This is a handy theorem: if its premise is satisfied & the calculations are manageable, you get <u>all</u> the moments of X just by computing $\psi_X(t)$ and differentiating it over & over.

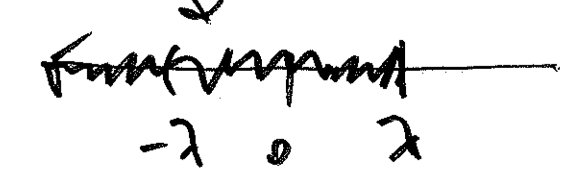
Example (16 May 19)

$X \sim \text{Exponential}(\lambda)$ $(\lambda > 0)$

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

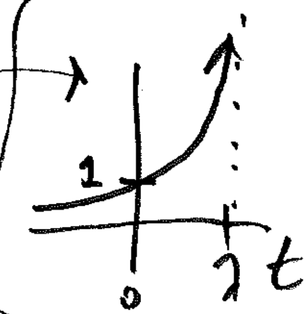
$$\psi_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(t-\lambda)x}\, dx$$

Now this integral is finite only if $t - \lambda < 0$, i.e. 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 we just crank out the derivatives:

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

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

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

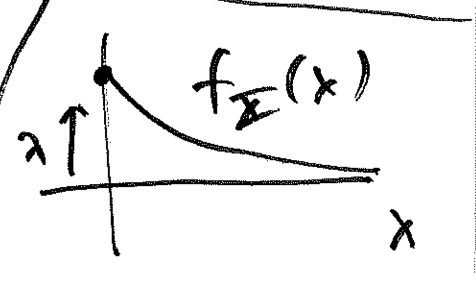
and $SD(X) = \frac{1}{\lambda}$

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

positive skew (long right-hand tail)

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

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



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: $\psi_{S_i}(t) = E\left(e^{t S_i}\right)$

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

$$= \left[p e^t + (1 - p)\right]$$

This uses the law of the unconscious statistician (LotUS)

② $X_1, \dots, X_n$ independent rv, MGF 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 $i = 1, \dots, n$, $\psi_Y(t) = \prod_{i=1}^{n} \psi_{X_i}(t)$.

(18 Aug 17)

~~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.

203

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

$$= np \quad ✓$$

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

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

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

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

$$= n(p - p^2) = np(1-p) \quad ✓$$

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

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

$\vdots$

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

Y has MGF $\psi_Y(t)$,

then $\psi_X(t) = \psi_Y(t) \overset{\text{iff}}{\longleftrightarrow}$ 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 with values in an interval I; (CDF F_X)

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

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

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

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?

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 *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 would be to find $\hat{x}$ such that $E(\hat{x}-X)^2$ is small. (Gauss)

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

Small theorem:

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

Small proof

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

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

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

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

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 $\hat{0}$ 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 $E[|\hat{x} - X|]$ 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

X rv with finite mean μ_X; let m_X be (a/the) <u>median</u> of X; → 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 → every number m such that

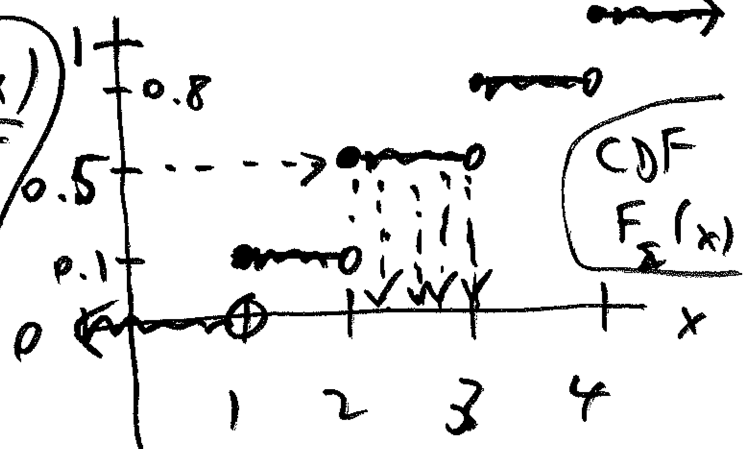
$$P(X \le m) \ge \frac{1}{2} \text{ and } P(X \ge m) \ge \frac{1}{2}$$

is <u>a 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



CDF $F_X(x)$

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 consequences of your prediction errors $(\hat{X} - X)$; quantifying these consequences involves the creation of a utility function, which we'll briefly examine later.

Covariance & correlation

bedrock

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

DS use $\mathrm{Cov}(X, Y)$

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

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

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

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

much 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 = education level (years of schooling completed)

Y = yearly income (\$)

$C(X, Y)$ comes out in (years) $\cdot$ (\$) ??

Example:

X = max daily temperature in $^\circ C$

Y = max daily relative humidity (%)

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

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

Easy to show that if a, b are fixed constants then $C(a\underline{X}+b, \underline{Y}) = a\,C(\underline{X}, \underline{Y})$ so

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

$\uparrow$ °F $\qquad \uparrow$ °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 $\underline{X}$ to standard units (SU) is achieved with the linear transformation $\underline{X}' = \dfrac{\underline{X} - E(\underline{X})}{SD(\underline{X})} = \dfrac{\underline{X} - \mu_{\underline{X}}}{\sigma_{\underline{X}}}$

(as long as $0 < \sigma_{\underline{X}} < \infty$, this is a meaningful definition)

$E(\underline{X}') = 0, \quad V(\underline{X}') = 1 = SD(\underline{X}')$

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 correlation of X and Y is

rho ("row")

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

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

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 definition

① Cauchy – Schwarz inequality:

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

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

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

Karl Schwarz (1843 – 1921) German mathematician

(associated)

Def | $\rho(X,Y) > 0 \leftrightarrow X, Y$ 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 $\begin{cases} 0 < \sigma_X^2 < \infty \\ 0 < \sigma_Y^2 < \infty \end{cases}$

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

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

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

$E(X) = 0$

→ 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 $\left\{ \begin{array}{l} a \neq 0 \\ b \end{array} \right\}$ constants → $(a > 0)$ $\rho(X, Y) = +1$

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

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

4 Important:

if X, Y rv, $\sigma^2_X < \infty$, $\sigma^2_Y < \infty$ $\longrightarrow$ then

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

(~~bedrock data science~~ formula)

5 (a, b, c any constants)

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

$\sigma^2_X < \infty$, $\sigma^2_Y < \infty \rightarrow V(aX + bY + c) =$

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

Special case:

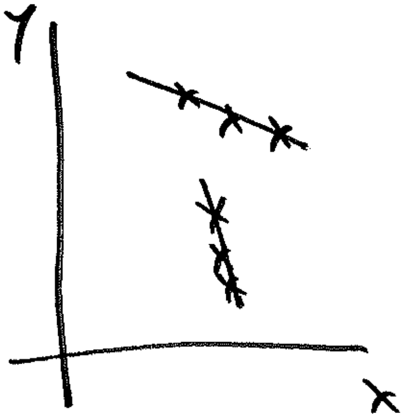
$$V(X - Y) = V(X) + V(Y) - 2C(X, Y).$$

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

---

(7)

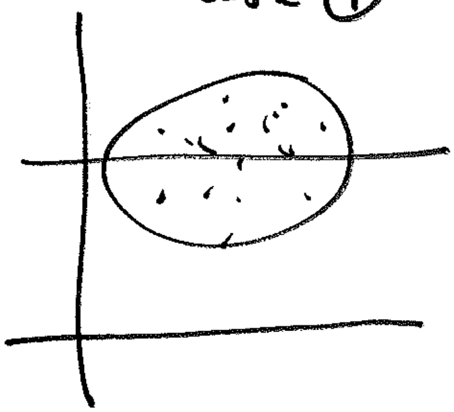
$\rho(X, Y) = -1$



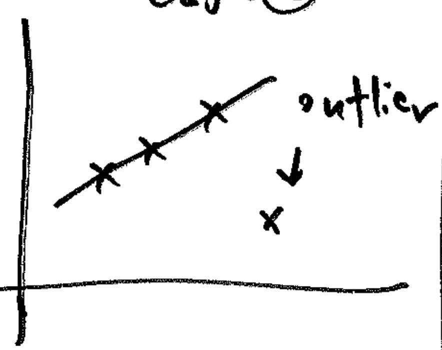
points in scatteplot sample from $f_{X,Y}(x, y)$ all fall on line with negative slope (not necessarily -1)

$\rho(X, Y) = 0$

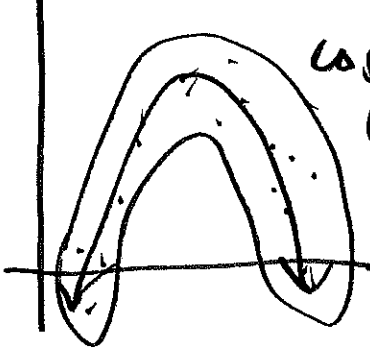
case (1)



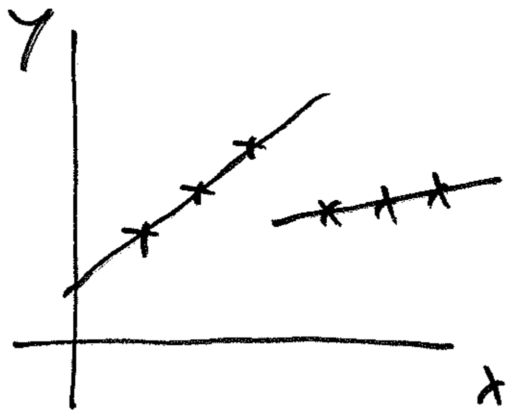
case (2)



non-linearity case (3)



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

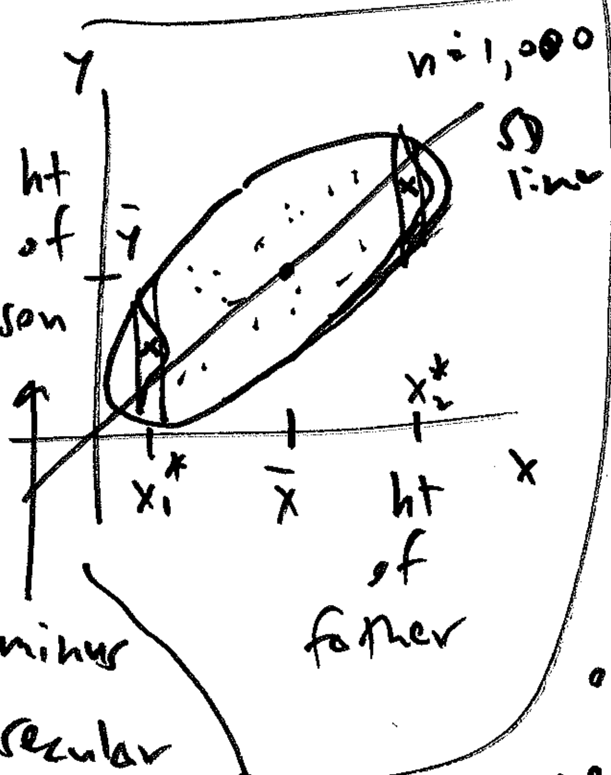


points in scatterplot sample from $f_{X,Y}(x, y)$ all fall on line with positive slope (not necessarily $+1$)

(21 Aug 17)

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?



minus secular trend of 1 inch

Galton example:

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 $X = x_2^*$, $f_{Y|X}(y \mid x = x_2^*)$

Galton knew about the small theorem last 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 conditional mean, or conditional expectation, of the rv $(Y \mid X = x_2^*)$

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

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

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 the constant x.

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

then the <u>rv</u> $E(Y|X) \hat{=} h(X)$ is the conditional expectation of Y given X. (21 Nov 19)