Prob Stats L18

Semester Review

April 25, 2023

Probability

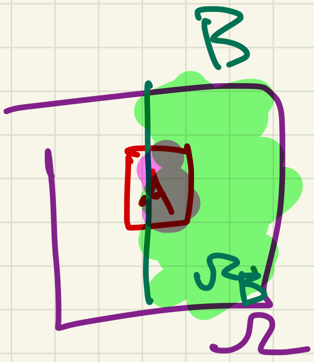
Events A, B sets of things that could happen

$A \subset \Omega$ ← sample space

$Pr(A) : 2^{\Omega} \to [0,1]$

conditional prob

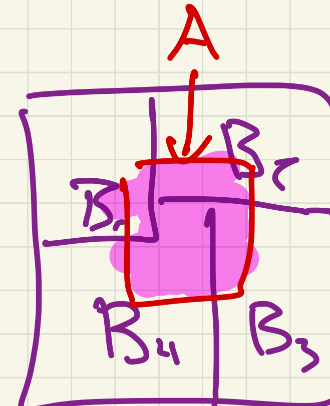
$$P(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)}$$



Total Probability

Partition $\Omega = \bigcup_{i=1} B_i$ $\quad B_i \cap B_j = \emptyset$

$$\Pr(A) = \sum_i \Pr(A \mid B_i) \cdot \Pr(B_i)$$



Tree Diagram

① flip coin

$1/2$

② Roll Die $\quad L\{\leq 2\} \quad G\{>2\}$

Coin $\quad$ Die

H $\xrightarrow{1/2}$ $\xrightarrow{1/3}$ L $= \Pr(H \cap L) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$

$\xrightarrow{2/3}$ G

T $\xrightarrow{1/2}$ $\xrightarrow{1/3}$ L $= \Pr(T \cap L) = \left(\frac{1}{3}\right) \cdot \left(\frac{1}{2}\right) = \frac{1}{6}$

$\xrightarrow{2/3}$ G

$\Pr(L) = \Pr(T \cap L) + \Pr(H \cap L) = \frac{1}{3}$

Independence A, B iff

- $Pr(A|B) = P(A)$
- $Pr(B|A) = P(B)$
- $\boxed{Pr(B \cap A) = Pr(A) \cdot Pr(B)}$

Bayes Rule

$$Pr(B|A) = \frac{Pr(A|B) \cdot P(B)}{Pr(A)}$$

B = model

A = data

Random Variables $X : \Omega \to \mathbb{R}$

Ω **discrete** (rolls of dice, flip coin)

$Pr(X = \text{head})$

$\frac{}{1}$

continuous (rainfall, time)

$Pr(\text{Rain} \leq 1 \text{inch})$

Probability density function

discrete $f_X(a) = P_X(X = a)$

continuous $P_r(x \in [a,b]) = \int_{x=a}^{b} f_X(x)\,dx$

Cumulative Density function

$F_X(a) = \int_{x=0}^{a} f_X(a)\,dx = P_r(X \leq a)$

Expectation = "average"

$$E[g(X)] = \sum_{i=1}^{n} \underbrace{g(a_i)}_{\text{value}} \cdot \underbrace{f_X(a_i)}_{Pr(X=a_i)}$$

$$= \int_X g(x) \cdot f_X(x)\, dx$$

$$= c \boxed{\int_X f_X(x)\, dx} = 1$$

$$= c$$

Variance X

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

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

$$= c^2 (1) - (c)^2 = 0$$

Linearity of Expectation

X, Y RVs $\quad a, b, c$ const.

- $E[aX + bY + c] = aE[X] + bE[Y] + c$

- $\text{Var}[aX + b] = a^2 \cdot \text{Var}[X]$

- $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}(X,Y)$

  if X, Y independent $\quad \text{Cov}(X,Y) = 0$

Distributions

- Bernoulli: $X \sim \text{Ber}(p)$

$X = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$

$E[X] = p$ $\quad$ $\text{Var}[X] = p(1-p)$

- Binomial $X \sim \text{Bin}(n, p)$

$\Pr(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}$

$k = \#$ heads

$E[X] = np$ $\quad$ $\text{Var} = np(1-p)$

- Geometric $X \sim \text{Geo}(p)$ $\quad$ $\Pr(X=k) = (1-p)^{k-1} \cdot p$

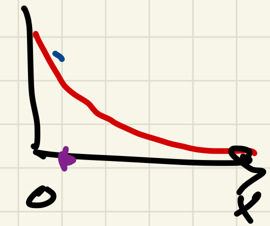
$\#$ trials until first heads.

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

Unif $\quad X \sim \text{Unif}(a,b) \quad f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

$\boxed{E[X] = \frac{b-a}{2}}$ $\quad Var[X] = \frac{1}{12}(b-a)^2$

Exponential $\quad X \sim \text{Exp}(\lambda)$



$f_X = \lambda \cdot \exp(-\lambda x)$

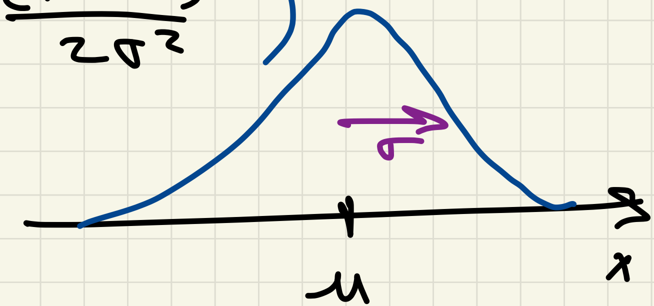
$F_X(a) = 1 - \exp(-\lambda \cdot a)$

$\boxed{E[X] = 1/\lambda}$ $\quad Var[X] = 1/\lambda^2$

Normal

$X \sim N(\mu, \sigma^2)$

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



$E[X] = \mu$

$Var[X] = \sigma^2$

cdf $F_X(a) = pnorm(a, \mu, \sigma^2)$

$\mu = 0$, $\sigma^2 = 1$: $= pnorm(a)$

$$F_X(a) = \int_{x=-\infty}^{a} f_X(x)\,dx$$

Joint Probabilities

$$\Pr(X=a, Y=b)$$

Independence

R.V. X, Y

if $\forall a,b$

$$\Pr[X=a]\cdot\Pr[Y=b] = \Pr[X=a, Y=b]$$

$$\begin{array}{c|cc|l}
 & A=0 & A=1 & \\
\hline
B=0 & 0.3 & 0.2 & = 0.5 \\
B=1 & 0.1 & 0.4 & = 0.5 = \Pr(B=1) \\
\hline
 & 0.4 & 0.6 &
\end{array}$$

$\Pr(A=0)$

marginal probability

$$\Pr(B=0 \mid A=1) = \frac{\overset{0.2}{\Pr(B=0, A=1)}}{\underset{0.6}{\Pr(A=1)}}$$

$$= \frac{0.2}{0.6} = \frac{1}{3}$$

$$E[g(X,Y)] = \sum_i \sum_j g(a_i, b_j) \Pr(X = a_i, Y = b_j)$$

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

$$= E[X \cdot Y] - E[X] \cdot E[Y]$$

if X, Y independent $\xrightarrow{\text{then}}$ $\mathrm{Cov}(X,Y) = 0$

but if $\mathrm{Cov}(X,Y) = 0$ $\xrightarrow{\text{then}}$ maybe X, Y indep
maybe <u>not</u>

Estimation

Random Sample

$X_1, X_2 \dots X_n \overset{iid}{\sim} f(\theta)$

Random Variables

Goal estimate θ

iid independent and identically distributed.

estimator $\hat{\theta} = T(X_1, X_2, \dots X_n)$

Random Variable

$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

unbiased : $\text{bias}(\hat{\theta}) = 0$

: $E[\hat{\theta}] = \theta$

$E[\bar{X}_n] = E[X_i] \overset{?}{=} \theta$

Sample variance

$$\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \qquad E[S_n^2] = \sigma^2$$

Central Limit Theorem

$X_1 \ldots X_n \overset{iid}{\sim} F$

$\bar{X}_n$

- $E[\bar{X}_n] = E[X_i] = \mu$
- $Var[\bar{X}_n] = \frac{Var[X_i]}{n} = \frac{\sigma^2}{n}$
- $\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$ as $n \to \infty$

Confidence Intervals

$100(1-\alpha)\%$ confid. interval $[L, R]$

$\hookrightarrow Pr(L \leq \theta < R) = 1-\alpha$

L, R random variables, from Random Sample

θ is unknown parameter

$$Z = \frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{if } X_i \sim N(\mu, \sigma^2)$$

$$\Delta_n = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \qquad L = \bar{X}_n - \Delta_n \quad R = \bar{X}_n + \Delta_n$$

$z_{\alpha/2} =$ qnorm$(1-\alpha/2)$

$$Pr(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = Pr(L_n \leq \mu \leq R_n) = 1-\alpha$$

if $X_i \sim N(\mu, \sigma^2)$

σ^2 unknown

or $X_i \sim Ber(p)$

<u>sample variance</u> $S_n^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$

$S_n^2 = p(1-p) = \bar{X}_n(1-\bar{X}_n)$

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim t\text{-dist}(df = n-1)$$

$\Delta_n = t_{\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$ $\quad t_{\alpha/2} = qt(1-\alpha/2, df = n-1)$

$L = \bar{X}_n - \Delta_n \qquad R = \bar{X}_n + \Delta_n$

$P_r(L_n \le \mu \le R_n) = 1 - \alpha$

Hypothesis Testing

Null Dist. H_0 : "boring" status quo

: model where $X_1 \dots X_n \sim H_0$

Alternative Hypothesis H_1 : guess of how H_0 is broken

$H_0 : \mu = 10$ $\quad$ $H_1 : \mu > 10$

$N(\mu, \sigma^2)$

$N(10, \sigma^2)$

don't know σ^2 $\rightarrow$ t-dist

$$T = \frac{\bar{X}_n - \mu_{=10}}{S_n/\sqrt{n}} = T(X_1, \dots X_n)$$

constant $\rightarrow$ realized test statistic $\rightarrow$ $t = T(x_1, \dots, x_n)$ real data

Critical Value

$t_\alpha = qt(1-\alpha, df = n-1)$

$Pr(T < t_\alpha) = 1-\alpha$

Random Variable

if t ← realization of data

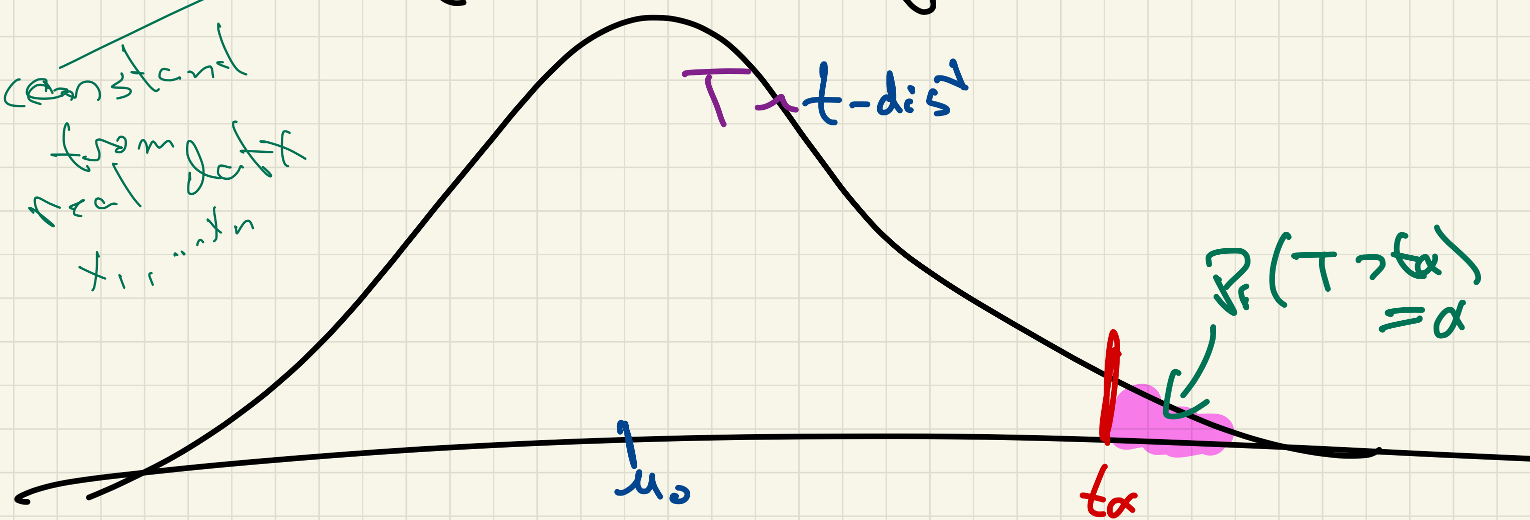
$t > t_\alpha \longrightarrow$ reject null hypothesis

constant from real data $x_1, \dots, x_n$

if $t < t_\alpha$

⇒ not reject null hypothesis

double negative on purpose



Actual Data $x_1, \ldots, x_n$

$\bar{x}_n \leftarrow \frac{1}{n} \sum_{i} x_i$

$H_0 \quad \mu_0 = 10$

$s_n^2 \leftarrow \frac{1}{n-1} \sum_{i} (x_i - \bar{x}_n)$

$$t = \frac{\bar{x}_n - \mu_0}{s_n / \sqrt{n}}$$

$$t_\alpha = \underbrace{qt(1-\alpha, df = n-1)}_{R}$$