概率基础

常见分布

分布	期望E	方差D	表达式
两点分布（伯努利分布、0-1分布）	p	$p(1-p)$	
二项分布(n重伯努利分布) $B(n,p)$	np	$np(1-p)$	$p=C_n^x p^x(1-p)^{n-x}$
泊松分布 $p(\lambda)$	λ	λ	$p=\frac{\lambda^x}{x!}e^{-\lambda}$
均匀分布 $U[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$p=\frac{x_2-x_1}{b-a}$
指数分布 $E(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$p=\frac{1}{e^{x_1\lambda}}-\frac{1}{e^{x_2\lambda}}$
正态分布 $N(\mu,\sigma^2)$	μ	σ^2	$\Phi(\frac{x_2-\mu}{\sigma})-\Phi(\frac{x_1-\mu}{\sigma})$

其中：

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

正态函数

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

期望

- 离散：$E(x)=\sum x_i p_i$
- 连续：$E(x)=\int_{-\infty}^{+\infty}xf(x)dx$

公式：

$E(C)=C$
$E(Cx)=CE(x)$
$E(x\pm y)=E(x)\pm E(y)$ —— X, Y独立
$E(xy)=E(x)E(y)$—— X, Y独立

方差

- 离散：$D(x)=\sum[x_i-E(x)]^2\cdot p_i$
- 连续：$D(x)=E(x^2)-E^2(x)$

公式：

$D(C)=0$
$D(Cx)=C^2D(x)$
$D(x\pm y)=D(x)\pm D(y)$

数理统计

抽样分布

正态分布的性质

- $X\sim N(\mu,\sigma^2)$，则$a+bx\sim N(a+b\mu,b^2\sigma^2)$
- $X\sim N(\mu,\sigma_1^2);Y\sim N(\mu,\sigma_2^2)$，则$X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$

卡方分布

$x_i\sim N(0,1)\quad\rightarrow\quad {x_1}^2+{x_2}^2+{x_3}^2+\ldots+{x_n}^2\sim\chi^2(n)$

$\chi^2\sim N(n,2n)$

当$n>45$时，$\chi_\alpha^2(n)=n+\sqrt{2n}\cdot u_\alpha$

T分布

$X\sim N(0,1);Y\sim\chi^2(n)$，则

$$T=\frac{X}{\sqrt{Y/n}}\sim t(n)$$

性质：

$$t_{1-\alpha}(n)=-t_\alpha(n)$$

因为t分布为偶函数，当$n>45$时，由$f_n(t)\rightarrow\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$，推出$t_\alpha(n)=u_\alpha$

F分布

$X\sim\chi^2(n_1);Y\sim\chi^2(n_2)$，则

$$F=\frac{X/n_1}{Y/n_2}\sim F(n1,n2)$$

性质：

$$F_{1-\alpha}(n1,n2)=\frac{1}{F_\alpha(n2,n1)}$$

[例] $x_1\dots x_9$**来自**$X\sim N(0,3^2)$，$Y_1\dots Y_9$**来自**$Y\sim N(0,3^2)$，**求**$U=\frac{x_1+\dots+x_9}{\sqrt{Y_1^2+\dots Y_9^2}}$**的分布**

$x_i\sim N(0,9)$，则$\sum_{i=1}^9 x_i\sim N(0,81)$

则$\frac{\left(\sum_{i=1}^9 x_i\right)-0}{\sqrt{81}}\sim N(0,1)$（正态分布标准化）

则$\frac{\sum_{i=1}^9 x_i}{9}=V$

又$Y_i\sim N(0,9)$，则$(\frac{Y_i}{3})^2\sim\chi^2(1)$，(正态分布标准化)

则$\sum_{i=1}^9(\frac{Y_i}{3})^2\sim\chi^2(9)=W$

最终：

$$U=\frac{9V}{\sqrt{9W}}=\frac{9V}{3\sqrt{W}}=\frac{3V}{\sqrt{W}}=\frac{V}{\sqrt{W/9}}\sim t(9)$$

[例]$X\sim N(0,1)$，有$x_1\sim x_6;Y=(x_1+x_2+x_3)^2+(x_4+x_5+x_6)^2$，试确定$C$，使得$CY\sim\chi^2$

解：

$$x_1+x_2+x_3\sim N(0,3)\quad;\quad x_4+x_5+x_6\sim N(0,3)$$

$$\text{标准化：}\quad\frac{(x_1+x_2+x_3)-0}{\sqrt{3}}\sim N(0,1)\,;\,\frac{(x_4+x_5+x_6)-0}{\sqrt{3}}\sim N(0,1)$$

$$\text{根据独立性，有：}$$

$$\left[\frac{1}{\sqrt{3}}(x_1+x_2+x_3)\right]^2+\left[\frac{1}{\sqrt{3}}(x_4+x_5+x_6)\right]^2=\frac{1}{3}Y\sim\chi^2(2)$$

$$\text{故：}\ C=\frac{1}{3}$$

假设检验

> 假设检验 与 区间估计，**先判断是 正态母体 还是 大样本母体**
>
> - 正态母体：
> - 题目说了服从正态
> - 数量n，小于50
> - 大样本母体：
> - 没说是正态分布，没说是啥分布，说了个其他什么分布二项分布之类的，有时候要自己

推是啥分布
- 数量n，大于等于50

正态假设检验

都用S^*，不用S。

$$S^* = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

假设一个H0，写出它的相反事件H1，根据公式计算**拒绝域**，落在拒绝域就拒绝H0，否则接收H0

单参数

双侧

假设H0	H1	已知条件	拒绝域
$\mu = \mu_0$	$\mu \neq \mu_0$	已知σ	$\lvert\frac{x-\mu_0}{\sigma/\sqrt{n}}\rvert \geq \mu_{\alpha/2}$
		未知σ	$\lvert\frac{x-\mu_0}{S^*/\sqrt{n}}\rvert \geq t_{\alpha/2}(n-1)$
$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	-	$\frac{(n-1)S^{*2}}{\sigma_0^2} \geq \chi^2_{2/\alpha}(n-1)$或$\frac{(n-1)S^{*2}}{\sigma_0^2} \leq \chi^2_{1-2/\alpha}(n-1)$

单侧

假设H0	H1	已知条件	拒绝域
$\mu \leq \mu_0$	$\mu > \mu_0$	已知σ	$\frac{x-\mu_0}{\sigma/\sqrt{n}} \geq \mu_\alpha$
		未知σ	$\frac{x-\mu_0}{S^*/\sqrt{n}} \geq t_\alpha(n-1)$
$\mu \geq \mu_0$	$\mu < \mu_0$	已知σ	$\frac{x-\mu_0}{\sigma/\sqrt{n}} \leq -\mu_\alpha$
		未知σ	$\frac{x-\mu_0}{S^*/\sqrt{n}} \leq -t_\alpha(n-1)$
$\sigma \leq \sigma_0$	$\sigma > \sigma_0$	-	$\frac{(n-1)S^{*2}}{\sigma_0^2} \geq \chi^2_\alpha(n-1)$
$\sigma \geq \sigma_0$	$\sigma < \sigma_0$	-	$\frac{(n-1)S^{*2}}{\sigma_0^2} \leq \chi^2_{1-\alpha}(n-1)$

双参数

$$S_w=\sqrt{\frac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}}$$

> 双侧

假设H0	H1	已知条件	拒绝域
$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	已知σ	$\lvert\frac{x_1-x_2-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\rvert\geq u_{\alpha/2}$
		未知σ	$\lvert\frac{x_1-x_2-\delta}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\rvert\geq t_{\alpha/2}(n_1+n_2-2)$
$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	-	$\frac{s_1^{*2}}{s_2^{*2}}\geq F_{\alpha/2}(n_1-1,n_2-1)$ 或 $\frac{s_1^{*2}}{s_2^{*2}}\leq F_{1-\alpha/2}(n_1-1,n_2-1)$

> 单侧

假设H0	H1	已知条件	拒绝域
$\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	已知σ_1,σ_2	$\frac{x_1-x_2-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\geq u_\alpha$
		未知σ_1,σ_2	$\frac{x_1-x_2-\delta}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\geq t_\alpha(n_1+n_2-2)$
$\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	已知σ_1,σ_2	$\frac{x_1-x_2-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leq -u_\alpha$
		未知σ_1,σ_2	$\lvert\frac{x_1-x_2-\delta}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\rvert\leq -t_\alpha(n_1+n_2-2)$
$\sigma_1^2\leq\sigma_2^2$	$\sigma_1^2>\sigma_2^2$	-	$\frac{s_1^{*2}}{s_2^{*2}}\geq F_\alpha(n_1-1,n_2-1)$
$\sigma_1^2\geq\sigma_2^2$	$\sigma_1^2<\sigma_2^2$	-	$\frac{s_1^{*2}}{s_2^{*2}}\leq F_{1-\alpha}(n_1-1,n_2-1)$

正态区间估计

> 求均值μ或σ^2的置信区间，置信度为$1-\alpha$
>
> - **实际上就是上面这一堆公式做一下移项，列出一些常用的置信区间公式，剩下的，根据对应情况移项即可**

要估计的值	已知条件	置信区间
μ	已知σ	$(\overline{x}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2},\overline{x}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2})$
	未知σ	$(\overline{x}-\frac{S^*}{\sqrt{n}}t_{\alpha/2}(n-1),\overline{x}+\frac{S^*}{\sqrt{n}}t_{\alpha/2}(n-1))$
σ^2	-	$(\frac{(n-1)S^{*2}}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^{*2}}{\chi^2_{1-\alpha/2}(n-1)})$

大样本假设检验

> - 只用S，不用S^*
> - 只用u检测，不用t和χ^2

> 如果已知具体概率分布，那么S就是概率分布的方差$D(x)$开方，μ就是概率分布的期望$E(x)$

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\overline{x}^2}$$

单参数

> 双侧

假设H0	H1	已知条件	拒绝域
$\mu=\mu_0$	$\mu\neq\mu_0$	已知σ	$\lvert\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\rvert\geq u_{\alpha/2}$
		未知σ	$\lvert\frac{\overline{x}-\mu_0}{S/\sqrt{n}}\rvert\geq u_{\alpha/2}$

> 单侧

假设H0	H1	已知条件	拒绝域
$\mu\leq\mu_0$	$\mu>\mu_0$	已知σ	$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\geq u_{\alpha}$
		未知σ	$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}\geq u_{\alpha}$
$\mu\geq\mu_0$	$\mu<\mu_0$	已知σ	$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\leq -u_{\alpha}$
		未知σ	$\frac{\overline{x}-\mu_0}{S/\sqrt{n}}\leq -u_{\alpha}$

单参数方差分析

> 问你某因素是否显著影响

$$S_T = \text{所有数据的平方和} - \frac{(\text{所有数据的和})^2}{\text{数据总数}}$$

$$S_A = \frac{(\text{类型}1\text{总和})^2}{\text{类型}1\text{数据数目}} + \frac{(\text{类型}2\text{总和})^2}{\text{类型}2\text{数据数目}} + \ldots + \frac{(\text{类型}n\text{总和})^2}{\text{类型}n\text{数据数目}} - \frac{(\text{所有数据的和})^2}{\text{数据总数}}$$

$$S_E = S_T - S_A$$

$$\overline{S_A} = \frac{S_A}{\text{类型总数目} - 1}$$

$$\overline{S_E} = \frac{S_E}{\text{数据总数} - \text{类型数}}$$

$$F_{\text{比}} = \frac{\overline{S_A}}{\overline{S_E}}$$

> 单因素方差分析表
>
> - 把上面这一堆填进来

方差来源	平方和	自由度	均方	F比
因素	S_A	类型数-1	$\overline{S_A}$	$F_{\text{比}} = \frac{\overline{S_A}}{\overline{S_E}}$
误差	S_E	数据总数-类型数	$\overline{S_E}$	
总和	S_T	数据总数-1		

$$F_\alpha = F_\alpha(\text{类型数} - 1,\ \text{数据总数} - \text{类型数})$$

最后，判断是否影响显著：

如果下式成立，则影响显著，否则不显著

$F_{\text{比}} \geq F_\alpha(\text{类型数} - 1,\ \text{数据总数} - \text{类型数})$

双参数方差分析（复习题有，考试没考过）

略。。。

一元线性回归

所求线性方程

$$y = \hat{a} + \hat{b}x$$

求线性方程

> 就是求$\hat{b}$和$\hat{a}$

参数	公式
$\hat{b}$	$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\overline{xy}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}^2}$
$\hat{a}$	$\hat{a} = \frac{y_1+\ldots+y_n}{n} - \frac{x_1+\ldots+x_n}{n}\hat{b} = \overline{y} - \hat{b}\overline{x}$

其中：

$$S_{xy} = (x_1y_1 + x_2y_2 + \ldots + x_ny_n) - \frac{(y_1+\ldots+y_n)(x_1+\ldots+x_n)}{n}$$
$$S_{xy} = n \cdot [\overline{xy} - \overline{x} \cdot \overline{y}]$$

问线性关系是否显著

> 就是求方差$\hat{\sigma^*}^2$

$$\hat{\sigma^*}^2 = \frac{S_{yy} - \hat{b}S_{xy}}{n-2}$$
$$or$$
$$\hat{\sigma^*}^2 = \frac{S_{yy} - \hat{b^2}S_{xx}}{n-2}$$
$$\hat{\sigma^*}^2 = \frac{n \cdot [(\overline{y^2} - \overline{y}^2) - \hat{b} \cdot (\overline{xy} - \overline{x} \cdot \overline{y})]}{n-2}$$

> 判断是否显著

判别式成立就显著

$$\frac{|\hat{b}|}{\sigma^*}\sqrt{S_{xx}} \geq t_{\alpha/2}(n-2)$$

方差区间估计

> 置信区间：

$$(\hat{b} - t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{b} + t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}})$$

点估计

矩估计

1. 写出$E(x)$与所求未知数的关系
2. 将1.的结果整理为 $未知数 = 关于E(x)的表达式$ 的形式
3. 根据样本，算出实际$E(x)$
4. 求未知数

双参数据估计

1. 写出$E(x)$与$E(x^2) = D(x) + E^2(x)$与戴求数的关系
2. 将1.的结果整理成 $未知数 = 关于E(x)和E(x^2)$ 的形式
3. 根据样本，算出实际的$E(x)$，$E(x^2)$
4. 算出未知数

极大似然估计

1. 将$x_1, x_2, x_3 \dots$)带入函数表达式，进行**连乘**
2. 对1的结果即取ln，变成求和
3. 依次对2的结果求导
4. 令3的结果导数为0

点估计的无偏估计

如果$\hat{\theta}$为估计值，θ为真实值

如果：

$$E(\hat{\theta}) = \theta$$

则称$\hat{\theta}$是θ的无偏估计