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

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

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

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 (LotUS)

(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 $i = 1, \ldots, n$,

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

(18 Aug 17)

~~MGF of~~ Binomial, continued

Since the S_i are IID,

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

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

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

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)

$$= np \checkmark$$

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

$$\text{So } V(X) = E(X^2) - \left[E(X) \right]^2$$

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

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

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

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

T-S PMF

(ugliest & ugliest)

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

(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

~~(21 Aug 20)~~

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

$$MSE(\hat{x}) = E(\hat{x} - \Sigma)^2$$

$$= V(\Sigma) + [bias(\hat{x})]^2$$

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

A different criterion

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

Definition

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

Another small theorem (Laplace)

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

→ the x that minimizes $MAE(x)$ is (a/the) median m_X.

Reminder: why a/the?

Careful definition of median

X rv → every number m such that

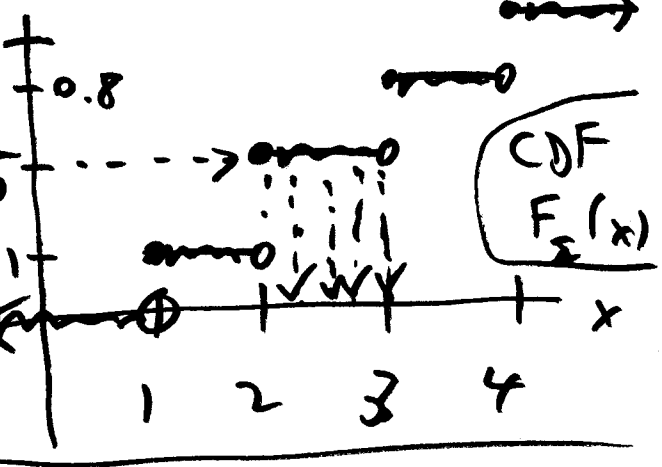
$P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$

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

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

(so long as this expectation exists)

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

Consequences of this definition

① $(X-\mu_X)\cdot(Y-\mu_Y) =$

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

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

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

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

(expectation of product − product of expectations)

much easier formula to compute with

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 daily 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 $+ 32$°,

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

Easy to show that if a, d 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 ("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 (either/both):

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

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

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

(19 Aug 19)

Consequences of the correlation definition

① Cauchy-Schwarz inequality:

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

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

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

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

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

(22 May 20)

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 ● 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 $\left\{\begin{matrix} a \neq 0 \\ b \end{matrix}\right\}$ 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, $\sigma_X^2 < \infty$, $\sigma_Y^2 < \infty$ → then

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

(bedrock data science formula)

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

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

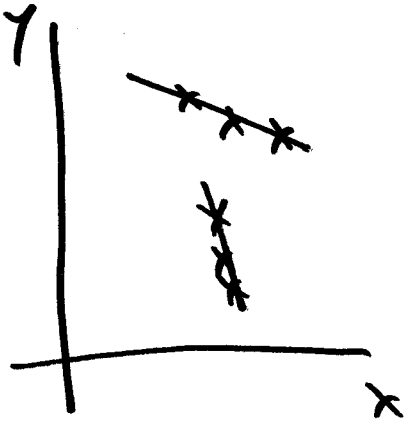
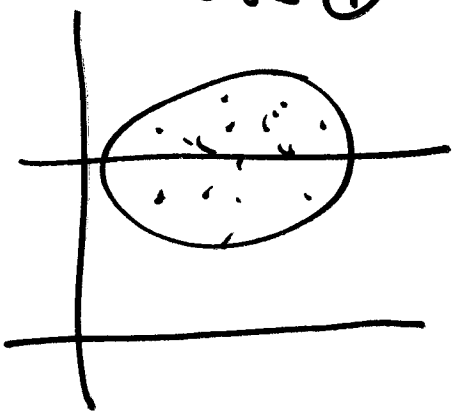
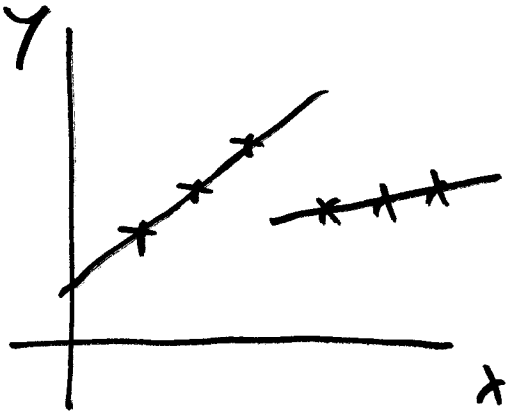
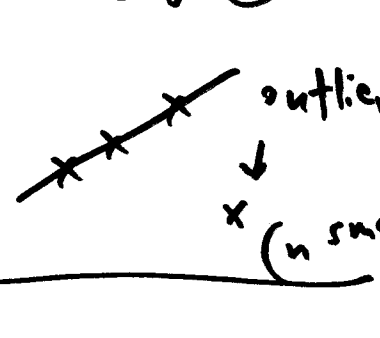
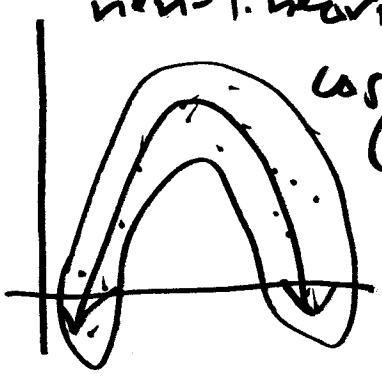
special case:

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

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

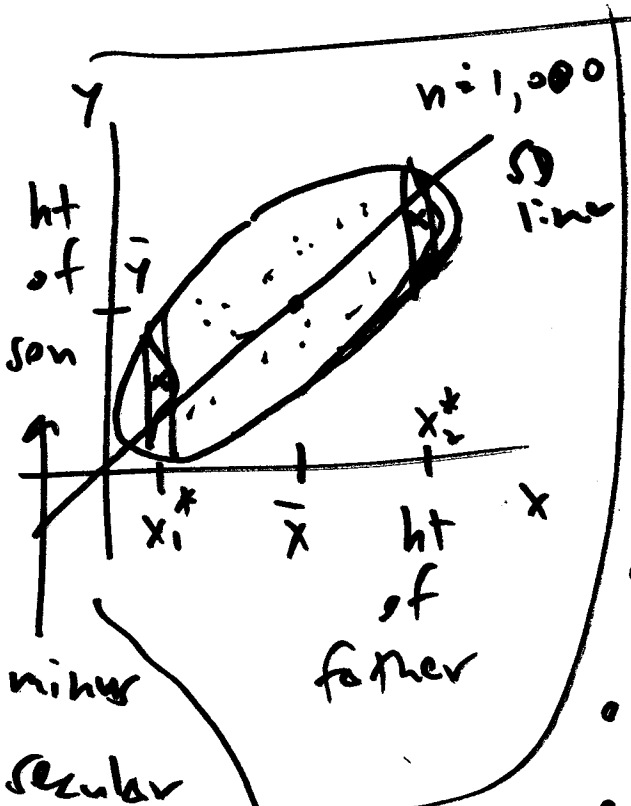
(6) (If) $X_1, \ldots, X_n$ such that (X_i, X_j) uncorrelated for all $1 \le i \ne j \le n$ (then) $V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$

(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 $+1$)
	non-linearity case ③ 	

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:

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

Galton knew about the small theorem

back 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 $x = x_2^*$ must be the conditional mean, or conditional expectation, of the rv $(Y|X=x_2^*)$

Def. X, Y rv, Y finite mean $\rightarrow$

{conditional expectation (mean) of Y given $X=x$} $= E(Y|x)$ 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 for, $E(Y|x)$ is just a constant, equal to the conditional mean of Y when X is the constant 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 18)

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) (C) and n_T to (the treatment group) (T).

Outcome of interest is dichotomous:

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

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 = { 1 if patient i in the actual (T) group had a success; 0 otherwise

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

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

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

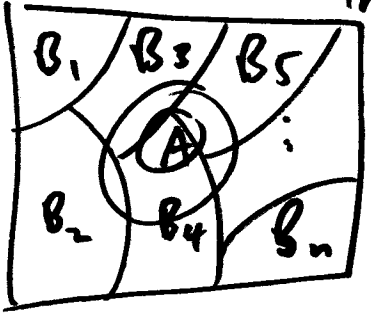
It's meaningful to talk about the conditional expectation rv. $E(S|\theta) = n_T\theta$ (a linear function of θ),

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.

for which all named densities exist →

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

Earlier we agreed that, by definition,

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

So watch the following slightly magical calculation:

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy$$

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

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

(ifok to interchange order of integration)

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

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

Recall that for any continuous r.v. W,

$$E(W) = \int_{-\infty}^{\infty} w f_W(w)\, dw$$

and

$$E[h(W)] = \int_{-\infty}^{\infty} h(w) f_W(w)\, dw \quad \text{(LOTUS)}$$

So $(*)$ is just

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

and we have shown that (Adam)

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

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

I need to postpone examples of these conditional expectation calculations until we've covered more standard distributions.

[redacted] X, Y r.v. such that $f_{Y|X}(y|x)$ exists → it makes sense to speak not only of $E(Y|x)$, the mean of $f_{Y|X}(y|x)$, but also of the variance of that dist.

Def the number $V(Y|x) \triangleq E_X\left\{\left[Y - E(Y|x)\right]^2 | x\right\}$ $= g(x)$ is called the conditional variance of Y given $X = x$, and the rv $V(Y|X)$ is just $g(X)$, the conditional variance of Y given X.

The payoff from all of this

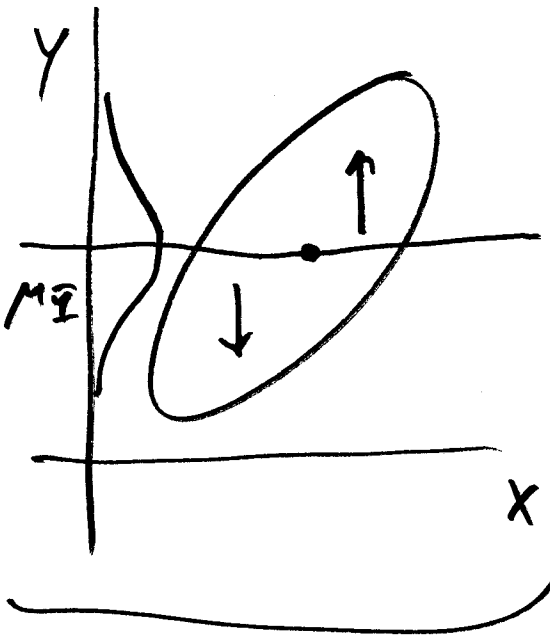
(formalizing Galton's intuition)

Theorem X, Y related rv's, want to use some function $\hat{Y} = d(X)$ to predict Y from X $\rightarrow$ the prediction $\hat{Y} = d(X)$ that minimizes the MSE $E(Y - \hat{Y})^2 = E\{[Y - d(X)]^2\}$ is $\hat{Y} = d(X) = E(Y|X)$, the conditional expectation of Y given X.

Part (2) of the double expectation theorem

X, Y r.v. such that all of the following expressions exist, $\rightarrow$

$$V(Y) = E_X[V(Y|X)] + V_X[E(Y|X)]. \quad \text{(Eve)}$$



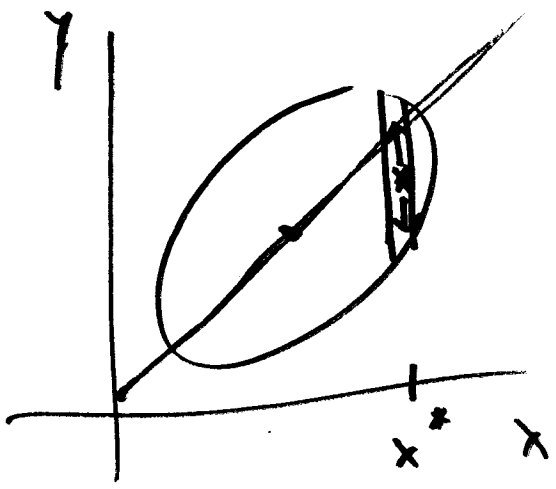
Imagine a 2-port game!

Stage 1 Predict I without knowing X. Well, if you buy into MSE as your measure of "goodness" of a prediction, we know that you should predict $\hat{I}_{\text{no } X} = \mu_I = E(I)$

and your resulting MSE will be

$$E\left[(I - \mu_I)^2\right] = V(I) = \sigma_I^2.$$

Stage 2 Observe X, now predict I



Let's say $X = x^*$. Then we know the MSE-optimal prediction is $\hat{I}_{X=x^*} = E(I|X=x^*)$

and your resulting MSE will be

$$E\left\{\left[\underline{Y} - E(\underline{Y} \mid \underline{X} = x^*)\right]^2\right\} = \underbrace{V(\underline{Y} \mid x^*)}_{(**)}.$$

From the vantage point of someone thinking about stage 2 before it happens, $\underline{X}$ is not yet known, so the expected value of (**), namely $E_{\underline{X}}[V(\underline{Y} \mid \underline{X})]$, is the best you can do to guess at how good the stage 2 prediction will be.

The second part of the double expectation theorem says

$$\underbrace{V(\underline{Y})}_{\text{MSE of } \hat{\underline{Y}}_{no\,\underline{X}}} = \underbrace{E_{\underline{X}}[V(\underline{Y} \mid \underline{X})]}_{\text{"E(MSE)" of } \hat{\underline{Y}}_{\underline{X}} = E(\underline{Y} \mid \underline{X})} + \underbrace{V_{\underline{X}}[E(\underline{Y} \mid \underline{X})]}$$

But since variances are always non-negative,

$$V_X[E(Y|X)] \geq 0, \text{ so}$$

$$E_X[V(Y|X)] + V_X[E(Y|X)] \geq E_X[V(Y|X)]$$

$$V(Y) \geq \text{"E(MSE)" of } \hat{Y}_X$$

MSE of $\hat{Y}_{no\ X}$

Thus you always expect your predictive accuracy to get better (or at least stay the same) when you use $E(Y|X)$ to predict Y.

Utility

Q: How to take action sensibly when the consequences are uncertain?

Another complete switch in subject

A: There is a theory of optimal action under uncertainty; it's called Bayesian decision theory — a concept called utility is central to this theory. The theory takes its simplest form when comparing gambles

Example X has discrete PMF $f_X(x) = \begin{cases} \frac{1}{2} & x = -\$350 \\ \frac{1}{2} & x = +\$500 \\ 0 & \text{else} \end{cases}$

Suppose X = your net gain from gamble (A),

Y has discrete PMF $f_Y(y) = \begin{cases} \frac{1}{3} & y = \$40 \\ \frac{1}{3} & y = \$50 \\ \frac{1}{3} & y = \$60 \\ 0 & \text{else} \end{cases}$

and Y = your net gain from gamble (B).

Turns out that $E(X) = \$+75$, $E(Y) = \$+50$ so is (A) automatically better than (B)?

Note that with (B) you're guaranteed to win at least $40, while (A) has no such guarantee; is (A) still automatically better for you than (B)? A risk-averse person would grab (B) quickly; a probably risk-seeking person would pick (A).

Evidently something more than just computing $E(X)$, $E(Z)$ is going on.

Def. of utility function

Your utility function $U(x)$ is that function which assigns to each possible net gain $-\infty < x < \infty$ a real # $U(x)$ representing the value to you of gaining x.

Q: If x is money, why not just use $u(x) = x$? (utility: linear in money)

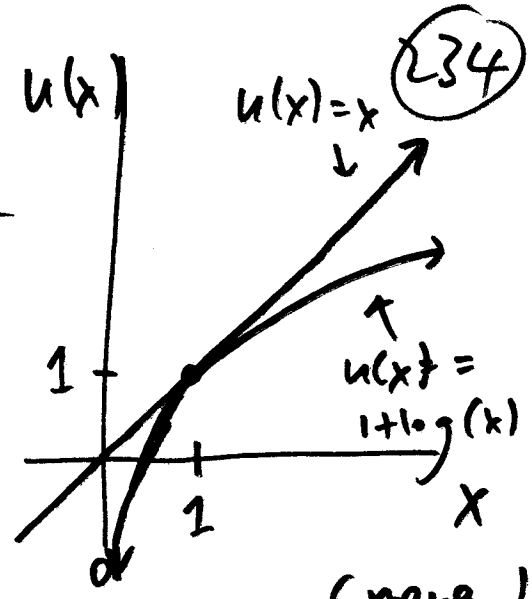
A: Lovely, subtle answer first supplied by Daniel Bernoulli (1700-1782), (Swiss mathematician) related to Jacob Bernoulli (1654-1705), for whom the Bernoulli distribution was named.

Daniel B: If your entire net worth is (say) $10, then the value to you of a new $1 is much greater than if your entire net worth is (say) $1,000,000; thus the utility of money is sublinear (meaning that it doesn't grow with x as fast as $f(x) = x$ does)

Daniel B proposed one particular sublinear function for utility,

namely $u(x) = 1 + \log(x)$ (for $x > 0$)

(Daniel B also invented the word utility) (Although the idea goes back at least to Aristotle (384-322 BCE))



Definition

(Principle of Expected Utility Maximization)

You are said to choose between gambles by maximizing expected utility (MEU) if, with $u(x)$ your utility function,

① you prefer gamble X to gamble Y if $E[u(X)] > E[u(Y)]$ and ① you're indifferent between X and Y if $E[u(X)] = E[u(Y)]$.

MEU first explored in depth by British (mathematician, philosopher, economist) Frank Ramsey (1903-1930), who died at age 26 of liver failure. (hepatitis)

Theorem (von Neumann-Morgenstern (1947))

Under 4 reasonable axioms, MEU is the best you can do.

John von Neumann
(1903-1957)
Hungarian-American
(mathematician, physicist, computer scientist, ⋮)
died at 53 of cancer

Oskar Morgenstern
(1902-1977)
German-economist
American

Simple example

Suppose you bought a single $2 ticket in the Powerball lottery examined in Take-Home Test 1 problem 2: the drawing on 30 Jul 2016 for which the Grand prize was $487 million. Let X be the unknown amount you will win (thinking about X before the drawing).

Match	x	$P(X=x)$	$x \cdot P(X=x)$
5w, 1R	\$487,000,000	$\frac{1}{292,201,338}$	\$1.667
5w, 0R	\$1,000,000	$\frac{1}{11,688,05352}$	0.086
4w, 1R	\$50,000	$\frac{1}{913,129.18}$	0.055
4w, 0R	\$100	$\frac{1/36,525.17}{}$	0.003
3w, 1R	\$100	$\frac{1/14,494.11}{}$	0.007
3w, 0R	\$7	$\frac{1/579.76}{}$	0.012
2w, 1R	\$7	$\frac{1}{701.33}$	0.010
1w, 1R	\$4	$\frac{1}{91.98}$	0.043
0w, 1R	\$4	$\frac{1}{38.32}$	0.104
			\$1.99 (!)

X has 9 possible values + (discrete),

so $E(X) = \sum_{\text{all 9 possibilities}} x \cdot P(X=x) = \$1.99$

Q: Before the drawing, someone offers you $\$x_0$ for your ticket; should you sell?

A: with $u(x)$ as your utility function, your expected gain if you keep the ticket is $E[u(X)]$; if for you $u(x) = x$ (utility $\triangleq$ money) then $E(u(X)) = \$1.99$

Action 1 (sell): you gain $\$x_0$ for sure

Action 2 (keep): your expected utility is $E[u(X)]$

Under MEU you should sell if $u(x_0) > E[u(X)]$

If $u(x) = x$ for you then your optimal action is (sell if offered more than $1.99).

Related but different problem

On the 13 Jan 2016 drawing the Powerball jackpot was $1.6 billion

X = your winnings

X uncertain before the drawing

Redo calculation on p. 236: $E(X)$ is now $5.80 on a $2 ticket

new 1st row in table is 1,600,000,000 / 292,201,338 ≈ $5.476

Q: If $u(x) = x$ for you, is it rational to sell all your assets & buy as many lottery tickets as possible! (under MEU) (*)

A: Yes, but that's a silly utility function; to be realistic you'd have to subtract from x the

monetary value (cost) to you of the disruption of your life that would ensue with action

(23 May 19)

(*) A catalog of useful distributions

(JSch.5) Case 1: Discrete Bernoulli

$X \sim \text{Bernoulli}(p)$, $0 < p < 1$, if

$$f_X(x) = p^x (1-p)^{1-x} \underbrace{I_{\{0,1\}}(x)}_{\text{support}(X)}$$

$$= \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{else} \end{cases}$$

$E(X) = p$ | $\psi_X(t) = pe^t + (1-p)$ for all $-\infty < t < \infty$

$V(X) = p(1-p)$ | $SD(X) = \sqrt{p(1-p)}$

Def If the X_i in $X_1, X_2, \ldots$ are IID Bernoulli (p), then $(X_1, X_2, \ldots)$ are called Bernoulli trials with parameter p; if the sequence $(X_1, X_2, \ldots)$ is infinite this defines a Bernoulli (stochastic) process.

Binomial $X \sim \text{Binomial}(n, p)$ (i.e., X follows the Binomial distribution with parameters n (positive integer) and $0 < p < 1$)

$$\leftrightarrow f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \underbrace{I_{\{0,1,\ldots,n\}}(x)}_{\text{support}(X)}$$

Consequences ① $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$

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

$X \sim \text{Binomial}(n, p)$ | $E(X) = np$ | $V(X) = np(1-p)$

$$\psi_X(t) = [pe^t + (1-p)]^n \quad \text{for all } -\infty < t < \infty$$

$$SD(X) = \sqrt{np(1-p)}$$

Case Study | Supreme Court case | Castaneda v. Partida (1977)

Grand juries in the U.S. judicial system have catchment areas: everybody 18 & over living in the judicial district for that grand jury (& a few other minor restrictions)

Hidalgo county, Texas — extreme southern border of TX with Mexico

eligible pool was 79.1% Mexican-American

$2\frac{1}{2}$ yr period at issue in Supreme Court case: 220 people called to serve on grand juries, but only 100 of them are Mexican-American

Q: Prima facie case of discrimination?

Before this 2½ yr period, let X be your prediction of # of Mexican-Americans among the 220 people

If no discrimination,

$X \sim \text{Binomial}(220, 0.791)$ $(X \mid T_1)$ → T_1 = theory! = no discrimination

$E(X \mid T_1) = (220)(0.791) \doteq 174.0$ ($n \cdot p$)

$SD(X \mid T_1) = \sqrt{n p (1-p)} \doteq 6.0$

Q: If you were expecting 174 give or take 6, would you be surprised to see 100?

A: You'd be astonished

Frequentist statistical answer: $P(X \leq 100 \mid T_1) = 8.0 \cdot 10^{-28}$ so T_1 looks ridiculous

Bayesian statistical answer: Need to compute $P(T_1 \mid X = 100)$, not the other way around (later)

Hypergeometric A finite population has A elements of type 1 and B elements of type 2; total population size (A+B).

You choose n elements at random without replacement from this population (i.e., you take a simple random sample (SRS) of size n)

Let X = (#elements of type 1 in your sample)

Then (as noted in take-Home TS+ 1 problem 2) X follows the hypergeometric distribution with parameters (A, B, n). As we saw in that problem, the PMF of X is

$$f_X(x \mid A, B, n) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} I[\max\{0, n-B\} \le x \le \min\{n, A\}]$$

support(X)

for (A, B, n) non-negative integers with $n \le A+B$

Consequences

① $E(X) = n \cdot \frac{A}{A+B}$

② $V(X) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{A+B-n}{A+B-1}\right)$

Note that if your sampling had been with replacement (i.e., you take an IID sample), X would have been Binomial with the same value of n and $p = \frac{A}{A+B}$; in that case $E(X) = np = n\frac{A}{A+B}$ and

$V(X) = np(1-p) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$ (compare)

If you let $T=(A+B)$ be the total # of elements in the population,

sampling method	mean of $\bar{X}$	variance of $\bar{X}$
with repl. (IID)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$
without repl. (SRS)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{T-n}{T-1}\right)$

$0 \le \alpha = \frac{T-n}{T-1} \le 1$ is called the finite population correction (21 Aug 19)

3 special cases worth considering

(a) $(n=1)$ $\alpha=1 \leftrightarrow$ SRS = IID with only 1 element sampled

(b) $(n=T)$ $\alpha=0 \leftrightarrow$ If you exhaust the entire population with SRS, you have no uncertainty left.

(c) (n fixed, $\uparrow\uparrow$) a [redacted] $\leftrightarrow$ with a small sample from a large population, SRS $\doteq$ IID

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

$\leftrightarrow X$ has PF $f_X(x) = \dfrac{\lambda^x e^{-\lambda}}{x!} I_{\{0,1,\ldots\}}(x)$

support of X

$E(X) = \lambda$

$V(X) = \lambda$

} thus for the Poisson dist.

$\dfrac{V(X)}{E(X)} = 1$

Def. If $E(X)$ and $V(X)$ both exist and $E(X) \neq 0$, $\dfrac{V(X)}{E(X)}$ is called the variance-to-mean ratio (VTMR)

$\psi_X(t) = e^{\lambda(e^t - 1)}$

$-\infty < t < \infty$

The Poisson can be unrealistic as a consequence of its VTMR of 1, because

many rvs that represent counts of occurrences of events in time intervals of fixed length have VTMR > 1.

The Poisson & Binomial distributions both count the number of "successes" in a process unfolding in time, so it should not be surprising to find out that these 2 dist. are related:

---

when (n is large, p is close to 0), PMF of Binomial(n,p) $\doteq$ PMF of Poisson$(n \cdot p)$

Theorem n positive integer, $0 < p < 1$, $X \sim$ Binomial(n,p), $\lambda > 0$, $Y \sim$ Poisson(λ) / Choose any sequence

$\{p_n\}_{n=1}^{\infty}$ of values between 0 and 1 with

$\lim_{n\to\infty} n \cdot p_n = \lambda$

Then $f_X(x \mid n, p_n) \to$ PMF Binomial $n\to\infty$

Poisson process, revisited

Def

$f_Y(y \mid \lambda)$ PMF Poisson

A poisson process with rate λ per unit (or space, or volume, or...) time, is a stochastic process with two properties:

(a) # arrivals in every interval of time of length $t \sim$ Poisson(λt)

(b) #s of arrivals in all disjoint (non-overlapping) time intervals are independent

~~Care Study~~ ~~Example~~ Parasitic Protozoa in drinking water

There's a kind of parasitic

organism called cryptosporidium that's capable of getting into the public drinking water supplies; at one stage in their life cycle they're called oocysts. They can make people sick at a concentration of only 1 oocyst per 5 liters = 1.3 gallons of water

One problem is that it can be hard to detect these oocysts with water filtration. Suppose that, in the water supply of your city, oocysts occur according to a Poisson process with rate λ oocysts per liter, & that the filtering system your water utility company uses can capture all the oocysts in a water sample but only has

probability p of detecting each oocyst that's actually there. (& counting events are independent)

Set Y = # oocysts in t liters of water, and $X_i = \begin{cases} 1 & \text{if oocyst } i \text{ gets counted} \\ 0 & \text{else} \end{cases}$

X = # counted oocysts | Then $(X \mid Y = y) = \sum_{i=1}^{y} X_i$

Under these assumptions, $(X \mid Y = y) \sim \text{Binomial}(y, p)$

Q: What's the dist. of X? | A: By the Law of total probability

$$f_X(x) = P(X = x) = \sum_{y=0}^{\infty} P(Y = y) P(X = x \mid Y = y)$$

for all $x = 0, 1, \ldots$

in which $P(Y = y) = \dfrac{(\lambda t)^y e^{-\lambda t}}{y!}$ for $y = 0, 1, \ldots$