Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Lecture 9

Tip + Paper

Tip Work on problems you genuinely enjoy thinking about/are passionate about. A lot of statistics (and science) is long periods of concentrated effort with no guarantee of success at the end. To be a really good statistician requires a lot of patience and effort. It is a lot easier to work hard on something you like or feel strongly about.

Paper of the Day:
"Short -term and long-term health risks of nuclear power plant accidents'
http://www.nejm.org/doi/full/10.1056/NEJMra1103676?query=featured_home

Outline For Today

- Bootstrapping
- Bootstrapping regression models

The bootstrap - the 30,000 foot view

Bootstrapping is a computational procedure for:

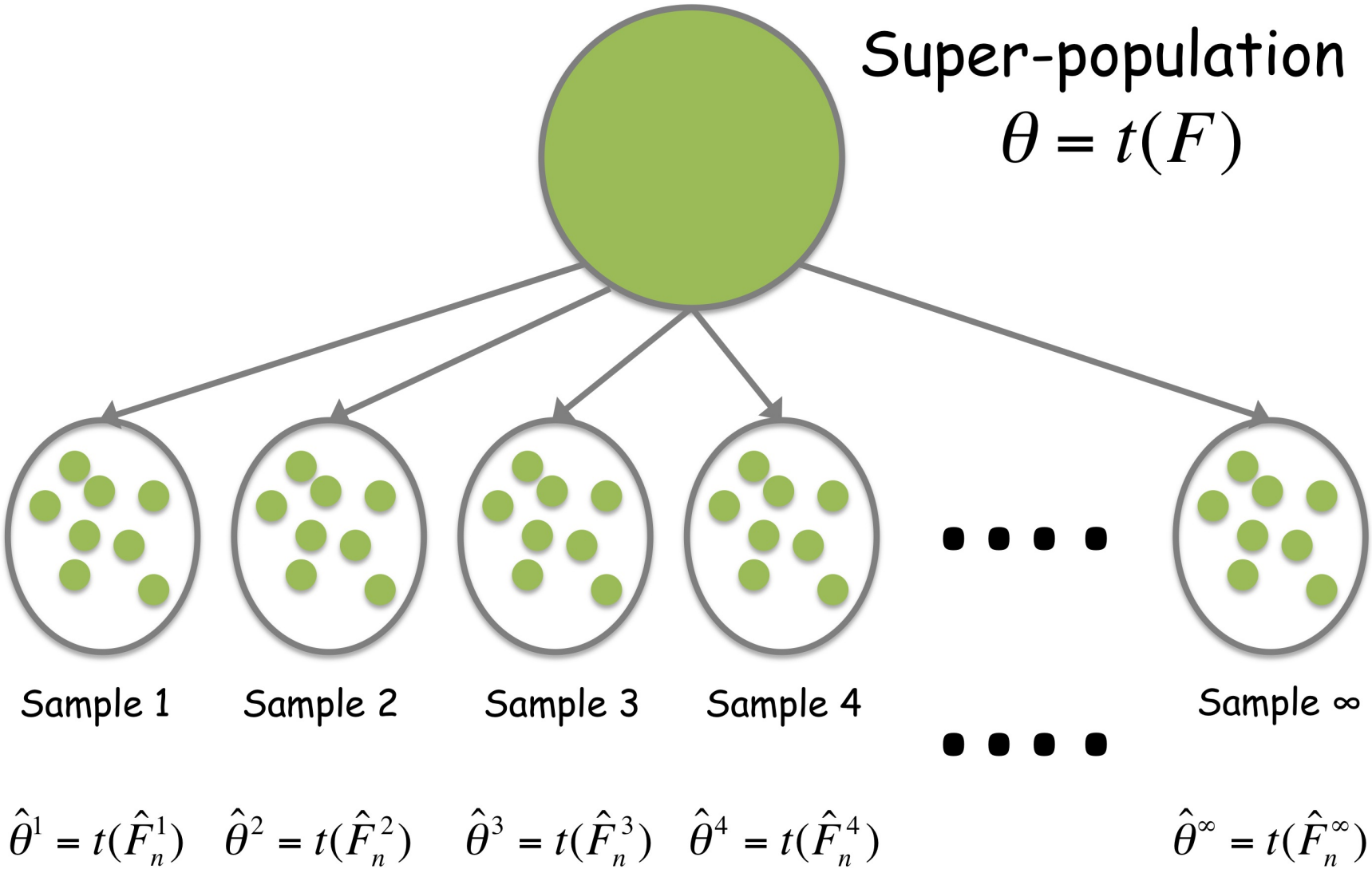
- Calculating standard errors
- Forming confidence intervals
- Performing hypothesis tests
- Improving predictors (called bagging)

The basic idea behind bootstrapping is to treat the sample of data you observe (or some appropriate function of the data) as if it was the super-population you were sampling from. Then you sample from the observed set of values to try to approximate the sampling variation in the whole population.

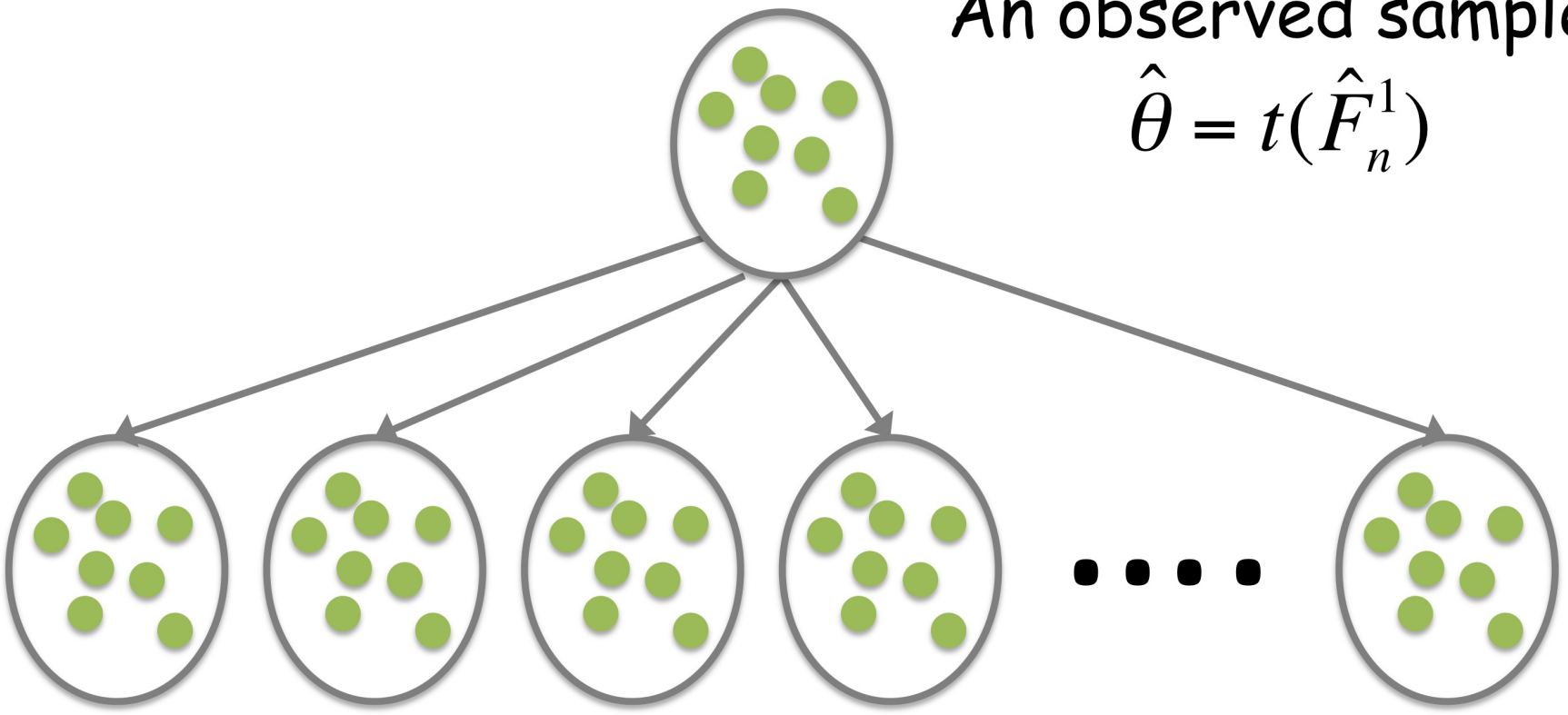
The idea of the bootstrap was originally proposed by Efron in 1979 http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176344552

Related ideas are very old by the standards of statistics (Quenouille, 1956 Notes on bias in estimation. Tukey, 1958 Bias and confidence in not-quite large samples)

The Frequentist "Central Dogma"



The Bootstrap "Central Dogma"



$\hat{\theta}^{1*} = t(\hat{F}_n^{1*})$ $\hat{\theta}^{2*} = t(\hat{F}_n^{2*})$ $\hat{\theta}^{3*} = t(\hat{F}_n^{3*})$ $\hat{\theta}^{4*} = t(\hat{F}_n^{4*})$ $\hat{\theta}^{\infty*} = t(\hat{F}_n^{\infty*})$

The ECDF & the plug-in principle

The **plug-in** principle states that if we have a parameter $\theta = t(F)$, then we estimate the parameter by applying the same functional to an estimate of the distribution function $\hat{\theta} = t(\hat{F}_n)$. Although other estimates can also be used

The default $\hat{F}_n = \mathbb{F}_n$ is the empirical distribution

$$\mathbb{F}_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1(Y_i \leq y)$$

A sample Y_i^* from $\mathbb{F}_n$ has the property that $Y_i^* = Y_j$ with probability $1/n$ for $1 \leq j \leq n$

Why $\mathbb{F}_n$?

- Glivenko-Cantelli ($||\mathbb{F}_n - F||_\infty = \sup_{y \in \mathbb{R}} |\mathbb{F}_n(y) - F(y)| \rightarrow_{a.s.} 0$)
- $\mathbb{F}_n$ is a maximum likelihood estimate (see e.g. http://www.cs.huji.ac.il/~shashua/papers/class3-ML-MaxEnt.pdf)
- It is sane

An example of the plug-in principle/the bootstrap †

Suppose $Y_i > 0$ are *i.i.d.* from F. We might be interested in:

$$\theta = \mathbb{E}_F \log(Y)$$

- the log of the geometric mean of F. A plug in estimate of θ is:

$$\hat{\theta} = \mathbb{E}_{\mathbb{F}_n} \log(Y^*)$$

which is actually available in closed form:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \log(Y_i)$$

We never specified F parametrically, hence this is a model-agnostic estimate and robustness can be expected, in large samples. (Also note $e^{\hat{\theta}}$ is the sample geometric mean)

Another example of the plug-in principle/the bootstrap

Let Y be a binary variable and suppose

$$\theta(F) = \Pr(\mathrm{Y} = 1) = \mathbb{E}_{\mathrm{F}}[1(\mathrm{Y_i} = 1)]$$

We can find the estimate of $\theta(F)$ using the plug-in principle:

$$\hat{\theta} = \mathbb{E}_{\mathbb{F}_n}[1(Y_i^* = 1)] = \bar{Y}$$

Suppose we wanted an estimate for the variance

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\bar{Y}) = \mathrm{Var}(Y_i)/n$$

We could use the plug in estimator

$$\mathrm{Var}_{\mathbb{F}_n}(Y_i)/n = \mathbb{E}_{\mathbb{F}_n}[(Y_i^* - \bar{Y})^2]/n = \frac{\bar{Y}(1-\bar{Y})}{n}$$

The algebra for $\mathrm{Var}_{\mathbb{F}_n}(Y_i)$

$$
\begin{aligned}
\mathrm{Var}_{\mathbb{F}_n}(Y_i)/n &= \mathbb{E}_{\mathbb{F}_n}[(Y_i^* - \bar{Y})^2]/n \\
&= \sum_{j=1}^{n} \frac{1}{n}(Y_j - \bar{Y})^2 \\
&= \frac{1}{n}\left[(1-\bar{Y})^2 \sum_{j=1}^{n} Y_j + \bar{Y}^2 \sum_{j=1}^{n}(1-Y_j)\right] \\
&= \frac{1}{n}\left[\sum_{j=1}^{n} Y_j - 2\bar{Y}\sum_{j=1}^{n} Y_j + n\bar{Y}^2\right] \\
&= \bar{Y}^2 - \bar{Y} = \bar{Y}(1-\bar{Y})
\end{aligned}
$$

When evaluating $\mathbb{E}_{\mathbb{F}_n}$, $\bar{Y}$ is a "parameter" and treated as fixed.

Bootstrap - no closed form

Usually, no closed form evaluation will exist (we sort of "got lucky" in the previous examples). How did we get lucky? The plug-in estimate ended up being an expectation of a single random variable Y_i^*.

What if we were unlucky and the plug in estimate was a function of all of the values $Y_1^*, \ldots, Y_n^*$?

For example, suppose we wanted to estimate the variance of the sample median $\hat{\theta} = \mathbb{F}_n^{-1}(1/2)$?

Bootstrap - no closed form

In this case, the variance $\mathrm{Var}_{\mathbb{F}_n}(\hat{\theta})$ is an expectation of a function of $Y_1^*, \ldots, Y_n^*$.

It isn't clear there is a pretty formula in this case, but if we let X_j denote the number of times Y_j occurs in a bootstrap sample, then: $(X_1, \ldots, X_n) \sim Mult(n; 1/n, \ldots, 1/n)$ so:

$$\sum_{Y^* \in \mathcal{S}} \left\{ \mathbb{F}_n^{*-1}(1/2) - \mathbb{E}_{\mathbb{F}_n}[\mathbb{F}_n^{-1}(1/2)] \right\}^2 \frac{n!}{\prod_{i=1}^n x_i!} (1/n)^n$$

where $\mathbb{F}_n^{*-1}(1/2)$ is the sample median for each bootstrap sample and $\mathcal{S}$ is the set of all unique bootstrap samples from $Y_1, \ldots, Y_n$.

There are $\binom{2n-1}{n}$ unique bootstrap samples.

For $n = 10$ there are 92,378 unique values. For $n = 25$ there are 63.2 trillion or so.

Bootstrap - Monte Carlo calculations

Most of the time, you'll use the bootstrap for parameters where a closed form doesn't necessarily exist.

Instead we use a **Monte Carlo** approach.

For the variance of the sample median you would:

1. Select B independent bootstrap samples Y^{*b} from $\mathbb{F}_n$.
2. Recalculate the statistic for each sample $\hat{\theta}^{*b} = \mathbb{F}_n^{*b-1}(1/2)$
3. Approximate $\sum_{Y^* \in \mathcal{S}} \left\{ \mathbb{F}_n^{*-1}(1/2) - \mathbb{E}_{\mathbb{F}_n}[\mathbb{F}_n^{-1}(1/2)] \right\}^2 \frac{n!}{\prod_{i=1}^n x_i!}(1/n)^n$ by

$$\frac{1}{B} \sum_{b=1}^{B} \left\{ \mathbb{F}_n^{*b-1}(1/2) - \bar{\mathbb{F}}_n^{*-1}(1/2) \right\}^2$$

where $\bar{\mathbb{F}}_n^{*-1}(1/2) = \frac{1}{B} \sum_{i=1}^{b} \mathbb{F}_n^{*b-1}(1/2)$

As $b \to \infty$ you get closer and closer to the exact or "ideal bootstrap".

Bootstrap - Monte Carlo standard error

The general form for calculating bootstrap standard errors is similar:

1. Select B independent bootstrap samples Y^{*b} from $\mathbb{F}_n$.
2. Recalculate the statistic for each sample $\hat{\theta}^{*b}$
3. Approximate $\mathbb{E}_{\mathbb{F}_n}[(\hat{\theta*} - \hat{\theta})^2]$ by

$$\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\hat{\theta}}^{*})$$

where $\bar{\hat{\theta}}^{*} = \frac{1}{B}\sum_{i=1}^{b}\hat{\theta}^{*b}$

Bootstrap - Monte Carlo coverage estimates †

For a confidence interval $a(Y), b(Y)$ based on a sample of size n, we may want to estimate coverage

$$\mathbb{E}_F 1_{[a<\theta<b]} = \Pr(\mathrm{a} < \theta < \mathrm{b}|\mathrm{F})$$

. The bootstrap estimate is

$$\mathbb{E}_{\mathbb{F}_n} 1_{[a<\hat{\theta}<b]} = \Pr(\mathrm{a} < \hat{\theta} < \mathrm{b}|\mathbb{F}_\mathrm{n})$$

The Monte-Carlo version of this calculation is:

1. Select B independent bootstrap samples Y^{*b} from $\mathbb{F}_n$
2. Approximate the coverage by:

$$\hat{\mathrm{Cover}}\mathrm{age}_{BOOT} = \frac{1}{B}\sum_{b=1}^{B} 1_{[\ell(Y^{*b})<\hat{\theta}<u(Y^{*b})]}$$

Note this means you need two steps of calculation

Bootstrap: getting intervals †

Think about these in very frequentist terms:

- Calculate 2.5%, 97.5% quantiles from many $\hat{\theta}_n(Y^*)$. This bootstrap percentile interval does contain $\hat{\theta}_n$ in 95% of the replicates under $\mathbb{F}_n$ and hence for large n should contain the true θ in $\approx 95\%$ replicates under F.
- Calculate bootstrap mean, variance of $\hat{\theta}_n$. For large n asymptotic normality means $\mathbb{E}_{\mathbb{F}_n}\hat{\theta}_n \pm 1.96 \times \sqrt{\mathrm{Var}_{\mathbb{F}_n}(\hat{\theta})_n}$ will contain $\theta_n(\mathbb{F}_n)$ in 95% of replicates from $\mathbb{F}_n$; hence approximately 95% under F.

Moments are more stable than quantiles; smaller B may be okay.

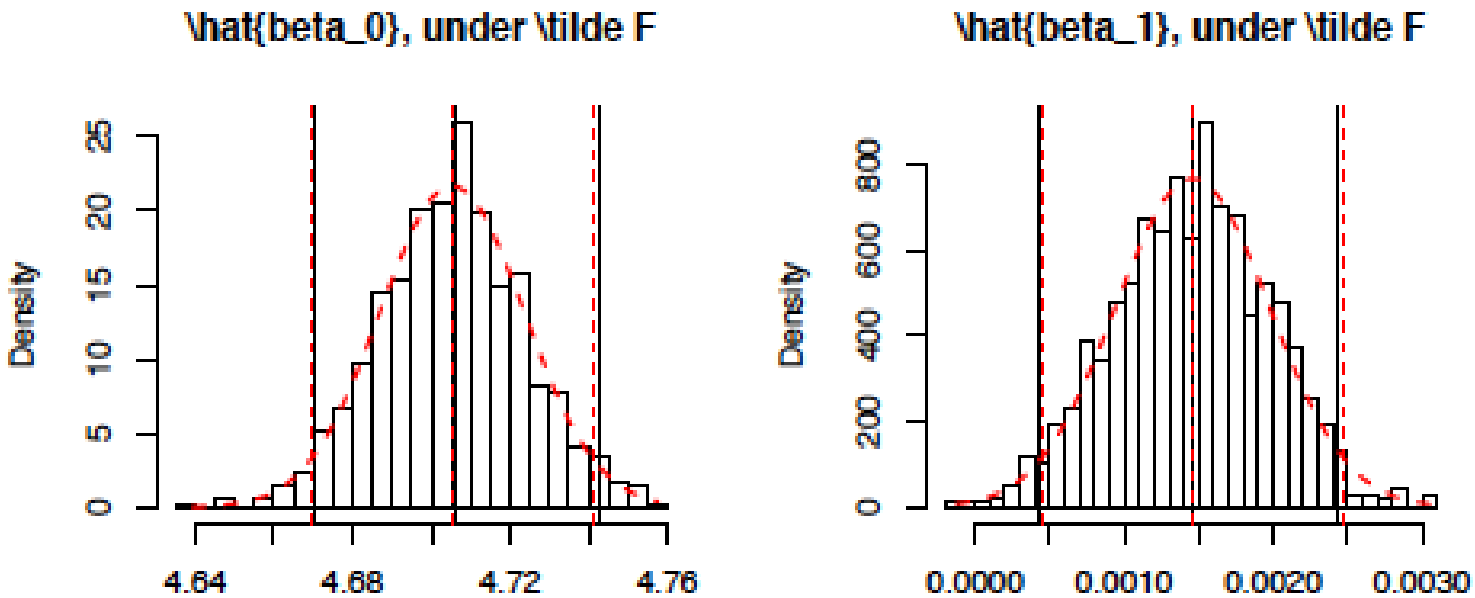
Draw a histogram of your $\hat{\theta}(Y^*)$ to informally check if asymptotic normality under $\mathbb{F}_n$ is reasonable. (If it's not, the quantile method is not guaranteed either!)

Bootstrap: non-parametric version, in R†

```
> n <- dim(airpol3)[1] #335 data points
> do.one <- function( ){
resample.rows <- sample(1:n, replace=T)
newdata <- airpol3[resample.rows,]
glm(Deaths∼ PM10, data=newdata, family=poisson)$coeff
}
> bigB <- 1000
> set.seed(4) #dont forget this!
> thetahat.star <- replicate(bigB, do.one( ) )
>
> apply(thetahat.star, 1, mean)
(Intercept) PM10
4.705230113 0.001458154
> apply(thetahat.star, 1, sd)
(Intercept) PM10
0.0184041313 0.0005198361
> apply(thetahat.star, 1, function(z){quantile(z, c(0.025, 0.975))})
(Intercept) PM10
2.5% 4.669952 0.0004297996

97.5% 4.742523 0.0024328609
```

Bootstrap: non-parametric version, in R†



- All the $\hat{\beta}$ from our B replicates of $\hat{F}$. Red/dashed lines indicate the moment approach, black/solid lines are the quantile-based estimates
- There is no interesting difference between them (here)

Bootstrap: important caveat†

We resampled Y **and** X; whole rows of the dataset

- This is effectively treating Y and X as random quantities, which may not be how the experiment was set up.
- Random X tends to lead to wider confidence intervals; inference is (typically) **conservative**, if X was actually fixed
- Remember the empirical average in the "default sandwich also treats X as random, in the approximation for $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$; "default" bootstrapping (a.k.a. resampling cases) is in general very similar
- In practice other choices often matter more (e.g. which $\hat{F}$).

Bootstrap & Sandwich[1]

The score contribution for the i^{th} subject for β_j is $U_{ij} = \frac{\partial \ell_i}{\partial \beta_j}$ so $\sum_i \frac{\partial \ell_i}{\partial \beta} = U_\beta^T \cdot \mathbf{1}$ where $\mathbf{1}$ is an $n \times 1$ vector of ones. Note that $\hat{\beta}$ satisfies $U_\beta^T \cdot \mathbf{1} = 0$.

Let $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = -A$ The score evaluated at $\beta = \hat{\beta}$ in the bootstrap sample is: $U_{\hat{\beta}}^T \cdot W \neq 0$ where W is a vector of random weights.

The one step approximation to $\hat{\beta}_W$ the ML estimate in the bootstrap sample, starting from $\hat{\beta}$ is given by

$$(\hat{\beta}^*_W - \hat{\beta}) \approx (\hat{\beta}^*_{one} - \hat{\beta}) \approx \hat{\mathbf{A}}^{-1} U_{\hat{\beta}}^T W$$

Now $\mathbb{E}[W] = \mathbf{1}$ and $\mathrm{Var}(W) = I - \frac{1}{n}\mathbf{1} \cdot \mathbf{1}^T$

[1] Slides modeled on lecture notes by N. Breslow at UW

Bootstrap & Sandwich[2]

Thus the approximate bootstrap mean is:

$$\mathbb{E}_W(\hat{\beta}^*_{one} - \hat{\beta}) \approx \hat{\mathbf{A}}^{-1} U_\beta^T \mathbb{E}(W) = \mathbf{0}$$

and the approximate variance is:

$$\begin{aligned} \mathbb{E}_W(\hat{\beta}^*_{one} - \hat{\beta})(\hat{\beta}^*_{one} - \hat{\beta})^T &\approx \hat{\mathbf{A}}^{-1} U_{\hat{\beta}}^T \mathrm{Var} W U_{\hat{\beta}} \hat{\mathbf{A}}^{-1} \\ &= \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{-1} \end{aligned}$$

where $\hat{\mathbf{B}} = \sum_{i=1}^n \hat{U}_i \hat{U}_i^T$

The sandwich!!!

[2] Slides modeled on lecture notes by N. Breslow at UW

Bootstrap: More advanced R code

Not actually recommended (assumes you're an expert)

```
> library(boot)
> set.seed(4)
> myboot <- boot(airpol3, function(mydata, useme){ glm(Deaths PM10, family=poisson,
data=mydata[useme,])$coeff
}, R=1000)
>
> boot.ci(myboot, index=2, type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Level Percentile
95% ( 0.0004, 0.0025 )
Calculations and Intervals on Original Scale
>
> boot.ci(myboot, index=2, type="norm")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates
Intervals :
Level Normal
95% ( 0.0004, 0.0025 )
Calculations and Intervals on Original Scale
```

Bootstrap: More advanced R code†

- `useme` (2nd argument) specifies the rows to use
- `R` is the number of replicates from $\hat{F}_n$ (i.e. B)
- `index =2` specifies the second element in `glm(...)$coeff` i.e. $\hat{\beta}_1$
- `type="norm", "perc"` do Normal approximations, percentile methods, respectively. There are *slight* tweaks compared to the methods given above. See the documentation.

Bootstrap: parametric version†

Suppose we believe the assumed model $F(\theta)$

Usually, the likelihood-based approaches give adequate intervals for parameters. We may be interested in analytically difficult functions of F, e.g. medians, ranks, .. for example, taking derivatives may be awkward.

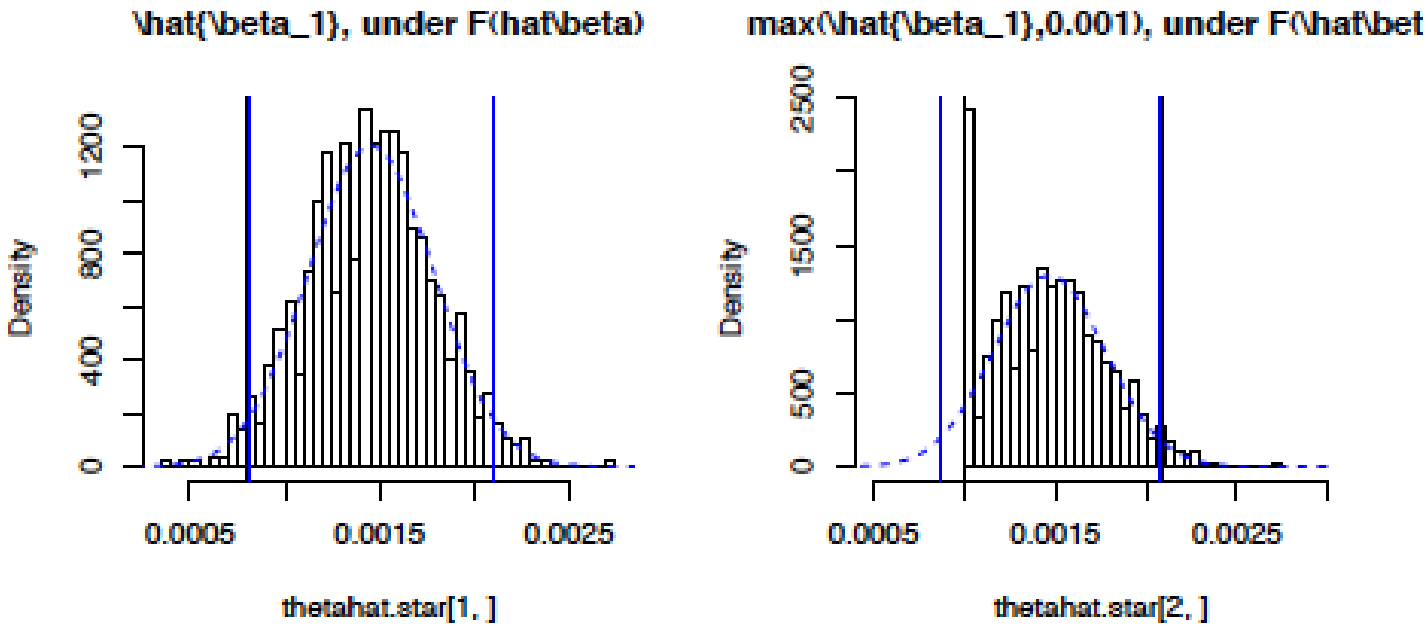
The parametric bootstrap is a useful approach; we take many replicates Y^* under $\hat{F}_n = F(\hat{\theta})$ and look at the long-run, frequentist behavior of $g(Y^*)$. Typically, $g()$ is some form of estimator.

Bootstrap: parametric version†

```
> mod1 <- glm(Deaths∼PM10, family=poisson, data=airpol3)
> fit1 <- fitted.values(mod1)
> bigB <- 5000
>
> do.one <- function( ){
+ NewDeaths <- rpois(335, lambda=fit1)
+ betahat <- coefficients(glm(NewDeaths ∼ PM10, family=poisson, data=airpol3))
+ c(betahat[2], max(betahat[2], 0.001) )
+ }
> set.seed(4)
> thetahat.star <- replicate(bigB, do.one() )
>
> apply(thetahat.star, 1, function(z){quantile(z, c(0.025, 0.975))})
PM10
2.5% 0.0008222455 0.001000000
97.5% 0.0021230006 0.002123001
```

See also `boot()`.

Bootstrap: parametric version, in R†



- Using $\hat{F}_n = F_{\hat{\beta}}$ for Poisson model. Blue/dashed lines indicate the moment approach, black/solid lines are the quantile-based estimates
- No difference in $\hat{\beta}_1$, but for $g(Y^*) = \max(\hat{\beta}_1(Y^*), 0.001)$ do not get approx Normal - caution required.

Bootstrap: parametric version, in R†

Point estimates from the quasi-likelihood, sandwich, and default bootstrap approaches are (here) identical. Parametric bootstrap and likelihood are also similar.

Inferential Method	$\widehat{\beta}_1$ ($\times 10^3$)	s.e.($\widehat{\beta}_1$) ($\times 10^4$)	95% confidence interval for $e^{10\beta_1}$
Poisson likelihood	1.5	3.3	(1.009,1.022)
Quasi-likelihood	1.5	5.6	(1.004,1.026)
Sandwich	1.5	5.3	(1.004,1.026)
Parametric bootstrap	1.5	3.3	(1.008,1.021)
Non-parametric bootstrap	1.5	5.4	(1.004,1.025)

The parametric bootstrap cannot be used with quasi-likelihood models, since we have no full probability distribution for the data.

Bootstrap: resampling residuals†

We discussed resampling cases, the default bootstrap method. In linear regression we can consider resampling residuals.

We illustrate the difference by considering the model

$$Y_i = \mu(x_i, \boldsymbol{\beta}) + \epsilon_i$$

where the residuals ϵ_i are such that $\mathbb{E}[\epsilon_i] = 0$ for $i = 1, \ldots, n$ and are independent.

We assume that F is the distribution of Y only. We consider the covariates x_i as fixed, for our data and all replications. The bootstrap datasets are formed as

$$Y_i^{*b} = \mu(x_i, \hat{\boldsymbol{\beta}}) + \epsilon_i^{*b}$$

where a number of options are available for sampling $\epsilon_i^{*(b)}$, $b = 1, \ldots, B$, $i = 1, \ldots, n$.

Bootstrap: resampling residuals†

The non-parametric version samples $\epsilon_i^{*(b)}$ with replacement from:

$$e_i = y_i - \mu(x_i, \hat{\beta}) - \left(\frac{1}{n} \sum_{j=1}^{n} y_j - f(x_j, \hat{\beta}) \right)$$

The second term centers the e_i, (should be minor, if $\mathbb{E}[\epsilon_i] = 0$).

If we assumed that $\epsilon_i \sim_{ind} N(0, \sigma^2)$ then a parametric resampling residuals method samples from

$$\epsilon_i^* \sim_{ind} N(0, \mathrm{Var}\{e_1, e_2, \ldots, e_n\})$$

- Both methods respect the "design" of the x_i
- Both methods break any relationship between x_i and ϵ_i, so they assume homoskedasticity. This is a big assumption, often violated by real data. It is violated by assumption of you believe $Y_i \sim Pois(\lambda_i)$.

Bootstrap: notes†

- Bootstrap methods do not work for all functions of interest. For example the maximum $Y_{(n)}$ if you want to estimate the range of F. Regularity conditions apply - roughly, asymptotic normality of $\hat{\theta}_n$ must hold.
- There are no small sample guarantees. Informally, if your data look nothing like F, you will get misleading results.
- The bootstrap is very good (i.e. competitive) at producing confidence intervals where the delta method is unhelpful but you must still give a scientific justification for your estimates (quantiles of F? number of modes in F?)

Bootstrap: notes†

- Theoretical validation of bootstrap methods is non-trivial; checking by simulating the bootstrap is quite slow
- Choosing B is not totally trivial, start with a few hundred and work your way up (the only penalty is computing time)
- Note this is a Monte Carlo method, we need to generate random numbers. Sandwich approach may be more appealing when possible on practical and rhetorical grounds.
- But the real advantage of the bootstrap is for complicated parameters (see for example the Felsenstein paper) where straightforward calculation of the sandwich/model based standard errors are not possible.

The jackknife

The jackknife, invented by Tukey and Quenouille is closely related to the bootstrap. The jackknife is a leave one out procedure. Let

$$Y_{(i)} = \{Y_1, Y_2, \ldots, Y_{i-1}, Y_{i+1}, \ldots Y_n\}$$

Similarly let $\hat{\theta}_{(i)}$ be the estimate of a parameter based on $Y_{(i)}$. Then the jackknife standard error estimate is:

$$\hat{\mathrm{Var}}_{jack} = \frac{n-1}{n}(\sum \hat{\theta}_i - \bar{\hat{\theta}})^2$$

The reason for the $n-1$ term is that the jackknife deviations are very close to each other (since we only leave out one observation), so they need to be inflated.

The jackknife can be thought of as a linear approximation to the bootstrap (see Efron 1979 or Efron and Tibshirani). For linear statistics:

$$\hat{\theta} = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(x_i)$$

they will agree up to a constant.

The jackknife

For non-linear statistics, particularly non-smooth statistics, this approximation can lead to problems. For example the sample median is not smooth. Consider the set of data

$$\{10, 27, 31, 40, 46, 50, 52, 104, 146\}$$

The median of the values is 46. If we start increasing the 4th largest value ($x = 40$). The median doesn't change at all until x becomes larger than 46, then the median is equal to x until x exceeds 50. So the median is not smooth (differentiable) function of x.

This lack of smoothness causes the jackknife estimate of the standard error to be inconsistent for the median. For the mouse data, the jackknife values for the median are:

$$\{48, 48, 48, 48, 45, 43, 43, 43\}$$

There are only 3 values, a consequence of the lack of smoothness of the median and the fact that the jackknife data sets only differ by 1 observation. The estimated jackknife variance is 44.62. A bootstrap estimate of the standard error based on $B = 100$ bootstrap samples is 91.72, considerably larger.

Vector outcomes

For univariate Y, nonparametric inference by sandwich methods is very close (in spirit and numerically) to use of the bootstrap; both use "plug-in" estimates of F, where the estimate is based on the data.

This result, and validity of bootstrap methods, also holds for analysis of vector outcomes; by studying an empirical, data-based estimate of F, we learn about e.g. $\mathrm{Var}_F[\hat{\boldsymbol{\beta}}_n]$.

Two competing strategies for estimating F are:

- Resample clusters with replacement, and within them sample observations $\{Y_{\cdot j}, X_{\cdot j}$ without replacement (S1)
- Resample clusters with replacement, and within each cluster sample observations $\{Y_{\cdot j}, X_{\cdot j}$ with replacement.

Vector outcomes

We[3] examine the properties of S1 and S2, in a tractable situation. The data generating mechanism is that

$$Y_{ij} = b_i + Z_{ij}, 1 \leq i \leq n, 1 \leq j \leq n_i$$

where n_i is constant for all i, and all b_i and Z_{ij} are independent with: $\mathbb{E}[b_i] = 0$, $\mathbb{E}[Z_{ij}] = 0$, $\mathrm{Var}[b_i] = \sigma_b^2$, $\mathrm{Var}[Z_{ij}] = \sigma_Z^2$ It follows that:

$$\begin{aligned} \mathrm{Var}[Y_{ij}] &= \sigma_b^2 + \sigma_Z^2 \\ \mathrm{Cov}[Y_{ij}, Y_{ij'}] &= \mathrm{Cov}[b_i + Z_{ij}, b_i + Z_{ij'}] \\ &= \mathrm{Var}[b_i] + \mathrm{Cov}[Z_{ij}, Z_{ij'}] = \sigma_b^2 \end{aligned}$$

Note the intraclass correlation is $\frac{\sigma_b^2}{\sigma_b^2+\sigma_Z^2}$; it can be interpreted as the correlation between two outcomes in the same cluster, or the proportion of total variance that is "between clusters".

[3] following Davison and Hinkley pg 100-102

Vector outcomes

Denoting the bootstrap distribution by a star, and denoting the random cluster number by I^*, under either S1 or S2 we get:

$$\begin{aligned}\mathbb{E}_{F^*}[Y_{I^*j}|I^*=i] &= n_i^{-1}\sum_l Y_{il} = \bar{Y}_i \\ \mathbb{E}_{F^*}[Y_{I^*j}^2|I^*=i] &= n_i^{-2}\sum_l Y_{il}^2\end{aligned}$$

and consequently:

$$\begin{aligned}\mathbb{E}_{F^*}[Y_{ij}] &= n_i^{-1}n^{-1}\sum_{ij} Y_{ij} \\ \mathbb{E}_{F^*}[Y_{ij}^2] &= n^{-1}\sum_i(\bar{Y}_i-\bar{Y})^2 + n^{-1}\sum_i n_i^{-1}(Y_{ij}-\bar{Y}_i)^2\end{aligned}$$

We see immediately that the expectation of the resampled outcomes is unbaised for $\mathbb{E}_F[Y_{\cdot j}]$. Slightly more work gives:

$$\mathbb{E}\left[\mathrm{Var}^*[Y_{\cdot j}^*]\right] = \frac{n-1}{n}\sigma_b^2 + \frac{nn_i-1}{nn_i}\sigma_Z^2$$

Vector outcomes

But for the cross-terms with $j \neq k$ we get:

$$\mathbb{E}_{F^*}[Y_{I^*j} Y_{I^*k} | I^* = i] = \frac{1}{n_i(n_i - 1)} \sum_{l \neq m} Y_{il} Y_{im}$$

$$\mathbb{E}_{F^*}[Y_{I^*j} Y_{I^*k} | I^* = i] = \frac{1}{n_i^2} \sum_{l \neq m} Y_{il} Y_{im}$$

for models (S1) and (S2).Given a particular data set the bootstrap(s) give

$$\mathrm{Cov}_{F^*}[Y_{ij} Y_{ik}] = n^{-1} \sum_i (\bar{Y}_i - \bar{Y})^2 - \frac{1}{nn_i(n_i - 1)} \sum_{ij} (Y_{ij} - \bar{Y}_i)^2$$

$$\mathrm{Cov}_{F^*}[Y_{ij} Y_{ik}] = \frac{1}{n} (\bar{Y}_i - \bar{Y})^2$$

for models (S1) and (S2) and this leads to:

$$\mathbb{E}[\mathrm{Cov}_{F^*}[Y_{ij} Y_{ik}]] = \frac{n-1}{n} \sigma_b^2 - \frac{1}{nn_i} \sigma_Z^2$$

$$\mathbb{E}[\mathrm{Cov}_{F^*}[Y_{ij} Y_{ik}]] = \frac{n-1}{n} \sigma_b^2 - \frac{n-1}{nn_i} \sigma_Z^2$$

for models (S1) and (S2). Recall that this covariance is just σ_b^2 in the true F; which formula above is the closest to that value?

Vector outcomes

Notes on this:

- Resampling whole clusters leads to better reconstruction of the first two moments of F
- Given that we didn't specify a within-cluster correlation structure, naïvely using a bootstrap that treats observations as i.i.d. within clusters seems unwise.
- With observations $\{Y_i, X_{i1}, X_{i2}, X_{i3}\}$ you'd resample whole observations. Now we have e.g. $\{Y_i, Y_{i2}, X_{i1}, X_{i2}\}$ ... so do the same.
- If n_i is very large or σ^2 is tiny, resampling within clusters won't hurt
- The default nonparametric bootstrap resamples whole clusters. In `R` you will have to write this yourself.