Bootstrap and Resampling Methods

- Often we have an estimator T of a parameter θ and want to know its sampling properties
  - to informally assess the quality of the estimate
  - for formal confidence intervals
  - for formal hypothesis tests
- In some cases we can obtain exact results; usually this requires
  - particularly convenient statistics (e.g. linear)
  - particularly convenient data models (e.g. normal)
- For more complex problems we can use approximations
  - based on large samples
  - use central limit theorem
  - use Taylor series approximations
- Alternative: use simulation in place of central limit theorem and approximations
- References:
  - Givens and Hoeting, Chapter 9.
  - Davison, A. C. and Hinkley, D. V. (1997) "Bootstrap Methods and their Application," Cambridge University Press.
  - Package `boot`; R port of S-Plus code written to support Davison and Hinkley.

Example: A Parametric Bootstrap

- Ordered times between failures of air conditioning unit on one aircraft:

3	5	7	18	43	85	91	98	100	130	230	487

- Suppose the data are exponentially distributed with mean μ and we want to estimate the failure rate $\theta = 1/\mu$.
- The MLE of μ is $\widehat{\mu} = \overline{X} = 108.0833$.
- The MLE of θ is $T = 1/\overline{X}$.
- The exact sampling distribution is inverse gamma with mean μ, which we do not know.
- We could compute the exact bias and variance of T,

$$b(\mu) = E_\mu[T] - \theta = E_\mu[T] - 1/\mu = \frac{1}{n-1}\frac{1}{\mu} = \frac{1}{n-1}\theta$$
$$v(\mu) = \mathrm{Var}_\mu(T) = \frac{1}{\mu^2}\frac{n^2}{(n-1)^2(n-2)} = \theta^2\frac{n^2}{(n-1)^2(n-2)}$$

and obtain plug-in estimates

$$B = b(\widehat{\mu})$$
$$V = v(\widehat{\mu})$$

- One alternative to exact computation of the sampling distribution is the delta method.
- The first order delta method approximates T by

$$T = 1/\overline{X} \approx 1/\mu - (\overline{X} - \mu)/\mu^2$$

So

$$\begin{aligned} b(\mu) &\approx 0 \\ v(\mu) &\approx \text{Var}(\overline{X})/\mu^4 = \frac{1}{n}\frac{\mu^2}{\mu^4} = \frac{\theta^2}{n} \end{aligned}$$

- A second order delta method for the bias uses

$$T = 1/\overline{X} \approx 1/\mu - (\overline{X} - \mu)/\mu^2 + (\overline{X} - \mu)^2/\mu^3$$

So

$$b(\mu) \approx \text{Var}(\overline{X})/\mu^3 = \frac{\theta}{n}$$

- Instead of working out $b(\cdot)$ and $v(\cdot)$ analytically, we can estimate B and V by simulation:
  - Draw a sample $X_1^*, \dots, X_{12}^*$ from an exponential distribution with mean $\mu = \widehat{\mu} = 108.0833$.
  - Compute T^* from this sample.
  - Repeat R times to obtain $T_1^*, \dots, T_R^*$
  - Estimate the bias and variance of T by

$$\begin{aligned} B^* &= \overline{T}^* - T \\ V^* &= \frac{1}{R-1}\sum_{i=1}^{R}(T_i^* - \overline{T}^*)^2 \end{aligned}$$

- The full sampling distribution can be examined using a histogram or other density estimate.

Bootstrap and Resampling Methods

Example: A Nonparametric Bootstrap

- Instead of assuming an exponential population in assessing the performance of T we can make the weaker assumption that $X_1,\dots,X_n$ is a random sample from an arbitrary distribution F.
- A delta method approximation to the variance is

$$\text{Var}(T) \approx \frac{S_X^2}{n\overline{X}^4}$$

- A non-parametric bootstrap approach takes the empirical distribution $\widehat{F}$ with

$$\widehat{F}(t) = \frac{\#\{X_i \leq t\}}{n}$$

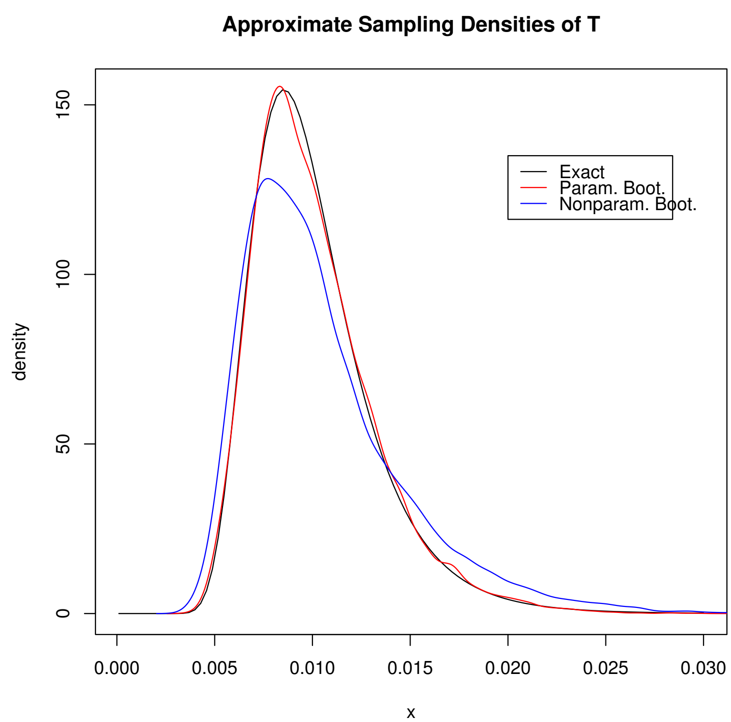
  as an approximation to F, and
  - draws a sample $X_1^*,\dots,X_{12}^*$ from $\widehat{F}$,
  - computes T^* from this sample,
  - repeats R times to obtain $T_1^*,\dots,T_R^*$,
  - estimate the bias and variance of T by

$$B^* = \overline{T}^* - T$$
$$V^* = \frac{1}{R-1}\sum_{i=1}^{R}(T_i^* - \overline{T}^*)^2$$

- Drawing a random sample from $\widehat{F}$ is the same as sampling with replacement from $\{X_1,\dots,X_n\}$

Some Results

Method	Bias	Stand. Dev.
Exact	0.000841	0.003192
Param. Delta (1st order)	0.0	0.002671
Param. Delta (2nd order)	0.0007710	
Nonparam. Delta (1st order)	0.0	0.003366
Nonparam. Delta (2nd order)	0.001225	
Parametric Bootstrap ($R = 10000$)	0.000857	0.003185
Nonparametric Bootstrap ($R = 10000$)	0.001244	0.004170



Notes

- There are two kinds of errors in the bootstrap approaches (Davison and Hinkley's terminology):
  - The *statistical error* due to using $\widehat{\mu}$ or $\widehat{F}$ instead of μ or F.
  - The *simulation error*.
- The simulation error can be reduced by
  - increasing R
  - using variance reduction methods if applicable
- The statistical error is reduced as the sample size n gets large.
- In some settings the statistical error can be reduced by working with a transformed parameter.
- For the parametric bootstrap, as $R \to \infty$

$$
\begin{aligned}
B^* &\to b(\widehat{\mu})\\
V^* &\to v(\widehat{\mu})
\end{aligned}
$$

  and

$$
\begin{aligned}
b(\widehat{\mu}) &= \frac{1}{n-1}\frac{1}{\widehat{\mu}} = b(\mu) + O_P(n^{-3/2})\\
v(\widehat{\mu}) &= \frac{1}{\widehat{\mu}}\frac{n^2}{(n-1)^2(n-2)} = v(\mu) + O_P(n^{-3/2})
\end{aligned}
$$

- Similar results hold, under suitable regularity conditions, for the nonparametric bootstrap.
- Suppose $X_1,\ldots,X_n$ are a random sample from F with mean μ_F and variance σ_F^2. Let

$$\begin{aligned} \theta_F &= \mu_F^2 \\ T &= \overline{X}^2 \end{aligned}$$

- The bias of T is

$$b(F) = E_F[T] - \theta(F) = E_F[\overline{X}^2] - \mu_F^2 = \text{Var}_F(\overline{X}) = \frac{1}{n}\sigma_F^2$$

- The bootstrap bias ($R = \infty$) is

$$b(\widehat{F}) = \frac{1}{n}\sigma_{\widehat{F}}^2 = \frac{1}{n}\frac{n-1}{n}S^2$$

- If the population has finite fourth moments, then

$$b(\widehat{F}) = b(F) + O_P(n^{-3/2})$$

- Except in some special cases the theoretical justification for the bootstrap is asymptotic in sample size.
- The motivation for the *nonparametric* bootstrap is often similar to the motivation for using the sandwich estimator of variance:
  - use a model to suggest an estimator T
  - do not assume the model in assessing the estimator's performance

A Simple Implementation

- The basic operation of bootstrapping is to generate samples and collect statistics:

```
b <- function(stat, gen, R)
    sapply(1:R, function(i) stat(gen()))
```

- A generator for an exponential model can be constructed by

```
makeExpGen <- function(data) {
    rate <- 1 / mean(data)
    n <- length(data)
    function() rexp(n, rate)
}
```

  This uses *lexical scope* to capture the variables `rate` and `n`.

- A parametric bootstrap sample of T is produced by

```
v <- b(function(data) 1 / mean(data),
       makeExpGen(aircondit$hours), 10000)
```

- A generator for a nonparametric bootstrap sample drawn from the empirical distribution is constructed by

```
makeEmpGen <- function(data) {
    if (is.vector(data)) {
        n <- length(data)
        function() data[sample(n, n, replace = TRUE)]
    }
    else {
        n <- nrow(data)
        function() data[sample(n, n, replace = TRUE), , drop = FALSE]
    }
}
```

  - The variables `data` and `n` are captured by lexical scope.
  - The `data` can be a vector, a matrix, or a data frame.

- A nonparametric bootstrap sample of T is produced by

```
vv <- b(function(data) 1 / mean(data),
        makeEmpGen(aircondit$hours), 10000)
```

A Cautionary Example

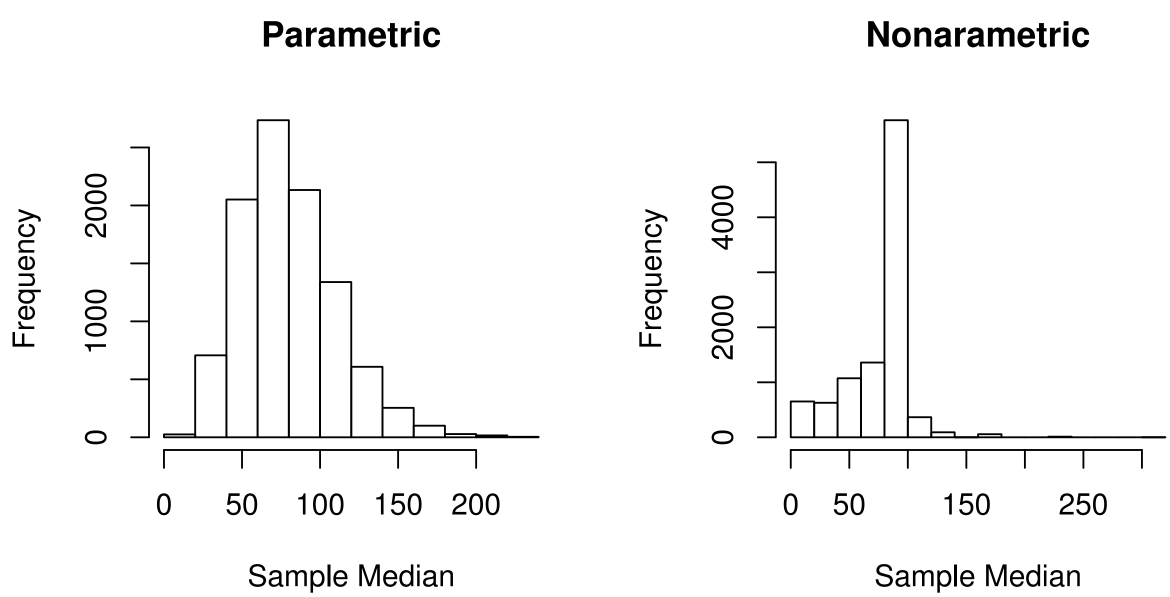
- In many elementary bootstrap applications we have, at least approximately,

$$T = T(\widehat{F})$$
$$\theta = T(F)$$

- For the basic nonparametric bootstrap to work, some level of smoothness of θ as a function of F is needed.
- Suppose we are interested in examining the sampling distribution of the median for the air conditioner data.
- Parametric and nonparametric bootstrap samples are produced by

```
medpb <- b(median, makeExpGen(aircondit$hours),10000)
mednpb <- b(median, makeEmpGen(aircondit$hours),10000)
```

- The resulting histograms are



- A more complex smoothed bootstrap can be used to address this issue.

- Suppose
  - $v_n(T) = \text{Var}(T)$ for a sample of size n
  - V_n^* is the boostrap estimate for $R = \infty$ for a sample of size n.

  Then, under suitable conditions,

$$\begin{aligned} nv_n(T) &\to \frac{1}{4f(\theta)^2} \\ nV_n^*(T) &\to \frac{1}{4f(\theta)^2} \end{aligned}$$

  In this sense the boostrap is *valid*.

- The mean square error $E[(nV_n^*(T) - nv_n(T))^2]$ tends to zero slower than for smooth functionals ($O(n^{-1/2})$ instead of $O(n^{-1})$).
- The mean square error for a suitable smooth bootstrap converges faster than for an unsmoothed one, but not as fast as for smooth functionals ($O(n^{-4/5})$ can be achieved).

Bootstrapping Regression Models

- Suppose we have independent observations $(x_1, y_1), \ldots, (x_n, y_n)$.
- Regression models model the conditional distribution $y|x$.
- Linear regression models assume $E[Y_i|x_i] = x_i\beta$.
- Often we assume $\text{Var}(Y_i|x_i)$ is constant.
- Parametric bootstrapping can be used for a parametric model.
- Two forms of non-parametric bootstrap are used:
  - Case-based bootstrapping: pairs (x^*, y^*) are selected with replacement from $(x_1, y_1), \ldots, (x_n, y_n)$. This is also called a *pairs bootstrap*.
  - Model-based bootstrapping: Residuals are formed using the model, and bootstrap samples are constructed by sampling the residuals.

Case-Based Bootstrap

- Assumes the (X_i, Y_i) are sampled from a multivariate population.
- Advantages:
  - Simplicity.
  - Does not depend on an assumed mean model.
  - Does not depend on a constant variance assumption.
  - Does not depend on the notion of a residual.
  - Similar in spirit to the sandwich estimator.
- Disadvantages:
  - Not appropriate for designed experiments.
  - Does not reproduce standard results conditional on x_i
  - For bivariate normal data the bootstrap distribution of $\widehat{\beta}_1$ is estimating a mixture over the x_i of

    $$N(\beta, \sigma^2 / \sum (x_i - \overline{x})^2)$$

    distributions, i.e. a non-central t_{n-1} distribution.
  - Does reproduce standard results for studentized quantities where the conditional distribution does not depend on the x_i.
  - Can lead with high probability to design matrices that are not of full rank.

Model-Based Bootstrap

- This is a semi-parametric approach.
- It assumes
  - $Y_i = \mu(x_i) + \varepsilon_i$
  - the ε_i have a common distribution G with mean zero
  - $\mu(x_i)$ has a parametric form (usually).
- The *raw residuals* are of the form $e_i = y_i - \widehat{\mu}(x_i)$.
- For generalized linear models one can use various definitions of raw residuals (original scale, linear predictor scale, deviance residuals, e.g.).
- The raw residuals usually do not have constant variance; for linear models
$$\text{Var}(e_i) = \sigma^2(1-h_i)$$
where h_i is the *leverage* of the i-th case, the i-th diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$.
- The *modified residuals* are
$$r_i = \frac{y_i - \widehat{\mu}(x_i)}{(1-h_i)^{1/2}} = \frac{e_i}{(1-h_i)^{1/2}}$$
Approximate leverages are used for generalized linear and nonlinear models.
- The modified residuals usually do not have mean zero, so need to be adjusted.
- The model-based resampling algorithm:
  - randomly resample ε_i^* with replacement from $r_1 - \overline{r}, \ldots, r_n - \overline{r}$.
  - set $y_i^* = \widehat{\mu}(x_i) + \varepsilon_i^*$
  - fit a model to the (x_i, y_i^*).

Example: Leukemia Survival

- For the leukemia survival data of Feigl and Zelen we can fit a model

$$\log \text{Time}_i = \beta_0 + \beta_1 \log(\text{WBC}_i / 10000) + \beta_2 \text{AG}_i + \varepsilon_i$$

  where the ε_i are assumed to have zero mean and constant variance.

- We can compute the fit and some diagnostics with

```
library(boot)
fz <-
    read.table(
        "http://www.stat.uiowa.edu/~luke/classes/STAT7400/feigzel2.dat"
        head = TRUE)
fz.lm <- glm(log(fz$Time) ~ log(fz$WBC / 10000) + fz$AG)
fz.diag <- glm.diag(fz.lm)
```

- A case-based bootstrap is computed by

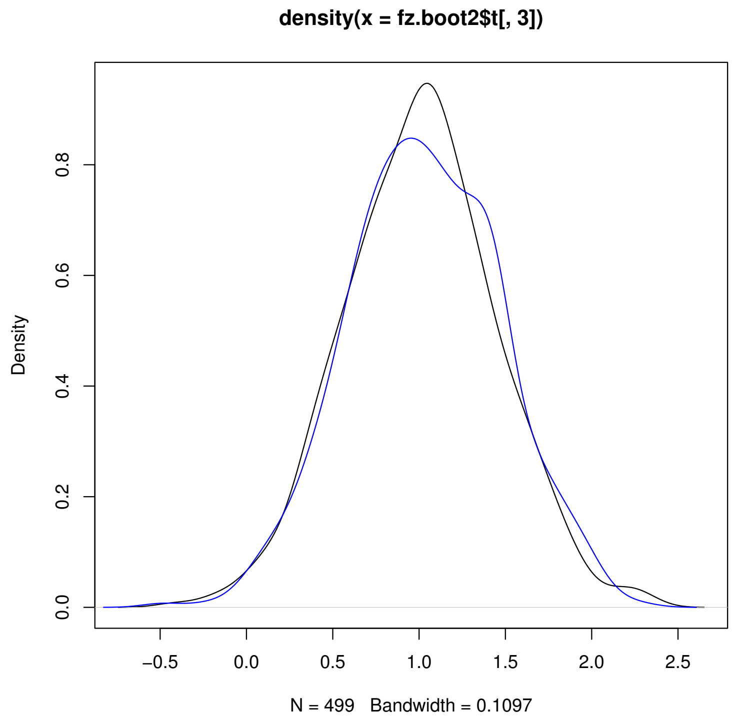
```
fz.fit <- function(d)
    coef(glm(log(d$Time) ~ log(d$WBC / 10000) + d$AG))
fz.case <- function(data, i)
    fz.fit(data[i,])
fz.boot1 <- boot(fz, fz.case, R=499)
```

- A model-based bootstrap is computed by

```
fz.res <- residuals(fz.lm) / sqrt(1 - fz.diag$h)
fz.res <- fz.res - mean(fz.res)
fz.df <- data.frame(fz, res = fz.res, fit = fitted(fz.lm))
fz.model <- function(data, i) {
    d <- data
    d$Time <- exp(d$fit + d$res[i])
    fz.fit(d)
}
fz.boot2 <- boot(fz.df, fz.model, R=499)
```

A comparison of the bootstrap distributions for $\widehat{\beta}_2$:

```
plot(density(fz.boot2$t[,3]))
lines(density(fz.boot1$t[,3]),col="blue")
```



Method	$\widehat{\beta}_2$ or $\overline{\widehat{\beta}}_2^*$	Stand. Error
Least Squares	0.9883	0.4361
Case-based Bootstrap	1.0187	0.4389
Model-based Bootstrap	0.9686	0.4353

Bootstrap Confidence Intervals

Simple Intervals

- A simple interval can be based on $T - \theta$ being approximately normal with mean B^* and variance V^*:

$$[\theta_L, \theta_U] = T - B^* \pm \sqrt{V^*} z_{1-\alpha/2}$$

- With the variance stabilizing logarithm transformation this becomes

$$[\theta_L, \theta_U] = \exp\{\log T - B_L^* \pm \sqrt{V_L^*} z_{1-\alpha/2}\}$$

  with B_L^* and V_L^* the bootstrap estimates of bias and variance for $\log T$.

- Results for the air conditioner data and $\alpha = 0.05$:

Parametric	Log Scale	θ_L	θ_U
yes	no	0.002152	0.01464
no	no	-0.000165	0.01618
yes	yes	0.004954	0.01585
no	yes	0.004238	0.01783

Basic Bootstrap Intervals

- If $P(T-\theta \leq a) = \alpha/2$ and $P(T-\theta \geq b) = \alpha/2$ then

$$P(a \leq T-\theta \leq b) = P(T-b \leq \theta \leq T-a) = 1-\alpha$$

- We can estimate a, b by $T^*_{((R+1)\alpha/2)} - T$ and $T^*_{((R+1)(1-\alpha/2))} - T$ to get the interval

$$\begin{aligned}[\theta_L, \theta_U] &= [T-(T^*_{((R+1)(1-\alpha/2))} - T), T-(T^*_{((R+1)\alpha/2)} - T)] \\ &= [2T - T^*_{((R+1)(1-\alpha/2))}, 2T - T^*_{((R+1)\alpha/2)}]\end{aligned}$$

- This can be done on a variance stabilizing scale as well; for the logarithm:

$$[\theta_L, \theta_U] = [T^2/T^*_{((R+1)(1-\alpha/2))}, T^2/T^*_{((R+1)\alpha/2)}]$$

- Results for the air conditioner data and $\alpha = 0.05$:

Parametric	Log Scale	θ_L	θ_U
yes	no	0.0006166	0.01286
no	no	-0.0026972	0.01325
yes	yes	0.0047855	0.01516
no	yes	0.0040375	0.01628

Studentized Bootstrap Intervals

- Suppose we have
  - an estimator T of θ
  - an estimator U of the variance of T, e.g. from the delta method
- From the bootstrap samples compute
  - estimates $T_1^*, \dots, T_R^*$
  - variance estimates $U_1^*, \dots, U_R^*$
- Compute and rank studentized values

$$Z_i^* = \frac{T_i^* - T}{\sqrt{U_i^*}}$$

- The studentized interval is then

$$[\theta_L, \theta_U] = [T - \sqrt{U} Z^*_{((R+1)(1-\alpha/2))}, T - \sqrt{U} Z^*_{((R+1)\alpha/2)}]$$

- For the air conditioner data the delta method variance estimate for T is

$$U = \frac{1}{n} \frac{S^2}{\overline{X}^4}$$

  For $\log T$ the variance estimate is

$$U_L = \frac{1}{n} \frac{S^2}{\overline{X}^2}$$

Other Bootstrap Confidence Intervals

- Percentile Methods
  - Basic: $[T^*_{((R+1)\alpha/2)}, T^*_{((R+1)(1-\alpha/2))}]$
  - Does not work very well with a nonparametric bootstrap
  - Modified: BC$_a$ method
- ABC method.
- Double bootstrap.
- ...

Bootstrap and Monte Carlo Tests

- Simulation can be used to compute null distributions or p-values.
- If the distribution of the test statistic under H_0 has no unknown parameters, then the p-value
$$P(T \geq t | H_0)$$
can be computed by simulation with no statistical error.
- If the distribution of the test statistic under H_0 does have unknown parameters but a sufficient statistic S for the model under H_0 is available, then it may be possible to compute the conditional p-value
$$P(T \geq t | S = s, H_0)$$
by simulation with no statistical error.
- In other cases, an approximate p-value
$$P(T \geq t | \widehat{F}_0)$$
is needed.
- Transformations and reparameterizations can help reduce the statistical error in these cases.

Parametric Tests: Logistic Regression

- Suppose $Y_1,\ldots,Y_n$ are independent binary responses with scalar covariate values $x_1,\ldots,x_n$.
- A logistic regression model specifies

$$\log \frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = \beta_0 + \beta_1 x_i$$

- Under $H_0 : \beta_1 = 0$, $S = \sum Y_i$ is sufficient for β_0.
- A natural test statistic for $H_1 : \beta_1 \neq 0$ is

$$T = \sum x_i Y_i$$

- Conditional on $S = s$ and H_0 the distribution of the Y_i is uniform on the $\binom{n}{s}$ possible arrangements of s ones and $n-s$ zeros.
- We can simulate from this distribution by generating random permutations.
- Another (in this case silly) option is to use MCMC in which each step switches a random y_i, y_j pair.

Parametric Bootstrap Tests: Separate Families

- For the air conditioner data we might consider testing

$$\begin{aligned} H_0 &: \quad \text{data have a Gamma}(\alpha,\beta) \text{ distribution} \\ H_1 &: \quad \text{data have a Log Normal}(\mu,\sigma^2) \text{ distribution} \end{aligned}$$

- A test can be based on

$$T = \frac{1}{n}\sum \log \frac{f_1(x_i|\widehat{\mu},\widehat{\sigma}^2)}{f_0(x_i|\widehat{\alpha},\widehat{\beta})}$$

- A normal approximation for the null distribution of T exists but may be inaccurate.
- The bootstrap null distribution of T is computed by:
  - Generate R random samples from the fitted null model, Gamma$(\widehat{\alpha},\widehat{\beta})$.
  - For each sample calculate the MLE's $\widehat{\alpha}^*$, $\widehat{\beta}^*$, $\widehat{\mu}^*$, and $\widehat{\sigma}^{*2}$.
  - Using these MLE's compute T^* for each sample
- For small samples this sort of test may not be useful:
  - For the air conditioner data there is no significant evidence for rejecting a Gamma model in favor of a log normal model.
  - There is also no significant evidence for rejecting a log normal model in favor of a Gamma model.

Both p values are quite large.

Nonparametric Permutation Tests: Correlation

- A test for independence of bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$ can be based on the sample correlation

$$T = \rho(\widehat{F})$$

- Under the null hypothesis of independence the two sets of order statistics are sufficient.
- The conditional null distribution of T, given the order statistics, is the distribution of

$$T^* = \rho((X_{(1)}, Y_1^*), \ldots (X_{(n)}, Y_n^*))$$

  where $Y_1^*, \ldots, Y_n^*$ is drawn uniformly from all permutations of $Y_1, \ldots, Y_n$.
- The null distribution of T can be simulated by randomly selecting permutations and computing T^* values.
- Any other measure of dependence can be used as well, for example a rank correlation.

Semi-Parametric Bootstrap Test: Equality of Means

- Suppose we have data

  $$Y_{ij} = \mu_i + \sigma_i \varepsilon_{ij}, \quad j = 1, \dots, n_i; i = 1, \dots, k$$

  and we assume the ε_{ij} are independent from a common distribution G.

- To test $H_0 : \mu_i = \cdots = \mu_k$ against a general alternative we can use

  $$T = \sum_{i=1}^{k} w_i (\overline{Y}_i - \widehat{\mu}_0)^2$$

  with

  $$\begin{aligned} \widehat{\mu}_0 &= \sum w_i \overline{Y}_i / \sum w_i \\ w_i &= n_i / S_i^2 \end{aligned}$$

- The null distribution would be approximately χ^2_{k-1} for large sample sizes.

- The null model studentized residuals are

  $$e_{ij} = \frac{Y_{ij} - \widehat{\mu}_0}{\sqrt{\widehat{\sigma}_{i0}^2 - (\sum w_i)^{-1}}}$$

  with

  $$\widehat{\sigma}_{i0}^2 = (n_i - 1) S_i^2 / n_i + (\overline{Y}_i - \widehat{\mu}_0)^2$$

- The bootstrap simulates data sets

  $$Y_{ij}^* = \widehat{\mu}_0 + \widehat{\sigma}_{i0} \varepsilon_{ij}^*$$

  with the ε_{ij}^* sampled with replacement from the pooled null studentized residuals.

Fully Nonparametric Bootstrap Tests: Comparing Two Means

- Suppose Y_{ij}, $i = 1,2$, $j = 1,\ldots,n_i$, are independent with $Y_{ij} \sim F_i$ and means μ_i.
- We want to test $H_0 : \mu_1 = \mu_2$.
- Nonparametric maximum likelihood estimates F_1 and F_2 by discrete distributions concentrated on the observed data values.
- A certain nonparametric maximum likelihood argument suggests that the maximum likelihood probabilities, under the constraint of equal means, are of the form

$$\widehat{p}_{1j,0} = \frac{\exp(\lambda y_{1j})}{\sum_{k=1}^{n_1} \exp(\lambda y_{1k})}$$
$$\widehat{p}_{2j,0} = \frac{\exp(-\lambda y_{2j})}{\sum_{k=1}^{n_2} \exp(-\lambda y_{2k})}$$

  with λ chosen numerically to satisfy

$$\frac{\sum y_{1j} \exp(\lambda y_{1j})}{\sum \exp(\lambda y_{1j})} = \frac{\sum y_{2j} \exp(-\lambda y_{2j})}{\sum \exp(-\lambda y_{2j})}$$

  These are called *exponential tilts* of the empirical distribution.

- The tilted bootstrap two-sample comparison:
  - Generate Y_{1j}^*, $j = 1,\ldots,n_1$ by sampling from y_{1j} with weights $\widehat{p}_{1j,0}$
  - Generate Y_{2j}^*, $j = 2,\ldots,n_2$ by sampling from y_{2j} with weights $\widehat{p}_{2j,0}$
  - Compute $T^* = \overline{Y}_2^* - \overline{Y}_1^*$
  - Repeat R times

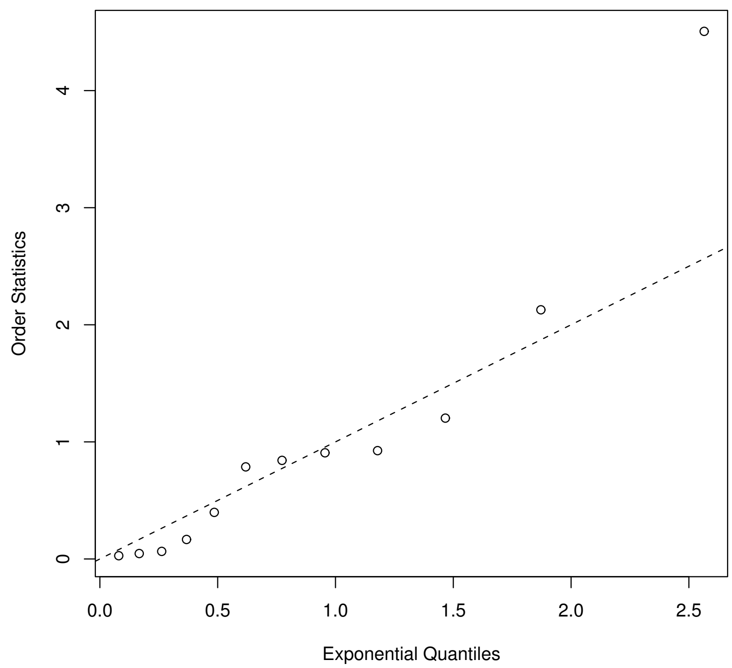
  Then compute a bootstrap p-value, say

$$p = \frac{1 + \#\{T_i^* \geq T\}}{R+1}$$

- It is a good idea to look at the two tilted distributions to make sure they are not too silly.

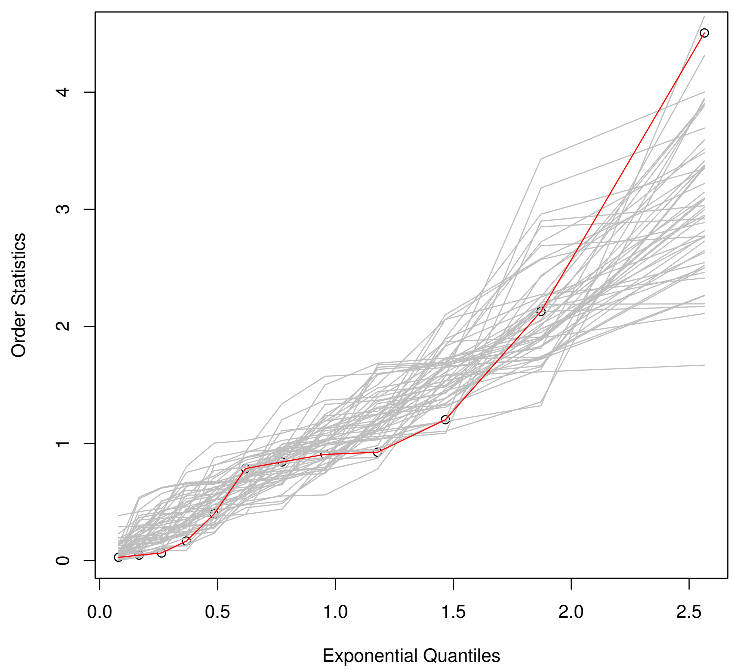
Graphical Tests

- Bootstrapping and resampling ideas can be used to help calibrate graphical procedures.
- As an example, a plot of $X_{(i)}/\overline{X}$ against the quantiles of a unit exponential distribution can be used to asses the appropriateness of the exponential model for the airconditioner data:



- The dashed line is the theoretical line for large samples.
- Is the departure more than might be expected by chance for a sample of this size from an exponential distribution?

- A bootstrap approach:
  - Generate a sample $Y_1^*, \dots, Y_n^*$ from an exponential distribution (the mean does not matter)
  - plot $Y_{(i)}^*/\overline{Y}^*$ against the unit exponential quantiles $-\log(1-i/(n+1))$
  - Repeat and add to the plot
- For the air conditioner data with $R = 50$ replications this produces



- Some references on similar ideas:
  - Andreas Buja (1999) Talk given at the Joint Statistics Meetings 1999, on the possibility of valid inference in exploratory data analysis, with Di Cook: Inference for Data Visualization
  - Andrew Gelman (2004) “Exploratory data analysis for complex models.” *Journal of Computational and Graphical Statistics*.
  - Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja (2010) “Graphical inference for infovis.” *IEEE Transactions on Visualization and Computer Graphics*