Some more probability

binomial processes: based on combinations

combination: the order doesn't matter (if it does, permutation)

↳ how many ways to pick 3 pairs of pants out of 7?

$\binom{7}{3}$ "7 choose 3" seems easier not to care about ordering but mathematically not quite

↳ to choose + order k objects:

choose k, order them $\longrightarrow \binom{n}{k} \cdot k! = \dfrac{n!}{(n-k)!}$

$= n \cdot \ldots \cdot (n-k+1)$

choose k ordered.

Binomial processes

divide a region $[0,1]$ into intervals $1/n$
and look at a binary $0/1$ thing.
$\hookrightarrow$ prob. λ/n for "on" and $1-\lambda/n$ for "off"

e.g. GW in some timeframe
star in a patch of sky

$\text{Bin}(n, \lambda/n)$ $\leftarrow$ distribution of things in n slots with prob. λ/n of being in a given slot

$$\mathbb{P}[\text{Bin}(n, \lambda/n) = k] = \binom{n}{k}\left(1-\frac{n}{k}\right)^{n-k}\left(\frac{\lambda}{n}\right)^{k}$$

choose any k to be "on" so $n-k$ are "off"

$$= \frac{\lambda^k}{k!} \cdot \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \left(1-\frac{\lambda}{n}\right)^{n-k}$$

looking messy... but let's do a limit

$\lim n\to\infty$

$$= \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda}$$

math fact!
$\lim \left(1+\frac{x}{n}\right)^n = e^x$

means tinier intervals!

$$= \mathbb{P}[\text{Poisson}(\lambda) = k]$$

woah!

let T be the first of n intervals that is "on".

how is T distributed? → good Q from Monday!

prob that T occurs at some $0 \le x \le 1$ → not interval #! we do this bc we're interested in the continuous limit.

↳ prob. everything is "OFF" until x.

$$P[T = x] = \left(1 - \frac{\lambda}{n}\right)^{nx} \cdot \frac{\lambda}{n} \longrightarrow \approx \underbrace{\lambda e^{-\lambda x}}_{\text{density of } \mathrm{Exp}(\lambda)}$$

x is a time

nx is # intervals → technically integer just below nx, so some rounding error here

on at x

some room for error here bc nx rounding etc but!

in the limit, density of T is $\mathrm{Exp}(\lambda)$

joint distribs

$F(x,y) = P\{(x'<x) \cup (y'<y)\}$ → $F(\infty,\infty)=1$

$F(x,-\infty) = F(-\infty,y) = 0$

PDF: $p(x,y) = \frac{\partial^2 F}{\partial x \partial y}$ with $\iint_{-\infty}^{\infty} p(x,y)\,dx\,dy = 1$

moments!

$\mu_x = E(x)$ $\quad \sigma_x^2 = E[(x-\mu_x)^2]$

$\mu_y = E(y)$ $\quad \sigma_y^2 = E[(y-\mu_y)^2]$

mixed moment

$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$

a note on calculating these: $E(g(x,y)) = \iint g(x,y)\cdot p(x,y)\,dx\,dy$

mixed moment is <u>covariance</u>.

↳ $E[xy - \mu_x y - \mu_y x + \mu_x\mu_y] = E[xy] - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y$

$= E[xy] - E[x]\,E[y]$

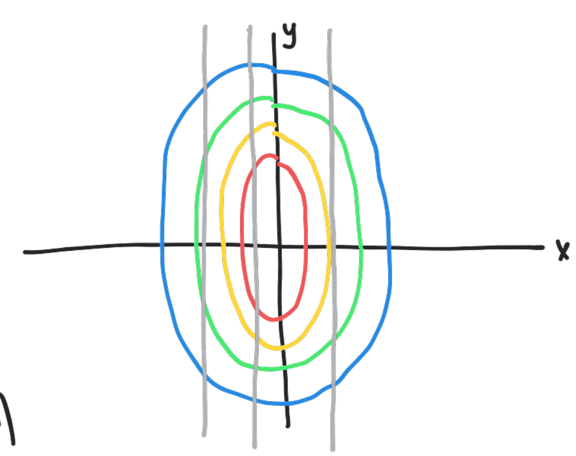
handy to use a ~matrix~

$M = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$ $\quad M_{ij} = \langle (x_i-\bar{x}_i)(x_j-\bar{x}_j)\rangle = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle$

on diag, just variance.

⇒ how far is x from its mean? how far is y from its?

how do these move together (hence product).



← contour plot of data

pick some x-vals.

look at 1D dist.

where maximized?

y=0 everywhere!

uncorrelated.

not statistical independence!

has ~0 to do w/ whether

$p(x,y) = p(x)\,p(y)$.

↳ 0 cov is subtle bc could be nonlinear relationship.

correlation

if covariance $\neq 0$, mean of y depends on x (vice versa)

↳ e.g. human height / weight

but cov isn't bounded — can be arbitrarily large/small.

↳ how to gauge degree of correlation?

introduce correlation coeff: $\rho_{xy} = \dfrac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$

not to get too mathy but:

Cauchy-Schwarz ineq: $|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$

$$\rightsquigarrow (\mathrm{cov}(x,y))^2 \leq \sigma_x^2 \sigma_y^2$$

$$|\rho_{xy}| \leq 1$$

measures how linear the data is basically

interpretation varies by field tbh. 0.8 could be good or garbage ┐(・_・)┌

chi-squared

how do we look for nonlinear relations?

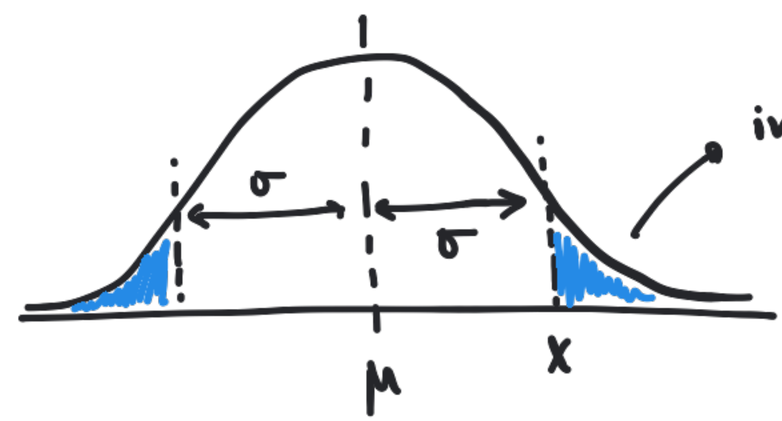
take a measurement and get $x \pm \sigma$.

model from theory predicted μ.

↳ prob of getting something <u>further</u> from pred than the obs value?

or, what's prob that obs is consistent w/ μ.

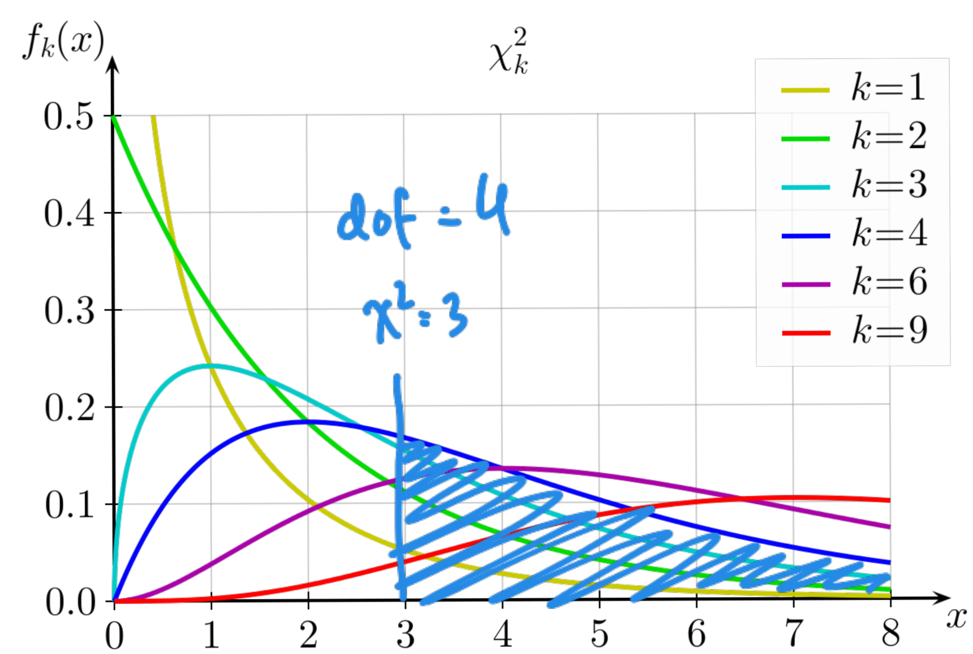
assume normal dist around μ.



define $\chi^2 = \dfrac{(x-\mu)^2}{\sigma^2}$

↳ normalisation of dist. from μ.

want $P(\chi^2 > \chi^2_{obs})$. this has distribution.



can extend this to higher dim! $k = \text{dof}$

so that was 1 measurement. in general:

$$y_i = f(x_i) + \epsilon$$

y_i = measurement

$f(x_i)$ = model pred at x_i

ϵ = gaussian error with σ_i

Standard error, comes from multiple measurements.

↳ expectation for data.

joint PDF for all the y_i? just one big gaussian!

$$P(y_1, y_2, \ldots, y_k) = \frac{1}{(2\pi)^{k/2} \prod_i \sigma_i} \exp\left(-\frac{1}{2} \sum_i \frac{(y_i - f(x_i))^2}{\sigma_i^2}\right)$$

so this is prob of data y_i given model $f(x)$.

↳ this is called LIKELIHOOD.

NB. $p(\text{obs} \mid \text{model}) \neq p(\text{model} \mid \text{obs})$ ← related tho!

we'll assume that if model maximizes prob of observing these, it's close to the real thing. "MAXIMUM LIKELIHOOD ESTIMATION" in general, model is a function of parameters $\alpha_1, \ldots, \alpha_m$.

so really we want

$$\alpha_j^{best} = \text{argmax}_{\alpha_j} \left[\frac{1}{(2\pi)^{k/2}} \cdot \frac{1}{\prod_i \sigma_i} \exp\left(-\frac{1}{2} \sum_i \frac{(y_i - f(x_i; \alpha_1, \alpha_2 \ldots))^2}{\sigma_i^2} \right) \right]$$

argmax of fct = argmax log fct bc log is monotonically ↗

So some algebra later...

$$\alpha_j^{best} = \text{argmin} \left[\sum_i \frac{(y_i - f(x_i; \alpha_1, \alpha_2 \ldots))^2}{\sigma_i^2} \right]$$

becomes min bc of $-\frac{1}{2}$ in front.

max likelihood soln comes from minimizing this sum.

sum is the χ^2!

$$\chi^2(\alpha_1, \alpha_2, \ldots, \alpha_m) = \sum_i \frac{(y_i - f(x_i; \alpha_1, \alpha_2, \ldots))^2}{\sigma_i^2}$$

Phil will discuss later.

"weighted least squares"

↳ bc of σ_i^2 term.

minimization can happen numerically! often gradient descent let's look at χ^2 on histogram.

genfromtxt returns list of tuples.

assuming model dist,
$p(\text{data}) = ?$

scipy.stats chisq

↳ uses Poisson error from pred, not yerr.

in principle, SHOULD use theoretical error like this.

↳ not too bad to do when it's Poisson! cts / bin = variance.

but if we're doing some complicated curve-fitting and our model is some $f(x; \alpha_1, \alpha_2, \ldots)$ ⟶ not as easy.

↳ fairly strong argument for using data σ then.

here, we let scipy use $\sqrt{N}$.

P is tiny? 0.5463% ⟶ v. far from pred. dodgy!

P is middly? 37.16% ⟶ looks OK! want to be in middle of χ^2 distrib.

P is huge? 99.04% ⟶ too close ... something weird in errors? like rolling dice 600 times, getting 100 of each exactly. OK to be a lil sus.