

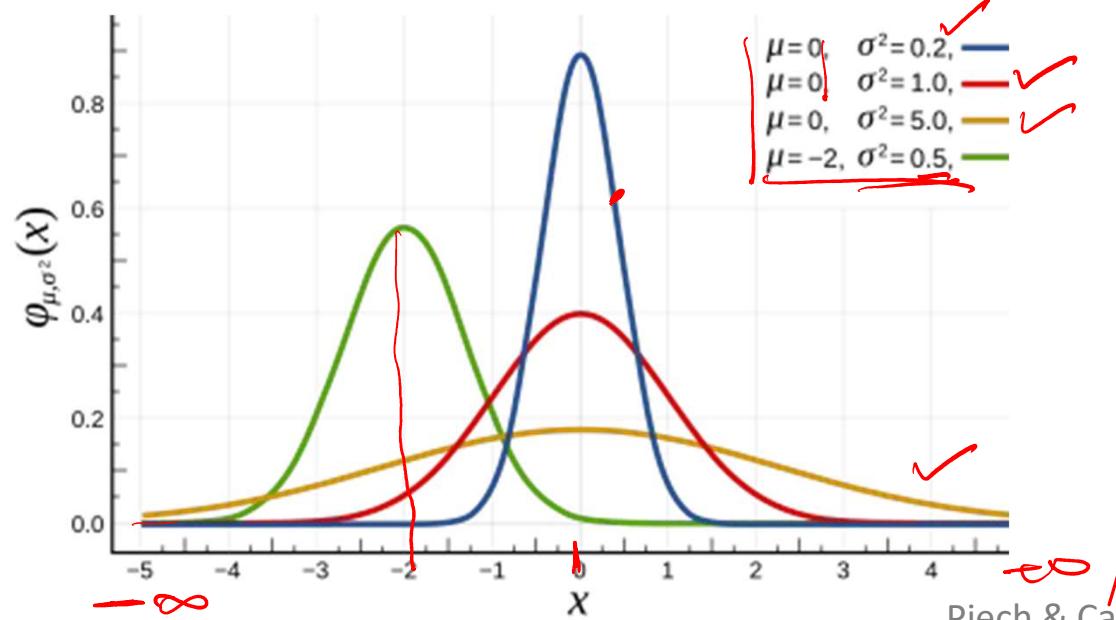
Normal (Gaussian) Random Variable

Support:
 $(-\infty, \infty)$

$$\underline{X} \sim \mathcal{N}(\underline{\mu}, \underline{\sigma}^2)$$

mean
variance

density



Normal (Gaussian) Random Variable

Support:
 $(-\infty, \infty)$

$$X \sim \mathcal{N}(\underline{\mu}, \sigma^2)$$

mean
variance

PDF:

$$f(X = \underline{x}) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Anatomy of a The Normal PDF

distance to the mean
(makes the PDF symmetric
around the mean)

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

a constant:
makes the integral
over all possible
outcomes sum to 1

...normalized by
the variance

Expected value of a normal distribution

Verify that μ is the expected value of
 $x \sim N(\mu, \sigma^2)$

$$\underline{E[(x-\mu)]} = E(x) - \mu$$

$$\int_{-\infty}^{\infty} \underline{(x-\mu)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \Big|_{-\infty}^{+\infty} = 0$$

$$E[(x-\mu)] = 0 \Rightarrow E(x) = \mu$$

Variance

$$\cancel{E((X - \mu)^2)} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/(2\sigma^2)} dx \quad \checkmark$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-(y)^2/(2)} dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (y)(ye^{-(y)^2/(2)}) dy$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left[\left(-ye^{-y^2/2} \right) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -e^{-y^2/2} dy \right] \quad \int udv = uv - \int vdu$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \underline{\sigma^2}$$

Properties

If $X \sim N(\mu, \sigma^2)$ and if $Y = aX + b$, then a & b are scalars -

Top ~~process~~

$Y \sim N(a\mu + b, a^2\sigma^2)$

Let F_Y be the cumulative density of Y

$$F_Y = P(Y \leq y) \quad f_Y = \frac{d}{dy} F_Y(y)$$

$$F_X = P(X \leq x) \quad f_X = \frac{d}{dx} F_X(x)$$

Let $a > 0 \Rightarrow P(ax + b \leq y) = P(X \leq \frac{y-b}{a})$

$$P(Y \leq y) = P(X \leq \frac{y-b}{a}) \Rightarrow \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right)$$

$$f_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) = \frac{d}{dx} F_X(x) \frac{d}{dy} \left[\frac{y-b}{a}\right]$$

$$= f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}$$

$$f_X\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\frac{y-b}{a}-\mu)^2}{2\sigma^2}}$$

$$f_x\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\mu a+b))^2}{2\sigma^2 a^2}}$$

$$f_y(y) = f_x\left(\frac{y-b}{a}\right) \cdot \frac{1}{a} = \frac{1}{\sqrt{2\pi}\sigma\sqrt{a^2}} e^{-\frac{(y-(\mu a+b))^2}{2a^2\sigma^2}}$$

$$\Rightarrow Y \sim N(\mu a + b; \frac{\sigma^2}{a^2}) \text{ if } a > 0$$

$$\begin{aligned} a < 0 \\ F_Y(y) &= P(Y \leq y) = P(ax + b \leq y) = P(X \geq \frac{y-b}{a}) \\ &= 1 - F_X\left(\frac{y-b}{a}\right) \end{aligned}$$

$$Y \sim N(\mu a + b; \sigma^2 a^2)$$

Question in class: Why do we substitute in CDF instead of directly in the density function?

- Density measures the rate of change of probability with value x of a random variable *but* under the constraint that the integration of density is 1.
- When we directly apply rule of calculus on the density, we are not ensuring that the integration is 1. See this example:

Suppose $X = U(0,1)$, $Y = 2X$.

Probability density function $f(X) = 1$.

If we directly substitute in above pdf, then $f(Y) = 1$. But Y takes values between 0 and 2, and integration of $f(Y)$ from 0 to 2 will be 2, not 1. We need to adjust $f(Y)$ to account for the stretch.

The correct way to substitute variables directly in the density function is much more involved. You can take a look here if you want to know about that.

https://en.wikipedia.org/wiki/Probability_density_function#Function_of_random_variables_and_change_of_variables_in_the_probability_density_function

Properties

- Median = mean (why?)
- Because of symmetry of the pdf about the mean
- Mode = mean – can be checked by setting the first derivative of the pdf to 0 and solving, and checking the sign of the second derivative.

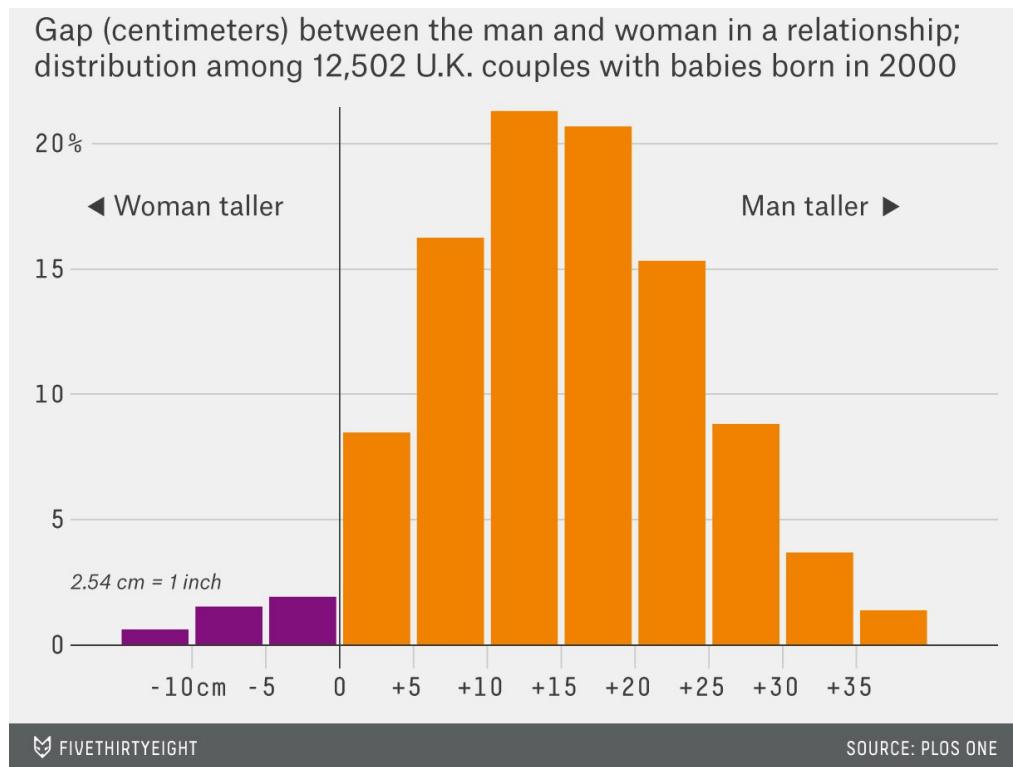
Carl Friedrich Gauss (1777-1855)

- German mathematician
- Sort-of invented the normal distribution
- Also astronomer, geologist, physicist
- Super influential in a lot of fields



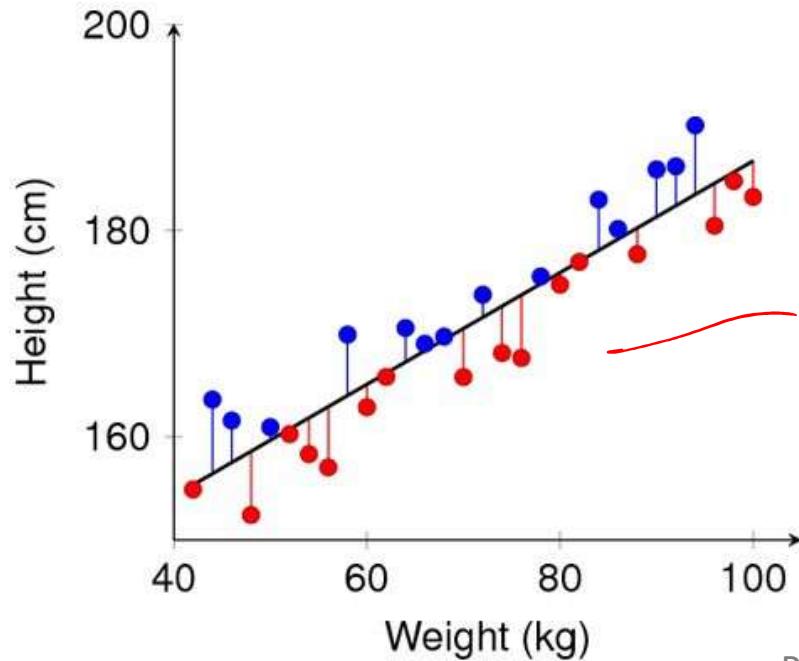
Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.



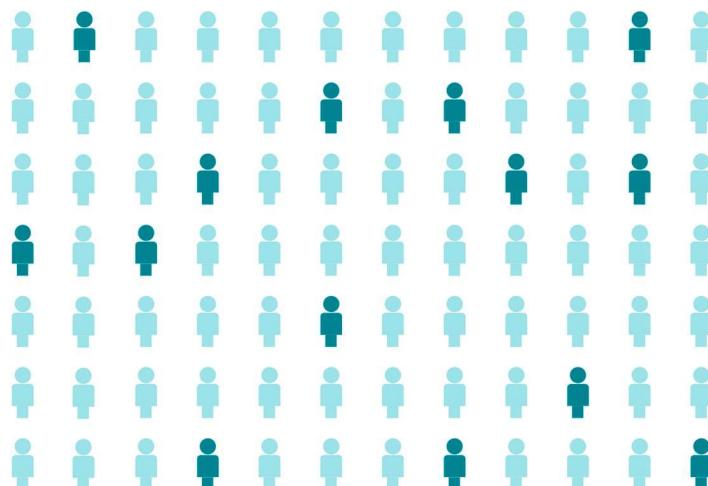
Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression



Why the Normal?

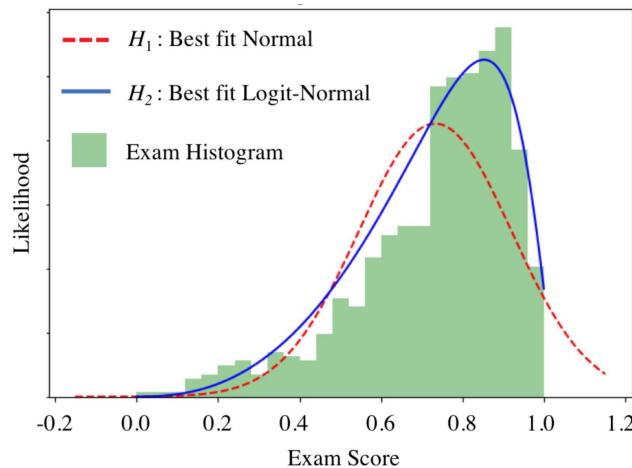
- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics



Piech & Cain, CS109, Stanford University

Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics
- Even things that aren't Normal might fit a normal-related distribution



Why the Normal?

- Common for natural phenomena: human height, weight, shoe sizes, etc.
- A lot of noise in the world is Normal
 - E.g. random errors in measurements, residuals in linear regression
- The sum of many random variables often looks Normal (spoilers)
- Sample means are distributed normally – important for statistics
- Even things that aren't Normal might fit a normal-related distribution

People also just assume things are normally distributed a lot.

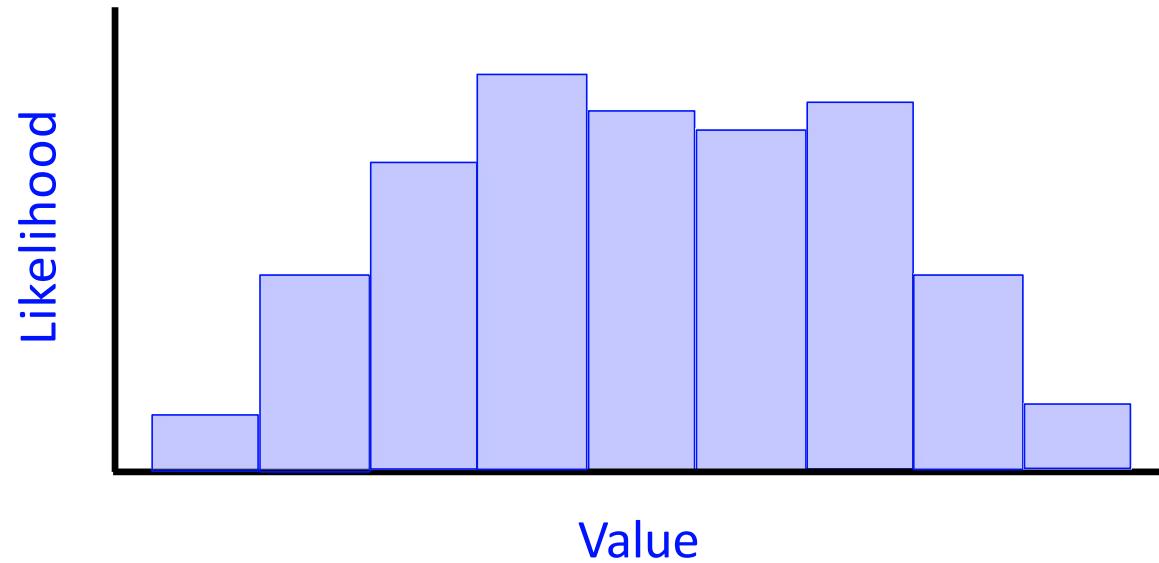
- They can do this in part because the Normal is so common
- But there's a deeper reason to it...

Ockham's razor

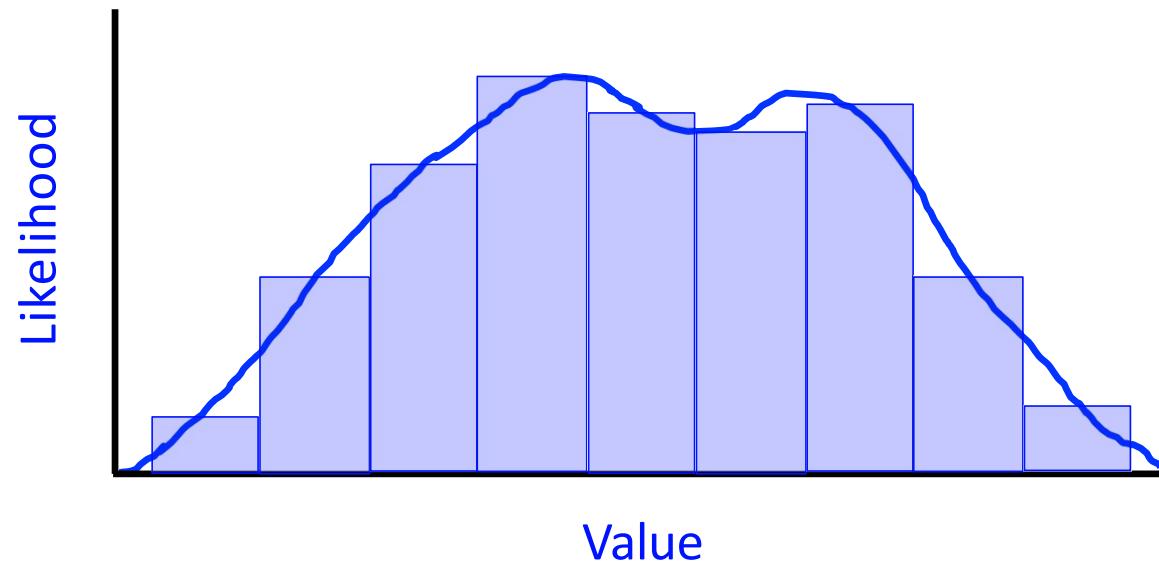
Shaving your hypothesis since 14th Century



When We Fit Models To Data, We Try To Keep It Simple

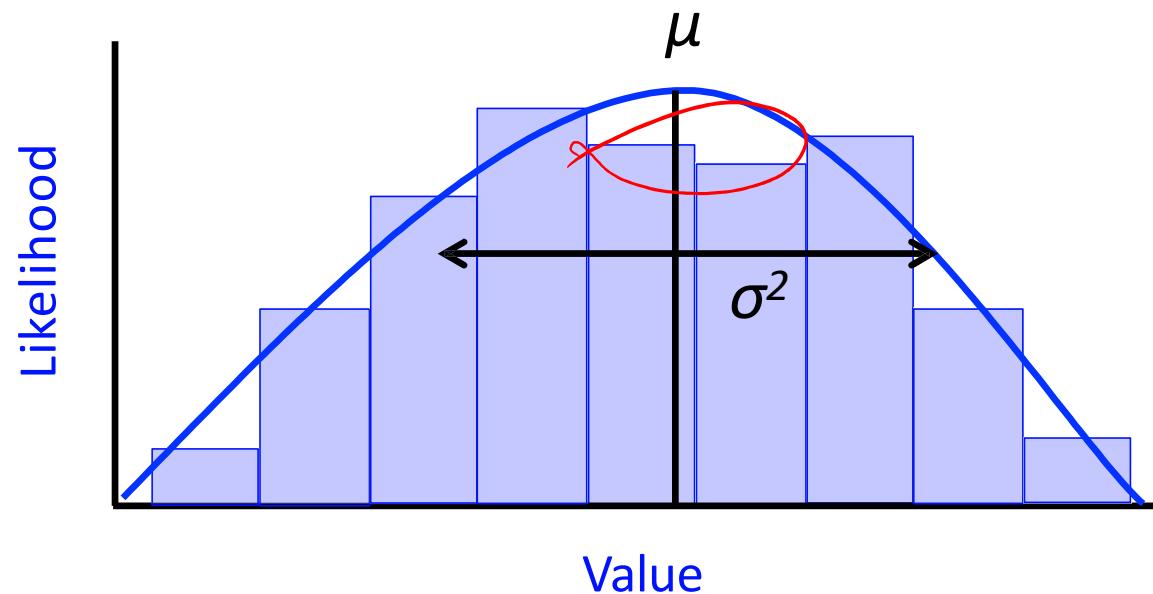


When We Fit Models To Data, We Try To Keep It Simple



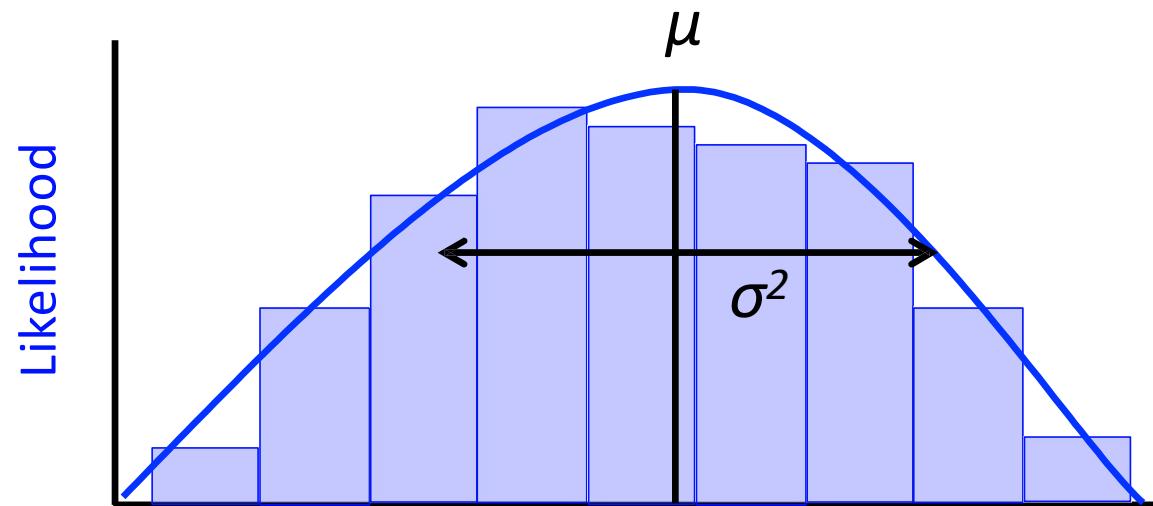
This curve fits the data well, but does it really represent the distribution?
Or is it “overfit”, so that the curve captures too much of the noise?

When We Fit Models To Data, We Try To Keep It Simple



This curve fits the data about as well, but appears to overfit less.
We could say that this simpler distribution makes fewer assumptions.
The formal concept for this idea is entropy

When We Fit Models To Data, We Try To Keep It Simple



For a fixed mean and variance, the unique distribution that maximizes the entropy is the normal distribution.

Entropy

- Measures the amount of uncertainty associated with a distribution.
- High entropy \rightarrow high uncertainty or chaos.
- Formula of entropy:

x is continuous.

Entropy: $x, f(x)$

$$\text{Entropy}(x) = - \int f(x) \log f(x) dx$$

$\text{Ent}(x) \xrightarrow{x \rightarrow \text{discrete}} E(x) \leq 0 \quad E(x) > 0$

Minimum entropy $p(x_i) = 1$ for any $i=1$

Maximum entropy: $p(x_i) = \frac{1}{k}$ for all i

Discrete $x, \text{ PMF } P(x)$
 $x \in \{x_1, x_2, \dots, x_k\}$

$$\text{Entropy}(x) = - \sum p(x_i) \log p(x_i)$$

Goal: find $p(x_i)$ s.t.
max $p(x_1) \cdot p(x_k) - \sum_{x_i} p(x_i) \log p(x_i)$
s.t $p(x_i) \geq 0$
 $\sum_{i=1}^k p(x_i) = 1$

Question in class

Optional information: Not in syllabus

- Example of a distribution with negative entropy: $X \sim U(0,1/2)$
- What is the interpretation of entropy for continuous R.V.
 - Entropy for continuous R.V is more precisely referred to as Differential entropy

The differential entropy describes the equivalent side length (in logs) of the set that contains most of the probability of the distribution.

This is nicely illustrated and explained in Theorem 8.2.3 in *Elements of Information Theory* by Thomas M. Cover, Joy A. Thomas

https://poincare.matf.bg.ac.rs/nastavno/viktor/Differential_Entropy.pdf

<https://stats.stackexchange.com/questions/256203/how-to-interpret-differential-entropy>

Entropy of Gaussian distribution.

$$\begin{aligned} \text{Ent}_s(x) &= - \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} f(x) \right] \left[\log e^{-\frac{(x-\mu)^2}{2\sigma^2}} - \log \sqrt{2\pi}\sigma \right] dx \\ &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} [f(x)] dx + \left[\log \sqrt{2\pi}\sigma \int f(x) dx \right]_{-\infty}^{\infty} + \log(\sqrt{2\pi}\sigma) \\ &= \frac{\sigma^2}{2\sigma^2} \\ &= \frac{1}{2} + \log \sqrt{2\pi}\sigma \end{aligned}$$

Proof that Gaussian distribution maximizes entropy given fixed mean and variance.

- Not in syllabus...
- For the interested, check out.

https://en.wikipedia.org/wiki/Differential_entropy

<https://medium.com/mathematical-musings/how-gaussian-distribution-maximizes-entropy-the-proof-7f7dcb2caf4d>

<https://statproofbook.github.io/P/norm-maxent.html>

Why is the Gaussian density defined so?

Optional topic: Not in syllabus.

- One student asked after class: how did the Gaussian density end up with such a non-intuitive form?
- It is possible to derive the Gaussian density function just starting from the desire to maximize entropy while matching a given mean μ , and variance σ^2
- Proof here: https://en.wikipedia.org/wiki/Differential_entropy

And here:

- [How Gaussian Distribution Maximizes Entropy — The Proof | by Freedom Preetham | Mathematical Musings | Medium](#)

CDF of a Gaussian distribution

- $X \sim N(0,1)$

↳ Standard normal distribution.

$$P(X \leq x) = \Phi(x) = F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-(z)^2/2} dz$$

$F_X(x)$

Not easy to compute in closed form: You can use libraries to access pre-computed values.

- CDF $F_Y(y)$ of a general $Y \sim N(\mu, \sigma^2)$

• Convert Y to standard form $X = \frac{Y-\mu}{\sigma} \sim N(0, 1)$

$$F_Y(y) = F_X\left(\frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

$$P(Y \leq y) = P(\sigma X + \mu \leq y) = P\left(X \leq \frac{y-\mu}{\sigma}\right) = F_X\left(\frac{y-\mu}{\sigma}\right)$$

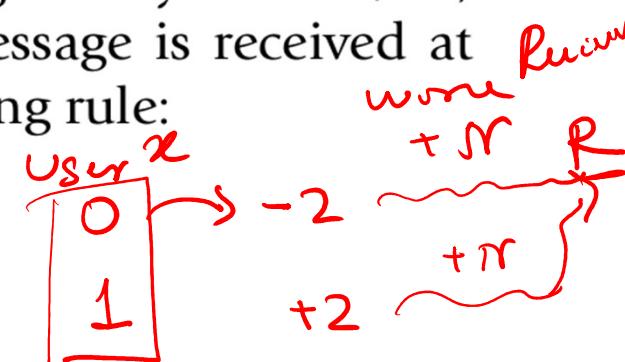
from a library



Example 5.5.b. Suppose that a binary message — either "0" or "1" — must be transmitted by wire from location A to location B. However, the data sent over the wire are subject to a channel noise disturbance and so to reduce the possibility of error, the value 2 is sent over the wire when the message is "1" and the value -2 is sent when the message is "0." If x , $x = \pm 2$, is the value sent at location A then R , the value received at location B, is given by $R = x + N$, where N is the channel noise disturbance. When the message is received at location B, the receiver decodes it according to the following rule: Review

if $R \geq .5$, then "1" is concluded

if $R < .5$, then "0" is concluded



Because the channel noise is often normally distributed, we will determine the error probabilities when N is a standard normal random variable.

$$N \sim \mathcal{N}(0, 1)$$

Let y denote the final 0/1 decoded value at the receiver.

phi represents here the cumulative density function.

$$P(y=0 | x=1) = P(N < -1.5) \quad N \sim \mathcal{N}(0, 1)$$

Symmetry around 0 in Normal Distribution.

$$= \Phi(-1.5) = 1 - \Phi(1.5) = 0.0668$$

It should be just $1 - \Phi(-1.5)$

$$P(y=1 | x=0) = P(N > 2.5)$$
$$= 1 - P(N \leq 2.5)$$
$$= 1 - \Phi(2.5) = 0.0062$$

$$P\{\text{error} | \text{message is "1"}\} = P\{N < -1.5\}$$
$$= 1 - \Phi(1.5) = .0668$$

and

$$P\{\text{error} | \text{message is "0"}\} = P\{N > 2.5\}$$
$$= 1 - \Phi(2.5) = .0062$$

Properties

MGF of $Z \sim N(0,1)$

$$E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2 + tz} dz$$
$$= \frac{1}{\sqrt{2\pi}} e^{t^2/2}$$

MGF of $X \sim N(\mu, \sigma^2)$

Properties

MGF of $Z \sim N(0,1)$

$$\begin{aligned} E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= e^{t^2/2} \end{aligned}$$

$$X = \sigma Z + \mu$$

MGF of $X \sim N(\mu, \sigma^2)$

$$\begin{aligned} E[e^{tX}] &= E[e^{t\mu + t\sigma Z}] \\ &= E[e^{t\mu} e^{t\sigma Z}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{t\mu} e^{(\sigma t)^2/2} \\ &= e^{\mu t + \sigma^2 t^2/2} \end{aligned}$$

Sum of Gaussian Random Variables

- Let $\underline{Y} = \underline{X_1 + X_2 + \cdots + X_n}$
 - Where each $\underline{X_i} \sim N(\underline{\mu_i}, \underline{\sigma_i^2})$
 - What is the distribution of \underline{Y} ?
-
- $\underline{Y} \sim N(\sum_i \underline{\mu_i}, \sum_i \underline{\sigma_i^2})$
 - Proof via MGF.

Sum of Gaussian random variable is also a Gaussian Random Variable

MGF of sum of Gaussians

$$\begin{aligned}
 \bullet E(e^{tY}) &= E_Y \left(e^{t(x_1 + x_2 + \dots + x_n)} \right) \\
 &= E_Y \left(e^{tx_1} \cdot e^{tx_2} \cdot \dots \cdot e^{tx_n} \right) = \\
 &= E_{x_1} \left(e^{tx_1} \right) \cdot \dots \cdot E_{x_n} \left(e^{tx_n} \right) \\
 &= \prod_{i=1}^n E_{x_i} \left(e^{tx_i} \right) = \prod_{i=1}^n e^{\mu_i + t^2 \sigma_i^2 / 2} \\
 &= e^{\sum_i \mu_i + t^2 \sum_i \sigma_i^2 / 2} \\
 &= \boxed{e^{\mu + t^2 \sigma^2}}
 \end{aligned}$$

where $\mu = \sum_i \mu_i$
 $\sigma^2 = \sum_i \sigma_i^2$
 since \rightarrow MGF $\mathcal{N}(\mu, \sigma^2)$ [\therefore MGF & density fn. have a 1-1 correspondence]