06 Sep 2019

Flaws in Frequentist Inference

e.g. (in frequentist inference) Say that an ongoing experiment is being run each month.

An independent normal variate is observed $x_t \sim N(\mu, 1)$ ← observation of month t

You're planning to run an hypothesis test to test $H_0: \mu = 0$ vs. $H_a: \mu > 0$

$$z_i = \frac{\sum_{t=1}^{i} x_t / i - 0}{1/\sqrt{i}} \quad \text{se} = \sigma = \frac{\sigma^2}{\sqrt{n}}$$

↑ up to month i

[if we know the variance σ^2, the test statistic is: $z = \frac{\bar{x} - \mu_0}{\sigma}$ variance σ^2 = standard deviation]

$$z_i = \sum_{t=1}^{i} x_t / \sqrt{i}$$

↳ is the z-score based on data up to month i

→ Say that at month 30 (the scheduled end of the experiment) $z_{30} = 1.66 > 1.645$

↑ one sided 95% for a $N(0,1)$ distribution

, Then reject H_0.

[$t_1(x)$
→ So in this case, our hypothesis testing algorithm is collect data for 30 months, once you have all the data points, compute z_{30} and compare to a critical value.

hypothesis testing

[$t_2(x)$
→ let's say we change the algorithm to: collect data for 20 months & check z_{20}, if it is significant stop (if $z_{20} > 1.645$, stop) if it's not, then collect data from 10 more months and compare z_{30} to a critical value.

→ in Frequentist, if the algorithm changes, even though the data points stay exactly the same, the significance level is different for each algorithm.
↳ which is a flaw.

→ for $t_1(x)$, $\alpha_1 = 5\%$

- for $t_2(x)$, $\alpha_2 = P(\text{rejecting } H_0 \text{ incorrectly})$

$$\alpha_2 = P(Z_{20} > 1.645 \cup Z_{30} > 1.645 \mid \mu = 0)$$
$$\alpha_2 = P(Z_{20} > 1.645 \mid \mu = 0) + P(Z_{30} > 1.645 \mid \mu = 0) - P(Z_{20} > 1.645 \cap Z_{30} > 1.645 \mid \mu = 0)$$

→ Since Z_{20} happens before Z_{30}; use conditional probability to write the $\cap$.

$$\alpha_2 = 0.05 + 0.05 - [1 - P(Z_{30} \leq 1.645 \mid Z_{20} > 1.645, \mu = 0)] \, P(Z_{20} > 1.645 \mid \mu = 0)$$

$$\alpha_2 = 0.1 - [1 - P(Z_{30} \leq 1.645 \mid Z_{20} > 1.645, \mu = 0)] \, 0.05$$
$$\alpha_2 = 0.1 - [1 - 0.7] \, 0.05$$
$$\alpha_2 = 0.085$$

[Under the algorithm $t_2(x)$, Frequentist would have said the result was not significant
but significant under algorithm $t_1(x)$

[On Bayesian inference, we always use the same algorithm called Bayes Rule in which the likelihood function of $x = (x_1, x_2, x_3, \ldots, x_{29}, x_{30})$ is always

$$L_x(\mu) = \prod_{i=1}^{30} e^{-\frac{1}{2}(x_i - \mu)^2}$$ regardless of stopping the experiment

early or not. Stopping early does not affect on the posterior as it only depends on x through the likelihood function.

II e.g. We have data of prostate cancer comparing 52 patients with 50 healthy controls. Each subject got genetic activity measured for a panel of $n = 6033$ genes.

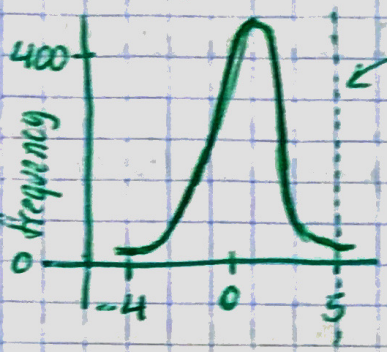
For each gene, the two-sample t-statistic is

$$t_i = \frac{\bar{x}_{ip} - \bar{x}_{ic}}{s^2} \quad \begin{pmatrix} \text{patients' mean for gene } i = \bar{x}_{ip} \\ \text{control's} \quad " \quad = \bar{x}_{ic} \end{pmatrix}$$

Then, (since large sample size) $t_i \sim N(\mu_i, 1)$ $i = 1, \ldots, 6033$

where μ_i is the true effect size for gene i
↳ from F

Let's say: the following histogram



← we are interested in genes that have a very significant healthy activity.

Say gene 610 with $t_i = t_{610} = 5.29$ (which is the largest t_i)

If try to estimate μ_{610}, Is t_{610} OK?
- 610 showed a very large value in this case, and that's the reason we picked it to estimate μ; However if the experiment is conducted again t_{610} might not be the largest
- We biased the estimate of μ by choosing the largest effect size t Since in our sample 610 was particularly high.
  - ↳ ("picking the highest value and take it as significant" is part of the algorithm)
    - ↳ & that creates "selection bias"
      - ↳ 610 might be an overestimate of μ_{610}
        - ↳ selecting genes randomly would ease the overestimation/bias.

Even though gene 610 is individually unbiased for μ_{610}, frequentists would worry that there is an upward bias on $t_{610} = 5.29$.

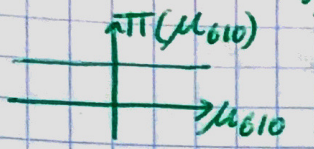
↳ As frequentist assumes this large value was obtained by choice.

↳ It's likely frequentists would recommend a procedure to correct the bias.

On Bayesian Inference would ignore whether t_{610} was picked because it was large.

However Bayesian Inference is sensitive to the choice of the prior

↳ for instance, the choice of a flat prior (aka "I have no idea"), the result might be $\mu_{610} = 5.29$.



flat prior. However a normal prior results in $\mu_{610} = 4.11$

☆ Attention shifts from

choosing an algorithm $t(x)$ in frequentist inference to

choosing a prior $\pi()$ in bayesian inference.

- is much easier to create biases in frequentist inference as any change in the algorithm has an effect on the statistical significance

↳ in Bayesian, the prior may be subjective, but it's the only thing that can be biased.

➢ Bayesian vs. Frequentist.

BAYESIAN	FREQUENTIST (specific question)
1) operates only on one sample with the whole parameter space	1) operates with one parameter in many samples.
2) requires a prior distribution (past experience)	2) replaces the choice of a prior with the choice of a method (algorithm $t(x)$)
	3) is much flexible than Bayes inf. as we can come up with many algorithms.

BAYESIAN	FREQUENTIST
4) Bayesian analyses answers all possible questions at once, (because the posterior is a distribution)	4) usually only computes the expected value and the variance. (each characteristic requires an specific algorithm.)

10 Sep 2019

Computer Age Statistical Inference CHAPTER 04

→ Fisherian Inference and Maximum Likelihood Estimation. (MLE)

▷ For a family of $\mathbb{N}$ probability densities $f_\theta(x)$, where θ is a vector of parameters. the log-likelihood function is defined as:

$$l_x(\theta) = \log\{f_\theta(x)\} \quad \text{for a fixed } x \text{ and variable } \theta.$$

(only one sample)

Get the most likely parameters that would have generated the sample we have.

▷ the MLE is the value of $\theta \in \Omega$ that maximizes the likelihood function.

θ → parameter vector; Ω → parameter space

$$\text{MLE: } \hat{\theta} = \underset{\theta \in \Omega}{\arg\max}\{l_x(\theta)\}$$

- we can also provide MLE estimates for a function $\gamma = T(\theta)$ using $\hat{\gamma} = T\hat{\theta}$. (aka. estimate functions of the true parameter)

- MLE properties

i) it's automatic: likelihood function + data → MLE → $\hat{\theta}^{MLE}$

ii) excellent frequentist properties (good bias & variance)

↳ [next page...]

↳ expected deviation from the true value.

- bias: $\mu - E(\hat{\mu})$

$E(\hat{\mu})$ → expected value of the estimation

μ → true value of the parameter

- unbiased estimater → bias = 0

- variance - is the deviation of the expected value from the true value

$$\text{variance} = \underbrace{\sum_{i=1}^{I}}_{\text{for a normal distribution}} (\hat{\mu}^{(i)} - E(\hat{\mu}))^2 \quad \left\{ \begin{array}{l} I = \text{number of samples} \\ \hat{\mu}^{(i)} = \hat{\mu} \text{ of sample } i \end{array} \right.$$

$$= \underbrace{E_F}_{\text{for any distribution's probability function } F.} \{ (\mu^{(i)} - E(\hat{\mu}))^2 \}$$
for any distribution's expected value for the density

iii) has a reasonable Bayesian justification.

$$P(\theta|x) = C_x \, \pi(\theta) \, e^{\ell_x(\theta)}$$

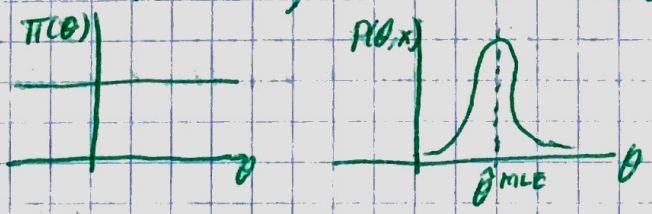
maximum likelihood estimation is the log likelihood function

$P(\theta|x) \rightarrow$ posterior
$C_x \rightarrow$ a constant
$\pi(\theta) \rightarrow$ prior

$\hat{\theta}^{MLE}$ is a maximizer of $P(\theta|x)$ if $\pi(\theta)$ is flat (aka. unknown)

* - If we assume we know nothing about the parameter we are to estimate, then the MLE will be the highest point of the posterior distribution.

- Remmember that in Bayesian we have a posterior distribution of θ.
  ↳ If $\pi(\theta)$ is flat, then $P(\theta|x)$ shall be as follows.



> as the algorithm (described in *) does not change, then the significance level is not affected by unexpected changes in the algorithm (as in frequentist)

e.g. we'll use the Glomecular Filtration Rate data:

- we'll consider 2 potential families

i) Normal: let $\theta = (\mu, \sigma)$, then

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \Rightarrow \text{density function}$$

ii) Gamma (shifted): let $\theta = (\lambda, \sigma, \nu)$, then

$$f_\theta(x) = \frac{(x-\lambda)^{\nu}}{\sigma^{\nu}\,\Gamma(\nu)} e^{-(x-\lambda)/\sigma} \quad \text{for } x \geq \lambda, \text{ 0 otherwise}$$

↳ we are shifting the density distribution with λ.

Since

$$f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i) \Rightarrow \text{The Likelihood Function.}$$

- Under iid sampling, we got.

$$\ell_x(\theta) = \sum_{i=1}^{n} \text{Log}\, f_\theta(x_i) = \sum_{i=1}^{n} \ell_{x_i}(\theta)$$

↳ Log-likelihood function

iid = independent and identically distributed.
↳ identically = every observation comes from the same distr.
independent = the value of a previous observation does not influences the next observation

- For Normal: $\hat{\mu}^{MLE} = \bar{x}$

$$\hat{\sigma}^{MLE} = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^{1/2}$$

But $\hat{\mu}^{MLE}$ and $\hat{\sigma}^{MLE}$ have a closed form solution.

- For Gamma; there is no closed form solution, so it needs to be numerically maximized in the computer. (Jupyter Lecture 7 part 1)

MLE can cause overfitting identification problems in high dimensions.

★ ↳ if we fit a lot of parameters in θ, that fit would become very specific to our current data set. (& would not represent the population)

- However regulized versions of MLE such as Lasso ~~MM~~ are a workaround to this issue.

[redacted]

> Permutation & Randomization

- Previously we conducted a hypothesis test on the activity of gene 136 on 2 groups of leukemia patients.
- We computed a 2-sample t-test to assess the significance of the effect on gene 136 of ALL diagnosis vs. AML diagnosis
  - ↳ whenever we conduct a t-test, we assume that the data samples come from a Normal distribution.
    - ↳ however, as the sample is small, it follows a different distribution
- So Fisher suggested the use of randomization to avoid the Normality assumption
  - ↳ Randomization is: taking groups from the data ~~dist~~ that are of the same size as the tested groups. (In our case $n_1 = 47$ & $n_2 = 25$). Computing the t-statistic for each randomly sampled pair of groups & get their histogram

Utilizing random generated groups, we expect the t values not to be very high, so we can construct an empical distribution of t values

13 Sep 2019

> Fisher Information and the MLE

- We'll go over the univariate case of Fisher Information with one parameter family of densities

$$F = \{ f_\theta(x),\ \theta \in \Omega,\ x \in X \}$$

- $f_\theta(x)$ ↳ parameter density function
- Ω ↳ parameter space
- X ↳ sample space

- will consider the case of continuous random variables.

▶ The Log-likelihood is defined as:

$$l_x(\theta) = \log f_\theta(x)$$

The derivative of $l_x(\theta)$ with respect to θ is the score function. $\dot{l}_x(\theta)$

$$\dot{l}_x(\theta) = \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{\dot{f}_\theta(x)}{f_\theta(x)}$$

↳ How higher or lower the likelihood value of our sample gets as θ varies.

① Let's compute the expectation of the score function.

$$E(x) = \int_x x \underbrace{f(x)}_{\text{density function}} dx$$

$$E[\dot{l}_x(\theta)] = \int_x \dot{l}_x(\theta) f_\theta(x)\, dx = \int_x \dot{f}_\theta(x)\, dx$$

$$= \int_x \frac{\partial}{\partial\theta} f_\theta(x)\, dx$$

if $f_\theta(x)$ is continuous & continuously differenciable,

$$= \frac{\partial}{\partial\theta} \int_x f_\theta(x)\, dx = \frac{\partial}{\partial\theta} \cdot 1 = 0$$

$$\underline{E[\dot{l}_x(\theta)] = 0}$$

② Let's compute the variance of $\dot{l}_x(\theta)$

$$V(x) = \int_x [x - E(x)]^2 f(x)\, dx$$

$$V[\dot{l}_x(\theta)] = \int_x [\dot{l}_x(\theta) - E(\dot{l}_x(\theta))]^2 f_\theta(x)\, dx$$

$$= \int_x [\dot{l}_x(\theta)]^2 f_\theta(x)\, dx$$

The Fisher Information is defined as: the variance of the score function

$$I_\theta \triangleq V[\dot{l}_x(\theta)] = \int_x [\dot{l}_x(\theta)]^2 f_\theta(x)\,dx$$

③ The MLE estimator of $\hat{\theta}$: $\hat{\theta}^{MLE}$

$$\hat{\theta}^{MLE} \overset{approx}{\sim} N\left(\theta, \frac{1}{I_\theta}\right)$$

(θ: mean; $\frac{1}{I_\theta}$: variance)

➢ Proof of ③

let $\ddot{l}_x(\theta) = \frac{\partial^2}{\partial\theta^2} \log f_\theta(x)$

$$= \frac{\partial}{\partial\theta}\left[\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right]$$

$$= \frac{\ddot{f}_\theta(x) f_\theta(x) - \dot{f}_\theta(x)\dot{f}_\theta(x)}{(f_\theta(x))^2}$$

$$= \frac{\ddot{f}_\theta(x)}{f_\theta(x)} - \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right)^2$$

$\ddot{l}_x(\theta)$ has expectation

$$E[\ddot{l}_x(\theta)] = \int_x \frac{\ddot{f}_\theta(x)}{f_\theta(x)} f_\theta(x)\,dx - \int_x \left(\frac{\dot{f}_\theta(x)}{f_\theta(x)}\right)^2 f_\theta(x)\,dx$$

$$= \int_x \ddot{f}_\theta(x)\,dx - \int_x [\dot{l}_x(\theta)]^2 f_\theta(x)\,dx$$

$$= \frac{\partial^2}{\partial\theta^2}\int_x\int_x f_\theta(x)\,dx\,dx - I_\theta$$

$$E[\ddot{l}_x(\theta)] = -I_\theta$$

➢ Now suppose $x = (x_1, x_2, \ldots, x_n)$ is a sample from $f_\theta(x)$