《应用统计学--公式及知识点汇总》

第一章 数据的整理与图形表示

1. **饼形图：**适用于分组个数比较少的，并且多用于描述和表现各成分或某一成分占全部的百分比。
2. **直方图：**根据从生产过程中收集来的质量数据分布情况，画成以组距为底边、以频数为高度的一系列连接起来的直方型矩形图。
3. **条形图和柱状图：**条形图是用宽度相同的条形的高度或长短来表示数据多少的图形。柱形图，又称长条图、柱状图，是一种以长方形的长度为变量的统计图表。
4. **并列条形图或柱状图**有利于对两组和两组以上的并列数据进行比较。
5. **茎叶图：**把一个数字分成两个部分，通常是以该组数据的高位数值作为茎，而叶上只保留该数值的最后一个数字。

公式名称	数学公式	说明
组距	（最大值－最小值）/组数	——
	全距/组数	
组中值	（上限＋下限）/2	——
	上限－相邻组的组距/2	开口组只有上限
	上限＋相邻组的组距/2	开口组只有下限

数据的描述性指标

公式名称		数学公式	说明
数据的集中趋势度量	平均数	简单算术平均数：$\bar{x}=\frac{x_1+x_2+\cdots\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$	x_1，x_2，……，x_n为数据； $\bar{x}$为数据的均值
		加权算术平均数：$\bar{x}=\frac{x_1q_1+x_2q_2+\cdots\cdots+x_nq_n}{q_1+q_2+\cdots\cdots+q_n}=\frac{\sum_{i=1}^{n}x_iq_i}{\sum_{i=1}^{n}q_i}$	x_1，x_2，……，x_n为数据； q_1，q_2，……，q_n为数据的权重
	众数	——公式不做要求	一组数据中出现次数最多的数值
	中位数	$M_d=\frac{x_{n+1}}{2}$　　(数据n为奇数)	先排序，再计算第几个数
		$M_d=\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$　　(数据n为偶数)	
	四分位数	上四分位数：$i=\left\lfloor\frac{75}{100}\times n\right\rfloor$	——
		下四分位数：$j=\left\lfloor\frac{25}{100}\times n\right\rfloor$	
数据的离散趋势度量	极差	$R=X_{max}-X_{min}$	一组数据内最大值与最小值之差，又称范围误差或全距，以R表示
	四分位差	——公式不做要求	上四分位数－下四分位数
	方差	样本方差：$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 总体方差：$\sigma^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2$	方差越大，离散程度越大；反之，越小。
	标准差	样本标准差：$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ 总体标准差：$\sigma=\sqrt{\sigma^2}=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2}$	标准差越大，离散程度越大；反之，越小。

❖ **为何度量数据的离散趋势？？**

①数据的离散趋势常常作为衡量数据稳定性的工作，如说明产品质量是否稳定；

②离散度量数值越小质量越稳定；

③离散度量数值越小说明风险越小，数值越大风险越大。

第二章 随机变量以及抽样分布

公式名称		数学公式	说明
数学期望	离散型随机变量	$\mu = E(X) = \sum_i x_i p_i$	① $p_i \geq 0$ $(i = 1,2, \ldots\ldots)$ ②$\sum_i p_i = 1$ $(i = 1,2, \ldots\ldots)$
	连续型随机变量	$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx$	$f(x) \geq 0$，且$\int_{-\infty}^{+\infty} f(x)dx = 1$
方差	离散型随机变量	$\sigma^2 = D(X) = \sum_i (x_i - \mu)^2 p_i$ ，此处，$\mu = E(X) = \sum_i x_i p_i$	标准差σ或$\sqrt{D(X)}$
	连续型随机变量	$\sigma^2 = D(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$ ，此处，$\mu = E(X) = \sum_i x_i p_i$	

随机变量分布类型

类型		相关公式	说明
离散型概率分布	0—1 分布	概率密度函数 $P\{X = k\} = p^k(1-p)^{1-k}$，$k = 0$，$1$ 数学期望：$\mu = p$　　方差：$\sigma^2 = p(1-p)$	① $p_k \geq 0$ $(k = 0, 1)$ ② $\sum_k p_k = 1$ $(i = 0, 1)$
	二项分布 $X \sim B(n, p)$	概率密度函数 $P\{X = k\} = C_n^k p^k (1-p)^{n-k}$　$k = 0, 1,2, \ldots\ldots, n$ 数学期望：$\mu = np$　　方差：$\sigma^2 = np(1-p)$	① $p_k \geq 0$ $(k = 1,2, \ldots\ldots)$ ② $\sum_k p_k = 1$ $(k = 1,2, \ldots\ldots)$
连续型概率分布	正态分布 $X \sim N(\mu, \sigma^2)$ μ为均值； σ^2为方差$(\sigma > 0)$	概率密度函数——公式不要求记忆 $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$　$-\infty < x < +\infty$ 或 $\Phi(x) = P\{X \leq x\} = \int_{-\infty}^{x} p(s)ds$	正态分布与标准正态分布之间的关系： $Z \sim N(0, 1)$，则$\frac{X-\mu}{\sigma} \sim N(0,1)$，即$\frac{X-\mu}{\sigma} = Z$ 或$\Phi(x) = \Phi_0(\frac{x-\mu}{\sigma})$
	标准正态分布 $Z \sim N(0, 1)$	概率密度函数——公式不要求记忆 $p_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$　$-\infty < x < +\infty$	

❖ **数学期望的性质**

1. 设C是常数，则$E(C) = C$；
2. 若k是常数，则$E(kX) = kE(X)$；
3. $E(X+Y) = E(X) + E(Y)$。

❖ **方差的性质及与期望联系**

1. 设C是常数，则$D(C) = 0$；
2. 若k是常数，则$D(kX) = k^2 D(X)$；
3. 若X与Y独立，则$D(X+Y) = D(X) + D(Y)$；
4. 设随机变量X的数学期望$E(X)$存在，若$E[(X - E(X))^2]$存在，则$E[(X - E(X))^2]$为随机变量X的方差，即$D(X) = E[(X - E(X))^2]$ → $D(X) = E(X^2) - [E(X)]^2$；
5. $E(kX + b) = kE(X) + b$，$D(kX + b) = k^2 D(X)$。

总体和样本

名称	内容
总体	研究对象的全体
个体	组成总体的每一个基本元素

样本	从总体中抽取若干个体组成的集合，如X_1，X_2，……，X_n
样本容量	样本中所含个体的个数，如N、n……

抽样方法

简单随机抽样	有放回或不放回
分层抽样	先分层（分组），后随机
整体抽样	抽取一个群，调查所有单位
系统抽样	再随机初始，后规则

样本统计量的分布

类型	相关公式	说明
t−分布 $t\sim t(n)$ 自由度为n	设$X\sim N(0，1)$，$Y\sim\chi^2(n)$，且X，Y相互独立，则称随机变量$t=\frac{X}{\sqrt{Y/n}}$	①$h(t)$关于 t=0(纵轴)对称。 ②$h(t)$的极限为 N(0，1)的密度函数 对称性：$t_{1-\alpha}(n)=-t_\alpha(n)$
χ^2−分布 $\chi^2\sim\chi^2(n)$ 自由度为n $\chi^2\sim N(n，2n)$	设X_1，X_2，……，X_n是来自总体$N(0,1)$的样本，即$Z=\frac{\overline{X}-\mu}{\sigma}\sim N(0,1)$，则随机变量 $\chi^2=X_1^2+X_2^2+\ldots\ldots+X_n^2=\sum_{i=1}^n X_i^2$ 数学期望（均值）：$E(\chi^2)=n$ 方差：$D(\chi^2)=2n$	χ^2－分布的可加性：$\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，且χ_1^2，χ_2^2相互独立，则有$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$。
F−分布 $F\sim F(n_1，n_2)$ 自由度为$(n_1，n_2)$	设$U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且U，V相互独立，则称随机变量 $F=\frac{U/n_1}{V/n_2}$	$F_{1-\alpha}(n_1，n_2)=\frac{1}{F_\alpha(n_2，n_1)}$

正态总体的样本均值和样本方差的分布

总体 X 服从$N(\mu，\sigma^2)$，则样本X_1，X_2，……，X_n服从$N(\mu，\sigma^2/n)$		$E(X)=\mu$，$D(X)=\sigma^2/n$
定理一	设X_1，X_2，……，X_n是来自总体$X\sim N(\mu，\sigma^2)$的一个样本，则样本服从$\overline{X}\sim N(\mu，\sigma^2/n)$，且$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$	
定理二	设X_1，X_2，……，X_n是来自总体$X\sim N(\mu，\sigma^2)$的一个样本	（1）$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$ （2）$\overline{X}$与S^2相互独立
定理三	设X_1，X_2，……，X_n是来自总体$X\sim N(\mu，\sigma^2)$的一个样本，$\overline{X}$与S^2分别为样本均值和样本方差	$\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$
定理四	设X_1，X_2，……，X_n与Y_1，Y_2，……，Y_n分别是具有相同方差的两个正态总体$N(\mu_1，\sigma^2)$，$N(\mu_2，\sigma^2)$的样本，且两个样本相互独立。	①均值分别为： $\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$，$\overline{Y}=\frac{1}{n_2}\sum_{j=1}^{n_2}Y_j$ ②方差分别为： $S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2$ $S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2$
中心极限定理 $\overline{X}\to N(\mu，\frac{\sigma^2}{n})$	设从均值为μ、方差为σ^2；有限的任意一个总体中抽取样本量为 n 的样本，当 n 充分大时，样本均值的抽样分布近似服从均值为μ、方差为$\frac{\sigma^2}{n}$ 的正态分布。	

第三章 参数的点估计

»» **参数估计：**指用样本统计量去估计总体的参数。

»» **点估计：**指用估计量的某个取值直接作为总体参数的估计值。

»» **矩估计法**（简称矩估计）：指用样本各阶原点矩的函数来估计总体各阶原点矩的统一个函数的方法，相应的估计量称为矩估计量。

»» **点估计量的评价标准：**①无偏性：$E(\hat{\theta})=\theta$；②有效性：若有$D(\hat{\theta}_1)<D(\hat{\theta}_2)$，则$\hat{\theta}_1$比$\hat{\theta}_2$更理想；③一致性：$\lim\limits_{n\to\infty}P(|\hat{\theta}-\theta|<\varepsilon)=1$。

		样本容量	σ已知	σ未知
置信区间	单一总体均值	大样本$n\geq 30$	$\overline{X}\pm\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$	$\overline{X}\pm\frac{S}{\sqrt{n}}z_{\frac{\alpha}{2}}$
		小样本$n<30$	$\overline{X}\pm\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$	$\overline{X}\pm\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)$
	两总体均值之差	大样本$n\geq 30$	$\overline{X}-\overline{Y}\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\overline{X}-\overline{Y}\pm z_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
		小样本$n<30$	$\overline{X}-\overline{Y}\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\overline{X}-\overline{Y}\pm t_{\frac{\alpha}{2}}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	单个总体比率	设X_1，X_2，……，X_n为来自总体的一个样本。由中心极限定理知$\frac{\overline{X}-p}{\sqrt{p(1-p)/n}}$		
		$(\overline{X}-z_{\frac{\alpha}{2}}\sqrt{\overline{X}(1-\overline{X})/n}，\overline{X}+z_{\frac{\alpha}{2}}\sqrt{\overline{X}(1-\overline{X})/n}$		
	两个总体比率差	对充分大的n和m（要求$n\overline{X}>5$，$n(1-\overline{X})>5$，$m\overline{Y}>5$，$m(1-\overline{Y})>5$），根据中心极限地理，有p_1-p_2的置信度$1-\alpha$的双侧置信区间为 $\overline{x}-\overline{y}\pm z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}+\frac{\overline{y}(1-\overline{y})}{m}}$		

第四章 假设检验

»» **假设检验的思路：**①对所关心的总体提出某种假设；②从待检验的总体中抽取一个随机样本并获得数据；③根据样本提供的信息判断假设是否成立。

»» **两类错误**

第I类错误（“以真为假”的错误）：由于作出判断的依据是一个样本，即由部分推断整体，因而假设检验不会绝对正确，在实际上H_0为真时仍可能作出拒绝H_0的错误判断。$P\{\text{拒绝}H_0|H_0\text{为真}\}=\alpha$

第II类错误：在实际上H_0不真时，也可能作出接受H_0的错误判断。$P\{\text{接受}H_0|H_0\text{不真时}\}=\beta$

»» **原假设与备择假设的选择**

①原假设一般代表一种久已存在的状态，而备择假设则反映改变；②样本观测值显示所支持的结论应作为备择假设；③应当尽量使后果严重的错误成为第一类错误。

»» **假设检验的一般步骤**

①根据实际问题要求，明确提出原假设H_0与备择假设H_0；②给定显著性水平α以及样本容量n；③确定检验统计量以及拒绝域的形式；④按$P\{\text{拒绝}H_0|H_0\text{为真}\}=\alpha$求出拒绝域；⑤取样，根据样本观测值确定接受还是拒绝H_0。

关于1个正态总体均值的假设检验				
已知方差σ 统计量为$z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$， H_0为真时，$z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$		**双侧检验“$\neq$”**	**左侧检验“$<$”**	**右侧检验“$>$”**
	假设	$H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$	$H_0:\mu\geq\mu_0$ $H_1:\mu<\mu_0$	$H_0:\mu\leq\mu_0$ $H_1:\mu>\mu_0$
	临界值	$z_{\alpha/2}$	$-z_\alpha$	z_α
	拒绝域	$\left\|\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right\|>z_{\alpha/2}$	$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}<-z_\alpha$	$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}>z_\alpha$
未知方差σ	假设	$H_0:\ \mu=\mu_0$ $H_1:\ \mu\neq\mu_0$	$H_0:\mu\geq\mu_0$ $H_1:\mu<\mu_0$	$H_0:\mu\leq\mu_0$ $H_1:\mu>\mu_0$

统计量为$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$， H_0为真时，$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t(0,1)$	临界值	$t_{\alpha/2}(n-1)$	$-t_{\alpha}(n-1)$	$t_{\alpha}(n-1)$
	拒绝域	$\left\|\frac{\overline{x}-\mu_0}{S/\sqrt{n}}\right\|>t_{\frac{\alpha}{2}}(n-1)$	$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}<-t_{\alpha}(n-1)$	$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}>t_{\alpha}(n-1)$

关于一个正态总体方差的假设检验

	选择的统计量为$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}$，$H_0$为真时，$\chi^2=\frac{(n-1)S^2}{\sigma_0^2}\sim\chi^2(n-1)$		
	双侧检验“≠”	**左侧检验“<”**	**右侧检验“>”**
假设	H_0：$\sigma^2=\sigma_0^2$　H_1：$\sigma^2\neq\sigma_0^2$	H_0：$\sigma^2\geq\sigma_0^2$　H_1：$\sigma^2<\sigma_0^2$	H_0：$\sigma^2\leq\sigma_0^2$　H_1：$\sigma^2>\sigma_0^2$
临界值	$\chi^2_{\frac{\alpha}{2}}(n-1)$，$\chi^2_{1-\frac{\alpha}{2}}(n-1)$	$\chi^2_{1-\alpha}(n-1)$	$\chi^2_{\alpha}(n-1)$
拒绝域	$\frac{(n-1)S^2}{\sigma_0^2}>\chi^2_{\frac{\alpha}{2}}(n-1)$或$\frac{(n-1)S^2}{\sigma_0^2}<\chi^2_{1-\frac{\alpha}{2}}(n-1)$	$\frac{(n-1)S^2}{\sigma_0^2}>\chi^2_{1-\alpha}(n-1)$	$\frac{(n-1)S^2}{\sigma_0^2}<\chi^2_{\alpha}(n-1)$

关于 2 个正态总体均值差的假设检验

已知方差σ_1^2，σ_2^2 未知方差σ_1^2，σ_2^2，但n_1，n_2都很大	选择统计量为$z=\frac{\overline{X}-\overline{Y}-a}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$，在$H_0$为真时，$z=\frac{\overline{X}-\overline{Y}-a}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$			
		双侧检验“≠”	**左侧检验“<”**	**右侧检验“>”**
	假设	H_0：$\mu_1-\mu_2=a$ H_1：$\mu_1-\mu_2\neq a$	H_0：$\mu_1-\mu_2\geq a$ H_1：$\mu_1-\mu_2<a$	H_0：$\mu_1-\mu_2\leq a$ H_1：$\mu_1-\mu_2>a$
	临界值	$z_{\alpha/2}$	$-z_{\alpha}$	z_{α}
	拒绝域	$\|z\|>z_{\alpha/2}$	$z<-z_{\alpha}$	$z>z_{\alpha}$
$\sigma_1^2=\sigma_2^2=\sigma^2$，但$\sigma^2$为未知	采用统计量$t=\frac{(\overline{X}-\overline{Y})-a}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$，在$H_0$为真时，$t\sim t(n_1+n_2-2)$，其中$S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}$。			
	临界值	$t_{\alpha/2}(n_1+n_2-2)$	$-t_{\alpha}(n_1+n_2-2)$	$t_{\alpha}(n_1+n_2-2)$
	拒绝域	$\|t\|>t_{\frac{\alpha}{2}}(n_1+n_2-2)$	$t<-t_{\alpha}(n_1+n_2-2)$	$t>t_{\alpha}(n_1+n_2-2)$

关于比率p的假设检验

	当$n\geq 30$，$n\overline{X}>5$，$n(1-\overline{X})>5$时，采用统计量$z=\frac{\overline{X}-p_0}{\sqrt{\overline{X}(1-\overline{X})/n}}$。在$H_0$为真时，z 近似服从$N(0,1)$分布。		
	双侧检验“≠”	**左侧检验“<”**	**右侧检验“>”**
假设	H_0：$p=p_0$　　H_1：$p\neq p_0$	H_0：$p\geq p_0$　H_1：$p<p_0$	H_0：$p\leq p_0$　H_1：$p>p_0$
临界值	$z_{\alpha/2}$	$-z_{\alpha}$	z_{α}
拒绝域	$\|z\|>z_{\alpha/2}$	$z<-z_{\alpha}$	$z>z_{\alpha}$

关于两个总体比率差的假设检验

	对于充分大的n和m（要求$n\overline{X}>5$，$n(1-\overline{X})>5$，$m>5$，$m(1-\overline{Y})>5$），采用统计量$z=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}+\frac{\overline{Y}(1-\overline{Y})}{m}}}$		
	双侧检验“≠”	**左侧检验“<”**	**右侧检验“>”**
假设	H_0：$p_1-p_2=0$ H_1：$p_1-p_2\neq 0$	H_0：$p_1-p_2\geq 0$ H_1：$p_1-p_2<0$	H_0：$p_1-p_2\leq 0$ H_1：$p_1-p_2>0$

第五章 线性回归分析

»» 回归分析研究的主要内容：

①确定变量之间的相互关系和相关程度；②建立回归模型；③检验变量之间的相关程度或回归模型的显著性；

⑤应用回归模型进行估计和预测等。

»» **简单线性回归模型的基本特征**

1. 由于$y_i=(a+bx_i)+\varepsilon_i$，其中$(a+bx_i)$为常量项（不是随机变量），$\varepsilon_i$是随机变量，因此$y_i$也是随机变量；
2. $E(y_i)=E(a+bx_i+\varepsilon_i)=E(a+bx_i)+E(\varepsilon_i)=a+bx_i$；
3. $D(y_i)=D(a+bx_i+\varepsilon_i)=D(\varepsilon_i)=\sigma^2$；
4. 因为$Cov(\varepsilon_i，\ \varepsilon_j)=0$，$Cov(y_i，\ y_j)=0$；
5. $y_i-E(y_i)=y_i-(a+bx_i)=\varepsilon_i$

以上特征表明，y_i是一个随机变量，它来自于$N(a+bx_i，\ \sigma^2)$分布，对于不同的i，y_i的均值随x_i的不同而不同，但方差不i随变化（同方差假设）。

»» **回归参数的最小二乘估计**

解方程组，得到参数a和b的最小二乘估计：$\begin{cases}\hat{a}=\overline{y}-\hat{b}\overline{x}\\ \hat{b}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum_{i=1}^{n}x_i{}^2-n\overline{x}^2}\end{cases}$

其中$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$，$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$。由此得到回归方程：$\hat{y}=\hat{a}+\hat{b}x$

»» **回归参数的最小二乘估计的统计特性：**

1. 线性性：$\hat{a}$和$\hat{b}$都是y_i的线性组合；
2. 无偏性：$\mathrm{E}(\hat{a})=a$，$E(\hat{b})=b$；
3. 方差最小性：a和b的最小二乘估计$\hat{a}$、$\hat{b}$分别是a、b的所有线性无偏估计量中方差最小的。

»» **σ^2的估计，可决系数与相关系数**

①总偏差平方和$S_T=\sum_{i=1}^{n}(y_i-\overline{y})^2$，$S_T$的自由度为$n-1$；

②误差平方和$S_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，$S_E$的自由度为$n-2$；

③回归平方和$S_R=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$，$S_R$的自由度为 1；

④总偏差平方和的分解：$S_T=S_E+S_R$。

由于 $\mathrm{E}(S_E)=(n-2)\sigma^2$，因此$MS_E=\frac{S_E}{n-2}$是$\sigma^2$的一个无偏估计。$MS_E$称为平均误差平方和。类似地，$MS_E=\frac{S_R}{1}$称为平均回归平方和。

定义：$r^2=\frac{S_R}{S_T}$为可决系数。$r=\pm\sqrt{r^2}$为相关系数

相关系数是度量两个变量之间线性关系强度的统计量。$0\le r^2\le 1$，r^2越大拟合程度越高；r^2越小，拟合程度越低。相关系数r不仅反映了y和x之间的线性相关密切程度，也反映了y和x之间的相关方向。

»» **回归效果的显著性检验与方差分析表**

（一）F检验法

采用的统计量为$F=\frac{MS_R}{MS_E}$。当H_0：$b=0$ 为真时，$F\sim F(1，\ n-1)$。

对于给定的显著性水平α，若$F>F_\alpha(1，\ n-1)$，则应拒绝H_0，即认为线性回归效果显著；若$F<F_\alpha(1，\ n-2)$，则接受H_0，认为线性回归效果不显著。

误差来源	自由度	平方和	均方和	F
回归R	1	S_R	MS_R	MS_R/MS_E
误差E	$n-2$	S_E	MS_E	
总和T	$n-1$	S_T		

（二）t检验法

若采用统计量$t=\frac{\hat{b}}{\sqrt{MS_E/l_{xy}}}$，其中$l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})^2$，则当$H_0$：$b=0$为真时，$t\sim t(n-2)$。

对于给定的显著性水平α，$|t|>t_{\frac{\alpha}{2}}(n-2)$时，拒绝$H_0$；$|t|<t_{\frac{\alpha}{2}}(n-2)$时，接受$H_0$。

»» 回归参数的假设检验与区间估计

（一）回归系数b的假设检验与区间估计

已知对于b的最小二乘估计$\hat{b}$，

$$\frac{\hat{b}-b}{s\{\hat{b}\}}\sim t(n-2)$$

其中$s^2\{\hat{b}\}=MS_E/l_{xy}$，$l_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})^2$

检验假设H_0：$b=0$，H_1：$b\neq 0$的检验统计量为$t=\frac{\hat{b}-b}{s\{\hat{b}\}}$。对于给定的显著水平α，如果$|t|>t_{\frac{\alpha}{2}}(n-2)$，则拒绝原假设否则接受原假设。回归参数$b$的置信度为100(1−α)%的置信区间为$\left[\hat{b}\pm t_{\frac{\alpha}{2}}(n-2)s\{\hat{b}\}\right]$

（二）回归参数α的置信区间

回归参数α的置信度为100(1−α)%的置信区间为$\left[\hat{a}\pm t_{\frac{\alpha}{2}}(n-2)s\{\hat{a}\}\right]$

多次统计软件在做线性回归分析时，会同时给出$s\{\hat{a}\}$和$s\{\hat{b}\}$的值，一般把分析结果记为

$$\hat{y}=\hat{a}+\hat{b}x$$
$$(s\{\hat{a}\}s\{\hat{b}\})$$

»» 多元线性回归模型的基本假设：

1. 随机误差项ε_i具有零均值和同方差，即$E(\varepsilon_i)=0\quad i=1,2,.....n$；$D(\varepsilon_i)=\sigma^2\quad i=1,2,.....n$
2. 随机误差项在不同样本点之间是相互独立的，不存在序列关系，即$Cov(\varepsilon_i,\ \varepsilon_j)=0\quad i\neq j\quad i,\ j=1,2,.....n$
3. 随机误差项ε_i应服从正态分布，即$\varepsilon_i\sim N(0,\sigma^2)$；
4. 自变量是确定性变量，且它们之间是相关的；
5. 因变量与自变量之间存在着显著的线性相关关系，即模型是线性的。

»» σ^2的估计，复可决系数

多元线性回归模型

1. 总偏差平方和$S_T=\sum_{i=1}^{n}(y_i-\overline{y})^2$，它的自由度是$n-1$；
2. 误差平方和$S_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，它的自由度为$n-p-1$；
3. 回归平方和$S_R=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$，它的自由度是$p$；
4. 总偏差平方和的分解：$S_T=S_E+S_R$；
5. 平均误差平方和$MS_E=\frac{S_E}{n-p-1}$；
6. 平均回归平方和$MS_R=\frac{S_R}{n}$。

复可系数$R^2=\frac{S_R}{S_T}=1-\frac{S_E}{S_T}$，反映线性拟合程度的指标。$0\leq R^2\leq 1$，且$R^2$越接近于1，说明线性拟合程度越高；$R^2$越接近于0，说明线性拟合程度越低。

调节的可决系数R_a^2的思路：在样本容量一定的情况下，增加解释变量必定使得自由度减少，所以要将残差平方和与总离差平方和分别除以各自的自由度，以剔除变量个数对拟合程度的影响：$R_a^2=1-\frac{S_E/n-p-1}{S_T/n-1}$

»» 多元线性回归模型的假设检验

（一）全检验

需要检验的假设是：H_0：$b_1=b_2=\ldots\ldots=b_p=0$　　H_1：b_1，b_2，……，b_p不全为零

若拒绝H_0，即认为b_1，b_2，……，b_p中至少有一个不为零，说明线性回归模型有意义，否则认为b_1，b_2，……，b_p全都与零没有显著性的差异，即y与x_1，x_2，……，x_p之间不存在显著线性相关关系，说明线性回归模型无意义。这个检验称全检验。

检验上述假设的统计量为$F=\frac{MS_R}{MS_E}$。当H_0为真时，$F\sim F(p,n-p-1)$。对于给定的显著性水平α，若$F>F_\alpha(p,n-p-1)$，则应拒绝H_0，即认为线性回归效果显著；若$F<F_\alpha(p,n-p-1)$，则接受H_0，认为线性回归效果不显著。

误差来源	自由度	平方和	均方和	F
回归R	p	S_R	MS_R	MS_R/MS_E
误差E	$n-p-1$	S_E	MS_E	
总和T	$n-1$	S_T		

（二）偏检验

偏检验要检验的假设是：对于某一$K(0\le K\le p)$

$$H_0：b_k=0 \quad H_1：b_k\neq 0$$

这一检验的目的是检验y与新纳入模型的x_k是否存在显著的偏相关或净相关关系，即x_k是否应当保留在回归模型中的问题。如果H_0成立，x_k不具有预测y的能力，应该将它从模型中剔除。反之，若H_0不成立，$b_k\neq0$，说明y与x_k之间存在显著的偏相关关系，x_k具有预测y的能力，应当保留在模型中。

当H_0为真时，$t\sim t(n-p-1)$。对于给定的显著性水平α，如果$|t|>t_{\frac{\alpha}{2}}(n-p-1)$，则拒绝假设。

»» 估计与预测

1. 对于给定的$x_0=(x_{10}，x_{20}，\ldots\ldots，x_{p0})$，$y_0$的点预测为$\hat{y}_0=\hat{b}_0+\hat{b}_1x_{10}+\hat{b}_2x_{20}+\ldots\ldots+\hat{b}_px_{p0}$
2. 回归系数b_x的置信度为1−α的置信区间为$\left[\hat{b}_k\pm t_{\frac{\alpha}{2}}(n-p-1)S(\hat{b}_k)\right]$

»» 多项式回归模型

则多项式回归模型可以写成如下多元线性回归模型：

$$\begin{cases} y_i=b_0+b_1x_{1i}+b_2x_{2i}+\ldots\ldots+b_px_{pi}+\varepsilon_i \quad i=1,2,\ldots\ldots n \\ \varepsilon_i\sim N(0,\sigma^2) \text{ 相互独立} \end{cases}$$

»» 非线性---散点图---如果总体是非线性的，则存在下列问题：

1. 回归参数b的估计量$\hat{b}$不是有效估计量；
2. 无法准确地估计σ^2；
3. 有关回归模型的推断、检验和应用都会失去准确性。

»» 异方差性---如果样本数据存在异方差现象，则存在以下问题：

1. 回归系数的最小二乘估计不是具有有效性；
2. 无法准确地确定回归参数的置信区间；
3. 假设检验的结论无效。

»» 序列相关性---如果随机误差项之间存在着序列相关，则存在以下问题：

1. 回归参数的最小二乘估计虽然是无偏的，但不是有效的；
2. 回归效果的显著性检验不再有效；
3. 预测失去准确性。

»» **非正态性---直方图**

一般情况下，随机误差项稍微偏差正态分布，不会产生严重问题。但是，如果严重偏差正态分布的话，那么正态假设条件下的统计推断、估计和预测就失去了意义。

»» **多重共线性：**是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。

»» **多重共线性带来的问题有：**

①一般情况下，严格的共线性不多见，经常存在的是近似共线性；

②多重共线性的存在也会导致错误的假设检验结论。因变量与全体或部分自变量之间本来十分显著的相关关系；

③可能会由于自变量之间的相关关系而检验不出来；

④增加或减少一个自变量将会导致回归参数的估计值发生大的变化，甚至发生符号变化。

»» **常用的检查多重共线性的方法有：**

1. 散点图法：由自变量之间的散点图观察它们之间是否存在显著的相关性；

2. 计算自变量之间的相关系数$r_{x_ix_j}$，$|r_{x_ix_j}|$越接近于 1 时，说明自变量x_i和x_j之间的高度相关。

第六章 时间序列分析

»» **时间序列：**是指将某种现象某一个统计指标在不同时间上的各个数值，按时间先后顺序排列而形成的序列。时间序列法是一种定量预测方法，亦称简单外延方法。

►构成要素：

1. 截面数据描述：在相同或近似相同时间点上收集的数据，描述事物在某一时刻的变化情况，即横向数据；

2. 时间序列数据描述：在不同时间上收集到的数据，描述事物在一定的时间范围内的变化情况，即纵向数据；

平行数据：是截面数据与时间序列数据的组合。

影响时间序列的因素：① 长期趋势——T；② 季节变动——S；③ 循环波动——C；④ 不规则波动——I。

►常用的时间序列的分解模型为乘法模型和加法模型。

乘法模型：$Y_i = T_i \times S_i \times C_i \times I_i$，即认为四个因素之间是相互影响的。

加法模型：$Y_i = T_i + S_i + C_i + I_i$，即认为时间序列的变动是四个因素的总和。

»» **滑动平均法**

设滑动平均的间隔长度为m，则滑动平均法序列为

$$\overline{Y}_1 = \frac{Y_1 + Y_2 + \ldots\ldots + Y_m}{m};\ \overline{Y}_2 = \frac{Y_2 + Y_3 + \ldots\ldots + Y_{m+1}}{m};\ \overline{Y}_i = \frac{Y_i + Y_{i+1} + \ldots\ldots + Y_{i+m-1}}{m}$$

其中m为大于 1 的正整数。

滑动平均的目的在于消除原时间序列数据中的短期（季节性和不规则）波动，因此滑动的间隔长度应适中。若时间序列是月份资料，m应采用 12；若是季度资料，m应为 4；若时间序列具有周期性变动，则m应为周期长度。

中心化滑动平均：滑动平均后的趋势值应放在各滑动项的中间位置。

»» **按月（季）平均法**

定义：根据时间序列，通过简单平均来计算季节指数的一种既简单又常用的方法。

步骤：① 用原始数据计算出同月（或同季）的平均数；② 将各同月（或同季）平均数除以数列的总平均数，得到的便是季节指数 S。

$$\text{季节指数 S} = \frac{\text{同月（季）平均数}}{\text{总月（季）平均数}} \times 100\%$$

»» **滑动平均趋势剔除法（趋势剔除法）**

假定时间序列的乘法模型中各时点的不规则波动I相互独立。因此，在对数据进行间隔长度m=12（如果是月资

料）或m=4（如果是季度资料）的滑动平均后，就能够消除季节波动和不规则波动的影响，从而得到滑动平均趋势值$T \times C$。将原数列除以滑动平均趋势值$T \times C$，得到的百分比称为滑动平均百分比，即$\frac{T \times C \times S \times I}{T \times C} = S \times I$

对其进行同月（或同季）平均，以便消除不规则波动因素 I 的影响，从而得到季节变动S，最后将其调整为以 100 为基准的季节指数。

»» 滑动平均趋势剔除法测定季节变动的步骤：

① 对原数据进行 12 个月或 4 个季度的滑动平均，求出滑动平均趋势；

② 将各实际值除以相应的趋势值，得到滑动平均百分比；

③ 将滑动平均百分比重新按月（或按季）排列，求出同月（或同季）平均数；

④ 将出同月（或同季）平均数除以总平均数，得到季节指数 S。

»» 季节调整

测定了季节变动之后，可以将它从时间序列中剔除，从而观察和分析时间序列的其他特征。用乘法模型，将原序列除以相应的季节指数，便得到调整的时间序列：

$$\frac{Y}{S} = \frac{T \times S \times C \times I}{S} = T \times C \times I$$

它反映的是在没有季节因素影响情况下，时间序列之变化趋势。

»» 用剩余法测定循环波动的做法：

①求出季节变动指数S

②在原时序数列中消除季节因素之影响，计算公式：

$$\frac{T \times S \times C \times I}{S} = T \times C \times I$$

③计算长期趋势值，并在无季节影响之时序数列中消除长期趋势的影响，计算公式：

$$\frac{T \times C \times I}{T} = C \times I$$

④用滑动平均法对时序数列$C \times I$进行滑动平均，消除不规则波动的影响，得到循环波动值，通常用百分数表示。