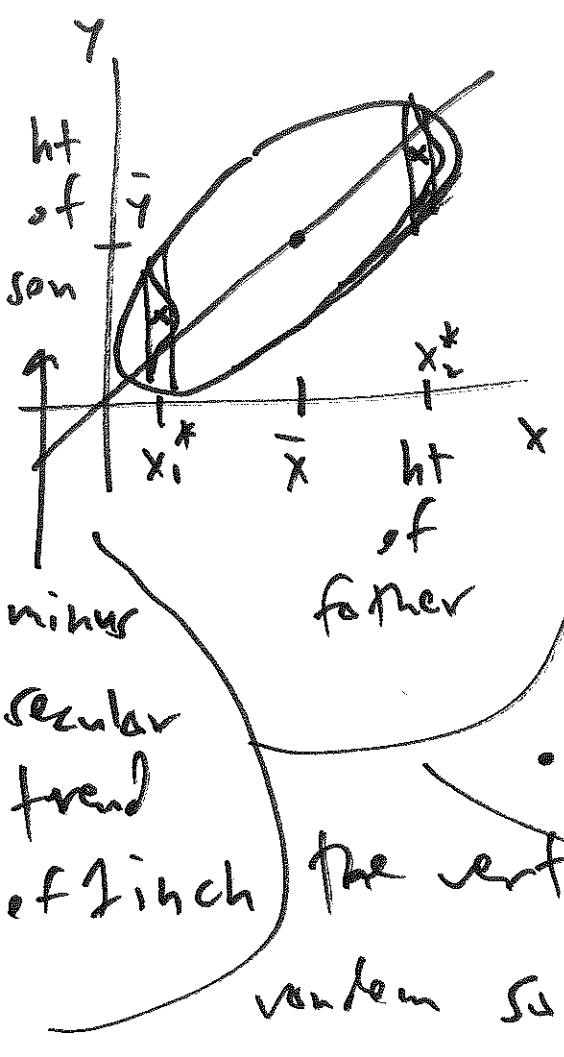
Conditional Expectation

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 $\left(Y | X = x_2^*\right)$

Def. X, Y rv, Y finite mean

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

the expectation of the conditional distribution

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

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.

Def. $h(x) \triangleq E(Y|X=x)$ the constant

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

$\left(\underset{\underline{P}}{\text{population}}\right)$ = {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)).

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.

outcome of interest is dichotomous:

(success) 1 = disease went into remission

(failure) 0 = didn't

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 $(S_i|\theta)$ are IID Bernoulli(θ) and the rv $S = \sum_{i=1}^{n_T} S_i$ has a conditional binomial dist: $(S|\theta) \sim \text{Binomial}(n_T, \theta)$

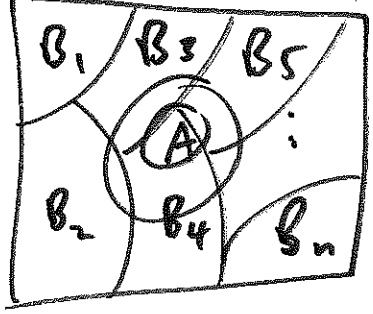
It's meaningful to talk about the conditional expectation rv. $E(S|\theta) = n_T\theta$ (a linear function of θ), and – via Bayes' Theorem – it's even more meaningful to talk about the conditional expectation <u>rv.</u> $E(\theta|S)$ (more about this later)

---

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

Important consequence of the def. of conditional expectation

Remember the law of total prob.!



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

Continuous version of LTP

X, Y continuous r.v.

for which all named densities exist →

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

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:

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

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

(Folk to interchange order of integration)

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

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

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

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

so $(*)$ is just

$$E_X\left[E(Y|X)\right]$$

and we have shown that (Adam)

$$E(Y) = E_X\left[E(Y|X)\right]$$

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.

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 $V(Y|x) \triangleq E_{Y}\{[Y - E(Y|x)]^2 | x\}$ the number

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. $= g(x)$

(formalizing Galton's intuition)

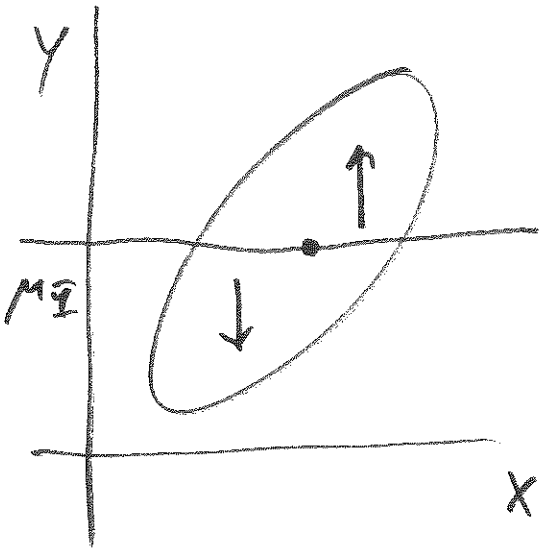
The payoff from all of this

Theorem X, Y related rv's, want to use some function $\hat{Y} = d(X)$ to predict Y from X → 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 rv. such that all of the following expressions exist, →

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

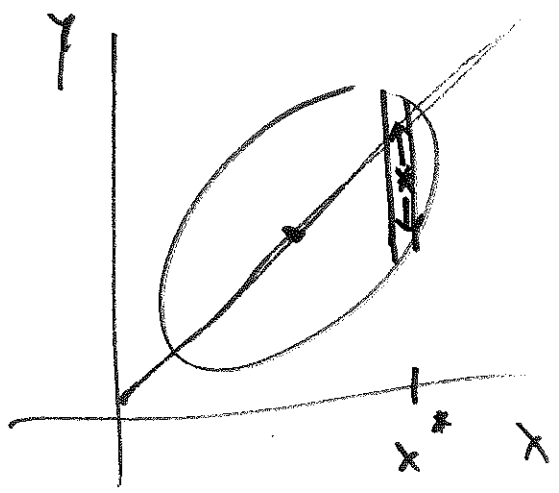


Imagine a 2-part game!

Step 1 Predict Y without knowing X. Well, if you buy into MSE as your measure of "goodness" of a prediction, we know that you should predict $\hat{Y}_{\text{no } X} = \mu_Y = E(Y)$ and your resulting MSE will be

$$E\left[(Y - \mu_Y)^2\right] = V(Y) = \sigma^2_Y.$$

Step 2 Observe X, now predict Y



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

and your resulting MSE will be

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

From the vantage point of someone thinking about stage 2 before it happens, Ξ is not yet known, so the expected value of (**), namely $E_\Xi[V(\Sigma \mid \Xi)]$, 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

$$\underbrace{V(\Sigma)}_{\substack{\uparrow \\ \text{MSE of} \\ \hat{\Sigma}_{no\,\Xi}}} = \underbrace{E_\Xi[V(\Sigma \mid \Xi)]}_{\substack{\text{"}E(\text{MSE})\text{" of} \\ \hat{\Sigma}_\Xi = E(\Sigma \mid \Xi)}} + \underbrace{V_\Xi[E(\Sigma \mid \Xi)]}$$

But since variances are always non-negative,

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

$$\underbrace{E_X[V(Y|X)] + V_X[E(Y|X)]}_{V(Y)} \geq \underbrace{E_X[V(Y|X)]}_{\text{"E(MSE)" of } \hat{Y}_X}$$

$$V(Y) \geq \text{MSE of } \hat{Y}_{\text{no } X}$$

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.

Another complete switch in subject

Utility

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

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

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

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 risk-seeking person would probably pick (A).

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

DS 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 $u(x) = x$? (1: utility = 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.

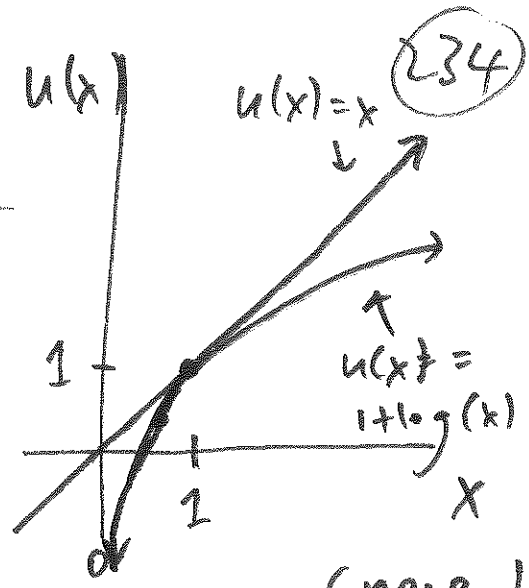
Daniel B: If your entire net worth is (say) \$10, then the value to you of a new \$1 is much greater than if your entire net worth is (say) \$1,000,000; thus the utility of money is sublinear (meaning that it doesn't grow with x as fast as $f(x) = x$ does)

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 $E[u(X)] = E[u(Y)]$.

EU 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, EU is the best you can do.

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

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

Oskar Morgenstern
(1902-1977)
German-economist
American

Match	x	$P(X=x)$	$x \cdot P(X=x)$
5w, 1R	\$47,000,000	$\frac{1}{292,201,338}$	\$1.667
5w, 0R	\$1,000,000	$\frac{1}{11,688,053.52}$	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 $E[u(\mathbb{X})]$; if for you $u(x) = x$ (utility $\triangleq$ money) then $E(u(\mathbb{X})) = \$1.99$

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

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

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

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. (238)

X = your winnings

X uncertain before the drawing

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

new 1st row in table is $\frac{1{,}600{,}000{,}000}{292{,}201{,}338} \doteq 5.476$

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

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

(*) A catalog of useful distributions

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

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

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

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

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

Case Study: Cortaneda 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 the 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	variance
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)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{T-n}{T-1}\right)$

$0 \leq \alpha = \frac{T-n}{T-1} \leq 1$ is called the finite population correction

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, $\uparrow\uparrow$) $\alpha \uparrow 1$ $\leftrightarrow$ with a small sample from a large population, SRS = IID

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

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

support of X

$E(X) = \lambda$

$V(X) = \lambda$

} thus for the Poisson dist.

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

Def. If $E(X)$ and $V(X)$ both exist and $E(X) \neq 0$, $\frac{V(X)}{E(X)}$ 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), Binomial(n, p) ≐ Poisson(n·p)

Theorem n positive integer, $0 < p < 1$, $X \sim \text{Binomial}(n, p)$

$\lambda > 0$, $X \sim \text{Poisson}(\lambda)$ / 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 $f_{\underline{X}}(x \mid n, p_n) \xrightarrow[n\to\infty]{}$

Poisson process, revisited

Def

$f_{\underline{Y}}(y \mid \cdot\lambda)$

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 $t \sim \text{Poisson}(\lambda t)$

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

~~Case Study~~

~~Parasitic~~ Protozoa in drinking water

Here'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)

Let 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 $y = 0, 1, \ldots$

and $P(X=x \mid Y=y) = \binom{y}{x} p^x (1-p)^{y-x}$

Notice that if $X=x$, $Y \geq x$ because the actual number of oocysts (Y) has to be at least as large as the number of oocysts detected (X). After a careful calculation you get,

$$f_X(x) = \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{(\lambda t)^y e^{-\lambda t}}{y!}$$

$$= \frac{e^{-p\lambda t} (p\lambda t)^x}{x!}$$

i.e., $X \sim \text{Poisson}(p\lambda t)$: losing a proportion $(1-p)$ of the oocysts to faulty counting just lowers the rate of the Poisson process from λ/liter to $\lambda \cdot p$/liter (making excellent sense).

In practice oocysts are hard to detect: p is small (not far from 0). **Q:** How much water (t liters) do you need to filter to achieve $P(\text{at least 1 oocyte detected}) \geq 1-\alpha$ for small α? **A:** Not hard to work out

$$P(\text{at least 1 detected}) = 1 - P(\text{none detected})$$

$$= 1 - P(X=0) = 1 - e^{-p\lambda t} \geq 1-\alpha$$

$$\leftrightarrow \alpha \geq e^{-p\lambda t} \leftrightarrow \ln\alpha \geq -p\lambda t \leftrightarrow$$

$$\boxed{t \geq -\frac{\ln\alpha}{p\lambda}}$$

Example $\alpha = .01$, $p = 0.1$, $\lambda = 0.2$ / liter (1 per 5 liters ↕ minimum sickness level)

to achieve prob 99%, t has to be at least 230.3 liters.

Negative Binomial Distribution

You're watching a potentially endless sequence of Bernoulli trials with constant success probability p. Let X = # failures before r^{th} success, r integer ≥ 1.

You can show that X follows what's called the Negative Binomial dist:

its pmf is $f(x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x$.

with parameters (r, p) & $(0 < p < 1)$

The name comes from the fact that, when you watch a sequence of Bernoulli trials with unknown constant success probability p unfold, there are two different ways to

estimate p: decide ahead of time to sample n (known constant) successes/failure trials, and record the (random) # S of successes you see (from which a reasonable estimate would be $\hat{p}_B = \frac{S}{n}$ ← Binomial).

(or) decide ahead of time that you're going to sample until you've seen s (known constant) successes & record the (random) # of trials N needed to accumulate that many successes (from which a reasonable estimate would be $\hat{p}_{NB} = \frac{s}{N}$ ← Negative Binomial).

Special Case of Negative Binomial

Set $r=1$ and record the number X of failures until the first success: X is said to follow the Geometric(p) distribution, with

$$f_X(x \mid p) = p(1-p)^x \, \mathbb{1}_{\{0,1,\dots\}}(x)$$

(parameter p) (support of X)

Sequence $X_1, \dots, X_n$ IID Geometric(p)

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

This is a direct analogue to the Bernoulli/Binomial story: $X_1, \dots, X_n$ IID Bernoulli(p) $\Rightarrow \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$.

$X \sim$ Negative Binomial (r, p)

$$\psi_X(t) = \left[\frac{p}{1-(1-p)e^t}\right]^r \quad \text{for } t < \log\left(\frac{1}{1-p}\right)$$

from which $E(X) = \frac{r(1-p)}{p}$, $V(X) = \frac{r(1-p)}{p^2}$

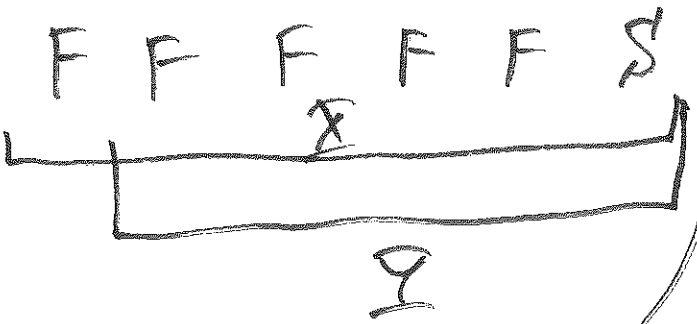
Consequence $X \sim$ Geometric $(p) \rightarrow$

$\left\{\begin{matrix} k \\ t \end{matrix}\right\}$ both non-negative integers

$$P(X = k+t \mid X \geq k) = P(X = t)$$

this is called the memoryless property of the Geometric distribution, and it turns out that this is the only

discrete distribution with this property.



X = # failures until first success = 5 (here)

Y = # failures, starting at trial $(k+1 = 2$ here$)$ until next success $(= 4$ here$)$

Then Y has the same dist. as X and is independent of what happened on the first k trials, i.e., "the process has no memory".

Case 2: Important Continuous Distributions

Normal (Gaussian) Distribution

$X \sim \text{Normal}(\mu, \sigma^2)$ mean μ, variance $0 < \sigma^2 < \infty$

PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

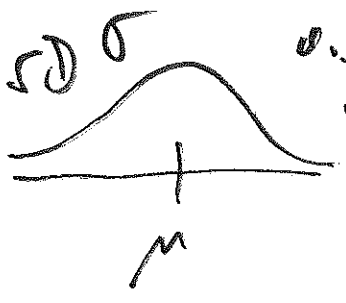
The Normal dist. is the single most important dist. in all of probability & statistics, mainly for 2 reasons: (1) many observable random processes have dist. shapes that are close to the "bell curve" (Normal PDF), and (2) the Central Limit Theorem (CLT), which we'll examine soon.

Properties of the Normal Dist.

$X \sim \text{Normal}(\mu, \sigma^2)$ $\quad E(X) = \mu$

$V(X) = \sigma^2$, $\quad SD(X) = \sigma$

$$\psi_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$



$\left(\begin{array}{c}\text{center of symmetry} \\ \text{mean} \\ \text{median} \\ \text{mode}\end{array}\right)$ all $= \mu$

Consequences (1) $X \sim \text{Normal}(\mu, \sigma^2)$,

$Y = aX + b$, ($a \neq 0$, b fixed constants) $\rightarrow$

$$Y \sim \text{Normal}(a\mu + b, a^2\sigma^2).$$

In other words, Normality is preserved under linear transformation | Def.

The Normal dist. with mean $\mu = 0$ and SD $\sigma = 1$ is called the standard Normal dist.

The PDF of $Z \sim \text{Normal}(0, 1)$ is

$$\phi_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$ and its

(phi (lower-case))

CDF is $\Phi(x) = \int_{-\infty}^{x} \phi_Z(t)\, dt$

(upper-case Phi: Φ)