The Bootstrap, Resampling Procedures, and Monte Carlo Techniques

Herwig Friedl
Graz University of Technology/Austria

25th May, 2005

Outline

- Spirit and Principle of the Bootstrap
- Estimating Bias & Standard Error
- Bootstrapping the Bootstrap (Iterating the Principle)
- Hypothesis Tests
- Linear Regression Models
- Generalized Linear Models (if timing permits?)

The Bootstrap Principle

Efron (1979), Efron & Tibshirani (1986):

$$\begin{tikzcd}
P \arrow[rr] \arrow[dr] & & \mathbf{y} \arrow[dl] & \Rightarrow & \hat{P} \arrow[rr] \arrow[dr] & & \mathbf{y}^* \arrow[dl] \\
& R(\mathbf{y}, P) & & & & R(\mathbf{y}^*, \hat{P}) &
\end{tikzcd}$$

- unknown probability mechanism (statistical model) P
- sample (observed data) $\mathbf{y} = (y_1, \ldots, y_n)$
- random variable $R(\mathbf{y}, P)$, which possibly depends on both, the data and the unknown P.

The Real World (left triangle) is described/estimated by the Bootstrap World (right triangle)

E.g., the expectation of $R(\mathbf{y}, P)$ is estimated by the bootstrap expectation of $R(\mathbf{y}^*, \hat{P})$

The double arrow indicates the crucial step in applying the bootstrap

The bootstrap 'estimates'

1) P by means of the data $\mathbf{y}$

2) distribution of $R(\mathbf{y}, P)$ through the conditional distribution of $R(\mathbf{y}^*, \hat{P})$, given $\mathbf{y}$

Spirit of the Bootstrap

Use sample behavior of the triple

$$(\hat{P}, \mathbf{y}^*, R(\mathbf{y}^*, \hat{P})),$$

to mimic the one of $(P, \mathbf{y}, R(\mathbf{y}, P))$,

where the relationship between

$$\hat{P}, \mathbf{y}^* \text{ and } R(\mathbf{y}^*, \hat{P})$$

has to equal that between

$$P, \mathbf{y} \text{ and } R(\mathbf{y}, P)$$

How to estimate P (iid)

- parametric (known likelihood): assume that $y_i \overset{iid}{\sim} F(\theta)$

$$\hat{P} = F(\hat{\theta}),$$

- nonparametric (unknown likelihood):

$$\hat{P} = \hat{F} : \qquad \text{puts mass } 1/n \text{ at every } y_i$$

Interpretation: draw a resample $y_1^*, \ldots, y_n^*$ again of size n with replacement from the original data $y_1, \ldots, y_n$.

Remarks:

The concept was introduced by Prof. Bradley Efron, Stanford University, in his 1977 Rietz lecture

The bootstrap is a resampling procedure (as the Jackknife or as cross-validation).

In Efron (1979), acknowledge part: his personal favorite name actually was **Shotgun**, *which, to paraphrase Tukey, can blow the head off any problem if the statistician can stand the resulting mess!*

In German language: **Muenchhausen-Trick** instead of pull strap.

Measures of Statistical Error

Error Measure:

characteristic of the sampling distribution

$$H_F(r) = P(R(\mathbf{y}, F) \le r)$$

Estimator:

respective characteristic of the estimated sampling distribution

$$H_{\hat{F}}(r) = P_*(R(\mathbf{y}^*, \hat{F}) \le r|\mathbf{y})$$

Choices of R:

Estimation of Bias: For a statistic $\hat{\theta}(\mathbf{y})$ and a parameter $\theta(F)$, let

$$R(\mathbf{y}, F) = \hat{\theta}(\mathbf{y}) - \theta(F)\,.$$

The bias of $\hat{\theta}$ for estimating θ is

$$\text{bias}(F) = \mathsf{E}_F(R(\mathbf{y}, F)) = \mathsf{E}_F(\hat{\theta}(\mathbf{y})) - \theta(F)\,.$$

The bootstrap estimate of bias is

$$\text{bias}(\hat{F}) = \mathsf{E}_{\hat{F}}(R(\mathbf{y}^*, \hat{F})|\mathbf{y}) = \mathsf{E}_{\hat{F}}(\hat{\theta}(\mathbf{y}^*)|\mathbf{y}) - \theta(\hat{F})\,.$$

Estimation of Standard Error: let $R(\mathbf{y}, F) = \hat{\theta}(\mathbf{y})$

$$\begin{aligned}
\Big(\mathsf{var}_F(R(\mathbf{y}, F))\Big)^{1/2} &= \Big(\mathsf{var}_F(\hat{\theta}(\mathbf{y}))\Big)^{1/2} \\
\Big(\mathsf{var}_{\hat{F}}(R(\mathbf{y}^*, \hat{F}))\Big)^{1/2} &= \Big(\mathsf{var}_{\hat{F}}(\hat{\theta}(\mathbf{y}^*)|\mathbf{y})\Big)^{1/2}
\end{aligned}$$

Mean Squared Error:

$$\mathsf{MSE}(\hat{\theta}(\mathbf{y})) = \mathsf{var}(\hat{\theta}(\mathbf{y})) + \mathsf{bias}^2(\hat{\theta}(\mathbf{y}))\,.$$

Example: Nonparametric Bootstrap $y_i \overset{iid}{\sim} F$ (unknown), $y_i^* \overset{iid}{\sim} \hat{F}$

$$
\begin{aligned}
\mathsf{E}_*(y_1^*|\mathbf{y}) &= \sum_{i=1}^{n} y_i P(y_1^* = y_i) = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{\mathbf{y}} \\
\mathsf{var}_*(y_1^*|\mathbf{y}) &= \mathsf{E}_*\Big((y_1^* - \overline{\mathbf{y}})^2|\mathbf{y}\Big) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{\mathbf{y}})^2 = s^2
\end{aligned}
$$

$\mathsf{E}_*(\cdot)$ and $\mathsf{var}_*(\cdot)$ are the bootstrap moments w.r.t. the edf $\hat{F}$

Example cont'd:
assess $\hat{\theta}(\mathbf{y}) = \overline{\mathbf{y}}$ as an estimate of $\mu(F) = \int y \, dF(y)$

Bias:

$$\begin{aligned} \mathsf{E}(\overline{\mathbf{y}}) - \mu(F) &= 0 \\ \mathsf{E}_*(\overline{\mathbf{y}}^*|\mathbf{y}) - \mu(\hat{F}) &= \mathsf{E}_*(y_1^*|\mathbf{y}) - \overline{\mathbf{y}} = 0 \end{aligned}$$

Standard Error:

$$\begin{aligned} \mathsf{var}(\overline{\mathbf{y}}) &= \sigma^2/n \\ \mathsf{var}_*(\overline{\mathbf{y}}^*|\mathbf{y}) &= \mathsf{var}_*(y_1^*|\mathbf{y})/n = s^2/n \end{aligned}$$

Example cont'd:
Now assess $\hat{\theta}(\mathbf{y}) = s^2$ as an estimate of $\sigma^2(F) = \int (y - \mu)^2 \, dF(y)$

Bias:

$$\begin{aligned} \mathsf{E}(s^2) - \sigma^2(F) &= -\sigma^2/n \\ \mathsf{E}_*(s^{*2}|\mathbf{y}) - \sigma^2(\hat{F}) &= -s^2/n \end{aligned}$$

Drawback: usually it's pretty hard to find explicitly the bootstrap distribution or even the bootstrap moments.

Bootstrapping the Bootstrap

E.g., bias correction of bootstrap calculations:

Estimator $T = t(\hat{F})$ for $\theta = t(F)$ has bias

$$\beta = \mathsf{bias}(F) = \mathsf{E}(T) - \theta = \mathsf{E}[t(\hat{F})|F] - t(F)$$

Bootstrap estimate of this bias is

$$B = \mathsf{bias}(\hat{F}) = \mathsf{E}_*(T^*) - T = \mathsf{E}_*[t(\hat{F}^*)|\hat{F}] - t(\hat{F})$$

$\hat{F}^*$ is the edf of the bootstrap sample $Y_1^*, \ldots, Y_n^*$ (drawn from $\hat{F}$).

As with $T = t(\hat{F})$ itself, so with $B = \mathsf{bias}(\hat{F})$: the bias can be estimated using the bootstrap. Write

$$c(F) = \mathsf{E}(B|F) - \mathsf{bias}(F)$$

then the simple bootstrap estimate is

$$\begin{aligned} c(\hat{F}) &= \mathsf{E}_*(B^*|\hat{F}) - \mathsf{bias}(\hat{F}) \\ &= \mathsf{E}_*\Big\{\mathsf{E}_{**}[t(\hat{F}^{**})|\hat{F}^*] - t(\hat{F}^*)|\hat{F}\Big\} - \Big\{\mathsf{E}_*[t(\hat{F}^*)|\hat{F}] - t(\hat{F})\Big\} \\ &= \mathsf{E}_*\Big\{\mathsf{E}_{**}(T^{**})\Big\} - 2\mathsf{E}_*(T^*|\hat{F}) + T \end{aligned}$$

$\hat{F}^{**}$ is the edf of a resample $Y_1^{**}, \ldots, Y_n^{**}$ drawn from $\hat{F}^*$.

Since there are 2 levels of bootstrapping here, this procedure is also called **nested** or **double bootstrap**.

The adjusted estimate of the bias of T is

$$B_{\mathsf{adj}} = \mathsf{bias}(\hat{F}) - c(\hat{F})\,.$$

Example: $T = n^{-1} \sum_i (y_i - \overline{y})^2$ to estimate var$(Y) = \sigma^2$. Thus

$$\beta = \text{bias}(F) = -\sigma^2/n\,,$$

which the bootstrap estimates by

$$B = \text{bias}(\hat{F}) = -T/n\,.$$

The bias of this bias estimate is $\mathsf{E}(B) - \beta = \sigma^2/n^2$,

which the bootstrap estimates by $c(\hat{F}) = T/n^2$.

Thus, the adjusted bias estimate is $B_{\text{adj}} = -T/n - T/n^2$.

Improvement: $\mathsf{E}(B_{\text{adj}}) = \beta(1 + n^{-2})$, whereas $\mathsf{E}(B) = \beta(1 + n^{-1})$.

Bootstrap Distribution

- direct calculation (often impossible)
- asymptotical methods (see e.g. Hall, 1992)
- Monte Carlo approximation (always a good choice ;)

For $b = 1, \ldots, B$ (B large) resample $\mathbf{y}_b^*$ from $\hat{P}$ and calculate $t(\mathbf{y}_b^*)$

Apply the Law of Large Numbers in the sense that

$$\mathsf{E}_{MC}(t(\mathbf{y}^*)|\mathbf{y}) = \bar{t}(\mathbf{y}^*) = \frac{1}{B}\sum_{b=1}^{B} t(\mathbf{y}_b^*) \xrightarrow{a.s.} \mathsf{E}_*(t(\mathbf{y}^*)|\mathbf{y})$$

Thus, $\mathrm{var}_*(t(\mathbf{y}^*)|\mathbf{y})$ is approximated by

$$\mathrm{var}_{MC}(t(\mathbf{y}^*)|\mathbf{y}) = \frac{1}{B}\sum_{b=1}^{B} \left(t(\mathbf{y}_b^*) - \bar{t}(\mathbf{y}^*)\right)^2 \xrightarrow{a.s.} \mathrm{var}_*(t(\mathbf{y}^*)|\mathbf{y})$$

Remarks:

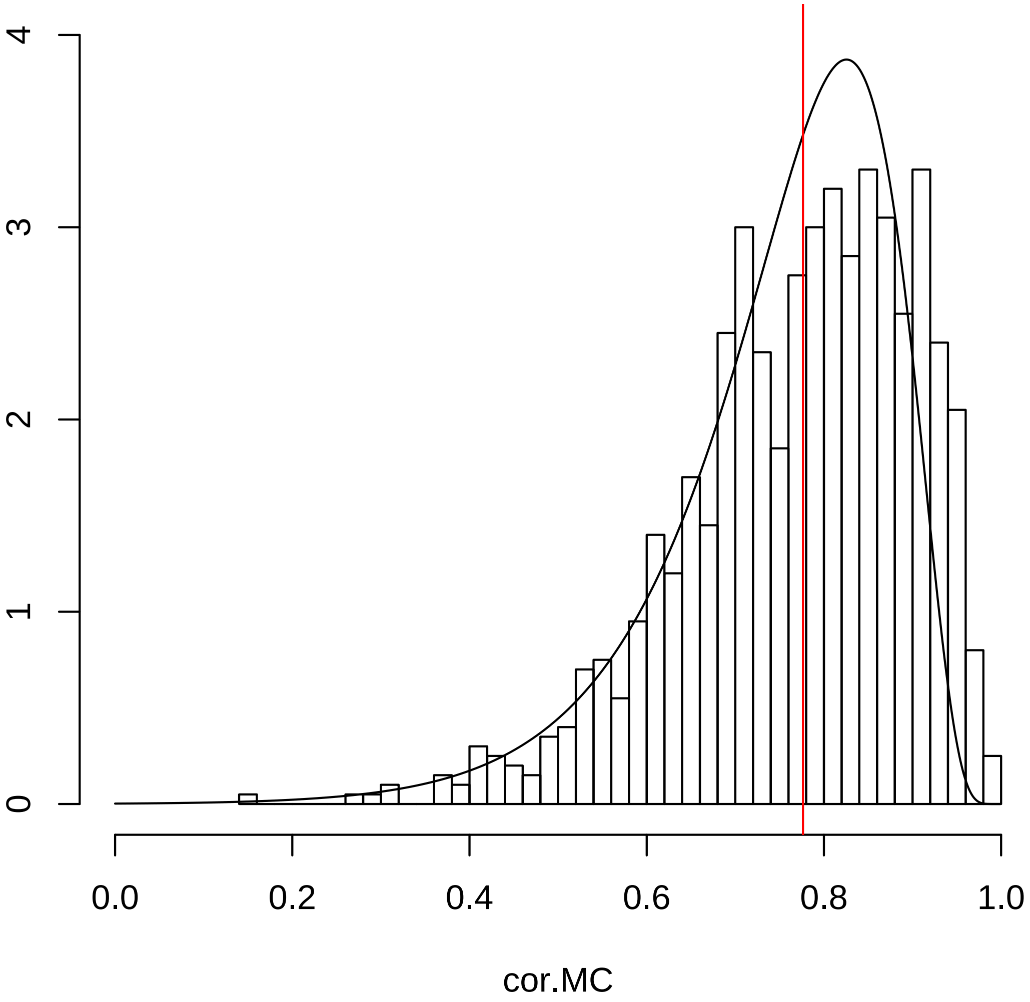
- MC results are used to approximate BT quantities.
- BT quantities are desired as estimates of population charact's.
- sometimes MC results are misleadingly called BT estimates
- generating large MC samples can be computationally expensive!

- Histogram of the MC resamples approximates the BT density (estimates the unknown density)
- empirical MC moments approximate the respective BT moments (estimates the unknown moments)

Using R for MC Simulations

Example 1: correlation coefficient, $n = 15$

```
> library(bootstrap);  data(law)
> R <- cor(law$LSAT, law$GPA);  R
  0.7763745

> B <- 1000;  cor.MC <- 1:B
> for (b in 1:B) {
+   i <- sample(15, replace=TRUE)
+   cor.MC[b] <- cor(law$LSAT[i], law$GPA[i])
+ }
> mean(cor.MC);  sd(cor.MC)
   Mean     StdDev
  0.7716   0.1309
```


0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
cor.MC

Example 2: BHCG blood serum levels for 54 breast cancer patients

```
> lev <- c(0.1, 0.1, ..., 4.4, 4.5, 6.4, 9.4)
> mean(lev);  mean(lev, trim=0.25)
[1] 2.3185  2.2393
```

We want to estimate the **true mean** $\mu = \mathsf{E}_F(y)$ of this population, using $\hat{\theta}$, the 25% trimmed mean (because of 2 large obs's 6.4, 9.4).

```
> for (b in 1:B) {
+   i <- sample(54, replace=TRUE)
+   m.MC[b,1] <- mean(lev[i])
+   m.MC[b,2] <- mean(lev[i], trim=0.25)
+ }
> sd(m.MC)
[1] 0.2116334 0.1617231
```

Thus, the standard error for $\hat{\theta}$ is much smaller.

Example 2 cont'd: consider the t-like statistic

$$R(\mathbf{y}, F) = \frac{\hat{\theta}(\mathbf{y}) - \mu(F)}{\widehat{iqr}(\mathbf{y})}$$

What about confidence intervals for $\mu(F)$?

Suppose we know the 5th and 95th percentiles of R, say $\rho^{(0.05)}(F)$ and $\rho^{(0.95)}(F)$, where

$$P_F(R(\mathbf{y}, F) \leq \rho^{(\alpha)}(F)) = \alpha\,.$$

This gives a central 90% interval for $\mu(F)$,

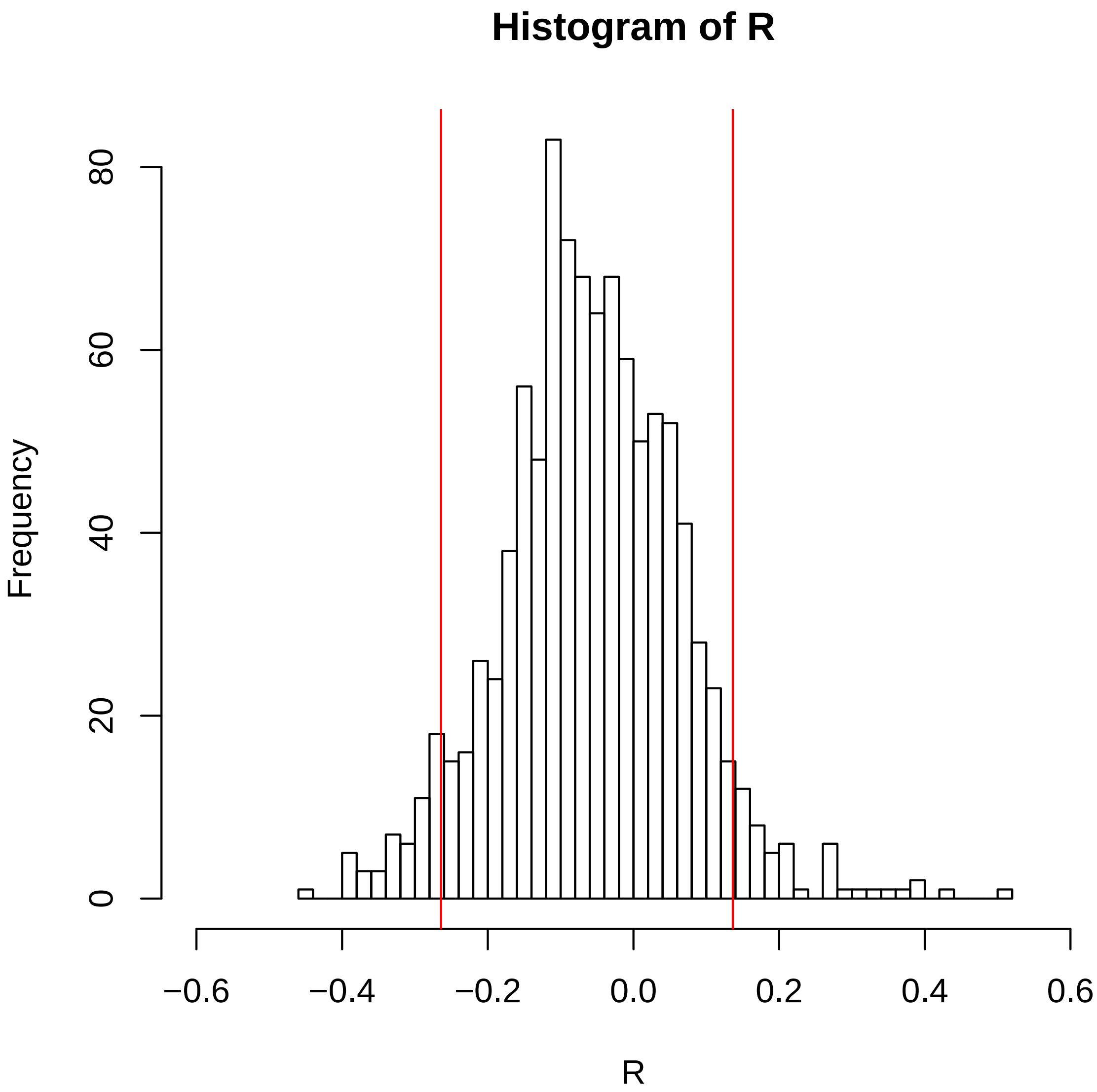
$$\mu(F) \in [\hat{\theta}(\mathbf{y}) - \widehat{iqr}(\mathbf{y}) \cdot \rho^{(0.95)}, \hat{\theta}(\mathbf{y}) - \widehat{iqr}(\mathbf{y}) \cdot \rho^{(0.05)}]$$

Because $\rho^{(\alpha)}(F)$ is unknown, a bootstrap sample gives

$$R(\mathbf{y}^*, \hat{F}) = \frac{\hat{\theta}(\mathbf{y}^*) - \mu(\hat{F})}{\widehat{iqr}(\mathbf{y}^*)}$$

MC estimate of $P_{\hat{F}}(R(\mathbf{y}^*, \hat{F}) < \rho)$ is $\#(R(\mathbf{y}_b^*, \hat{F}) < \rho)/B$

```
> for (b in 1:B) {
+   i <- sample(54, replace=TRUE)
+   R[b] <- (mean(lev[i], trim=0.25) - mean(lev))/IQR(lev[i])
+ }
> mean(lev, trim=0.25) - IQR(lev)*sort(R)[950]
[1] 2.05
> mean(lev, trim=0.25) - IQR(lev)*sort(R)[ 50]
[1] 2.60
```


Histogram of R
Frequency
0
20
40
60
80
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
R

This interval $(2.05, 2.60)$ is smaller than the usual t interval

$$\overline{\mathbf{y}} \pm \sqrt{\widehat{\mathsf{var}}(\mathbf{y})/n} \cdot t_{0.95,53} = (1.97, 2.67)$$

MC Resampling Libraries:

Efron & Tibshirani (1993): `bootstrap`

Davison & Hinkley (1997): `boot`

Hypothesis Tests

Question: How to use resampling methods for significance tests in parametric & nonparametric settings.

Simplest situation: simple null hypothesis H_0 completely specifies the distribution of the data; e.g. $H_0 : F = F_0$, where F_0 contains **no unknown parameters**; $exponential\ with\ \lambda = 1.$

Situation in practice: composite null hypothesis H_0; some aspects of F **unknown** when H_0 is true; $normal\ with\ \mu = 1$ (σ^2 not specified).

Test statistic T measures discrepancy between data and H_0. Convention: large values of T are evidence against H_0.

H_0 simple, $T = t$ observed: level of evidence against H_0 measured by the significance probability, the **P-value**

$$p = \Pr(T \geq t | H_0) \, .$$

Critical value t_p for t, associated with testing at level p: if $t \geq t_p$ we reject H_0 at level p. Thus, $\Pr(T \geq t_p | H_0) = p$.

p is called **error rate** and $\{(y_1, \ldots, y_n) : t \geq t_p\}$ level p **critical region** of the test. The distribution of T under H_0 is called the **null distribution**.

Choice of test statistic in parametric setting:

Explicit form of sampling distribution is known, with a finite number of unknown param's. H_0 specifies relationships between param's.

Likelihood function: $L(\theta) = f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n|\theta)$.

When $H_0 : \theta = \theta_0$, $H_A : \theta = \theta_A$ are both simple, the best test statistic is the likelihood ratio $T = L(\theta_A)/L(\theta_0)$.

Different situation when we test **goodness of fit** of the parametric model. This can be done by embedding the model into a larger model (add'al param's), to check departures from the original model.

Choice of test statistic in nonparametric setting:

No particular form specified for the sampling distribution. Choice of T is less clear. Usually T based on a **statistical function** $s(\hat{F})$, for which H_0 specifies a value.

We test $H_0 : X, Y$ are independent, sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. $\rho = \text{corr}(X, Y) = s(F)$ measures dependence, and $\rho = 0$ under H_0. If H_A is *positive dependence* (one-sided), we can use

$$T = s(\hat{F})$$

the sample correlation. If H_A is *dependence*, then

$$T = s^2(\hat{F})\,.$$

Conditional tests:

H_0 is often **composite**, leaves param's **unknown**, F not completely specified. P-value not well defined, since $\mathrm{Pr}(T \geq t|F)$ may depend on which $F \in H_0$ is taken.

a) choose T, so that its distribution is the same for all $F \in H_0$ (e.g. Student-t test for normal mean with σ^2 unknown).

b) Eliminate param's which are unknown under H_0, by **conditioning** on the **sufficient statistic** under H_0. Let S denote this statistic, then the conditional P-value is defined by

$$p = \mathrm{Pr}(T \geq t|S = s, H_0).$$

(e.g., Fisher's exact test for 2×2 tables, Student-t test)

A less satisfactory approach (which can give good approximations) is to estimate F by a cdf $\hat{F}_0$, which satisfies H_0 and then calculate

$$p = \Pr(T \geq t|\hat{F}_0).$$

Typically this will not satisfy the definition of the error rate exactly.

Pivot tests:

For $H_0 : \theta = \theta_0$, use equivalence between tests and confidence sets. If θ_0 is outside a $1 - \alpha$ confidence set for θ, then θ differs from θ_0 with P-value less than α. **Pivot tests** based on this equivalence.

Let T be an estimator for scalar θ with **estimated variance** V. Suppose that the studentized form

$$Z = (T - \theta)/V^{1/2}$$

is a **pivot**, meaning that its distribution is the same for all relevant F, and in particular for all θ (e.g. Student-t statistic).

For $H_0 : \theta = \theta_0$ vs $H_A : \theta > \theta_0$, and $z_0 = (t - \theta_0)/v^{1/2}$ observed

$$p = \Pr\Big((T - \theta_0)/V^{1/2} \geq (t - \theta_0)/v^{1/2} | H_0\Big).$$

But because Z is a pivot,

$$\Pr\Big(Z \geq (t - \theta_0)/v^{1/2} | H_0\Big) = \Pr\Big(Z \geq (t - \theta_0)/v^{1/2} | F\Big),$$

and therefore

$$p = \Pr(Z \geq z_0 | F).$$

No special null sampling distributions needed to resample from.

Distinguish param's of interest ψ and **nuisance** param's λ.
H_0 concerns only ψ. Thus, conditional p-value is independent of λ.

How to construct a general test statistic T?

Generalize the likelihood ratio and define

$$LR = \frac{\max_{H_A} L(\psi, \lambda)}{\max_{H_0} L(\psi, \lambda)} \,.$$

For $H_0 : \psi = \psi_0$ vs $H_A : \psi \neq \psi_0$, this is

$$LR = \frac{L(\hat{\psi}, \hat{\lambda})}{L(\psi_0, \hat{\lambda}_0)} = \frac{\max_{\psi,\lambda} L(\psi, \lambda)}{\max_{\lambda} L(\psi_0, \lambda)} \,.$$

Often $T = 2 \log LR \sim \chi^2_d$ under H_0 (approx.), where d is the dimension of ψ, so that

$$p \doteq \Pr(\chi^2_d \geq t)$$

independently of λ. Thus the LR is an approximate pivot.

Approximations for p exist in many cases (behavior for $n \to \infty$).

Resampling alternatives, if such approximations fail to give appropriate accuracy or do not exist at all.

Resampling for Parametric Tests

Monte Carlo Tests

Null distribution of T does not include nuisance parameters (conditioning). Often it is impossible to calculate the conditional P-value, but MC tests provide approximations to the full tests.

Basic MC test compares the observed t to R independent values of T, e.g. $t_1^*, \ldots, t_R^*$, obtained from samples, which are independently simulated under H_0.

Under H_0, all $R+1$ values $t, t_1^*, \ldots, t_R^*$ are equally likely values of T. Thus, if T is continuous,

$$\Pr(T < T_{(r)}^* | H_0) = \frac{r}{R+1}.$$

If exactly k of the simulated t^* values exceed t (and none equal it), the **MC P-value** is

$$p = \Pr(T \geq t | H_0) \doteq p_{\mathsf{mc}} = \frac{k+1}{R+1}\,.$$

If T is **continuous**, then the distribution of P_{mc} is uniform on $(1/(R+1), \ldots, R/(R+1), 1)$ under H_0 (error rate interpretation). The full test corresponds to $R \to \infty$.

If T is **discrete**, then repeated values of t^* can occur. If exactly l of the t^* values equal t, then

$$\frac{k+1}{R+1} \leq p_{\mathsf{mc}} \leq \frac{k+l+1}{R+1}\,.$$

We (have to) use the upper bound

$$p_{\mathsf{mc}} = \frac{1 + \#(t_r^* \geq t)}{R+1}\,.$$

Example: $n = 50$ counts of fir seedlings in 5 feet square quadrats.

0	1	2	3	4	3	4	2	2	1
0	2	0	2	4	2	3	3	4	2
1	1	1	1	4	1	5	2	2	3
4	1	2	5	2	0	3	2	1	1
3	1	4	3	1	0	0	2	7	0

Test the null that the data are iid Poisson(μ).

Concern: **overdispersion** relative to Poisson, var$(Y_i) = \psi\mu$, $\psi > 1$. Take dispersion index as test statistic

$$T = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{\overline{Y}}\,.$$

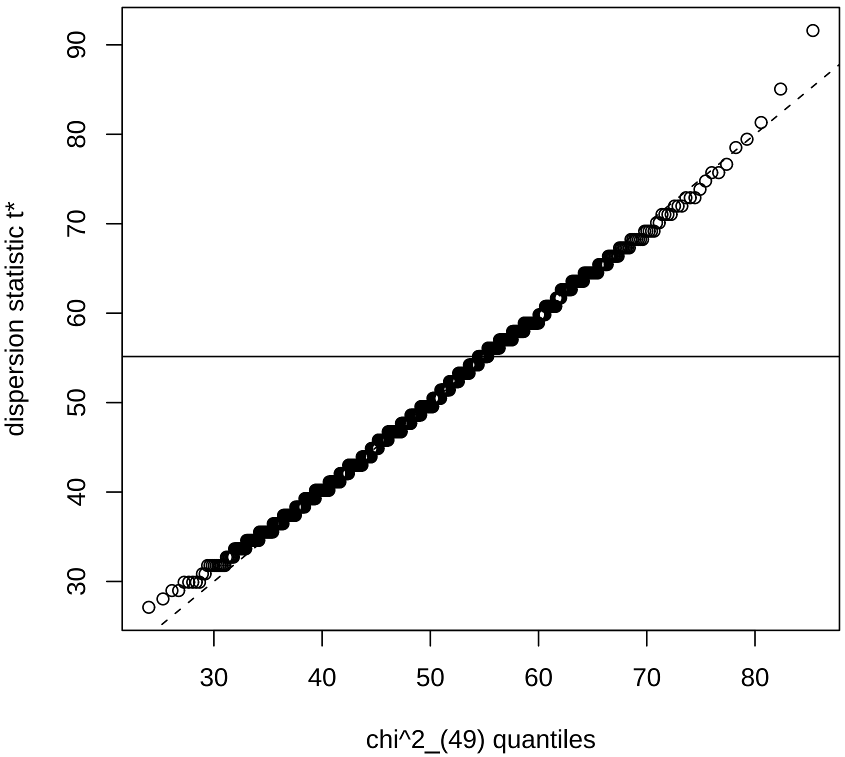
Under H_0, $S = \sum_{i=1}^{n} Y_i$ is **sufficient** for μ.

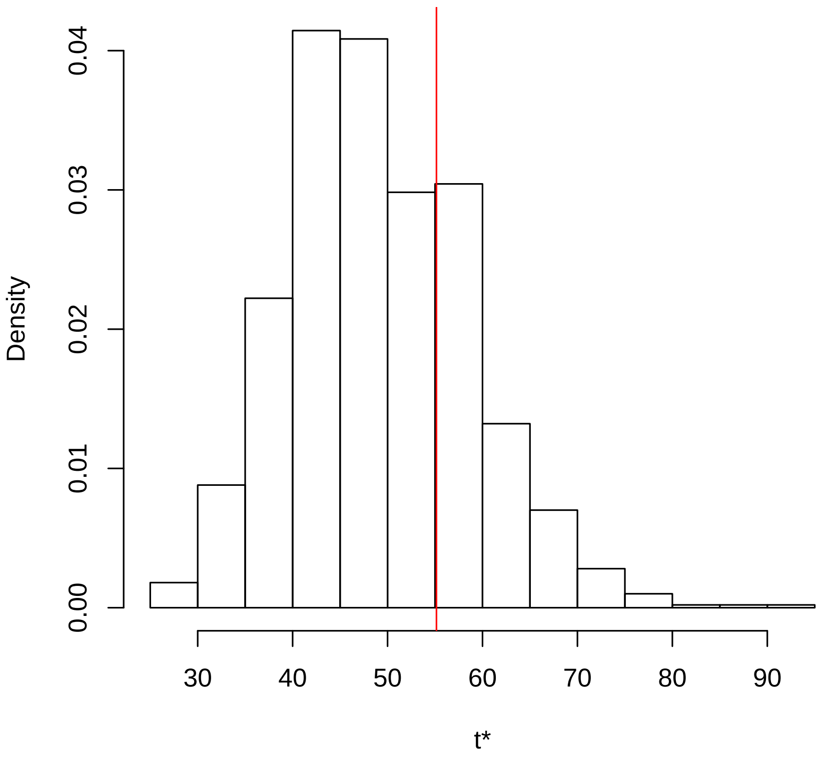
Conditional test: $(Y_1, \ldots, Y_n)|(S = s)$ **multinomial** with denominator s and n categories, each having probability $1/n$.
It is easy to simulate from this.

We further know that $T|(S = s) \overset{H_0}{\sim} \chi^2_{n-1}$ (approximately).

```
> library(boot);  attach(fir)
> fir.mle <- c(sum(fir$count), nrow(fir));  fir.mle    # s & n
[1] 107  50

> fir.fun <- function(data)   # test statistic t
+    ((nrow(data) - 1) * var(data$count))/mean(data$count)
> fir.fun(fir)
[1] 55.14953
```

```
> fir.gen <- function(data, mle) {
+   d <- data
+   y <- sample(x=mle[2], size=mle[1], replace=T)
+   d$count <- tabulate(y, mle[2]);  d
+   }
> fir.boot <- boot(fir, fir.fun, R=999, sim="parametric",
+                  ran.gen=fir.gen, mle=fir.mle)

> summary(fir.boot$t)
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 27.11   42.07  48.61 49.08   55.15 91.60

> pmc <- (sum(fir.boot$t>fir.boot$t0)+1)/(fir.boot$R+1);  pmc
[1] 0.249
> 1 - pchisq(fir.boot$t0, fir.mle[2]-1)
[1] 0.2534432
```


dispersion statistic t*
30
40
50
60
70
80
90
chi^2_(49) quantiles
30
40
50
60
70
80


Density
0.00
0.01
0.02
0.03
0.04
t*
30
40
50
60
70
80
90

Parametric Bootstrap Tests

If null distribution of T depends on nuisance param's, which cannot be conditioned away so MC tests cannot be applied. Fit $\hat{F}_0$ and calculate

$$p = \Pr(T \geq t|\hat{F}_0)\,.$$

In a parametric model test $H_0 : \psi = \psi_0$ with λ a nuisance parameter, $\hat{F}_0$ is the cdf of $f(y|\psi_0, \hat{\lambda}_0)$, with $\hat{\lambda}_0$ the MLE of λ when $\psi = \psi_0$.

If p cannot be computed, draw R iid samples $y_1^*, \ldots, y_n^*$ from $\hat{F}_0$ and calculate t_r^*. Significance probability is approximated by

$$p_{\text{boot}} = \frac{1 + \#\{t_r^* \geq t\}}{R+1}\,.$$

Example: (Separate Families Test)

Choose between alternative families $f_0(y|\eta)$ and $f_1(y|\zeta)$. Nuisance parameter $\lambda = (\eta, \zeta)$.

Indicator ψ with null value $\psi_0 = 0$ and alternative value $\psi_A = 1$. Likelihood ratio is

$$T = \frac{1}{n} \log \frac{L_1(\hat{\zeta})}{L_0(\hat{\eta})} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_1(y_i|\hat{\zeta})}{f_0(y_i|\hat{\eta})}\,.$$

$(\hat{\zeta}, \hat{\eta})$ are the MLE's under f_1 and f_0. If families are strictly separated (not nested), then the χ^2 approximation does not apply!

Generate R samples (size n) by random sampling from the fitted null model $f_0(y|\hat{\eta})$. For each sample calculate $\hat{\eta}^*$ and $\hat{\zeta}^*$ by maximizing the simulated log-likelihoods

$$\ell_1^*(\zeta) = \sum_{i=1}^{n} \log f_1(y_i^*|\zeta)\,, \qquad \ell_0^*(\eta) = \sum_{i=1}^{n} \log f_0(y_i^*|\eta)\,.$$

and compute the simulated log-likelihood ratio statistic

$$t^* = 1/n\{\ell_1^*(\hat{\zeta}^*) - \ell_0^*(\hat{\eta}^*)\}\,.$$

Example: Failure times of air-conditioning equipment ($n = 12$).

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

Plausible models for $y > 0$:

Gamma: $$f_0(y|\eta) = \frac{\kappa(\kappa y)^{\kappa-1}\exp(-\kappa y/\mu)}{\mu^\kappa \Gamma(\kappa)}$$

Lognormal: $$f_1(y|\zeta) = \frac{1}{\beta y} \ \phi\left(\frac{\log y - \alpha}{\beta}\right)$$

Gamma mean: $\hat{\mu} = \overline{y} = 108.1$
Gamma index: solves $\log(\hat{\kappa}) - d\log\Gamma(\hat{\kappa})/d\hat{\kappa} = \log(\overline{y}) - \overline{\log(y)}$
giving $\hat{\kappa} = 0.707$

Normal mean: $\hat{\alpha} = \overline{\log y} = 3.829$
Normal variance: $\hat{\beta}^2 = (n-1)s^2_{\log y}/n = 2.339.$

```
> data(aircondit);  attach(aircondit)

> gamma.estim(hours)
 $kappa: 0.7064932  $mu: 108.0833

> lognormal.estim(hours)
 $alpha: 3.828588  $beta2: 2.33853
```

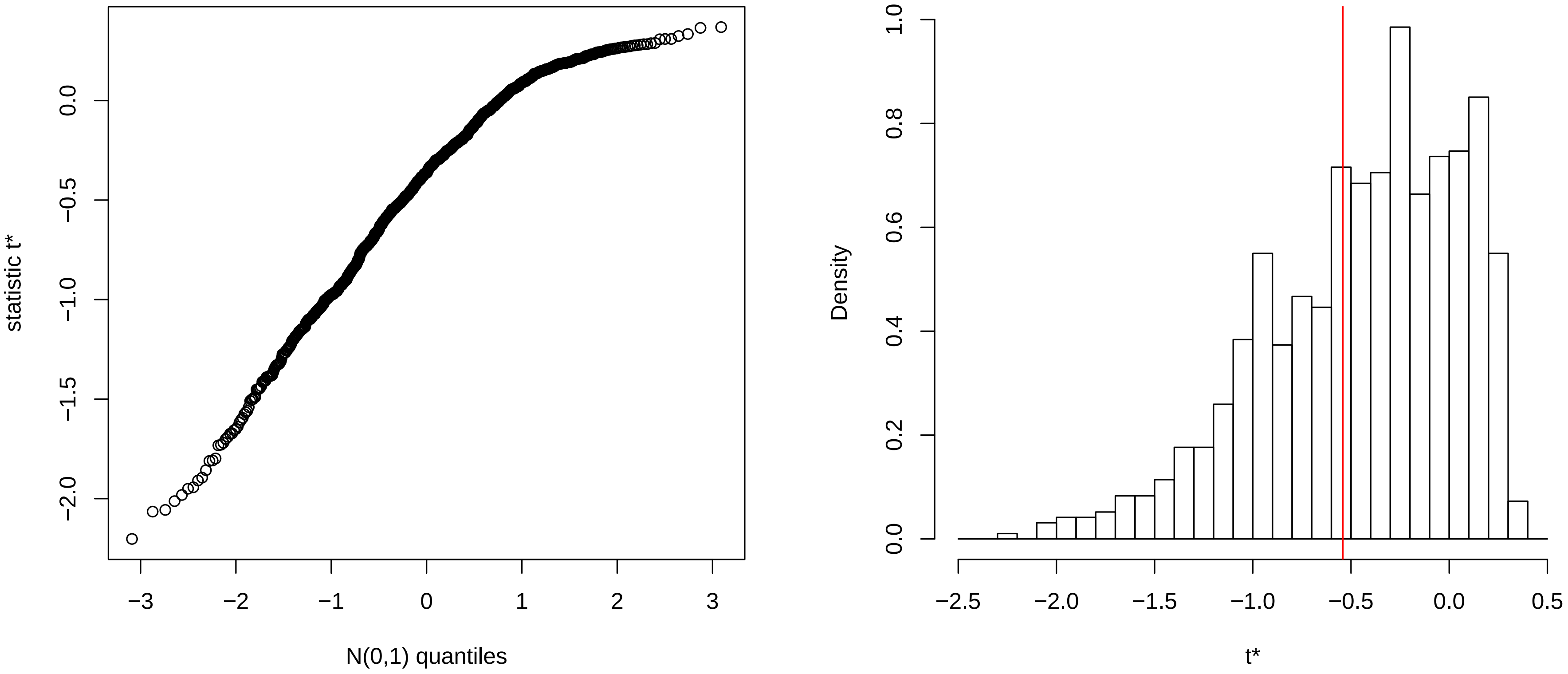
The observed test statistic is

$$\begin{aligned} t &= -\hat{\kappa}\Big(\log(\hat{\kappa}/\hat{\mu}) + \hat{\alpha} - 1\Big) - \log\Gamma(\hat{\kappa}) - \log(2\pi\hat{\beta}^2)/2 - 1/2 \\ &= -0.465 \end{aligned}$$

```
> air.mle <- c(gamma.estim(hours)$kappa, gamma.estim(hours)$mu)

> air.gen <- function(data, mle) {
+   d <- data
+   d$hours <- rgamma(nrow(data), mle[1], rate = mle[1]/mle[2])
+   d
+ }

> air.fun <- function(data) {
+   k <- gamma.estim(data$hours)$kappa
+   mu <- gamma.estim(data$hours)$mu
+   alpha <- lognormal.estim(data$hours)$alpha
+   beta2 <- lognormal.estim(data$hours)$beta2
+   -k*(log(k/mu)+alpha-1)-log(gamma(k))-log(2*pi*beta2)/2-1/2
+ }
```

```
> air.boot <- boot(aircondit, air.fun, R=999, sim="parametric",
+                  ran.gen=air.gen, mle=air.mle)

> summary(air.boot$t)
  Min.  1st Qu.  Median    Mean  3rd Qu.   Max. NA’s
 -2.20  -0.7515 -0.3627 -0.4409  -0.0422 0.3688   44

> (1+sum(air.boot$t>=air.boot$t0,na.rm=T))
+ /(air.boot$R+1-sum(is.na(air.boot$t)))
[1] 0.6310881
```

Histogram has fairly non-normal shape. Thus, a normal approximation will not be very accurate!


statistic t*
0.0
−0.5
−1.0
−1.5
−2.0
−3
−2
−1
0
1
2
3
N(0,1) quantiles
Density
1.0
0.8
0.6
0.4
0.2
0.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
t*

Graphical Tests

Popular in model checking:
(half-) normal plots of residuals, plots of Cook distances, ...

Reference shape is straight line. Detect deviations from null model.

Requires notion of the $plots\ probable\ variation$ under a null model.

Superimpose a **probable envelope** to which the original plot is compared.

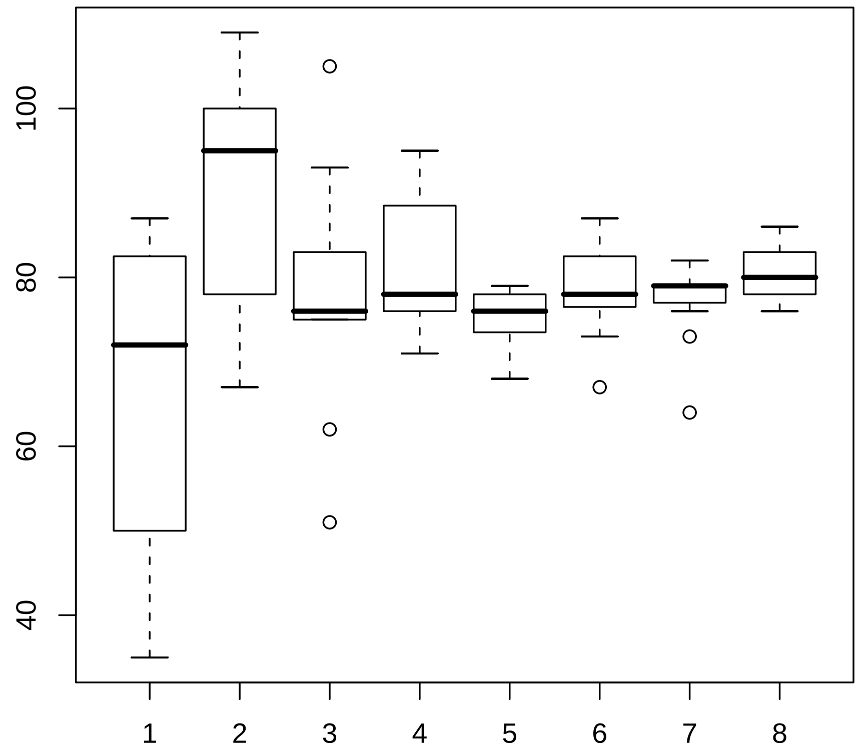
Probable envelope is obtained by MC or parametric resampling.

Suppose that the graph plots $T(a)$ vs $a \in \mathcal{A}$, a bounded set. The observed plot is $\{t(a) : a \in \mathcal{A}\}$.

In a normal plot $\mathcal{A}$ is a set of normal quantiles and the values of $t(a)$ are the ordered values of a sample.

The idea now is to compare $t(a)$ with the probable behavior of $T(a)$ for all $a \in \mathcal{A}$, when H_0 is true.

Example: (Normal plot of gravity data)



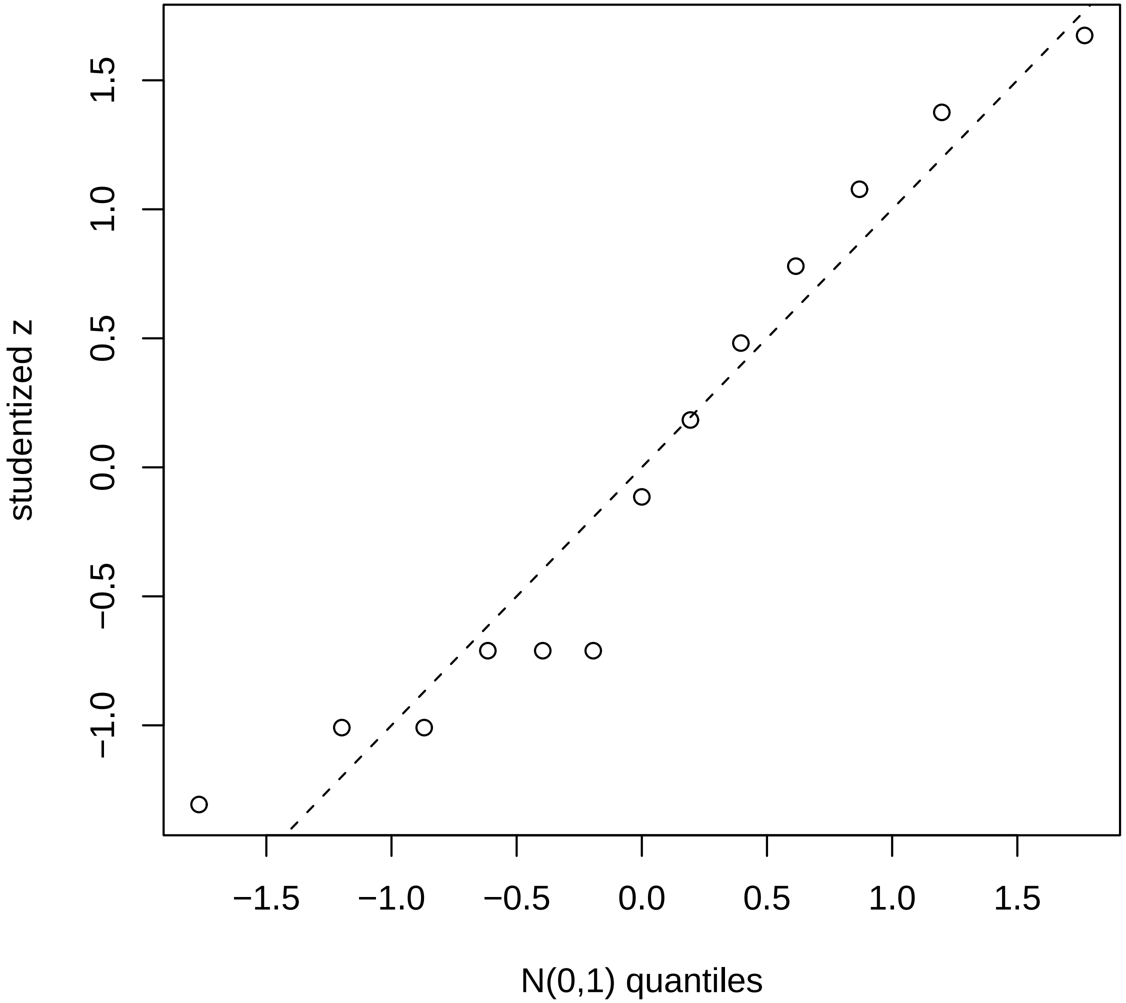
Check if the last series of $n = 13$ measurements of the acceleration due to gravity can be assumed normal.

Plot ordered studentized data values against N(0,1) quantiles, i.e.

$$z_{(i)} = (y_{(i)} - \overline{y})/s \qquad \text{vs.} \qquad a_i = \Phi^{-1}(i/(n+1))\,.$$

$\mathcal{A}$ is the set of normal quantiles, and $t(a_i) = z_{(i)}$.

```
> attach(gravity);  g <- grav$g[grav$series==8]
> grav.z <- (g-mean(g))/sqrt(var(g))
> qqnorm(grav.z, xlab="N(0,1) quantiles", ylab="studentized z")
> abline(0, 1, lty=2)
```



Dotted line is the expected pattern, approximately, and the question is whether or not the points deviate sufficiently from this to suggest that the sample is non-normal.

Assume the joint null distribution of $\{T(a) : a \in \mathcal{A}\}$ is free of nuisance param's (as for z_i's in normal plot). For any fixed a we can undertake $t(a)$ a MC test. For each of R indep. sets of data $y_1^*, \ldots, y_n^*$ (from null model) compute simulated plot

$$\{t^*(a) : a \in \mathcal{A}\}$$

Under H_0, $T(a), T_1^*(a), \ldots, T_R^*(a)$ are iid for any fixed a, so that

$$\Pr(T(a) < T_{(r)}^*(a) | H_0) = \frac{r}{R+1}\,.$$

applies. This leads to the one-sided MC P-value at given a, i.e.

$$p_{\textsf{mc}} = \frac{1 + \#(t_r^*(a) \geq t)}{R+1}\,.$$

Graphical test should rather look at all $a \in \mathcal{A}$ simultaneously. At each a compute lower and upper critical values (one-sided levels p) and plot them against a (critical curves).

Procedure: choose integers R and k with $k/(R+1) = p$ and calculate (from the R simulated plots) critical values

$$t^*_{(k)}(a), t^*_{(R+1-k)}(a)\,.$$

If $t(a)$ is outside, the one-sided P-value is at most p. A two-sided test, which rejects H_0 if $t(a)$ falls outside, has level $2p$.

The set of all lower and upper critical values defines the **test envelope**

$$\mathcal{E}^{1-2p} = \{[t^*_{(k)}(a), t^*_{(R+1-k)}(a)] : a \in \mathcal{A}\}$$

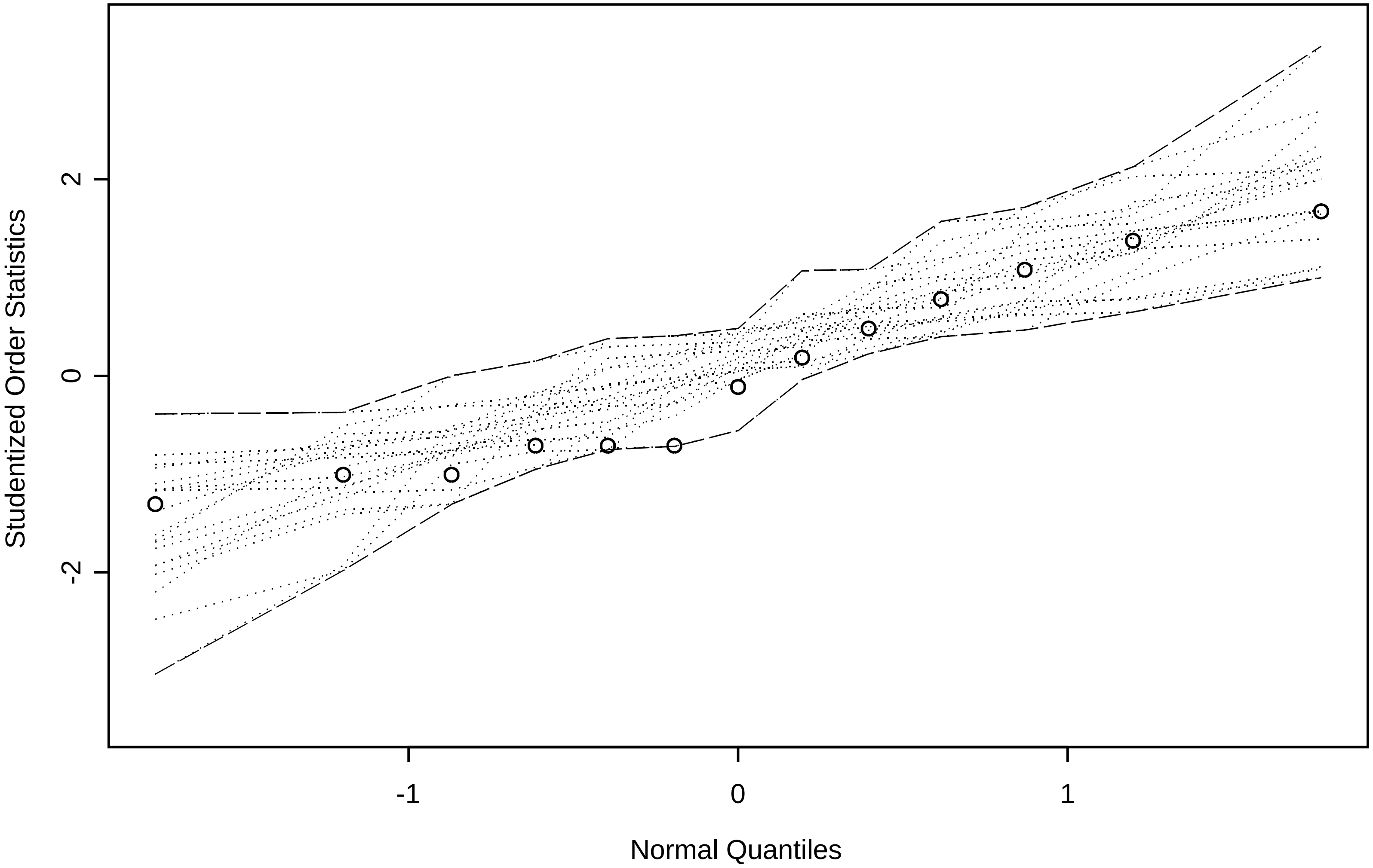
$Excursions$ of $t(a)$ outside $\mathcal{E}^{1-2p}$ give evidence against H_0.

Example normal plot cont'd:
For $p = 5$%, use at least $R = 19$ (take $k = 1$). Test envelope: lines connecting minima and maxima.

Since studentized values are plotted, simulation is done with $N(0, 1)$. Each sample $y^*_1, \ldots, y^*_n$ is studentized to give $z^*_i = (y^*_i - \overline{y}^*)/s^*$, whose ordered values are plotted against $a_i = \Phi^{-1}(i/(n+1))$.

Graphical test of normality:

```
> grav.gen <- function(dat, mle)  rnorm(length(dat))
> grav.qqboot <- boot(grav.z, sort, R=19, sim="parametric",
+                     ran.gen=grav.gen)
> grav.env <- envelope(boot.out=grav.qqboot,
+                      mat=grav.qqboot$t, level=0.90,
+                      index=1:ncol(grav.qqboot$t))
> grav.qq <- qqnorm(grav.z, plot=F)
> grav.qq <- lapply(grav.qq, sort)
> plot(grav.qq, ylim=c(-3.5,3.5),
+      ylab="Studentized Order Statistics",
+      xlab="Normal Quantiles", lty=1)
> lines(grav.qq$x, grav.env$point[1,], lty=4)
> lines(grav.qq$x, grav.env$point[2,], lty=4)
> for (i in 1:19) lines(grav.qq$x, grav.qqboot$t[i,], lty=18)
```

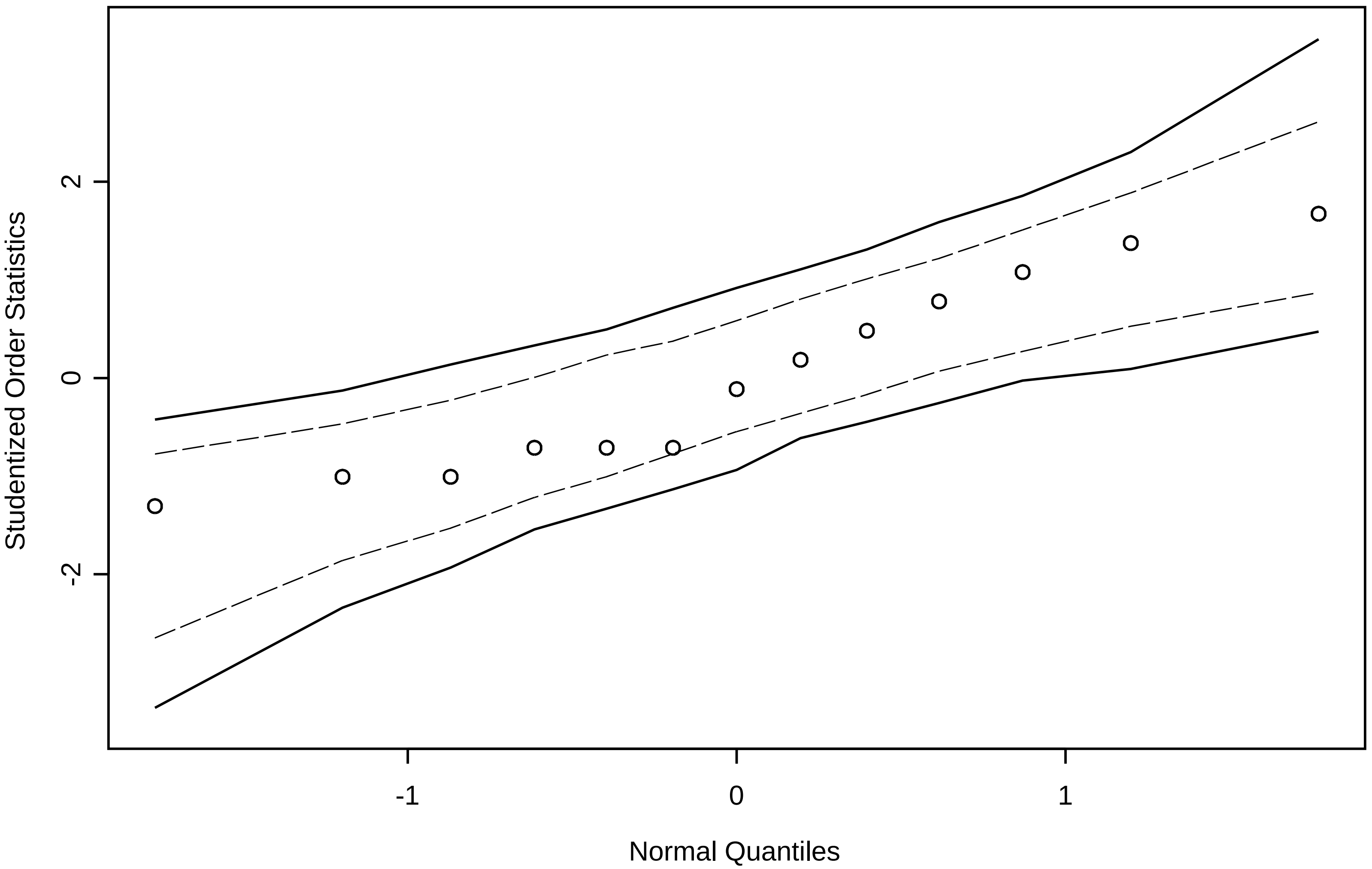

Studentized Order Statistics
-2
0
2
-1
0
1
Normal Quantiles

Levels p hold pointwise only. Chance that $\mathcal{E}^{1-2p}$ captures entire plot is smaller than $1 - 2p$.
Evidence against the null model, if 1 point falls outside? The chance for this is about $1/2$ (in contrast to the pointwise chance 0.1).

Overall error rate: (empirical approach) Given R simulated plots, compare $\{t_r^*(a), a \in \mathcal{A}\}$ to $\mathcal{E}_{-r}^{1-2p}$ (from the other $R-1$ plots). Repeat this simulated test for all r yields resample estimate

$$\frac{\#\{r : \{t_r^*(a), a \in \mathcal{A}\} \text{ exits } \mathcal{E}_{-r}^{1-2p}\}}{R}.$$

```
> grav.qqboot <- boot(grav.z, sort, R=999, sim="parametric",
+                     ran.gen=grav.gen)
> grav.env <- envelope(boot.out=grav.qqboot,
+                      mat=grav.qqboot$t, level=0.90,
+                      index=1:ncol(grav.qqboot$t))

> grav.env$k.pt # Quantiles used for pointwise env
  50 950
> grav.env$err.pt # pt, ov error rate for pt-env
  0.100 0.491
> grav.env$k.ov # Quantiles used for overall env
  7 993
> grav.env$err.ov # pt, ov error rate for ov-env
  0.014 0.095
> grav.env$err.nom # nom. error rates for pt- and ov-env
  0.1 0.1
```


Studentized Order Statistics
-2
0
2
-1
0
1
Normal Quantiles

Nonparametric Permutation Tests

Statistical methods which do not depend on specific parametric models (sign test, Wilcoxon test).

Simplest form of nonparametric resampling tests is the permutation test, which is a comparative test (compares edf's).

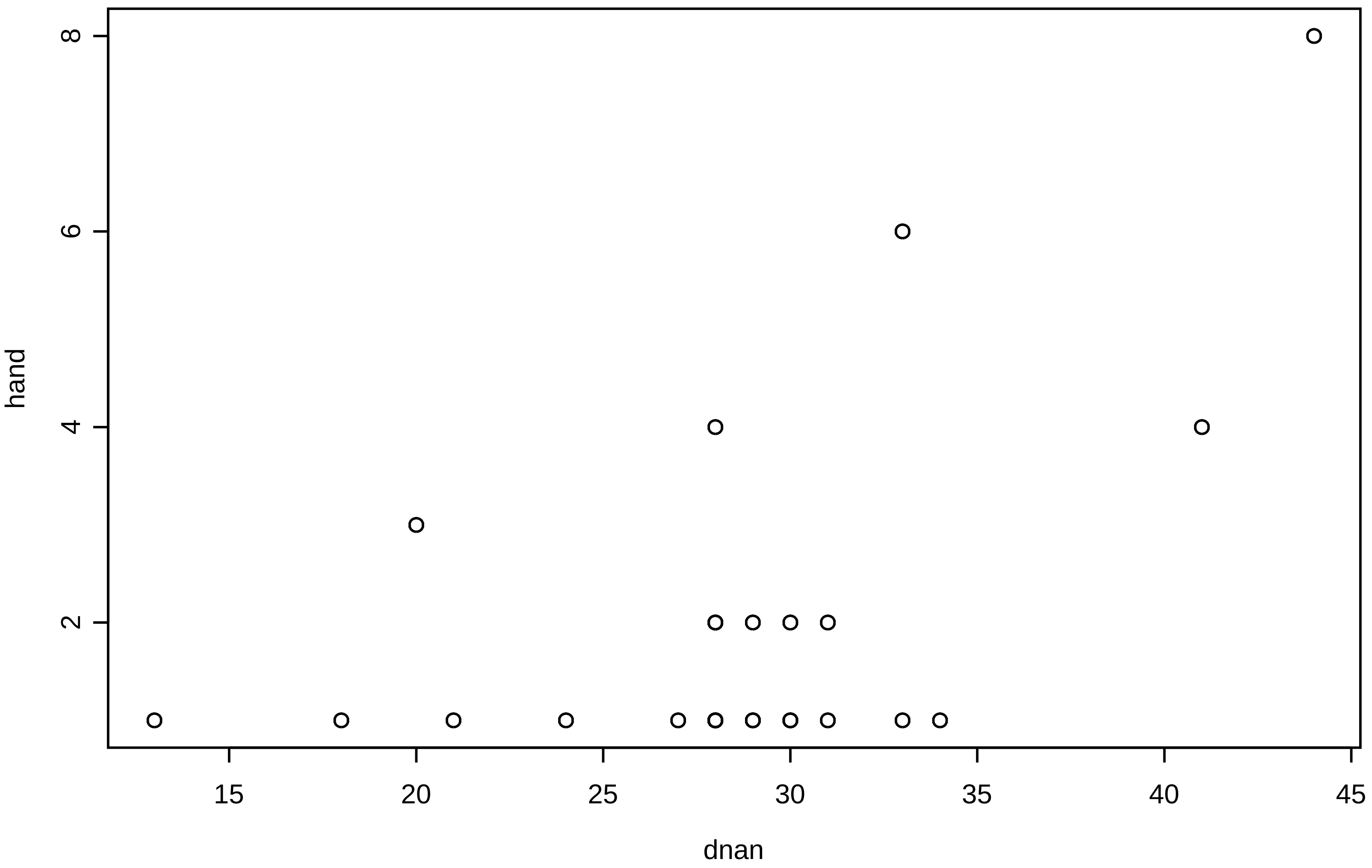
Example: (correlation test)
Random pair $Y = (U, X)$. Are U and X independent (H_0)?
Alternative (H_A): x tends to be larger when u larger.

Illustrative data set, $n = 37$ pairs: $u =$`dnan` is a generic measure and $x =$`hand` is an integer measure of left-handedness.

Simple test statistic $T = \rho(\hat{F})$, the sample correlation. Note that the joint edf $\hat{F}$ puts mass $1/n$ at each (u_i, x_i).

Correlation is zero for any distribution satisfying H_0.

```
> data(claridge);  attach(claridge)
> cor(dnan, hand)
  0.5087758
```


hand
2
4
6
8
15
20
25
30
35
40
45
dnan

F unspecified: $S = \hat{F}$ is **minimal sufficient** for F. Under H_0, S consists of the marginal edf's, $s = (u_{(1)}, \ldots, u_{(n)}, x_{(1)}, \ldots, x_{(n)})$.

A conditional test is applied with $p = \Pr(T|S, H_0)$, which is independent of the marginal distributions of U and X.

When $S = s$, the random sample $(U_1, X_1), \ldots, (U_n, X_n)$ is equivalent to $(u_{(1)}, X_1^*), \ldots, (u_{(n)}, X_n^*)$, with $(X_1^*, \ldots, X_n^*)$ a random **permutation** of $x_{(1)}, \ldots, x_{(n)}$.

Under H_0 all $n!$ permutations are equally likely. One-sided P-value

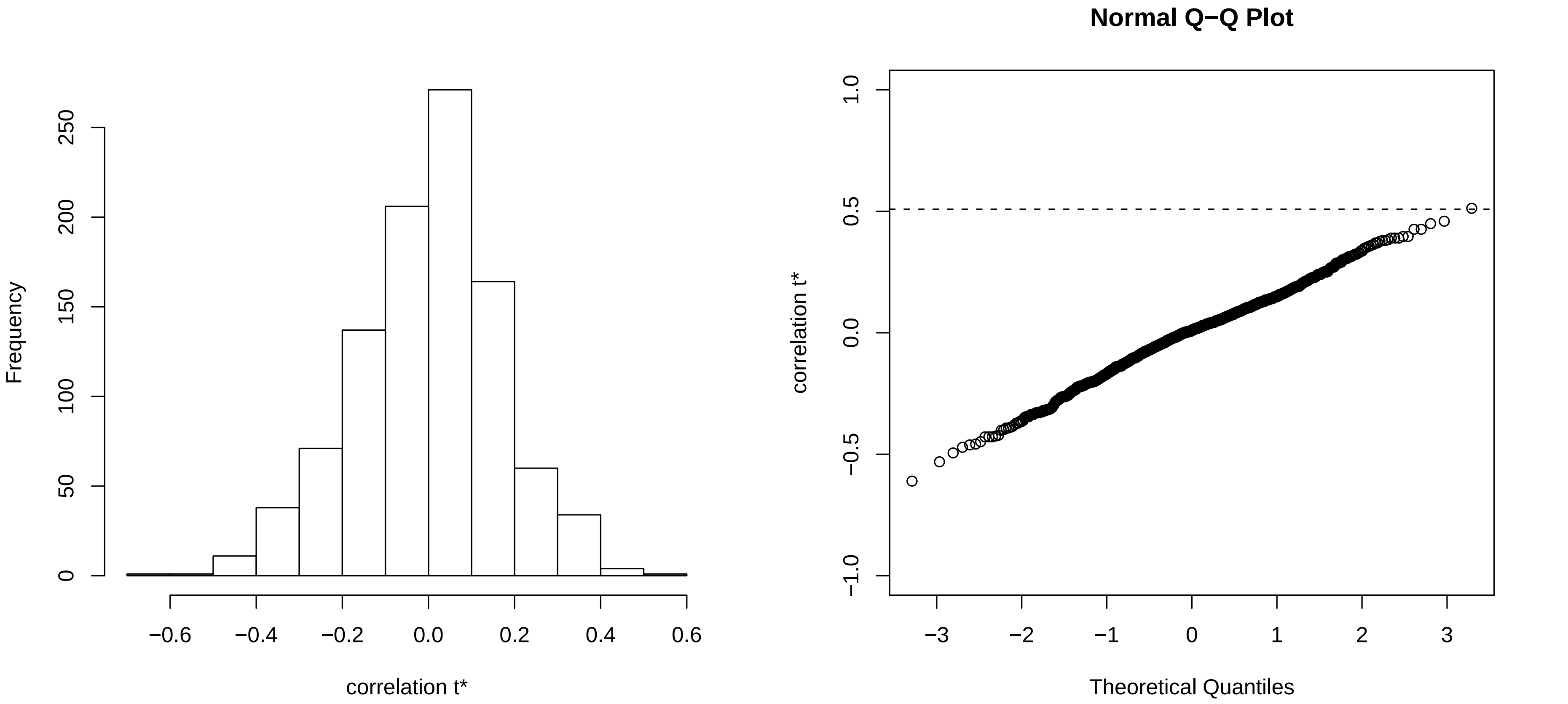
$$p = \frac{\#\text{ permutations such that } T^* \geq t}{n!}.$$

All marginal sample moments are constant across permutations. This implies that $T \geq t$ is equivalent to $\sum_i X_i U_i \geq \sum_i x_i u_i$.
Problem: large number of permutations. Make use of MC !

Take R random permutations, calculate $t_1^*, \ldots, t_R^*$, approximate p

$$p \doteq p_{\mathsf{mc}} = \frac{1 + \#\{t_r^* \geq t\}}{R+1}.$$

```
> data(claridge);  attach(claridge)
> cor.fun <- function(data, i) cor(data[ ,1], data[i, 2])
> cor.boot <- boot(claridge, cor.fun, R=999, sim="permutation")
> (1 + sum(cor.boot$t>cor.boot$t0))/(cor.boot$R + 1)
  0.002
```


Frequency
0
50
100
150
200
250
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
correlation t*
Normal Q−Q Plot
correlation t*
−1.0
−0.5
0.0
0.5
1.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Example: compare means μ_1, μ_2 of 2 populations samples $(y_{11}, \ldots, y_{1n_1})$, $(y_{21}, \ldots, y_{2n_2})$.

$H_0 : \mu_1 = \mu_2$ alone does not reduce sufficient statistic from the 2 ordered samples. We also assume that F_1, F_2 have one of the forms

$$F_1(y) = G(y - \mu_1), \qquad F_2(y) = G(y - \mu_2)$$
$$F_1(y) = G(y/\mu_1), \qquad F_2(y) = G(y/\mu_2)$$

Now H_0 implies common cdf F for both populations, and the H_0 sufficient statistic s is the set of ordered statistics for the *pooled sample* $u_1 = y_{11}, \cdots, u_{n_1} = y_{1n_1}, u_{n_1+1} = y_{21}, \cdots, u_{n_1+n_2} = y_{2n_2}$, that is $s = (u_{(1)}, \ldots, u_{(n_1+n_2)})$.

Suppose we use $t = \overline{y}_2 - \overline{y}_1$ to test against $H_A : \mu_2 > \mu_1$. If H_0 implies a common cdf for Y_{1i} and Y_{2j}, then the exact significance probability is

$$p = \Pr(T \geq t | S = s, H_0)\,.$$

When $S = s$, the concatenation $(Y_{11}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2})$ must form a permutation of s.

The first n_1 elements of a permutation will give the first sample and the last n_2 components will give the second sample. Under H_0, all $\binom{n_1+n_2}{n}$ permutations are equally likely, i.e.

$$p = \frac{\#\text{ permutations such that } T^* \geq t}{\binom{n_1+n_2}{n}}$$

Nonparametric Bootstrap Tests

Permutation tests are special nonparametric resampling tests without replacement.

Significance tests needs the calculation of P-values under H_0.

We must resample from $\hat{F}_0$, satisfying H_0. The basic bootstrap test is to compute the P-value as

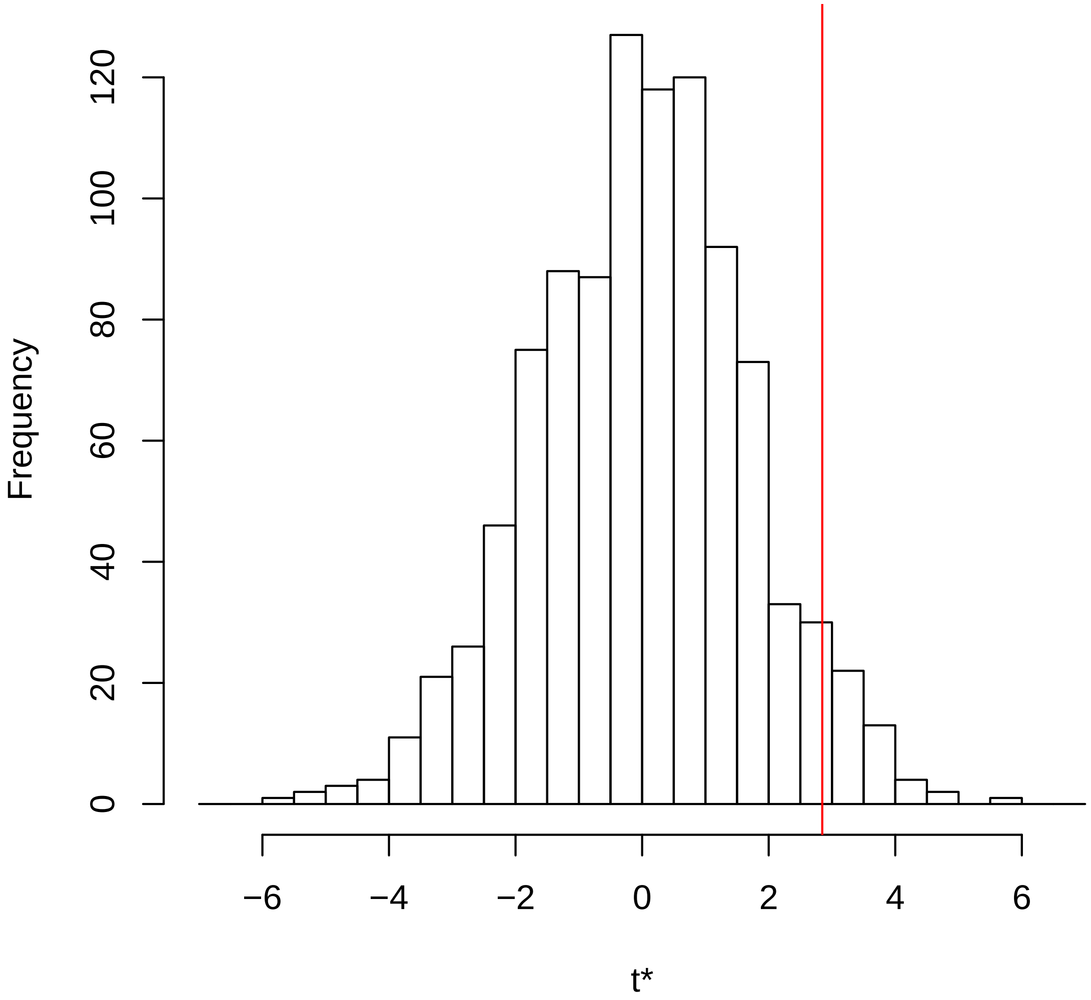
$$p_{\textsf{boot}} = \Pr{}^*(T^* \geq t | \hat{F}_0) \doteq \frac{\#\{t_r^* \geq t\} + 1}{R + 1}\,.$$

Example: (compare 2 means, cont'd)

$\hat{F}_0$: pooled edf of $(y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2})$.

Take random samples with replacement of size $n_1 + n_2$ from pooled data.

```
> grav <- gravity[as.numeric(gravity$series) >= 7, ]
> grav.fun <- function(data, i) {
+   d <- data[i, ]
+   m <- tapply(d$g, data$series, mean); m[8]-m[7]
+ }
> grav.boot <- boot(grav, grav.fun, R=999)
> (sum(grav.boot$t > grav.boot$t0) + 1)/(grav.boot$R + 1)
  0.036
```


Frequency
0
20
40
60
80
100
120
−6
−4
−2
0
2
4
6
t*

Question: do we lose anything by assuming that the two distributions have the same shape?

$\hat{F}_0$ partly motivated by permutation test – this is clearly not the only possibility.

More reasonable null model would be one which allows for different variances, too. Generally, there are many candidates for null model with different restrictions imposed in addition to H_0.

Semiparametric null models: Some features of underlying distributions are described by parameters.

Example: (compare several means)

H_0: means of all 8 series are equal, but allow for heterogeneity, i.e.

$$y_{ij} = \mu_i + \sigma_i \epsilon_{ij}\,, \qquad j = 1, \ldots, n_i\,,\ i = 1, \ldots, 8\,.$$

$\epsilon_{ij} \sim G$. H_0: $\mu_1 = \cdots = \mu_8$ with general alternative. Appropriate test statistic

$$t = \sum_{i=1}^{8} w_i (\overline{y}_i - \hat{\mu}_0)^2\,, \qquad w_i = n_i / s_i\,,$$

with $\hat{\mu}_0 = \sum w_i \overline{y}_i / \sum w_i$ the null estimate of the common mean. The null distribution of T would be approximately χ^2_7.

Question: what about the effect of small sample sizes?

Answer: a bootstrap approach is sensible.

The null model fit includes $\hat{\mu}_0$ and the estimated variances

$$\hat{\sigma}_{i0}^2 = (n_i - 1)s_i^2/n_i + (\overline{y}_i - \hat{\mu}_0)^2 \,.$$

The plot of the H_0 studentized residuals

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

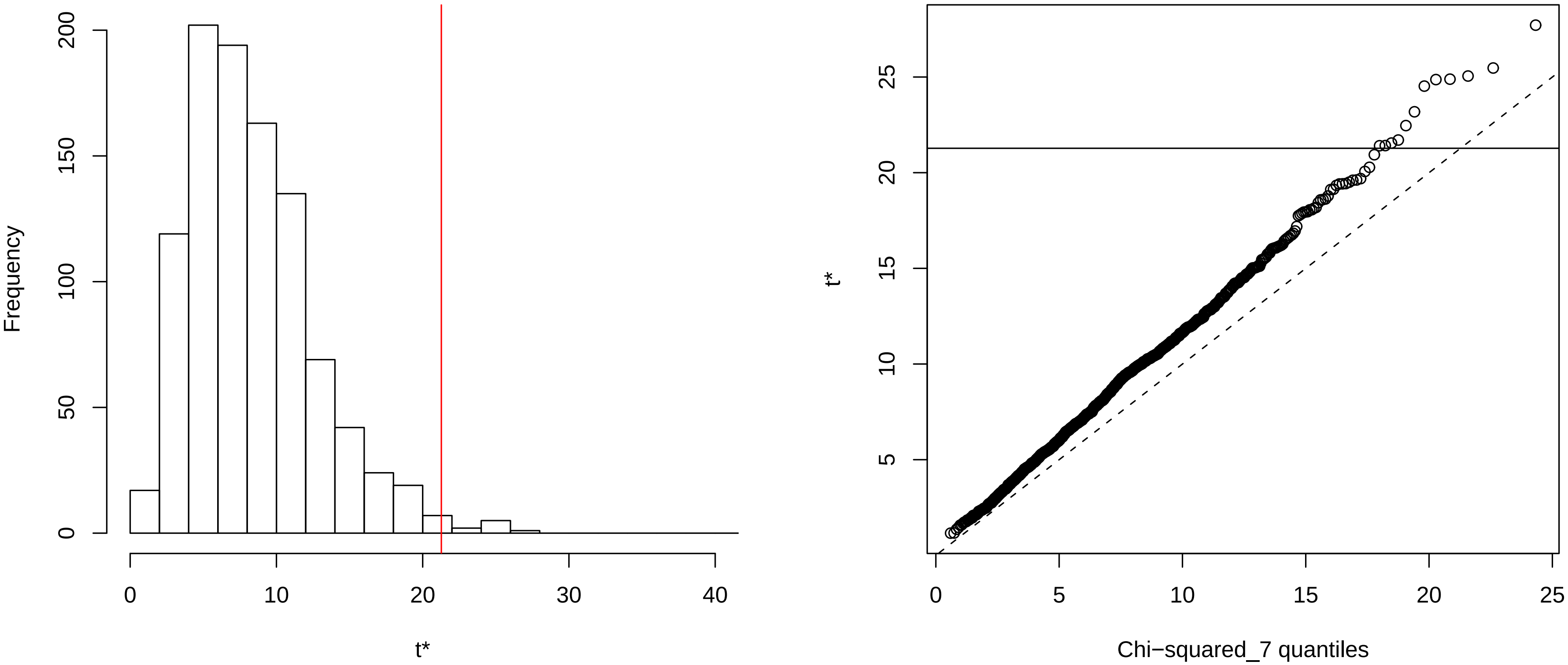
against normal quantiles shows mild non-normality.

Apply nonparametric bootstrap and simulate data under H_0

$$y_{ij}^* = \hat{\mu}_0 + \hat{\sigma}_{i0}\epsilon_{ij}^* \,, \qquad \epsilon_{ij}^* \overset{iid}{\sim} \hat{F}_e$$

```
> data(gravity);  grav8 <- gravity

> grav8.fun <- function(data, i) {
+   d <- data[i, ]
+   mi <- tapply(d$g, data$series, mean)
+   si <- tapply(d$g, data$series, var)
+   ni <- summary(data$series); wi <- ni/si
+   mu0 <- sum(wi*mi/sum(wi))
+   sum(wi*(mi-mu0)^2)   # test statistic
+ }
```

```
> grav8.gen <- function(data) {
+   d <- data
+   d.g <- data$mu0
+          + sqrt(data$sigmai0)
+          * sample(x=data$e, size=length(data$e), replace=T)
+   d
+ }

> grav8.boot <- boot(grav8, grav8.fun, R=999, ran.gen=grav8.gen)

> (sum(grav8.boot$t > grav8.boot$t0)+1)/(grav8.boot$R+1)
  0.013
```


Frequency
0
50
100
150
200
0
10
20
30
40
t*
t*
5
10
15
20
25
0
5
10
15
20
25
Chi−squared_7 quantiles

Regression Models

Assume: $y_i|x_i \overset{ind}{\sim} F_y(\mu(x_i,\beta),\sigma^2)$, $i=1,\ldots,n$ or (equivalently)

$$y_i = \mu(x_i,\beta) + \epsilon_i\ ,$$

$$\epsilon_i \overset{iid}{\sim} F_\epsilon(0,\sigma^2)$$

where F_ϵ is centered but unknown. Covariates $x_1,\ldots,x_n$ are fixed.

$P = F_y$ identified through (β, F_ϵ)

Least-Squares estimate $\hat{\beta}$ minimizes the criterion

$$\mathsf{SSE}(y, \beta) = \sum_{i=1}^{n} \Big(y_i - \mu(x_i, \beta)\Big)^2$$

Question: How accurate is $\hat{\beta}$ generally as an estimate of β?

Residual Resampling

Estimate β by the LSE $\hat{\beta}$ and F_ϵ by

$$\hat{F}_\epsilon : \text{Mass } 1/n \text{ at each } r_i - \overline{r}$$

with residuals

$$r_i = y_i - \mu(x_i, \hat{\beta})$$

Thus, estimate P by $\hat{P} = (\hat{\beta}, \hat{F}_\epsilon)$.

The bootstrap sample

$$y_i^* = \mu(x_i, \hat{\beta}) + \epsilon_i^*, \qquad \epsilon_i^* \overset{iid}{\sim} \hat{F}_\epsilon$$

gives again a LSE $\hat{\beta}^*$, the minimizer of SSE(y^*, β)

For ordinary linear models, $\mu(x_i, \beta) = x_i^t \beta$, we have

$$y_i^* = x_i^t \hat{\beta} + \epsilon_i^*, \qquad \epsilon_i^* \overset{iid}{\sim} \hat{F}_\epsilon$$

This gives

$$\hat{\beta}^* = (X^t X)^{-1} X^t y^*$$

in closed form with

$$\mathsf{E}_*(\hat{\beta}^* | y) = \hat{\beta}, \quad \mathsf{var}_*(\hat{\beta}^* | y) = \tilde{\sigma}^2 (X^t X)^{-1}$$

and

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (r_i - \overline{r})^2.$$

Let $R(y, P) = \hat{\beta} - \beta$.

A familiar measure of accuracy is the MSE matrix

$$\mathsf{E}_P\Big((\hat{\beta} - \beta)(\hat{\beta} - \beta)'\Big) = \mathsf{E}_P\Big(R(y, P)R(y, P)'\Big)$$

The bootstrap estimate of this matrix is

$$\mathsf{E}_{\hat{P}}\Big(R(y^*, \hat{P})R(y^*, \hat{P})'\Big)$$

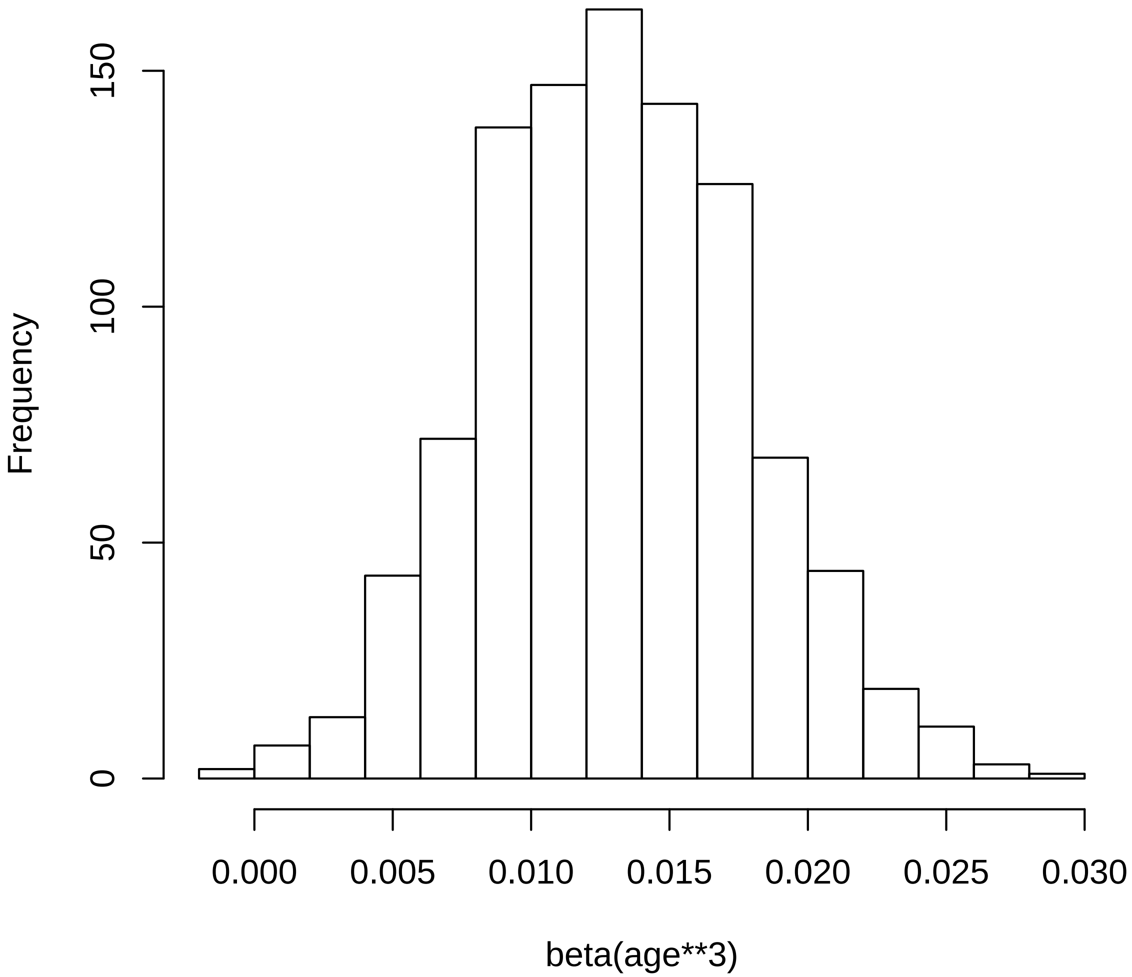
with $\hat{P} = (\hat{\beta}, \hat{F}_\epsilon)$

Example 3: mean vital capacity (lung volume) linearly depends on age (powers) and height, really?

```
> summary(model <- lm(VC ~ height+age+I(age**2)+I(age**3)))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.106e+03  2.055e+02  -5.381 8.33e-07 ***
height       6.660e+00  9.106e-01   7.314 2.55e-10 ***
age          4.888e+01  1.575e+01   3.105  0.00270 **
I(age^2)    -1.462e+00  4.979e-01  -2.935  0.00443 **
I(age^3)     1.318e-02  4.945e-03   2.665  0.00944 **
---
Residual standard error: 51.59 on 74 degrees of freedom
Multiple R-Squared: 0.566,      Adjusted R-squared: 0.5426
F-statistic: 24.13 on 4 and 74 DF,  p-value: 8.468e-13
```

```
> r <- model$residuals
> f <- model$fitted
> for (b in 1:1000) {
+   i <- sample(79, replace=TRUE)
+   VC.star <- f + r[i]
+   b.MC[b, ]<-lm(VC.star ~ height+age+I(age**2)+I(age**3))$coef
+ }

> cov(b.MC)
       [,1]    height         age      age**2   age**3
[1,] 40149.83 -1.1797e+02 -1.9235e+03 59.50235 -5.8042e-01
[2,]  -117.97  7.7866e-01 -2.1327e+00  0.06546 -5.8798e-04
[3,] -1923.54 -2.1327e+00  2.3616e+02 -7.38972  7.1890e-02
[4,]    59.50  6.5463e-02 -7.3897e+00  0.23406 -2.3024e-03
[5,]    -0.58 -5.8798e-04  7.1890e-02 -0.00230  2.2900e-05
```

```
> summary(beta.MC[,2:5])
     height           age            age**2          age**3
 Min.   :3.56   Min.   : 1.84   Min.   :-2.98   Min.   :-0.002
 1st Qu.:6.06   1st Qu.:38.77   1st Qu.:-1.78   1st Qu.: 0.010
 Median :6.67   Median :48.37   Median :-1.44   Median : 0.013
 Mean   :6.64   Mean   :48.88   Mean   :-1.46   Mean   : 0.013
 3rd Qu.:7.22   3rd Qu.:59.15   3rd Qu.:-1.12   3rd Qu.: 0.016
 Max.   :9.62   Max.   :96.37   Max.   : 0.06   Max.   : 0.029
```


Frequency
0
50
100
150
0.000
0.005
0.010
0.015
0.020
0.025
0.030
beta(age**3)

Generalized Linear Models (GLM's)

Assume: McCullagh & Nelder (1989), Fahrmeir & Tutz (1994)

1) $y = (y_1, \ldots, y_n)^t$ independent random sample from an
2) exponential dispersion family (normal, Poisson, Binomial, Gamma) with mean μ_i and variance $\phi_i V(\mu_i)$. Let $\phi_i = \phi a_i$, with fixed a_i and possibly unknown ϕ.
3) link function $g(\mu_i) = \eta_i$ (linear predictor)
4) $\eta_i = x_i^t \beta$ with explanatory variables $x_i = (x_{i1}, \ldots, x_{ip})^t$ and unknown parameters $\beta = (\beta_1, \ldots, \beta_p)^t$.

Special GLM's

- Linear regression: $y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$,

$\mathsf{E}(y_i) = \mu_i$, $\mathsf{var}(y_i) = \sigma^2$, $g(\mu_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$

- Linear logistic regression: $y_i \overset{ind}{\sim} \mathsf{Binomial}(m_i, \pi_i)$,

$\mathsf{E}(y_i/m_i) = \pi_i$, $\mathsf{var}(y_i/m_i) = \frac{1}{m_i}\pi_i(1 - \pi_i)$

$g(\mu_i) = \log(\mu_i/(1 - \mu_i)) = \sum_{j=1}^{p} x_{ij}\beta_j$

- Binary dispersion model: y_i independent with

$\mathsf{E}(y_i/m_i) = \pi_i$, $\mathsf{var}(y_i/m_i) = \phi\frac{1}{m_i}\pi_i(1 - \pi_i)$

An Example: Tumor Prognosis

Find relevant Risk factors for Recurrence of Cervical Carcinoma.

Data: 313 patients, 123 recurrences within 5 years.

Risk factors:
LN: lymph node metastases (0-3),
BZ: border zone involvement (0-2),
MA: mitotic activity (0-2).

$4 \times 3 \times 3$ Contingency Table with entries:

$$\#\text{recurrences } y_i \;/\; \#\text{patients } m_i$$

Standard Model: $y_i \sim$ Binomial(m_i, π_i), $i = 1, \ldots, 36$

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_1 + \beta_2 LN + \beta_3 BZ + \beta_4 MA.$$

What about allowing for additional dispersion parameter $\phi \neq 1$?

Problems:

- find MLE $\hat{\beta}$ and an estimate $\hat{\phi}$ (if $\phi \neq 1$),
- properties of these estimates?

Quasi-Likelihood Estimate

Wedderburn (1974): GLM assumptions only based on moments

E$(y_i) = \mu_i$, var$(y_i) = \phi V(\mu_i)$

Dispersion ϕ

Note: Exponential family log-likelihood gives

$$\frac{\partial \ell_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi a_i V(\mu_i)}.$$

ϕ sometimes fixed (Poisson, binomial).

The quasi log-likelihood q is defined as

$$\frac{\partial q_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi a_i V(\mu_i)}.$$

Important: derivatives of ℓ and q have equal mean and covariance.

Assuming $g(\mu_i) = \eta_i = x_i^t \beta$ gives

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{1}{g'(\mu_i)} x_i.$$

g canonical $\Leftrightarrow V(\mu_i) = 1/g'(\mu_i)$.

Only canonical links will be considered in the following. So we have for $\phi = 1$

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial q_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i} x_{ij}.$$

Standard Estimators

Remember var$(y_i) = \phi a_i V(\mu_i)$. The MLE $\hat{\beta}$ solves

$$u_j(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i} x_{ij} = 0. \tag{1}$$

Expanding $u(\beta)$: Iteratively Weighted Least Squares procedure

$$\hat{\beta} = (X'W(\mu)X)^{-1} X'W(\mu) z(\mu), \tag{2}$$

with $W(\mu) =$ diag$(V(\mu_i)/a_i)$ and some pseudo observations

$$z(\mu_i) = \eta(\mu_i) + (y_i - \mu_i)/V(\mu_i). \tag{3}$$

Start the iteration at some $\mu = \mu^0$.

In case of unknown ϕ we use the mean Pearson statistic

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)} = (n-p)X^2 \tag{4}$$

Properties of the MLE: $\hat{\beta}$: asymptotically normal

$$\hat{\beta} - \beta \sim AN\left(0, \phi(X'W(\mu)X)^{-1}\right).$$

Finite samples:

- estimate bias and variance of $\hat{\beta}$.

Example cont'd: $\log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 LN_i + \beta_3 BZ_i + \beta_4 MA_i$.

ML-estimation:

risk factor	$\hat{\beta}$	Std.Error	Wald Prob.
1	-1.918	0.258	< 0.001
LN	1.036	0.156	< 0.001
BZ	0.530	0.180	0.003
MA	0.440	0.178	0.014
$X^2 = 44.86,\ df = 32$			

Note: $X^2/df = \hat{\phi} = 1.402$. Is there Overdispersion: $\phi > 1$?
Then var$(\hat{\beta})$ should be multiplied by $\hat{\phi}$!

Non-Parametric Residual Resampling

Motivation in GLM's: Start in the true μ

$$\begin{aligned} z(\mu_i) &= \eta(\mu_i) + \sqrt{a_i/V(\mu_i)}\frac{y_i - \mu_i}{\sqrt{a_i V(\mu_i)}} \\ &= \eta(\mu_i) + W(\mu_i)^{-1/2}\epsilon(\mu_i), \end{aligned}$$

so $\mathrm{E}(\epsilon(\mu_i)) = 0$, $\mathrm{var}(\epsilon(\mu_i)) = \phi$.

Ordinary linear models (OLM): $\mathrm{var}(y_i) = \phi$:

$$z(\mu_i) = y_i = \mu_i + (y_i - \mu_i) = \mu_i + \epsilon(\mu_i).$$

$\epsilon(\mu_i)$ are exchangeable. $\hat{F}_\epsilon$ describes F_z. Residual resampling:

$$\hat{F}_\epsilon : \text{mass } 1/n \text{ at } \epsilon(\hat{\mu}_i) = z_i - \hat{\mu}_i.$$

Very good (well known) properties.

- known z in OLM's but unknown in GLM's
- exact additivity in OLM's but approx. additivity in GLM's.

Resample $e_1^*, \ldots, e_n^*$ from the edf $\hat{F}_e$ based on $e_i = \epsilon(\hat{\mu}_i) - \overline{\epsilon}(\hat{\mu})$ then $\mathsf{E}^*(e_i^*) = 0$ and $\mathsf{var}^*(e_i^*) = \hat{\phi}_e$

$$\hat{\phi}_e = \frac{1}{n}\sum_{i=1}^{n} \epsilon^2(\hat{\mu}_i) - \overline{\epsilon}^2(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 < \hat{\phi}.$$

Build resampled quantities

$$z_i^* = \eta(\hat{\mu}_i) + W(\hat{\mu}_i)^{-1/2} e_i^*,$$

$\mathsf{E}^*(z_i^*) = \eta(\hat{\mu}_i), \mathsf{var}^*(z_i^*) = \hat{\phi}_e a_i / V(\hat{\mu}_i)$, to bootstrap z_i with

$$\mathsf{E}(z_i) = \eta(\mu_i), \quad \mathsf{var}(z_i) = \phi a_i / V(\mu_i).$$

- known z in GLM's wrt. the bootstrap.

Alternative Residuals

Moulton & Zeger (1991) resampled from

$$e_{h,i} = \frac{e_i}{\sqrt{1 - h_{ii}(\hat{\mu})}}.$$

where $H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}$.

Friedl & Tilg (1995): Because $\overline{h}(\hat{\mu}) = p/n$, it is easier to use

$$e_{\overline{h},i} = \frac{e_i}{\sqrt{1 - p/n}} = e_i \sqrt{n/(n-p)}.$$

This gives

$$\hat{\phi}_{\overline{h}} = \hat{\phi} - \frac{n}{n-p} \bar{\epsilon}^2(\hat{\mu}).$$

No matter from what type of residuals we are resampling from,

$$\mathsf{E}^*(z^*) = X\hat{\beta}, \quad \mathsf{var}^*(z^*) = \hat{\phi}_{(\cdot)} W(\hat{\mu})^{-1}.$$

Parameter replications

Define Bootstrap replications like in the IWLS procedure (2):

$$\hat{\beta}^* = (X'W(\hat{\mu})X)^{-1}X'W(\hat{\mu})z^*.$$

Since $\hat{\mu}$'s are 'known', this defines a one-step procedure, with

$$\mathsf{E}^*(\hat{\beta}^*) = \hat{\beta},$$

and

$$\mathsf{var}^*(\hat{\beta}^*) = \hat{\phi}_{(\cdot)}(X'W(\hat{\mu})X)^{-1}.$$

Therefore, no bias correction is available but we get some new dispersion estimates depending on the residuals used.

Wild Bootstrap

Consider the mean (quasi)-scorevector

$$\overline{u}(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i} x_{ij}.$$

Then $\mathsf{E}(u_i(\beta)) = 0$ and

$$\mathsf{var}(u_i(\beta)) = x_i x_i' \mathsf{var}(y_i)/a_i^2.$$

From the IWLS procedure (2) it follows that

$$\mathsf{var}(\hat{\beta}) = (X'WX)^{-1}\mathsf{var}(u(\beta))(X'WX)^{-1}.$$

Estimate var$(u(\beta))$ by the wild bootstrap

$$u_i^* = \overline{u}(\hat{\beta}) + (u_i(\hat{\beta}) - \overline{u}(\hat{\beta}))t_i^* = u_i(\hat{\beta})t_i^*,$$

with $t_i \overset{iid}{\sim} F_t(0,1)$. Hence

$$\mathsf{E}^*(u_i^*) = 0, \quad \mathsf{var}^*(u_i^*) = x_i x_i'(y_i - \hat{\mu}_i)^2 / a_i^2.$$

Let $u^* = \sum_i u_i^* = \sum_i u_i(\hat{\beta})t_i^*$. Then

$$\begin{aligned} \hat{\beta}^* &= \hat{\beta} + (X'W(\hat{\mu})X)^{-1}u^* \\ &= (X'W(\hat{\mu})X)^{-1}X'W(\hat{\mu})z^* \\ \text{with } z_i^* &= \hat{\eta}_i + W(\hat{\mu}_i)^{-1/2}\epsilon(\hat{\mu}_i)t_i^*. \end{aligned}$$

Moreover, $\mathsf{E}^*(z^*) = X\hat{\beta}$ and

$$\mathsf{var}^*(z^*) = W(\hat{\mu})^{-1/2}\epsilon(\hat{\mu})\epsilon(\hat{\mu})'W(\hat{\mu})^{-1/2}.$$

For the i-th term this is

$$\mathsf{var}^*(z_i^*) = \hat{\phi}_i/W(\hat{\mu}_i), \quad \hat{\phi}_i = \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}$$

is the i-th contribution to $\hat{\phi}$. Therefore,

$$\mathsf{var}^*(\hat{\beta}^*) = (X'W(\hat{\mu})X)^{-1}X'S(\hat{\mu})X(X'W(\hat{\mu})X)^{-1}$$

with $S(\hat{\mu}) = \mathsf{diag}((y_i - \hat{\mu}_i)^2/a_i^2)$.

Example cont'd:

Standard Errors:

	ML	MQL				Wild
		$\hat{\phi}$	$\hat{\phi}_e$	$\hat{\phi}_h$	$\hat{\phi}_{\overline{h}}$	BT
1	.26	.31	.29	.31	.30	.24
LN	.16	.18	.17	.19	.18	.18
BZ	.18	.21	.20	.21	.21	.20
MA	.18	.21	.20	.21	.21	.21

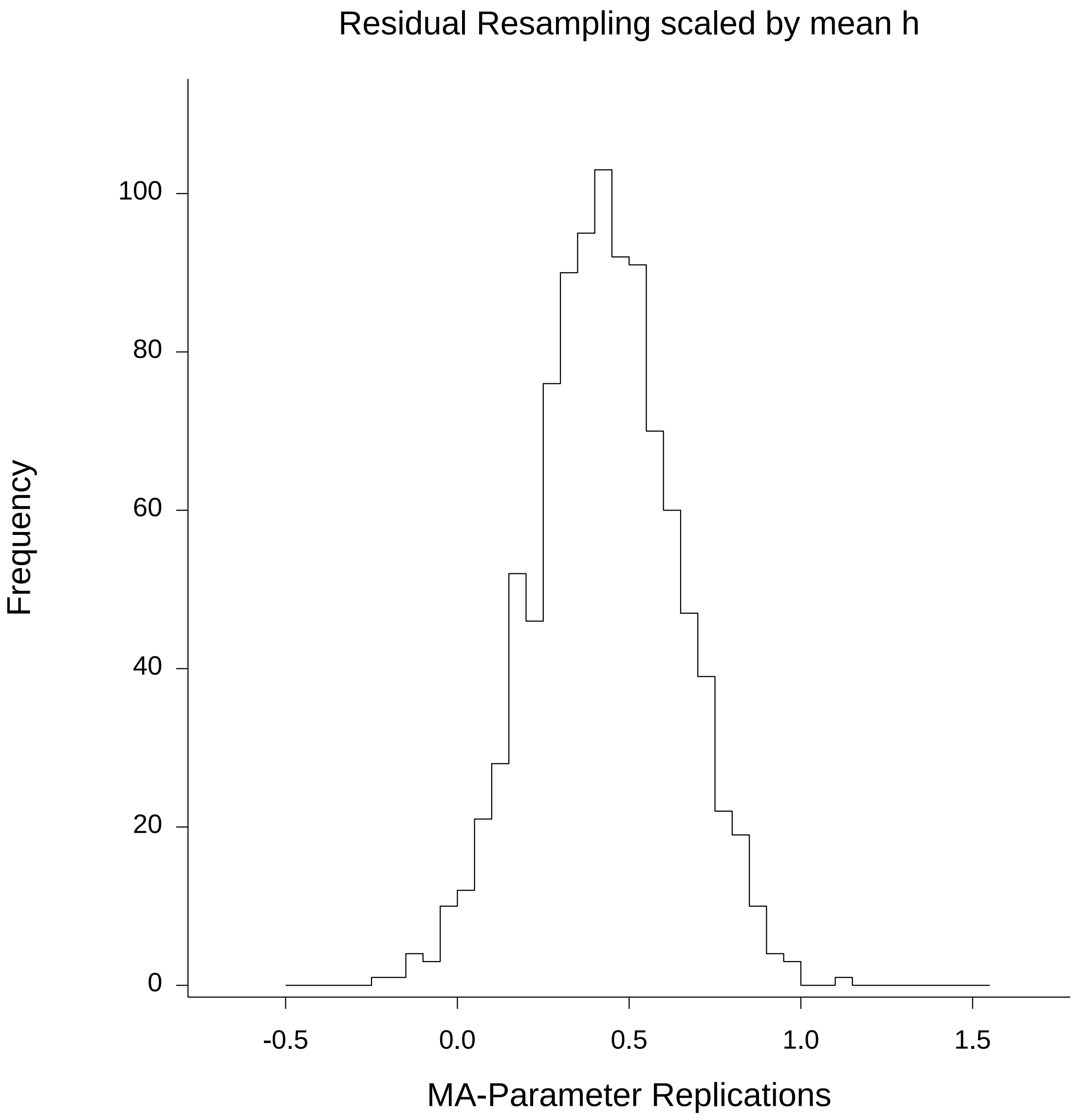
Overdispersion:

ML–model:	$\phi = 1.000,$
MQL–model:	$\hat{\phi} = 1.402,$
Res. unscaled:	$\hat{\phi}_e = 1.241,$
Res. h-scaled:	$\hat{\phi}_h = 1.418,$
Res. p/n-scaled:	$\hat{\phi}_{\overline{h}} = 1.396.$

Monte-Carlo Results

1000 β_4 (MA) Replications

	mean	S.E.	Pr> 0
unscaled	0.432	0.192	1.2%
p/n scaled	0.432	0.203	1.9%
Wild BT	0.446	0.204	1.4%


Residual Resampling scaled by mean h
Frequency
100
80
60
40
20
0
-0.5
0.0
0.5
1.0
1.5
MA-Parameter Replications

Discussion

- no bias reduction
- alternative variance estimators
- robust against misspecifications
- not computationally intensive
- theoretical results
- also applicable to study $h(\hat{\beta})$ via Monte Carlo Simulation
- works without defining some y^*

References

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer.

Efron B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

Shao, J., Tu, D. (1995). *The Jackknife and Bootstrap*, Springer.

Davison, A.C., Hinkley, D.V. (1997). *Bootstrap Methods and their Application*, Cambridge University Press.

Efron B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1-26.

Efron B., Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 54-77.