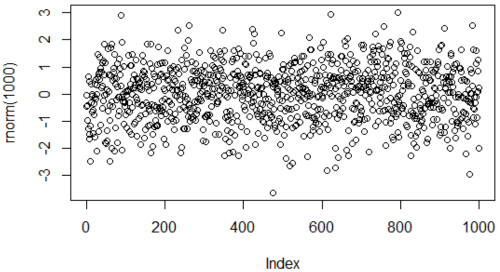
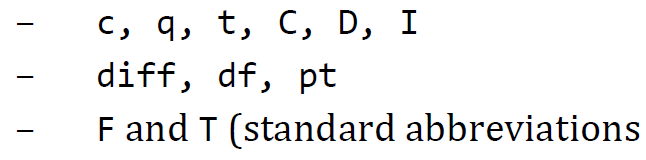
**STAT1301 STATISTICAL PACKAGES**

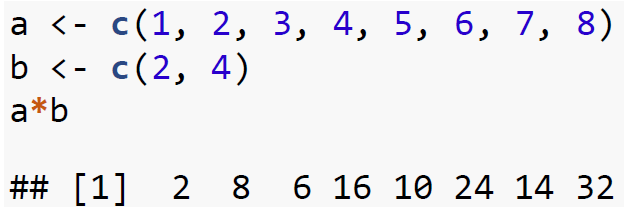
**Lec1 Intro to R**

一些R的基本用法

1. **plot**(**rnorm**(1000)) - 标绘1000个标准正态随机变量的观测值：
2. **sqrt**(2) – 求根号2
3. **factorial**(7) – 求7的阶乘
4. **rnorm**(20) – 产生20个随机数
5. x <- 6; x - r语言可以给变量赋值；<-是赋值运算符
6. y <- x**^**2; y - 用变量给变量赋值
7. z = x**+**y; z - =也可以作为赋值运算符
8. 一些规则：变量名不得以数字或dot为头，一些变量是不可以用的
9. 向量（data vector）是一组数字，可以将a vector of numbers 定义给向量变量vector variable

weight <- **c**(60, 72, 57, 90, 95, 72); weight

function c用来定义向量 - concatenate

1. 向量运算如同数字 height <- **c**(1.75, 1.80, 1.65, 1.90, 1.74, 1.91) ；BMI <- weight**/**height**^**2; BMI
2. 如果两个向量长度不一样，较短的将会被“recycled”

即output的长度由长的向量input决定

1. 用内置函数计算weight的标准差

x\_bar <- **sum**(weight)**/length**(weight); x\_bar

weight **-** x\_bar

**sqrt**(**sum**((weight**-**x\_bar)**^**2)**/**(**length**(weight)**-**1))

1. Object 对象：被赋予了变量；Expression 表达式：包含变量、运算符、函数
2. Vector支持以下数据类型：numeric，integer，complex；logical；character/string，factor

Vector有三种：numeric vectors；character vectors；logical vectors

1. Logical Vector：可以把TURE, FALSE, NA赋给向量

L <- **c**(T,F,T,T); L; BMI **>** 25

1. Character Vector: 给予字符，单双引号都无所谓

students <- **c**("Abby", "Becky", "Cathy"); students

1. R有categorical variables，它们被指定为factors。Factors的类别被称为level。

A **factor-type** vector contains a set of numeric codes with character-valued levels.

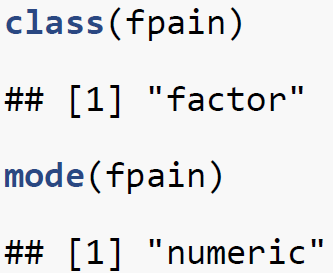
pain <- **c**(0, 3, 2, 2, 1); fpain <- **factor**(pain, levels=0**:**3); fpain

函数factor()用于将vector编成factor

函数as.factor() coerces its argument to a factor

1. 用levels()更方便 - **levels**(fpain) <- **c**("none","mild", "medium", "severe"); fpain

或者 - fpain2<-**factor**(pain,labels=**c**("factornone","mild", "medium", "severe"));fpain2;

这种方式是不可以 - **factor**(pain,levels=**c**("none","mild", "medium", "severe"))

1. class() – class是分配给对象的属性，函数根据class操作（不排斥

mode() – mode结果分为numeric, complex, charter and logical（相互排斥

1. 缺失值NA – x <- c(3,1,NA)

如何处理NA？ **mean**(x) ； **mean**(x,na.rm=TRUE)

NA is not a character value（需要加引号才能作为NA）

1. 向量的基本属性

length(x): vector或list的元素的数量

class(x): mode(x):

is.logical(x), is.numeric(x), is.character(x), is.integer(x), is.na(x) 判断是否是xx

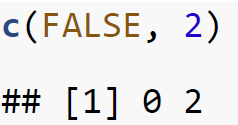
is.null(x): 确定对象是否为空。NULL表示长度为零的列表，通常由值未定义的表达式和函数返回。

1. 能够提供vector信息的函数 - **summary**(x)

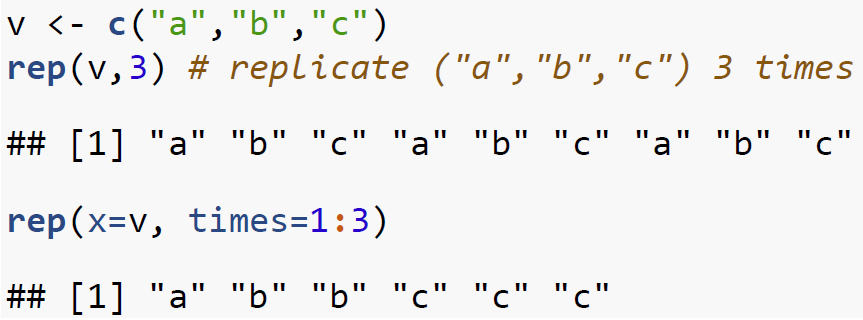
有的函数可以改变对象的属性 - as.numeric, as.integer, as.character, as.logical, etc.

例. pain.num <- **as.numeric**(fpain); **is.numeric**(pain.num)

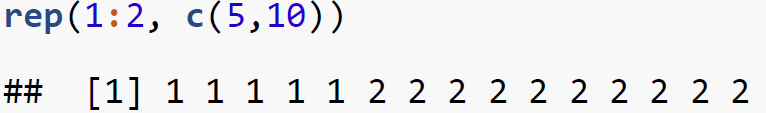
1. c()函数的功能 ：1.将字符或数字聚合成向量；2.将向量聚合

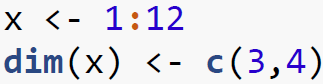
c()函数还可以用来生成vector。元素类型不同，转换成限制最小的。

1. seq()函数的功能：用于生成数列 - **seq(0,10)**

生成等距数列 - **seq(0,10,2)； seq**(from = 0, to = 10, by = 2)

1. rep()函数的功能：replicate的缩写，用于生成重复值



rep（）函数也可以用于分组

1. matrix是个二维数组。dim函数用于设置维度，存储是column-major

matrix()函数也可以用于生成矩阵 - M1 <- **matrix**(1**:**12,3,4); M1

矩阵默认以列方式填充：vector nrow ncol；如果想以行方式填充 M2 <- **matrix**(1**:**12,3,4, byrow = T); M2

1. 添加行名 – **colnames**(M1) <- LETTERS[1**:**4]; M1

添加列名 - **rownames**(M3) <- LETTERS[1**:**4]; M3

1. 矩阵的转置（行列互换）- **t(M1)**
2. cbind() – 列融合 - M4 <- **cbind**(A,B,C); M4

rbind() – 行融合 - M5 <- **rbind**(A,B,C); M5

以上两个函数的输入也可以是数字 - M4\_new <- **cbind**(M4,D=13**:**16); M4\_new

1. list() – 列表是对象的组合。对象的种类并不限制 L <- **list**(x,lv, M)
2. 列表切片 – 查询子列表信息 L[2]

查询子列表具体信息 – L[[2]]

更改子列表元素 - L[[2]][1]="W"

但是原lv数组并没有影响！！！

1. data frame是具有相同长度的元素的列表，这些元素交叉相关 - df <- **data.frame**(height,weight); df

查询其中组成部分的方法 – df $ weight & df[,2]

1. subset [ ]用于选择数据 - weight[2]

一次选择多个 - weight[c(1,3,5)] 或者 weight[1:3]

[ ]也可以直接对类矩阵搜索

条件搜索 - df[df**$**height**>**1.75 **&** df**$**weight**<**80,]

1. head() – 展示前几行数据**head**(df,x) ；tail() – 展示末尾几行
2. Sort a single vector - **sort**(df**$**weight)

另compute an ordering of a variable - o <- **order**(df**$**height);o

df**$**height[o]; df**$**weight[o]

height is sorted the way just as the result of sort(). weight is sorted according to the ordering of height.

1. df\_sorted <- df[o,] df\_sorted -以order（即先前height）的顺序进行排序

可以再添加一个属性筛选条件

1. **ls() –** 可以查看环境中储存了哪些变量、对象

**rm(a,b) –** 可以删除具体指定的对象

**rm**(list=**ls**()) – 相当于清空workspace

但是以上方法不会删除名字中有.的variable，得用**ls(all=T)**

保存输出就用save & save as

还可以新建new script以及run line or selection

1. install.packages("abc")安装包；library("abc")载入包
2. read.table() – 把数据写入R

thuesen2 <- **read.table**("C:/Users/jub69/Dropbox/A-Pitt/STAT 1293/Lectures/Lecture1\_IntroR/thuesen.txt", header=T) **head**(thuesen2,3)

**Lec2 Probability & Distribution 概率与分布**

1. 集合的定义：1. 补集complement -；2. 交集intersection -；

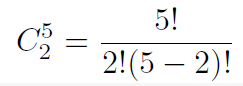
3. 并集union -

1. **union**(A,B) - Union of the sets A and B

**intersect**(A,B) - Intersection of the sets A and B

1. **setdiff**(A,B) – 用来find elements that are in A but not in B

**setdiff**(B,A) – 用来find elements that are in B but not in A

1. **setequal**(A,B) – 检验A与B之间的相等性
2. A <- **c**(2,4,6,8,10);B <- 1**:**5; B **%in%** A – B的元素是否在A里有
3. 阶乘 **factorial**(3) 

5选2 **choose**(5,2)

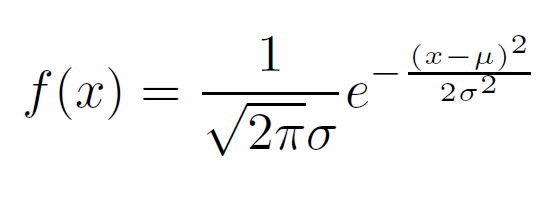
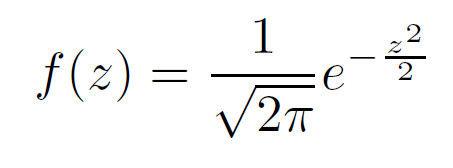
1. 离散随机变量：Binomial Distribution二项分布；Geometric Distribution几何分布；Negative Binomial Distribution负二项分布；Hypergeometric Distribution超几何分布；Poisson Distribution泊松分布

连续随机变量：Uniform Distribution均匀分布；Exponential Distribution指数分布；t distribution t分布；Normal Distribution正态分布； Chi-square Distribution卡方分布；F distribution F分布

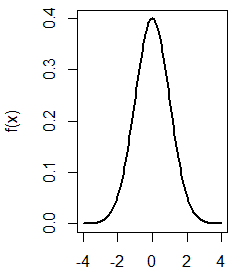
1. 每个统计分布，都有四个基本项：1. Probability density function (pdf) 概率密度函数；2. Cumulative probability distribution function (cdf) 累积概率分布函数；3. Quantiles (percentiles) 分位数；4. Pseudo-random numbers伪随机数

以正态分布而言，展现这几个项的函数是dnorm, pnorm, qnorm, and rnorm

1. **正态分布**！

若随机变量X服从一个数学期望为μ、方差为σ2的正态分布，记为N(μ，σ2)。其概率密度函数为正态分布的期望值μ决定了其位置，其标准差σ决定了分布的幅度。

当μ = 0,σ = 1时的正态分布是标准正态分布。

9.1 概率密度函数；对于标准正态分布 - dnorm(x)

一个连续型随机变量某个确定的取值点附近的可能性

**par**(mfrow=**c**(1,2))

x <- **seq**(**-**4,4,0.1)

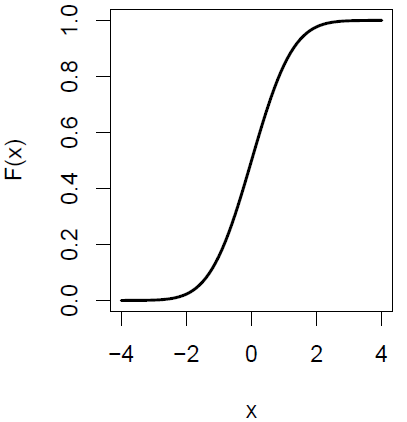
**plot**(x, **dnorm**(x), type="l", ylab="f(x)",lwd=2)

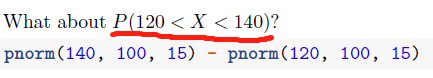
- mfrow图形参数，“multi-frame, rowwise, 1-by-2 layout”.

9.2 Cumulative Distribution Function（CDF是概率密度函数的积分）- pnorm()函数

P(Z < −2) : **pnorm**(**-**2) ; P(Z > 3.15): 1**-pnorm**(3.15)

除了标准，其他正态分布也可以找



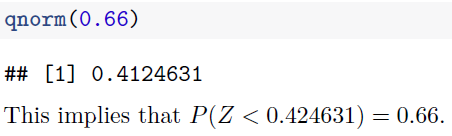
x <- **seq**(**-**4,4,0.1)

**plot**(x, **pnorm**(x), type="l", ylab="F(x)",lwd=2)

CDF的几个例题

Ex1 a student slept less than 4.5 hours: **pnorm**(4.5,7,1.5)

Ex2 a student slept more than 10.5 hours: 1**-pnorm**(10.5,7,1.5)

* 1. Quantiles/Percentiles分位数- 它是CDF的倒数，是一个具有以下性质的值：概率p小于或等于它 (x为该分位数)

Ex1 如果，求z值

（实际问题是标准正态分布的第66百分位）

Ex2 IQ是一个随机连续变量X，，查找5%和1%聪明的人智商是多少

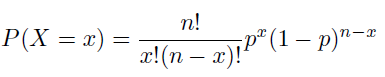
（x is the 95th percentile of X）

**qnorm**(0.95, 100, 15) - 124.6728 (智商top5%的人起码125)

**qnorm**(0.99, 100, 15) - 134.8952 (智商top1%的人起码135)

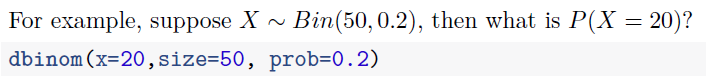
* 1. Random Numbers随机数

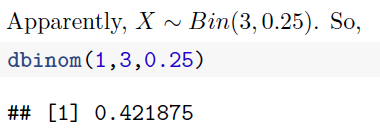
**rnorm**(5,5,2) *# generate 5 numbers from a normal dist. with mean=5 and sd=2.*

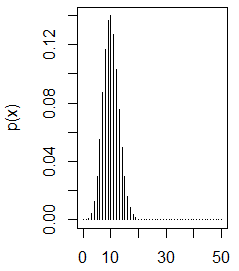
1. **Binomial Distribution二项分布！**

X是n次实验中成功的次数

（precon是结果只有成功或失败；实验相互独立；实验成功概率相同）

10.1 Mass Function/Point Probability（质量函数、点概率） - dbinom(x,n,p)

Ex1

Ex2 一对父母有三个孩子，每个孩子o型血的概率为0.25，

只有一个孩子是o型血的概率是多少？

绘图 *n* = 50; *p* = 0*.*2; x从0到50

**plot**(x,**dbinom**(x,size=50,prob=0.2),type="h", ylab="p(x)")

type="h"指画的是histogram-like（类似直方图）

10.2 Cumulative Distribution Function累积分布函数 （各点相加即可）

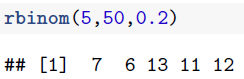


Ex. 至少有两个孩子是O型血的概率？

1**-pbinom**(1,3,0.25)

* 1. Binomial Quantiles 二项分位数

Ex. X~Bin(50, 0.2), 算一下14的分位 - **pbinom**(14,50,0.2)

* 1. 随机数

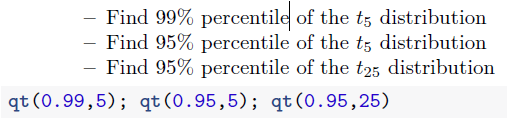
****

1. **t分布**

其基本属性：size n; mean *μ;* standard Deviation标准差 σ; 自由度为n-1

11.1 t分布的概率密度函数 dt(x,df) df是自由度

11.2 t分布的累计密度函数 pt(x,df)

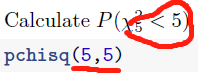
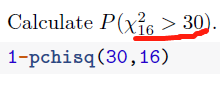
11.3 t分布的分位数

11.4 t分布的随机数 rt(n,df)

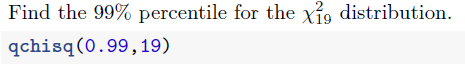
1. **Chi-square distributions卡方分布**

12.1 概率密度函数 dchisq(x,df)

12.2 cpf累计概率函数pchisq(x,df)

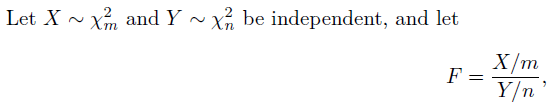
 

* 1. 分位数qchisq(p,df)



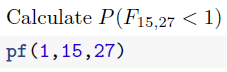
12.4 绘图

1. F分布：两个服从卡方分布的独立随机变量，各除以其自由度后的比值的抽样分布

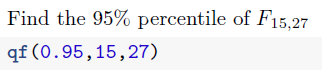
（m是分子自由度，n是分母自由度）

The set of possible values of F is (0,∞).

13.1 cdf函数：pf(x,df1,df2)



13.2 分位数：qf

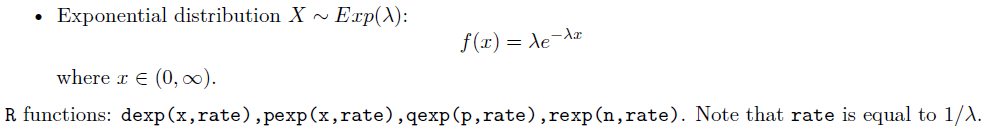


1. Poisson distribution泊松分布



函数：dpois(x,lambda), ppois(x,lambda), qpois(p,lambda), rpois(n,lambda)

1. Exponential distribution指数分布



**Lec3 数值变量的描述性统计和图形显示**

1. 1.1 Individual: 由一组数据描述的一个对象

Variable: characteristic of the individual

* categorical variable: 将individual置于几个组或类别之一
* quantitative variable: 采用算术运算的值（如加法/平均）

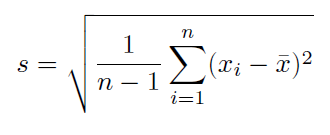
1.2 以上两者的图形显示

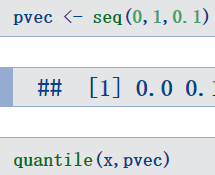
- categorical variable：Pie chart(饼状图); Bar graph(条形图)

- quantitative variable: Stem-and-leaf plot(茎叶图); Histogram(直方图); Boxplot(箱线图); Empirical CDF; Q-Q plot

1. 单数值变量的汇总统计

平均值 **mean**(x)

方差**var**(x)

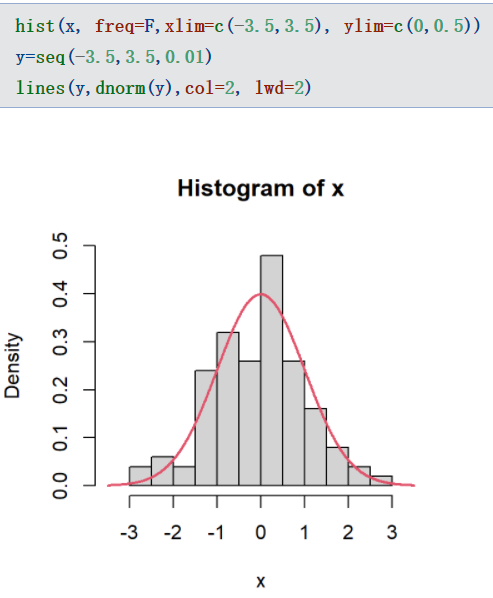
标准差**sd**(x)

中位数**median**(x)

分位数**quantile**(x,p) 第二个变量p-分位

Summary()提供数值汇总；fivenum()直接将其向量化

1. 缺失值（？ppt有问题）

na.rm = T 用该参数来请求忽略

1. **Histograms直方图：对分布形状有大致印象**

3.1 par(mfrow=c(1,2)) //一页两幅图

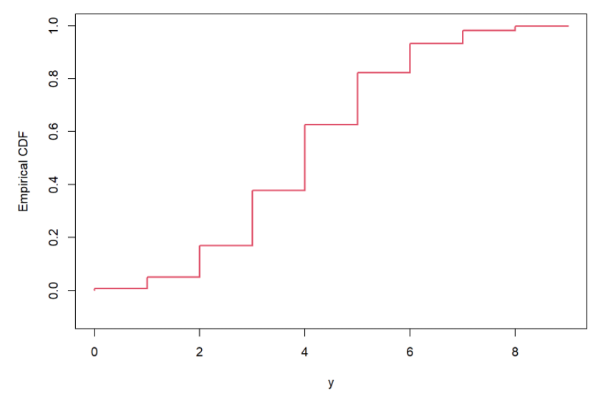
hist(x) //x的直方图

hist(x, break=30) //x的直方图，30个柱子

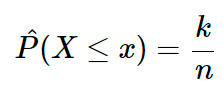
3.2 显示密度，插入参数freq=F

绘制理论密度曲线

1. **Stem-and-Leaf Plot 茎叶图：**
   1. w <- sample(100,30) //sample函数抽样：1~100的vector，抽30次
   2. stem(w) //将以上绘制成茎叶图



1. **Empirical CDF 累积经验分布：**<=x的数据部分

假设x是n个观测值中的第k个最小观测值，

则数据的k/n<=x

6.1 rbinom(n, size, prob): 产生n个b(size,prob)的二项分布

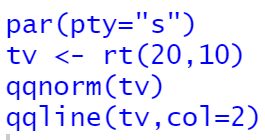
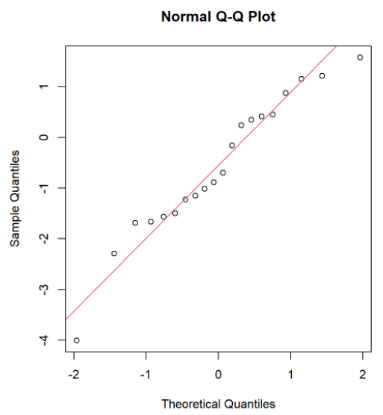
6.2 绘图：**plot(y.sorted,(1:n)/n, type="s", ylim=c(-0.1,1),lwd=2,col=2,**

**ylab="Empirical CDF", xlab="y")**

// **y.sorted** 指x轴数据；**(1:n)/n** 指y轴数据；type指线型#s意味楼梯状；ylim指y轴范围；lwd指线条宽度；col指颜色；ylab指y轴命名；xlab指x轴命名

1. **Q-Q plots**

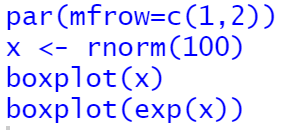
计算empiricalCDF的目的是查看数据是否normally distributed

Q-Q plot通过绘制第 k 个最小观测值，与标准正态分布n个值中的第k个最小观测值对比来评估正态性

标准正态分布是个直线 qqline



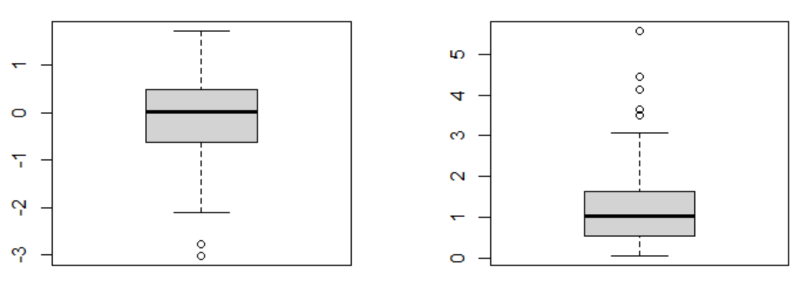
我们的观测值是点 qqnorm



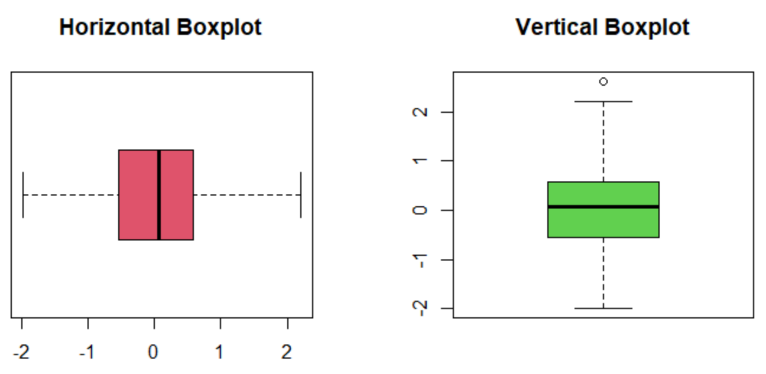
1. Boxplots：箱线图



底部是Q1四分之一位数，顶部是四分之三位数

盒子中间那q根线对应中值

如果观测值不在范围内，以点分布

**boxplot(x,outline=F,col=2,horizontal=T,main="Horizontal Boxplot" )**

**boxplot(x,outline=T,col=3,horizontal=F,main="Vertical Boxplot")**

水平、竖直箱线图

1. 计算汇总统计信息：**tapply**（要出数据的元素；数据类名；求值类型）

tapply()允许根据某些变量的值，把原始数据分割为若干组，然后对每一组数据应用特定的操作。

第二个参数定义的子组上创建函数值的表

Ex. tapply(age,sex,median) 把数据根据性别分类，求每组的中位数

tapply(age,list(sex,tanner),mean) -也可以考虑两个因素

1. **by function**

工作于整个data frame, 产生的结果也为data frame

**by(juul,sex,summary)**

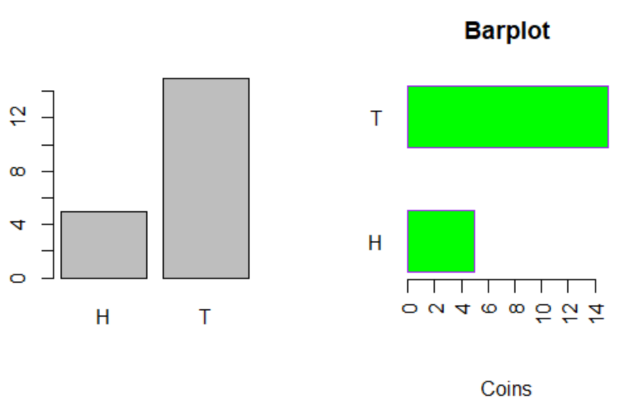
取整个数据框作为变量

By也可以多个因素组合

**detach(juul) //detach解除存储绑定**

**by(juul,list(sex,tanner),summary)**

1. 分组数据图形：**Histograms** -比较变量在不同组之间的分布

**Lec4 分类变量的描述性统计和图形显示**

1. 条形图（单变量）

**coins<-sample(c("H","T"),size=20,replace=T)**

**coins.df <- data.frame(coins)**

**coins.tb <- table(coins); coins.tb**

//抛硬币二十次；将结果转为data frame；用表格总结结果

**par(mfrow=c(1,2),pty="s")**

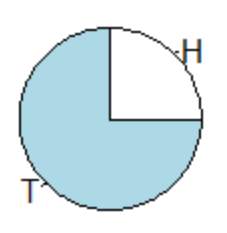
**barplot(coins.tb)**

**barplot(coins.tb, col="green",border="purple", main="Barplot",sub="Coins",horiz=T,space=1, width=1,las=2)**

//生成条形图

* 1. 条形图的细节设置（axis labels）-axis()用于改变轴

axis(side, at=, labels=, pos=, lty=, col=, las=, tck=, ...)

* side: 指示要绘制轴的图形侧面(1=bottom, 2=left, 3=top, 4=right)
* at: numeric, 表示何处绘制刻度线
* labels: 放置于刻度线的标签
* pos: 调整文字的方向位置, 1,2,3和4分别对应坐标的下，左，上和右
* lty: line type
* col: 颜色
* las: label轴平行 (=0) 或垂直 (=2)
* tck: 刻度线长度作为绘图区域的分数（图外为负数，图内为正数，0抑制刻度，1创建网格线）默认值为-0.01

1. 饼图（单变量） （用的不多）

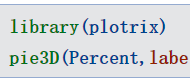
**pie(coins.tb)**

细节设置：pie(x, labels = names(x), clockwise = FALSE, col = NULL, main = NULL, ...)

* x: non-negative vector, 其值显示为饼图中切片的区域
* labels: slice的名称
* clockwise: slice是顺时针还是逆时针绘制
* col: 颜色
* main: 图的名称
* //画图前可以先o. 先排序再画图
* **lb=paste(Discipline,perc,"%") //添加百分比**

**attach(major);par(mar=c(4,6,2,6))**

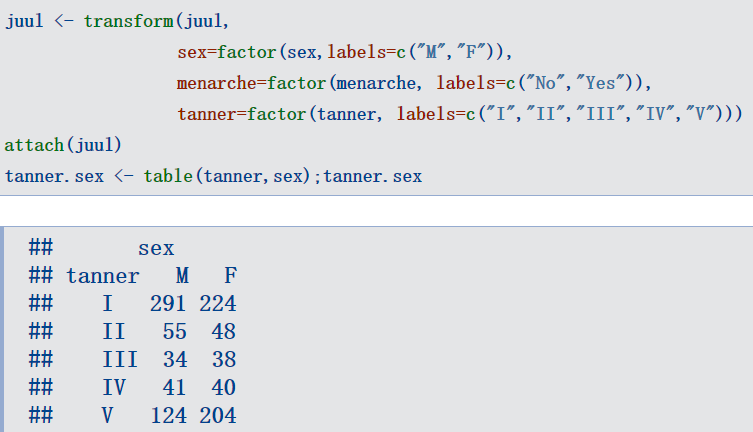
**pie(Percent,Discipline,col=rainbow(length(Discipline)),main="Distribution of Major",clockwise = T)**

如何画3D饼图？？

1. 两个分类变量：two-way table

Twoway table 可以作为矩阵对象输入

**lens<- matrix(c(121,42,32,37,129,85),nrow=2,ncol=3); lens**

* 1. 存在一个class为table，用as.table（）将矩阵转换为table

as.data.frame() 也能用

* 1. 双向表的建立过程（右）

1. **Marginal Tables边际表**

margin.table()函数

1. 双变量条形图

研究两个分类变量

Barplot第一个参数改为x.y

1. 相对频率表Tables of relative frequencies

**prop.table(tanner.sex,1) //计算行的相对比例**

**prop.table(tanner.sex,2) //计算列的相对比例**

1. 条件比例条形图，

# LEC5&6 单样本与双样本测试

1. Standard Error：s/(n^1/2); s为标准差，n为样本量
2. t分布
   1. 密度曲线类似正态分布，关于0对称，single-peaked单峰, bell-shaped钟型；扩散性大于正态分布（尾部概率更大，中心概率更小）
3. 手动测量CI：
   1. 先算自由度（样本量-1）

**n <- length(Percent)**

**df = n -1 ; df**

* 1. 默认置信水平95，计算97.5分位数

**t <- qt(0.975, df); t**

* 1. 计算上下限

**x\_bar <- mean(Percent)**

**s <- sd(Percent)**

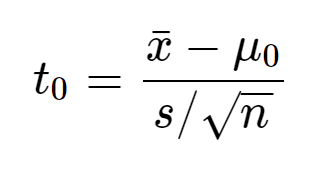
**CI = x\_bar +c(-1,1)\*t\*s/sqrt(n); CI**

* 1. 简易法

**t.test(Percent, conf.level=.95)$conf.int**

4. 用R语言进行t检验的流程例子（天气好的时候，小费超过20%，置信水平95）

4.1 陈述假设：

4.2 计算检验统计量

**x\_bar=mean(Percent); s=sd(Percent); n=length(Percent);df= n-1**

**t\_0 = (x\_bar-20)/(s/sqrt(n)); t\_0**

4.3 决策规则

因为假设的是大于，所以是右侧检验

求得p值，非常小，拒绝h0，求拒绝域，与t0做比较

4.4 结论成立

5. t.test

5.1 t.test(x, alternative, mu)重要参数：x为向量；alternative是替代假设的字符串one of “two.sided” (default), “greater” or “less”；mu是原假设下的均值

6. Wilcoxon test in R

wilcox.test(x, alternative, mu)

6.1 除非样本量为 6 或以上，否则符号秩检验在 5% 的水平上根本不会变得显着

精确的 p 值exact=T。

默认情况下，wilcox.test具有连续性校正。您可以通过指定将其关闭correct=F。

1. 例题
   1. 计算99%置信下限

**t <- qt(.99, df)**

**d\_bar - t\*s\_d/sqrt(n)**

//d\_bar是mean; s\_d是方差

1. 配对检验