Bootstrap

Set up: Data $X_1, \ldots, X_n$ — i.i.d. as X with population cdf F (which is not completely known).

Parameter of interest: θ, estimated by $\hat{\theta}$

Examples: Mean, variance, median, quantiles, etc.

Issue: Need to get the *sampling distribution* of $\hat{\theta}$ so that we can compute, e.g., standard error of $\hat{\theta}$, or confidence interval for θ?

Q: Why not use the methods that we have learnt?

Basics

Bootstrap: A simulation based technique that allows us to approximate the sampling distribution of $\hat{\theta}$. Assumes large n, but its value needed for validity of bootstrap is typically less than that for the usual large-sample procedure.

Original sample: $X_1, \ldots, X_n \sim$ i.i.d. with cdf F

Bootstrap (re)sample: $X_1^*, \ldots, X_n^* \sim$ i.i.d. with cdf $\hat{F}$, where $\hat{F} =$ estimated cdf (which is completely known)

Parametric bootstrap:

- Functional form of F is known (e.g., normal), but F may depend on unknown parameter θ.
- $\hat{F}$ is same as F but with θ replaced by its MLE $\hat{\theta}$. In other words, $\hat{F}$ is the cdf of the fitted model.
- Ex: $F = N(\mu, \sigma^2)$, $\hat{F} =$
- Often easy to simulate i.i.d. draws $X_1^*, \ldots, X_n^*$ from $\hat{F}$.

Nonparametric bootstrap:

- Functional form of F is unknown.
- $\hat{F}$ = empirical cdf, where

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) =$$

- Think of $\hat{F}$ as a discrete distribution that assigns $1/n$ probability to each of the sample observations, $X_1, \ldots, X_n$.
- Get $X_1^*, \ldots, X_n^*$ by sampling n times **with replacement** from $X_1, \ldots, X_n$.

Bootstrap distribution of estimator $\hat{\theta}$ of θ:

Original sample: $X_1, \dots, X_n$ — gives $\hat{\theta}$

- Simulate a large number b of *bootstrap resamples*, and compute $\hat{\theta}^*$ from each resample exactly the way $\hat{\theta}$ is computed from original sample.

- This process gives a large number of draws, $\hat{\theta}_1^*, \dots, \hat{\theta}_b^*$.
- These draws are coming from the *bootstrap distribution* of $\hat{\theta}$.
- How to see this distribution?
- It approximates the *sampling distribution* of $\hat{\theta}$.
- Use the draws $\hat{\theta}_1^*, \dots, \hat{\theta}_b^*$ to estimate features of sampling distribution of $\hat{\theta}$ that may be of interest.

Estimating a feature η of distribution of $\hat{\theta}$:

- Get a large number b of draws, $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$.
- $\hat{\eta}^*$ = same feature computed from these draws.

Ex 1: $\eta = E(\hat{\theta})$. $\hat{\eta}^* =$

Ex 2: $\eta = var(\hat{\theta})$. $\hat{\eta}^* =$

Ex 3: $\eta =$ bias of $\hat{\theta} = E(\hat{\theta}) - \theta$. $\hat{\eta}^* =$

Ex 4: $\eta = \alpha$-th quantile of $\hat{\theta}$. $\hat{\eta}^* =$

Ex 5: $\eta = \alpha$-th quantile of $\hat{\theta} - \theta$. $\hat{\eta}^* =$

Bootstrap Confidence Intervals

Set up: $\hat{\theta} \approx N(\theta, \hat{V})$ when n is large.

- For example, when $\hat{\theta}$ is MLE and $\hat{V} = \hat{I}^{-1}$.
- Don't need population to be normal.

Recall: The standard (approximate) $100(1-\alpha)\%$ CI for θ is:

$$[\hat{\theta} - z_{1-\alpha/2}\,\widehat{SE},\ \hat{\theta} - z_{\alpha/2}\,\widehat{SE}],$$

where $z_\alpha = \alpha$-th percentile of $T = (\hat{\theta} - \theta)/\widehat{SE} \approx N(0,1)$.

Why?

Issues: This CI may not be accurate because n may not be large enough for

- normal approximation for T to be good, implying that

  (The distribution of T may not even be symmetric.)
- bias in $\hat{\theta}$ to be negligible, implying that
- $\hat{V}$ to be a good estimate of true V, implying that

  (Often ML-theory based $\hat{V}$ underestimates V.)

Bootstrap CI overcomes these issues to a large extent.

Four Bootstrap CIs for θ

1. <u>Normal approximation CI:</u> Use z critical point but correct $\hat{\theta}$ for bias and use $\hat{V}^*$ to estimate V.

- Estimated bias of $\hat{\theta} =$
- CI: $\left[\hat{\theta} - \hat{B}^* - z_{1-\alpha/2}\,\widehat{SE}^*,\ \hat{\theta} - \hat{B}^* - z_{\alpha/2}\,\widehat{SE}^*\right]$.

2. <u>Studentized bootstrap CI:</u> Use bootstrap critical point of T instead of z critical point.

- Get $T_1^* = (\hat{\theta}_1^* - \hat{\theta})/\widehat{SE}_1^*, \ldots, T_b^* = (\hat{\theta}_b^* - \hat{\theta})/\widehat{SE}_b^*$
- Estimated α-th percentile of $T =$
- CI: $\left[\hat{\theta} - t^*_{\left((b+1)(1-\alpha/2)\right)}\widehat{SE},\ \hat{\theta} - t^*_{\left((b+1)(\alpha/2)\right)}\widehat{SE}\right]$.

3. <u>Basic bootstrap CI:</u> Based on percentiles of $\hat{\theta} - \theta$ rather than $(\hat{\theta} - \theta)/\widehat{SE}$. Use bootstrap to estimate them. Notice

$$1 - \alpha = P(a_{\alpha/2} \leq \hat{\theta} - \theta \leq a_{1-\alpha/2})$$
$$=$$

- Estimated $a_\alpha =$
- CI: $\left[2\hat{\theta} - \hat{\theta}^*_{\left((b+1)(1-\alpha/2)\right)},\ 2\hat{\theta} - \hat{\theta}^*_{\left((b+1)(\alpha/2)\right)}\right]$.
- Doesn't require $\widehat{SE}$.

4. Percentile bootstrap CI: Works as in basic bootstrap but uses "magic." Suppose there exists a transformation h so that the distribution of $h(\hat{\theta}) - h(\theta)$ is symmetric about zero. Let $U = h(\hat{\theta})$. As before, we can write

$$\begin{aligned} 1 - \alpha &= P\left(a_{\alpha/2} \leq U - h(\theta) \leq a_{1-\alpha/2}\right) \\ &= P\left(-a_{1-\alpha/2} \leq U - h(\theta) \leq -a_{\alpha/2}\right) \\ &= P\left(U + a_{\alpha/2} \leq h(\theta) \leq U + a_{1-\alpha/2}\right) \end{aligned}$$

- Estimated $a_\alpha =$
- $U + a_{\alpha/2} \approx U + \left\{U^*_{\left((b+1)(\alpha/2)\right)} - U\right\} = U*_{\left((b+1)(\alpha/2)\right)}$
- Similarly, $U + a_{1-\alpha/2} = U^*_{\left((b+1)(1-\alpha/2)\right)}$

Therefore,

$$1 - \alpha \approx P\left(U^*_{\left((b+1)(\alpha/2)\right)} \leq h(\theta) \leq U^*_{\left((b+1)(1-\alpha/2)\right)}\right)$$

$$=$$

- CI: $\left[\hat{\theta}^*_{\left((b+1)(\alpha/2)\right)},\ \hat{\theta}^*_{\left((b+1)(1-\alpha/2)\right)}\right]$.
- Magic:

Q. Which method to use?

Research shows that studentized bootstrap is the best choice, but it requires $\widehat{SE}$. However, if $\widehat{SE}$ is not available, then percentile bootstrap is often the next best choice. More accurate versions of this method are available.

Example: Recall the CPU time data from Example 8.12 on page 217. We had seen that a gamma distribution fit well to these data. Suppose we would like to perform inference on median cpu time.

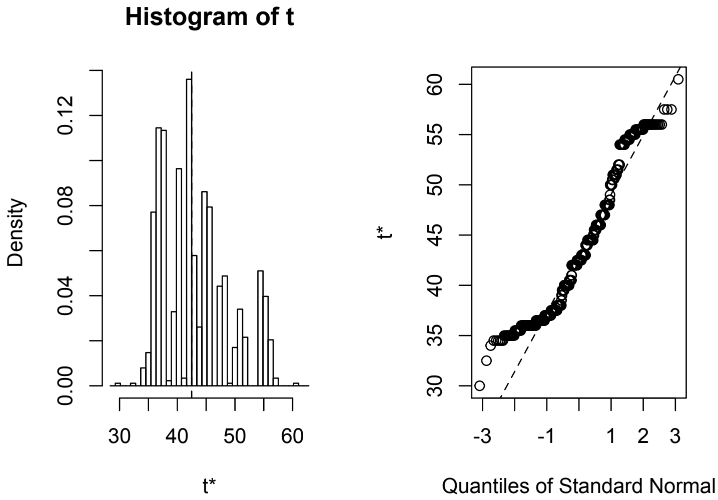
R code:

```
# use install.packages("boot") to first install
# the package and then load it

library(boot)

# read the cpu data (we have seen these before)

> (cpu <- scan(file="cputime.txt"))
Read 30 items
 [1]  70  36  43  69  82  48  34  62  35  15  59 139
 46  37  42  30  55  56
```

```
[19]  36  82  38  89  54  25  35  24  22   9  56  19
>

# Parameter of interest: Median

###########################
# Nonparametric Bootstrap #
###########################

median.npar <- function(x, indices) {
  result <- median(x[indices])
  return(result)
  }

> (median.npar.boot <- boot(cpu, median.npar, R=999,
sim="ordinary", stype="i"))

ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
Call:
boot(data = cpu, statistic = median.npar, R = 999,
sim = "ordinary",  stype = "i")

Bootstrap Statistics :
    original     bias    std. error
t1*     42.5 0.6721722    5.876943
>

# Let's verify the calculations
# See what's else is stored in median.npar.boot

> names(median.npar.boot)
 [1] "t0"          "t"           "R"           "data"
 "seed"        "statistic"
 [7] "sim"         "call"        "stype"       "strata"
  "weights"
```

```
>

> median(cpu)
[1] 42.5
>
> median.npar.boot$t0
[1] 42.5
>
> mean(median.npar.boot$t)-median.npar.boot$t0
[1] 0.6721722
>
> sd(median.npar.boot$t)
[1] 5.876943
>
# See the bootstrap distribution of median estimate

plot(median.npar.boot)
```


Histogram of t
Density
0.00
0.04
0.08
0.12
30
40
50
60
t*
t*
30
35
40
45
50
55
60
-3
-1
1
2
3
Quantiles of Standard Normal

```
# Get the 95% confidence interval for median
> boot.ci(median.npar.boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates
CALL :
boot.ci(boot.out = median.npar.boot)

Intervals :
Level      Normal              Basic
95%   (30.31, 53.35 )   (29.50, 49.50 )

Level     Percentile            BCa
95%   (35.5, 55.5 )   (35.0, 55.5 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(median.npar.boot) :
  bootstrap variances needed for studentized intervals
```

```
# Let's verify
# Normal approximation method

> c(42.5 - 0.6721722 - qnorm(0.975) * 5.876943,
    42.5 - 0.6721722 - qnorm(0.025) * 5.876943)
[1] 30.30923 53.34642
>

# Percentile bootstrap method
> sort(median.npar.boot$t)[c(25, 975)]
[1] 35.5 55.5
>

# Basic bootstrap method
> c(2*42.5-55.5, 2*42.5-35.5)
[1] 29.5 49.5
>
```