13 Aug 2019

Computer Age Statistical Inference. CHAPTER 01

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

i) Inference - measuring the uncertanty around these estimates

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

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

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

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

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)~~ ang 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: $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

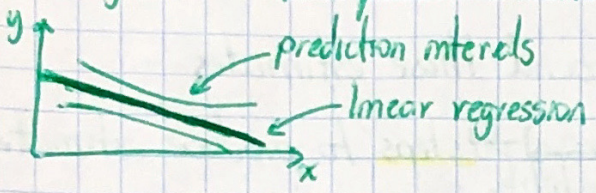
y_i are the actual observed y's

> 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 dont 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 possition 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.

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

d is the distance from x_i to the kth neighboring point
The largest value is attained at point x_i (the local 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: $\hat{y}_k = a + b x_k$

LOESS: $\hat{y}_k = a + b x_k + c x_k^2$

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 (aka 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 $\hat{y}_{j(b)} = E[y_{j(b)} \mid x_{j(b)}]$
  ↳ 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 doesnt assume any distribution.

> The OLE confidence intervals width are minimized in the mean.

20 Aug 2019

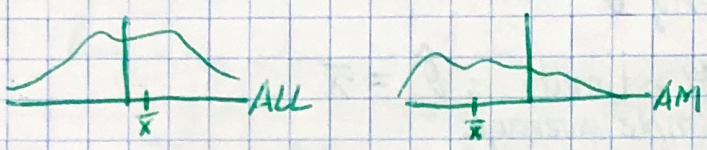
☆ Big Data Implications on HYPOTHESIS TESTING:

e.g. following the example from the book,

72 leukemia patients, where:
47 have ALL (acute lymphoblastic 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:

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

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 $x = (x_1, \dots, x_n)$, which is a realization of a vector of random variables $X = (X_1, \dots, X_n)$, 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:

$$\text{bias} \stackrel{d}{=} \mu - \theta \quad \& \quad \text{var} \stackrel{d}{=} E_F\{(\hat{\Theta} - \mu)^2\}$$

→ variance: if we follow algorithm $t()$, 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 → $se(\bar{x}) = [var_F(X)/n]^{1/2}$ $\quad$ / $se(\bar{x}) = \sigma_{\bar{x}}$
↳ $\bar{x} = \sum x_i / n$

However, $var_F(x) = \hat{var}_F = \left[\frac{\sum (x_i - \bar{x})^2}{(n-1)}\right]$ 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 Θ is ~~the~~ 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 one $\bar{x}$ using the data

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

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

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,

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

we can compute the estandard 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 ~~the~~:

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

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

$$P(x_1 | N(\hat{\mu}, 5)) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \Longrightarrow \text{probability to observe } x_1$$

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

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

$$\downarrow$$

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

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

↳ The Likelihood Function.

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

for normal dist

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

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

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 $\hat{\theta}^{(k)} = t(X^{(b)})$ 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 $\mu^* = 0$, $\sigma^{*2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$$\left[\begin{array}{l} \sigma^2_{\bar{x}_2} = se(\bar{x}_2) = \frac{\sigma^2}{n_2} \\ \sigma^2_{\bar{x}_1} = \frac{\sigma^2}{n_1} \end{array}\right.$$

* 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...