医学统计学笔记

统计工作的基本步骤：

- 设计：指统计设计，应包括对资料收集、整理、分析全过程的设想和安排。
- 收集资料：采取措施取得准确可靠的原始数据。
- 整理资料：将原始的数据进行净化、系统化和条理化，以便为下一步计算和分析打好基础的过程。
- 分析资料：又称统计分析，包括统计描述和统计推断
    - 统计描述：用恰当的统计指标，通常称为统计量（statistic），选用合适的统计图表，对资料的数量特征及其分布规律进行测定和描述
    - 统计推断：指如何在一定的可信程度下由样本信息推断总体特征。包括参数估计和假设检验
        - 参数估计：如何用样本统计指标（统计量）推断总体相应指标（参数，parameter）
        - 假设检验：如何由样本间的差异推断总体之间是否存在差异

统计学的发展简史

分为三个阶段：

- 古典统计学
- 近代统计学
- 现代统计秀儿

17世纪中叶，Pascal和Fermat创始了概率论，法国的Laplace和德国的Gauss相继发现了正态分布的过程，并用于行星轨迹的预测。在医学统计领域，法国的医师P.C.A. Louis提出医学观察中的“混杂”问题和疗效比较的“数量化”方法，被尊称为“临床统计之父”；1837年英国成立了出生、死亡登记中心，为描述流行病学发展提供了广阔的舞台，1840年发过数学家S.D.poisson的学生J.Gavarret出版了世界第一部医学统计学，1834年英国统计学家成立了伦敦统计学会。1885年成立了全球性的统计学术组织—国际统计学会 K.Pearson 1893年提出标准差，1900年提出了最早的假设检验方法—- χ^2检验，创立最权威的神物统计杂志“Biometrika“,创办了世界上第一所统计学校。 1908年W.S.Gosset以‘Student'的笔名在Biometrika上发表了经验分布（t分布）后经R.A.Fisher等人完善，形成了当今广为使用的假设检验方法—t检验，开创了小样本统计的新纪元。 R.A.Fisher创立了用于随机化实验设计和方差分析的理论和方法，发现了许多小样本统计量的精确分布，对小样本统计方法做出来重要贡献，被誉为现代统计学的奠基人之一。 在我国生物统计方法在医学界的传播与运用始于20世纪初，1948年，郭祖超编著《医学与生物统计方法》为我国第一部医学统计方法教材

计量资料的统计描述

频数分布：通常是针对样本而言， - 对于连续变量（continuous variable），频数分布为n个变量值在各变量区间内的变量值个数的分配 - 对于离散变量（discrete variable），频数分布为n个变量值在各（或各几个）变量值处的变量个数的分配 频数分布表编制： 例2.1某医院用随机抽样的方法检测了138名正常成年鱼子的红细胞数(×10^{12}/L),其结果如下，试编制频数分布表：

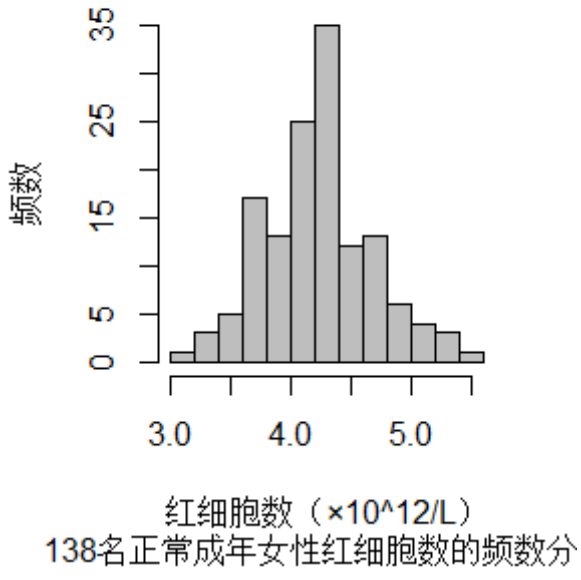
3.96,4.23,4.42,3.59,5.12,4.02,4.32,3.72,4.76,4.16,4.61,4.26,
3.77,4.20,4.36,3.07,4.89,3.97,4.28,3.64,4.66,4.04,4.55,4.25,
4.63,3.91,4.41,3.52,5.03,4.01,4.30,4.19,4.75,4.14,4.57,4.26,
4.56,3.79,3.89,4.21,4.95,3.98,4.29,3.67,4.69,4.12,4.56,4.26,
4.66,4.28,3.83,4.20,5.24,4.02,4.33,3.76,4.81,4.17,3.96,3.27,
4.61,4.26,3.96,4.23,3.76,4.01,4.29,3.67,3.39,4.12,4.27,3.61,
4.98,4.24,3.93,4.20,3.71,4.03,4.34,4.36,3.62,4.18,4.26,4.36,
5.28,4.21,4.42,4.36,3.66,4.02,4.31,4.83,3.59,3.97,3.96,4.49,
5.11,4.20,4.36,4.54,3.72,3.97,4.28,4.76,3.21,4.04,4.56,4.25,
4.92,4.23,4.47,3.60,5.23,4.02,4.32,4.68,4.76,3.69,4.61,4.26,
3.89,4.21,4.36,3.42,5.01,4.01,4.29,3.68,4.71,4.13,4.57,4.26, 4.03,5.46,4.16,3.64,4.16,3.76

1.求极差： RBCC <- c(~) range(RBCC)

2.确定组段数和组距 rbcc<- cut(RBC,c(seq(3.07,5.47,0.2)))

3.根据组距统计频数 table(rbcc)

4.画频数分布图 hist(RBC,main = '',sub = '138名正常成年女性红细胞数的频数分布',xlab= '红细胞数（×10^12/L）',ylab = '频数',col = 'grey')



5.频数表和频数分布图的用途

- 描述频数分布的类型 频数分布的类型可分为对称分布和偏态分布两种
  - 对称分布：各组段频数以频数最多组段为中心左右两侧大体对称
  - 偏态分布：非对称分布，哪侧拖尾就为哪侧偏态分布

- 描述频数分布的特征
- 便于发现一些特大或特小的离群值（outlier）
- 便于进一步做统计分析和处理

集中趋势的描述

1.算术均数

- 直接计算法

$$X=\frac{X_1+X_2+X_3+\ldots+X_n}{n}=\frac{\Sigma X}{n}$$

- 频数表法

$$X=\frac{f_1X_1+f_2X_2+f_3X_3+\ldots+f_kX_k}{f_1+f_2+f_3+\ldots f_k}=\frac{\Sigma fX}{\Sigma f}$$

2.几何均数 geometric mean：可用于反映一组经对数转换后呈堆成分布的变量值在数量上的平均水平，在医学研究中常适用于免疫学的指标。其计算公式为：

$$G=\sqrt[n]{X_1X_2\ldots X_n}\text{或}G=lg^{-1}(\frac{\Sigma lgX}{n})$$

例2-4 某地5例微丝蚴血症患者治疗7年后用间接荧光抗体试验测得其抗体滴度倒数分别为10,2,40,40,160,求几何均数。

$$G=\sqrt[5]{10\times20\times40\times40\times160}=34.8$$

或

$$G=lg^{-1}(\frac{lg10+lg20+lg40+lg40+lg160}{5})=34.8$$

故5份血清抗体效价的平均滴度为1:34.8 对于频数表资料，几何均数的计算公式为：

$$G=lg^{-1}(\frac{\Sigma flgX}{\Sigma f})$$

3.中位数与百分位数

- 中位数 中位数（median）是将n个变量值从小到大排列，位置居于中间的那个数。n为奇取中间的变量值，n为偶数时，取中间两个变量值的均数。

n为奇数：

$$M=X_{\frac{n+1}{2}}$$

n为偶数：

$$M=\frac{1}{2}(X_{\frac{n}{2}}+X_{(\frac{n}{2}+1)})$$

例： 7名患某病的潜伏期分别为2,3,4,5,6,9,16天，求其中位数。 expn <- c(2,3,4,5,6,9,16) median(expn) 例： 8名患者食物中毒的潜伏期分别为1,2,2,3,5,8,15,24小时，求其中位数 expn2 <- c(1,2,2,3,5,8,15,24) median(expn2)

- 百分位数 百分位数（percentile）是一种位置指标，用P_x来表示，读作第X百分位数。一个百分位数P_x将全部变量分为两部分，在P_x处若无相同的变量值，则在不包含P_x的全部变量值中有X%的变量值小于它（100-X）%变量值大于它。故百分位数是一个界值，其重要的用途是却行医学参考值范围（reference range）。中位数实际上市第50百分位数。

quantile（x,probs = [0,1])

离散趋势的描述

1.极差 range：最大最小值之差 range(x)

2.四分位数间距

$P_{75} - P_{25}$

x

quantile(x,0.75)-quantile(x,0.25)

3.方差与标准差

- 方差 variance也称均方差（mean square deviation），反映一组数据的平均离散水平。

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

- 标准差 standard deviation是方差的平方根，其单位与原变量值的单位相同。总体的标准差用σ表示，计算公式为：

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

一般情况下，总体均数μ未知，需用样本均数$\overline{X}$估计。数理统计证明：若用样本个数n代替N，计算出的样本方差对μ^2的估计偏小，需将n用n-1代替。样本方差标记为S^2,其标准差S的计算公式为：

$$S = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}}$$

简化后可以表示为：

$$S = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}}$$

例2-11 试计算下面三组同龄男孩的身高（CM）均数和极差 甲组：90 95 100 105 110 乙组：96 98 100 102 104 丙组：96 99 100 101 104

a <- c(90, 95, 100, 10, 110) b <- c(96 ,98, 100, 102 ,104) c <- c(96, 99, 100, 101, 104) d <- data.frame(a,b,c) sapply(d,range) sapply(d,mean) sapply(d,sd)

求例2.1红细胞的标准差为 sd(RBC)

- 变异系数

（coefficient of variation）记为CV，多用于观察指标单位不同时，如身高与体重的变异程度的比较；或均数相差较大时，如儿童身高与成人身高的变异程度的比较。其计算公式为：

$$CV = \frac{S}{\overline{X}} \times 100\%$$

正态分布

1.正态分布的概念和特征

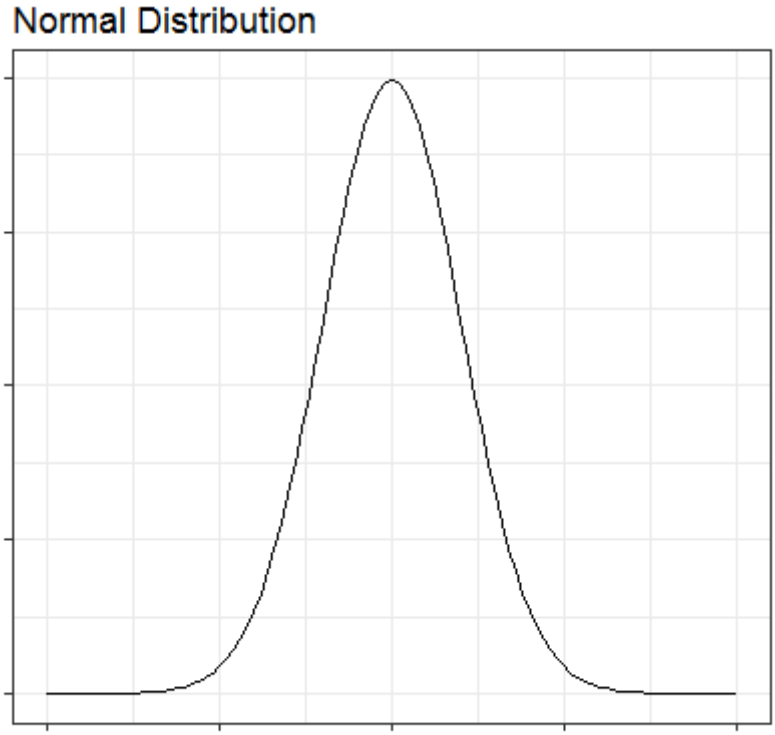
1. 正态分布曲线的数学函数表达式 如果随机变量X的分布服从概率密度函数

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}, -\infty < X < +\infty$$

则称X服从正态分布，记作X~N（μ, σ^2),μ为X的总体均数，σ^2为总体方差。

2. 正态分布的特征

- 在直角坐标的横轴上方呈钟形曲线，两端与X轴永不相交，且以X=μ为对称轴，左右完全对



称。

- 在X=μ 处， $f(x)$取最大值，其值为$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$;X越远离$\mu$， $f(x)$值越小。
- 正态分布有两个参数，即位置参数μ和形态参数σ
- 正态曲线下的面积分布有一定的规律。
  - X轴与正态曲线所夹的面积恒等于1
  - 区间$\mu \pm \sigma$的面积为68.27%，区间$\mu \pm 1.96\sigma$的面积为95%，区间$\mu \pm 2.58\sigma$的面积为99.00%

2.标准的正态分布

正态分布是一个分布族，对应于不同的参数μ和σ会产生不同位置不同形状的正态分布。

为了应用方便，令

$$u = \frac{X - \mu}{\sigma}$$

则有

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}, -\infty < u < +\infty$$

即将X~N(μ, σ^2)的正态分布转化为μ $N(0, 1^2)$的标准正态分布

若用R求例2-1中的在4.0（ $\times 10^{12}/L$）以下者占正常成年女性总人数（总体）的百分比；4-5（ $\times 10^{12}/L$）之间者占总体的百分比；在5（ $\times 10^{12}/L$）以上者占总体的百分比

X <- mean(RBC) Xs <- sd(RBC) p4 <- pnorm(4,X,Xs) p5 <- pnorm(5,X,Xs) q1 <-p4 q2 <- p5-p4 q3 <- 1-p5;q1;q2;q3

医学参考值范围的制定

医学参考值指包括绝大多数正常人的人体形态、功能和代谢产物等各种生理及生化指标常数，也称正常值。 由于个体存在差异，生物医学数据，并不是常数，而是在一定范围内波动，故采用医学参考值范围作为判定正常还是异常的参考标准。

其采用单侧界值还是双侧界值应视实际情况而定。方法有正态分布法和百分位数法

- 正态分布法 在采用此法前一般要对资料进行正态性检验，且要求样本量足够大（n>100），其计算公式为： 双侧1-α参考值范围：$\overline{X} \pm u_{\alpha/2} S$ 单侧1-α参考值范围： $> \overline{X} u_{\alpha} S$或 $< \overline{X} + u_{\alpha} S$

用R可以直接使用 qnorm(0.025,X,Xs);qnorm(0.975,X,Xs)

- 百分位数法 偏态分布的资料通常采用百分位数法，所要求的样本含量比正态分布法要多（n>100），其计算公式为

双侧1-α参考值范围：$P_{100\alpha/2}$ $P_{100-100\alpha/2}$

单侧1-α参考值范围：$> P_{100\alpha}$或 $< P_{100-100\alpha}$

R中可使用quantile(x,probs = [0,1])函数