

03. Fisher Information: Statistics

1. Motivation: Curve Fitting

- Task: To fit a probability distribution $p(x; \theta_j)$ characterized by one or more parameters θ_j to a set of data $\{x_1, \dots, x_\ell\}$.

Idea: For a given data set $\{x_i\}$:

Step 1: Chose a general form of p that matches the data.

Step 2: Find a concrete form of p by using the data to construct an estimate of p 's parameter.

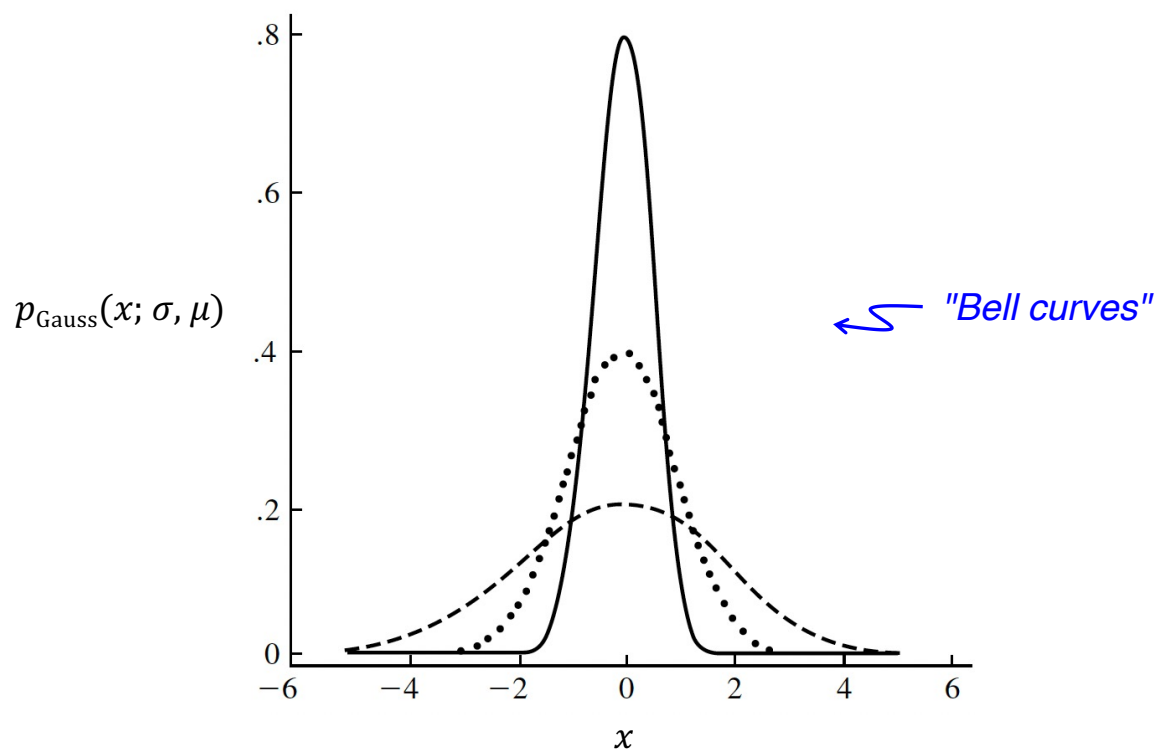
- Question: Is there a measure of how accurately a data set estimates a parameter θ of a probability distribution p ?
 - Answer: Yes! And it's called the Fisher information $I_{\text{Fisher}}(p)$ of $\{x_i\}$ with respect to $p(x; \theta)$.

Examples of probability distributions

Ex. 1. Gaussian (or Normal) distribution.

$$p_{\text{Gauss}}(x; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

Parameters: μ = mean, σ = standard deviation



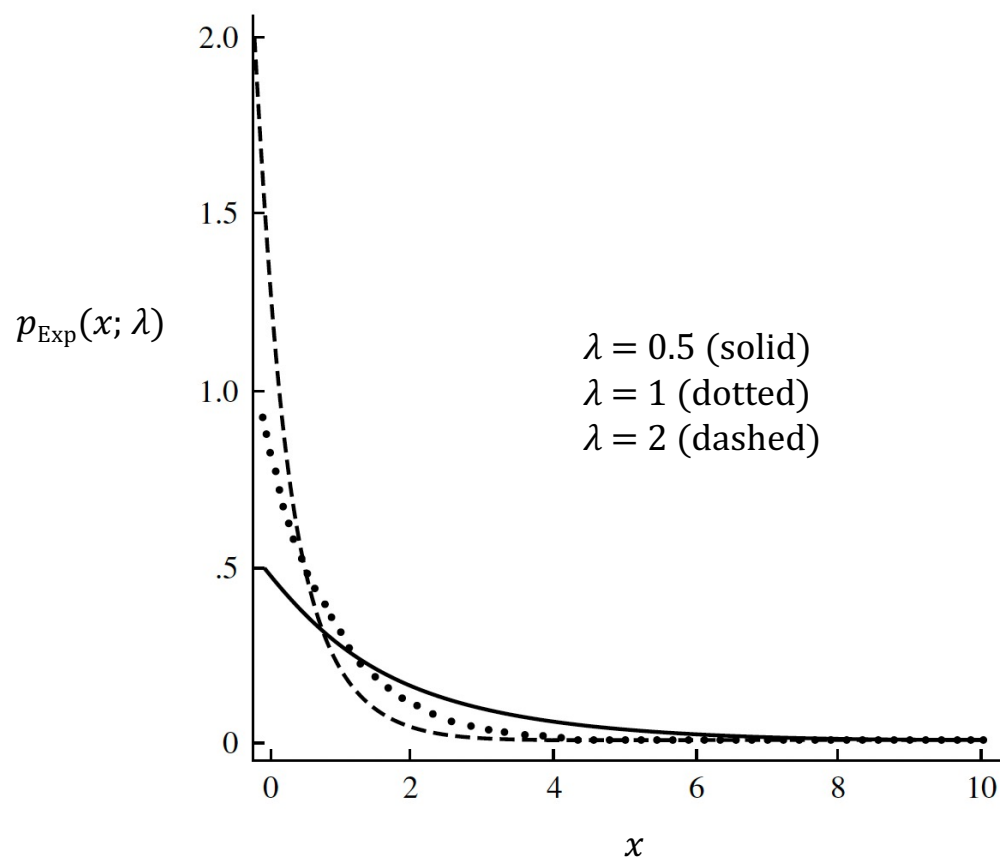
$\mu = 0$ and $\sigma = 0.5$ (solid)
 $\mu = 0$ and $\sigma = 1$ (dotted)
 $\mu = 0$ and $\sigma = 2$ (dashed)

Examples of probability distributions

Ex. 2. Exponential distribution.

$$p_{\text{Exp}}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Parameter: $\lambda = \text{rate}$



Examples of probability distributions

Ex. 3. Gamma distribution.

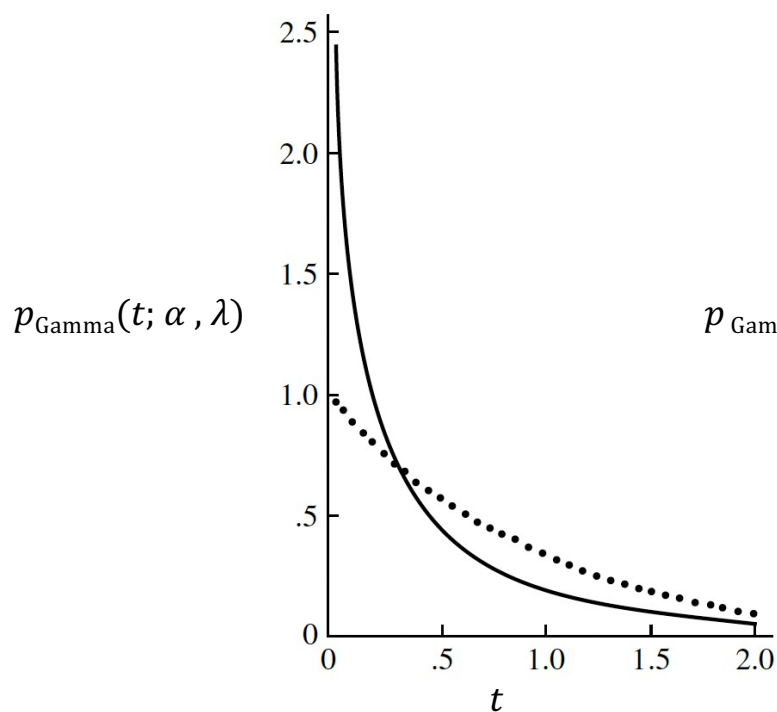
$$p_{\text{Gamma}}(t; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, \quad t \geq 0,$$

Gamma function:

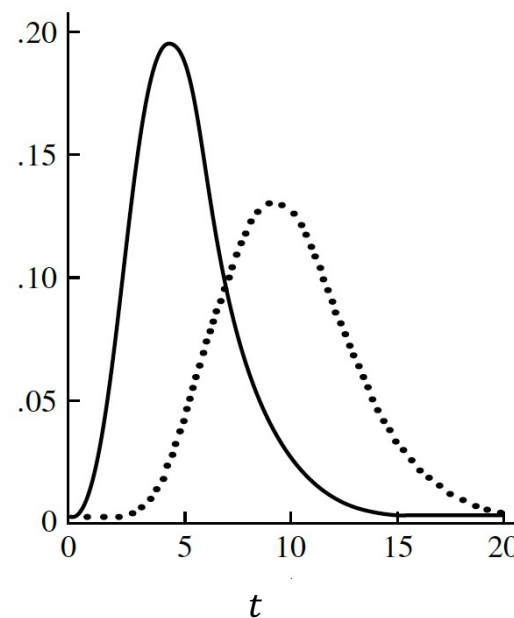
$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du, \quad \alpha > 0$$

Parameters: α = shape, λ = rate

↪ $\alpha = 1$ reproduces
exponential distribution!



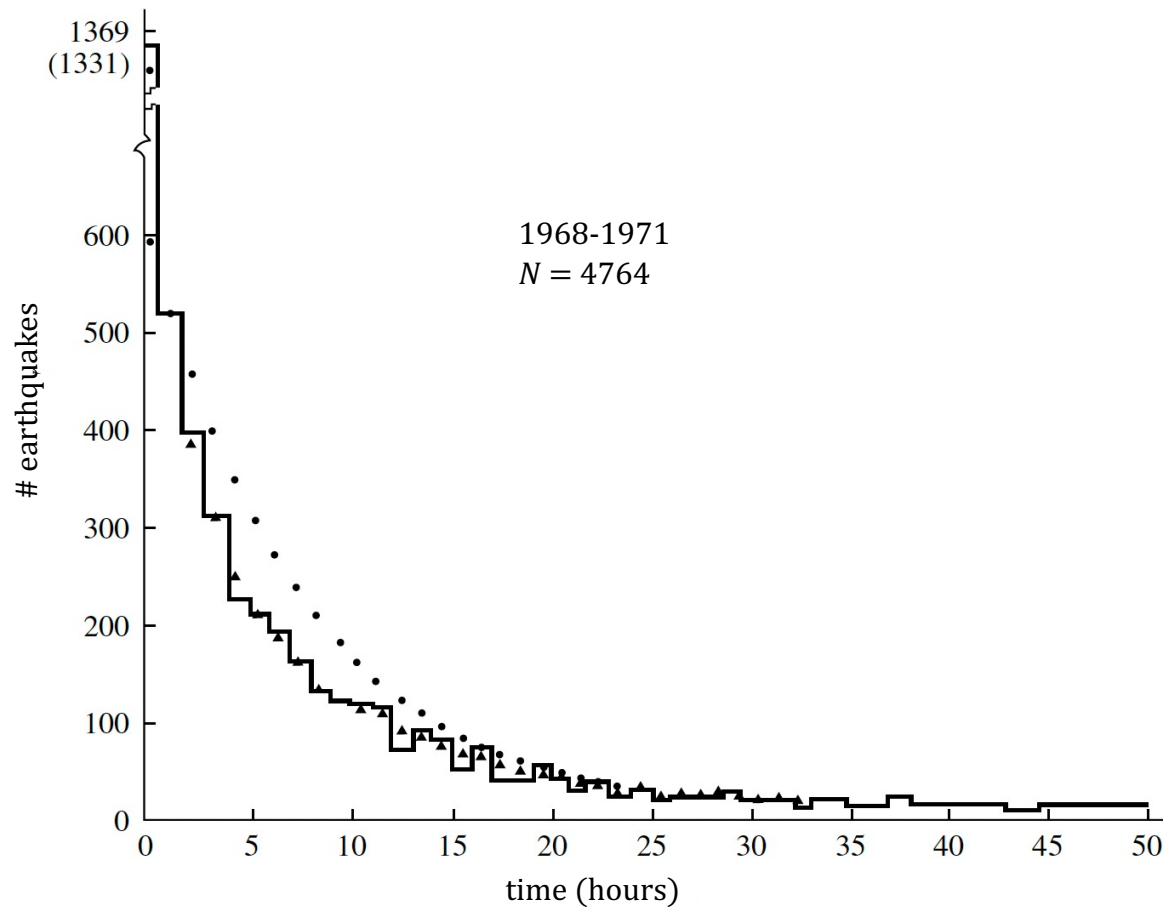
$\alpha = 0.5, \lambda = 1$ (solid)
 $\alpha = 1, \lambda = 1$ (dotted)



$\alpha = 5, \lambda = 1$ (solid)
 $\alpha = 10, \lambda = 1$ (dotted)

Choosing a probability distribution for a data set

Ex. Data set = Observed times separating a sequence of small earthquakes, Hollister, CA.*



Observed data (solid line)

p_{Exp} (dots)

p_{Gamma} : $\alpha = 0.509$, $\lambda = 0.00115$ (triangles)

- How good of a fit are this particular parameter values for the gamma distribution?
- How much "information" about the data set do they contain?

*Udias, A. & J. Rice (1975) "Statistical Analysis of Microearthquake Activity Near San Andreas Geophysical Observatory, Hollister, CA", *Bulletin of the Seismology Society of America* 65, pp. 809-28.

2. Fisher Information

- For a probability distribution $p(x; \theta)$ with unknown parameter θ , we can use the data set $\{x_i\}$ to construct an *estimate* $\hat{\theta}$ of θ .
Many ways to do this: sample mean, ratio, method of moments, method of max likelihood.
- Task: Find a measure of how good $p(x; \hat{\theta})$ fits the data.

Def. 1 (*Expected error*). Let $\hat{\theta}$ be an estimate of a parameter θ in a probability distribution $p(x; \theta)$. The **expected error** $E(\hat{\theta} - \theta)$ of $\hat{\theta}$ is given by:

$$E(\hat{\theta} - \theta) = \int (\hat{\theta} - \theta)p(x; \theta)dx$$

The average error with respect to the distribution p .

Def. 2 (*Mean squared error*). Let $\hat{\theta}$ be an estimate of a parameter θ in a probability distribution $p(x; \theta)$. The **mean squared error** e^2 of $\hat{\theta}$ is given by,

$$e^2 = \int (\hat{\theta} - \theta)^2 p(x; \theta)dx$$

The average squared error with respect to the distribution p .

- Question: Is there a lower bound on the mean squared error of an estimate $\hat{\theta}$?
If a given estimate achieved this bound, we would know it could not be improved upon!

Answer: Yes! The "*Cramer-Rao Inequality*"...

Claim (*Cramer-Rao Inequality*). Let $\hat{\theta}$ be an estimate of a parameter θ in a probability distribution $p(x; \theta)$. If $\hat{\theta}$ is "unbiased"; i.e., $E(\hat{\theta} - \theta) = 0$, then the mean squared error has a lower bound:

$$e^2 \geq \frac{1}{I_{\text{Fish}}(p; \theta)}$$

where $I_{\text{Fish}}(p; \theta) = \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right)^2 p(x; \theta) dx$

- So: As $I_{\text{Fish}}(p; \theta)$ increases, e^2 decreases.
 - Which means: The greater $I_{\text{Fish}}(p; \theta)$ is, the more accurate an unbiased estimate $\hat{\theta}$ of the parameter θ in $p(x; \theta)$ is with respect to the data set $\{x_i\}$.

$I_{\text{Fish}}(p; \theta)$ is called the "Fisher information" of the probability distribution $p(x; \theta)$.

- Note: $I_{\text{Fish}}(p; \theta)$ is independent of any particular estimate $\hat{\theta}$.
- So: Regardless of the estimate you're using for θ , $I_{\text{Fish}}(p; \theta)$ tells you how accurate it is!

Proof of Cramer-Rao Inequality

For an unbiased estimate, $E(\hat{\theta} - \theta) = \int (\hat{\theta} - \theta)p(x; \theta)dx = 0$

So: $\frac{\partial}{\partial \theta} \int (\hat{\theta} - \theta)pdx = - \int pdx + \int (\hat{\theta} - \theta) \frac{\partial p}{\partial \theta} dx = 0$

Or: $\int (\hat{\theta} - \theta) \frac{\partial p}{\partial \theta} dx = 1 \quad \leftarrow \int pdx = 1$

Or: $\int (\hat{\theta} - \theta) \frac{\partial \log p}{\partial \theta} p dx = 1 \quad \leftarrow \frac{\partial p}{\partial \theta} = \left(\frac{\partial \log p}{\partial \theta} \right) p$

Or: $\int \left(\frac{\partial \log p}{\partial \theta} \sqrt{p} \right) \left((\hat{\theta} - \theta) \sqrt{p} \right) dx = 1$

Or: $\langle f, g \rangle = 1 \quad \leftarrow f \equiv \left(\frac{\partial \log p}{\partial \theta} \sqrt{p} \right), g \equiv \left((\hat{\theta} - \theta) \sqrt{p} \right), \text{ and } \langle , \rangle \text{ is the inner product}$

So: $\underbrace{\int \left(\frac{\partial \log p}{\partial \theta} \right)^2 p dx}_{I_{\text{Fish}}(p; \theta)} \underbrace{\int (\hat{\theta} - \theta)^2 p dx}_{e^2} \geq 1 \quad \leftarrow \text{by the Cauchy-Schwarz Inequality: } \langle f, f \rangle \langle g, g \rangle \geq |\langle f, g \rangle|^2$

Def. 3 (Fisher Information). The **Fisher Information** $I_{\text{Fish}}(p; \theta)$ of a probability distribution $p(x; \theta)$ is given by:

$$I_{\text{Fish}}(p; \theta) = \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right)^2 p(x; \theta) dx$$

- Note 1: As an inverse lower-bound on e^2 , $I_{\text{Fish}}(p; \theta)$ is a measure of how accurately a probability distribution $p(x; \theta)$ fits a data set $\{x_i\}$ with respect to (unbiased) estimates of the parameter θ .
- Note 2: Let $\left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right)^2$ be the *sensitivity* of $p(x; \theta)$ to θ with respect to the data set $\{x_i\}$.
 $\partial/\partial\theta \log p$ indicates how $\log p$ changes as θ changes
- Then: $I_{\text{Fish}}(p; \theta)$ is the *average sensitivity* of $p(x; \theta)$ to θ with respect to the data set $\{x_i\}$.

Def. 3 (Fisher Information). The **Fisher Information** $I_{\text{Fish}}(p; \theta)$ of a probability distribution $p(x; \theta)$ is given by:

$$I_{\text{Fish}}(p; \theta) = \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right)^2 p(x; \theta) dx$$

- Note 3:

Claim. The Fisher information $I_{\text{Fish}}(p; \theta)$ can also be expressed by

$$I_{\text{Fish}}(p; \theta) = - \int \left(\frac{\partial^2}{\partial \theta^2} \log p(x; \theta) \right) p(x; \theta) dx$$

- So: The *sensitivity* of $p(x; \theta)$ to θ with respect to the data set $\{x_i\}$ can also be identified with $-\frac{\partial^2}{\partial \theta^2} \log p(x; \theta)$.



The "gradient content" of p : how "spread out" it is.

Proof of Claim

Note: $\int p(x; \theta) dx = 1$

So: $\frac{\partial}{\partial \theta} \int p(x; \theta) dx = 0$

So: $0 = \int \frac{\partial}{\partial \theta} p(x; \theta) dx = \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right) p(x; \theta) dx$

So: $0 = \frac{\partial}{\partial \theta} \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right) p(x; \theta) dx$

$$= \int \left(\frac{\partial^2}{\partial \theta^2} \log p(x; \theta) \right) p(x; \theta) dx + \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right) \frac{\partial p(x; \theta)}{\partial \theta} dx$$

$$= \int \left(\frac{\partial^2}{\partial \theta^2} \log p(x; \theta) \right) p(x; \theta) dx + \int \left(\frac{\partial}{\partial \theta} \log p(x; \theta) \right)^2 p(x; \theta) dx$$

$$\frac{\partial p}{\partial \theta} = \left(\frac{\partial \log p}{\partial \theta} \right) p$$




Example: Gaussian distribution

$$p_{\text{Gauss}}(x; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Note: $\log[p_{\text{Gauss}}] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$

So: $\frac{\partial}{\partial \mu} \log[p_{\text{Gauss}}] = -\frac{x-\mu}{\sigma^2}$

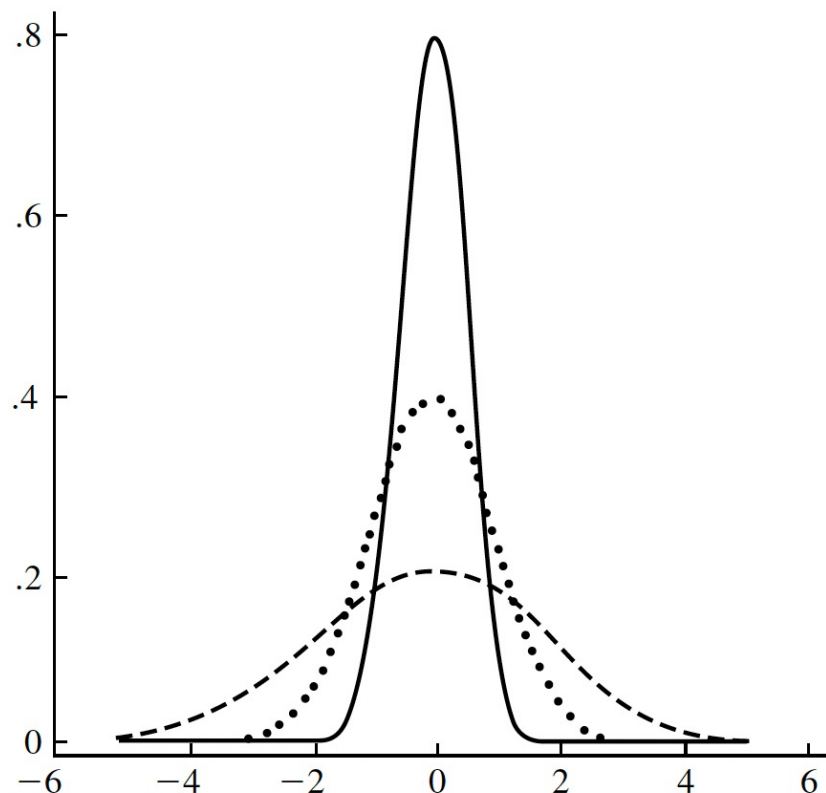
And: $\frac{\partial^2}{\partial \mu^2} \log[p_{\text{Gauss}}] = -\frac{1}{\sigma^2}$

So: $I_{\text{Fish}}(p_{\text{Gauss}}; \mu) = -\int \left(\frac{\partial^2}{\partial \mu^2} \log[p_{\text{Gauss}}] \right) p_{\text{Gauss}} dx$

$$= \frac{1}{\sigma^2} \int p_{\text{Gauss}} dx$$
$$= \frac{1}{\sigma^2}$$



The "variance" σ^2 of p_{Gauss} :
how "spread out" it is.



$\mu = 0$ and $\sigma = 0.5$ (solid)

$\mu = 0$ and $\sigma = 1$ (dotted)

$\mu = 0$ and $\sigma = 2$ (dashed)

- Suppose these Gaussians are plots of data sets produced by three different thermometers (random variables) that measure the temperature of a room (unknown parameter).
- The solid set has the smallest σ^2 (sharpest "peak"), hence the largest I_{Fish} , and hence is the most sensitive to changes in the temperature of the room.

2. Fisher Information vs. Shannon Entropy

Fisher Information

$$I_{\text{Fish}}(p; \theta) = - \int \left(\frac{\partial^2}{\partial \theta^2} \log p(x; \theta) \right) p(x; \theta) dx$$

- The average value of the sensitivity $-\frac{\partial^2}{\partial \theta^2} \log p(x; \theta)$ of a probability distribution $p(x; \theta)$ to θ .
- How spread out p is with respect to θ .

Shannon entropy

$$S_{\text{Shan}}(p) = - \int (\log_2 p(x)) p(x) dx$$

- The average value of the Shannon information content $-\log_2 p(x)$ of a probability distribution $p(x)$.
- How spread out p is.

• Recall: $I_{\text{Fish}}(p_{\text{Gauss}}; \mu) = \frac{1}{\sigma^2}$

The "variance" σ^2 of p_{Gauss} :
how "spread out" it is.

Claim. $S_{\text{Shan}}(p_{\text{Gauss}}) = \log_2 \sqrt{2\pi e} + \log_2 \sigma$

The "standard deviation" σ of p_{Gauss} :
also how "spread out" it is.

Note: The greater the spread of p ,
the more its Shannon entropy, and
the less its Fisher information.

Proof of Claim:

$$p_{\text{Gauss}}(x; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\underline{\text{So:}} \log_2[p_{\text{Gauss}}] = -\frac{1}{2}\log_2(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\log_2(e)$$

$$\begin{aligned}\underline{\text{So:}} S_{\text{Shan}}(p_{\text{Gauss}}) &= -\int \log_2(p_{\text{Gauss}}) p_{\text{Gauss}} dx \\ &= \frac{1}{2}\log_2(2\pi\sigma^2) \underbrace{\int p_{\text{Gauss}} dx}_1 + \frac{\log_2(e)}{2\sigma^2} \underbrace{\int (x-\mu)^2 p_{\text{Gauss}} dx}_{\sigma^2} \\ &= \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{\log_2(e)}{2\sigma^2} \sigma^2 \\ &= \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{1}{2}\log_2(e) \\ &= \frac{1}{2}\log_2(2\pi\sigma^2 e) \\ &= \log_2\sqrt{2\pi e} + \log_2\sigma\end{aligned}$$