STAT5002

RQUIZS	week4, week8, week12	12%
ASSIGNMENT	week12	8%
Mid-Term		20%
FINAL		60%

WEEK 1: Introduction

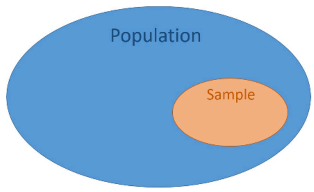
1. Population（集合） and sample（样本）

Sample: Subset of Population,

Large enough

Not Biased

Observations independent



Ensure sample representative: Make sure sample is Random

2. Bias

偏见的来源：不是绝对随机 -> Not Random

a) Selection bias (选择偏见)

例：往篮球队里统计平均身高

b) Measurement bias (测量偏见)

例：用耳内体温计统计人的体温(耳内体温比体外高)

c) Response bias (回应偏见)

例：调查很少得到回应(说明有回应的结果往往存在内在联系)

注意 回应率 response rate 一定要记录

d) Confounding (混淆偏见)

例：查看 冰激凌销量和溺亡人数的关系，两者没有直接关系， 两者都可能是因为气温。准确的说：confounder影响了(升高或降低)一个变量对另一个变量的影响。

3. Study Design

Observational study： No treatment, simply observe, collect data record data -> done

Observational study support infer **association.** 内在关联

需要注意object之间的联系是否合理(预防confounding)

Experimental study: Impose treatment on subject,

Explanatory variable ->Dependent variable(response)

Experimental study support infer **causation.** 内在因果关系

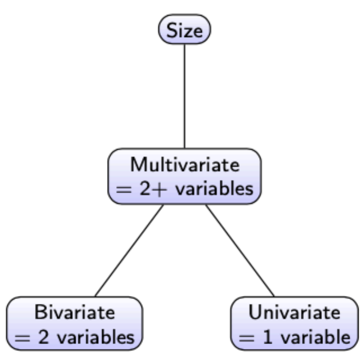
需要注意要考虑到所有变量的联系

4.EDA (Exploratory data analysis 探索性数据分析)

Size of Variables:

p: how many variables

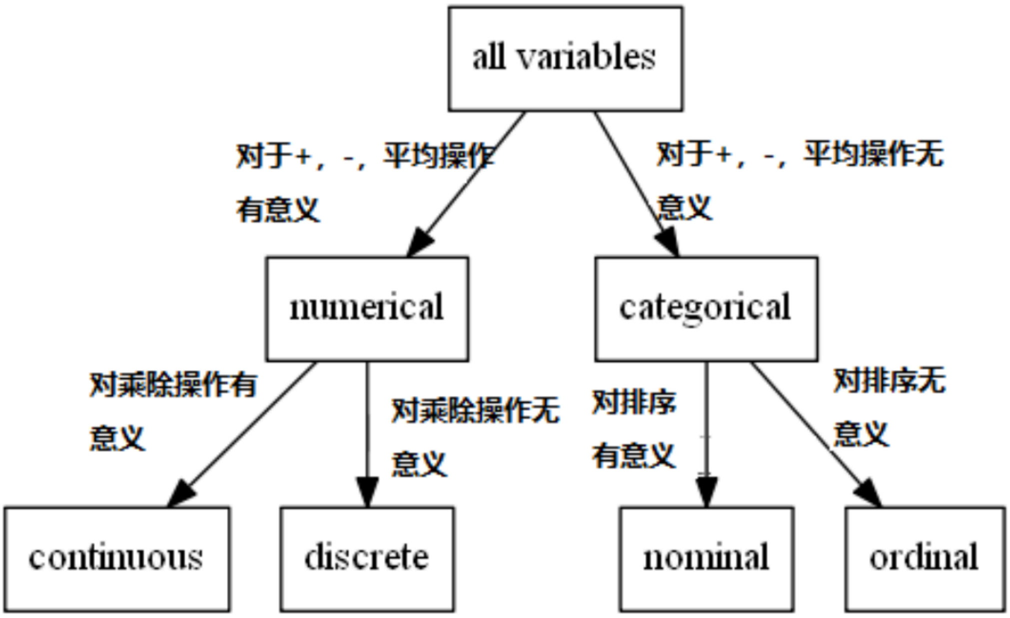
n: how many observations



Multivariate: variables > 2

Bivariate: variable = 2

Univariate:variables = 1



Type of variables

Type = Numerical data(数字数据) + Categorical data(种类数据)

Numerical data: measurement 一般是数字

= Discrete separated (year) + continuous separated (age)

离散变量(**Discrete data):** 不连续 在自然数中取值

连续性变量(**continuous data):** 一定区间内可以任意取值，两个节点之间可以任意取值

Categorical data: coded category 一般是字符串

= Ordinal data(orderable) + Nominal data(non-orderable)

有序分类变量：描述一个事物的等级

无序分类变量：仅作出分类

有序分类变量比较是有意义的 无序分类变量 则没有

无序分类变量(nominal) <有序分类变量(ordinal)< 离散型数值变量(discrete)< 连续型数值变量(continuous)

5. Data summary

Categorical data:

already summarised by category, we care about the most common category or any trend within category

Numerical data:

We focus on centre and spread

The main two types of summaries: numerical and graphical summary

Graphical summary -> sum data and then produce some plots

Categorical data we use bar plot or line plot

Discrete data we use frequency table

Continuous data we use tables or histogram

1. Use equal bins -> regular histogram
2. Use unequal bins -> probability histogram

Bin	Frequency	Relative Frequency	Height
[-10,18)	31	$31/442 = 0.07$	0.0025
[18,25)	72	$72/442 = 0.16$	0.0232
[25,70)	259	$259/442 = 0.59$	0.0130
[70,100)	80	$80/442 = 0.18$	0.0060
Total	442	1	

where:
Relative Frequency = Frequency/442
Height = Relative Frequency/Bin length
Eg For bin [-10,18): height = 0.07/28 =3.6.

6.Basic notations

描述一个数据集合:

Data = {Xi} where I=1,2,3,4……….n

Or

Data = {X1, X2, X3, ……….. Xn}

描述一个有序数据集合(asending)

Data = {X(i)} where I=1,2,3,4……….n

Or

Data = {X(1), X(2), X(3), ……….. X)n)}

一个集合的和

- The sum of the data is $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots x_n$.

7. Summary centre(or location)

Mean(平均数):

均值不是一个稳健的衡量工具，因为它会被异常值大大的影响到。

Outlier 会影响mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median(中位数):

中间值非常适合偏态分布，因为它源自集中趋势（central tendency），因此它是更稳健和明智的(robust)

不会收到outlier影响

- If n is odd, the unique median is the middle value:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

- If n is even, the median is the average of the 2 middle values (by convention):

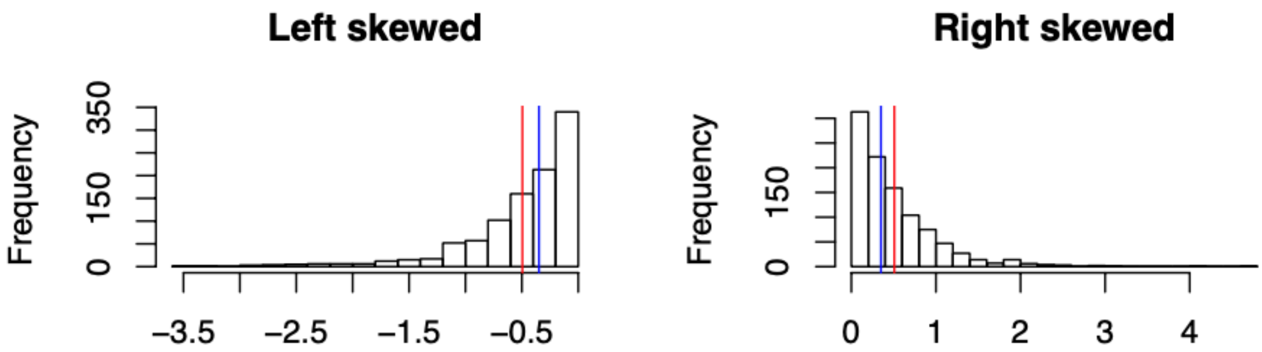
$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

需要注意的是 中位数找位置 而不是具体数值

在symmetric system中 mean=median

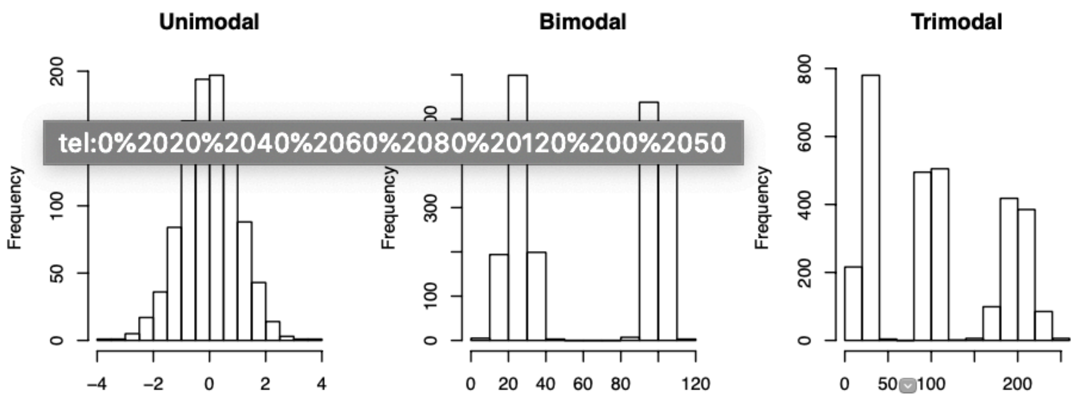
left skew分布中 median小于 mean

right skew分布中 median大于 mean



对于outlier比较少的数据我们可以使用mean，mean的统计性很强 因为我们是拿具体的数据计算的，得到的也是可计算的数据

median的话不受 outlier影响，但是我们得到的是一个位置，统计性不如mean



Unimodal multimodal trimodal bimodal 分布

8. Variance(方差)

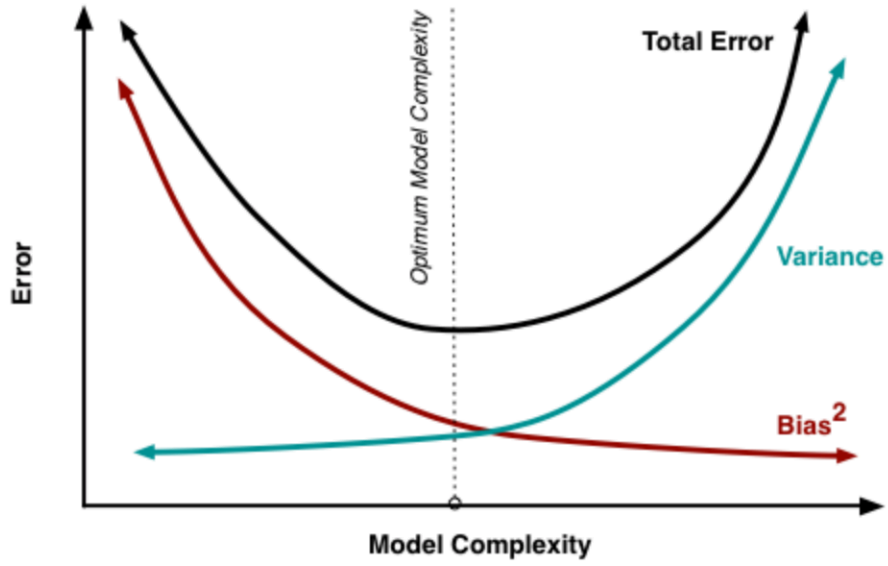
例子：Spread in data :{-1,0,1} / {-100,0,100}

$$
\begin{aligned}
Var(x) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \\
&= \frac{1}{n-1}\sum_{i=1}^{n}x_i^2 - n\bar{x}^2.
\end{aligned}
$$

一个有趣的问题：Variance: Var(x)-> ^2 >0 !^2 ==0

方差这一概念的目的是为了表示数据集中数据点的离散程度

Variance 和mean的scale不一样 所以我们还需要 标准差



Standard deviation：是各数据偏离平均数的平均距离

SD 就是 Variance开根号

为什么使用标准差：更好理解，方便后续运算

假如学习算法训练不足时，此时学习器的拟合能力不够强，此时数据的扰动不会对结果产生很大的影响（可以想象成由于训练的程度不够，此时学习器指学习到了一些所有的数据都有的一些特征），这个时候偏差主导了算法的泛化能力。随着训练的进行，学习器的拟合能力逐渐增强，偏差逐渐减小，但此时不同通过数据学习得到的学习器就可能会有较大的偏差，即此时的方差会主导模型的泛化能力。若学习进一步进行，学习器就可能学到数据集所独有的特征，而这些特征对于其它的数据是不适用的，这个时候就可能会发生过拟合。

原文链接：https://blog.csdn.net/wuzqChom/article/details/75091612

Quartiles：

{1,2,4,6,7,8} -> median = 5 = Q2

{1,2,4} + {6,7,8} -> Q1 = 2 Q3 = 7

So: quartiles = {Q1,Q2,Q3} = {2,5,7}

IQR is robust and perform well on many outlier data or skewed data.

SD is not robust and will be impacted by the **outliers**

(Fivenum(x)) -> (x(1), Q1, Q2, Q3, x(n))

Interquartile Range (IQR)：

Interquartile Range = $Q_3 - Q_1$

We couple ($\tilde{x}$,IQR) as a summary of centre and spread.

Deal with Outliers

1. IQR method
2. 3-σ method

3*sd(data) is a standard used to identify outliers, example:

heights1[abs(heights1-mean(heights1))>3*sd(heights1)]

就是说具体的一个值和mean的距离超过了三倍方差 就被视为outlier

9.Boxplot

Boxplot -> use to compare datasets and identify outliers

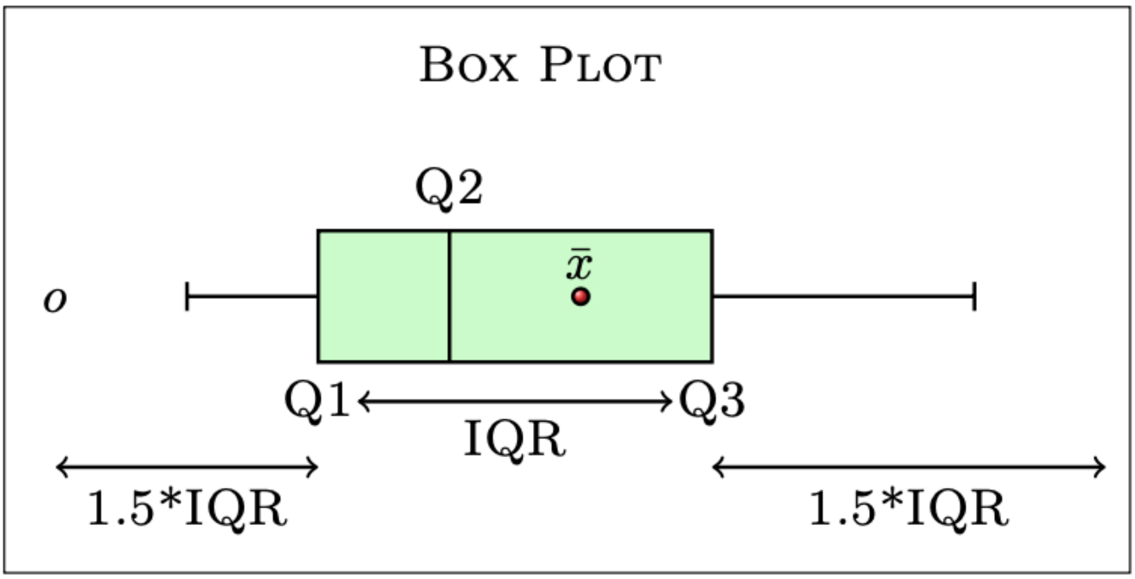
Lower Threshold (LT) = $Q_1 - 1.5$IQR;

Upper Threshold (UT) = $Q_3 + 1.5$IQR;

Inter quartile range is (IQR)=$Q_3 - Q_1$.

所以 boxplot的 底线就是LT = Q1-1.5*IQR 封顶是RT = Q3+1.5*IQR

IQR就是Q3-Q1的绝对值



Boxplot原理:

算出 Q1 Q2 Q3 和 IQR 然后画出盒子

算出LT RT 确定 margin 和底线

Any points outside the thresholds are outliers, designated by circles.

WEEK 2: Probability

概率论核心

什么是概率论: 现实世界中的现象分为两大类：分为确定性的和随机性现象；而概率论研究的是在随机性的现象中的规律的预测和决策。概率实际上使用少量样本(当然，也不能太少)随机映射整体的情况，从而以最小的成本的预测整体的走向和行为

概率论的定义： The probability of an event is a measure of the likelihood of that event occurring. 就是一个事件有多可能发生.

概念

Mutually exclusive(互相排斥的事件):

outcomes cannot occur at the same time.

一个样本空间可以两个或更多结果完全不同的事件，多个事件共同填满完整的样本空间

Collectively Exhaustive(完全穷尽事件):

One outcome in sample space must occur.

The set of events covers the entire sample space.

样本空间中必须出现一个结果. 事件集覆盖了整个样本空间, 多个事件共同填满完整的样本空间

Simple event(简单事件):

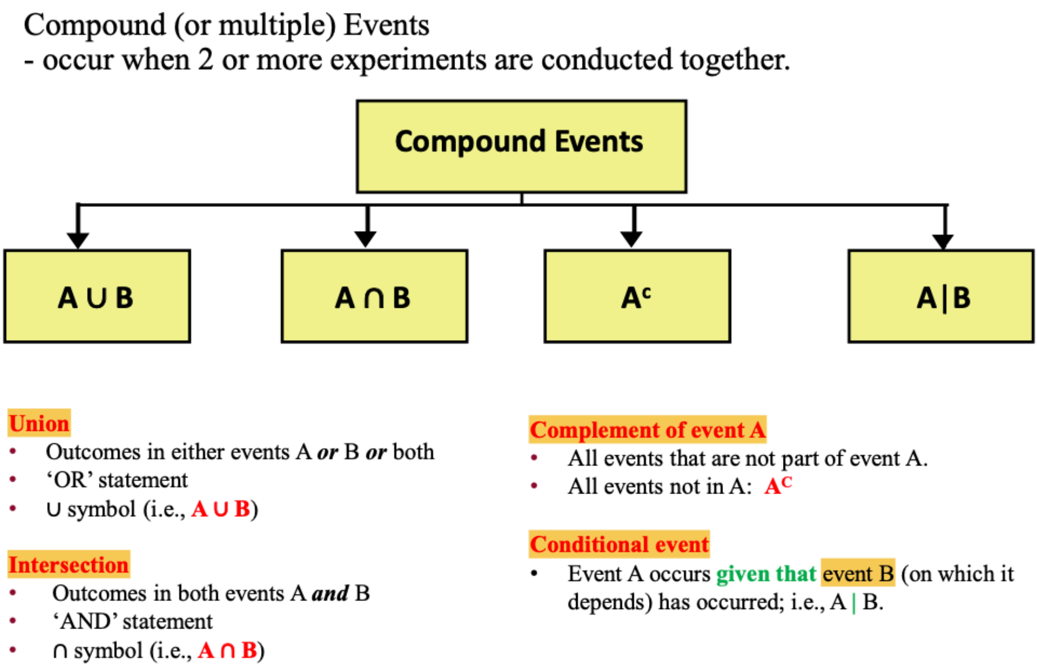
outcome with one characteristic 具有一个特征的结果

Compound event(复合事件):

collection of outcomes or simple events 具有多种特征的结果

joint event is a special case: two events occurring simultaneously

联合事件是特殊的 两种事件同时发生



Conditional event: A|B 也就是说 - B是条件，在B发生的时候，发生 A 的概率有多大，也可以通过一个等式理解 P(A|B) = P(A^B)/P(B). 以上这些表达式的形成 大部分都基于复合事件

Visualising Event(概率事件可视化):

有三种方法 - Contingency Tables 联列表 Decision Trees 决策树

还有 Venn diagram

一些统计学基本运算

DeMorgan's Law:

$P(A^c \cup B^c)=P(A\cap B)^c = 1-P(A\cap B)$ $P(A^c \cap B^c)=P(A\cup B)^c = 1-P(A\cup B)$

Complement rule:

$P(A) = 1 - P(A^c)$

Addition rule – union of events:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multiplication rule – probability of intersection events:
$P(A \cap B) = P(A|B)*P(B)$

Joint probability = conditional probability * P(condition)

$P(A \cap B) = P(B|A)*P(A)$

需要注意的一个公式变换

$P(B|A) = P(A \cap B)/P(A)$

如何甄别Mutual exclusive event 和 Independent event

Let A and B be 2 probability event

Mutual exclusive event:

If $P(A \cap B) = 0$ OR If $P(A \cup B) = P(A) + P(B)$

Independent event:

$P(A \cap B) = P(A)*P(B)$ OR $P(A|B) = P(A)$ OR $P(B|A) = P(B)$

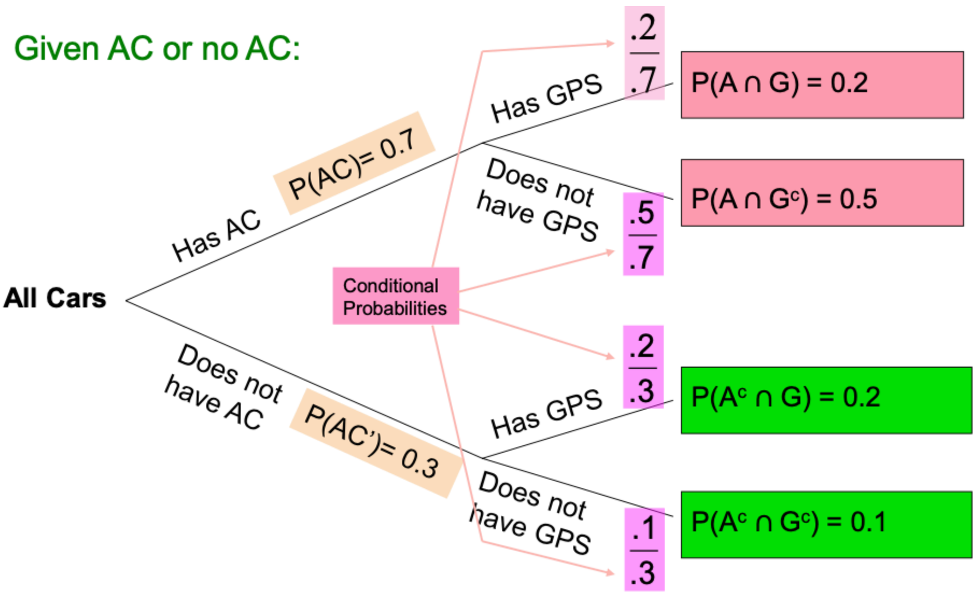
conditional probability = unconditional probability

所以:

$P(A|B) = P(A)$ OR $P(B|A) = P(B)$

两种图表的使用方法:

决策树:



联列表:

Event	Event B$_1$	Event B$_2$	Total
A$_1$	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	$P(A_1)$
A$_2$	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	$P(A_2)$
Total	$P(B_1)$	$P(B_2)$	1

Joint Probabilities

Marginal (Simple) Probabilities

大同小异基本上一个意思 拆开算概率 综合结果为1

Bayes' Theorem贝叶斯理论

Bayes' Theorem is used to revise previously calculated probabilities based on new information.

贝叶斯定理用于根据新的信息修正先前计算的概率，比如说患病问题. 尤其是某一事件已经发生 假设成立的概率

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)}$$

where

B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$

例子: 患病问题 - 假设被诊断出患有非常罕见的疾病，这种病患的比例仅是人口的0.1％. 参加的检查这种疾病的检测能正确地找出99%的患者，将健康的人错误分类的几率只有1%. 设患病概率为P(E) 设检测出来为阳性为P(H)

⚠️The Law of Total Probability:

我们想知道P(H｜E) 即在知道确诊的情况下，假设成立的概率有多大，那我们有如下等式:

P(H|E) = P(E|H)*P(H)/P(E)

而根据 情况(Decision Tree) 我们知道:

P(E) = P(E)*P(H|E)+P(~H)*P(E|~H)

所以 P(H|E) = P(E|H)*P(H) / P(E)*P(H|E)+P(~H)*P(E|~H)

随机变量

A random variable is a variable whose numerical value is determined by the outcome of a random trial.

随机变量是由随机试验结果决定其数值的变量。

随机变量有两种:

Discrete 就是说 整数形式的数值，例如(number of car)

continuous 就是说 可连续数值 一般可以为小数，例如(年薪，体重)

Distribution(分布)

一些大致分布 可以根据背景区分

Population distribution (总体分布): 感兴趣的总体的一个分部也就是说所有样本的集合

Sample distribution (样本分布): 从总体中提取的样本的一个分部，因为我们很少知道总体的分布

Sampling distribution (抽样分布): 我们可能感兴趣从样本中估计均值。我们可以进行另一项研究取另一组样本计算样本均值。重复这样做，就能得到样本均值的抽样分布。

可以根据性质分布: **离散分布，连续分布** 见上一个知识点

Counting Techniques

Multiplication Rules:

If there are m ways of doing one thing and n ways of doing another thing, there are **m*n** ways of doing both.

The rule can be extended to more than 2 events. For 3 events, p, q, and r, the total number of arrangements = **p*q*r**

但是这只适用于独立情况

Permutation:

Repetition is allowed

The number of permutations of n objects taken r at a time when repetition is allowed is ${}_nP_r = n^r$

Repetition is not allowed

The number of permutations of n objects taken r at a time is ${}_nP_r = \frac{n!}{(n-r)!}$ where

${}_nP_r$ is read as "n permute r."

P is the number of permutations (or ways) the objects can be arranged.

n is the total number of objects.

r is the number of objects to be used at one time.

n! = n*(n – 1)*(n – 2)*(n – 3)*4*3*2*1

Combination:

Repetition is allowed

A **combination** of a set of objects is a subset of the objects disregarding their order; i.e., the **order is not important**. {a, b} is the **same** as {b, a}.

There are two types of combination as follows:

- **Repetition is not allowed.**

  The number of combinations of n distinct taken r at a time is $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ where

  $\binom{n}{r}$ is read as "n choose r".

  n is the total number of objects.

  r is the number of objects to be used at one time.

  $n!$ = n*(n – 1)*(n – 2)*(n – 3)*4*3*2*1

If a set has n elements, a total of 2^n subsets can be formed from those elements; i.e., $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$

- **Repetition is allowed.**

  The number of combinations of n distinct taken r at a time when repetition is allowed is

  $\binom{n+r-1}{r} = \binom{n+r-1}{n-1} = \frac{(n+r-1)!}{r!(n-1)!}$

Two types of Probability

Data-based Probability(基于数据的概率): 实验的频率处于试验的总次数，例子: 抛硬币一万次，计算出正面朝上的频率。

Model-based Probability(基于模型的概率): 这是一个数学结构，为每个可能的事件分配一个数字，例子: 我们提出一个模型，得到正面的概率是0.5。

Data-based Probability

The probability of an event is the proportion of times that event would occur in a large number of repeated experiments

一个事件发生的概率是该事件在大量重复实验(模拟)中发生的次数的比例。也就是说硬币朝上的概率=P(H)=大量扔硬币，head的概率

R相关练习代码见 tut2.R

Table函数主要有两个用途:1, 总结出频率 2, 实现混淆矩阵

Sample函数主要起到一个抽样作用

Week 3: Random Variable(随机变量)

主要探讨三个问题:

1. 离散随机变量(PMF) - 映射discrete data - categorical
2. 连续随机变量(PDF) - 映射continuous data - numerical
3. 组合随机变量

简单的例子

例如 x = {-2, 0, 10, 14}	如果 x 属于 1<x<14
x is discrete random variable	x is continuous random variable

PDF：概率密度函数（probability density function）是一个描述这个随机变量的输出值，在某个确定的取值点附近的可能性的函数。（概率）是个方程

PMF：概率质量函数（probability mass function) 是离散随机变量在各特定取值上的概率。（趋势）**表示成{x, P(X=x)}** 是个方程

CDF：累积分布函数 (cumulative distribution function)，又叫分布函数，是概率密度函数的积分，能完整描述一个实随机变量X的概率分布。（概率）**表示成{x,P(X<=x)}** 是个面积

Discrete random variable 使用 probability math function(pmf)

Continuous random variable 使用 probability density function(pdf)

PDF 是一种趋势对连续的值(或一个区间的值)积分后才是概率

PDF积分之后就是CDF 所以说 同事探讨CDF和PMF才有意义

而 PMF是该值的概率

Discrete Distribution (离散随机变量)

离散随机变量的Mean 和 variance:

Definition (Mean or Expectation)

The mean of X is

$$\mu = E(X) = \sum_{\text{all } x} x P(X = x)$$

Definition (Variance)

The variance of X is

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

Definition (Expectation of a Function)

The expectation of $g(X)$ is

$$E(g(X)) = \sum_{\text{all } x} g(x) P(X = x)$$

For example: $E(X^2) = \sum_{\text{all } x} x^2 P(X = x)$.

Discrete random variable - we sum

Continuous random variable - we integrate

CDF的概念和variance的公式适用于两者(X <= x)

在Binomial中 CDF = P(X <= x) = sum(P(X=0), P(x=1), ... P(X=x))

(Probability Distribution Function) only works for 离散随机变量

实质上 mean就是 sum variance就是sum的差(小练习见讲义)

If c is a constant P(X=c) = 0 那就是 X不能等于 c

性质:

countable number of possible values

可能性的总和是1

离散随机变量我们有两种分布:

1. 二项分布 Binomial Distribution（有放回)

练习week2 最后一单元的讲义

2. 超几何分布 Hypergeometric Distribution（无放回)

练习week2 最后一单元的讲义

Binomial distribution （二项式分布）

基础性质:

1. Characteristic we have n identical trials (有n个实验(n个值))
2. Trials have 2 outcomes success or failure (只有两个可能的值)
3. Probability of success (p) stays same from one trail to another

   Probability of fail (q) = 1- p stays same from one trail to another
4. All trials are independent. Since p+q = 1

因为这玩意不是1就是0 而且有放回

⚠️: ~ = ‘is distributed as’

一般地 二项分布可以表示为x~Bin(n, p)

parameters: n is trails, p is success probability

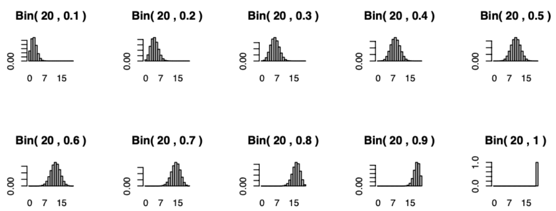
The PMF of x is

If X = the number of successes in n trials, then $X \sim Bin(n, p)$ with

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x = 0, 1, 2, \ldots, n.$$

P(X=0) +P(X=1) +……+P(X=n) = 1 ⚠️ **0! =1**

$X \sim Bin(n = 20, p)$, for different $p = 0.1, 0.2, \ldots, 1$.



if p = 0.5 hist is symmetrical 也就是说 p=0.5的时候: 单峰

二项分布的均值(mean of Binomial distribution)

E(X) = n*p

二项分布的方差(Variance of Binomial distribution)

Variance(X) = n*p*(1-p) = n*p*q

⚠️需要注意的是 这两个公式只适用于二项分布

R 代码

```
#dbinom(x, n, p) - calculate PMF P(X=x)
dbinom(0, 10, 0.2)
dbinom(1, 9, 0.2)
#CDF  - calculate P(X<=x)
pbinom(1, 10, 0.2)
```

记不住怎么办

Discrete we use pxxxx and dxxxxx. For continuous random variable we don’t use dxxxx. If continuous we always use pxxxxx

(小练习见 Exercise 2)

对于二项分布 我们需要知道 x， n， p (x 是研究的个数，n是总数， p 是可能性)

Sampling with replacement => events are **independent**

=> **binomial distribution (有放回)**

Sampling without replacement => events are **not independent**

=> **hypergeometric distribution (无放回)**

Hypergeometric distribution （超几何分布）

基本性质:

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$$

where *x* that satisfies the inequalities:

- $x \le N_1$
- $x \le n$
- $n - x \le$ n

三个决定性变量 N N1 n (x是你选几个 自己定)

R 代码

方法1: choose (10,2)* choose(6,1)/choose(16/3)

方法2: dhyper(2,10,6,3) -> dhyper(x,N1,N2,n)

三个里边两个符合要求 然后诶分别有N1个 N2个

(小练习见 Exercise 3)

Continuous Distribution (连续随机变量)

连续随机变量和离散随机变量对比

	Discrete	Continuous
Values	Countable	Infinite
Plot	Histogram $P(X = x)$ probability distribution function	Smooth curve $f(x)$ probability density function (pdf)
$P(X = x)$	$0 \le P(X = x) \le 1 \ \ \forall x$	$P(X = x) = 0 \ \ \forall x$
Sum of Probabilities	$\sum_x P(X = x) = 1$ Area of histogram	$\int_x f(x)dx = 1$ Area under density
$F(x) = P(X \le x)$	$\sum_{y=min(x)}^{x} P(X = y)$	$\int_{-\infty}^{x} f(y)dy$

连续随机变量的性质:

1. there is an infinite number of possible values
2. may be within a fixed interval
3. probability density function (pdf), must be 1

一般地: continuous P(a<X <b)=P(a≤X ≤b)

⚠️ 因为 P (X=a) = P(X=b) = 0 更直观地讲: 一条线的面积为0

Normal distribution (正态分布)

The Normal distribution models a symmetric, bell-shaped variable with 2 parameters mean μ and variance σ^2 and points of inflection at $\mu \pm \sigma$. We say the variable $X \sim N(\mu, \sigma^2)$.

The probability density function (pdf) is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in (-\infty, \infty)$$

The cumulative distribution function (CDF) is

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$$

计算可能性的两种方法 (已知 mean是161.8， variance是6^2)

方法1： 带入pdf公式 u=mean o^2 is variance

R 代码:

```
f<- function(x) {dnorm(x,161.8,6)}

integrate(f,189,200)
```

方法2: R 代码 `pnorm(189, 161.8, 6) # x, mean, var`

我们得到的是P<=189 需要拿1 - P(x<=189)

如果求 区间的可能性 比如说(170<P<175) 我们需要转化一下

P(170<x<175) = P(170<=x<=175) = P(x<=170) - P(x<=175)

带入R: pnorm(175, 161.8, 6) - pnorm(170, 161.8, 6)

如果是一个标准正态分布(variance = 1, mean = 0)

在R代码中我们就不需要指定var和mean

正态分布标准化:

Standardise a normal random variable (标准化正态分布的方法):

对x进行处理

Definition (Standardardising a Normal)

If $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0,1)$, then
$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

x = x-u/standard deviation(o)

Z has a standard normal distribution (所得的Z就是一个SND)

Normal Percentiles:

Given X ~ N(161.8,6^2), what is the 90% percentile for heights of Australian women.

We need to find x such that P (X ≤ x) = 0.9.

R 代码:

Pnorm(0.9, 161.8, 6)

Linear Function of a Random Variable:
随机变量的线性函数

Definition (Linear Function of Random Variable)

Given a random variable X, then $Y = a + bX$ has moments

$$\boxed{E(Y) = a + bE(X)}$$

and

$$\boxed{Var(Y) = b^2 Var(X)}$$

for all 2 constants a and b.

Special Case: If $X \sim N(\mu, \sigma^2)$, then $Y \sim N(a + b\mu, b^2\sigma^2)$.

Liner function of normal is a normal

小练习: Suppose the weight of an Australian women in kg, W ~ N (71.1, 12^2). Find the distribution of the weight of an Australian women in pounds, given 1kg = 1 pound/2.2046.

答:

Let P = Weight of an Australian women in pounds = 2.2406W .

This is a linear function where a = 0 and b = 2.2406.

E(P) = 0 + 2.2406E(W) = 2.2406 × 71.1 = 159.3067

V ar(P) = 2.2406^2V ar(W) = $2.2406^2 \times 12^2$ = 722.9215 So P ~ N(159.3067,26.8872^2)

随机变量的独立

如果说 有两个随机变量x和y 那么 如果他们满足:

P(x < X, y<Y) = P(x < X) * P(y < Y)

那么我们可以说 他们是相对独立的(the joint CDF splits into the 2 individual CDFs.)

这个理论由 Cov(X, Y) = E(XY) − E(X)E(Y) = 0 而来

随机变量的和(Total of random variables)

有几个随机变量数据集 X1, X2, X3 ………., Xn。那么有 E(X1), E(X2), ………E(Xn) 那么他们的total就是 **sum(E(X1), E(X2)…E(Xn))**

他们的方差 也就是**sum(Var(X1), Var(X2) … Var(Xn))**

随之而来的 如果我们想知道他们的:

随机变量的样本和(sample total of random variables)

那么 一般的 他们样本和的mean就是:

1/n* sum(E(X1), E(X2)…E(Xn))

相同的 他们样本和的方差就是:

1/n^2*sum(Var(X1), Var(X2) … Var(Xn))

以上的这些公式针对的是所有的随机变量，也就是说任何随机变量都适用

正态随机变量的和与样本和(Total and Sample Mean of Normal RVs)

在正态分布随机变量集中: $X_i \sim N(\mu_i,\sigma_i^2)$

Given a sequence of random variables $X_i \sim N(\mu_i, \sigma_i^2)$ (for $i = 1, 2 \ldots, n$)
then

$$\boxed{T = \sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)}$$

and

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\frac{1}{n}\sum_{i=1}^{n} \mu_i, \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2)}$$

Summary: for constants a_i,

$$\boxed{T = \sum_{i=1}^{n} a_i X_i \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)}$$

(通用结论)

(小练习见Exercise6)

⚠️ 一个需要注意的点是: 在计算概率相关问题时如果遇到正态分布问题，先标准化正态分布，然后带入系统计算。

Sum of iid normal (iid 正态分布问题)

Definition (Total and Sample Mean of iid Normal RVs)

Given a sequence of iid random variables $X_i \sim N(\mu, \sigma^2)$ (for $i = 1, 2 \ldots, n$)
then

$$\boxed{T = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)}$$

and

$$\boxed{\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})}$$

也就是说 X1，X2，……… Xn 有相同的range 和相同的数据模式 所以我们直接是用乘法

(小练习见Exercise7)

What is the Distribution of the Sample Mean for Any Population - Central Limit Theorem

那么有没有一种sample mean的分布可以适用于 任何一种数据集

这就要提到 <u>中心极限定理 - **Central limit theorem (CLT)**</u>

If $X_i \sim (\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$ then

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

⚠️中心极限定理是统计学中最重要的概念之一

CLT 的要求和前提是:

数据量足够大 - n足够大

我们的方差必须小于无限 - 方差不能太大