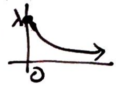
5/21/19

Lecture 15

Thm: If its premise is satisfied and the calculations are manageable, you get all the moments of X just by computing $\psi_X(t)$ and differentiating it over and over.

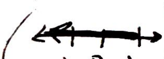
Ex: $X \sim$ Exponential (λ)

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



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

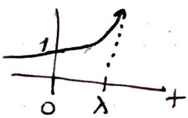
This integral is finite only if $t - \lambda < 0 \rightarrow$ for $t < \lambda$ (



) but this means (since $\lambda > 0$) that it's definitely finite in an open interval around 0 (e.g. $(-\lambda, \lambda)$)

So $\psi_X(t)$ exists for $t < \lambda$ and equals $\psi_X(t) = \lambda \int_0^\infty e^{(t-\lambda)x} dx$

$$= \frac{\lambda}{\lambda - t}$$



Now take the derivatives

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

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

$$E(X^3) = \left(\frac{d^3}{dt^3} \left(\frac{\lambda}{\lambda - t} \right) \right)\bigg|_{t=0} = \frac{6}{\lambda^3}$$

$\leftarrow$ positive skew (long right-hand tail)

$$E(X^4) = \left(\frac{d^4}{dt^4} \left(\frac{\lambda}{\lambda - t} \right) \right)\bigg|_{t=0} = \frac{24}{\lambda^4}$$

Evidently $E(X^k) = \dfrac{k!}{\lambda^k}$

So $V(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

and $SD(X) = \frac{1}{\lambda}$.

Consequences of the MGF definition

1) X rv with MGF $\psi_X(t)$

$Y = aX + b$

(a, b constants)

then at every value of t for which $\psi_X(at)$ is finite,

$$\psi_Y(t) = e^{bt}\,\psi_X(at)$$

Ex:

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

$S_i \overset{iid}{\sim} \text{Bernoulli}(p) \quad (i=1, \dots, n)$

MGF of S_i is easy:

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

$$= [pe^t + (1-p)]$$

This uses the Law of the unconscious Statistician

2) $X_1, \dots, X_n$ independent rv, MGF of X_i is $\psi_X(t)$,

$Y = \sum_{i=1}^{n} X_i$, MGF of Y is $\psi_Y(t)$ for every t such that

$\psi_{X_i}(t)$ is finite for all $i = 1, \dots, n$

$$\psi_Y(t) = \prod_{i=1}^{n} \psi_{X_i}(t)$$

MGF of Binomial cont.

Since the S_i are IID, $\psi_X(t) \overset{IID}{=} \prod_{i=1}^{n} \psi_{S_i}(t)$

$$\overset{IID}{=} \prod_{i=1}^{n} [pe^t + (1-p)]$$

Take derivatives

$$\overset{IID}{=} [pe^t + (1-p)]^n$$

$$E(X) = \left(\frac{d}{dt} \psi_X(t)\right)\Big|_{t=0}$$

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

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

So $V(X) = E(X^2) - (E(x))^2$

$$= np + n(n-1)p^2 - n^2p^2$$

$$= np + n^2p^2 - np^2 - n^2p^2$$

$$= n(p-p^2)$$

$$= np(1-p) \checkmark$$

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

$$= np[1 + (n-2)(n-1)p^2 + 3p(n-1)]$$

3) X has MGF $\psi_X(t)$ → finite in an open interval around $t=0$

Y has MGF $\psi_Y(t) \overset{iff}{\longleftrightarrow}$ X, Y have identical probability distributions

So the MGF (if it exists) uniquely characterizes an rv

Mean versus Median

1) X rv (CDF F_X) with values in an interval I

$h(x)$ 1-1 function on I

$Y = h(x)$

if m_X is a median of X (i.e if $m_X = F_X^{-1}(\frac{1}{2})$)

then $h(m_X)$ is a median of $Y = h(X)$

This is not in general true of the mean:

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

Prediction Bedrock

X rv with mean μ_X, SD σ_X

Before X is observed, suppose your job is to predict what its value will be. What should you do? How can you tell if a prediction is good?

Say you picked the number $\hat{x}$ (a fixed known constant) before X is observed.

Then after X arrives, your prediction error would be $(\hat{x} - X)$ which might be either positive or negative

- One possible criterion for goodness would be to find $\hat{x}$ such that $E(\hat{x} - X) = 0$

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

Def: Your prediction $\hat{x}$ is unbiased if $bias(\hat{x}) = 0$

Clearly, to achieve this just choose $\hat{x} = E(x)$

- Another possible criterion for goodness would be to find $\hat{x}$ such that $E(\hat{x} - X)^2$ is small

Def: $E[(\hat{x}-X)^2]$ is called the mean squared error (MSE) of $\hat{x}$ as a prediction for X

Small theorem: The $\hat{x}$ that minimizes MSE is $\hat{x}=E(X)$

Proof: $E((\hat{x}-X)^2) = E(\hat{x}^2 - 2\hat{x}X + X^2)$

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

This is a quadratic function of $\hat{x}$

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

$$\frac{d^2}{dx^2} = 2 > 0 \quad \text{So } E(X) \text{ is a minimum}$$

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

So the choice $\hat{x}=E(X)$ both minimizes $MSE(\hat{x})$ and achieves 0 bias, and with this choice $MSE(\hat{x}) = V(X) = \sigma_X^2$

- A different criterion: find $\hat{x}$ such that $E[|\hat{x}-X|]$ is small

Def: $E|\hat{x}-X|$ is called the mean absolute error (MAE) of $\hat{x}$ as a prediction for X

Thm: X rv w/ finite mean μ_X

Let m_X be (a/the) median of X

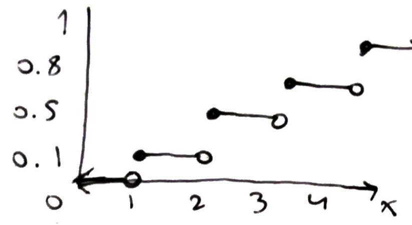
$\rightarrow$ the $\hat{x}$ that minimizes $MAE(\hat{x})$ is (a/the) median m_X

Median def:

$X \sim \rightarrow$ every number m such that $P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$ is a median of the dist. of X

Ex: Non unique median

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



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

Most people would say median = 2.5

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

Which is a better criterion, MSE or MAE?

There is no universal right answer. It depends on the real-world consequences of your prediction errors $(\hat{x} - X)$

Quantifying these consequences involves the creation of a utility function, which we'll briefly examine later

Bedrock Covariance and Correlation

Independence of 2 or more rvs is a special case of a more general reality, in which (your uncertainty about something) and (your uncertainty about something else) are related.

Def: X, Y rv with finite means $\mu_X = E(X)$ and $\mu_Y = E(Y)$

The Covariance of X and Y, written $C(X, Y)$ is defined as

$C(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$ as long as this expectation exists

Consequences of this def:

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

So $C(X, Y) = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$

$= E(XY) - \mu_X \mu_Y - \cancel{\mu_X \mu_Y} + \cancel{\mu_X \mu_Y}$

$C(X, Y) = E(XY) - \mu_X \mu_Y$ ← easier formula to compute w/

Expectation of product − product of expectations

2) Sufficient condition for $C(X, Y)$ to exist

$\sigma^2_X < \infty$ and $\sigma^2_Y < \infty$

3) Covariance's value depends on the units of measurement of X and Y.

Ex: X = education level (years of schooling completed)

Y = yearly income ($)

$C(X, Y)$ comes out in (years)·($)

Ex: X = max daily temperature in °C

Y = max daily relative humidity (%)

If you change your mind & measure temperature X in

$°F = \frac{9}{5} °C + 32°$

$C(X', Y) = C(\frac{9}{5}X + 32, Y) \neq C(X, Y)$

Easy to show that if a, b are fixed constants then

$C(aX + b, Y) = aC(X, Y)$ so $\underset{\uparrow \atop °F}{C(X', Y)} = 1.8 \cdot \underset{\uparrow \atop °C}{C(X, Y)}$

You can make the association between temperature & relative humidity seem larger just by switching from °C to °F

Easy fix: Def: The process of converting a rv X to standard units (SU) is achieved with the linear transformation

$$X' = \frac{X - E(X)}{SD(X)} = \frac{X - \mu_X}{\sigma_X}$$

As long as $0 < \sigma_X < \infty$ this is a meaningful definition

$E(X') = 0 \quad V(X') = 1 = SD(X')$

Def: X, Y rv with finite variances σ_X^2 and σ_Y^2 (and therefore finite means μ_X and μ_Y → the correlation of X and Y is $\rho(X, Y) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\cdot\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$

$$= \frac{C(X, Y)}{\sigma_X \cdot \sigma_Y}$$

With this definition, the correlation is invariant to linear transformation of either variable (or both):

for any constants $a, c > 0$ and b, d,

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

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

Consequences of the correlation def:

1) Cauchy-Schwartz inequality:

For all rv X, Y for which $E(XY)$ exists,

$[E(XY)]^2 \leq [E(X)]^2 \cdot [E(Y)]^2$

from which $[C(X, Y)]^2 \leq \sigma_X^2 \cdot \sigma_Y^2$ and $-1 \leq \rho(X, Y) \leq +1$

Def:

$\rho(X, Y) > 0 \longleftrightarrow X, Y$ positively correlated ← associated

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

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

2) X, Y independent rv with $\left\{\begin{array}{l} 0 < \sigma_X^2 < \infty \\ 0 < \sigma_Y^2 < \infty \end{array}\right\}$

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

So independence implies 0 correlation, but (interestingly) not the converse:

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

$\rightarrow X, Y$ clearly dependent since X completely determines Y, but $E(XY) = E(X^3) = E(X) = 0$ since X and X^3 are identically distributed

$$C(X, Y) = \underbrace{E(XY)}_{0} - \underbrace{E(X)}_{0} \cdot E(Y) = 0$$

So $\rho(X, Y) = \dfrac{C(X, Y)}{\sigma_X \sigma_Y} = 0$ and X, Y are uncorrelated

3) $X \sim$ with $0 < \sigma_X^2 < \infty$, $Y = aX + b$ for $\left\{ \begin{matrix} a \neq b \\ b \end{matrix} \right\}$ constants $\rightarrow (a > 0)$ $\rho(X, Y) = 1$

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

So $\rho(X, Y)$ measures the strength of linear association between X and Y

4) Important: If $X, Y \sim$, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$ then

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

5) (a, b, c any constants)

$$C(aX, bY) = ab\, C(X, Y)$$

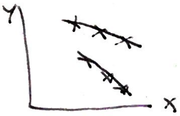
$\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty \rightarrow V(aX + bY + c)$

$$= a^2 V(X) + b^2 V(Y) + 2ab\, C(X, Y)$$

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

6) If $X_1, \dots, X_n$ such that (X_i, X_j) uncorrelated for all $1 \le i \ne j \le n \xrightarrow{\text{then}} V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i)$

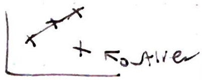
7) $\rho(X, Y) = -1$



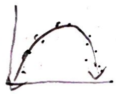
Points in scatterplot sample from $f_{XY}(x, y)$ all fall on line with negative slope (not necessarily -1)

$\rho(X, Y) = 0$

case 1

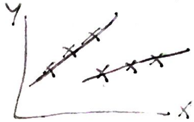


case 2



Nonlinearity
case 3

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



points in scatterplot sample from $f_{X,Y}(x, y)$ all fall on line with positive slope (not necessarily $+1$)

Conditional Expectation

X, Y related rvs (not independent): then there is information in X for predicting Y

We should be able to find some function $d: \mathbb{R} \to \mathbb{R}$ such that $d(X)$ is "close" in some sense to Y - what is the optimal d?

Galton ex:

Galton divided the elliptical scatterplot up into a bunch of narrow vertical strips, e.g.. the one over x_1^* or the other one over x_2^*

The points in the vertical strip over x_2^* are a random sample from the conditional distribution of Y given $X = x_2^*$
$f_{Y|X}(y | x = x_2^*)$

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

So he adopted MSE as his measure of "close" and concluded that the $\hat{y}$ that minimizes the MSE $E[(\hat{y}-Y)^2]$ in the vertical strip defined by $x \doteq x_2^*$ must be the conditional mean, or conditional expectation, of the rv $(Y \mid X = x_2^*)$

<u>Def:</u> X, Y rv, Y finite mean

$$\left\{\frac{\text{Conditional expectation}}{\text{(mean) of } Y \text{ given } X = x}\right\} = E(Y \mid x) \text{ is just the}$$

expectation of the conditional distribution $f_{Y \mid X}(y \mid x)$ of Y given $X = x$

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

and
$$E(Y \mid x) = \sum_{\text{all } y} y\, f_{Y \mid X}(y \mid x) \quad \text{for discrete } (Y \mid X = x)$$

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

<u>Def:</u> $h(x) \triangleq E(Y \mid X = x)$ then the rv $E(Y \mid X) \triangleq h(X)$ is the conditional expectation of Y given X