Then $\dot{\ell}_x(\theta) = \sum_{i=1}^{n} \dot{\ell}_{x_i}(\theta)$, as $\dot{\ell}_x(\theta)$ is in Log space.

$$\ddot{\ell}_x(\theta) = \sum_{i=1}^{n} \ddot{\ell}_{x_i}(\theta)$$

Since $\hat{\theta}^{MLE}$ for the full sample x satisfies the maximizing condition

$\dot{\ell}_x(\theta) = 0$ $\longrightarrow$ as we are maximizing the likelihood function, and every maximum has a 1st derivative of 0.

Then, we can get a Taylor Series approximation.

↳ (local linearization)

$$\dot{\ell}_x(\hat{\theta}^{MLE}) = 0 \overset{approx}{\approx} \dot{\ell}_x(\theta) + \ddot{\ell}_x(\theta)(\hat{\theta}^{MLE} - \theta)$$

Solve for $\hat{\theta}^{MLE}$

$$\hat{\theta}^{MLE} = \frac{\theta \ddot{\ell}_x(\theta) - \dot{\ell}_x(\theta)}{\ddot{\ell}_x(\theta)}$$

$$\hat{\theta}^{MLE} = \theta - (\dot{\ell}_x(\theta) / \ddot{\ell}_x(\theta))$$

④

$\dot{\ell}_x(\theta) = \sum_{i=1}^{n} \dot{\ell}_{x_i}(\theta)$ is a sum of a function of random variables.

- Means of random variables are preferable over Sums of random variables.

↳ because means (~~are random variable~~) of very large number of random variables follow asymptotically a Normal distribution.

↳ which is the central limit theorem. CLT

So: by the CLT, $\frac{\dot{\ell}_x(\theta)}{n} \sim N(E[\dot{\ell}_x(\theta)], V[\dot{\ell}_x(\theta)]/n)$

$\sim N(0, I_\theta/n)$ $\quad$ ↯ $\frac{I_\theta}{n}$ = variance

> then $\hat{\theta}^{MLE} = \theta + \frac{\dot{\ell}_x(\theta)/n}{\ddot{\ell}_x(\theta)/n}$

As $n \to \infty$, the numerator $\sim N(0, I_\theta/n)$

and the mean of the function will converge to the expected value

, the denominator

$$\frac{\ddot{\ell}_x(\theta)}{n} \to \frac{E_\theta[\ddot{\ell}_x(\theta)]}{n} = -I_\theta$$

$\lightning$ $V(cx) = c^2 V(x)$

$$\frac{\dot{\ell}_x(\theta)/n}{\ddot{\ell}_x(\theta)/n} \sim N\left(0, \frac{I_\theta/n}{I_\theta^2}\right)$$

$$\sim N\left(0, \frac{1}{n I_\theta}\right)$$

> thus $\hat{\theta}^{MLE} \sim N(\theta, 1/nI_\theta)$ $\lightning$ nI_θ is the fisher Information for n iid observations

---

$$E\left[\theta + \frac{\dot{\ell}_x(\theta)}{\ddot{\ell}_x(\theta)}\right] = E(\theta) + E\left(\frac{\dot{\ell}_x(\theta)}{\ddot{\ell}_x(\theta)}\right)$$

$$= \theta - 0 = \theta$$

So, $\hat{\theta}^{MLE} \sim N\left(\theta, \frac{1}{I_{\theta(n)}}\right)$ > end of Proof ③

17 Sep 2019

▷ Fisher Information Implementation.

to compute the Fisher Information I_θ ...

Let $x_i \overset{ind}{\sim} N(\theta, \sigma^2)$, where σ^2 is known.

→ we compute log-likelihood $\ell_x(\theta)$

density function for a normal distribution $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

normal likelihood function $f_\theta(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$$\Downarrow$$

$$\ell_\theta(x) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

→ the score function

$$\dot{\ell}_\theta(x) = -\frac{1}{2}\frac{1}{\sigma^2} 2(-1)\sum_{i=1}^{n}(x_i-\theta)$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)$$

→ $\ddot{\ell}_\theta(x)$

$$\ddot{\ell}_\theta(x) = \frac{1}{\sigma^2}(-1)\sum_{i=1}^{n} 1 = -\frac{n}{\sigma^2}$$ as $E[\ddot{\ell}_\theta(x)] = -I_\theta$

→ Fisher Information

$$I_\theta = \frac{n}{\sigma^2}$$ $E[\dot{\ell}_x(\theta)] = 0$

→

$$E[\dot{\ell}_x(\theta)] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) = 0$$ such that

$$\sum_{i=1}^{n} x_i = n\theta \Rightarrow \hat{\theta}^{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

→ Finally; for a large enough n,

$$\hat{\theta}^{MLE} \approx N\left(\theta, \frac{1}{I_\theta}\right) \Rightarrow \hat{\theta}^{MLE} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

↳ standard error of the MLE.

→ To finish our discussion about the properties of MLE in a 1-dimensional setting, suppose that

$\tilde{\theta} = t(x)$ $\tilde{\theta}$ is an unbiased estimator computed by some algorithm $t(x)$, based on iid sample $x = (x_1, \dots, x_n)$ from $f_\theta(x)$.

↳ that is: $E_\theta\{t(x)\} = \theta$

▶ Then the Cramér-Rao lower bound says that the variance of $\tilde{\theta}$ exceeds the Fisher Information bound

$$Var_\theta(\tilde{\theta}) \geq \frac{1}{nI_\theta}$$

↳ MLE variance.

- In other words, MLE has variance at least as small as the best unbiased estimate of θ.
- Note that MLE is not unbiased in general, but its bias is small. (in the order of $1/n$) → making the comparison with unbiased estimates appropriate.

▶ Conditional Inference

Say we have an iid sample $x_i \overset{iid}{\sim} N(\theta, 1)$ which has produced estimate $\hat{\theta} = \bar{x}$.

However people conducting the sampling initially disagreed on what the sample size would be. So they flipped a coin to decide:

$$n = \begin{cases} 25 & \text{prob } 1/2 \\ 100 & \text{prob } 1/2 \end{cases}$$ and $n = 25$ was decided

Q: what is $\sigma_{\bar{x}}$?

Classical Frequentist rationale would have resulted in

$$\left[\frac{1}{2}\frac{\sigma^2}{100} + \frac{1}{2}\frac{\sigma^2}{25}\right]^{1/2} = 0.158 = \sigma_{\bar{x}} = sd(\bar{x})$$

↳ variance of $\bar{x}$ with $n=25$

↳ variance of $\bar{x}$ with $n=100$

However Conditional Inference would resulted in

$$\left[\frac{\sigma^2}{25}\right]^{1/2} = 0.2 = \sigma_{\bar{x}}$$

- we have the likelihood function with out the prior
  ↓ based on the observations
- "just take the sample you have"

} it's a mix of Freq. and Bayes thinking

⊃ Fisher's arguments for conditional inference:

i) More relevant inferences
(as the inferences only have to do with what really happend)

ii) Simpler inferences
(no need to assess any correlation between the result and the sample size selection)

e.g. Observed Fisher Information

Rather than using $\hat{\theta}^{MLE} \overset{a}{\sim} N(\theta, 1/nI_\theta)$, Fisher suggested using

where $I(x)$ is the observed Fisher Information

$$\hat{\theta}^{MLE} \overset{a}{\sim} N\left(\theta, \frac{1}{I(x)}\right) \;/\; I(x) \triangleq -\ddot{\ell}_x(\hat{\theta}^{MLE}) = -\frac{\partial^2}{\partial\theta^2}\ell_x(\theta)\Big|_{\hat{\theta}^{MLE}}$$

$E[I(x)] = nI_\theta$, in large samples, the observed Fisher Information is the same as the Fisher Information

↳ However Fisher suggested that in smaller samples $I(x)$ gives a better idea of $\hat{\theta}$'s accuracy.

> We can confirm Fisher's suggestion by sampling from a distribution.

- for instance, let's use the Cauchy distribution.
  ↳ doesn't have an expected value, no variance

1) Compute the log-likelihood
2) get its 1st derivative
3) then the 2nd one
4) get the Fisher Information
5) get the variance of the estimate, even if the distribution doesn't have a variance or an expected value

$$f_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x+\theta)^2} \Rightarrow$$ Cauchi distribution's density function

We are to estimate θ.

→ Compute the log likelihood function $\ell_\theta(x)$

$$\ell_\theta(x) = \log\left(\frac{1}{\pi}\right) + \cancel{\log(1)}^{0} - \log(1+(x+\theta)^2)$$

$$\dot{\ell}_\theta(x) = \frac{2(x-\theta)}{1+(x+\theta)^2}$$

$$\ddot{\ell}_\theta(x) = \frac{-2[1+(x-\theta)^2] - [2(x-\theta)(-1)(+2)(x-\theta)]}{[1+(x-\theta)^2]^2}$$

$$\ddot{\ell}_\theta(x) = \frac{-2[1+(x-\theta)^2] + 4(x-\theta)^2}{[1+(x-\theta)^2]^2}$$

$$I(x) = \ell_x(\hat{\theta}^{MLE})$$

→ 10000 samples of size = 20 drawn with $\theta = 0$
We computed $1/I(x)$ for each sample.

→ Grouped the 10000 $\hat{\theta}^{MLE}$ values according to quantiles of $1/I(x)$ and calculated the empirical variance for each group.
↳ Rough estimate of conditional variance of $\hat{\theta}^{MLE}$ given $1/I(x)$

for All samples, the unconditional variance $1/nI_\theta$ is the same.

↳ Specifically $I_\theta = 1/2$ for a single cauchy observation.

Thus $\frac{1}{20 \cdot (1/2)} = \frac{1}{10} = 0.1$

↳ $\frac{1}{nI_\theta}$ ⇒ all the samples have the same Fisher Information because they are of the same size.

↳ The standard Fisher Information is defined as an expected value.

On the other hand, the observed Fisher Information will varie from sample to sample. Because the $\hat{\theta}^{MLE}$ are different for each sample. (jupyter Lecture 11)

- The observed Fisher Information is related to the variance

20 Sep 2019

➤ Permutation & Randomization

→ permutation (shuffling data)

- recall our dataset where we had data for activity of 7128 genetic markers on Leukemia patients, with 2 different types of diagnosis, AML & ALL.
  ↳ 47 patients
  ↳ 25 patients.

- we first attempted obtaining t-values for the group means of each marker. then we got an histogram of the 7128 t-values.

$t_i = \frac{\bar{X}_{ALL_i} - \bar{X}_{AML_i}}{\hat{se}_i}$ ⟶ assumes that the AML & ALL are distributed Normaly.

- an early conclusion we got was that the empirical distribution of t-values didn't match the t-student with 70 degrees of freedom it was supposed to follow (as it had fatter tails)
(jupyter Lecture 2 part 2)

→ The permutation method we will not need to assume normality.

- one major assumption to use the two-sample t-test that compared our t-scores vs. the t-student with 70 degrees of freedom was that our observations of gene activity come from a normal distribution

- To get rid of the normality assumption, Fisher suggested:
  - to take random permutations (of a single gene's activity from our 72 individuals).
    - pick the 1st 72 permuted individuals as "having ALL" and the next 25 individuals as "having AML"
    - The intention is to make any differences in the mean activity for each gene across the two groups be completely random
    - Take gene 136 & let's compute 10,000 random permutations of the 72 observations of gene 136. (Jupyter lecture 12 part 1)
  - the t-value distribution tells us if that t-value could or couldnt be generated by chance.

- In the classical approach, we compare the t-values of all the 7128 genes against the same t-distribution with 70 degrees of freedom
  ↳ using permutation, we have a distribution for each gene.

→ **Randomization**

Randomly assigning experimental units (individuals) to possible treatment groups.

In the AML/ALL study, there might have existed other reasons (variables) that may influence the diagnosis.

for istance, from an homogeneous group, choose test samples and control samples randomly.

Computer Age Statistical Inference CHAPTER 05

Univarriate Parametric Models

Philosophical differences aside, the three classical approaches we covered relied on low-dimentional (the number of parameters we are to learn / fit) parametric models.

A brief table of useful univariate families

Name, Notation	Density	$\mathcal{X}$ Sample Space	Ω Parameter space	Expectation	Var
Normal $N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	$\mathbb{R}^{(1)}$	$\mu \in \mathbb{R}^{(1)}$ $\sigma^2 \in \mathbb{R}^+$	μ	σ^2

Useful to model quantities that • take positive and negative continuous values • if the distribution is symmetric or • if there are no to many extreme values

Poisson $Poi(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	$\mathbb{N}_0$	$\lambda \in \mathbb{R}^+$	λ	λ

↳ not a density function but a probability mass function, as $Poi(\lambda)$ is discrete (not continuous)

(if the mean grows/shrinks, the variance grows/shrinks proportionaly)

λ must stay positive,
λ is mean interval of time of an exponential distribution, which is continuous ⟶ meaning that the expected number of successes can have decimals.

Use to model a quantity that is discrete, it's the number of counts of something. It's not very flexible, as only has one parameter to tweak.

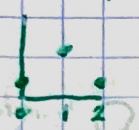
Note, in the Normal we have 2 parameters to tweak, but they are very specific on what they describe. μ is the location parameter and σ^2

Name, Notation	Density	x	Ω	Expectation, Variance
Binomial $Bin(n,\theta)$	$\binom{n}{x}\theta^{n}(1-\theta)^{n-x}$	$\{0,1,\dots,n\}$	$0\le\theta\le 1$	$n\theta$ $n\theta(1-\theta)$

We also count the number of successes, however we know the number of trials (unlike the Poisson)

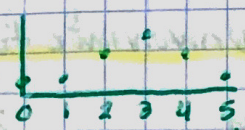
The n is constraining what the x can be, giving some flexibility.
↳ if n is 2 the probability needs to be concentrated between 0 and 2.

↳ $n=2$



(less spread)

$n=5$



(more spread)

Gamma $Ga(v,\sigma)$	$\frac{x^{v-1}e^{-x/\sigma}}{\sigma^{v}\,\Gamma(v)}$	$\mathbb{R}^{+}$	$v>0$ $\sigma>0$	σv $\sigma^{2}v$

$v-1$ shall stay positive & $-x/\sigma$ shall stay negative

The Gamma is used to model positive quantities. It's common to use the inverse Gamma to model variances.

Beta $Be(\alpha,\beta)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$	$0\le x\le 1$	$\alpha>0$ $\beta>0$	$\alpha/(\alpha+\beta)$ $*$

As the x goes from 0 to 1, it's mostly used to talk about probabilities (as a probability distribution)

$$* = \frac{\alpha\beta}{(\alpha+\beta)^{2}(\alpha+\beta+1)}$$

- Both the Gamma & Beta have 2 parameters that convey some degree of flexibility.
- Gamma is flexible but not as flexible as the Beta
- The Binomial with a large n and small probability can approximate the Poisson distribution and viceversa