alabama	8.2	34.3	141.4	247.8
Alaska	4.8	81.1	80.9	465.1
Arizona	7.5	33.8	144.4	327.4
Arkansas	6.7	42.9	91.1	386.8
California	6.9	26.0	176.1	317.3
Colorado	3.7	43.4	84.6	264.7

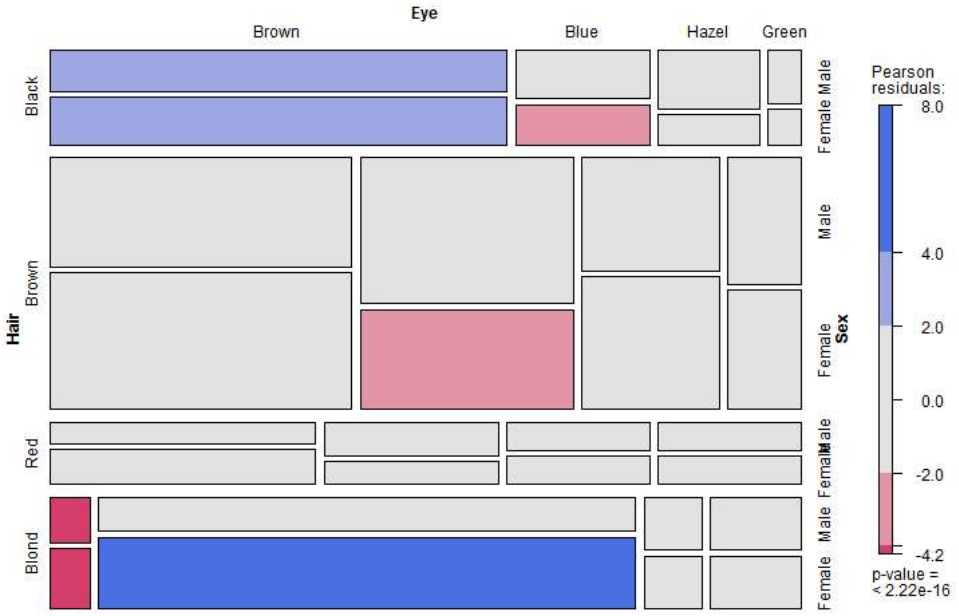


Rose Plot



Mosaic Plot

Mosaic plot常常用来展示多个Categorical data(分类数据)



The Bernoulli distribution

- 伯努利分布(Bernoulli distribution) 是一个两值输出。
- 伯努利随机变量只输出1或者0，对应概率为 p and $1-p$。
- 伯努利随机变量 X ,其PMF

$$P(X=x)=p^x(1-p)^{1-x}$$

- 伯努利随机变量均值为 p ，方差为 $p(1-p)$
- 我们一般把 $X=1$ 认为是 "success" ，而 $X=0$ 认为是 "failure"
- 记为 $X\sim Ber(p)$

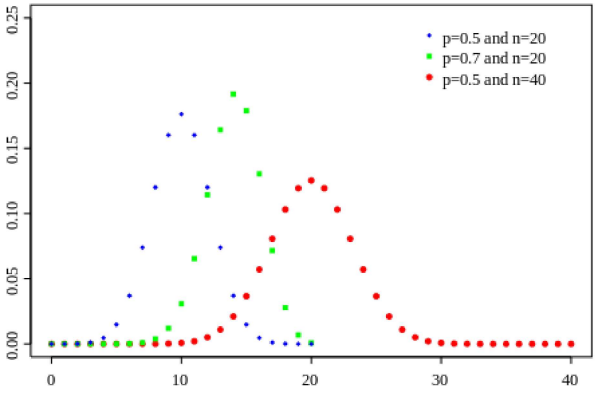
二项式分布(Binomial distribution

- 二项随机变量(binomial random variables)是由独立不相关（iid)的伯努利变量的和得到的。
- 令 $X_1, \ldots, X_n$ 为 iid Bernoulli(p); 那么 $X = \sum_{i=1}^{n} X_i$ 是二项随机变量。
- The binomial mass function is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

对于 $x = 0, \ldots, n$

- 记为 $X \sim Bin(n, p)$
- 考虑扔十次硬币，出现正面的次数



正态分布(The normal distribution)

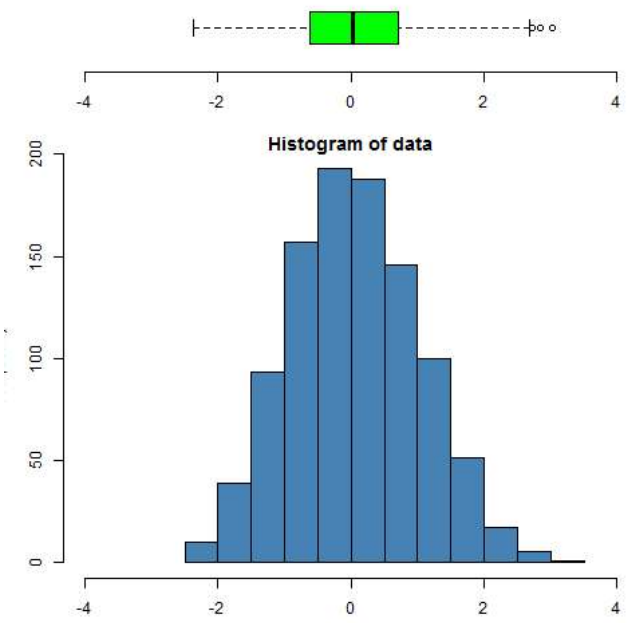
- 一个 正态(normal) 或者 高斯(Gaussian) 分布随机变量均值为 μ ,方差为 σ^2 ，其 概率密度函数:

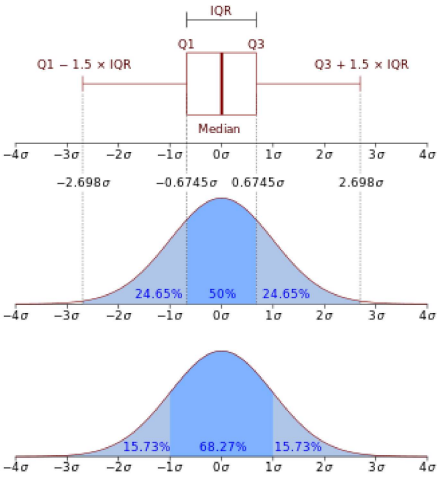
$$(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}$$

  如果 X 是符合此密度函数的随机变量，那么 $E[X]=\mu$ 且 $Var(X)=\sigma^2$

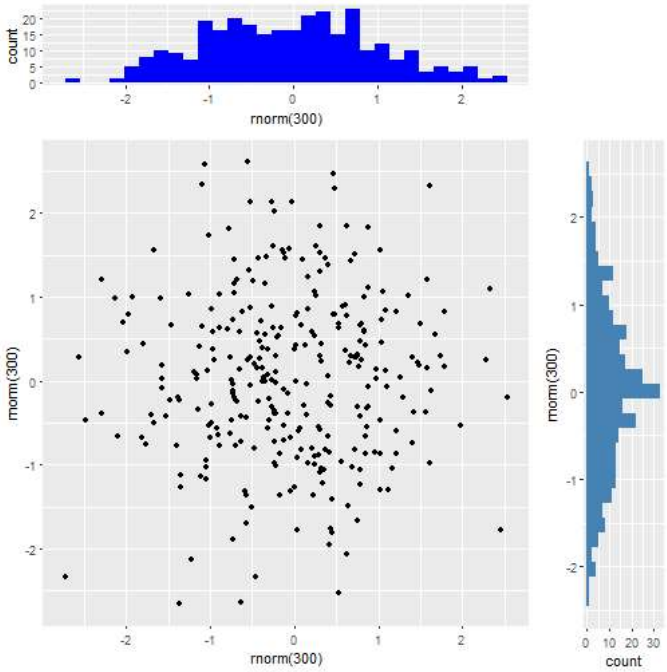
- 记为 $X\sim \mathrm{N}(\mu,\sigma^2)$
- 当 $\mu=0$ 且$\sigma=1$ ，称为 标准正态分布(the standard normal distribution)
- 标准正态分布经常用 Z 来表示

Normal Distribution 特性





Example for Norm and ScatterPlot



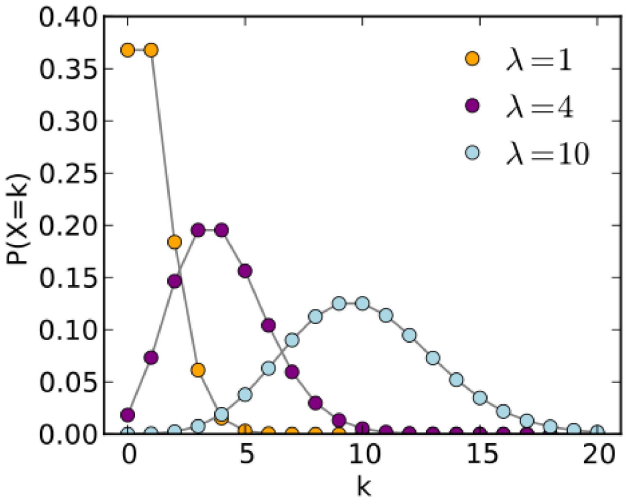
泊松分布(The Poisson distribution)

- 用来对次数进行建模
- Poisson PMF

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

对于 $x = 0, 1, \ldots$

- 均值 λ
- 方差也为 λ
- 注意这里 x 取值范围是从 0 到 ∞
- 记为 $X \sim \pi(\lambda)$



Example Poisson distribution

- Modeling count data
- Modeling event-time or survival data
- Modeling contingency tables
- Approximating binomials when n is large and p is small

Rates and Poisson random variables

- 泊松随机变量经常对单位时间或面积某事件发生的次数建模
- $X \sim Poisson(\lambda t)$,这里
  - $\lambda = E[X/t]$ is the expected count per unit of time
  - t is the total monitoring time

当 n 很大同时 p 很小时，泊松分布是二项式分布的一个精确近似。

- Notation
  - $X \sim \text{Binomial}(n, p)$
  - $\lambda = np$
  - n gets large while p gets small

泊松分布的例子

1. 在一个时间间隔内某电话交换台收到的电话呼叫次数。
2. 一本书一页中的印刷错误数。
3. 某地区一天内邮递丢失的信件数。
4. 某医院一天内的急诊病人数。
5. 某城市一个时间间隔内发生交通事故的次数。

- Question: 在一公共汽车站出现的人数符合Poission分布，均值为2.5 每小时.
  如果我们观察这个车站4小时, 整个时间内小于等于3个人出现的概率是多少?