(2) $X_1, \ldots, X_n$ independent rv, mgf of X_i is $\psi_{X_i}(t)$, $Y = \sum_{i=1}^{n} X_i$, mgf of Y is $\psi_Y(t)$ + 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 (18 Aug 17)

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.

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

(203)

$$= np \checkmark$$

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

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

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

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

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

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

T-S PF

(ugliest & uglier)

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

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

Y has MGF $\psi_Y(t)$

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

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

small theorem:

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

small proof

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

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

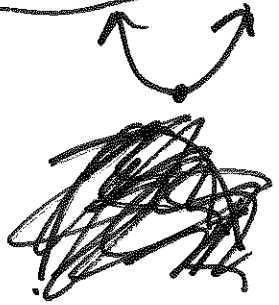
This is a quadratic function of $\hat{x}$;

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

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 $E[|\hat{x} - X|]$ 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) median of X;

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

Reminder: why a/the?

Careful definition of median

X rv → every number m such that

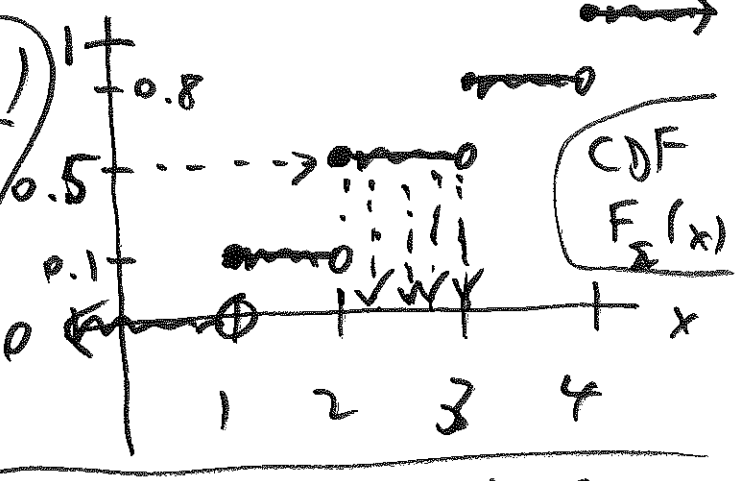
$P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$

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

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 "$E(X)$"

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

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

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

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

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

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 = temperature (max in °C)

Y = humidity (% relative max)

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 $+ 32$°,

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

Easy to show that if a, b are fixed constants then $C(aX+b, Y) = aC(X, Y)$ so

$C(X', Y) = 1.8 \cdot C(X, Y)$, 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 $X' = \dfrac{X - E(X)}{SD(X)} = \dfrac{X - \mu_X}{\sigma_X}$

(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) → 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 $a, c \neq 0$ 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,

$$\left(E(XY)\right)^2 \le \left(E(X)\right)^2 \cdot \left(E(Y)\right)^2$$

Karl Schwarz (1843 - 1921) German mathematician

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

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

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

$\rho(X,Y) < 0 \leftrightarrow X, Y$ 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 a correlation,

but (interestingly) not the converse:

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

$E(X) = 0$

→ 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

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

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

$(a<0)\ \rho(X,Y) = -1$ | so $\rho(X,Y)$ (21)

measures the strength of linear association between X and Y.

④ Important:

(if) X, Y rv, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$ → (then)

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

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

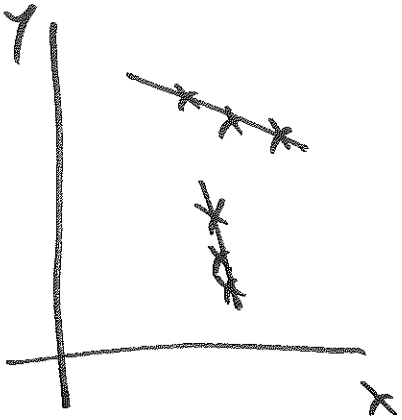
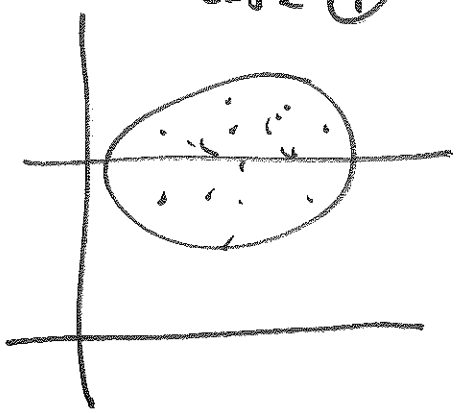
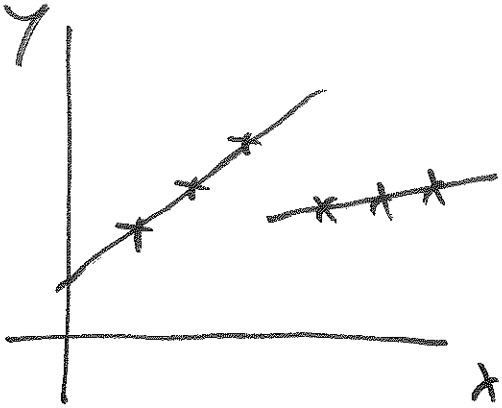
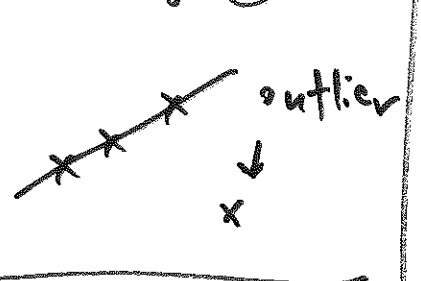
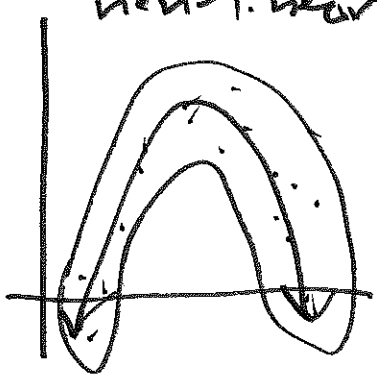
$\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty \rightarrow V(aX+bY+c) =$

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

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