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 \(T(x^*, y)\)
where \(x^*\) is a permuted version of \(x\).
P-value = \(\Pr [T(x^*, y) > T(x, y)]\)

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)

Example

```r
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 \( x_1, x_2, \ldots, x_n \sim \text{iid } f(\cdot, \theta) \) where \( f \) is known.

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

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

**Parametric bootstrap:**

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

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

**Note:** We may have $(\theta, \psi)$ rather than just $\theta$ (and these could all be vectors), where $\psi$ 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)

```r
# 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[i] <- 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)
  \[ \text{SE}(\hat{\theta}) \]

  versus bootstrap est \( \widehat{\text{SE}}(\hat{\theta}) \) \((n \to \infty)\)

  versus obs bootstrap est \( \text{SD}\{\hat{\theta}^*_i\}\)

• Bias correction: bias ↓ ⇒ var ↑

• Transformations

• Balanced bootstrap
Bootstrap in regression

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

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon
\]

\[
E(\epsilon) = 0 \quad \text{var}(\epsilon) = \sigma^2 \quad \epsilon_i \text{ independent}
\]

Approaches:

- Parametric bootstrap:
  - Obtain \(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\)
  - Sample \(\epsilon_i^* \sim \text{iid normal}(0, \sigma^2)\)
  - Take \(y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \epsilon_i^*\)
  - Obtain \(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*\) 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 \(\epsilon_i^*\) by drawing with replacement from \(\{\hat{\epsilon}_i\}\)
  - Obtain \(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*\) and repeat many times
Bootstrap in regression (continued)

Approaches (continued)

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