定量研究方法期中考试材料

第三讲 概率与条件概率

概率公理

$P(A) \geq 0$

$P(\Omega = 1)$

加法法则 $P(A或B) = P(A) + P(B) - P(A和B)$

全概率公式

$$P(A或B) = P(A和B) + P(A和B^c) + P(B和A^c)P(A) = P(A和B) + P(A和B^c)$$

条件概率

$$P(A|B) = \frac{P(A和B)}{P(B)}$$

$P(A和B)$ ← 联合概率；$P(B)$ ← 边际概率

乘法法则 $P(A和B) = P(A|B)P(B) = P(B|A)P(A)$

全概率公式与条件概率 $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

贝叶斯法则

后验概率 → $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$（$P(A)$：先验概率）

独立性

事件A、B独立的充要条件 $P(A和B) = P(A)P(B)$

条件独立 给定事件C，事件A、B条件独立的充要条件

$P(A和B|C) = P(A|C)P(B|C)$ 即 $P(A|B和C) = P(A|C)$

第四讲 随机变量与概率分布

随机变量的种类

离散随机变量： 数值为有限个，如家庭成员个数、态度（支持/反对）等。
→概率质量函数（PMF）、累积分布函数（CDF）

连续随机变量： 在某个实线区间内有无限个取值，如长度、GDP等。
→概率密度函数（PDF）、累积分布函数（CDF）

二项分布

二项随机变量： n个独立并且相同分布（i.i.d.）的伯努利随机变量之和。

$\binom{n}{x} = C_n^x = \frac{n!}{x!(n-x)!}$

PMF:

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

CDF:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$$

分布

伯努利分布

伯努利随机变量： 一个二元随机变量，该二元随机变量的数值为两个不同的数。

PMF:

$$f(x) = \begin{cases} p & (x=1) \\ 1-p & (x=0) \\ 0 & 其他 \end{cases}$$

CDF:

$$F(x) = \begin{cases} 0 & (x<0) \\ 1-p & (0 \leq x < 1) \\ 1 & (x \geq 1) \end{cases}$$

均匀分布

均匀随机变量： 均匀随机变量在给定区间[a, b]内取一值的可能性相同。

PDF: 在[a, b]内

$$f(x) = \begin{cases} \frac{1}{b-a} & (a \leq x \leq b) \\ 0 & 其他 \end{cases}$$

CDF: 在[a, b]内

$$F(x) = \begin{cases} 0 & (x<a) \\ \frac{x-a}{b-a} & (a \leq x \leq b) \\ 1 & (x \geq b) \end{cases}$$

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P\left(-k \leq \frac{X-\mu}{\sigma} \leq k\right) = P(-k \leq Z \leq k)$$

正态分布

关键指标：
平均数μ
标准差σ
$X \sim N(\mu, \sigma^2)$，σ^2表示方差

PDF:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

CDF:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

CDF代表PDF下方从-∞到x所形成的面积

- 正态分布PDF形状为钟型，中心为平均数，分布范围由标准差大小决定。
- 正态分布PDF下的面积为1。
- 不同的平均数只移动PDF和CDF，但不改变形状。
- 标准差增加可造成PDF更平坦，CDF增长趋势更平缓。

z分数

定义为Z=X+c的随机变量服从$Z \sim N(\mu+c, \sigma^2)$的正态分布；
定义为Z=cX的随机变量服从$Z \sim N(c\mu, (c\sigma)^2)$的正态分布。

$$z分数 = \frac{X-\mu}{\sigma} \sim N(0,1)$$

计算正态随机变量位于平均值k (k > 0)个标准差之间的概率（下方左式）
其中，$P(-k \leqslant Z \leqslant k)$也可以被转写成CDF在-k和k取值时的差（下方右式）

$$P(-k \leq Z \leq k) = P(Z \leq k) - P(Z \leq -k) = F(k) - F(-k)$$

第五讲 点估计

估计相关概念

估计目标（estimand）是我们需要估计的总体的参数（parameter）。通常用希腊字母表示（如μ、θ）。
估计量（estimator）是我们估算估计目标的方法。通常用希腊字母加hat表示（如$\hat{\mu}$、$\hat{\theta}$）。
估计值（estimate）是我们运用估计量从所得样本中进行估算所得的具体数值。

期望和方差

期望

$$E(X) = \begin{cases} \sum_x x \cdot f(x) & 如果X是离散的 \\ \int x \cdot f(x)dx & 如果X是连续的 \end{cases}$$

其中，f(x)是离散变量X的概率质量函数（PMF）或连续变量X的概率密度函数（PDF）

$E(a) = a$ $E[E(X)] = E(X)$
$E(aX) = aE(X)$
$E(aX+b) = aE(X) + b$
$E(aX+bY) = aE(X) + bE(Y)$

X与Y独立的必要非充分条件：
$E(XY) = E(X)E(Y)$

方差

$$V(X) = E[\{X - E(X)\}^2] = E(X^2) - \{E(X)\}^2$$

$V(a) = 0$ $V(aX) = a^2V(X)$
$V(X+b) = V(X)$
$V(aX+b) = a^2V(X)$

X与Y独立的必要非充分条件：
$V(X+Y) = V(X) + V(Y)$

协方差

$$Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E(XY) - E(X)E(Y)$$

X与Y独立的必要非充分条件：
$Cov(X,Y) = 0$

适用于任何样本容量的有限样本特征

无偏性
偏误：$Bias(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$
估计量无偏的充要条件：$Bias(\hat{\theta}) = 0 \Leftrightarrow E(\hat{\theta}) = \theta$

相对有效性 比较$\hat{\theta}_1$和$\hat{\theta}_2$，如果的$\hat{\theta}_1$方差小于$\hat{\theta}_2$的方差，那么$\hat{\theta}_1$更有效 $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

平均数平方误差（均方误，MSE）
$MSE(\hat{\theta}) = E\left\{(\hat{\theta} - \theta)^2\right\} = V(\hat{\theta}) + Bias(\hat{\theta})^2$
用于估计量都存在偏误时，选取合适的估计方案。

适用于大样本容量的特征

一致性 如果一个估计量$\hat{\theta}_n$的抽样分布$\hat{\theta}_1, \cdots, \hat{\theta}_n$随着样本容量n的增加越来越集中于估计目标θ，那么$\hat{\theta}_n$就是一致的。 $\hat{\theta}_n \xrightarrow{p} \theta$ 即 $p - \lim_{n\to\infty} \hat{\theta}_n = 0$

一致性体现在估计量的期望值越来越接近估计目标以及估计量的方差逐渐趋近于0。
一致的估计量不一定无偏，无偏误的估计量不一定具有一致性。

渐进正态性

中心极限定理： 样本平均数的分布随着样本数的增加接近正态分布。
假设我们有独立并且相同分布（i.i.d）的随机变量的样本$X_1, X_2, \ldots, X_n$，形成一个均值为μ，方差为σ^2的概率分布。样本平均数用$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$表示，那么中心极限定理为：

$$Z_n = \frac{\overline{X_n} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \to N(0, 1)$$

大数定理： 在样本容量足够大的情况下

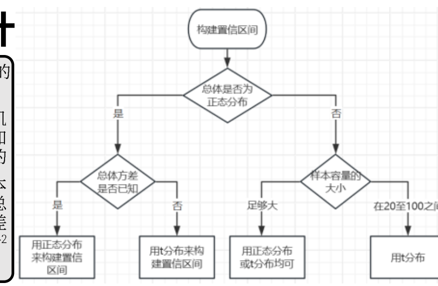
$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu = E(X)$$

其中，μ可以用E(X)来表示，σ^2可以用V(X)来表示。
这一公式的含义是：随着样本规模的增加，分布样本平均数的z分数收敛成标准正态分布，即N(0,1)。如果样本容量n足够大，我们可以用样本标准差S_n来代替σ。

在样本容量足够大的情况下： $\overline{X_n} \to N\left(\mu, \frac{\sigma^2}{n}\right)$

这一公式的含义是：随着样本规模的增加，样本平均数的分布收敛成正态分布，即$N\left(\mu, \frac{\sigma^2}{n}\right)$

第六讲 区间估计



标准误

用于刻画随机抽样所得的样本平均值的离散程度。

假设我们有独立且相同分布的随机变量样本$\{X_1, X_2, \ldots, X_n\}$，并且已知总体方差$\sigma^2$的时候，样本平均数的标准误为：$\frac{\sigma}{\sqrt{n}}$

可以通过样本方差来估计总体方差σ^2

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

置信水平（置信度）

决定了在多大程度上确信区间内包括估计目标的真实值。研究者通常选取95%、90%和99%作为置信水平（其中95%置信水平最为常见）。
置信水平通常被写为$(1-\alpha) \times 100\%$。

σ^2已知的置信区间

假设我们有独立并且相同分布的随机变量的样本$X_1, X_2, \ldots, X_n$，并且来自于总体平均值为μ，方差为σ^2的正态分布中，总体方差已知的情况下，对于样本平均数$\bar{X}$的$(1-\alpha) \times 100\%$置信区间为：

$$CI(\alpha) = \left[\bar{X} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]$$

其中$\frac{\sigma}{\sqrt{n}}$为标准误，$z_{\frac{\alpha}{2}}$为临界值

临界值$z_{\frac{\alpha}{2}}$等于标准正态分布N(0, 1)的$(1-\frac{\alpha}{2})$分位数。最大临界值$z_{\frac{\alpha}{2}}$和最小临界值$-z_{\frac{\alpha}{2}}$围成的面积为置信水平$(1-\alpha) \times 100\%$。置信水平越高，最大临界值越高，在其他条件不变的情况下，所得置信区间的范围越大。

n%置信区间的含义是：在重复数据产生的过程中，有n%的数据的区间内包含估计目标。

σ^2未知的置信区间

学生t分布

$S_n^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n-1}$ 是σ2无偏且具有一致性的估计量，此时标准误为$\frac{S_n}{\sqrt{n}}$

假设我们有独立并且相同分布的随机变量的样本$X_1, X_2, ..., X_n$，并且来自于总体平均值为μ，方差为σ²的正态分布中，则样本平均数$\overline{X_n}$和标准误$\frac{S_n}{\sqrt{n}}$的t统计量服从自由度为n-1的学生t分布：

$$T_n=\frac{\overline{X_n}-\mu}{\frac{S_n}{\sqrt{n}}}\sim t_{n-1}$$

t分布与正态分布相似，但是尾部更大。随着自由度（df）的增加，t分布接近正态分布。一般而言，当自由度大于100的时候，t分布和正态分布是没有显著区别的。自由度为观测值的个数（样本容量）减去待估参数的个数；在估计μ的例子中，待估参数为1。

在总体方差未知的情况下，对于样本平均数$\bar{X}$的(1−α)×100%的大样本置信区间为：

$$CI(\alpha)=\left[\bar{X}-t_{\frac{\alpha}{2},n-1}\times\frac{S_n}{\sqrt{n}},\bar{X}+t_{\frac{\alpha}{2},n-1}\times\frac{S_n}{\sqrt{n}}\right]$$

其中为$\frac{S_n}{\sqrt{n}}$标准误，$t_{\frac{\alpha}{2},n-1}$为临界值
在相同置信水平的情况下，t分布的临界值要稍大于正态分布的临界值

随着自由度（df）的增加，t分布接近正态分布。
随着n→∞，则（见右式）：

$$T_n=\frac{\overline{X_n}-\mu}{\frac{S_n}{\sqrt{n}}}\rightarrow N(0,1)$$

这意味着当样本容量足够大的时候我们不需要假设总体服从正态分布。

第七讲 假设检验

1.设置零假设H_0；
2.设置备选假设H_1并确定采取单边检验还是双边检验；
3.设定检验水平α；
4.选取检验统计量Z_n。如果n足够大，$Z_n\sim N(0,1)$；
5.计算统计量Z_n；
6.根据检验水平选取临界值z；
  a) 检验水平α=0.05；
  b) 临界值z——单边：1.64；双边：1.96。
7.比较Z_n与z；
  a) 如果统计量$|Z_n|>z$，拒绝H_0；
  b) 如果统计量$|Z_n|\leqslant z$，不拒绝（保持）H_0。
8.同时我们也可以计算p值，再比较p值与α值。
  a) p值<α值，拒绝H_0；
  b) p值⩾α值，不拒绝H_0。

零假设、备选假设

$H_0:\mu=0.45$
$H_1:\mu>0.45$

μ>或<x → 单边
μ≠x → 双边

检验水平α
（显著性水平）

α值代表着我们对犯第一类错误（弃真）的容忍程度，常用的α值有0.10、0.05和0.01。

检验统计量Z_n

$$Z_n=\frac{\overline{X_n}-\mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

样本容量足够大时，可以用样本方差S_n^2来估计$\hat{\sigma}$且$Z_n\sim N(0,1)$

*临界值z会在卷面上给出。

p值检验

p值指的是在零假设正确的情况下，至少观测到一次检验统计量的概率；p值越小，拒绝零假设的证据越强

单边：$p值=P(Z_n\geq Z_n|H_0)$

双边：$p值=P(Z_n\geq Z_n或Z_n\leq Z_n|H_0)$

双样本检验

步骤与单样本检验基本一致

$H_0:\mu_a=\mu_b$
$H_1:\mu_a\neq\mu_b$

$$Z_n=\frac{\mu_a-\mu_b}{\sqrt{\frac{\hat{\sigma}_a^2}{n_a}+\frac{\hat{\sigma}_b^2}{n_b}}}$$

第八讲 简单线性回归

相关性

样本协方差

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

当Cov(X, Y)为正时，整体数据云的趋势向上；为负时，则向下。

相关系数

$$Cor(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{Cov(X,Y)}{S_xS_y}$$

其中，S_x、S_y分别是变量X和Y的标准差

相关性的性质

1.相关系数是一种刻画两个随机变量线性关系的指标；
2.相关系数为正（负）时，一个观测值的X值大于（小于）其样本平均数，那么这个观测值所对应的Y值很可能也大于其样本平均值；
3.相关系数为正（负），散点图中的数据云呈上升（下降）趋势；
4.相关系数的绝对值大，则数据点紧紧环绕在直线周围；
5.相关系数的范围是[−1,1]，1（-1）为完全正（负）相关；
6.X和Y相关并不意味着二者具有因果关系。

最小二乘法

简单线性模型

$\hat{Y}=\hat{\alpha}+\hat{\beta}X$

$\hat{\alpha}$、$\hat{\beta}$：α和β的估计值
$\hat{Y}$：结果变量的预测值（或拟合值）

残差（预测误差）：观测结果值Y与预测值$\hat{Y}$之间的差

$\hat{\epsilon}=Y-\hat{Y}=Y-\hat{\alpha}-\hat{\beta}X$

以最小二乘法计算相关系数

残差平方和（SSR）：

$$SSR=\sum_{i=1}^{n}\hat{\epsilon}_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)^2$$

残差平方和最小时，最小二乘估计的$\hat{\alpha}$和$\hat{\beta}$：

$$\begin{cases}\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}\\ \hat{\beta}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\end{cases}$$

根据协方差和样本方差的公式有：

$$\hat{\beta}=\frac{Cov(X,Y)}{S_X^2}$$

根据相关系数的公式有：

$$\hat{\beta}=\rho_{XY}\times\frac{S_Y}{S_X}$$

其中，ρ_{XY}是X和Y的相关系数

第九讲 最小二乘法的性质与应用

回归线与均值

回归线总是通过数据的中心点$(\bar{X},\bar{Y})$
残差平均值为0

均方根误差（RMSE）：
表示回归预测误差的平均幅度

$$RMSE=\sqrt{\frac{1}{n}SSR}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$

用于预测的回归模型

E[Y|X]描述了在所有X可以取到的值的条件下Y的平均值是如何变化的。我们可以用观察到的样本来拟合出E[Y|X]，从而运用拟合模型来进行预测推断。

可以将1~2个新的X值x_{new}代入回归模型算出Y的期望值：

$$\hat{E}[Y|X=x_{new}]=\hat{\alpha}+\hat{\beta}x_{new}$$

$\hat{\beta}$可以被解释成当X改变一个单位时两个数据点的Y值的平均/期望差异：若X增加（减少）n个单位，则Y的预期变化为$n\hat{\beta}$

$$\hat{E}\left[Y\middle|X=x_{new,1}\right]-\hat{E}\left[Y\middle|X=x_{new,2}\right]=\hat{\beta}(x_{new,1}-x_{new,2})$$

总平方和、解释平方和与残差平方和

在一个回归模型中，解释平方和是可被回归模型解释的部分，残差平方和则是没有被回归模型预测到的部分。

$$TSS=ESS+SSR \qquad S_Y^2=S_{\hat{Y}}^2+S_{\hat{\epsilon}}^2$$

总平方和（TSS）：$TSS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=(n-1)S_Y^2$

解释平方和（ESS）：$ESS=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2=(n-1)S_{\hat{Y}}^2$

残差平方和（SSR）：$SSR=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=(n-1)S_{\hat{\epsilon}}^2$

决定系数R^2

为预测值X可以解释的总平方和（TSS）的比例，表明线性模型与数据的拟合程度

$$R^2=\frac{ESS}{TSS}=1-\frac{SSR}{TSS}$$

性质：
- 0⩽R²⩽1；
- 当R²=1时，所有的数据点都在回归线上；
- 当R²=0时，X和Y之间没有相关性。

应注意，R²越大，数据的拟合程度不一定越好

类别	语法/代码	参数解释	备注
数学	`a + b`		加法
数学	`a - b`		减法
数学	`a * b`		乘法
数学	`a / b`		除法
数学	`a ^ b`		幂运算
数学	`sqrt(a)`		开方
逻辑	`a > b, a < b`		大于、小于
逻辑	`a >= b, a <= b`		大于等于、小于等于
逻辑	`a == b`		等于
逻辑	`a != b`		不等于
逻辑	NA		缺失项
逻辑	TRUE		真
逻辑	FALSE		假
逻辑	`as.numeric(TRUE)`		数字1
逻辑	`as.numeric(FALSE)`		数字0
逻辑	`rep(a)`	a：需要复制的内容	复制；可用于大规模赋值，如dat$col <- rep(a)
对象	`obj <- a`	obj：对象名，不建议以数字或特殊字符为首字符 a：变量，可以是数值，也可以是字符串（此时应注意双引号）	赋值
对象	`obj = a`		赋值，同上
对象	`print(obj)`		显示对象
对象	`obj`		直接键入对象名，效果同print(obj)
对象	`class(obj)`		对象的类型，例如数值numeric，字符（串）character，函数function
函数	`name <- function(x){ ... }`	name：函数名 x：变量名	定义函数
函数	`return(obj)`	obj：对象	返回对象
向量	`c(x, y, z)`	x, y, z：对象；当对象均为向量时，即为合并向量	创建、显示以及合并向量
向量	`vec[a]`	vec：向量 a：索引号	向量可以参与数学运算，效果为对其中的每个元素都执行运算
向量	`vec[c(x, -y)]`	vec：向量 x, y：在vec中存在的索引号；若前面加上-，则意味着排除	
向量	`length(vec)`	vec：向量	向量中的元素数量
向量	`min(vec)`	vec：向量	向量中的最小值
向量	`max(vec)`	vec：向量	向量中的最大值
向量	`range(vec)`	vec：向量	向量中的最小值和最大值
向量	`mean(vec[, na.rm = TRUE/FALSE])`	vec：向量 na.rm：是否忽略缺失项；默认为FALSE	向量中各元素的平均值
向量	`round(obj[, a])`	obj：对象，可以是数值，也可以是向量（此时对向量中每个元素执行） a：小数点后保留位数，默认为0	四舍五入
向量	`sum(vec)`	vec：向量	向量中各元素求和
向量	`median(vec)`	vec：向量	向量中各元素的中位数
向量	`seq([from = ]x, [to = ]y[, by = z])`	x：起始值 y：终止值 z：步长，默认为1；为-1时倒序	创建一个[x, y]，步长为z的向量
向量	`c(x:y)`	x：起始值 y：终止值	创建一个[x, y]，步长为1的向量
向量	`names(vec1) <- vec2`	vec1：被命名的向量名 vec2：命名来源的向量名	为vec1中的每个元素命名
向量	`var(x[, na.rm = TRUE/FALSE])`	x：一个向量，或数据表中的一列，或某个numeric对象 na.rm：是否忽略缺失项；默认为FALSE	方差
向量	`sd(x[, na.rm = TRUE/FALSE])`	x：一个向量，或数据表中的一列，或某个numeric对象 na.rm：是否忽略缺失项；默认为FALSE	标准差
矩阵	`cbind(vec1, vec2)`	vec1, vec2：作为列的向量	以列绑定向量（即向量为列，序号为行），返回一个矩阵matrix（数列array）对象
矩阵	`rbind(vec1, vec2)`	vec1, vec2：作为行的向量	以行绑定向量（即向量为行，序号为列），返回一个矩阵matrix（数列array）对象
矩阵	`mat[[a], [b]]`	mat：矩阵 a：行的序数；若留空，则表示整列 b：列的序数；若留空，则表示整行	矩阵中的元素
矩阵	`colMeans(mat[, na.rm = TRUE/FALSE])`	mat：矩阵 na.rm：是否忽略缺失项；默认为FALSE	计算矩阵中每一列的平均值，并将其输出为一个向量
矩阵	`apply(mat, x, fun)`	mat：矩阵 x：当x = 1时，作用于每一行；当x = 2时，作用于每一列 fun：函数	将函数fun作用于矩阵mat的每一行（或每一列）
数据表	`as.data.frame(mat)`	mat：矩阵	将矩阵转换为数据表，返回一个数据表data.frame对象
数据表	`head(dat, a)`	dat：数据表 a：显示行数	数据表的首a行
数据表	`tail(dat, a)`	dat：数据表 a：显示行数	数据表的末a行
数据表	`dat[[a], [b]]`	a：行；可以是行的序数，也可以是行的名称（字符串）；留空时，显示全行的元素 b：列；可以是列的序数，也可以是列的名称（字符串）；留空时，显示全列的元素	数据表指定行指定列的元素
数据表	`dat$name`	name：行的名称或列的名称（作为字符串，此时不用加引号）	数据表指定行或指定列的所有元素
数据表	`as.numeric(obj)`	obj：非numeric对象	返回对象的数字numeric对象
数据表	`colnames(dat)`	dat：数据表	所有列的名称；同names(dat)
数据表	`colnames(dat) <- vec`	dat：数据表 vec：向量	重命名列的名称；同names(dat) <- vec
数据表	`colnames(dat)[a]`	a：列的序数	指定列的名称；可同上一项语法一样，为其赋值以重命名
数据表	`ncol(dat)`	dat：数据表	总列数
数据表	`nrow(dat)`	dat：数据表	总行数
数据表	`dim(dat)[[a]]`	dat：数据表 a：1或2；当为1时，为行数；当为2时，为列数	总行数与总列数
数据表	`summary(dat)`	dat：数据表；亦可为dat$name格式	按列显示数据表概述；当其为dat$name格式时，仅显示指定列的概述
数据表	`unique(col)`	col：行或列的名称	该行或列不重复的所有元素
数据表	`subset(dat, con)`	dat：数据表 con：条件，一般是逻辑对象；可有多个条件，条件之间以,分隔	满足指定条件的数据表子集；常用条件包括： 1)行列条件：如col == name 2)选择条件：如select = c("colname1", "colname2", "colname3") 3)子集条件：如subset = (col == name)
数据表	`quantile(col)`	col：数据表内的行或列，即向量，如dat$name	上下四分位数
数据表	`table(dat$col, exclude = a)`	dat：数据表 col：列的名称 a：排除的值，一般填NULL（排除空值）	对于一个二元的数据表的列，显示各值的频数
文件	`setwd(dir)`	dir：目录，为字符串形式；其中的反斜杠（\）需要替换为斜杠（/）	设置工作目录
文件	`getwd()`		显示工作目录
文件	`write.csv(dat, file = fname)`	dat：数据表 fname：文件名，应是一个带.csv后缀的字符串	将数据表存储为csv文件
文件	`saveRDS(dat,file = fname)`	dat：数据表 fname：文件名，应是一个带.rds后缀的字符串	将数据表存储为rds文件
文件	`readRDS(fname)`	fname：文件名，应是一个带.rds后缀的字符串	读取rds文件，将其作为数据表data.frame对象；通常将其赋值给一个对象，方便之后调用
图表	`plot(x = dat$col1, y = dat$col2, col = colour, type = typename, xlim = vec1, ylim = vec2, xlab = lab1, ylab = lab2, main = title)`	dat：数据表 col1, col2：列的名称 colour：颜色（字符串） typename：图表类型（字符串） vec1, vec2：用于划定x轴和y轴范围的向量 lab1, lab2：用于命名x轴和y轴的字符串 title：图表标题（字符串）	绘制折线图或散点图 图表类型包括： 1)"l"：折线（曲线）图； 2)"p"：散点图。
图表	`abline([v = x], [h = y], lty = t)`	x：横轴上的值 y：纵轴上的值 t：直线线型（数字）；1为实线，2为断线，3为虚线，4为点横线	在图表上的指定值处作一条垂直于坐标轴的直线
图表	`lines(x, a, col = colour, lty = t, lwd = b)`	a：数据表中的一列，或某个分布函数 colour：颜色（字符串） t：线条的线型（数字）（见第65行） b：线条的粗细（数字）；默认为2，数字越大线条就越粗	在图表上显示指定的曲线
图表	`segments(x1, y1, x2, y2)`	x1, y1：线段起点的横纵坐标 x2, y2：线段终点的横纵坐标	绘制一个指定坐标间的线段
随机	`set.seed(a)`	a：种子码	设置随机种子
概率	`sample(vec, size = a, replace = b)`	vec：作为样本空间的向量 a：试验次数 b：布尔值；为TRUE时允许重复的事件（有替换地取样），为FALSE时不允许重复的事件（无替换地取样）	在样本空间内进行一定次数的随机取样，返回一个包含结果的向量
概率	`replicate(n = x, sample(vec, size = a, replace = b))`	x：重复次数 vec：作为样本空间的向量 a：试验次数 b：布尔值（同上项）	一个常用的重复抽样的函数
概率	`dbinom(x = a, size = b, prob = c)`	a：成功次数 b：试验次数 c：每次试验的成功概率	二项随机变量的概率质量函数（PMF）的值
概率	`pbinom(q = a, size = b, prob = c)`	q：成功次数 b：试验次数 c：每次试验的成功概率	二项随机变量的累积分布函数（CDF）的值
概率	`dnorm(x, mean = a, sd = b)`	a：平均数 b：标准差	正态随机变量的概率密度函数（PDF）
分布	`density(col)`	col：数据表中的一列，或某个numeric对象	生成一个density对象
分布	`qnorm(x)`	x：置信水平的对应数值（99%→x=0.995，95%→x=0.975，90%→x=0.95）	置信水平在正态分布的对应临界值
分布	`qt(x, df = y)`	x：置信水平的对应数值（99%→x=0.995，95%→x=0.975，90%→x=0.95） y：自由度；自由度为n-1时，此处填写length(n-1)	置信水平在学生t分布的对应临界值
分布	`pnorm(z, lower.tail = a)`	z：检验统计量 a：布尔值，当TRUE时为双边检测，当FALSE时为单边检测	进行单样本p值假设检验（计算p值）
分布	`t.test(x, mu = a, alternative = b)`	x：一个向量，或数据表中的一列，或某个numeric对象 a：预先指定值（零假设中正确的概率） b：字符串，当为"greater"或"less"时为单边检测	进行单样本t检验（计算t值）。在样本容量足够大（不少于100）的情况下，t检验和z检验的结果差距不大
回归	`cor(x, y)`	x, y：随机变量	两个随机变量的相关性
回归	`lm(y ~ x, data = dat)`	y：数据dat中的随机变量名，因变量 x：数据dat中的随机变量名，自变量 dat：数据	以最小二乘法拟合线性回归模型
回归	`fitted(lm)`	lm：线性回归模型	返回线性模型拟合后的预测值
回归	`coef(lm)`	lm：线性回归模型	先后返回线性回归模型的截距(α_hat)和回归系数(β_hat)
回归	`resid(lm)`	lm：线性回归模型	返回线性回归模型的残差(ε_hat)