Statistical Processes Lecture 2

Physics 129AL

Zihang Wang
10/18/2023

Two Schools of Statistics

Recall what we discussed last time,

Frequentist Goal:

Construct procedures with frequency guarantees

Bayesian Goal:

Describe and update degree of belief in propositions

https://www.psi.ch/en/media/our-research/observation-of-a-new-particle-with-a-mass-of-125-gev

Probabilities

Recall that discrete or continuous event are represented as random variables,

- Random Variable: x
- Outcomes: $S \equiv \{x_1, x_2, ...\}$

As an example, let consider the probabilities of a dice with the following properties,

$$p(E): \quad p_{\text{dice}}(1) = \frac{1}{6}$$
$$p(E) \geq 0$$
$$p(A\&B) = p(A) + p(B)$$
$$p(S) = 1: \sum_i p(x_i) = 1$$

Probabilities

We can define **joint probability** of two random variables, x and y (not necessarily independent),

$$P(x, y) \quad x \in S_x,\ y \in S_y \qquad P(x, y) = P(x)P(y)$$

Independent random variables

Or **marginal Probability** of two random variables, x and y

$$P(x = x_i) = \sum_{y \in S_y} P(x = x_i, y)$$

And **conditional Probability** of two random variables, x and y

$$P(x = x_i \mid y = y_j) = \frac{P(x = x_i, y = y_j)}{P(y = y_j)}$$

Marginal

Probabilities

Product rule (Chain Rule, giving conditional),

$$P(x, y \,|\, \mathcal{H}) = P(x \,|\, y, \mathcal{H}) P(y \,|\, \mathcal{H}) = P(y \,|\, x, \mathcal{H}) P(x \,|\, \mathcal{H}).$$

where $\mathcal{H}$: Generative Model

Sum rule (giving marginal)

$$\begin{aligned} P(x \,|\, \mathcal{H}) &= \sum_y P(x, y \,|\, \mathcal{H}) \\ &= \sum_y P(x \,|\, y, \mathcal{H}) P(y \,|\, \mathcal{H}). \end{aligned}$$

And **Bayes Theorem**

$$\begin{aligned} P(y \,|\, x, \mathcal{H}) &= \frac{P(x \,|\, y, \mathcal{H}) P(y \,|\, \mathcal{H})}{P(x \,|\, \mathcal{H})} \\ &= \frac{P(x \,|\, y, \mathcal{H}) P(y \,|\, \mathcal{H})}{\sum_{y'} P(x \,|\, y', \mathcal{H}) P(y' \,|\, \mathcal{H})}. \end{aligned}$$

Probabilities

Example 2.3. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

$$\begin{array}{ll} a = 1 & \text{Jo has the disease} \\ a = 0 & \text{Jo does not have the disease.} \end{array} \tag{2.12}$$

The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

OK – Jo has the test, and the result is positive. What is the probability that Jo has the disease?

From Information%Theory, Inference%and Learning Algorithms, by David MacKay

Probabilities

Step 1: Write down all probabilities

We are given conditional probabilities

$$P(b=1\,|\,a=1) = 0.95 \qquad P(b=1\,|\,a=0) = 0.05$$
$$P(b=0\,|\,a=1) = 0.05 \qquad P(b=0\,|\,a=0) = 0.95;$$

And marginal probability of a

$$P(a=1) = 0.01 \qquad P(a=0) = 0.99.$$

We want $P(a=1|b=1)$

Probabilities

Step 2: Deduce joint probability $P(a,b)$

$$P(a,b)=P(a|b)P(b)=P(b|a)P(a)$$

Step 3: conditional probability $P(a|b)$

$$\begin{aligned} P(a=1\,|\,b=1) &= \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1\,|\,a=1)P(a=1)+P(b=1\,|\,a=0)P(a=0)} \\ &= \frac{0.95\times 0.01}{0.95\times 0.01+0.05\times 0.99} \\ &= 0.16. \end{aligned}$$

Lots of false positives!

Statistical Properties: Moments

- Expectation value $\langle F(x)\rangle = \int_{-\infty}^{\infty} \mathrm{d}x\ p(x)F(x).$
- Moments $m_n \equiv \langle x^n \rangle = \int \mathrm{d}x p(x) x^n.$
- Characteristic function generates moments:

$$\tilde{p}(k) = \left\langle \mathrm{e}^{-\mathrm{i}kx} \right\rangle = \int \mathrm{d}x p(x)\ \mathrm{e}^{-\mathrm{i}kx}.$$

Fourier Transform of the PDF: contains same information as PDF *p(x)*

Recovering the PDF from the characteristic function through the inverse F.T.:

$$p(x) = \frac{1}{2\pi} \int \mathrm{d}k \tilde{p}(k)\, \mathrm{e}^{+\mathrm{i}kx}.$$

$$\int e^{-i(w-w')\tau} d\tau = 2\pi\delta_D(w-w')$$

$$\int \frac{dw'}{2\pi} 2\pi\delta_D(w-w') = 1$$

Moments of the distribution:

$$\tilde{p}(k) = \left\langle \sum_{n=0}^{\infty} \frac{(-\mathrm{i}k)^n}{n!} x^n \right\rangle = \sum_{n=0}^{\infty} \frac{(-\mathrm{i}k)^n}{n!} \langle x^n \rangle .$$

Infinite series that in principle contains same information as p(x)

Statistical Properties: Moments

PDF moments around x_0:

$$e^{ikx_0}\tilde{p}(k) = \left\langle e^{-ik(x-x_0)} \right\rangle = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \left\langle (x-x_0)^n \right\rangle .$$

- Cumulant generating function defines cumulants $\langle x^n \rangle_c$

$$\ln \tilde{p}(k) = \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c .$$

We can obtain relations between moments and cumulants using

$$\ln(1+\epsilon) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\epsilon^n}{n}.$$

We expand ε into a sum of moments and ln(1+ε) into a sum of cumulants and match powers of k to obtain

Mean $\langle x \rangle_c = \langle x \rangle ,$

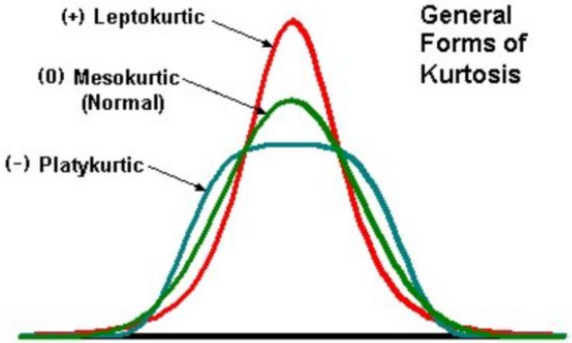
Variance $\langle x^2 \rangle_c = \langle x^2 \rangle - \langle x \rangle^2 ,$

Skewness $\langle x^3 \rangle_c = \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3 ,$

Curtosis $\langle x^4 \rangle_c = \langle x^4 \rangle - 4\langle x^3 \rangle \langle x \rangle - 3\langle x^2 \rangle^2 + 12\langle x^2 \rangle \langle x \rangle^2 - 6\langle x \rangle^4 .$

Statistical Properties: Moments

Moments characterize the probability (mass) density function.

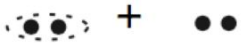


Statistical Properties: Cumulants

Moments are related to clusters of cumulants: useful pictorial that relates cumulants to moments. Cumulants are also called connected moments.

$< x > =$ 

$$\langle x \rangle = \langle x \rangle_c$$

$< x^2 > =$ 

$$\langle x^2 \rangle = \langle x^2 \rangle_c + \langle x \rangle_c^2$$

$< x^3 > =$ 

$$\langle x^3 \rangle = \langle x^3 \rangle_c + 3 \langle x^2 \rangle_c \langle x \rangle_c + \langle x \rangle_c^3$$

$< x^4 > =$ 

$$\langle x^4 \rangle = \langle x^4 \rangle_c + 4 \langle x^3 \rangle_c \langle x \rangle_c + 3 \langle x^2 \rangle_c^2 + 6 \langle x^2 \rangle_c \langle x \rangle_c^2 + \langle x \rangle_c^4$$

Example: Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

- Characteristic Function

$$\tilde{p}(k) = \int_{-\infty}^{\infty} \mathrm{d}x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2} - \mathrm{i}kx\right] = \exp\left[-\mathrm{i}k\lambda - \frac{k^2\sigma^2}{2}\right]$$

- Cumulants

From $\ln \tilde{p}(k) = -\mathrm{i}k\lambda - k^2\sigma^2/2$, we can identify:

$$\langle x \rangle_c = \lambda, \quad \langle x^2 \rangle_c = \sigma^2, \quad \langle x^3 \rangle_c = \langle x^4 \rangle_c = \cdots = 0.$$

- Moments from cluster expansion

$$\langle x \rangle = \lambda, \qquad \langle x^3 \rangle = 3\sigma^2\lambda + \lambda^3,$$

$$\langle x^2 \rangle = \sigma^2 + \lambda^2, \qquad \langle x^4 \rangle = 3\sigma^4 + 6\sigma^2\lambda^2 + \lambda^4,$$

Many Random Variables

- Joint PDF $p(\mathbf{x})$ where $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$

$$p_{\mathbf{x}}(\mathcal{S}) = \int \mathrm{d}^N \mathbf{x} p(\mathbf{x}) = 1.$$ (*S* is the set of all outcomes)

- Joint characteristic function $\tilde{p}(\mathbf{k}) = \left\langle \exp\left(-\mathrm{i}\sum_{j=1}^{N} k_j x_j\right)\right\rangle$

Example

$$\langle x_1 x_2 \rangle = \overset{1}{\bullet}\ \overset{2}{\bullet} + \overset{1}{\circ}\ \overset{2}{\circ} \qquad \langle x_1 x_2 \rangle = \langle x_1 \rangle_c \langle x_2 \rangle_c + \langle x_1 * x_2 \rangle_c$$

$$\langle x_1^2 x_2 \rangle = \overset{2}{\bullet}\ \overset{1}{\bullet} + \overset{2}{\circ}\ \overset{1}{\circ} + 2\ \overset{1}{\circ}\ \overset{2}{\circ} + \overset{2}{\triangle}\ \overset{1}{\triangle}$$

$$\langle x_1^2 x_2 \rangle = \langle x_1 \rangle_c^2 \langle x_2 \rangle_c + \langle x_1^2 \rangle_c \langle x_2 \rangle_c + 2 \langle x_1 * x_2 \rangle_c \langle x_1 \rangle_c + \langle x_1^2 * x_2 \rangle_c$$

$\langle x_\alpha * x_\beta \rangle_c$ is zero if x_α and x_β are independent (symbol * here means product)

Multi-variate Gaussian

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det[C]}} \exp\left[-\frac{1}{2}\sum_{mn}(C^{-1})_{mn}(x_m - \lambda_m)(x_n - \lambda_n)\right]$$

where C is the covariance matrix, and C^{-1} is its inverse

$$\tilde{p}(\mathbf{k}) = \exp\left[-ik_m\lambda_m - \frac{1}{2}C_{mn}k_mk_n\right]$$

Note: repeated indices are summed

- Cumulants

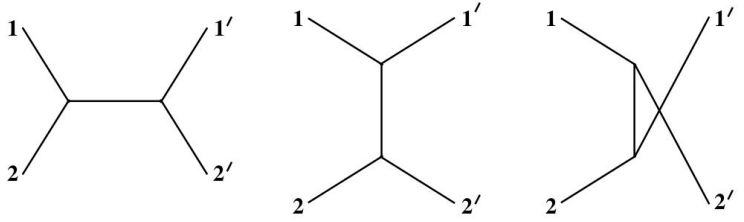
$$\langle x_m \rangle_c = \lambda_m, \quad \langle x_m * x_n \rangle_c = C_{mn}, \text{ rest are zero}$$

Wick's Theorem

4 point function for a multi-variate gaussian with zero mean depends only on products of 2 point functions, $\langle x_a x_b x_c x_d \rangle = C_{ab}C_{cd} + C_{ac}C_{bd} + C_{ad}C_{bc}$

The three tree-level Feynman diagrams that contribute to the connected correlation function are related to the Wick's theorem,

$$
\begin{aligned}
&\langle 0|\mathrm{T}\varphi(x_1)\varphi(x_2)\varphi(x_1')\varphi(x_2')|0\rangle_{\mathrm{C}} \\
&\qquad = (ig)^2 \left(\tfrac{1}{i}\right)^5 \int d^4y\, d^4z\, \Delta(y-z) \\
&\qquad\quad \times \Big[\, \Delta(x_1-y)\Delta(x_2-y)\Delta(x_1'-z)\Delta(x_2'-z) \\
&\qquad\qquad + \Delta(x_1-y)\Delta(x_1'-y)\Delta(x_2-z)\Delta(x_2'-z) \\
&\qquad\qquad + \Delta(x_1-y)\Delta(x_2'-y)\Delta(x_2-z)\Delta(x_1'-z) \,\Big] \\
&\qquad\qquad + O(g^4)\,.
\end{aligned}
$$



Sum of variables

- PDF $p_X(x) = \int \mathrm{d}^N \mathbf{x} p(\mathbf{x}) \delta\left(x - \sum x_i\right)$
- Characteristic function

$$\tilde{p}_X(k) = \left\langle \exp\left(-\mathrm{i}k \sum_{j=1}^{N} x_j\right)\right\rangle = \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right)$$

- Cumulants

$$\ln \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right) = -\mathrm{i}k \sum_{i_1=1}^{N} \left\langle x_{i_1}\right\rangle_c + \frac{(-\mathrm{i}k)^2}{2} \sum_{i_1,i_2}^{N} \left\langle x_{i_1} x_{i_2}\right\rangle_c + \cdots$$

$$\left\langle X\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i\right\rangle_c, \quad \left\langle X^2\right\rangle_c = \sum_{i,j}^{N} \left\langle x_i x_j\right\rangle_c, \cdots$$

If variables are independent cross-cumulants vanish ->

$$\left\langle X^n\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i^n\right\rangle_c \longrightarrow \left\langle X^n\right\rangle_c = N \left\langle x^n\right\rangle_c$$

If all drawn from p(x)

Central Limit Theorem

For large N $\quad < x > \propto N,\ \langle (x - \langle x \rangle)^2 \rangle \propto N$

$$y = \frac{x - N\langle x \rangle_c}{\sqrt{N}},\ \langle y^n \rangle_c \propto N^{1-n/2}$$

$$N \to \infty,\ \langle y^n \rangle_c \to 0 \text{ for } n > 2$$

$$\lim_{N\to\infty} p\left(y = \frac{\sum_{i=1}^{N} x_i - N\langle x \rangle_c}{\sqrt{N}}\right) = \frac{1}{\sqrt{2\pi \langle x^2 \rangle_c}} \exp\left(-\frac{y^2}{2\langle x^2 \rangle_c}\right)$$

Gaussian Distribution

Poisson Distribution

Example: Radioactive decay,

Probability of one and only one event (decay) in [t, $t+dt$] is proportional to dt as dt -> 0.
Probabilities of events are independent.

Poisson $p(M|T)$ M events in time interval T

- Limit of binomial: $N = \frac{T}{dt} >> 1$

$$p_1 = \alpha dt,\ p_0 = 1 - \alpha dt,\ p_2 \propto dt^2 \to 0$$

- We imagine many binomial events over time interval T

Poisson Distribution

Assume stars randomly distributed around us with density n, what is probability that the nearest star is at distance R ?

$$p(r < R, 0 \text{ star}|n) \sim \frac{(\gamma T)^0}{0!} e^{-n\frac{4\pi}{3}R^3}$$

$$p(r \approx R, 1 \text{ star}|n) \sim \frac{(n4\pi R^2 dR)^1}{1!} e^{-n4\pi R^2 dR}$$

$$p(r < R, 0 \text{ star}, r \approx R, 1 \text{ star}|n) = p(r < R, 0 \text{ star}|n)p(r \approx R, 1 \text{ star}|n)$$

$$\sim n4\pi R^2 dR e^{-n4\pi R^2 dR} e^{-n\frac{4\pi}{3}R^3} \approx n4\pi R^2 dR e^{-n\frac{4\pi}{3}R^3}$$

What do we mean by Probability

1) Frequency of outcomes for repeated random experiments
2) Degrees of belief in propositions not involving random variables (quantifying uncertainty)

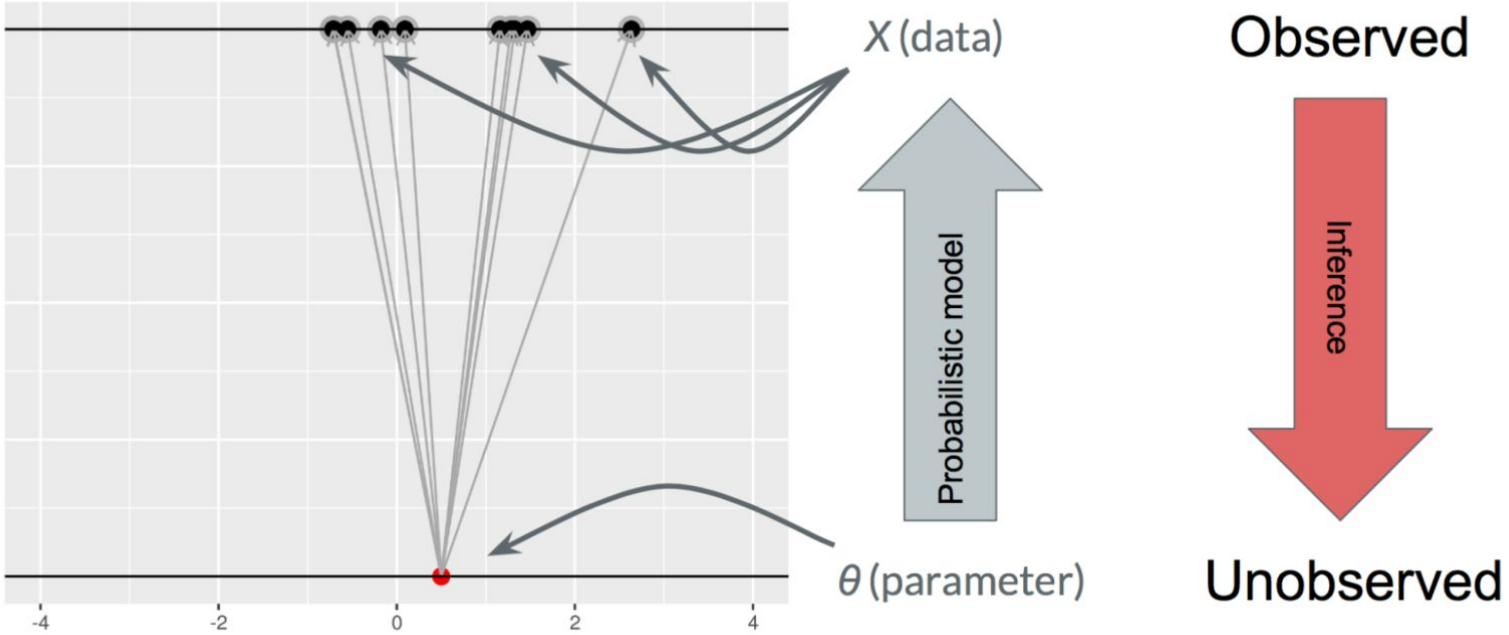
Forward Probability

Generative model describing a process giving rise to some data.

Inverse Probability

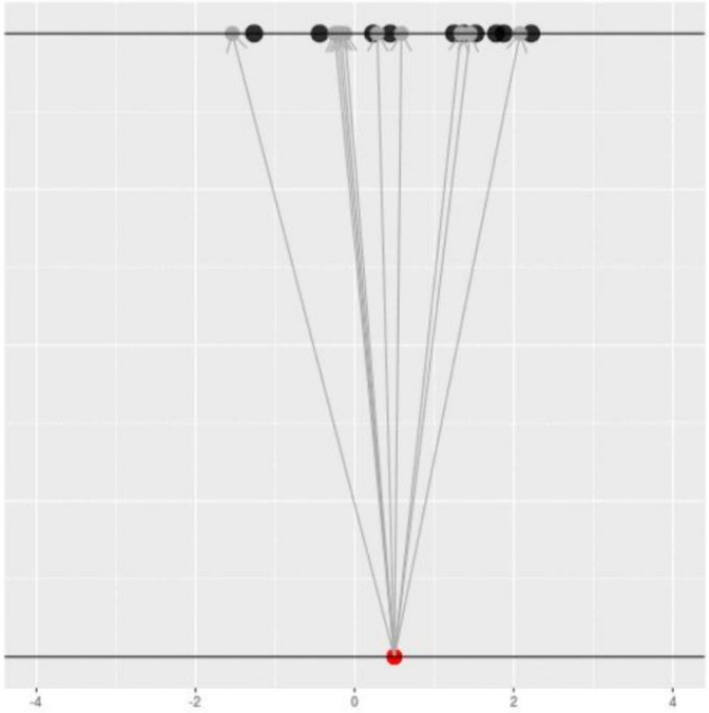
We compute probability of some unobserved quantity, given the observed variables, i.e. Bayes theorem.

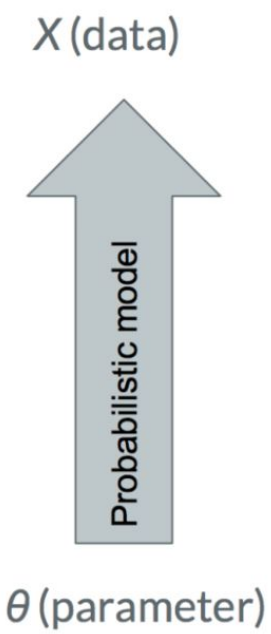
Parameters and Data



Slides'credit: Ryan Giordano

Frequentist Inference



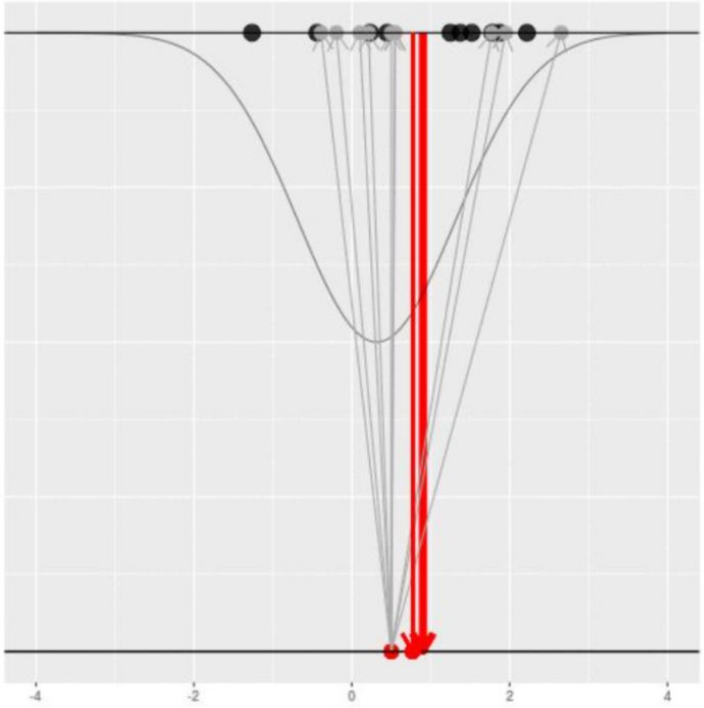


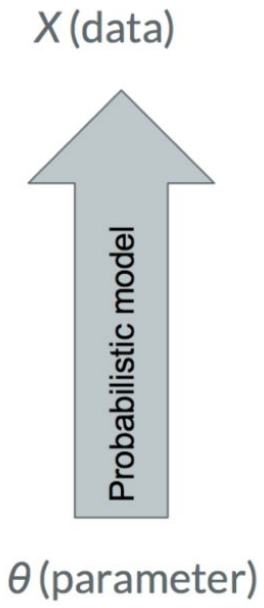
Frequentist idea:

We got the parameter indicated by the red dot and saw the dataset in black.

But the same parameter could have given us lots of other datasets.

Frequentist Inference





Frequentist idea:

For each dataset, we might pick some summary function and call it an “estimate”.
It will be different each time because the data will be different each time.

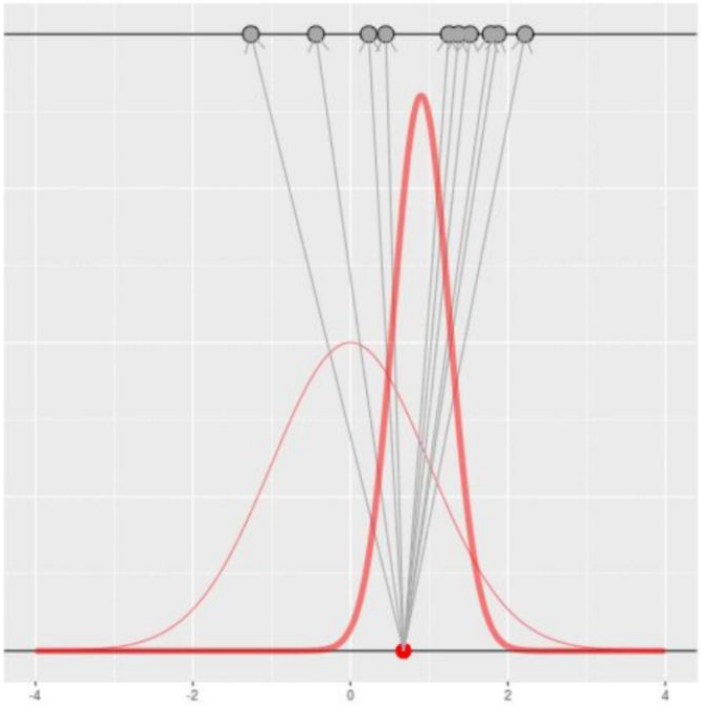
A typical estimate -- but not the only one -- is the value that maximizes the likelihood of the data.

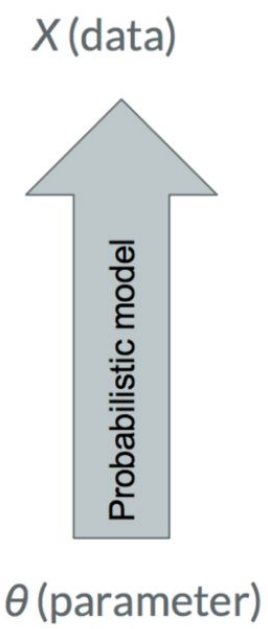
We hope the estimate is usually near the true parameter in some sense.

Frequentist Inference

The frequentist uses the concept of frequency or repeated sampling. It focuses on estimating statistical parameters, e.g. means and variances and making statistical inferences based on the given data. In frequentist statistics, parameters governing the underlying distribution are treated as **fixed, unknown values**, and the goal is to estimate these parameters using **point estimates** (e.g., maximum likelihood) or confidence intervals. The frequentist approach does not incorporate prior beliefs on the parameter p (it is given and fixed), and **it relies solely on the data at hand.**

Parameters and Data





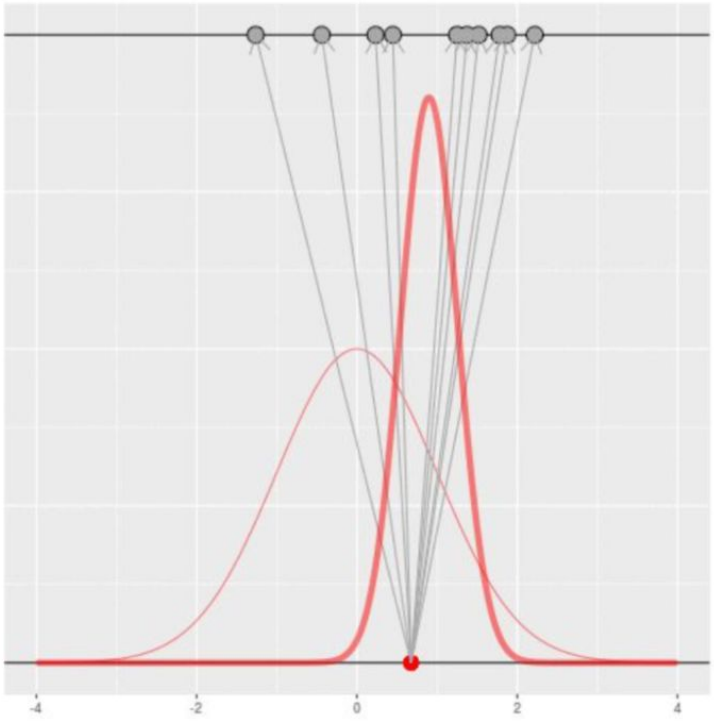
Bayesian idea:

Suppose we draw a bunch of parameters and datasets, and then throw out every pair where the data doesn't match what we observed.

The distribution of the parameters that are left represents which parameters could have given us the dataset we saw.

We hope the prior is reasonable and the model is accurate.

Parameters and Data



Of course, in practice you don't usually generate parameters and data hoping to get your original dataset.

Instead, you use Bayes' rule:

$$p\left(\theta|x\right) = \frac{p\left(x|\theta\right)p\left(\theta\right)}{p\left(x\right)}$$

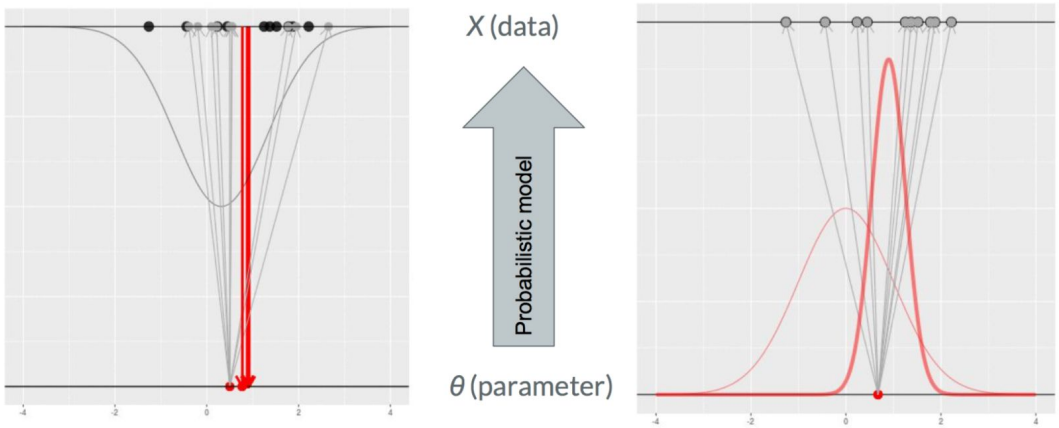
This is intractable in general (the denominator is a problem). Turn to approximations schemes like MCMC, variational Bayes, &c.

Bayesian Inference

Bayesian inference treats both observed data and parameters as random variables with probability distributions, in contrast to the frequentist approach. Bayesian inference typically begins with **prior beliefs or knowledge about the controlling parameters**. The likelihood function quantifies the probability of observing the data with a specific set of controlling parameters and often relies on certain assumptions. The posterior probability represents the **updated beliefs about controlling parameters after integrating the observed data**. It can then be employed as the new prior for subsequent observations, enabling the continuous refinement of beliefs in the presence of additional data.

Frequentist v.s. Bayesian Inference

Bayesian inference incorporates subjective prior beliefs and provides posterior probability distributions of the controlling parameters, while frequentist inference focuses on objective measures based solely on observed data. The choice between the two approaches often depends on the specific problem and available data.



Acknowledgement

The slides are partially developed or inspired by Professor Uros Seljak at UC Berkeley. For more information, please visit the github page.

A short story: When I was doing my undergraduate, I took his class (the exact same one, linked below).

Uros Seljak (Berkeley)

https://phy151-ucb.github.io/seljak-phy151-fall-2018/#course-syllabus