THE ABNORMAL NORMAL DISTRIBUTION.

1. Gaussians are ubiquitous
2. Information Entropy
3. Gauss' derivation
4. Error cancellation
5. Ubiquity explained
6. Properties - Ganssians are special
7. Note on naming
8. History?
9. Herschel derivation

The extreme abnormality of the normal distribution

(1)

Gaussians are <u>everywhere</u>. At the heart of modern physics:

Path integral: $\int e^{\frac{i}{\hbar}(KE - PE)} d\phi$, with $PE = \frac{1}{2} m \phi^2$

Feynman diagrams are: Taylor expansion of the PE term...
ie can do Gaussian integral, interaction terms are tricky...

Gaussian distribution is also very common model - stock price, swap rate... but most common application in history of science is as an "error law" (erf(x) !).
└ assumed frequency distribution of errors.

But, <u>are</u> errors Gaussian? "Willy" Feller ~1955 after dinner speech anecdote (Jaynes pg 198)... Even 1983... (George Barnard)
└ quality control guy.

(2) Groundwork: Gaussian form has higher "information entropy" than any other for a given variance. Information entropy is really "amount of uncertainty". Want to assign a distribution that reflects our state of knowledge as fully as possible, which is ignorance, other than, say,

rms ~ average size of the errors.

how "spread out" a dist'n is... *

Amount of uncertainty is related to # ways of getting a given distribution - we should assign one that is realizable in a number of ways so much greater than that of nearest neighbour that it is overwhelmingly the most likely.

* want to be as "impartial" as possible, as true to the principle of ~~ignorance~~ indifference as possible...

namer → Keynes 1921

· Amount of uncertainty in distribution $p(x)$ is:

$$H_p = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx. \qquad \text{(Shannon 1948)}$$

If know nothing other than range (or finite set of possibilities in discrete case), variational calculus gives us same dist'n as intuition: $p(x) = \frac{1}{b-a}$.

$$L_0[p(x)] = -\int_a^b p(x) \log p(x)\, dx + \lambda \left(\int_a^b p(x)\, dx - 1 \right)$$

↑ Lagrange multiplier

$$\delta L_0[p(x)] = -\int_a^b \left(\log p(x) + \frac{p(x)}{p(x)} - \lambda \right) \delta(p(x))\, dx = 0$$

$$\log p(x) = \lambda - 1$$

$$p(x) = e^{\lambda - 1}$$

$$\int_a^b e^{\lambda - 1}\, dx = 1$$

$$\therefore e^{\lambda - 1} = \frac{1}{b-a} \quad \checkmark\checkmark$$

Add more constraints, still have the $\log p(x)$ term which leads to a family of exponential distributions including just about everything. Poisson, Γ, χ^2, binomial etc etc ....

So, Gaussian:

$$L_2[p(x)] = \int_{-\infty}^{\infty} \Big[-p(x) \log p(x) + \alpha \big(p(x) - 1 \big) + \beta \big(x p(x) - \mu \big) + \gamma \big(x^2 p(x) - \nu \big) \Big] \times dx$$

(actually $\delta(x)$)

$$\delta L_2[p(x)] = 0 \Rightarrow -\log p(x) - 1 + \alpha + \beta x + \gamma x^2 = 0$$

$$\therefore p(x) = e^{\gamma x^2 + \beta x + \alpha - 1}$$

Have "derived" Gaussian form as dist'n best representing knowledge of variance. Exhibit A. Now for Gauss' derivation:

(3) Estimate θ from $n+1$ observations $(x_0 \dots x_n)$ by max Likelihood,

$$p(\underline{x}|\theta) = \prod_{i=0}^{n} f(x_i|\theta)$$

θ and x have same scale

Write as $\frac{\partial}{\partial\theta} \log p(\underline{x}|\theta) = \sum_{i=0}^{n} \frac{\partial \log f(x_i|\theta)}{\partial\theta} = 0$.

"location" parameter

then write $\log f(x|\theta) = g(\theta - x) = g(u)$ $\quad u = \theta - x$

so MLE condition becomes $\sum_i \frac{dg}{d\theta}(\hat{\theta} - x_i) = 0$ for MLE $\hat{\theta}$.

This is all very fancy, depends on unknown dist'n f (ie g)...

What if $\hat{\theta} = \frac{1}{(n+1)} \sum_{i=0}^{n} x_i$? Turns out requiring arithmetic mean is MLE uniquely determines $f(x|\theta)$ as Gaussian.

Haven't constrained to a particular set of values $\underline{x}$ - can try it out with an extreme, if unlikely, realization, to probe things:

$$x_0 = (n+1)u, \quad x_1 \dots x_n = 0$$

Then $\hat{\theta} = u$ and $\hat{\theta} - x_0 = -nu$ and MLE condition is:

$\hat{\theta} - x_i = u \quad i > 0.$ $\quad \frac{d}{d\theta} g(-nu) + n\frac{dg}{d\theta}(u) = 0$

$n=1$ case: $\frac{dg}{d\theta}(-u) = -\frac{dg}{d\theta}(u)$, ie $\frac{dg}{d\theta}$ is antisymmetric,

so have $\frac{dg}{d\theta}(nu) = n\frac{dg}{d\theta}(u)$ so $\frac{dg}{d\theta}$ is <u>linear</u>.

and $\frac{d}{d\theta} = \frac{d}{du}$, so $\frac{dg(u)}{du} = au + b \rightarrow g(u) = \frac{1}{2}au^2 + bu$

redefine!

$$\log f(x|\theta) = g(u) = g(\theta - x) = \frac{1}{2}au^2 + b$$ (scale param)

$$\text{so } \underline{f(x|\theta) = e^{\frac{1}{2}a(x-\theta)^2 + b}}$$ solve for $b(a)$ via normalization, which will diverge unless $a < 0$.

But assumed a special sample. Necessary form for MLE $\longleftrightarrow$ arithmetic mean.

Suppose instead it is "the general solution"? No, assume the distribution in MLE condition and solve for $\hat{\theta}$:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\theta)^2}{2\sigma^2}}$$

$$\sum_i \left.\frac{\partial g(\theta - x_i)}{\partial \theta}\right|_{\hat{\theta}} = \sum_i \frac{x_i - \hat{\theta}}{2\sigma^2} = 0 = \frac{1}{2\sigma^2}\left(\sum_i x_i - \hat{\theta}(n+1)\right)$$

$$\therefore \hat{\theta} = \frac{1}{n+1}\sum_{i=0}^{n} x_i$$

(remember Daniel Bernoulli's arrows...)

ie Gaussian form is sufficient for MLE $\longleftrightarrow$ arithmetic mean, as well.

(4) Nothing 100% fundamental about arithmetic mean as an estimator though. There are lots of possible estimators, varying by distribution. Can consider: $\beta(\underline{y}) = \sum_{i=1}^{n} w_i(\underline{y})\, y_i$ [easier now if $n+1 = n'$ $x_0 \dots x_n = y_1 \dots y_{n'}$ $n' \to n$ ⌟

and assess quality of estimate given $f(y|\theta)$, Gaussian

I.I.D. — $p(\underline{y}|\theta) = f(y_1) f(y_2) \dots f(y_n)$

How assess quality? ↘

$$\Psi = \left\langle \left(\beta(\underline{y}) - \theta\right)^2 \right\rangle \quad \text{where} \quad \langle h(\underline{y}) \rangle = \int_{-\infty}^{\infty} h(\underline{y})\, p(\underline{y}|\theta)\, d\underline{y}$$

Define i'th error: $\varepsilon_i = y_i - \theta$ $\longrightarrow$ $\left\langle \left(\sum_{i=1}^{n} w_i(\underline{y})\, e_i\right)^2 \right\rangle$

$$\therefore \Psi = \sum_{i,j=1}^{n} w_i w_j \underbrace{\langle e_i e_j \rangle}$$

$\rightarrow \sigma^2 \delta_{ij}$ IID Gaussian,

drop $\underline{y}$-dependence of weights for now.

$$= \sigma^2 \sum_i w_i^2$$

Variational problem:

$$L(\underline{w}) = \sigma^2 \sum_{i=1}^{n} \left(w_i^2 + \lambda \left(w_i - \frac{1}{n} \right) \right) \rightarrow 2w_i + \lambda = 0$$

$$w_i = -\frac{\lambda}{2} \text{ (constant)}$$

$$\sum_i w_i = -\frac{\lambda}{2} n = 1 \text{ so } w_i = \frac{1}{n}$$

$$\therefore \boxed{\beta(\underline{y}) = \frac{1}{n} \sum_i y_i}$$

(5) Take stock:

(i) Gaussian is maximum entropy distribution under mean and variance constraints. Must spread out a distribution can be under the constraints. Assigns maximum possible probability to each value, indifferently, while conserving variance. Must honestly represents our state of knowledge. Ignores distribution of errors beyond first 2 moments...

(ii) Requiring the arithmetic mean to maximize likelihood uniquely determines a Gaussian form.

(iii) Estimate of parameter via arithmetic mean is more accurate (minimizes expected square error Ψ) for a Gaussian than for any other distribution. more opportunity for error cancellation

So a Gaussian is the distribution we assign to errors if we want the most accurate estimates from arithmetic means, given the general magnitude of the noise.

(6) Gaussian form really is odd:

A: Product of 2 Gaussians is Gaussian

B: Fourier Transform of G. is G.

C: Convolution of 2 G's is G

D: G. is limiting form of various others: binomial, Poisson, beta...

E: Any hump raised to higher and higher powers → G.

F: Any hump under repeated convolution → G (central limit theorem)

CLT is usually given as justification for Gaussian noise assumption, but its really a special case of something deeper, as as E, D. Gaussian is max-ent dist'n, so any transformation which discards information (detail, certainty, preference of some possibilities over others) while preserving variance, leads closer to Gaussian form.

(7) Can't say this about any distribution except Gaussian. It really is weird! Not very "normal". Origin of term is unclear but probably geometrical in origin, in context of regression - Gauss described something as perpendicular to surfaces of const. probability or similar.

Jaynes suggests "central".

(8) History?!

History

1741 Laplace born
1749 Euler tries to resolve Great Inequality of Jupiter + Saturn ✱.
- 8 orbital parameters. ✱

82 — 75 obs. of longitudes — 1745 — 1783 ▲ — 1800 — 1860 □ — 1900 — 1941 — 2000

Q.F.T.! 7.

1600 — 1700 — 1676 Halley noted problem — 1733 de Moivre — 1777 ⊛ ✱ Gauss derivation / 1809 passing remark — 1850 Herschel derivation — Landon derivation

Tycho Brahe averaging obs as routine.

made no use

in a work on astronomy

1787 ✱ Problem of Jupiter and Saturn solved by Laplace

Jupiters drift would reverse...

Laplace 1810, 1812 liked it, demonstrated useful properties, central limit theorem, linear regression

→ "Gaussian" ie called so

Euler, Lagrange tried to solve the problem

✱ combining the ~~errors~~ observations ⟶ "multiply the errors".

Which 8 equations to pick? Got 2, known anyway, French Ac. of Sciences prize 1749

yet: ⊕

⊛ Daniel Bernoulli says: obviously probability of distance of arrow from vertical line target is not uniform, so nonsense to weight uniformly when estimating target location from arrow positions.

▲ Laplace shows binomial → Gaussian
derived main properties
suggested tabulation

but failed to see Gaussian was the natural error law
- still used $p(x) \propto e^{-a|x|}$.

but got the combination of observations idea so ✱

□ Maxwell 3rd version of same ~~thing~~ argument (Maxwellian) → distribution of velocities of gas molecules

Herschel - Maxwell

P1: No preferred direction: $p(x,y)\,dx\,dy = f(x)\,dx \cdot f(y)\,dy$

P2: No preferred angle: $g(r,\theta) = g(r)$ in $p(x,y)\,dx\,dy = g(r,\theta)\,r\,dr\,d\theta$

so $f(x)\,f(y) = g\left(\sqrt{x^2+y^2}\right)$

Must still be true if $y=0 \rightarrow g(x) = f(x)\,f(0)$

$$f(x)\,f(y) = f\left(\sqrt{x^2+y^2}\right)\cdot f(0)$$

or $$\frac{f(x)}{f(0)}\,\frac{f(y)}{f(0)} = \frac{f\left(\sqrt{x^2+y^2}\right)}{f(0)}$$

$$\frac{f(x)}{f(0)} = e^{ax^2}$$ does the trick.

1850

Maxwell 1860 - 3-d version of above arg

$\rightarrow$ Maxwellian distribution of gas molecules.