08/19/19

Lecture 10 AMS 131

$X_1, \dots, X_n$ independent R.V. $a_1, \dots, a_n, b$ constants

$$V\left[\left(\sum_{i=1}^{n} a_i X_i\right) + b\right] = \sum_{i=1}^{n} a_i^2 V(X_i)$$

$V(cX) = c^2 V(X)$

$V(X + c) = V(X)$

$X \sim$ Binomial (n, p) $E(X) = np$ $V(X), SD(X)$?

$$S_i = \begin{cases} 1 & \text{success on } i\text{TH trial} \\ 0 & \text{else} \end{cases} \quad i = 1, 2 \dots n$$

$S_1, \dots, S_n$ are IID Bernoulli(p)

then $X = \sum_{i=1}^{n} S_i$ $V(X) = V\left[\sum_{i=1}^{n} S_i\right] = \sum_{i=1}^{n} V(S_i)$ with independ.

$V(S_i) = E(S_i^2) - [E(S_i)]^2$ mean of Bernoulli $= E(S_i) = p$

$$S_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with prob. } (1-p) \end{cases} \quad S_i^2 = \begin{cases} 1^2 = 1 & \text{with prob. } p \\ 0^2 = 0 & \text{with prob } (1-p) \end{cases}$$

$V(S_i) = E(S_i^2) - [E(S_i)]^2 = p - p^2 = p(1-p)$ for individual trial

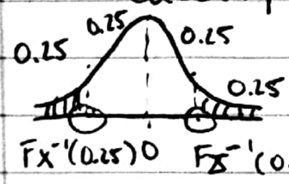
$$V(X) = \sum_{i=1}^{n} V(S_i) = \sum_{i=1}^{n} p(1-p) = np(1-p) \quad SD(X) = \sqrt{np(1-p)}$$

T-S. study $Y \sim$ Binomial (n, p) = Binomial $(5, \frac{1}{4})$

$E(Y) = np = 1.25$ $SD(Y) = \sqrt{np(1-p)} = \sqrt{(5)(\frac{1}{4})(\frac{3}{4})} = 0.97$

"The number of T-S babies this couple will have will be around $\underset{\mu_Y}{1.26}$ give or take $\underset{\sigma_Y}{1}$" Y

$X \sim$ (standard) Cauchy $f_X(x) = \begin{cases} \frac{1}{\pi(1+x^2)} \\ -\infty < x < \infty \end{cases}$ $E(X)$ DNE, SD $V(X)$ DNE



interquartile range (IQR) measure of spread.

$= F_X^{-1}(0.75) - F_X^{-1}(0.25)$

median = 50TH percentile

CDF: $F_X(x) = \int_{-\infty}^{x} \frac{1}{\pi(1+t^2)}\, dt = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi}$

Scanned by CamScanner

AMS 131 Lecture 10 (cont.)

$\tan^{-1}(x)$ is the principal inverse of tan (arctangent) ranging from $-\frac{\pi}{2}$ to $+\frac{\pi}{2}$ as $-\infty < x < \infty$

$$F_X(x) = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi} = p \qquad x = F_X^{-1}(p) = \tan\left(\frac{p-\frac{1}{2}}{\pi}\right) = -\cot(p\pi)$$

$$IQR = F_X^{-1}\left(\tfrac{3}{4}\right) - F_X^{-1}\left(\tfrac{1}{4}\right) = \tan\left(\tfrac{\pi}{4}\right) - \tan\left(-\tfrac{\pi}{4}\right)$$

$E(X) = E(X^1)$ — moments of a R.V.

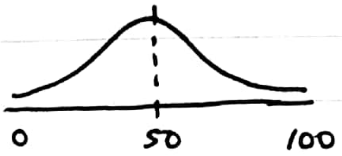
$V(X) = E(X^2) - [E(X^1)]^2$
$= E(X^1 - \mu)^2$

X R.V, $k \geq 1$ integer $E(X^k) \triangleq k^{TH}$ moment of X

$\left(k^{TH} \text{ moment of } X \text{ exists}\right) \longleftrightarrow E(|X|^k) < \infty$

If $E(|X|^k) < \infty$ for integer $k \geq 1$, then $E(|X|^j) < \infty$ for all integers $j < k$. If the k^{TH} moment exists, so does all of the previous ones

- $E[(X-\mu)^k]$ is the k^{TH} central moment, or the k^{TH} moment of X around its mean.
- $E[(X-\mu)^1] = E(X) - \mu = \mu - \mu = 0$
  ↳ every R.V. has 1st central moment $= 0$
- If the dist. of X is symmetric around μ_X, then $E[(X-\mu)^k] = 0$ for all ODD integers – where it exists

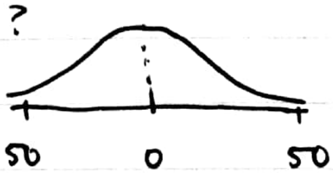


$E(X) = 50 = \mu$

$E[(X-\mu)^1] = 0$

$E[(X-\mu)^3] = 0$

- maybe relates to skewness?



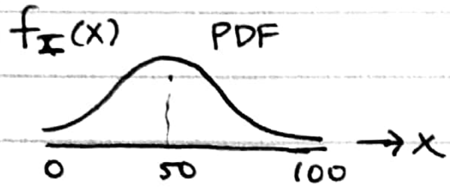
first moment $= 0$

$\mu = 0$

Scanned by CamScanner

If the third moment exists and is finite: $\text{skewness}(X) \triangleq E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right]$

σ_X converts X to standard units

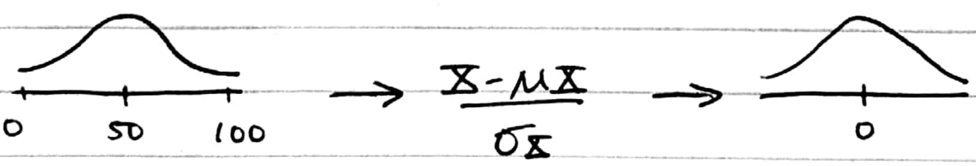


empirical rule (graphical interpretation) of SD

1. start at mean μ_X, go 1 SD σ_X either way: $(\mu_X - \sigma_X, \mu_X + \sigma_X)$, you will usually enclose about $\frac{2}{3}$ of the total probability.
   ↳ exact for bell curve, 68%.
2. go 2 SD: $(\mu_X - 2\sigma_X, \mu_X + 2\sigma_X)$, you will usually enclose most (95%) of the total probability.
3. go 3 SD: $(\mu_X - 3\sigma_X, \mu_X + 3\sigma_X)$, you will usually enclose almost all (99.7%) of the total probability.

$\sigma_X = 2$ (48, 52) way too small
$\sigma_X = 45$ (5, 95) way too big.
$50 - 3\sigma_X = 0 \rightarrow \sigma_X \approx 17$

STANDARD NORMAL DISTRIBUTION



does not change the shape

$\mu = 50, \sigma \doteq 17$ $\mu = 0, \sigma = 1$

skewness > 0 skewness = 0 skewness < 0

moment generating functions X R.V, t a real #

$$e^x = \frac{1}{0!} + \frac{x}{1!} + \frac{x^2}{2!} + \dots = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

$$e^{tX} = 1 + (tX) + \frac{(tX)^2}{2} + \dots$$

$$E(e^{tX}) = 1 + E(tX) + E\left(\frac{(tX)^2}{2}\right) + \dots = 1 + tE(X) + \frac{t^2}{2}E(X^2) + \dots$$

AMS 131 Lecture 10 (cont.)

$$\frac{d}{dt} E(e^{tX}) = E(X) + t\,E(X^2) + \frac{1}{2}t^2 E(X^3) + \dots$$

$t=0, \quad = E(X)$ $\quad \downarrow \frac{d}{dt}$

$$\frac{d^2}{dt^2} E(e^{tX}) = E(X^2) + t\,E(X^3) + \dots$$

$t=0, \quad = E(X^2)$

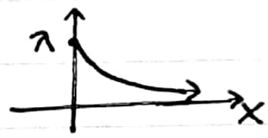
(MGF)

$\psi_X(t) \triangleq E(e^{tX})$ moment generating function of X

X. R.V. with MGF $\psi_X(t)$, finite for all values t on the open interval $(-a, b)$ around 0, $a, b < \infty$, then for all integers $n > 0$:

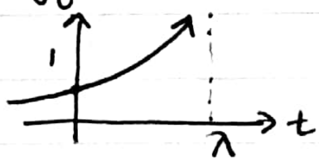
$$E(X^n) = \frac{d^n}{dt^n} \psi_X(t) \Big|_{t=0}$$ n^{TH} derivative of $\psi_X(t)$, evaluated at $t=0$

$X \sim$ Exponential(λ) $\quad f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{else} \end{cases}$
$\lambda > 0$



$$\psi_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(t-\lambda)x}\, dx$$

Only finite for $t < \lambda$ $\quad = \frac{\lambda}{\lambda - t}$
otherwise $\rightarrow$ infinity



$$E(X^1) = \left(\frac{d}{dt} \frac{\lambda}{\lambda - t}\right)\Big|_{t=0} = \frac{1}{\lambda}$$

$$V(X) = E(X^2) - [E(X^1)]^2$$

$$E(X^2) = \left(\frac{d^2}{dt^2} \frac{\lambda}{\lambda - t}\right)\Big|_{t=0} = \frac{2}{\lambda^2}$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

$E(X^3) = \frac{6}{\lambda^3}$, $E(X^4) = \frac{24}{\lambda^4}$ $\quad SD(X) = \sqrt{V(X)} = \frac{1}{\lambda}$

$\uparrow$ $\qquad \searrow$ $E(X^k) = \frac{k!}{\lambda^k}$

positive skew, ✓
long right tail

Scanned by CamScanner

Lecture 10 (cont.) AMS 131

X R.V. with MGF $\psi_X(t)$, $Y = aX + b$ a, b constants

then at every value of t for which $\psi_X(at)$ is finite: $\psi_Y(t) = e^{bt}\psi_X(at)$

$X \sim$ Binomial(n,p) $\quad X = \sum_{i=1}^{n} S_i$

$S_i \overset{IID}{\sim}$ Bernoulli(p) $i = 1, 2, \ldots, n$

$$\psi_{S_i}(t) = E(e^{tS_i}) = e^{t \cdot 1} \cdot P(S_i = 1) + e^{t \cdot 0} \cdot P(S_i = 0)$$

$X_1, \ldots, X_n$ independent R.V., MGF $\psi_{X_i}(t)$, $Y = \sum_{i=1}^{n} X_i$

MGF of Y is $\psi_Y(t)$ for every t such that $\psi_{X_i}(t)$ is finite for all $i = 1, \ldots, n$ $\quad \psi_Y(t) = \prod_{i=1}^{n} \psi_{X_i}(t)$

MGF of the sum = product of individual MGFs

Since S_i are IID $\psi_X(t) = \prod_{i=1}^{n} \psi_{S_i}(t) = \prod_{i=1}^{n} [pe^t + (1-p)]$

$$E(X) = \left(\frac{d}{dt}\psi_X(t)\right)\Big|_{t=0} \qquad \text{IID} = [pe^t + (1-p)]^n$$

$$= \frac{d}{dt}[pe^t - (1-p)]^n \Big|_{t=0} = np \checkmark$$

$$E(X^2) = \frac{d^2}{dt^2}[pe^t - (1-p)]^n \Big|_{t=0} = np[1 + (n-1)p]$$

$$V(X) = E(X^2) - [E(X^1)]^2$$
$$= np + n(n-1)p^2 - n^2p^2$$
$$= np + n^2p^2 - np^2 - n^2p^2 = np(1-p) \checkmark$$

$$E(X^3) = \frac{d^3}{dt^3}[pe^t - (1-p)]^n \Big|_{t=0} = np[1 + (n-2)(n-1)p^2 + 3p(n-1)]$$

AMS 131

Lecture 10 (cont.)

X has MGF $\psi_X(t)$ ⎫ finite in an open
Y has MGF $\psi_Y(t)$ ⎭ interval around $t=0$

then $\psi_X(t) = \psi_Y(t) \overset{\text{iff}}{\longleftrightarrow}$ X, Y have identical probability distributions

So the MGF uniquely characterizes a R.V.

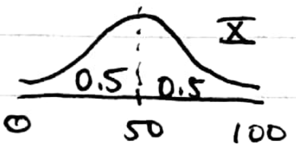
X R.V with CDF $F_X(x)$ on interval I

$h(x)$ is 1-1 on I, $Y = h(X)$

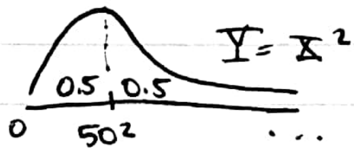
$E[h(X)] \neq h[E(X)]$ unless $h(x) = ax+b$

if m_X is a median of X, $m_X = F_X^{-1}(\frac{1}{2})$,
then $h(m_X)$ is a median of $Y = h(X)$

↳ median preserved under transformation



$\longrightarrow$



probability preserved under transformation

How to predict a value, is it a good prediction?

$\hat{x}$: value picked before observation X.

$(\hat{x}-X)$: prediction error, either pos./neg.

$E(\hat{x}-X)=0$: average over repititions

$$\begin{array}{ccc} \hat{x} & X & \hat{x}-X \\ 50 & 57 & -7 \\ 50 & 39 & +11 \\ \cdots & \cdots & \cdots \\ 50 & & \end{array}$$

The bias of $\hat{x}$ as a prediction for X is bias$(\hat{x}) \triangleq E(\hat{x}-X)$

The prediction $\hat{x}$ is unbiased if bias$(\hat{x}) = 0 \rightarrow \hat{x} = E(X)$

$E(\hat{x}-X)^2$, $E|\hat{x}-X|$ want to be small

↳ mean square error (MSE) of $\hat{x}$ as a prediction

$\hat{x}$ that minimizes MSE is $E(X)$

$$E[(\hat{x}-X)^2] = E(\hat{x}^2 - 2\hat{x}X + X^2) = \hat{x}^2 - 2\hat{x}E(X) + E(X^2)$$

$$\frac{d}{d\hat{x}} E[(\hat{x}-X)^2] = 2\hat{x} - 2E(X) = 0 \text{ iff } \hat{x} = E(X) \qquad \frac{d^2}{d\hat{x}^2} = 2 > 0$$

minimum

Lecture 10 (cont.) AMS 131

$MSE(\hat{x}) = E(\hat{x} - X)^2 = V(X) + [bias(\hat{x})]^2$

$\hat{x} = E(X)$ minimizes $MSE(\hat{x})$ and achieves 0 bias, thus $MSE(\hat{x}) = V(X)$

Root MSE (RMSE) $(\hat{x}) = SD(X) = \sigma_X$

- Always predict 50, expect to be wrong by 17

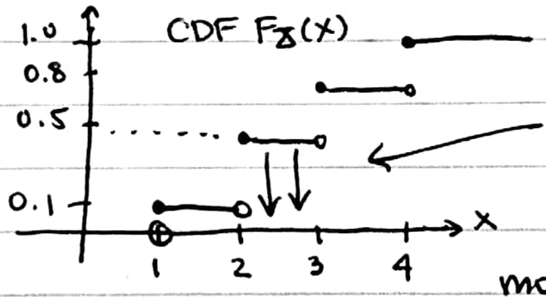
For $E[|\hat{x} - X|]$ Mean absolute error (MAE), the $\hat{x}$ that minimizes this is the median m_X

- if the dist. is skewed, use μ_X / m_X responsibly.

median: X R.V. → every number m such that
$P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$

x	$P(X=x)$
1	0.1
2	0.4
3	0.3
4	0.2

X discrete on $\{1,2,3,4\}$



$2 < x < 3$ all have $F_X(x) = \frac{1}{2}$

most say median = 2.5

There is no right answer to MSE vs MAE, depends on the real world consequences of the prediction error $(\hat{x} - X)$

↳ quantifying requires a utility function, seen later.

$$\begin{array}{c} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{array} \begin{bmatrix} y_1 & x_1 \\ y_2 & x_2 \\ \cdots & \cdots \\ y_n & x_n \end{bmatrix} \begin{array}{c} X_1 \\ X_2 \\ \cdots \\ X_n \end{array}$$

$n = 1000$

1 row for each British family ~1885 with ≥ 1 son

x_i: height of father

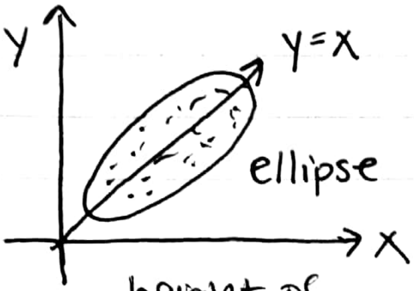
y_i: height of son (≥1 son, at random) height

	mean	SD
X	58 in	2.5 in
Y	59 in	2.5 in

secular trend in height: better nutrition

AMS 131 Lecture 10 (cont.)

height of son -1 inch.

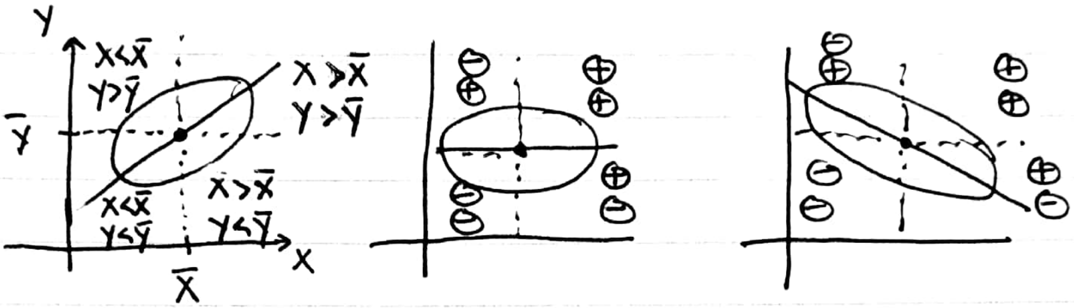


ellipse

scatterplot

height of father

positive association: tall fathers usually have tall sons
↓
but how strong?



pos. assoc. zero assoc. neg. assoc.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\cdot\left(\frac{y_i-\bar{y}}{s_y}\right) = r$$ correlation coefficient

$$E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\cdot\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$ R.V. version

numerator $E[(X-\mu_X)\cdot(Y-\mu_Y)]$

↳ $C(X,Y)$ = covariance of X and Y

Independence of 2 or more R.Vs is a special case of a more general reality, where
(your uncertainty about something) and
(your uncertainty about something else) are related

X, Y R.V.s with finite means $\mu_X = E(X)$, $\mu_Y = E(Y)$
The covariance of X, Y $C(X,Y)$ is defined as:
$E[(X-\mu_X)\cdot(Y-\mu_Y)]$ as long as it exists.

Lecture 10 (cont.) AMS 131

$(X-\mu_X)\cdot(Y-\mu_Y) = X\cdot Y - \mu_X Y - \mu_Y X + \mu_X \mu_Y$

so, $C(X,Y) = E(XY) - \underline{\mu_X E(Y)} - \underline{\mu_Y E(X)} + \underline{\mu_X \mu_Y}$

$C(X,Y) = E(XY) - \mu_X \mu_Y$ much easier formula

(expectation of product) $\cdot$ (product of expectations)

$C(X,Y)$ exists: $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$

Covariance depends on the units of measurement of X, Y

X = education level $C(X,Y) = (\text{years})\cdot(\$)$ $\underline{?}$

Y = yearly income

X = Max daily temp in °C $\rightarrow$ °F $= \frac{9}{5}(°C) + 32°$ = not

Y = max daily relative humidity [%] the same $C(X,Y)$

a, b fixed constants $C(aX+b, Y) = aC(X,Y)$

↳ changes scale → changes amount of association

converting to standard units = unitless, divide by σ_X, σ_Y

$$X' = \frac{X - E(X)}{SD(X)} = \frac{X - \mu_X}{\sigma_X} \qquad 0 < \sigma_X < \infty$$

$E(X') = 0$, $V(X') = 1 = SD(X')$

X, Y R.V. with finite variances σ_X^2, σ_Y^2 (and therefore finite means μ_X, μ_Y), the correlation is:

$$(\text{rho})\ \rho(X,Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\cdot\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$

- invariant to linear transformations

$$\rho(aX+b, cY+d) = \rho(X,Y)$$

if $a<0$, $\rho(aX+b, Y) = -\rho(X,Y)$