Extending the Basic Bootstrap (II)

Ex: Ten subjects were randomly assigned to each of two treatments. Find

(a) A bootstrap SE for $\bar{X}_1 - \bar{X}_2$.

(b) A bootstrap 95% CI for $\mu_1 - \mu_2$

(c) A bootstrap 95% CI for σ_1^2 / σ_2^2.

How would we do this?

Note: There are multiple possible approaches that one might use. A key thing for success of the bootstrap is to mimic the original sampling as much as possible.

Two-Sample Example

Treatment	Values				
1	9	12	12	14	17
	19	21	22	26	31
2	8	9	10	11	13
	13	19	21	22	24

Two methods based on Kernels

https://www.youtube.com/watch?v=x5zLaWT5KPs
第一个网址一开始就说的不错啊
https://en.wikipedia.org/wiki/Kernel_density_estimation

① **Kernel density estimation** — Allows us to estimate the [probability density function] for a distribution w/o 不用 assuming a parametric form for the density.

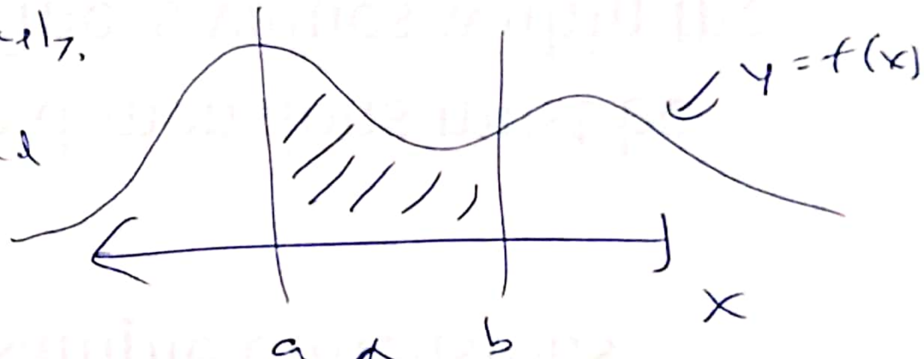
pdf

② **Kernel regression** — Allows us to estimate the [expected value] of y as a function of x w/o assuming a particular form for the regression curve. (It doesn't have to be linear, quadratic, or any other particular form.)

ex: linear relationship 不用

The Probability Density Function pdf

Recall: The pdf $f(x)$ is the function that, for a continuous distribution indicates which values are likely and which are unlikely. (It integrates to 1) and it completely determines the distribution, as does the distribution function.



$P(a < X < b) = \int_a^b f(x)\,dx$

One estimate: Create a histogram with equal-sized bins and standardize it so that it integrates to 1.

Some problems with this: Not smooth, sensitive to the choice of bins for the histogram.

Parametric Density Estimation 有参法

Q: How would we estimate a probability density within a particular parametric family?

A: We would estimate the parameters & then use the corresponding density as our estimate.

For example, given $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, we could compute $\bar{X}$ & s (sufficient statistics) and then use the $N(\bar{X}, s^2)$ pdf as our estimate. (Such estimators would be consistent as long as the parameter estimators are consistent.) (for familiar families of distributions)

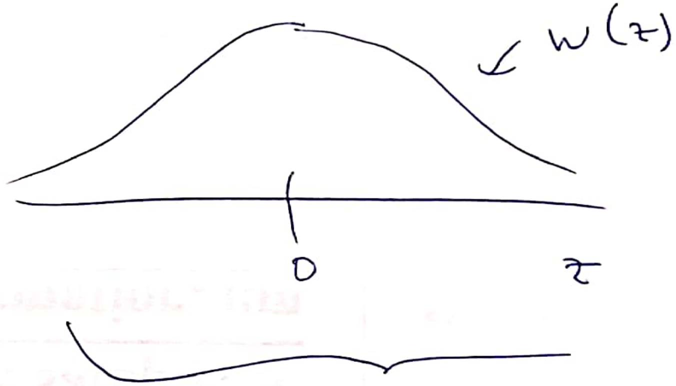
Kernel Density Estimation ← nonparametric! 非参

Let $X_1, X_2, \dots, X_n$ be the data, a random sample from the cts distribution with pdf $f(x)$.

Let $w(z)$ be a kernel, a ① symmetric density function ② with mean 0 ③ and standard deviation 1.

We then estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\Delta} w\left(\frac{x - X_i}{\Delta}\right),$$



One possible kernel.

note: where $\Delta > 0$ is the bandwidth

Note: The book uses h for the bandwidth, but h looks too much like n to work well in these notes.

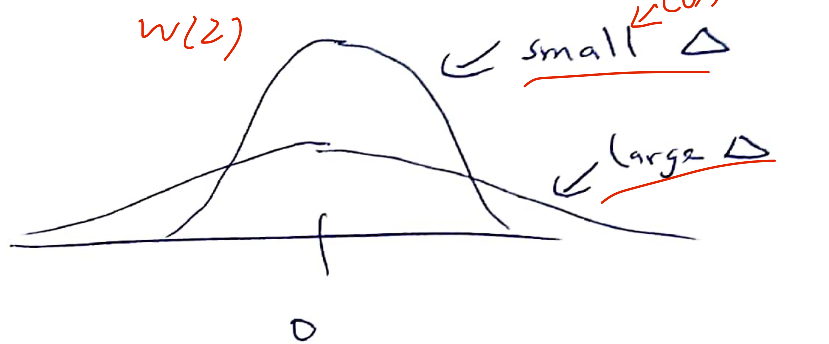
∴ 后面都把 Δ 作为 bandwidths

The Bandwidth Δ

If $w(z)$ has mean 0 and standard deviation 1,

then $\frac{1}{\Delta} w\left(\frac{z}{\Delta}\right)$ has mean 0 and st. dev. Δ.

Big $\Delta \Longrightarrow$ more $\boxed{\text{spread}}$ out.



kernal density estimator

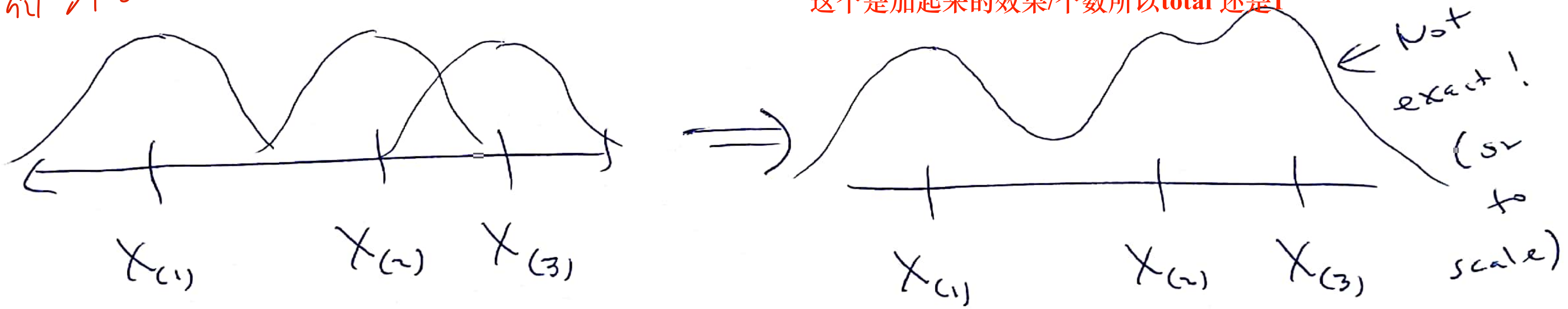
Interpreting $\hat{f}(x)$:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\Delta} w\left(\frac{x - x_i}{\Delta}\right):$$

We average together n pdfs, one centered at each data value.

假设分布于 data

这个是加起来的效果/个数所以total 还是1



Properties of $\hat{f}(x)$: kernel density

① $\hat{f}(x)$ is a valid density. In particular, it is non-negative and integrates to 1.

Q: Why? A: Note that $\frac{1}{\Delta} w\left(\frac{x - x_i}{\Delta}\right)$ integrates to 1 for each i.

② $\hat{f}(x)$ is smooth if $w(\cdot)$ is smooth. ⟸ w need not be smooth, though!

③ If we let $\Delta \to 0$ sufficiently slowly as $n \to \infty$, then $\hat{f}(x)$ is a consistent estimator for $f(x)$ at a particular point.

Scanned with CamScanner

Choosing the Bandwidth

Bigger Δ leads to more averaging, smaller Δ to a more jagged 锯齿 estimate.

Research indicates that Δ should go like $\frac{1}{n^{1/5}}$ to minimize the MSE for $\hat{f}(x)$.

One simple choice: (too simple?)

$$\Delta = \frac{1.06\, s}{n^{1/5}} \leftarrow \text{sample st. dev.}$$

More sophisticated: We could choose Δ using leave-one-out cross-validation. We could even let Δ be different for different x values.

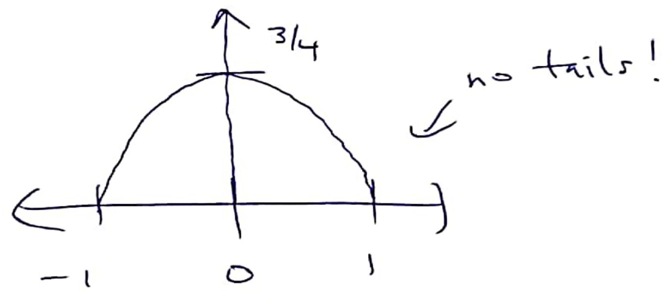
Some Kernel Choices

① Gaussian: Here the standard normal pdf is $w(z)$.

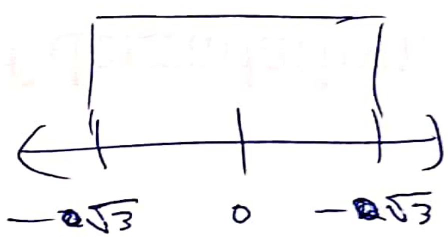
② Epanechnikov: The kernel $w(z)$ is a standardized version of the function $K(x) = \frac{3(1-x^2)}{4}$, $-1 \le x \le 1$.

总结来说
相对最好
好像

asymptotically optimal in one sense

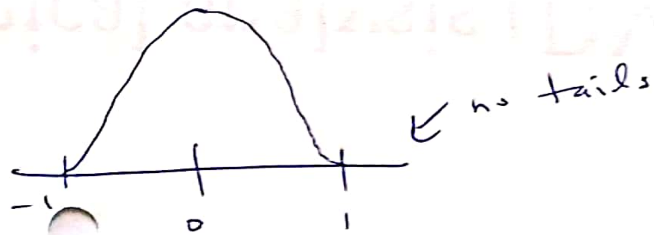


③ Rectangular: Here $w(z)$ is the Uniform$(-\sqrt{3}, \sqrt{3})$ pdf.



④ Biweight: Here $w(z)$ is a standardized version of

$$K(x) = \frac{15}{16}(1-x^2)^2, \quad -1 \le x \le 1.$$



Baseball Data Example

Try out the density function in R w/ different kernel, and different bandwidths.

large：越smooth
small:越锯齿状

Q: What happens when the bandwidth becomes either very large or very small?

Q: Do you see any limitations for the kernel estimators that we are discussing?

In particular, do the estimators have any undesirable properties?

Kernel Regression (I)

Recall: The regression of y on X is the function $E[Y \mid X = x]$, the conditional expected value.

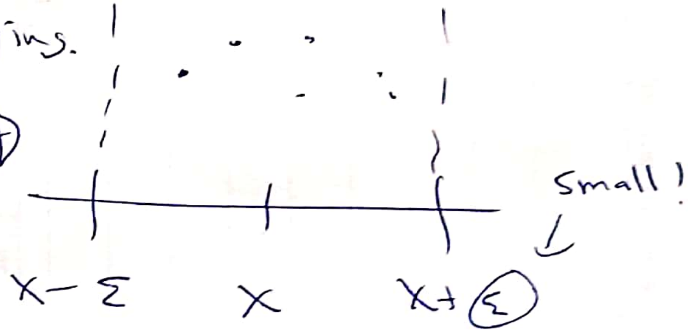
If the relationship is linear, then we get a regression line, but the relationship need not be linear.

One can add higher-order terms like X^2, X^3, etc., but that also gives only a limited amount of flexibility.

One idea: If $E[Y \mid X = x]$ is continuous then it is roughly constant over a small interval of x values.

This suggests doing some averaging.

Average these Y values to estimate $E[Y \mid X = x]$.



Kernel Regression (II)

One problem with this crude averaging is that $\widehat{E[Y|X=x]}$ won't be continuous in x at places where one point drops out of the interval or joins the interval.

就是说中间断了一下

Better idea: Use smoothly - changing weights based on a kernel, say $w(z)$.

Kernel regression estimate: $$\widehat{E[Y|X=x]} = \frac{\sum_{i=1}^{n} Y_i \, w\left(\frac{x-x_i}{\Delta}\right)}{\sum_{i=1}^{n} w\left(\frac{x-x_i}{\Delta}\right)}.$$

weight for Y_i in the average

where $\Delta > 0$ is the bandwidth.

Looking at the Weights

x_1 x_2 x_3

Suppose that we have X values 1, 3, 8 and that we use a Gaussian kernel with $\Delta = 2$.

Using R, plot the weights for Y_1, Y_2, + Y_3 as a function of x for x between -5 and 15.

Q: What will happen to the estimates as x becomes very small or very large?

Baseball Example:

Use kernel regression to estimate $E(Y|X=x)$ when X = at bats and Y = batting average (success rate).

Try out different choices of the bandwidth.

Does it appear that the relationship is linear or not?

Note on Bandwidth in R

In <u>density</u>, bw is the standard deviation of the kernel to be used.

In <u>ksmooth</u>, bw is set up so that the quartiles of the kernel are ± 0.25 bw. This means that the standard deviation σ satisfies

$$bw \cong 2.7\,\sigma.$$

<u>Why?</u> $0.25\,bw \cong 0.674\,\sigma$

$$\Rightarrow bw \cong 2.7\,\sigma.$$

Tests for Contingency Tables (I)

Recall: The chi-square test is used to test for association in a two-way table.

有多
的话

Test statistic: $X^2 = \sum_{\text{cells}} \dfrac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$. If the expected counts are all large, then X^2 is approximately distributed $\chi^2((r-1)\cdot(c-1))$, where r = # of rows and c = # of columns.

Q: How large must the expected counts be?

A: Cochran suggests that all be at least 1.0, with no more than 20% less than 5.0.

Tests for Contingency Tables (II)

Q: What do we do if the expected counts are too small for the chi-square approximation to work well?

A: We can do an exact permutation-based version of the test.

Another possibility: Combining cells which is likely to lead to loss of power. It is also not always obvious which cells should be combined.

We condition on the row + column totals and find the permutation distribution of the test statistic.

The possibilities are not all equally likely!

An Example

For the table $\begin{pmatrix} 4 & 0 \\ 1 & 4 \end{pmatrix}$, list out all the possible tables & find their probabilities under the null hypothesis of independence between the variables.

__Solution:__

$$\begin{array}{ccc} \begin{pmatrix} 4 & 0 \\ 1 & 4 \end{pmatrix} & \begin{matrix} 4 \\ 5 \end{matrix} \\ \begin{matrix} 5 & \ \ 4 \end{matrix} & \end{array} \quad \begin{pmatrix} 3 & 1 \\ 2 & 3 \end{pmatrix} \quad \begin{pmatrix} 2 & 2 \\ 3 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ 4 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 4 \\ 5 & 0 \end{pmatrix}$$

$$\frac{\binom{4}{4}\binom{5}{1}}{\binom{9}{5}} = \frac{5}{126} \qquad \frac{\binom{4}{3}\binom{5}{2}}{\binom{9}{5}} = \frac{40}{126} \qquad \frac{\binom{4}{2}\binom{5}{3}}{\binom{9}{5}} = \frac{60}{126} \qquad \frac{\binom{4}{1}\binom{5}{4}}{\binom{9}{5}} = \frac{20}{126} \qquad \frac{\binom{4}{0}\binom{5}{5}}{\binom{9}{5}} = \frac{1}{126}$$

加起来 $=1$

Now find X^2 for each table, and do an upper-tailed test (like the chi-square test).

Another Example

List out all the possibilities and probabilities for the table

$$\begin{array}{cccc} \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} & & & \begin{array}{c} 3 \\ 2 \end{array} \\ 3 \quad\; 1 \quad\; 1 & & & \end{array}$$

<u>Solution:</u>

$$\begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 0 \end{pmatrix}$$

$$\underbrace{\frac{\binom{3}{3}\binom{2}{0}\binom{2}{1}}{\binom{5}{3}\binom{2}{1}} = \frac{2}{20} \qquad \frac{\binom{3}{2}\binom{2}{1}\binom{1}{1}\binom{1}{1}}{\binom{5}{3}\binom{2}{1}} = \frac{6}{20} \qquad \frac{\binom{3}{2}\binom{2}{1}\binom{1}{1}}{\binom{5}{3}\binom{2}{1}} = \frac{6}{20} \qquad \frac{\binom{3}{1}\binom{2}{2}\binom{2}{1}}{\binom{5}{3}\binom{2}{1}} = \frac{6}{20}}_{}$$

These must sum to 1!

Fisher's Exact Test

If the contingency table is 2×2, it is also possible to test independence (or equality of proportions) using Fisher's exact test.

The listing is done exactly as for the chi-square test, and the test statistic is the value in the (1,1) cell or some equivalent test statistic.

Here the test can be one-sided, unlike for the chi-square test.

Ex: In a game against South Florida in 2010, Corey Fisher made 1 three-point shot and missed 3, while Corey Stokes made 3 and missed 3. Find the p-values for (a) a two-sided test and (b) a one-sided test where the alternative is that Stokes succeeds more often.

Fisher's Exact Test Example (I)

Observed table:

$$
\begin{array}{cccc}
 & \text{made} & \text{missed} & \\
\text{Fisher} & 1 & 3 & 4 \\
\text{Stokes} & 3 & 3 & 6 \\
 & 4 & 6 &
\end{array}
$$

Possibilities:

$$
\begin{pmatrix} 0 & 4 \\ 4 & 2 \end{pmatrix} \quad
\begin{pmatrix} 1 & 3 \\ 3 & 3 \end{pmatrix} \quad
\begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix} \quad
\begin{pmatrix} 3 & 1 \\ 1 & 5 \end{pmatrix} \quad
\begin{pmatrix} 4 & 0 \\ 0 & 6 \end{pmatrix}
$$

$$
\frac{\binom{4}{0}\binom{6}{4}}{\binom{10}{4}} \quad
\frac{\binom{4}{1}\binom{6}{3}}{\binom{10}{4}} \quad
\frac{\binom{4}{2}\binom{6}{2}}{\binom{10}{4}} \quad
\frac{\binom{4}{3}\binom{6}{1}}{\binom{10}{4}} \quad
\frac{\binom{4}{4}\binom{6}{0}}{\binom{10}{4}}
$$

210

$$
= \frac{15}{210} \quad = \frac{80}{210} \quad = \frac{90}{210} \quad = \frac{24}{210} \quad = \frac{1}{210}
$$

probability of a table as likely or less likely than the observed table

(a) add all but $\frac{90}{210}$ to get $\frac{15+80+24+1}{210} \approx \boxed{0.57}$

(b) add the observed prob. and those to the left

to get $\frac{80+15}{210} \approx \boxed{0.45}$

Fisher's Exact Test Example (II)

Verify our results for this example in R.

<u>Note:</u> The way we computed the two-tailed p-value is valid in general, but in symmetric situations, it reduces to doubling the shorter of the two one-tailed p-values.