13 Aug 2019

Computer Age Statistical Inference. CHAPTER 01

* **Statistics:** Science about learning from data.

i) **Inference** - measuring the uncertainty around these estimates

ii) **Algorithm** - following a set of data ~~(processed by)~~ probability steps to produce estimates

e.g. Say you have observations $X_1, \ldots, X_n$

> The mean $\bar{X} = \sum_{i=1}^{n} x_i / n$ summarizes $X_1, \ldots, X_n$ in a number.

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

↳ is the standard error of the mean $\bar{x}$. aka. $\hat{se}$

- σ^2 is the standard deviation of the sample.

* Here, averaging is the algorithm, while the standard error provides an inference of the algorithm's accuracy.

> $\hat{se}$ is itself an algorithm, which could be subject to further inferential analysis.

e.g. Linear Regression

> both linear regression & the mean minimize the error.

↳ however, ~~(linear)~~ any regression is a conditional mean. $\bar{y} = E(y)$ expectation of y

$\hat{y} = E(y \mid x)$ ⟶ conditional mean ↓ unconditional mean.

e.g: $y = \hat{\beta}_0 + \hat{\beta}_1 x$ ⟶ is the regression model

↳ Liver. ↳ Age

> Least-Squares Algorithm: $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

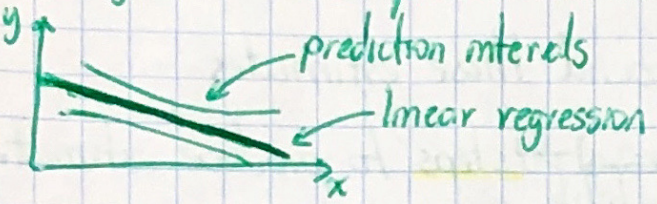
y_i are the actual observed y's

> Inference is:

$$\hat{se}_{LSQ} = \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 / (6(n-2)) \right]^{1/2}$$

→ Since we are estimating two parameters from the data

Q: Why aren't the prediction intervals the same width across the data?



A i) the local variance is not the same
ii) we don't have the same data points at every x^*

e.g. Lowess (aka Local Regression) Algorithm

This method uses a weighting function with the effect that the influence of a neighboring value on the smoothed value at a certain position decreases with their distance to that position.

Step 1: The range of the points to be included is determined. the number n of these points is specified.

- the larger n is, the smoother the adapted curve will be.
- the range shall be exactly n values, including the selected point. and the selected point is in the center of the selected range.
- it's possible that the number of points to the left and right may differ; i.e. the 1st value only has neighboring point to the right.

Step 2: Stablish the weighting for the locally regression smoothing.

$$w(x_k) = \left(1 - \left|\frac{x_i - x_k}{d_i}\right|^3\right)^3 \quad \text{for } k = 1, \ldots, n \qquad \text{(Proximity weighting)}$$

d_i is the distance from x_i to the k^{th} neighboring point
The largest value is attained at point x_i (the focal point)
The value 0 is at the range limits.

Step 3: A linear regression function based on the least squares method is estimated for the LOWESS procedure. The LOESS procedure uses a quadratic function

the parameters are calculated in such a way the the following function is minimized.

LOWESS: $\hat{y}_k = a + bx_k$

LOESS: $\hat{y}_k = a + bx_k + cx_k^2$

Step 4: (optional) It is possible that the estimated regression function is influenced by outliners.

Therefore, robust weightings can be determined.

$$G(x_k) = \begin{cases} \left(1 - \left(\frac{|y_i - \hat{y}_i|}{6 \text{ median } (|y_i - \hat{y}_i|)}\right)^2\right)^2 & \left|\frac{|y_i - \hat{y}_i|}{6 \text{ median } (|y_i - \hat{y}_i|)}\right| < 1 \\ 0 & \left|\frac{|y_i - \hat{y}_i|}{6 \text{ median } (|y_i - \hat{y}_i|)}\right| \geq 1 \end{cases}$$

The robust weighting is 0 if a residual is greater than or equal to $6m$ (m = median residuals). This eliminates the outliners' influence.

The robust weightings are multiplied with the proximity weightings and reused to re-estimate a regression function within the individual ranges.

LOWESS: $\sum_k w(x_k)\, G(x_k)\, (y_k - a - b x_k)^2$

LOESS: $\sum_k w(x_k)\, G(x_k)\, (y_k - a - b x_k - c x_k^2)^2$

Step n: A series of new smoothed values is the result. The procedure can be repeated several times. The higher the number of iterations, the more precise the curve fitting.

> Bootstrapping (Cake sampling with replacement) - INFERENCE (Bootstrap Confidence Intervals)

16 Aug 2019

- Under the bootstrap principle: $\sigma^2_{(\text{sampling w/replacement})} \sim \sigma^2_{(\text{across samples})}$

- let B (usually in the hundreds or thousands) be the number of bootstrap iterations.

- If our original sample is $(x_i, y_i)_{i=1}^N$, at each bootstrap iteration, we'll have:

$(x_{j(b)}, y_{j(b)})_{j \in I}$ for $b = 1, \ldots, B$ & $I = \{1, \ldots, N\}$

j is the index that is randomly sampled from the i indexes.

- For each b, compute $\hat{y}_{j(b)} = E[y_{j(b)} | x_{j(b)}]$
  ↳ using our LOWESS model (or any other model)

- The purpose is to output a table such as the following:

j \ b	1	2	3	...	B
1	$\bar{\hat{y}}_{1(1)}$	$\bar{\hat{y}}_{1(2)}$	$\bar{\hat{y}}_{1(3)}$	...	$\bar{\hat{y}}_{1(B)}$
2	$\bar{\hat{y}}_{2(1)}$	$\bar{\hat{y}}_{2(2)}$	$\bar{\hat{y}}_{2(3)}$	...	$\bar{\hat{y}}_{2(B)}$
⋮	⋮	⋮	⋮	⋮	⋮
N	$\bar{\hat{y}}_{N(1)}$	$\bar{\hat{y}}_{N(2)}$	$\bar{\hat{y}}_{N(3)}$	...	$\bar{\hat{y}}_{N(B)}$

→ we are calculating $\bar{\hat{y}}$ (the mean) since j can be repeated
→ also, due to replacement, j is not continuous.

- for each j in the table, compute $\hat{se}_{boot}$ or σ_j^{boot}

$$\sigma_j^{boot} = \sqrt{\left(\bar{\hat{y}}_j - \bar{\bar{\hat{y}}}_j\right)^2 / (B-1)}$$

} Then compute the confidence intervals assuming a statistic distribution.

OR - then sort by value of $\bar{\hat{y}}_{i(b)}$ from min to max
  ↳ then take the 5th & 95th values to get a 90% confidence interval.
  ↳ sorting does not assumes any stat. distribution, but assumes that the min and max values are indeed the min & max that we can ever get.

> the OLS confidence intervals work asymptotically, that means they assume that the number of available observations is infinite
  ↳ but it assumes normality.
> In LOWESS, n is not infinite, but it doesn't assume any distribution.
> The OLE confidence intervals width are minimized in the mean.

20 Aug 2019

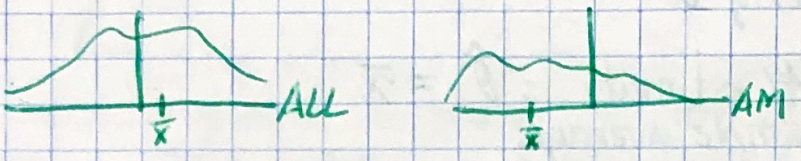
☆ Big Data Implications on HYPOTHESIS TESTING.

e.g. following the example from the book,

72 leukemia patients, where:
47 have ALL (acute lypholastic leukemia) - aka bad leukemia
25 have AML (acute myeloid leukemia) - aka worse leukemia

↳ measured 7128 genes ⟹ find whose genes are relevant to the diagnosis.

→ let's take gene 136.
→ let's say the distribution of gene 136 looks like:



So let's find out if the mean of the two diagnosis is different.

→ the standard classical answer is to use a two-tailed t-test for hypothesis

$H_0: \mu_{ALL} \neq \mu_{AML}$

→ the t-statistic is:

$$t = \frac{\bar{y}^{AML}_{136} - \bar{y}^{ALL}_{136}}{\hat{se}} \quad ; \quad \hat{se} = \hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2} \quad ; \quad n_1 = 47 \ \& \ n_2 = 25$$

if we assume same variances.

The degrees of freedom ν is given by

$$\nu = n_1 + n_2 - 2 = 70$$

Computer Age Statistical Inference CHAPTER 02

- For Frequentist Inference, we assume that the observed sample comes from an probability distribution F

- we call our data vector $x = (x_1, \ldots, x_n)$, which is a realization of a vector of random variables $X = (X_1, \ldots, X_n)$, which are distributed according to F.

X is language to say "a sample" & x is the sample's value

So: F is a probability distribution
X are individual draws of F
x are the very particular values for each of the X's we got from our sample.

➢ We want to learn properties from F, based on x (vector of realizations)
↓
Usually the expectation of a single random draw X_i from F.

→ If a get a number from that distribution, what can I expect that number to be, and how much it deviates (se)

→ let's call the expectation property θ

$\theta = E_F(X_i)$ ⟩ The obvious estimate is $\hat{\theta} = \bar{x}$ equal to the sample average.

[θ = the true expectation value of any draw X_i
$\hat{\theta}$ = is the best guess of what θ is
$\bar{x}$ = is the mean.
$\bar{x}$ is the algorithm; $\hat{\theta}$ is the estimation; θ is what is being estimated

$\hat{\theta}$ is calculated from x following an algorithm $\hat{\theta} = t(x)$

for instance, $t(x) = \Sigma x_i / n = \bar{x}$
↳ t is any algorithm (like sample mean $\bar{x}$ or sample std dev

➢ $\hat{\theta}$ is sample specific and is a realization of $\hat{\Theta} = t(X)$
↓
computing the algorithm to any sample

[$\hat{\theta}$ is a number
$\hat{\Theta}$ is accurate only to a degree which we want to quantify

➢ Define $\mu = E_F(\hat{\Theta})$ the expected value (across samples) of producing an estimate using algorithm $t(x)$ when samples come from distribution F

➢ Bias & variance of estimate $\hat{\theta}$ of parameter θ to be:

$\text{bias} \triangleq \mu - \theta$ & $\text{var} \triangleq E_F\{(\hat{\Theta} - \mu)^2\}$

→ variance: if we follow algorithm $t(\cdot)$, how much are the values of $\hat{\Theta}$ going to vary with respect to the expected value, μ

→ bias: expected value − true value
↳ given an algorithm, is that algorithm estimating the parameter properly.

➢ Bias-variance trade-off:

↳ • the estimator can be very precise, but it can change a lot from sample to sample (no bias, high variance)
• the estimator can be a simple algorithm that does not vary a lot from sample to sample. But are not able to describe all the data points (biased, low variance)

➢ Frequentist usually define quantities (parameters) with infinite sequence of trials.

In other words: Hypothetical datasets $X^{(1)}, X^{(2)}, \ldots$ generate infinite $\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \ldots$ samples

→ So frequentist calculate properties of estimator $\hat{\theta} = t(x)$

↳ However we don't know F...

• Frequentist principles to conduct inference.

i) Plugin principle → $se(\bar{x}) = [\text{var}_F(X)/n]^{1/2}$ / $se(\bar{x}) = \sigma_{\bar{x}}$

↳ $\bar{x} = \sum x_i / n$

However, $\text{var}_F(X) = \hat{\text{var}}_F = \left[\frac{\sum (x_i - \bar{x})^2}{(n-1)}\right]$ without bias

aka. relate the sample standard error with the true variance

23 Aug 2019

Recall: θ is the true parameter of F
$\hat{\theta}$ is the estimate of θ
$\hat{\theta}^t$ is the estimate of θ, using algorithm $t(\cdot)$

If F is $N(\mu, \sigma^2)$ aka a normal distribution
↳ θ would be μ or σ^2
→ for a gamma distribution $\Gamma(\alpha, \beta)$. θ would be α or β.

- The purpose of Frequentist Inference is to asses the inference of the estimates $\hat{\theta}$. So we need the $\hat{se}$ of $\hat{\theta}$
  ↳ However the $\hat{se}$ calculation requires many samples $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots$ but we only have one...

- Plug-in principle: (in human words)
  if we took one x from the distribution F, I can link the variance of that x to the variance of my sample mean.
  → the variance around the mean of a sample $\hat{\theta}$ is an estimator of the variance of F.

$$\hat{var}_F = \left[\frac{\sum (x_i - \bar{x})^2}{n-1}\right] \iff \sigma_{\bar{x}} = se(\bar{x}) = \left[\frac{var_F(X)}{n}\right]^{1/2} \quad \ldots (1)$$

↳ (-1) because we already estimated $\hat{\bar{x}}$ one using the data

- Taylor Series Approximations:
  relate $t(x)$ by local linear approximations.

e.g. if $\hat{\theta} = \bar{x}^2 \longrightarrow \frac{d\hat{\theta}}{d\bar{x}} = 2\bar{x}$

Since $\bar{x}$ is constant, $\sigma_{\bar{x}^2} = 2|\bar{x}|\,\sigma_{\bar{x}}$

To compute the $se(\bar{x}^2)$ of an estimator that is not $\bar{x}$ and the estimator a differential function of the original estimator, we can compute the standard error of the transformed estimator.

In other words: you can get the standard error of $\hat{\theta}$ by taking $d\hat{\theta}/d\bar{x}$ so that:

$$\sigma_{\hat{\theta}} = \sigma_{\bar{x}} \frac{d\hat{\theta}}{d\bar{x}}$$

3 Parametric Families & MLE (maximum likelihood estimate)

e.g. → assume family $N(\mu, \sigma^2)$, where $\sigma^2 = 5$
→ given x, find μ such that the probability of $N(\hat{\mu}, 5)$ is maximized.

where $x = [x_1, x_2, \dots, x_n]$

$$P(x_i | N(\hat{\mu}, 5)) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \Longrightarrow \text{probability to observe } x_i$$

If each element x was sample independently, we can multiply their probabilities.

↳ $P(x_1, x_2 | N(\mu, 5)) = P(x_1 | N(\mu, 5)) \cdot P(x_2 | N(\mu, 5))$

↓

$$P(x | N(\mu, \sigma^2)) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = L(x)$$

↳ The Likelihood Function. ← for normal dist

$L(x)$ is the probability of observe the whole sample x_i.

> MLE → [redacted], such that $L(x)$ is maximized.
Tune $\hat{\theta}$ (in this case μ)

$$\max_{\hat{\theta}} L(x) \Rightarrow \max_{\mu} \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^n e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\cdot 5}} \Rightarrow \hat{\mu}^{MLE}$$

4 Simulation & Bootstrap
Implement a repeated sequence of trials

→ Start with an estimate $\hat{F}$ of F, then simulate values from $\hat{F}$ to get a prior sample $\hat{\theta}^{(b)} = t(X^{*(b)})$ by computing the algorithm $t()$ on each of the bootstrap iterations, for $b = 1, \dots, B$

→ Then the empirical standard deviation of the $\hat{\theta}$'s is the frequentist estimate for $\sigma_{\hat{\theta}}$

Pivotal Statistics.

e.g. say we got two samples $x_1 = [x_{11}, x_{12}, \ldots, x_{1n}]$ and $x_2 = [x_{21}, x_{22}, \ldots, x_{2n}]$.

→ Assume $x_{1i} \overset{ind}{\sim} N(\mu_1, \sigma^2)$ $i = 1, 2, \ldots, n_1$

$x_{2i} \overset{ind}{\sim} N(\mu_2, \sigma^2)$ $i = 1, 2, \ldots, n_2$

→ Test the null hypothesis $H_0: \mu_1 = \mu_2$

$\hat{\theta} = \bar{x}_2 - \bar{x}_1$ ⟶ difference of means test-statistic

$\bar{x}_2 - \bar{x}_1 \sim N(\mu', \sigma'^2)$; where $\mu' = 0$, $\sigma'^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$\sigma^2_{\bar{x}_2} = se(\bar{x}_2) = \frac{\sigma^2}{n_2}$

$\sigma^2_{\bar{x}_1} = \frac{\sigma^2}{n_1}$

* σ^2 is unknown

So, estimate σ^2 using some sample variances.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{(n_1 + n_2 - 2)}$$

→ However, if we have more occurances of this problem, say $x_3, x_4, \ldots$ for each hypothesis test iteration,

$\hat{\sigma}^2$ changes ~~(every time)~~ for each sample x_j

→ The issue. In hypothesis testing we compute a sample statistic, and we compare that value to a critical value on a theoretical distribution

↳ The theoretical distribution will be different for each x_j set of samples

↳ The statistician Student tried to solve that...

t-student

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\hat{sd}}$$

$\hat{sd}$ is the estimation of the standard deviation.

$$\hat{sd} = \hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad \text{... use ①}$$

↳ Using t-student, the estimator does not depend on the other parameters of the distribution

↳ it allows to test $\bar{X}_1 - \bar{X}_2$ (and come with confidence intervals) without the estimation of $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^{2?}$

t-test statistic is a pivotal statistic as it does not depend on other parameters the distribution might have.

★ So. Frequentist use pivotal statistics whenever they are available to conduct stat tests.

27 Aug 2019

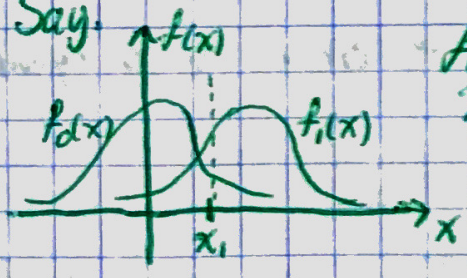
Frequentist Optimality

- Neyman-Pearson lemma provides an optimum hypothesis-testing algorithm.

→ NP lemma states:

1) if we are to decide between 2 possible density functions for observed data x
   → a null hypothesis density $f_0(x)$
   → and an alternative density $f_1(x)$

Say:



for the simple case $x = x_1$

So, with what confidence level can we say x_1 does not belong to f_0 (regeacting H_0)?

Determine the hypothesis test that is powerful at:
regeacting, but not regeacting when it's the appropriate thing.
(no false positives)

ii) A testing rule $t(x)$ says which choice (0 or 1) we'll make given x.

For any decision we make, there will be two associated errors α and β

$\alpha = Pr_{f_0}\{t(x)=1\}$ ⇒ the probability of rejecting the null hypothesis when x belongs to the null hypothesis (incorrectly rejecting f_0)

$\beta = Pr_{f_1}\{t(x)=0\}$ ⇒ choosing f_0 when actually f_1 generated x (probability of incorrectly failing to reject)

iii) Let $L(x)$ be the Likelihood ratio

$$L(x) = \frac{f_1(x)}{f_0(x)}$$

what is the likelihood of the alternative hypothesis $f_1(x)$ giving the data x vs. what is the likelihood of the null hypothesis $f_0(x)$ giving your data x

iv) Testing rule $t_c(x)$ by:

$$t_c(x) = \begin{cases} 1 & \text{if } \ln(L(x)) \geq c' \\ 0 & \text{if } \ln(L(x)) < c' \end{cases}$$

As only rules of the $t_c(x)$ form can be optimal

there is one such rule for each choice of the cutoff c'.

→ that is $\alpha_c < \alpha$ & $\beta_c < \beta$ for some c and where α & β are generated by another rule.

if $t_c(x)$ has an α_c with a c' such that the α error α_c is equal to α, then $\beta_c < \beta$ (and vice versa)

e.g.

We want to get c such that α is equal to Type I error rate and Type II error β is minimized.

$f_0 \sim N(0,1)$ ⚡ $f_1 \sim (N(1/2, 1))$

$f_0(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}}$ ⚡ $f_1(x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-1/2)^2}{2}}$ ⚡ for normal dists density functions.

$$f_0(x) = \left[\frac{1}{\sqrt{2\pi}}\right]^n \prod_{i=1}^{n} e^{\frac{-x_i^2}{2}} = \left[\frac{1}{\sqrt{2\pi}}\right]^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}$$

$$f_1(x) = \left[\frac{1}{\sqrt{2\pi}}\right]^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-1/2)^2}$$

} the likelihood functions

$$L(x) = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-1/2)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}}$$ ⚡ likelihood ratio

$$L(x) = e^{-\frac{1}{2}\left[\sum_{i=1}^{n} x_i + \frac{n}{4}\right]}$$

$$L(x) = e^{-\frac{1}{2}\left[n\bar{x} + \frac{n}{4}\right]}$$

$L(x) > c \Rightarrow e^{-\frac{1}{2}\left[n\bar{x}+\frac{n}{4}\right]} > c$ ⇒ $\bar{x}$ shall be greater than some constant.

$-\frac{1}{2}\left[n\bar{x}+\frac{n}{4}\right] > c$

$n\bar{x} + n/4 > c_2$

$\bar{x} > c_3$ ⟶ only the mean depends on the sample.

» If two likelihood functions come from a normal distribution, then the only thing we need for the hypothesis testing is the sample mean $\bar{x}$

» Most powerful hypothesis test at any Type I error rate α is to compare $\bar{x}$ to a constant.

→ let's relate c, α & β...

$\alpha = P(\bar{x} > c \mid \theta = 0)$ where $\theta = \mu = 0$

↳ probability under $f_0 \sim N(0, \sigma^2)$ where $\sigma^2 = 1$

↳ $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{1}{n} \Rightarrow \sigma_{\bar{x}}^2 = \frac{1}{\sqrt{n}}$

then $\frac{\bar{x}-0}{1/\sqrt{n}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}^2} = \bar{x}\sqrt{n} \sim N(0,1)$

(normalize)

- $\alpha = P(\bar{x} > c \mid \mu = 0)$

$\alpha = P(\bar{x}\sqrt{n} > c\sqrt{n} \mid \mu = 0)$

$\alpha = 1 - P(\bar{x}\sqrt{n} \leq c\sqrt{n} \mid \mu = 0)$ ↯ Φ is the cumulative density function (CDF) of a normal distribution.

$\alpha = 1 - \Phi(c\sqrt{n})$

↳ function for the CDF of a $N(0,1)$ distribution.

$\Phi(c\sqrt{n}) = 1 - \alpha$

$c\sqrt{n} = \Phi^{-1}(1-\alpha)$

$c = \frac{1}{\sqrt{n}} \Phi^{-1}(1-\alpha)$ ↯ In general: $c = \mu_0 + \frac{1}{\sqrt{n}} \Phi^{-1}(1-\alpha)$ for two normal distributions with the same variance

- $1-\beta$: Power of hypothesis test

↳ Probability of correctly rejecting

- A hypothesis test for which $1-\beta$ is maximized (β is minimized) at α level is called a Most Powerful Test (MPT)

On the other hand:

- $\beta = P(\bar{x} \leq c \mid \mu = 1/2)$

↳ fail to reject the null hypothesis when the alternative hypothesis is true.

$\frac{\bar{x} - \frac{1}{2}}{\frac{1}{\sqrt{n}}} \sim N(0,1)$ } normalize to a $N(0,1)$ distribution

$\beta = P[(\bar{x} - 1/2) \cdot \sqrt{n} \leq (c - 1/2) \cdot \sqrt{n} \mid \mu = 1/2]$

$\beta = \Phi((c - 1/2)\sqrt{n})$

Smallest Type II error ratio for hypothesis test with α Type I error rate is given by:

$$\beta = \Phi[(c - \tfrac{1}{2})\sqrt{n}]$$

30 Aug 2019

Computer Age Statistical Inference CHAPTER 03

→ Bayesian Inference

Bayes Rule: $P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B)}$

- like in Frequentist Inference, the fundamental unit of inference is a family of probability densities.

$F = \{ f_\mu(x) ; x \in \mathcal{X}, \mu \in \Omega \}$ — x is a point in the sample space $\mathcal{X}$; μ is an unobserved point in the parameter space Ω

→ we observe x from $f_\mu(x)$ & infer μ.

→ Bayesian Inference also assumes the knowledge of a prior density $g(\mu)$, $\mu \in \Omega$

e.g. if μ is the average height of a room, I would need to state a priori (before observing the sample x)

→ Bayesian works for cases where we don't have a lot of data but we have expert knowledge.

↳ in frequentist, with few data the confidence intervals would be very thick. (wide)

- Bayes Theorem (aka Bayes Rule) is a rule to combine prior knowledge in $g(\mu)$ with current evidence on x.

→ How to consistently update our belief of μ?

Let $g(\mu|x)$ denote the posterior density of μ

↳ after observing the data vector x

Bayes Rule → posterior distribution

$$g(\mu|x) = \frac{g(\mu) \cdot f_\mu(x)}{f(x)}$$ where $f(x)$ is the marginal density of x

$$f(x) = \int_\Omega f_\mu(x) \cdot g(\mu)\, d\mu$$

- $g(\mu)$ → prior distribution
- $f_\mu(x)$ → likelihood that μ is the true parameter given x
- $f(x)$ → probability of having observed the vector (data) x after averaging across all the possible values that μ can take.

Frequentists assumes that there is a true distribution where the data comes from. Also assumes that just like the sample x you got, there are an infinite number of them

In Bayesian, x is fixed (only one sample)
The only thing that changes is the belief of μ

Bayes Rule can be written as:

$$g(\mu|x) = c_x\, L_x(\mu)\, g(\mu) \quad \text{where } f_\mu(x) = L_x(\mu) = f(x|\mu)$$

- c_x → constant that only depends on x
  - ↳ aka. normalization constant of the posterior dist.
    - ↳ as any probability distribution integrates to 1

▷ for any two μ_1, μ_2 values on Ω, the ratio of posterior densities is given by:

$$\frac{g(\mu_1|x)}{g(\mu_2|x)} = \frac{g(\mu_1)}{g(\mu_2)} \cdot \frac{f_{\mu_1}(x)}{f_{\mu_2}(x)} \Rightarrow$$ the posterior odds ratio is the prior odds ratio times the likelihood ratio.

↳ the relative probability (aka. the odds ratio) of μ_1 being the right value of the parameter μ, given the sample x

is equal to the relative odds of the prior belief times the odds ratio of the likelihood function from the observed data x.

e.g. An engineer knows she's having twins.
She asks what's the probability that they'll be identical.
The doctor says:
$\frac{1}{3}$ of twin births are identical } prior belief

Say X is a sonogram result (either same sex or opposite sex)
and same sex is observed.
(identical twins always are of the same sex, while fraternals have 0.5 probability of same or different sex)

- How does the sonogram modify our prior of 1/3 probability on the twins being identical?

D = different twins.
I = identical twins.
S = same sex.
F = fraternal.

$$\frac{g(I,S)}{g(F,S)} = \frac{g(I)}{g(F)} \cdot \frac{f_I(S)}{f_F(S)}$$

$$= \frac{1/3}{1-1/3} \cdot \frac{1}{1/2}$$

1 → if they are identical, what's the prob. of being same sex
1/2 → " fraternal, " same sex

$$= 1$$

↳ g(I,S) & g(F,S) are equaly likely.

» Fraternal & Identical are equaly likely

	S	D	joint probability distribution
I	1/3	0	1/3 = g(I)
F	1/3	1/3	2/3 = 1 − 1/3 = g(F)

e.g. The binomial distribution counts the number of successes or failures in n trials

If a random variable is distributed binomialy

$$X \sim Bin(n, p) \Rightarrow P(X = k \mid p, n)$$

↳ number of successes being k

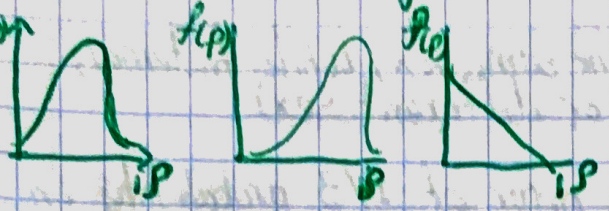
p = probability of success
n = number of observations (trials)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\cdot\Gamma(n-k+1)}$$

$P(X=k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$

- $P(X=k \mid n, p)$: Probability that the # of successes is exactly k
- p^k: observing k successes
- $(1-p)^{n-k}$: observing n-k failures

> let's assume that before observing our data, we belief p to look either of these ways:



$\Rightarrow$ all of these can be modeled using the Beta distributions. (family of)

So let's model our prior distribution on P as:

$$\pi(p \mid \alpha, \beta) \sim \text{Beta}(\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \quad ; 0 \leq p \leq 1$$

(π: prior)

where $B(\alpha, \beta)$ is the Beta function

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$ where Γ is the gamma function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-z}\, dx \quad ; z > 0$$

↳ is the continuous generalization of the factorial function.

$$\Gamma(n+1) = n!$$

$\frac{1}{B(\alpha,\beta)}$ is the normalizing constant

03 Sep 2019

[also: $B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\, dp$

> if $\pi(p \mid \alpha, \beta) \sim \text{Beta}(\alpha, \beta)$, then the mean of the Beta distribution is given by:

$$\underbrace{E(p \mid \alpha, \beta)}_{\text{The mean}} = \frac{\alpha}{\alpha+\beta}$$

For the following analysis is convenient to parameterize our Beta distribution such that the mean can be represented using only one parameter...
So, let.

$\mu = \frac{\alpha}{\alpha+\beta}$; $M = \alpha+\beta$ $\Rightarrow$ new prior $\pi(\rho|\mu,M) = Beta(M\cdot\mu, M(1-\mu))$

↳ under this parametrization the expected value E is.

$E(\rho|\mu,M) = \mu$; $V(\rho|\mu,M) = \mu(1-\mu)/M+1$

↳ expected value of ρ ↳ variance of ρ

▷ Using this new prior, we can proceed to get the posterior distribution.

$P(\rho|k) \propto$ (aka. is proportional to) $\ell(k|\rho)\cdot\pi(\rho|\mu,M)$

↳ posterior dist. our belief about ρ after observing k successes

↳ likelihood function ↳ prior

* → we are applying the Bayes rule, without the normalizing constant $1/B(\alpha,\beta)$ because it doesn't alter the distribution "shape".
- Bayes rule only needs the proportionality.

▷ then $P(\rho|k) \propto (\rho^k(1-\rho)^{n-k})(\rho^{M\mu-1}(1-\rho)^{M(1-\mu)-1})$

↳ without $\binom{n}{k}$ because the same reason as *

$\propto \rho^{(k+M\mu)-1}(1-\rho)^{n-k+M(1-\mu)-1}$

$\propto Beta[k+M\mu, n-k+M(1-\mu)]$ $\Rightarrow$ probability density function

---

▷ if the prior for ρ is Beta & the likelihood is Binomial, then the posterior for ρ is also Binomial.

↳ that is a conjugate prior (aka. the posterior is of the same family as the prior)

▷ the expected value of the posterior ρ given we have observed k success is:

$E(\rho, \frac{k}{n}) = \frac{k+M\mu}{n+M}$

↳ size of the sample

$x = [\underbrace{0,0,1,0,1,\ldots,0,1}_{n}]$

$k = \sum x_i$

k = number of successes

$M \cdot \mu$ = the two prior parameters, which is ~~(equivalent to have observed k)~~ the # of success in the past (before observing the sample)

M = is equivalent to having had a sample size in the past. (aka. the size of the prior sample)

μ is the expectation for p

> we can say our prior assumed having seen $M\mu$ successes in M trials before running the current experiment.

> What is our prediction of the number of successes given our posterior for p?

- Using conditional probability, we can write:

$$m(k|\mu, M, n) = \int_0^1 \ell(k|p) \cdot P(p|\mu, M, k) \cdot dp$$

↓ marginal distribution of the # of successes in n future trials, given parameters μ, M & n

$$= \frac{\Gamma(M)}{\Gamma(M\mu) \cdot \Gamma(M(1-\mu))} \binom{n}{k} \int_0^1 p^{k+M\mu-1} \cdot (1-p)^{n-k+M(1-\mu)-1} dp$$

This is the Beta function

$$= \frac{\Gamma(M)}{\Gamma(M\mu) \cdot \Gamma(M(1-\mu))} \cdot \frac{\Gamma(n+1)}{\Gamma(k+1) \cdot \Gamma(n-k+1)} \cdot \frac{\Gamma(k+M\mu) \cdot \Gamma(n-k+M(1-\mu))}{\Gamma(n+M)}$$

⇒ Probability mass function

= the probability distribution for the number of successes you are expecting to observe.

= is the posterior predicted distribution as the random variable (number of successes k) is discrete.

successes

$\alpha = k - M\mu$

$\beta = n-k + M(1-\mu)$

failures

> look for "Conjugate Prior" tables for more examples.

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 at 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 N/1 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.

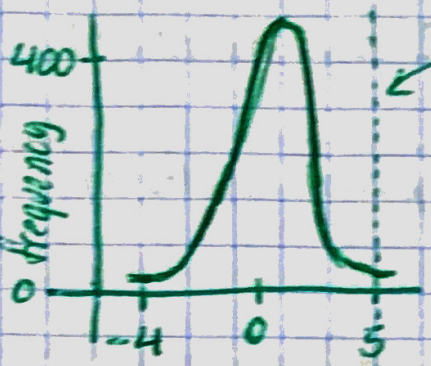
For each gene, the two-sample t-statistic is

$$t_i = \frac{\bar{x}_{ip} - \bar{x}_{ic}}{s_i^2} \quad \left(\begin{array}{l}\text{patients' mean for gene } i = \bar{x}_{ip} \\ \text{control's} \quad\quad\;\; " \quad\quad\;\; = \bar{x}_{ic}\end{array}\right)$$

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 9 largest
- We biased the estimate of μ_{610} 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 $t_{610} = 5.29$.

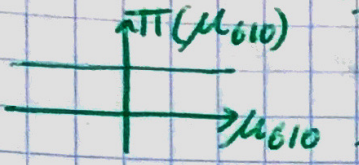
↳ As frequentist assumes this large value was obtained by choice.

↳ 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 in 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 posible 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 ℓℓ probability densities $f_\theta(x)$, / where θ is a vector of parameters. the log-likelihood function is defined as:

$\ell_x(\theta) = \log\{f_\theta(x)\}$ for a fixed x and variable θ.
(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.
↳ parameter space
↳ parameter vector

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

- 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 → MLE ← data → $\hat{\theta}^{MLE}$

ii) excellent frequentist properties (good bias & variance)
↳ [next page...]

- bias: $\mu - E(\hat{\mu})$
  - $E(\hat{\mu})$: expected value of the estimation
  - μ: true value of the parameter
- unbiased estimater → bias = 0

↳ expected deviation from the true value.

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

$$\text{Variance} = \underbrace{\sum_{i=1}^{I}}_{\text{for a normal distribution}} (\hat{\mu}^{(i)} - E(\hat{\mu}))^2 \quad \begin{matrix} I = \text{number of samples} \\ \hat{\mu}^{(i)} = \hat{\mu} \text{ of sample } i \end{matrix}$$

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

iii) has a reasonable Bayesian justification.

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

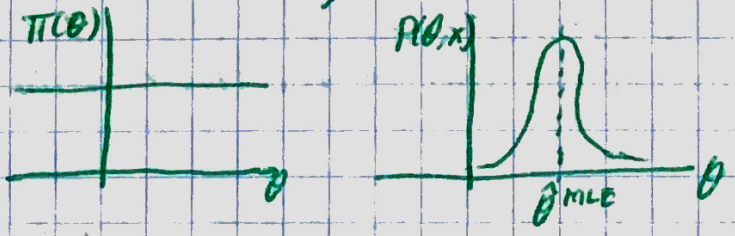
maximum likelihood estimation is the log likelihood function

$P(\theta|x) \rightarrow$ 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

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

↳ 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.

$$l_x(\theta) = \sum_{i=1}^{n} \text{Log } f_\theta(x_i) = \sum_{i=1}^{n} l_{x_i}(\theta)$$

↳ 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 on 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 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 that are random 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 valve 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 differenuable,

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

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

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 $\hat{\theta}$: $\hat{\theta}^{MLE}$

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

(mean; 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 $f_\theta(x)$

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

$\dot{\ell}_x(\theta) = 0$ ⟶ 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 $\hat{\theta}^{MLE}$

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

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

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 variables~~) of very large number of random variables follow asymptotically a Normal distribution.
    ↳ which is the central limit theorem. CLT

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

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

---

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

$$\overset{a}{=} \theta - 0 = \theta$$

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

$$\Downarrow$$

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

→ the score function

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

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

→ $\ddot{\ell}_{\underset{\sim}{\theta}}(\underset{\sim}{x})$

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

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

→

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

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

→ 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 estimator computed by some algorithm $t(x)$, unbiased based on iid sample $x = (x_1, \ldots, x_n)$ 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

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

↳ 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:

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

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

Classical Frequentist rationale would have resulted in

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

- $\frac{\sigma^2}{100}$: variance of $\bar{x}$ with $n = 100$
- $\frac{\sigma^2}{25}$: variance of $\bar{x}$ with $n = 25$

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

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

$E[I(x)] = nI_\theta$, 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 it's 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

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

We are to estimate θ.

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

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

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

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

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

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

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

↳ $\frac{1}{nI_\theta} \Rightarrow$ 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 couldn't 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

Univariate 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	Variance
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.

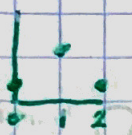
Use to model a quantity that is discrete, it's the number of counts of something. It's not very flexible, as only has one parameter to tweak.

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 $Bin(n,\theta)$	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$\{0,1,\dots,n\}$	$0\leq\theta\leq 1$	$n\theta$ $n\theta(1-\theta)$

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

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

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