This time: bootstrap, MCMC

next time: MCMC

read: JS ch. 12

first draft of take-home final is available at course web site (absolute dead date Fri 24 Mar 2017 5pm) CA time

AMS 132 14 Mar 17

①

please complete online course evaluation form for AMS 132

The bootstrap

Brad Efron (1979)

(1938-)

So far we have always assumed a parametric sampling distribution for the data. Ex. Kaiser ICU case study

$$(i=1,\dots,n)$$

$$(Y_i \mid \theta, B) \overset{IID}{\sim} \text{Bernoulli}(\theta)$$

Ex. NB10 case study

$$(i=1,\dots,n)$$

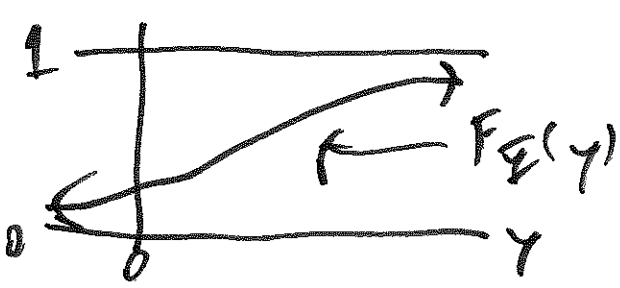
$$(Y_i \mid \mu, \sigma, r, B) \overset{IID}{\sim} t_r(\mu, \sigma^2)$$

Q: Is it possible to draw valid inferences without assuming a specific parametric form for the sampling distribution? A: Yes, in some cases.

This part of statistics is called <u>nonparametric</u> inference, & it has both frequentist & Bayesian versions. Efron developed the <u>bootstrap</u> as one frequentist nonparametric approach. Math Fact: Every probability distribution is uniquely characterized by its <u>cumulative distribution function</u> (CDF) F.

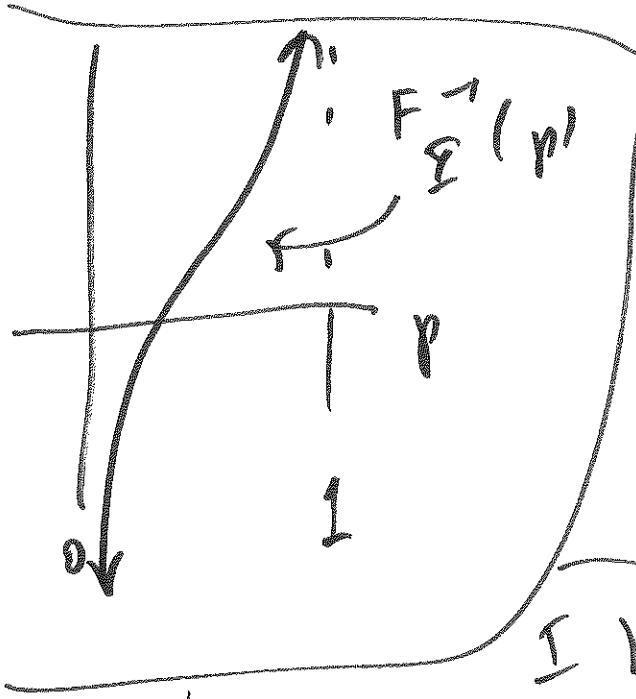
Def. (Ars 131) Y a random variable (RV) taking values on the real number line $\mathbb{R}$. The CDF of Y is defined to be $F_Y(y) = P(Y \leq y)$.

CDFs of continuous RVs look like this:

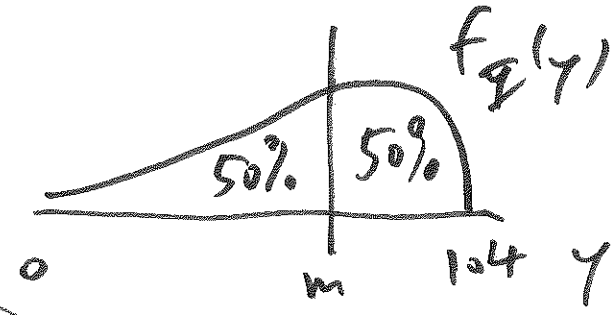


$0 \leq F_Y(y) \leq 1$ and $F_Y(y)$ is nondecreasing in y

Setting $F_Y(y) = p$, if F_Y is continuous it will have an inverse: $y = F_Y^{-1}(p)$



Ex. $m = F_Y^{-1}(\frac{1}{2})$ is the median of the distribution of Y:



I have

Case study: A random sample of $n = 101$ final exam scores from the population of {all UCLA students who took introductory chemistry in 2013.} = p

Q: Can we estimate the population median m & build a 95% confidence interval for m, without assuming a specific parametric form for F?

A: yes

There are actually two frequentist nonparametric methods to solve this problem, one specific to the <u>median</u> (or some other <u>percentile</u>/<u>quantile</u>) and one that's <u>much</u> more general.

Specific method:

Suppose the population CDF F is continuous.

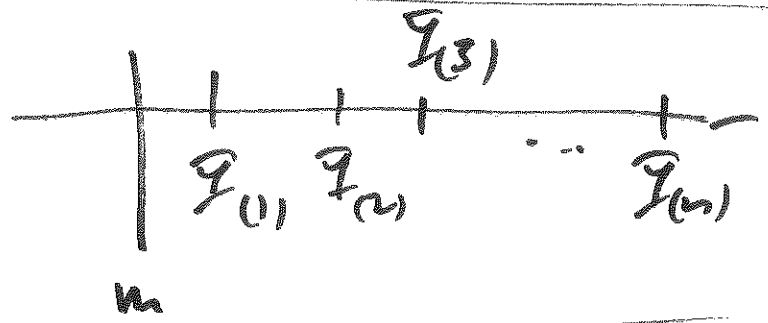
Take your random sample $Y_1, \ldots, Y_n$ from F and <u>sort</u> it from smallest to largest: let's call the sorted data set

$$\underbrace{Y_{(1)}}_{\substack{\text{minimum} \\ \text{of } \underset{\sim}{Y} = (Y_1, \ldots, Y_n)}}, \underbrace{Y_{(2)}, \ldots, Y_{(i)}, \ldots,}_{\text{order statistics}} \underbrace{Y_{(n)}}_{\text{maximum of } \underset{\sim}{Y}}$$

Q: What's the probability that any given data value Y_i falls below or above the pop. median m?

A: $P(Y_i < m) = P(Y_i > m) = \frac{1}{2}$

Q: What's the probability that $Y_{(1)}$ falls above the median?



Let $S = \#$ of $Y_i > m$

Then S follows the Binomial distribution with parameters n and $\frac{1}{2}$, so

$$\boxed{S \sim \text{Binomial}(n, \tfrac{1}{2})}$$

$$P(Y_{(1)} > m) = P(S = n) = \binom{n}{n} \frac{1}{2}^n \frac{1}{2}^0 = \left(\frac{1}{2}\right)^n$$

Similarly, $P(Y_{(1)} < m < Y_{(2)})$

$$= P(S = n-1) = \binom{n}{n-1}\left(\frac{1}{2}\right)^{n-1}\left(\frac{1}{2}\right)^1 = n\left(\frac{1}{2}\right)^n$$

$$= P(S = 1)$$

So you can find a $100(1-\alpha)\%$ CI for m just by finding the two integers ℓ and h (low) (high) such that

$$P(Y_{(\ell)} < m) = \frac{\alpha}{2}$$

and

$$P(Y_{(h)} > m) = \frac{\alpha}{2}$$

UCLA case study

R computes that

qbinom(.025, 101, 0.5) = 41

qbinom(.975, 101, 0.5) = 60

So $P(Y_{(41)} < m < Y_{(60)}) = 95\%$

$Y_{(41)} \to 77$, $Y_{(60)} \to 83$ test scores

As our point estimate of m we will of course use the sample median: $\hat{m} = Y_{(50)} = 81$, with 95% CI for m $(77, 83)$ (asymmetric about $\hat{m}$ ✓)

General method | **The bootstrap**

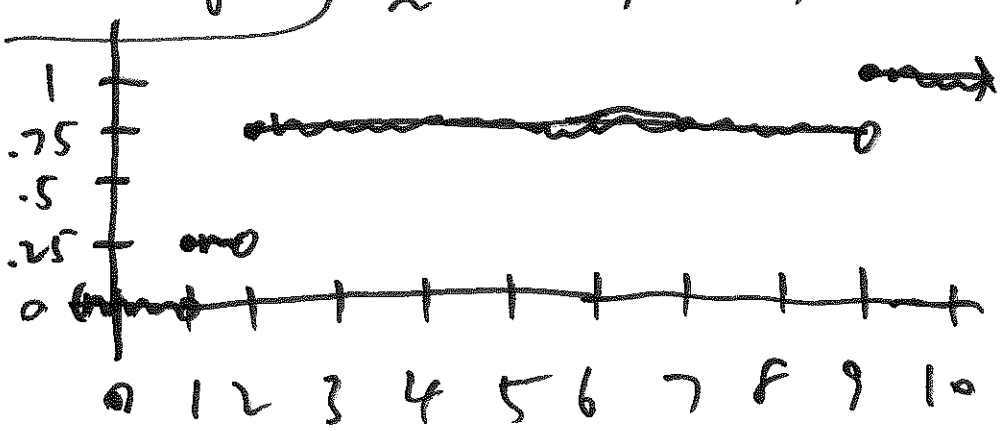
Our nonparametric sampling model is $(Y_i \mid F) \overset{IID}{\sim} F$ $(i=1,\dots,n)$ ← cont.

The population quantity of interest is $m = F^{-1}(\frac{1}{2})$ = population median

Math fact | The best frequentist estimate of F on the basis of $\underset{\sim}{Y} = (Y_1, \dots, Y_n)$ is the <u>empirical CDF</u> $\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y)$

$= \left(\begin{array}{c}\text{\% of data} \\ \text{values} \le y\end{array}\right)$

Example | $\underset{\sim}{y} = (1, 2, 2, 9)$



$\hat{F}_n$ is a step function with jumps at the data values.

P

UCLA chem. students

sample

the observed students

repeated sampling (RS)

data set

final exam score

$N =$ big

actual IID

final exam score

$n = 101$

median $\hat{m} = 81$

median $m = ?$

CDF $F = ?$

hypothetical IID

ECDF $\hat{F}_n$

$n = 101$

median $\hat{m} = ?$ (ex. 80)

$\vdots$

81
80
$\vdots$

n

?

long run dist. of $\hat{m}$

This was our basic frequentist diagram earlier in the course

Efron had an interesting idea: why not pretend

the sample = the population and take repeated samples from the sample?

"$\mathcal{P}$" = sample

the observed students

bootstrap samples

RS bootstrap data set

final exam score

$n = 101$

IID

y_{11}^* ⋮ y_{n1}^* $n = 101$

median $\hat{m}_1^*$

median $\hat{m} = 81$

CDF = ECDF

$\hat{F}_n$

IID

y_{12}^* ⋮ y_{n2}^* $n = 101$

median $\hat{m}_2^*$

⋮

80
79
82
⋮
1

m

dist. of $\hat{m}^*$

2.5%

77 83

approx. 95% bootstrap CI for m

This approach to building bootstrap CIs is called the <u>percentile</u> method; fancier small-n refinements exist but we don't have time to look at them

It has recently been shown (Draper, 2016) (also see Hjort (1986)) that the bootstrap is actually a special case of a Bayesian nonparametric method based on Dirichlet process priors on CDFs.

Back to MCMC

A single Bayesian model comparison method

DIC = Deviance Information Criterion (Spiegelhalter et al. (2002))

DIC is one of a number of "information criteria" (AIC, BIC, ...) that trade off (model goodness of fit) (I) against (model complexity) (II) – there is a tug of war between (I) and (II)

You can make a model fit the data arbitrarily well by increasing model complexity, but the resulting model won't do a good job of <u>predicting</u> new data because you have <u>overfit</u> it to the current data set – DIC tries to resolve this by of way

$$DIC = \bar{D} + p_D$$

$\bar{D}$ is an estimate of the <u>deviance</u> of your current model, defined to be

$-2 \log(\text{likelihood})$

we want the log likelihood (ll) to be <u>big</u> (the model fits the data well), so we want the deviance ($-2\,ll$) to be small.

p_D is an estimate of the <u>effective number of parameters</u> in your current model: $\boxed{\text{ex.}}$ $(Y_i \mid \mu, \sigma^2, B) \overset{IID}{\sim} N(\mu, \sigma^2)$

$p_D = 2$ for this model (μ, σ^2)

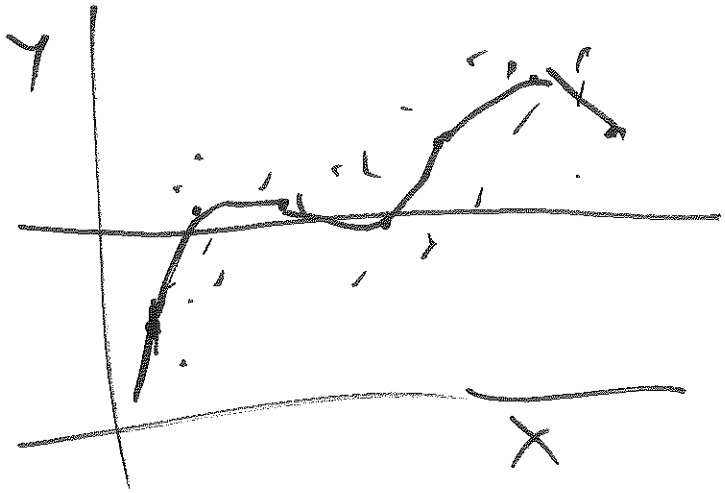
---

we want p_D to be <u>small</u> to avoid needlessly complicated models

But you can't have $\bar{D}$ and p_D simultaneously small because increasing model complexity (if you do it right) will <u>increase</u> the log(likelihood) & <u>decrease</u> the deviance

$$\boxed{DIC = \bar{D} + p_D}$$

DIC is a compromise between $\bar{D}$ and p_D.



M_2 is better than M_1 if

$DIC(M_2) < DIC(M_1)$

How much smaller? See DIC FAQ page on course website.