1. 基本信息

Baye's Theorem: $P[A_k|B]=\frac{P[B\cap A_k]}{P[B]}=\frac{P[B|A_k]\cdot P[A_k]}{\sum_{j=1}^{n}P[B|A_j]\cdot P[A_j]}$

$Var[X]=E[X^2]-E[X]^2$

If X and Y independent, $Var[X+Y]=Var[X]+Var[Y]$

$E[\varphi\circ X]=\sum\varphi(x)\cdot f_X(x)$

符号: μ 期望, σ: 标准差

Moments: $E[X^n]$ nth moments

$E[(\frac{X-\mu}{\sigma})^n]$ nth central moments

Moment generating function: $m_X(t)=\sum_{n=0}^{\infty}\frac{E[X^n]}{n!}t^n$

$m_X(t)=E[e^{tX}]$ (有条件, 课件111页)

$\left.\frac{d^k m_X(t)}{dt^k}\right|_{t=0}=E[X^k]$, 可用于计算期望和方差

2. 一些分布

① Bernoulli: 一次被选中的概率

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

② Binomial: n次中出现x次的概率

$f_X(x)=\binom{n}{x}p^x(1-p)^{n-x}$

$E[X]=np$ $Var[X]=np(1-p)$

$m_X(t)=(1-p+pe^t)^n$

③ Geometric: 第1次成功位于第x次的概率

$f_X(x)=(1-p)^{x-1}p$

CDF: $F(x)=1-(1-p)^{[x]}$

$E[X]=\frac{1}{p}$ $Var[X]=\frac{1-p}{p^2}$

$m_X(t)=\frac{pe^t}{1-(1-p)e^t}$

④ Pascal：第r次成功位于第x次的概率

$$f_X(x)=\binom{x-1}{r-1}p^r(1-p)^{x-r}$$

$$E[X]=\frac{r}{p}\quad Var[X]=\frac{(1-p)r}{p^2}$$

$$m_X(t)=\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$$

⑤ Negative Biomial：获得r次成功时，失败的次数

$$f_X(x)=\binom{x+r-1}{r-1}p^r(1-p)^x$$

$$E[X]=\frac{(1-p)r}{p}\quad Var[X]=\frac{(1-p)r}{p^2}$$

$$m_X(t)=\left[\frac{p}{1-e^t(1-p)}\right]^r$$

Rate of arrivals

Δt很短，Δt中有到达的概率是$\lambda\cdot\Delta t$（即$\frac{1}{\lambda}$为平均到达时间）

$-\lambda p_0(t)=p_0'(t)$ ①（$p_0(t)$代表t之前没有到达）

$p_X(t)=\frac{(\lambda t)^x}{x!}e^{-\lambda t}$ ② （x次到达）

---

⑥ Poisson：连续时间状态下，t时间内出现x次的概率，参考②，k为λt（$\frac{1}{\lambda}$为平均到达时间）

$$f_X(x)=\frac{k^x e^{-k}}{x!}$$

$$E[X]=k\quad Var[X]=k$$

$$m_X(t)=e^{k(e^t-1)}$$

☆ Posion approximation：预估binomial分布

$\binom{n}{m}p^m(1-p)^{n-m}=\frac{k^m}{m!}e^{-k}$ （$k=pn$）（当$n\to\infty$且$p<0.1$可使用）

⑦ Exponential：核元素衰变，参考①，β即为λ

$$f_\beta(x)=\begin{cases}\beta e^{-\beta x}, & x>0\\ 0, & x\le 0\end{cases}$$

$E[X]=\frac{1}{\beta}$ $Var[X]=\frac{1}{\beta^2}$

$m_x(t)=\frac{\beta}{\beta-t}$

CDF: $F_X(x)=1-e^{-\beta x}$

⑧ Gamma：exponential是出现第1次的时间分布，这是出现第α次

$$f_{T_r}(t)=\frac{\lambda^r}{(r-1)!}t^{r-1}e^{-\lambda t}$$

$$f_{\alpha,\beta}(x)=\begin{cases}\frac{\beta^\alpha}{\Gamma(\alpha)}\cdot x^{\alpha-1}\cdot e^{-\beta x}, & x>0\\ 0, & x\le 0\end{cases}$$

其中$\Gamma(\alpha)=\int_0^\infty z^{\alpha-1}e^{-z}dz$ (Euler gamma function，α为整数时是$(\alpha-1)!$)

$\alpha=r$，$\beta=\lambda$

$E[X]=\frac{\alpha}{\beta}$ $Var[X]=\frac{\alpha}{\beta^2}$

$m_x(t)=(1-\frac{t}{\beta})^{-\alpha}$

⑨ Chi-squared：多个正态分布的平方和

$$f_\gamma(x)=\begin{cases}\frac{1}{\Gamma(\gamma/2)\cdot 2^{\gamma/2}}\cdot x^{\frac{\gamma}{2}-1}e^{-\frac{x}{2}}, & x>0\\ 0, & x\le 0\end{cases}$$

γ: degrees of freedom

$E[X_\gamma^2]=\gamma$, $Var[X_\gamma^2]=2\gamma$

3、正态分布

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

$E[X]=\mu$ $Var[X]=\sigma^2$

$m_X(t)=e^{\mu t+\frac{1}{2}\sigma^2t^2}$

Transformation

$$f_Y(y) = f_X(\varphi^{-1}(y)) \cdot \left| \frac{d\varphi^{-1}(y)}{dy} \right| \quad (Y = \varphi(X))$$

因为密度变化，要乘上导数

最简单的例子是 $Y = \frac{X-\mu}{\sigma}$，使得 $E = 0$，$Var = 1$

$P[-\sigma < X-\mu < \sigma] = 0.68$

$2\sigma: 0.95$

$3\sigma: 0.997$

$P[-m\sigma < X-\mu < m\sigma] \geq 1 - \frac{1}{m^2}$ (Chebyshev Inequality)

CDF: $(E=0 \ Var=1)$ $\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$ ③ (表第7页)

De Moivre-Laplace: $\lim_{n\to\infty} P\left[a < \frac{S_n - np}{\sqrt{np(1-p)}} < b\right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2} dx$

(p 为成功率，S_n 为 n 次中成功的次数)

☆对 Binomial 分布前 n 项和的近似

$$P[X \leq y] = \sum_{x=0}^{y} \binom{n}{x} p^x (1-p)^{n-x} \approx \phi\left(\frac{y + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \quad (\phi \text{ 为③中的})$$

条件：p 接近 $\frac{1}{2}$，$n > 10$
或 $np > 5$，$p \leq \frac{1}{2}$
或 $n(1-p) > 5$，$p > \frac{1}{2}$

4、Multivariate random varibles：多个变量

$X = (X_1, X_2, \cdots, X_n)$

joint density function: $f_X(x)$ (X 为 $X_1 \sim X_n$)

Marginal density: $f_{X_k}(x_k) = \sum_{x_1 \sim x_n}^{x \neq x_k} f_X(x_1, \cdots, x_n)$

$$= \int_{R^{n-1}} f_X(x)\, dx_1 \cdots dx_{k-1}\, dx_{k+1} \cdots dx_n$$

Independent: $f_X(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$ (n个也一样)

Conditional density: $f_{X_1|X_2}(x_1)=\frac{f_{X_1X_2}(x_1,x_2)}{f_{X_2}(x_2)}$ $(f_{X_2}(x_2)>0)$

$$E[X_k]=\int_{\mathbb{R}} x_k f_{X_k}(x_k)dx_k=\int_{\mathbb{R}^n} x_k f_X(x)dx$$

$$E[X]=\begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

Covariance (协方差): $Cov[X,Y]=E[(X-\mu_X)(Y-\mu_Y)]=E[XY]-E[X]E[Y]$

$$Cov[X,X]=Var[X]$$

$$Var[X\pm Y]=Var[X]+Var[Y]\pm 2Cov[X,Y]$$

若X与Y线性独立，则 $Var[X+Y]=Var[X]+Var[Y]$

$$Var[X]=\begin{pmatrix} Var[X_1] & Cov[X_1,X_2] & \cdots & Cov[X_1,X_n] \\ Cov[X_1,X_2] & Var[X_2] & & \vdots \\ \vdots & & \ddots & Cov[X_{n-1},X_n] \\ Cov[X_1,X_n] & \cdots & Cov[X_{n-1},X_n] & Var[X_n] \end{pmatrix}$$

Standardize: $\tilde{X}=\frac{X-\mu_X}{\sigma_X}$ $(E[\tilde{X}]=0,\ Var[\tilde{X}]=1)$

$$Cov[\tilde{X},\tilde{Y}]=E[\tilde{X}\tilde{Y}]-E[\tilde{X}]E[\tilde{Y}]=\frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$$

Pearson coefficient of correlation: $\rho_{XY}=\frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$

$$-1\le\rho_{XY}\le 1$$

若 $\rho_{XY}=\pm1$，则X经过线性变换可与Y完全相同，即 $Y=aX+b$

$|\rho_{XY}|$越接近1，X与Y越线性相关

Fishor Transformation

$$\sqrt{\frac{Var[\tilde{X}+\tilde{Y}]}{Var[\tilde{X}-\tilde{Y}]}}=\sqrt{\frac{1+\rho_{XY}}{1-\rho_{XY}}}\in(0,+\infty)\quad \ln\left(\sqrt{\frac{Var[\tilde{X}+\tilde{Y}]}{Var[\tilde{X}-\tilde{Y}]}}\right)=\frac{1}{2}\ln\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right)=\underbrace{\tanh^{-1}}_{(\text{arctanh})}(\rho_{XY})\in\mathbb{R}$$

Bivariate normal distribution 双变量正态分布，见课件

5、Weak law of large numbers

$$P\left[\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right]\xrightarrow{n\to\infty}0$$ （μ为$E[X]$，$X_1\sim X_n$为X中随机量）

$$P\left[\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right]\leq\frac{\sigma^2}{\varepsilon^2}\frac{1}{n}$$

6、Hypergeometric distribution

$$f_X(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$ $(r>n,\ N-r>n)$

r个红球，$(N-r)$个黑球，随机选n个，选到x个红球的概率

$E[X]=\frac{nr}{N}$ $Var[X]=\frac{nr(N-n)(1-\frac{r}{N})}{N(N-1)}$

$Var[X_i]=\frac{r}{N}(1-\frac{r}{N})$ $Cov[X_i,X_j]=-\frac{1}{N}\cdot\frac{r(N-r)}{N(N-1)}$

需要注意，$p_i\neq p_j$，因为拿走一个，总数和红球数会变化

Binomial approximation：没有别的，就是当作取走后概率仍然恒定，变为binomial的情况

7、Transformation of variables

$Y=\varphi\circ X\ (R^n\to R^n)$

$f_Y(y)=f_X\circ\varphi^{-1}(y)\cdot|\det D\varphi^{-1}(y)|$ （D没有含义，就是$\varphi^{-1}(y)$的行列式）

与1个变量类似

具体怎么转换见课件248-249页

8、Reliability

failure density f_A，F_A：到目前为止出现过问题的概率

reliability function R_A：到目前为止一直都好的的概率

hazard rate：ρ_A：$\rho_A(t)=\frac{f_A(t)}{R_A(t)}=\lim\limits_{\Delta t\to 0}\frac{P[t\leq T\leq t+\Delta t\mid T\geq t]}{\Delta t}$

failure density是减，hazard rate要除

Weibull density

$\rho(t)=\alpha\beta t^{\beta-1}$ $R(t)=e^{-\alpha t^{\beta}}$ $f(t)=\alpha\beta t^{\beta-1}e^{-\alpha t^{\beta}}$

Weibull distribution: $f(x)=\begin{cases}\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, & x>0\\ 0, & x\le 0\end{cases}$

☆uniform distribution: $f_X(x)=\begin{cases}1, & 0\le x\le 1\\ 0, & \text{otherwise}\end{cases}$

串并联：series：只要有1个坏了就不行

parallel：只要有1个是好的就没问题

9、Sample and data

population: a large collection of individuals，每个个体都有给定的值，个体的选取是完全随机的

sample：里面的每个都是独立的，每个和整体X的分布均相同

percentiles: X%

quartiles: q_1: 25%数据 $\le q_1$

q_2: 50% $\le q_2$ (median) 中位数

q_3: 75% $\le q_3$

$$q_2=\begin{cases}X_{\frac{n+1}{2}}, & n\text{ odd}\\ \frac{1}{2}\left(X_{\frac{n}{2}}+X_{\frac{n}{2}+1}\right), & n\text{ even}\end{cases}$$

Interquartile range (IQR): $IQR=q_3-q_1$

Histogram: bin: number of categories 列的数量 k

Sturges: $k=\lceil \log_2(n)\rceil+1$

Excel: $k=\lceil\sqrt{n}\rceil$

Freedman-Diaconis: $h=\frac{2\cdot IQR}{\sqrt[3]{n}}$ (h为列的宽度)

sample range: $\max\{x_i\}-\min\{x_i\}$

$$h=\frac{\max\{x_i\}-\min\{x_i\}}{k}$$

分布情况，对应描述：见课件305页

用Mathematica画：见课件301~304页

Stem-and-Leaf diagrams：每个数取第1位，为stem，将其后面1位作为leaf，课件307页

Box-and-Whirper Plots (Boxplots)

Inner fences：$f_1 = q_1 - \frac{3}{2}IQR$，$f_3 = q_3 + \frac{3}{2}IQR$

Adjacent values：$a_1 = \min\{x_k : x_k \ge f_1\}$ $a_3 = \max\{x_k : x_k \le f_3\}$

即不能超出f_1, f_3，但只延伸到有元素的地方，不留空白

Outer fences：$F_1 = q_1 - 3IQR$ $F_3 = q_3 + 3IQR$

Inner与outer之间：near outliers，outer之外：far outliers

图见课件310页

10、Parameter estimation

estimator $\hat{\theta}$：评价指标（人为选取，比如平均数，中位数等）

Bias：$\theta - E[\hat{\theta}]$．若$E[\hat{\theta}] = \theta$，则$\hat{\theta}$ unbiased.（比如样本平均值）

（θ为population parameter，对于一个群体为固定值）

Mean square error：$MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2]$

$$MSE[\hat{\theta}] = Var[\hat{\theta}]^2 + (bias)^2$$

$\hat{\theta}$的要求：接近θ，方差要小

$Var[\bar{X}] = \frac{\sigma^2}{n}$（$\bar{X}$表示样本的平均数，$\sigma$为population的方差）

Method of moments：$\widehat{E[X^k]} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$，为kth moment的unbiased estimator

但这在某些情况下会biased

unbiased sample variance：$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2$

Likelihood function：$L(\theta) = \prod_{i=1}^{n} f_{X_\theta}(X_i)$（$X_\theta$为density函数）

要使其最大化，$X_1=X_2=\cdots=X_n=$ location of maximum（即 density 最大的地方）

11、Interval estimation

需要根据样本数据推测原始数据的信息

$Var[\bar{X}]=\mu$，$Var[\bar{X}]=\frac{\sigma^2}{n}$，且 $\bar{X}$ 满足正态分布（若 X 本身为正态分布）

（μ、σ 为全体数据的，$\bar{X}$ 为样本平均值，n 为样本大小）

Two sided confidence interval for θ ($100(1-\alpha)\%$):

$P[L_1\le\theta\le L_2]=1-\alpha$ and $P[\theta<L_1]=P[\theta>L_2]=\frac{1}{2}\alpha$（若分布对称，则 $\hat{\theta}=\frac{L_1+L_2}{2}$，在中间）

upper confidence bound: $P[\theta\le L]=1-\alpha$

lower confidence bound: $P[\theta\ge L]=1-\alpha$

θ is not random（全体数据的属性）, but L_1 and L_2 random（根据样本不同而不同）

$$\alpha/2=P[Z\ge z_{\alpha/2}]=\frac{1}{\sqrt{2\pi}}\int_{z_{\alpha/2}}^{\infty}e^{-\frac{1}{2}x^2}dx$$

① 已知全体的方差，计算置信区间（$100(1-\alpha)\%$） 平均值的

two sided: $\bar{X}\pm\frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}$

upper: $\bar{X}+\frac{z_{\alpha}\cdot\sigma}{\sqrt{n}}$ lower: $\bar{X}-\frac{z_{\alpha}\cdot\sigma}{\sqrt{n}}$

Sample from normal distribution ($n\ge2$ 且固定):

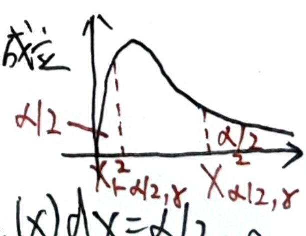
(i) $\bar{X}$、S^2 independent

(ii) $\bar{X}$ 正态分布，平均为 μ，方差 $\frac{\sigma^2}{n}$

(iii) $\frac{(n-1)S^2}{\sigma^2}$ is chi-squared distributed with $n-1$ degrees of freedom

(i) 反过来，若 $\bar{X}$、S^2 为独立，则说明"全体"正态分布，也成立

② 计算方差的置信区间



定义 $\chi^2_{1-\alpha/2,\gamma}$ 和 $\chi^2_{\alpha/2,\gamma}$: $\int_0^{\chi^2_{1-\alpha/2,\gamma}}f_{\chi^2_\gamma}(x)dx=\int_{\chi^2_{\alpha/2,\gamma}}^{\infty}f_{\chi^2_\gamma}(x)dx=\alpha/2$

$$1-\alpha = P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

upper{100(1-α)%}: $\sigma^2 \le \dfrac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}$

lower(100(1-α)%): $\sigma^2 \ge \dfrac{(n-1)S^2}{\chi^2_{\alpha,n-1}}$

③ 全体的方差未知，计算平均值的置信区间

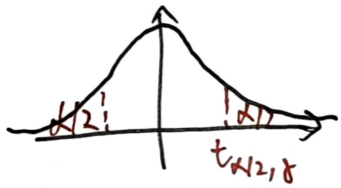
Student T-distribution

Z为标准的正态分布，χ_γ^2 为 chi-squared (γ degrees of freedom)

则 $T_\gamma = \dfrac{Z}{\sqrt{\chi^2_\gamma / \gamma}}$

大小为n的样本，平均值 $\underset{\text{全体}}{\mu}$，$\underset{\text{全体}}{\text{方差}}$ σ^2，则 $T_{n-1} = \dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 满足n-1自由度的 T distribution

定义 $t_{\alpha/2,\gamma}$: $\int_{t_{\alpha/2,\gamma}}^{\infty} f_{T_\gamma(t)}\,dt = \alpha/2$



则 100(1-α)% 为 $\bar{X} \pm t_{\alpha/2,n-1}\, S/\sqrt{n}$