08/21/19

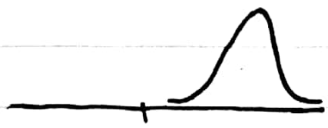
AMS 131 Lecture 11

X = # of Latinx people selected for grand jury duty

T_1: no discrimination $\quad n = 220$, 79.1% Latinx

Frequentist: if T_1 true, $X \sim \text{Binomial}(n, 0.791)$

X = 100



100

incorrect assumption that there is no discrimination

$P(X \le 100 \mid T_1) = 10^{-27}$

Bayesian: $P(T_1 \mid X \le 100)$ how probable is the theory based on how the data came out

θ = actual probability of an eligible Latinx person being chosen $(0 < \theta < 1)$

$(S \mid \theta) \sim \text{Binomial}(n, \theta)$

PMF: $f_{S|\theta}(s \mid \theta) = P(S = s \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$

$I(s = 0, 1, \dots, n)$

Information internal to the dataset about θ is summarized by the likelihood (un-normalized) density defined: $\ell(\theta \mid s) = c\, P(S = s \mid \theta)$

$$f_{\Theta|S}(\theta \mid s) = c \cdot f_\Theta(\theta) \cdot \ell(\theta \mid s)$$

$$\begin{pmatrix}\text{posterior}\\ \text{information}\end{pmatrix} = \begin{pmatrix}\text{normalizing}\\ \text{constant}\end{pmatrix} \cdot \begin{pmatrix}\text{prior}\\ \text{information}\end{pmatrix} \cdot \begin{pmatrix}\text{likelihood}\\ \text{information}\end{pmatrix}$$

$$f_{\Theta|S}(\theta \mid s) = c \cdot f_\Theta(\theta) \cdot \theta^s (1-\theta)^{n-s}$$

$\downarrow$

$c \cdot \theta^{x_1} (1-\theta)^{x_2}$ makes calculations easier

$X \sim \text{Beta}(\alpha, \beta) \quad \alpha > 0, \beta > 0$

$f_X(x) = \underbrace{c\, \theta^{\alpha - 1} (1-\theta)^{\beta - 1}}_{\text{plug into } f_{\Theta|S}(\theta \mid s)}$ prior PDF

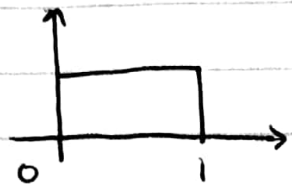
$= c \cdot \theta^{(\alpha + s) - 1} (1-\theta)^{(\beta + n - s) - 1} = \text{Beta}(\alpha + s, \beta + n - s)$

Lecture 11 (cont.)

$$\left.\begin{array}{l}\theta \sim \text{Beta}(\alpha,\beta) \\ (S\mid\theta) \sim \text{Binomial}(n,\theta)\end{array}\right\} (\theta\mid S) \sim \text{Beta}(\alpha+S,\ \beta+n-S)$$

Neutral prior

Uniform(0,1)



doesn't favor any value from (0,1)

$$\text{Uniform}(0,1) = \theta^{1-1}(1-\theta)^{1-1}$$

$$\theta \sim \text{Uniform}(0,1) \leftrightarrow \theta \sim \text{Beta}(1,1)$$

$$E(cX) = cE(X) \qquad V(X+c) = V(X)$$

$$E(X+c) = E(X)+c \qquad V(cX) = c^2V(X)$$

$C(X+c, Y) = C(X,Y)$ shift left / right

$C(cX, Y) = cC(X,Y)$

For all R.V. X, Y for which $E(XY)$ exists, — Cauchy-Schwarz Inequality

$(E[XY])^2 \le (E[X])^2 \cdot (E[Y])^2$ from which

$[C(X,Y)]^2 \le \sigma_X^2 \cdot \sigma_Y^2$ and

$-1 \le \rho(X,Y) \le 1$

$\rho(X,Y) > 0 \leftrightarrow X, Y$ positively correlated

$\rho(X,Y) < 0 \leftrightarrow X, Y$ negatively correlated

$\rho(X,Y) = 0 \leftrightarrow X, Y$ uncorrelated

X, Y independent with $\begin{cases} 0 < \sigma_X^2 < \infty \\ 0 < \sigma_Y^2 < \infty \end{cases}$

$\rightarrow C(X,Y) = \rho(X,Y) = 0$

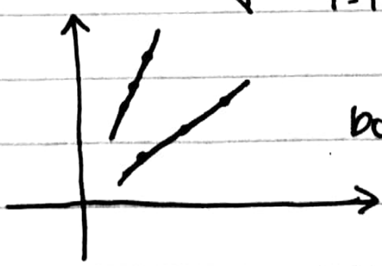
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_x^*}\right)\left(\frac{y_i-\bar{y}}{S_y^*}\right) \qquad S_x^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$-1 \le r \le 1$

perfect linearity

+/- depends on slope

fits all data



both $r=1$

AMS 131 Lecture 11 (cont.)

Independence → 0 correlation, but not the converse

$X \sim \text{Uniform}\{-1,0,1\}$, $Y \triangleq X^2$ $E(X)=0$

X, Y dependent but:

$E[XY] = E[X^3] = E[X] = 0$ since X and X^3 are identically distributed

$C(X,Y) = E(XY) - E(X)E(Y) = 0$

$\rho(X,Y) = \dfrac{C(X,Y)}{\sigma_X \sigma_Y} = 0$ X, Y uncorrelated, but dependent.

X R.V. with $0 < \sigma_X^2 < \infty$, $Y = aX + b$

for $a \neq 0$, b constants → $(a>0)$ $\rho(X,Y) = +1$

$(a<0)$ $\rho(X,Y) = -1$

so $\rho(X,Y)$ measures strength of linear association

If X, Y R.V, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$: if independent,

$V(X+Y) = V(X) + V(Y) + 2C(X,Y)$ $C(X,Y) = 0$

$C(aX, bY) = abC(X,Y)$ ← $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$

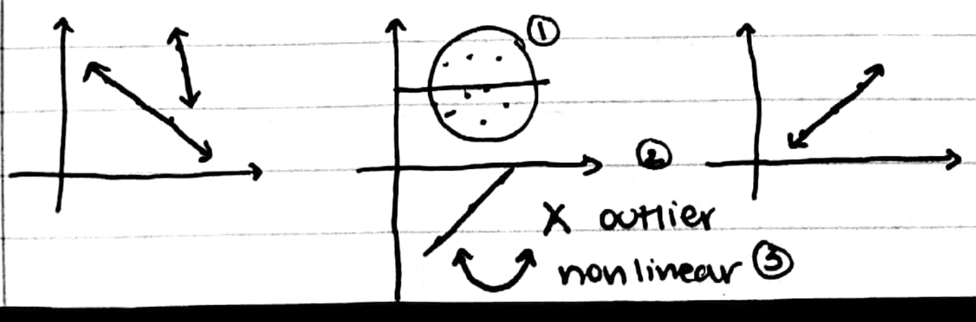
$V(aX + bY + c) = a^2 V(X) + b^2 V(Y) + 2abC(X,Y)$

$V(X-Y) = V(X) + V(Y) - 2C(X,Y)$

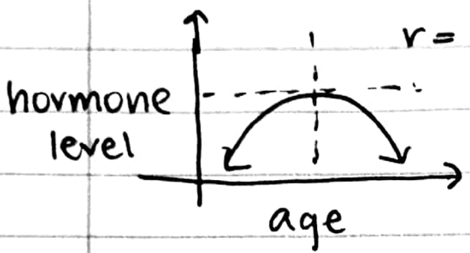
If $X_1, \ldots, X_n$ such that (X_i, X_j) uncorrelated for all $1 \le i \neq j \le n$ then:

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

$\rho(X,Y) = -1$ $\rho(X,Y) = 0$ $\rho(X,Y) = +1$

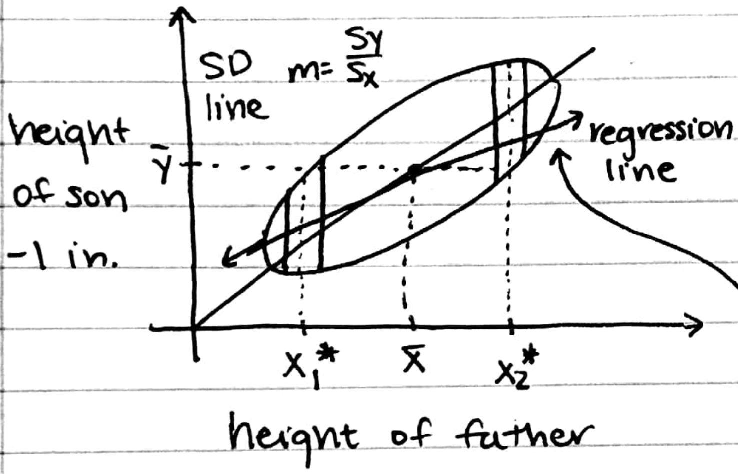


Lecture 11 (cont.) AMS 131



r can be fooled by outliers and/or non linearity

Conditional expectation: X, Y related R.V.s (not independent) then there is information in X for predicting Y. i.e. some function $d: \mathbb{R} \to \mathbb{R}$ so $d(X)$ is close to Y, optimal d?



divide elliptical scatterplot into a bunch of vertical strips - took mean over all the strips

the height of son is less than height of father

$E(Y \mid X = x)$ = approx. linear.

This effect is called regression effect, regression to the mean

The points in the vertical strip over x_2^* are a distribution of Y given $X = x_2^*$: $f_{Y|X}(y \mid x = x_2^*)$

The number $\hat{w}$ that minimizes the mean squared error $E[(\hat{w} - W)^2]$ - of $\hat{w}$ as a prediction for W is $\hat{w} = E(W)$

So Galton adopted MSE as a measure of "close", the $\hat{y}$ that minimizes MSE $E[(\hat{y} - Y)^2]$ in the vertical strip on $x = x_2^*$ must be the conditional mean/expectation of the RV. $(Y \mid X = x_2^*)$

AMS 131 Lecture 11 (cont.)

$\left\{\begin{array}{l}\underline{\text{Conditional expectation}} \\ \text{(Mean) of } Y \text{ given } X = x\end{array}\right\} = E(Y \mid X = x)$ is just the expectation of the conditional distribution $f_{Y|X}(y|x)$ of Y given $X = x$, namely:

$$E(Y|X) = \int_{\mathbb{R}} y f_{Y|X}(y|x)\,dy \quad \text{for continuous } (Y|X=x)$$

$$= \sum_{\text{all } y} y f_{Y|X}(y|x) \quad \text{for discrete}$$

$E(Y|X)$ is just a constant, equal to the conditional mean of Y when X is the constant x.

$h(x) \triangleq E(Y|X=x)$, then $h(X) \triangleq E(Y|X)$ is the conditional expectation of Y given X

$(n_C + n_T)$ people who are similar to population $p = \{$adults with disease A$\}$ and who are randomized n_C (control), n_T (treatment)

$S_i \begin{cases} 1 = \text{disease in remission} \\ 0 = \text{did not} \end{cases}$ θ = proportion of success if everyone was in p, θ is unknown.

The R.V.s $(S_i|\theta)$ are IID Bernoulli(θ) and the R.V. $S = \sum_{i=1}^{n_T} S_i$ has the conditional Binomial dist.

$$(S|\theta) \sim \text{Binomial}(n_T, \theta)$$

conditional expectation R.V. $E(S|\theta) = n_T\theta$ (linear)

also $E(\theta|S)$

and the constant $E(\theta|S=s)$

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A|B_i) \rightsquigarrow \underbrace{f_Y(y)}_{P(A)} = \int_{-\infty}^{\infty} \underbrace{f_X(x)}_{P(B_i)} \cdot f_{Y|X}(y|x)\,dx$$

Lecture 11 (cont.) AMS 131

$$E(Y|x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)\,dx$$

$$E(Y) = \int_{-\infty}^{\infty} y \cdot f_Y(y)\,dy = \int_{-\infty}^{\infty} y \cdot \left[\int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x)\,dx\right] dy$$

$$= \int_{-\infty}^{\infty} f_X(x)\left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy\right] dx$$ if ok to change order of integration.

$$= \int_{-\infty}^{\infty} f_X(x) \cdot E(Y|x)\,dx$$ → weighted average of $E(Y|X)$ with $f_X(x)$ as the weights

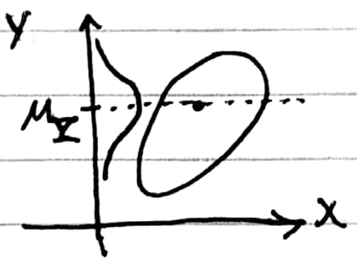
$E(Y) = E_X[E(Y|X)]$ Double Expectation Theorem

The number $V(Y|x) \triangleq E_X[(Y - E(Y|x))^2 \,|\, x] = g(x)$ is the conditional variance of Y given $X = x$ - and the R.V. $V(Y|X)$ is $g(x)$ for Y given X

X, Y related R.V.s we want some function $\hat{Y} = d(X)$ to predict Y from X → The prediction that minimizes the MSE $E[Y - \hat{Y})^2 = E[(Y - d(X))^2]$ is $\hat{Y} = d(X) = E(Y|X)$

$V(Y) = E_X[V(Y|X)] + V_X[E(Y|X)]$ Part 2 of Double Expectation Theorem.

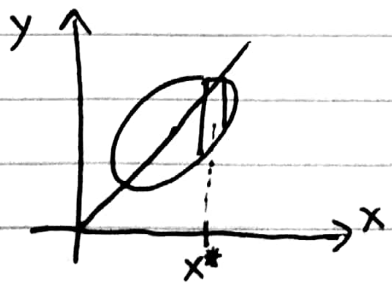
Stage 1: Predict Y without knowing X



$\hat{Y}$, no X $= \mu_Y = E(Y)$

MSE: $E[(Y - \mu_Y)^2] = V(Y) = \sigma_Y^2$

Stage 2: Observe X, now predict Y



$X = x^*$, then the MSE-optimal prediction is

$\hat{Y}_{X=x^*} = E(Y|X = x^*)$

MSE: $E\left[\left(Y - E(Y|X = x^*)\right)^2\right] = V(Y|x^*)$

AMS 131 Lecture 11 (cont.)

Before stage 2, $E_X[V(Y|X)]$ is the best guess

The second part of the Double Expectation Theorem:

$$\underset{\substack{\uparrow \\ \text{MSE of} \\ \hat{Y} \text{ no } X}}{V(Y)} = E_X[\underset{\substack{\uparrow \\ E(\text{"MSE"}) \text{ of} \\ \hat{Y}_X = E(Y|X)}}{V(Y|X)}] + V_X[E(Y|X)]$$

$V[E(Y|X)] \geq 0$ so,

$$\underbrace{E_X[V(Y|X)] + V_X[E(Y|X)]}_{V(Y) \text{ MSE of } \hat{Y}_{no X}} \geq \underbrace{E_X[V(Y|X)]}_{E(MSE) \text{ of } \hat{Y}_X}$$

You expect your predictive accuracy to get better when you bring in an X to predict Y

Bayes Decision Theory: optimal action under uncertainty

X has discrete PMF $f_X(x) = \begin{cases} \frac{1}{2} & x = -\$350 \\ \frac{1}{2} & x = +\$500 \end{cases}$ 0 else

X = net gain from gamble A.

Y has discrete PMF $f_Y(y) = \begin{cases} \frac{1}{3} & y = +\$40 \\ \frac{1}{3} & y = +\$50 \end{cases}$ $\frac{1}{3}$ $y = +\$60$ 0 else

Y = net gain from gamble B

$E(X) = +\$75$, $E(Y) = +\$50$ not necesarily better

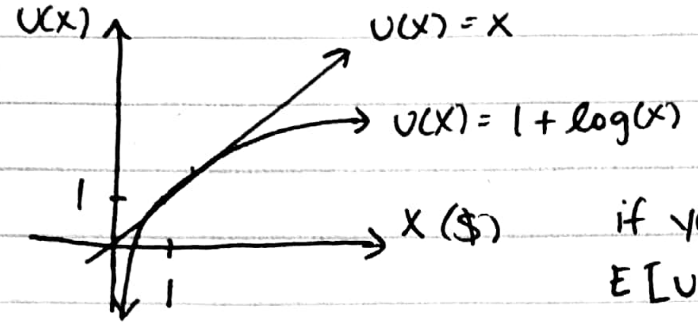
risk averse would pick B

risk seeking would pick A

Utility $U(x)$ is a function which assigns to each possible net gain $-\infty < x < \infty$ a real # $U(x)$ that represents the <u>value to you</u> of gaining x

Lecture 11 (cont.) AMS 131

Net worth: \$10] getting \$1 won't mean
Net worth: \$1M as much to the richer one
(sublinear)



maximizing expected utility (MEU)
if you prefer gamble X over Y if $E[U(X)] > E[U(Y)]$ and you're ok with X / Y if $E[U(X)] = E[U(Y)]$

Single \$2 ticket, grand prize \$487 million
X = the unknown amount you'll win, thinking about X before the drawing.
$x \cdot P(X = x)$ = \$1.99 weighted expectance

Before drawing someone offers \$$x_0$ to sell your ticket.
$E[U(X)]$ if you keep ticket = \$1.99
Sell if $U(x_0) > E[U(X)]$ under MEU
If $U(x) = x$, sell if offered more than \$1.99

grand prize now \$1.6 billion
$E(X)$ is now \$5.80 on a \$2 ticket
Difference between doing once vs. repeating.
Using this utility, you would have to subtract from x the monetary cost of the disruption of selling everything to buy tickets.

$U(a, \theta) = U(\text{action, unknown}) = g[B(a, \theta), C(a, \theta)]$
Cost benefit analysis benefit cost.

AMS 131

Lecture 11 (cont.)

Bernoulli: $X \sim$ Bernoulli(p) $0<p<1$ if DISCRETE

$f_X(x) = p^x(1-p)^{1-x} \, I_{\{0,1\}}(x)$ ← support

$= \begin{cases} p & \text{for } x=1 \\ 1-p & x=0 \end{cases}$ 0 else

$E(X)=p$ $\psi_X(t) = pe^t + (1-p)$ for all $-\infty<t<\infty$

$V(X) = p(1-p)$ $SD(X) = \sqrt{p(1-p)}$

If X_i are IID Bernoulli(p) → Bernoulli Trials with parameter p, if infinite = Bernoulli (stochastic) process

Binomial: $X \sim$ Binomial(n,p) $n>0$ integer $0<p<1$

$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \, I_{\{0,1,\dots,n\}}(x)$ ← support

$X_1, \dots, X_n$ $\sim$ IID Bernoulli(p) $X = \sum_{i=1}^{n} X_i \sim$ Binomial(n,p)

$X \sim$ Binomial(n,p) $E(X) = np$ $V(X) = np(1-p)$

$\psi_X(t) = [pe^t + (1-p)]^n$ for all $-\infty<t<\infty$

$SD = \sqrt{np(1-p)}$

Hypergeometric: finite population, $S = A+B$

A elements of type 1, B elements of type 2

n elements at random without replacement

↳ SRS: simple random sample

X: # of type 1 elements $X \sim (A,B,n)$

$f_X(x \mid A,B,n) = \dfrac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \, I[\max(0,n-B) \le x \le \min(n,A)]$

for $(A,B,n) > 0$ $n \le A+B$

$E(X) = n \cdot \frac{A}{A+B}$ $V(X) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{A+B-n}{A+B-1}\right)$

if with replacement → IID Binomial

$p = \frac{A}{A+B}$ $E(X) = np = n\frac{A}{A+B}$ $V(X) = np(1-p) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$