

Consequences
of the
MGF definition

① X rv with MGF $\psi_X(t)$, (20)

$$Y = aX + b, \quad (a, b \text{ constants})$$

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

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

Example

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

$S_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$
($i=1, \dots, n$)

MGF of S_i

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

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

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

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

This was the
Law of the
unconscious
Statistician

(2) X_1, \dots, X_n independent r.v., MGF (202)

of X_i is $\psi_{X_i}(t)$, $Y = \sum_{i=1}^n X_i$,

MGF of Y is $\psi_Y(t) \rightarrow$ for every

t such that $\psi_{X_i}(t)$ is finite for all

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

MGF of
Binomial,
continued

(18 Aug 17)

Since the S_i are IID,

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

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

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

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

$$E(X) = \left(\frac{d}{dt} \psi_X(t) \right) \Big|_{t=0} = \frac{d}{dt} [pe^t + (1-p)]^n \Big|_{t=0} \quad \text{203}$$

$$= np \checkmark$$

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

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

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

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

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

$$E(X^3) = \left(\frac{d^3}{dt^3} [pe^t + (1-p)]^n \right) \Big|_{t=0} = \text{graph} \quad \begin{matrix} \text{T-S} \\ \text{PF} \end{matrix}$$

(using
& by lier)

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

∴

③ X has MGF $\psi_X(t)$, finite in an open interval around $t=0$ (204)

Y has MGF $\psi_Y(t)$

then $\psi_X(t) = \psi_Y(t) \iff$ 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.

(CDF F_X)

① X rv with values in an interval I ;
 $h(x)$ 1-1 function on I , $Y = h(X)$;

if m_X is a median of X (ie, 205)

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~~
~~Structure~~
Structure
 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?

Let's say you pick the number \hat{x} ⁽²⁰⁶⁾ $\leftarrow x\text{-hat}$
(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 bias of \hat{x} as a prediction for X is $\text{bias}(\hat{x}) \triangleq E(\hat{x} - X)$.

Def) Your prediction \hat{x} is unbiased

if $\text{bias}(\hat{x}) = 0$. Clearly, to achieve this just choose $\hat{x} = E(X)$.

Another possible criterion for goodness ⁽²⁰⁷⁾
would be to find \hat{x} such that $E(\hat{x} - X)^2$

is small.
(Gauss) Def. $E[(\hat{x} - X)^2]$ is called the

mean/squared error (MSE) of \hat{x} as

a prediction for X . Small ~~the~~ 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} ;

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

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

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

so $E(X)$ is a minimum

Also easy
to show

$$\begin{aligned} \text{MSE}(\hat{x}) &= E(\hat{x} - X)^2 \quad \text{(LOF)} \\ &= V(X) + [\text{bias}(\hat{x})]^2 \end{aligned}$$

So the choice $\hat{x} = E(X)$ ^{both} minimize, $\text{MSE}(\hat{x})$ and achieves 0 bias, and

with this choice $\text{MSE}(\hat{x}) = V(X) = \sigma_X^2$

A different
criterion

Yet another possible criterion for a good prediction \hat{x} would be to find \hat{x} such

that $E[|\hat{x} - X|]$ is small.

(Laplace)

Definition

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

Another small theorem (Laplace)

X rv with finite mean μ_X ; (209)
 let m_X be (a/the) median of X ;

\rightarrow the x that minimizes $MAD(x)$

is (a/the) median m_X . Reminder: why (a/the)?

Careful definition of median

X rv \rightarrow every number m such that

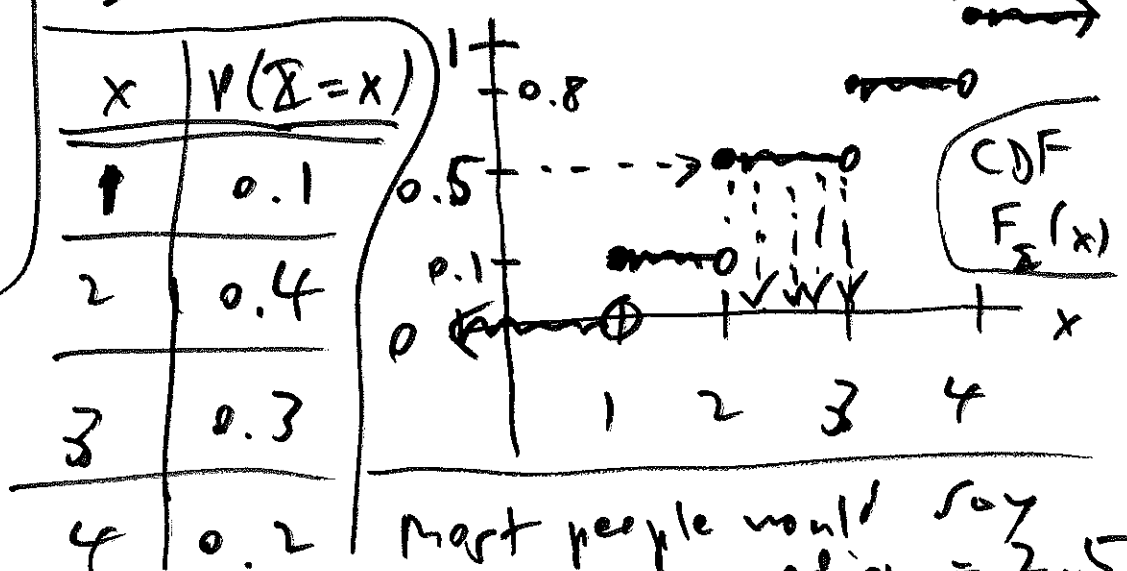
$P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq \frac{1}{2}$

is a median of the dist. of X

Example of nonunique median

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

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



Most people would say median = 2.5

which is
a better
criterion,
MSE or
MAE?

There is ^{universal} no right answer (210)
to this question: 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
& 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.

Let's see how to quantify
such relationships.

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

If use $Cov(X, Y)$ $C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, 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 $C(X, Y) = E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$

$C(X, Y) = E(XY) - \mu_X \mu_Y$ (expectation of product - product of expectations) easier formula to compute with

② Sufficient condition for $C(X, Y)$ to (212)

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 daily} temperature

in $^{\circ}\text{C}$ ^{max daily}

$Y =$ ^{relative} humidity (%)

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

$Y =$ yearly income (\$)

$C(X, Y)$ comes out in

(years) \cdot (\$) (??)

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

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

Easy to show that if a, b are ^{fixed} constants (23)

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

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

of C make the association between temperature & relative humidity seem larger just by switching from $^{\circ}C$ to $^{\circ}F$ (???)

Easy fix:

Def The process of converting a var X to standard units (SU) is achieved with

the linear transformation $X' = \frac{X - E(X)}{SD(X)}$

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

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

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

Def. / X, Y rv with finite variances (214)
 σ_X^2 and σ_Y^2 (and therefore finite means μ_X and μ_Y) \rightarrow the correlation of X
 and Y is $\rho(X, Y) = E \left[\left(\frac{X - \mu_X}{\sigma_X} \right) \cdot \left(\frac{Y - \mu_Y}{\sigma_Y} \right) \right]$
 (labeled "rho" ("row"))

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

for any constants $a, c \geq 0$ and $b, d,$

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

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

(19 Aug 19)

Consequences
of the
correlation
definition

① Cauchy - Schwarz inequality:

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

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

or $-1 \leq \rho(X, Y) \leq +1$

Karl Schwarz
(1843-1921)
German
mathematician
(associated)

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

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

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

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

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

So independence implies ρ correlation, (2/16)
but (interestingly) not the converse:

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 $E(XY) = E(X^3)$

(since X and X^3 are
identically distributed) $= E(X) = 0$
and thus

$$C(X, Y) = \underbrace{E(XY)}_0 - \underbrace{E(X)}_0 \cdot E(Y) = 0$$

$$\therefore \rho(X, Y) = \frac{C(X, Y)}{\sigma_X \sigma_Y} = 0 \quad \text{and } X, Y \text{ are uncorrelated!}$$

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

$$(a < 0) \rho(X, Y) = -1 \quad \text{so } \rho(X, Y) \quad (217)$$

measures the strength of linear association between X and Y .

(4) Important:

(if)

$$X, Y \text{ rv, } \sigma_X^2 < \infty, \sigma_Y^2 < \infty \quad \text{then}$$

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

(bedrock data science formula)

(5) $\begin{matrix} a, b, c \\ \text{any} \\ \text{constants} \end{matrix}$ $C(aX, bY) = ab C(X, Y)$

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

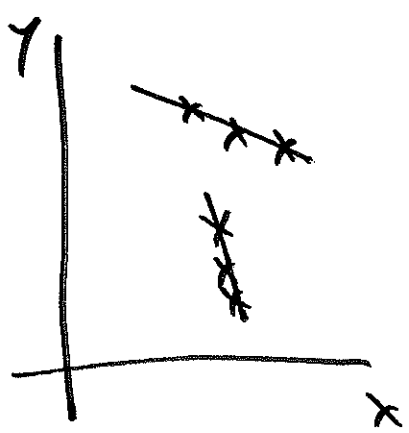
Special case:

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

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

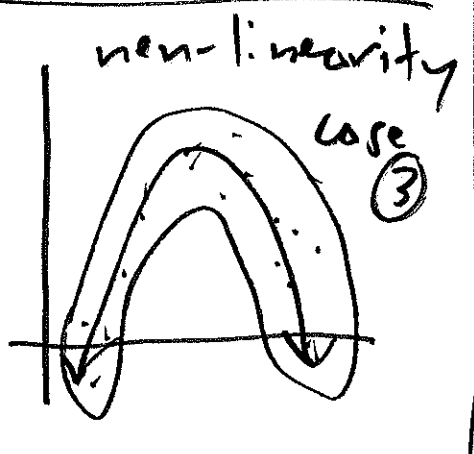
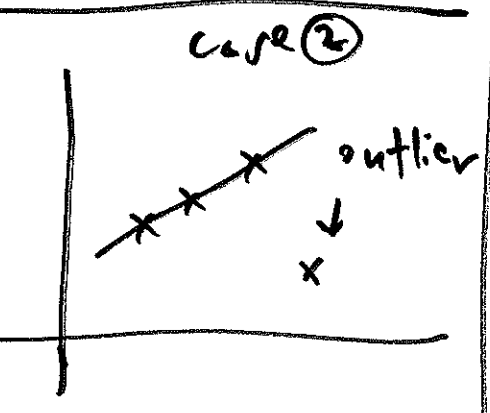
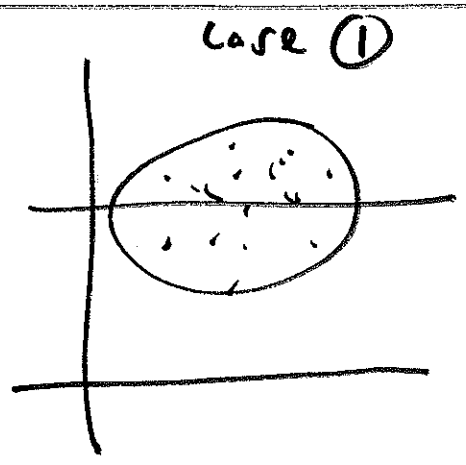
⑥ ⁽²¹⁸⁾ X_1, \dots, X_n such that (X_i, X_j) uncorrelated
 for all $1 \leq i \neq j \leq n \rightarrow$ (then) $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i)$

⑦ $\rho(X, Y) = -1$

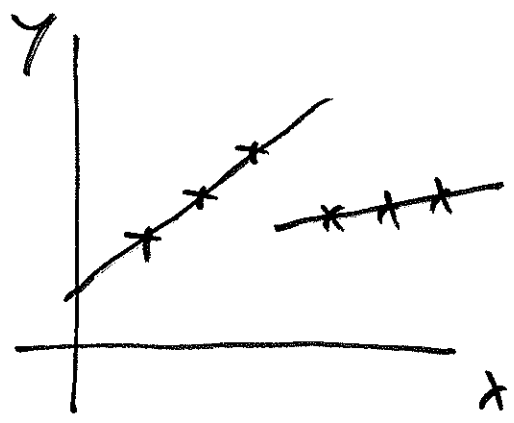


points in
 scatter plot
 sample from
 $f_{X,Y}(x,y)$
 all fall on line
 with negative
 slope (not
 necessarily
 -1)

$\rho(X, Y) = 0$



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

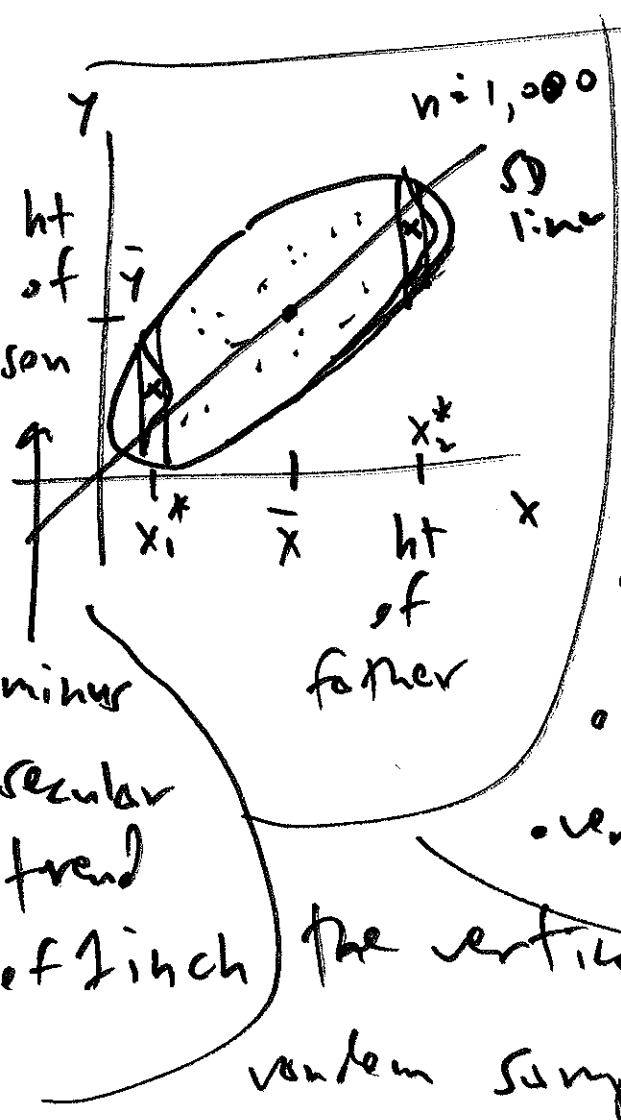


points in
 scatter plot
 sample from
 $f_{X,Y}(x,y)$
 all fall on line
 with positive
 slope (not
 necessarily
 +1)

(21 Aug 17)
 Conditional
 Expectation

X, Y related vrs (not independent): then there is

information in X for predicting Y ;
 i.e., we should be able to find some function
 $d: \mathbb{R} \rightarrow \mathbb{R}$ such that $d(X)$ is "close" in
 some sense to Y — what is the optimal d ?



Galton example ~~graph~~:

Galton divided the elliptical
 scatterplot up into a bunch
 of 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 $X = x_2^*$, $f_{Y|X}(y|x=x_2^*)$ (220)

Galton knew about the small theorem

but on p. (207): 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 "closeness" and concluded that the \hat{y} that minimizes the MSE $E[(\hat{y} - Y)^2]$ in the vertical strip defined by $x = x_2^*$ must be the conditional mean, or conditional expectation, of the

$v(Y | X = x_2^*)$ Def. E, Y, v, Y finite mean \rightarrow

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

the expectation of the conditional distribution (221)

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

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

for continuous ($Y|X=x$)

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

for discrete ($Y|X=x$)

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

the constant

when X is x . Def. $h(x) \triangleq E(Y|X=x)$

then the rv $E(Y|X) \triangleq h(X)$ is the

conditional expectation of Y given X . (21
Nov
19)

Clinical trial example, continued

$(n_C + n_T)$ people ^(a) who are similar in all relevant ways to

(population P) = { all adult patients with disease A }

and (b) who consent to participate in your clinical trial are randomized, n_C to ^{the} (control group) and n_T to ^{the} (treatment group) (c)

outcome of interest is dichotomous:

let θ be the proportion of successes you would have seen if you

(success)	1 = disease went into remission
(failure)	0 = did not

could have put (everybody in P) into your treatment group; θ is unknown.

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

Then the rvs $(S_i | \theta)$ are IID Bernoulli(θ) ⁽²³⁾

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

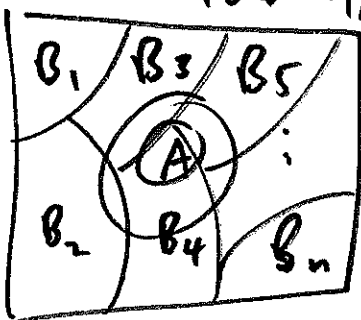
Binomial dist: $(S | \theta) \sim \text{Binomial}(n_T, \theta)$

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

and - via Bayes' Theorem - it's even more meaningful to talk about the conditional expectation rv. $E(\theta | S)$ (more about this later)

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

Remember the Law of Total Prob.!



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

(LTP)

Important consequence of the def. of conditional expectation

Continuous version of LTP

X, Y continuous r.v. (224)

for which all named densities exist \rightarrow

$$\frac{f_X(y)}{P(A)} = \int_{-\infty}^{\infty} \frac{f_X(x)}{P(B_i)} \cdot f_{Y|X}(y|x) dx$$

Earlier we agreed that, by definition,

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

So watch the following slightly magical calculation:

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

ifok to interchange order of integration

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

$$= \int_{-\infty}^{\infty} f_X(x) \cdot E(Z|x) dx, \quad (*)$$

is of the form { weighted average of $E(Z|x)$,
with $f_X(x)$ as the weights }

Recall that
continuous
for any r.v. \underline{W} ,

$$E(\underline{W}) = \int_{-\infty}^{\infty} w f_{\underline{W}}(w) dw$$

and

$$E(h(\underline{W})) = \int_{-\infty}^{\infty} h(w) f_{\underline{W}}(w) dw \quad (\text{LOTUS})$$

so (*) is just

$$E_X[E(Z|X)]$$

and we have shown that (Adam)

$$E(Z) = E_X[E(Z|X)]$$

This is referred to as part (1) of the
double expectation theorem; strangely, I
don't even mention that name, calling it instead
the LTP for expectations.