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,

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

Example

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

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

MGF of S_i

is easy: $\psi_{S_i}(t) = E(e^{tS_i})$

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

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

This uses the Law of the Unconscious Statistician

(2) $X_1, \ldots, X_n$ <u>independent</u> rv, MGF of X_i is $\psi_{X_i}(t)$, $Y = \sum_{i=1}^{n} X_i$,

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

~~MGF of~~ Binomial, continued

Since the S_i are IID,

$$\psi_X(t) \overset{\text{IID}}{=} \prod_{i=1}^{n} \psi_{S_i}(t)$$

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

$$\overset{\text{IID}}{=} \left[p e^t + (1-p)\right]^n$$

Now, as before, we just crank out the derivatives.

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

$$= np \checkmark$$

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

so $V(X) = E(X^2) - \left[E(X)\right]^2$

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

$$= np + \cancel{n^2p^2} - np^2 - \cancel{n^2p^2}$$

$$= n(p - p^2) = np(1-p) \checkmark$$

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

(ugliest of ugliest)

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

$\vdots$

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

Y has MGF $\psi_Y(t)$,

then $\psi_X(t) = \psi_Y(t) \leftrightarrow$ iff X, Y have identical probability distributions

So the MGF (if it exists) uniquely characterizes a random variable.

Mean versus median	we've already made some contrasts between the mean and the median of a distribution;

here are 2 more things worth saying.

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,

as we have already seen:

$$E[h(X)] \neq h[E(X)]$$

unless $h(x) = ax + b$

Prediction

X rv with

mean μ_X, SD σ_X

Before X is observed, suppose your job

is to <u>predict</u> what its value will be;

what should you do? How can you tell

if a prediction is <u>good</u>?

Let's say you pick the number $\hat{x}$ ← x-hat (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 <u>bias</u> of $\hat{x}$ as a prediction for X is $\text{bias}(\hat{x}) \triangleq E(\hat{x} - X)$.

Def Your prediction $\hat{x}$ is <u>unbiased</u> if $\text{bias}(\hat{x}) = 0$.

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 <u>mean squared error</u> (MSE) of $\hat{x}$ as a prediction for X.

Small theorem:

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

Small 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 $\hat{x}$;

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

iff $\boxed{\hat{x} = E(X)}$

$$\frac{d^2}{d\hat{x}^2} = 2 > 0$$

so $E(X)$ is a minimum

Also easy to show

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

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

Yet another possible criterion for a good prediction $\hat{x}$ would be to find $\hat{x}$ such that $E\left[|\hat{x} - X|\right]$ is small.

Definition

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

Another small theorem

X rv with finite mean μ_X; let m_X be (a/the) <u>median</u> of X;

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

Reminder: why a/the?

Careful definition of median

X rv $\rightarrow$ every number m such that

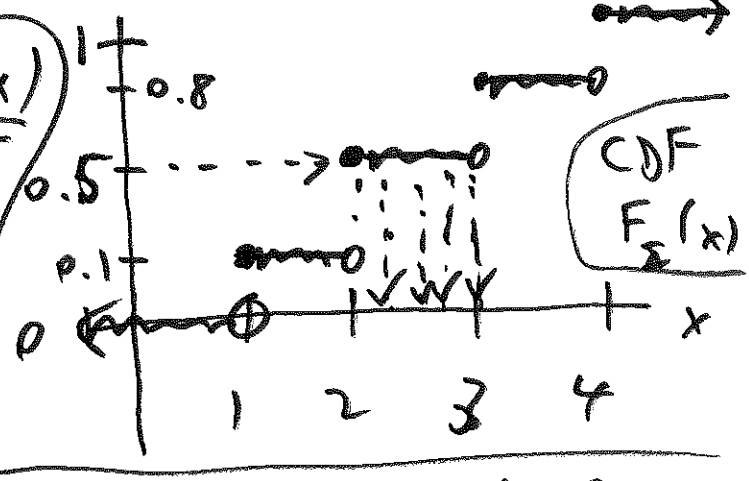
$P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$

is a <u>median</u> of the dist. of X

Example of nonunique median

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

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



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

Most people would say median $= 2.5$

which is a better criterion, MSE or MAE?

There is no universal right answer to this question: it depends on the real-world <u>consequences</u> of your prediction errors $(\hat{X} - X)$; quantifying these consequences involves the creation of a <u>utility function</u>, which we'll examine later.

Covariance & correlation

<u>Independence</u> 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 <u>related</u>.

Let's see how to <u>quantify</u> such relationships.

Def. X, Y rv with finite means $\mu_X = E(X)$ and $\mu_Y = E(Y)$. The <u>covariance</u> of X and Y, written $C(X,Y)$, is defined as

If use $Cov(X,Y)$

$$C(X,Y) = E\left[(X-\mu_X)\cdot(Y-\mu_Y)\right],$$

as long as this expectation exists

Consequences of this definition

① $(X-\mu_X)\cdot(Y-\mu_Y) =$

$$X\cdot Y - \mu_X \cdot Y - \mu_Y \cdot X + \mu_X \mu_Y$$

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

$$= E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y$$

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

← easier formula to compute with

(expectation of product − product of expectations)

② Sufficient condition for $C(X, Y)$ to exist: $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$.

③ Covariance is a good start at measuring strength of relationship, but it has a big flaw: its value depends on the units of measurement of X and Y

Example:

X = max temperature in °C

Y = relative humidity (%)

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

Y = yearly income ($)

$C(X, Y)$ comes out in (years)·($) ??

If you change your mind & measure temperature X' in °F $= \frac{9}{5}$°C $+ 32$°,

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

Easy to show that if a, b are fixed constants

then $C(aX+b, Y) = a\,C(X, Y)$ so

$C(X', Y) = 1.8 \cdot C(X, Y)$, i.e. you can

°F °C

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 $X' = \dfrac{X - E(X)}{SD(X)}$

$$= \frac{X - \mu_X}{\sigma_X}$$

(as long as $\sigma_X < \infty$, this is a meaningful definition)

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

Def. X, Y rv with finite variances σ_X^2 and σ_Y^2 (and therefore finite means μ_X and μ_Y) $\rightarrow$ the <u>correlation</u> 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 <u>invariant to linear transformation</u> 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 definition

① <u>Cauchy-Schwarz inequality</u>:

For all rv X, Y for which $E(XY)$ exists, $(E(XY))^2 \le (E(X))^2 \cdot (E(Y))^2$,

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

and $-1 \le \rho(X,Y) \le +1$.

Karl Schwarz (1843-1921) German mathematician

(associated)

Def $\rho(X,Y) > 0 \leftrightarrow X, Y$ <u>positively correlated</u>

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

$\rho(X,Y) = 0 \leftrightarrow X, Y$ <u>uncorrelated</u>

② X, Y <u>independent</u> rv with $0 < \sigma_X^2 < \infty$, $0 < \sigma_Y^2 < \infty$

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

So independence implies correlation, (216)

but (interestingly) not the converse:

$\boxed{\text{Example:}}$ $X \sim \text{Uniform}\{-1, 0, +1\}$, $Y \triangleq X^2$.

$E(X) = 0$

$\rightarrow$ 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)

and thus

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

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

---

(3) X rv with $0 < \sigma_X^2 < \infty$, $Y = aX + b$

for $\left\{\begin{matrix} a \neq 0 \\ b \end{matrix}\right\}$ constants $\rightarrow$ $(a > 0)$ $\rho(X, Y) = +1$

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

measures the strength of <u>linear</u> association between X and Y.

④ Important:

X, Y rv, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty \longrightarrow$

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

⑤ a, b, c (any constants) $\rightarrow$ $C(aX, bY) = ab\,C(X, Y)$

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

Special case:

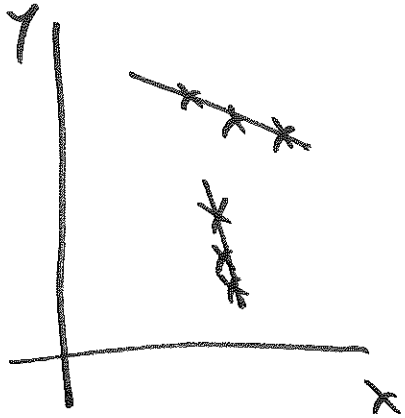
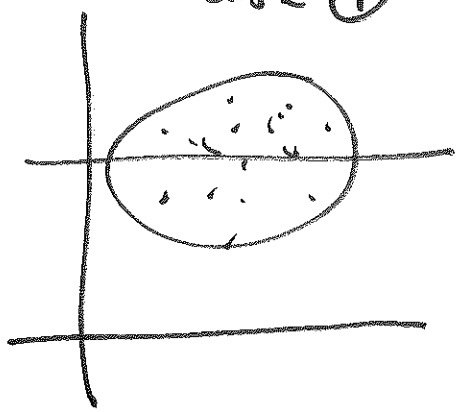
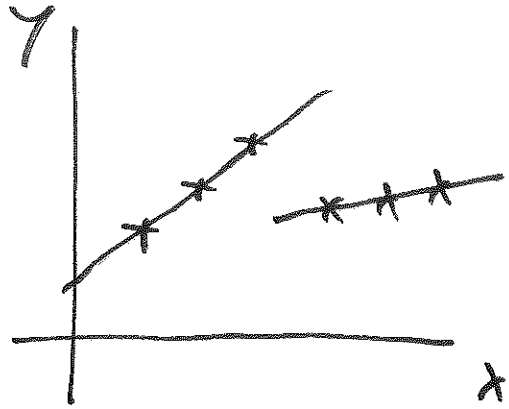
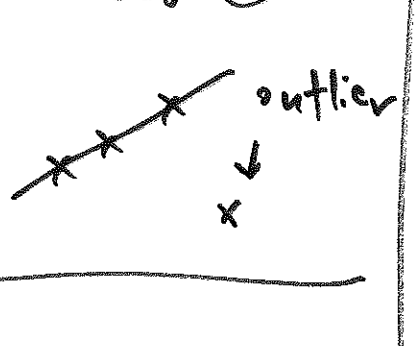
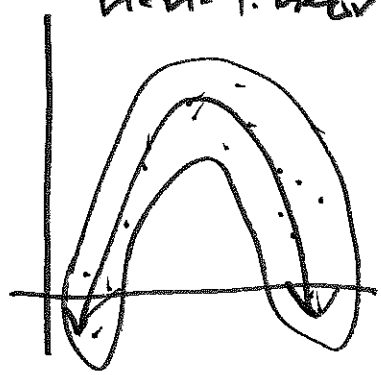
$$a^2 V(X) + b^2 V(Y) + 2ab\,C(X, Y)$$

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

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

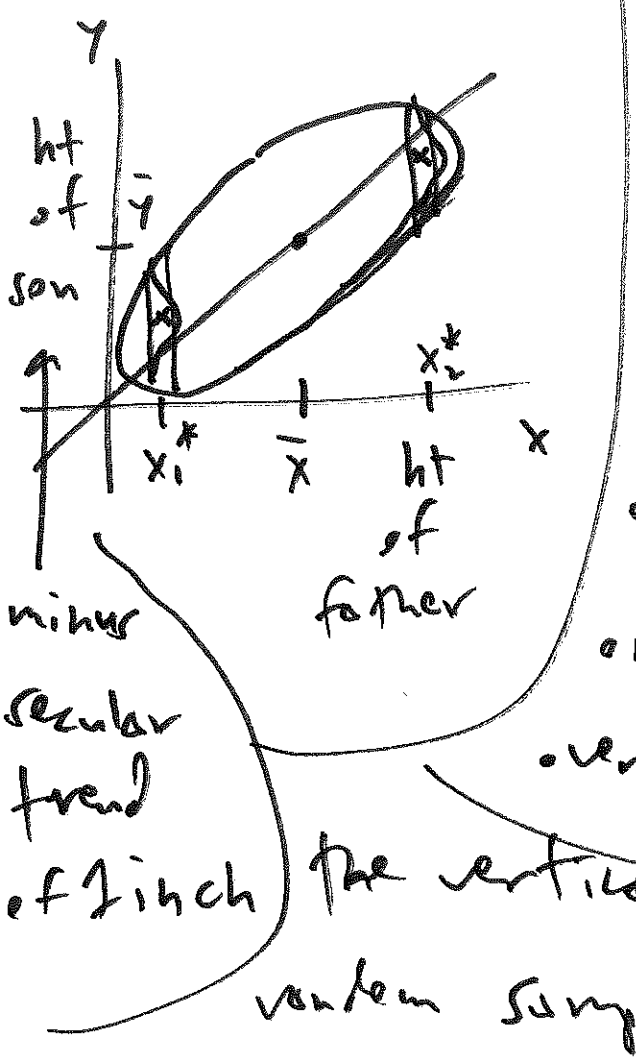
---

(7)

$\rho(X,Y) = -1$	$\rho(X,Y) = 0$	$\rho(X,Y) = +1$
	case (1) 	
points in scatterplot sample from $f_{X,Y}(x,y)$ all fall on line with negative slope (not necessarily -1)	case (2)  non-linearity 	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; i.e., 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 example again:

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 <u>conditional</u>

<u>distribution</u> of Y given $X = x_2^*$, $f_{Y|X}(y \mid x = x_2^*)$

Galton knew about the small theorem

text on p. 207: the number $\hat{w}$ that minimizes

the mean squared error (MSE) $E\left[(\hat{w} - W)^2\right]$ 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 $E\left[(\hat{y} - Y)^2\right]$ in the vertical strip

defined by $x \doteq x_2^*$ must be the <u>conditional</u>

<u>mean</u>, or <u>conditional expectation</u>, of the

rv $(Y \mid X = x_2^*)$

$\boxed{\text{Def.}}$ X, Y rv, Y finite mean $\rightarrow$

$\left\{\begin{array}{l}\underline{\text{conditional expectation}} \\ \text{(mean) of } Y \text{ given } X = x\end{array}\right\} = E(Y \mid x)$ is just

the expectation of the conditional distribution

$f_{Y|X}(y|x)$ of Y given $X = x$,

namely $E(Y|x) = \int_{\mathbb{R}} y\, f_{Y|X}(y|x)\, dy$

for continuous $(Y|X=x)$

and $E(Y|x) = \sum_{\text{all } y} y\, f_{Y|X}(y|x)$

for discrete $(Y|X=x)$

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

$\boxed{\text{Def.}}$ $h(x) \triangleq E(Y|X=x)$

then the <u>rv</u> $E(Y|X) \triangleq h(X)$ is the conditional expectation of Y given X

Clinical trial example, continued

$(n_c + n_T)$ people (a) who are similar in all relevant ways to $\left(\begin{matrix}\text{population} \\ \underline{\underline{P}}\end{matrix}\right) = \{$all adult patients with disease A$\}$

and (b) who consent to participate in your clinical trial are randomized, n_c to (the control group) (C) and n_T to (the treatment group) (T).

outcome of interest is dichotomous:

1 = disease went into remission (success)	
0 = didn't (failure)	

Let θ be the proportion of successes you would have seen if you could have put (everybody in P) into your treatment group; θ is unknown.

Let $S_i = \begin{cases} 1 & \text{if patient } i \text{ in the actual (T) group had a success} \\ 0 & \text{otherwise} \end{cases}$

Then the rvs $(S_i \mid \theta)$ are IID Bernoulli(θ)

and the rv $S = \sum_{i=1}^{n_T} S_i$ has a conditional

binomial dist: $(S \mid \theta) \sim \text{Binomial}(n_T, \theta)$

It's meaningful to talk about the conditional expectation rv. $E(S \mid \theta) = n_T \theta$ (a linear function of θ),

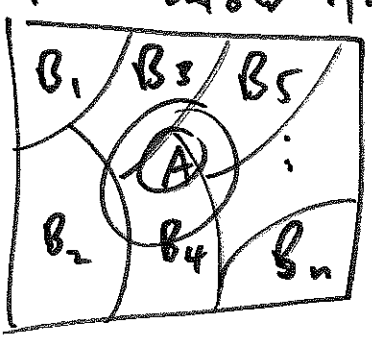
and – via Bayes' Theorem – it's even more meaningful to talk about the conditional expectation <u>rv</u> $E(\theta \mid S)$ (more about this later)

---

and the <u>constant</u> $E(\theta \mid S = s)$.

(15 Aug)

Remember the Law of Total Prob.!



$$P(A) = \sum_{i=1}^{n} P(B_i)\, P(A \mid B_i)$$

(LTP)

Important consequence of the def. of conditional expectation