Permutation tests

Data: $(x, y)_i$ for $i = 1, \ldots, n$

Question: Are the x's and y's associated?

Statistic: $T(x, y)$ (T large $\Rightarrow$ association)

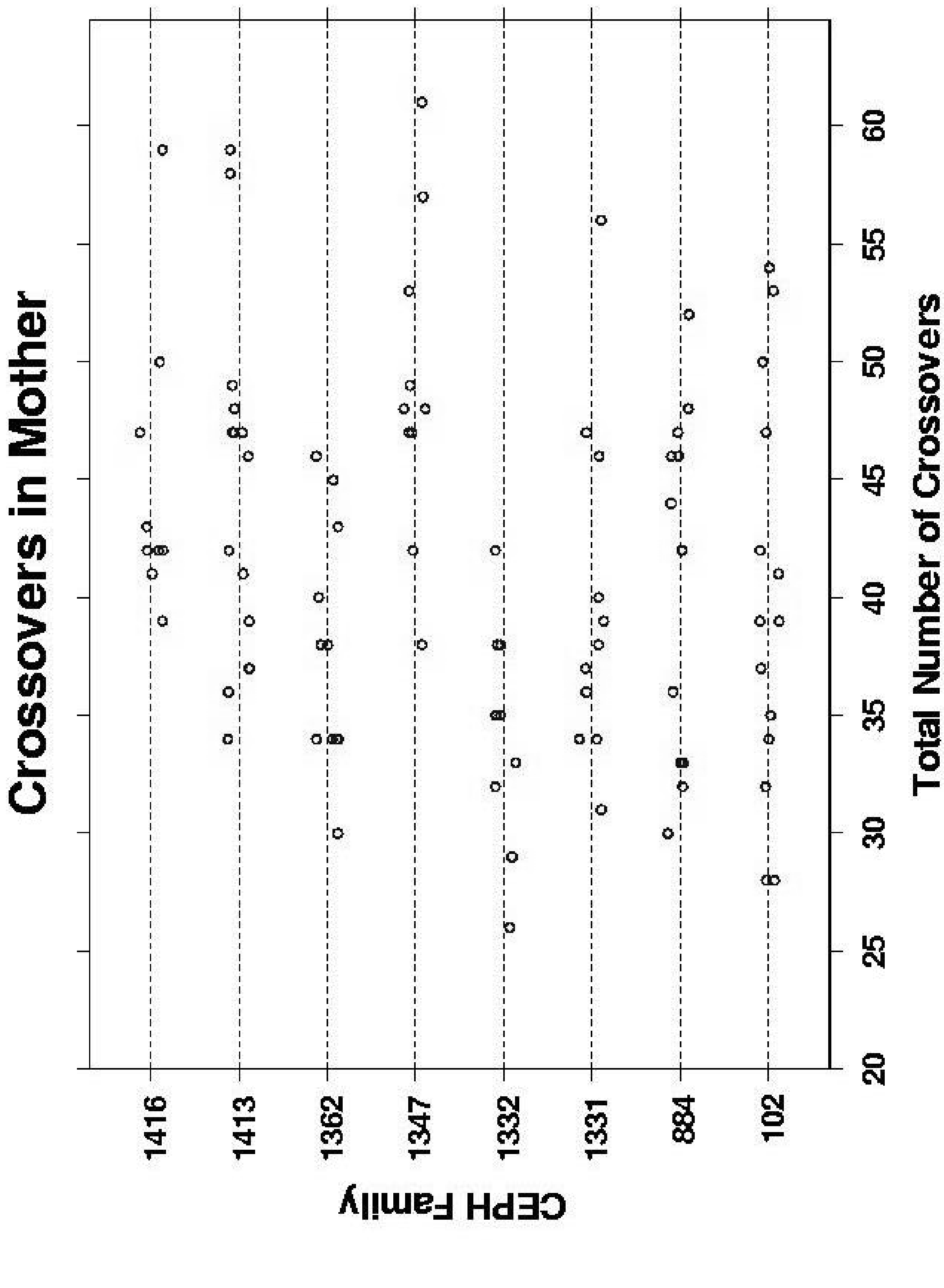
Permutation distribution:

Look at the distribution of $T(x^\star, y)$
where $x^\star$ is a permuted version of x.
P-value $= \Pr\left[T(x^\star, y) > T(x, y)\right]$

Important issues:

- Sampling vs systematic enumeration
- Choice of test statistic
- Conditional vs unconditional test
  See Lehmann (1986) TSH, 2nd ed, chapter 10
- Normal approximation (esp for ANOVA)

Book: BFJ Manly (1997) Randomization, bootstrap and Monte Carlo methods in biology, 2nd ed. Chapman & Hall, London.


Crossovers in Mother
CEPH Family
1416
1413
1362
1347
1332
1331
884
102
20
25
30
35
40
45
50
55
60
Total Number of Crossovers

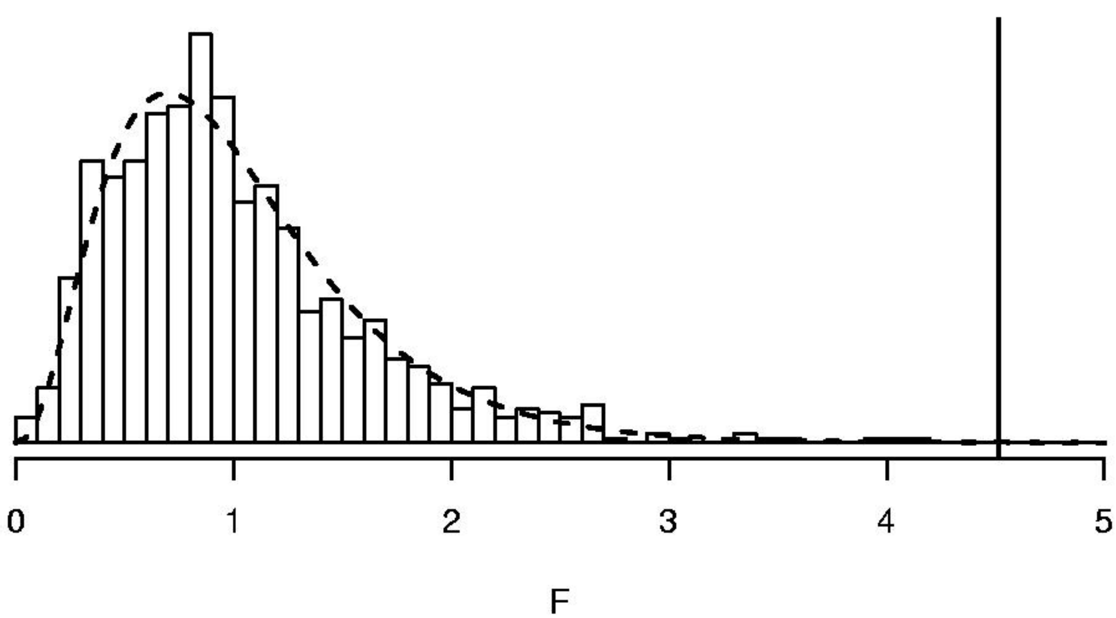
Example

```
y <- c(28, 28, 32, ...)
x <- factor(c(1,1,1, ..., 2, ..., 8, ...))

f <- anova(aov(y ~ x))$F[1]

f0 <- 1:1000
for(i in 1:1000)
  f0[i] <- anova(aov(y ~ sample(x)))$F[1]

mean(f0 > f)
```



Parametric bootstrap

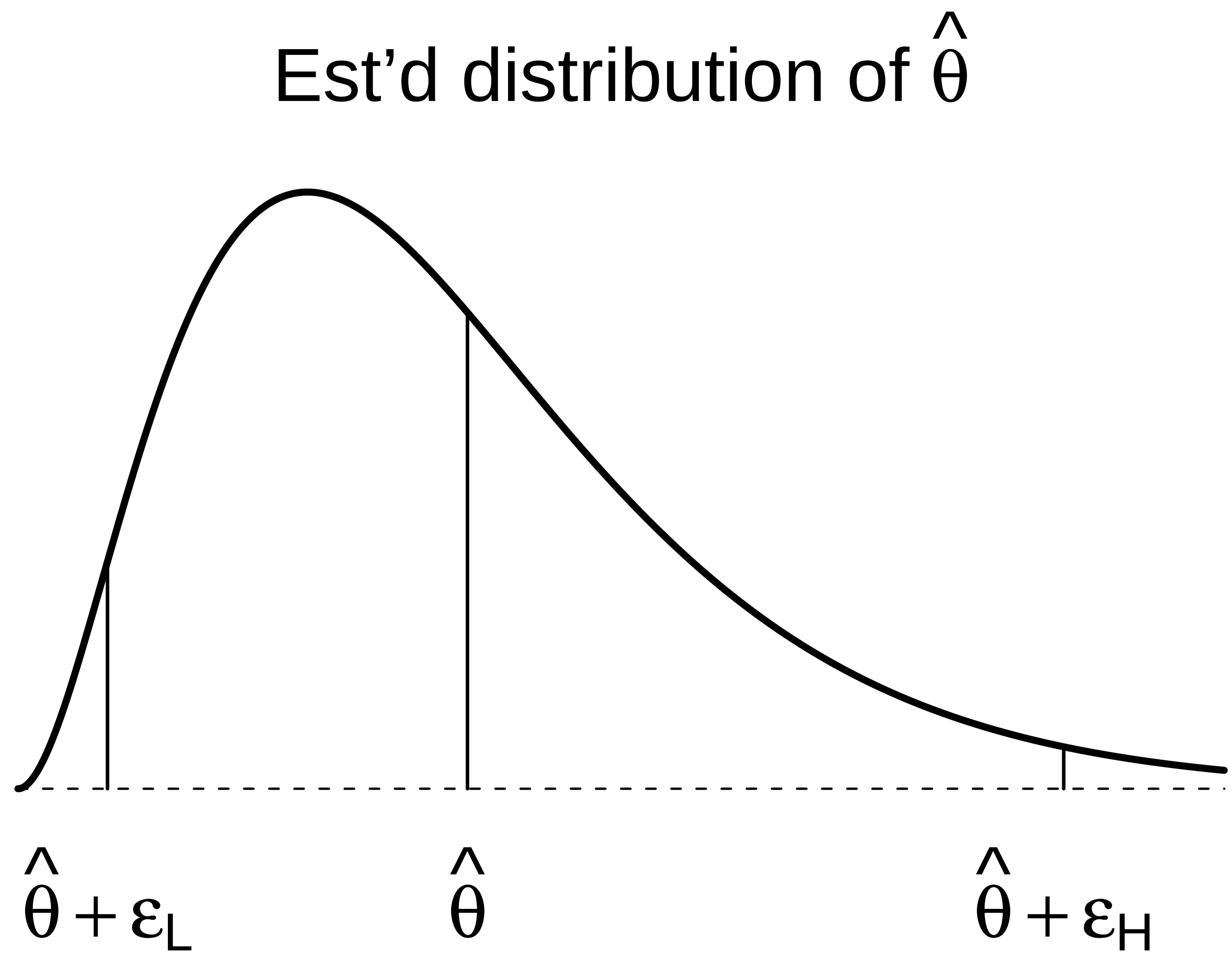
Suppose $x_1, x_2, \ldots, x_n \sim$ iid $f(\cdot, \theta)$ where f is known.

Let $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ be our estimator of θ (eg, the MLE).

We wish to estimate the SE of $\hat{\theta}$ and get a confidence interval for θ.

Parametric bootstrap:

1. Simulate $x_1^\star, x_2^\star, \ldots, x_n^\star \sim$ iid $f(\cdot, \hat{\theta})$.
2. Obtain $\hat{\theta}^\star = \hat{\theta}(\boldsymbol{x}^\star)$
3. Repeat steps (1) and (2) m times to obtain
   $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_m^\star$
4. Estimate SE$(\hat{\theta})$ by SD$\{\hat{\theta}_i^\star\}$
5. Estimate the bias of $\hat{\theta}$ by $\hat{\theta} -$ ave$\{\hat{\theta}_i^\star\}$
6. Calculate the confidence interval for θ by either
   (a) 2.5%ile to 97.5%ile of $\{\hat{\theta}_i^\star\}$
   (b) $(\hat{\theta} - \epsilon_H,\ \hat{\theta} - \epsilon_L)$ where ϵ_L and ϵ_H are the 2.5%ile and 97.5%ile, respectively, of $\{\hat{\theta} - \hat{\theta}_i^\star\}$


Est’d distribution of $\hat{\theta}$
$\hat{\theta}+\varepsilon_L$
$\hat{\theta}$
$\hat{\theta}+\varepsilon_H$

Nonparametric bootstrap

Suppose $x_1, x_2, \ldots, x_n \sim$ iid $f(\cdot, \theta)$ where the form f is perhaps unknown. Let $F(\cdot, \theta)$ be the corresponding cdf.

We estimate F by the empirical cdf $\hat{F}_n$.

Nonparametric bootstrap:

1. Simulate $x_1^\star, x_2^\star, \ldots, x_n^\star \sim$ iid $\hat{F}_n$. In other words, draw n values *with replacement* from the set
   $$\{x_1, x_2, \ldots, x_n\}$$
2. Obtain $\hat{\theta}^\star = \hat{\theta}(\boldsymbol{x}^\star)$
3. Repeat steps (1) and (2) m times to obtain
   $$\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_m^\star$$
4. Everything else is as before.

Note: We may have (θ, ψ) rather than just θ (and these could all be vectors), where ψ is a nuisance parameter. But that's really no big deal.

Example

```
options(digits=3)
x <- rexp(30, 2)

print(theta <- sd(x))

    0.425


# parametric bootstrap
lambda <- 1/mean(x)
thetas <- 1:1000
for(i in 1:1000)
  thetas[i] <- sd(rexp(30,lambda))


c(mean(thetas),sd(thetas))

    0.469 0.116

quantile(thetas,c(0.025,0.975))

    0.284 0.739

theta - rev(quantile(thetas-theta,
                     c(0.025,0.975)))
    0.111 0.566
```

Example (continued)

```
# nonparametric bootstrap
thetas2 <- 1:1000
for(i in 1:1000)
  thetas2[i] <- sd(sample(x,repl=T))


c(mean(thetas2),sd(thetas2))

    0.417 0.042

quantile(thetas2,c(0.025,0.975))

    0.330 0.494

theta - rev(quantile(thetas2-theta,
                     c(0.025,0.975)))
    0.356 0.520


# Monte Carlo estimate (knowing the truth)
thetas3 <- 1:100000
for(i in 1:100000)
  thetas3 <- sd(rexp(30,2))

c(mean(thetas3),sd(thetas3))

    0.486 0.120


Compare: 0.5/sqrt(30) = 0.091
```

Further issues

- How many bootstrap replicates?
  - As many as you can
  - $n = 100 – 1000$
  - Bootstrap to estimate the Monte Carlo error

- (Related to the above)

$$
\begin{array}{l}
\qquad\qquad \mathsf{SE}(\hat{\theta}) \\
\textit{versus} \text{ bootstrap est } \widehat{\mathsf{SE}}(\hat{\theta}) \; (n \to \infty) \\
\textit{versus} \text{ obs bootstrap est } \mathsf{SD}\{\hat{\theta}_i^\star\}
\end{array}
$$

- Bias correction: bias $\downarrow \;\Rightarrow\;$ var $\uparrow$

- Transformations

- Balanced bootstrap

Bootstrap in regression

Consider $(x_1, x_2, y)_i$ for $i = 1, \ldots, n$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$\mathsf{E}(\epsilon) = 0 \qquad \text{var}(\epsilon) = \sigma^2 \qquad \epsilon_i$ independent

Approaches:

- Parametric bootstrap:
  - Obtain $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$
  - Sample $\epsilon_i^\star \sim$ iid normal$(0, \sigma^2)$
  - Take $y_i^\star = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \epsilon_i^\star$
  - Obtain $\hat{\beta}_0^\star$, $\hat{\beta}_1^\star$, $\hat{\beta}_2^\star$ and repeat many times

- Nonparametric bootstrap I:
  - Obtain $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$
  - Calculate $\hat{\epsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$
  - Sample $\epsilon_i^\star$ by drawing with replacement from $\{\hat{\epsilon}_i\}$
  - Obtain $\hat{\beta}_0^\star$, $\hat{\beta}_1^\star$, $\hat{\beta}_2^\star$ and repeat many times

Bootstrap in regression (continued)

Approaches (continued)

- Nonparametric bootstrap II:
  - Sample $(x_1^\star, x_2^\star, y^\star)_i$ by drawing with replacement from $(x_1, x_2, y)_i$
  - Obtain $\hat{\beta}_0^\star$, $\hat{\beta}_1^\star$, $\hat{\beta}_2^\star$ and repeat many times