统计学公式 Statistics formula

包括：数据的概括性度量 Summary measure of data，概率与概率分布 Probability and probability distribution，抽样分布 **sampling distribution**，参数估计 **parameter estimation**，假设检验 **hypothesis testing**，列联分析 **contingency analysis**，方差分析 **variance analysis**，相关与线性回归分析 Correlation and linear regression analysis，时间序列分析和预测 Time series analysis and forecasting，指数 index。

数据的概括性度量 Summary measure of data

名称	公式	名称	公式
中位数 Median	$M_e=\begin{cases} x_{\left(\frac{n+1}{2}\right)} & n为奇数 \\ \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right\} & n为偶数 \end{cases}$	标准序 Z-score	$z_i=\frac{x_i-\bar{x}}{s}$
样本平均数 Sample Mean	$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$	四分位差 interquartile range	$Q_d=Q_U-Q_L$
样本加权平均数 weighted mean	$\bar{x}=\frac{\sum_{i=1}^{k}M_i f_i}{n}$	异众比率 variation ratio	$V_r=\frac{\sum f_i-f_m}{\sum f_i}=1-\frac{f_m}{\sum f_i}$
几何平均数 geometric mean	$G_m=\sqrt[n]{x_1\times x_2\times\cdots\times x_n}=\sqrt[n]{\prod_{i=1}^{n}x_i}$	极差 Range	$R=\max(x_i)-\min(x_i)$
平均差 average deviation	$M_d=\frac{\sum_{i=1}^{n}\lvert x_i-\bar{x}\rvert}{n}$	离散系数 coefficient of variation	$v_s=\frac{s}{\bar{x}}$
加权平均差 Weighted average deviation	$M_d=\frac{\sum_{i=1}^{k}\lvert M_i-\bar{x}\rvert f_i}{n}$	加权样本方差 Weighted sample variance	$s^2=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^2 f_i}{n-1}$
样本方差 Sample variance	$s^2=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^2 f_i}{n-1}$	样本标准方差 Sample standard deviation	$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
数据偏态系数 skewness coefficient	$SK=\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$	分组数据的偏态系数 skewness coefficient of grouped data	$SK=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^3 f_i}{ns^3}$
峰态系数 kurtosis coefficient	$K=\frac{n(n+1)\sum(x_i-\bar{x})^4-3\left(\sum(x_i-\bar{x})^2\right)^2(n-1)}{(n-1)(n-2)(n-3)s^4}$	分组数据的峰态系数 kurtosis coefficient of grouped data	$K=\frac{\sum_{i=1}^{k}(M_i-\bar{x})^4 f_i}{ns^4}-3$

概率与概率分布 Probability and probability distribution

名称	公式	名称	公式
概率的古典定义 Classical probability	$P(A)$ =(事件 A 的事件个数 m)/(样本空间事件个数 n)	概率的统计定义 Statistical probability	$P(A)=\frac{m}{n}=p$
互斥事件 mutual exclusive	$P(A\cup B)=P(A)+P(B)$	n 个两两互斥事件	$P(A_1\cup A_2\cup\cdots\cup A_n)=P(A_1)+P(A_2)+\cdots+P(A_n)$
事件 A 与其补事件 A^c Complement	$P(A)+P(A^c)=1$	两个任意事件 any pair event	$P(A\cup B)=P(A)+P(B)-P(A\cap B)$
概率的乘法法则 multiplication rule	$P(AB)=P(B)P(A/B)$ or $P(AB)=P(A)P(B/A)$	全概率公式 Law of Total Probability	$P(B)=\sum_{i=1}^{n}P(A_i)P(B\mid A_i)$
两个独立事件 independent	$P(A\cap B)=P(A)P(B)$	n 个相互独立事件 each other independent	$P(A_1\cap A_2\cap\cdots\cap A_n)=P(A_1)*P(A_2)\cdots*P(A_n)$
贝叶斯概率公式 Bayesian probability formula	$P(A_i\mid B)=\frac{P(A_i)P(B\mid A_i)}{\sum_{j=1}^{n}P(A_j)P(B\mid A_j)}$	泊松分布的概率 poisson distribution	$P(X)=\frac{\lambda^x e^{-\lambda}}{x!}$
离散型随机变量期望 The expectation of a discrete random variable	$E(X)=x_1p_1+x_2p_2+\cdots+x_np_n=\sum_{i=1}^{n}x_ip_i$	离散型随机变量方差 The variance of a discrete random variable	$\sigma^2=D(X)=\sum_{i=1}^{\infty}[x_i-E(X)]^2\cdot p_i$
连续型随机变量期望 Expectation of continuous random variable	$E(X)=\int_{-\infty}^{+\infty}xf(x)d(x)$	连续型随机变量方差 The variance of continuous random variable	$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2f(x)d(x)=\sigma^2$
正态分布的概率密度函数 normal distribution PDF	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	标准正态分布率密度函数 The standard normal distribution PDF	$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
标准正态分布分布函数 CDF	$\Phi(x)=\int_{-\infty}^{x}\varphi(t)d(t)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$	正态随机变量的 a≤X≤b 概率	$P(a\le X\le b)=\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})$
二项分布 Binomial distribution	$P\{X=x\}=C_n^x p^x q^{n-x}$	二项分布期望和方差 Expectation and variance of binomial distribution	E(x)=np and Var(x)=np(1-p)

抽样分布 sampling distribution

$\overline{X}$ 抽样分布期望和方差 Expectation and variance of the sampling distribution of the sample mean	$E(\overline{X})=\mu$ $D(\overline{X})=\frac{\sigma^2}{n}$	两个样本均值之差抽样分布的期望和方差 Expectation and variance of the sampling distribution of the difference of two population mean	$E(\overline{X}_1-\overline{X}_2)=\mu_1-\mu_2$ $D(\overline{X}_1-\overline{X}_2)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_2}$

参数估计 parameter estimation

<table>
<tr><td rowspan="3">一个总体均值的置信区间
confidence interval</td><td>正态总体，σ 已知
normal population known σ</td><td colspan="3">$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$</td></tr>
<tr><td>σ 未知，大样本
Unknown σ, large sample</td><td colspan="3">$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$</td></tr>
<tr><td>正态总体，σ 未知，小样本
normal population , Unknown σ, small sample</td><td colspan="3">$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$</td></tr>
<tr><td colspan="2">一个总体比例的置信区间</td><td colspan="3">$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$</td></tr>
<tr><td rowspan="5">两个总体均值之差的置信区间
Confidence interval for the difference of two population mean</td><td>正态总体或独立大样本，σ_1和σ_2 已知
normal population or Independent large sample, known σ_1, σ_2</td><td colspan="3">$(\bar{x}_1-\bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$</td></tr>
<tr><td>独立大样本，σ_1和σ_2 未知
Independent large sample, unknown σ_1, σ_2</td><td colspan="3">$(\bar{x}_1-\bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$</td></tr>
<tr><td>独立小样本，σ_1和σ_2 未知但相等 Independent small sample, unknown σ_1, σ_2, assume $\sigma_1=\sigma_2$</td><td colspan="3">$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2}(n_1+n_2-2)\sqrt{s_p^2\frac{1}{n_1}+\frac{1}{n_2}}$, $s_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$</td></tr>
<tr><td>独立小样本，σ_1和σ_2 未知且不相等，$n_1 \neq n_2$
Independent small sample, unknown σ_1, σ_2 and $\sigma_1 \neq \sigma$, $n_1 \neq n_2$</td><td colspan="3">$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2}(\upsilon)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$, $\upsilon=\frac{S_1^2/n_1+S_2^2/n_2}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^2}{n_2^2(n_2-1)}}$</td></tr>
<tr><td>配对大样本
Paired large sample</td><td>$\bar{d} \pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}}$</td><td>配对小样本
Paired small sample</td><td>$\bar{d} \pm t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}$</td></tr>
<tr><td colspan="2">两个总体比例之差的区间估计
Confidence interval for the difference of two proportions</td><td colspan="3">$(p_1-p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$</td></tr>
<tr><td colspan="2">两个总体方差比的置信区间
Confidence interval for the difference of two variance</td><td colspan="3">$\frac{s_1^2/s_2^2}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_{1-\alpha/2}}$</td></tr>
<tr><td>估计总体均值时的样本容量
Sample size for estimating the population mean</td><td>$n=\frac{(z_{\alpha/2})^2\sigma^2}{E^2}$</td><td colspan="2">估计总体比例时的样本容量 Sample size for estimating the population proportion</td><td>$n=\frac{(z_{\alpha/2})^2\cdot\pi(1-\pi)}{E^2}$</td></tr>
</table>

假设检验 hypothesis testing

总体均值检验的统计量 The statistic of the population mean test	正态总体，σ 已知 normal population known σ	$z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	正态总体，σ 未知，小样本 normal population, unknown σ ,small sample	$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$	
	σ 未知，大样本 Unknown σ ,large sample	$z=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$			
总体比例检验的统计量 The statistic of the population proportion test	$z=\dfrac{p-\pi_0}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}$	总体方差检验的统计量 The statistic of the population variance test	$\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}$		
两个总体均值之差的检验统计量 A test statistic for the difference between the means of two populations	独立大样本 Independent large sample	$z=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$			
	独立小样本，σ 1 和σ 2 未知但相等 Independent small sample, unknown σ1, σ2 , assume σ1=σ2	$t=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$			
	独立小样本，σ 1 和σ 2 未知且不相等 Independent small sample, unknown σ1, σ2 , assume σ1≠σ2	$t=\dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}\sim t(f)$			
两个总体比例之差的检验统计量 Test statistic for the difference between the proportions of two populations	$z=\dfrac{(p_1-p_2)}{\sqrt{p(1-p)(\dfrac{1}{n_1}+\dfrac{1}{n_2})}}$	两个总体方差比的检验统计量 Test statistic for the ratio of variance of two populations	$F=\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$		

列联分析 contingency analysis 独立性检验

χ2 统计量 statistic	$\chi^2=\sum\dfrac{(f_0-f_e)^2}{f_e}$ f_0:观测值，f_e：期望值	列联相关系数 contingency coefficient	$c=\sqrt{\dfrac{\chi^2}{\chi^2+n}}$
φ 统计量 statistic	$\varphi=\sqrt{\chi^2/n}$	ν 相关系数 correlation	$v=\sqrt{\dfrac{\chi^2}{n\times\min[(R-1),(C-1)]}}$

方差分析 variance analysis

组间方差 Groups mean square	$MSA=\dfrac{组间平方和}{自由度}=\dfrac{SSA}{k-1}$	方差分析的检验统计量 Test statistics for ANOVA	$F=\dfrac{MSA}{MSE}\sim F(k-1,n-k)$
组内方差 Error mean square	$MSE=\dfrac{组内平方和}{自由度}=\dfrac{SSE}{n-k}$	关系强度的测量 Measurement of Correlation	$R^2=\dfrac{SSA(组间SS)}{SST(总SS)}$
多重比较的 LSD multiple comparisons LSD	$LSD=t_{\alpha/2}\sqrt{MSE(\dfrac{1}{n_i}+\dfrac{1}{n_j})}$		

相关与线性回归分析 Correlation and linear regression analysis

相关系数 correlation	$r_{xy}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}=\frac{\sum xy}{(n-1)s_xs_y}$		
相关系数检验的统计量 Correlation coefficient test statistic	$t=\lvert r\rvert\sqrt{\frac{n-2}{1-r^2}}\sim t(n-2)$		
一元线性回归估计模型 Unary linear regression estimation model	$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$	截距 Intercept	$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$
多元线性回归估计模型 Multiple linear regression estimation model	$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_kx_k$		
回归方程的斜率（回归系数） The slope of regression equation (regression coefficient)	$\hat{\beta}_1=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$	判定系数 R Square	$R^2=\frac{SSR}{SST}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$
一元回归估计标准误差 Standard errors of unary regression estimates	$s_e=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}$		
多元回归估计标准误差 Multiple regression estimates standard errors	$s_e=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-k-1}}=\sqrt{\frac{SSE}{n-k-1}}=\sqrt{MSE}$		
一元回归线性关系检验统计量 Unary regression line test statistic	$F=\frac{SSR/1}{SSE/(n-2)}=\frac{MSR}{MSE}\sim F(1,n-2)$		
多元回归线性关系检验统计量 Multiple regression line test statistic	$F=\frac{SSR/k}{SSE/(n-k-1)}\sim F(k,n-k-1)$		
一元回归系数检验统计量 Unary regression coefficient test statistic	$t=\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\sim t(n-1)$	多元回归系数检验统计量 Multiple regression coefficient test statistic	$t_i=\frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}\sim t(n-k-1)$
y 的平均值的置信区间 The confidence interval for the mean of y	$\hat{y}_0\pm t_{\alpha/2}s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$		
y 的个别值的预测区间 The prediction interval for individual values of y	$\hat{y}_0\pm t_{\alpha/2}s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$		
残差 residual	$e_i=y_i-\hat{y}_i$	标准化残差 Standardization of residuals	$z_{e_i}=\frac{e_i}{s_e}=\frac{y_i-\hat{y}_i}{s_e}$
修正的多重判定系数 adjusted R square (multiple coefficient of determination)	$R_a^2=1-(1-R^2)\times\frac{n-1}{n-k-1}$		

时间序列分析和预测 Time series analysis and forecasting

平均增长率 Average growth rate	$\overline{G}=\sqrt[n]{\frac{Y_1}{Y_0}\times\frac{Y_2}{Y_1}\times\cdots\times\frac{Y_n}{Y_{n-1}}}-1=\sqrt[n]{\frac{Y_n}{Y_0}}-1$		
环比增长率 Chain growth rate	$G_i=\frac{Y_i}{Y_{i-1}}-1$	定基增长率 Growth rate of a fixed base	$G_i=\frac{Y_i-Y_0}{Y_0}=\frac{Y_i}{Y_0}-1$
年度化增长率 Annualized growth rate	$G_A=(\frac{Y_i}{Y_{i-1}})^{m/n}-1$	平均预测误差 mean prediction error	$ME=\frac{\sum_{i=1}^{n}(Y_i-F_i)}{n}$
均方预测误差 mean square prediction error	$MSE=\frac{\sum_{i=1}^{n}(Y_i-F_i)^2}{n}$	平均百分比预测误差 Mean percentage forecast error	$MPE=\frac{\sum\left(\frac{Y_i-F_i}{Y_i}\times 100\right)}{n}$
简单平均法预测 Simple Average forecast Method	$F_{t+1}=\frac{1}{t}(Y_1+Y_2+\cdots+Y_t)=\frac{1}{t}\sum_{i=1}^{t}Y_i$		
移动平均法预测 moving average forecast method	$F_{t+1}=\overline{Y}_t=\frac{Y_{t-k+1}+Y_{t-k+2}+\cdots+Y_{t-1}+Y_t}{k}$		
指数平滑法预测 index smoothing forecasting method	$F_{t+1}=\alpha Y_t+(1-\alpha)F_t$		
线性趋势方程的截距和斜率 The intercept and slope of the linear trend equation	$\begin{cases} b=\frac{n\sum tY-\sum t\sum Y}{n\sum t^2-(\sum t)^2} \\ a=\overline{Y}-b\overline{t} \end{cases}$	二次曲线的标准方程组 Standard system of equations for quadratic curve	$\begin{cases} \sum Y=na+b\sum t+c\sum t^2 \\ \sum tY=a\sum t+b\sum t^2+c\sum t^3 \\ \sum t^2Y=a\sum t^2+b\sum t^3+c\sum t^4 \end{cases}$
指数曲线的标准方程组 Standard system of equations for exponent curve	$\begin{cases} \sum\lg Y=n\lg b_0+\lg b_1\sum t \\ \sum t\lg Y=\lg b_0\sum t+\lg b_1\sum t^2 \end{cases}$	修正指数曲线的未知数 modified exponential curve	$\begin{cases} b_1=\left(\frac{S_3-S_2}{S_2-S_1}\right)^{\frac{1}{m}} \\ b_0=(S_2-S_1)\frac{b_1-1}{b_1(b_1^m-1)^2} \\ K=\frac{1}{m}\left(S_1-\frac{b_0b_1(b_1^m-1)}{b_1-1}\right) \end{cases}$
Gompertz 曲线的待定系数 Undetermined coefficients of the Gompertz curve	$\begin{cases} b_1=\left(\frac{S_3-S_2}{S_2-S_1}\right)^{\frac{1}{m}} \\ \lg b_0=(S_2-S_1)\frac{b_1-1}{b_1(b_1^m-1)^2} \\ \lg K=\frac{1}{m}\left(S_1-\frac{b_1(b_1^m-1)}{b_1-1}\times\lg b_0\right) \end{cases}$	Logistic 曲线待定系数 Undetermined coefficients of the Logistic curve	$\begin{cases} b_1=\left(\frac{S_3-S_2}{S_2-S_1}\right)^{\frac{1}{m}} \\ b_0=(S_2-S_1)\frac{b_1-1}{b_1(b_1^m-1)^2} \\ K=\frac{1}{m}\left(S_1-\frac{b_0b_1(b_1^m-1)}{b_1-1}\right) \end{cases}$
k 阶曲线方程 Curve equation of order k	$\hat{Y}_t=b_0+b_1t+b_2t^2+\cdots+b_kt^k$		

指数 index。

名称	公式	名称	公式
加权综合销售量指数	$I_q=\frac{\sum q_1 p}{\sum q_0 p}$	加权综合价格指数	$I_p=\frac{\sum q p_1}{\sum q p_0}$
拉式数量指标指数	$I_q=\frac{\sum q_1 p_0}{\sum q_0 p_0}$	拉式质量指标指数	$I_p=\frac{\sum q_0 p_1}{\sum q_0 p_0}$
帕式数量指标指数	$I_q=\frac{\sum q_1 p_1}{\sum q_0 p_1}$	帕式质量指标指数	$I_p=\frac{\sum q_1 p_1}{\sum q_1 p_0}$
基期总量加权的加权平均数量指标	$A_q=\frac{\sum \frac{q_1}{q_0} q_0 p_0}{\sum q_0 p_0}$	基期总量加权的加权平均质量指标	$A_p=\frac{\sum \frac{p_1}{p_0} q_0 p_0}{\sum q_0 p_0}$
报告期总量加权的加权平均数量指标	$H_q=\frac{\sum q_1 p_1}{\sum \frac{q_0}{q_1} q_1 p_1}$	报告期总量加权的加权平均质量指标	$H_p=\frac{\sum q_1 p_1}{\sum \frac{p_0}{p_1} q_1 p_1}$
总量指数	$v=\frac{\sum p_1 q_1}{\sum p_0 q_0}$	总量指数体系	$\frac{\sum q_1 p_1}{\sum q_0 p_0}=\frac{\sum q_1 p_0}{\sum q_0 p_0}\times\frac{\sum q_1 p_1}{\sum q_1 p_0}$