Resampling

**Definition**

- sampling or scrambling the original (sample) data numerous times
- nonparametric method

**Methods**

- **jackknifing**
  - computing the statistics of interest (medians, variances, percentiles) for all combination of the original data or subsets of available data
- **bootstrapping**
  - drawing randomly with replacement from a set of data points to estimate the sampling distribution
- **permutation tests**
  - Exchanging labels on data points (or scrambling the order of the original data) when performing significance tests

**Randomization (permutation) test**

**procedure**

- take original (sample) data
- reordering the data (reshuffling the subscript)
- calculate the test statistic, $\Lambda$ for testing the null hypothesis
- do the same thing a numerous times to get the sampling dist. of $\Lambda$ and test the null with p-value

*(example)*

\[ W = \frac{\text{height}}{(\text{weight})^{1/3}} \]

- Test the difference of $W$ by gender
- **Data**: $W$ of size $n_m$ and $W$ of size $n_f$
- The dist. of $W$ is not normal, but very skewed.
  - computing the mean difference of $W$ from the original data, for example 10 unit.
  - Take $n_m$ from all $W$ of size $(n_m+n_f)$ and assign “male” to the sample observation.
  - Assign “female” to the remaining and Computing the means of $W$ separately.
- do the same thing 5,000(N) times.
- $k$ times the mean difference of (male–female) out of N went over 10 units. Then $k/N$ is p-value for the one-sided test.
Tukey (1958) suggest Jackknife idea based on removing data and then recalculating the estimator.

**Idea**
- sample data of size $n$: $(x_1, x_2, ..., x_n)$
  $$\bar{x}_{-j} = \frac{1}{n-1} \sum_{i \neq j} x_i$$
- we know both the sample mean and $x_{(-j)}$
- $j$th data obs.
  $$x_j = n\bar{x} - (n-1)\bar{x}_{-j}$$
- similarly,
  $$\hat{\theta}_j^* = n\hat{\theta} - (n-1)\hat{\theta}_j$$, where $\hat{\theta} = \phi(x_1, x_2, ..., x_n)$

- jackknife estimator of $\theta$ is
  $$\hat{\theta}^* = \frac{\sum \hat{\theta}_j^*}{n}$$
  
- an approximate sampling error of $\hat{\theta}^*$
  $$V(\hat{\theta}^*) = V(\frac{\sum \hat{\theta}_j^*}{n}) = \frac{\sum (\hat{\theta}_j^* - \hat{\theta}^*)^2}{n(n-1)}$$

- the $100(1-\alpha)\%$ conf. interval for $\theta$
  $$\hat{\theta}^* \pm t_{\alpha/2} \sqrt{V(\hat{\theta}^*)}$$

**Motivation of Jackknife**
- reduce the bias

**Example**
- estimate the species diversity based on $n$ samples?
- total number $S=120$ of species in the sample?
  - it should be under-estimate
  - Burn and Overton (1978) suggest a Jackknife estimator to reduce the bias.
  $$S^* = S + \frac{(n-1)}{n} \cdot f_1 = 120 + \frac{9}{10} \cdot 25 = 147.5$$
Bayesian Inference

Bootstrap

- Idea
  - re-sampling of obtained single sample for generating the sampling dist.
  - robustness is depending on the sample size, in general, more than 20
  - with replacement sampling
    - randomized sampling = without replacement

- Procedure for estimating bias
  - Let $\hat{\theta} = \phi(x_1, x_2, ..., x_n)$ be the estimator for $\theta$
  - Let $\hat{\theta}_0$ denote the estimate using the original data and $\hat{\theta}_i$ be the estimator using $i$th bootstrap sample.
  - The mean of all the bootstrap $\hat{\theta}_B = \frac{1}{P} \sum_i \hat{\theta}_i$ is used for estimating the bias.
  - Bias $\hat{B} = E(\hat{\theta}) - \theta$ and $\hat{\theta}_B$ is the bootstrap estimator of $E(\hat{\theta})$.
  - Then, the bootstrap estimator of bias is $\hat{B} = \hat{\theta}_B - \hat{\theta}_0$

- Standard bootstrap confidence intervals
  - $S^2 = \frac{1}{p-1} \sum_i (\hat{\theta}_i - \hat{\theta}_B)^2$
  - then $\hat{\theta} \pm z_{\alpha/2} S$

- Efron’s percentile confidence intervals
  - using bootstrapping sampling dist.
  - $(\hat{\theta}_{L, \alpha/2}, \hat{\theta}_{H, \alpha/2})$
### Bootstrap in R

- `boot` [boot]
  - Generate R bootstrap replicates of a statistic applied to data. Both parametric and nonparametric resampling are possible. For the nonparametric bootstrap, possible resampling methods are the ordinary bootstrap, the balanced bootstrap, antithetic resampling, and permutation. For nonparametric multi-sample problems stratified resampling is used. This is specified by including a vector of strata in the call to `boot`. Importance resampling weights may be specified.

- `boot(data, statistic, R, sim="ordinary", stype="i", strata=rep(1,n), L=NULL, m=0, weights=NULL, ran.gen=function(d, p) d, mle=NULL, simple=FALSE, ...)`

### Jackknife in R

- `jackknife` [bootstrap]
  - See Efron and Tibshirani (1993) for details on this function.

- `jackknife(x, theta, ...)`

### Bayesian and Resampling

- Bayes vs. Resampling: A Rematch, Campbell Garvey et. al. (2006)
  - summary of 3 pages.