Chapter 7: Bootstrap, Jackknife, and Permutation

- Discusses different ways of computing variance, percentiles, confidence intervals, and bias of an estimator, including parametric and nonparametric bootstrap, and jackknife.
- Discusses different ways of computing p-values of hypothesis tests, including (semi-)parametric and nonparametric bootstrap, and permutation.

Section 7.1: Introduction

- Discusses some motivation behind resampling methods.
- Discusses how to generate a random sample parametrically and nonparametrically.
- Introduces random generators of common probability distributions in R.

In real-life applications of statistics, we are often required to estimate the variance of the estimated parameter of interest, Var$(\hat{\theta})$. One motivation is to know the (statistical) efficiency of the estimator when there are two or more competing methods for estimating θ.

However, there are also other cases where Var$(\hat{\theta})$ is necessary. The most popular cases are:

- Calculation of confidence intervals (interval estimates of θ);
- Calculation of p-values of the hypothesis tests about θ;
- Calculation of bias of the estimated θ.

When θ is the population mean μ, then, it is straightforward to estimate Var$(\hat{\theta}) = $ Var$(\bar{X}) = \frac{\sigma^2}{m}$, where $\bar{X}$ is the sample mean from m observations.

$\tilde{\mu}$ = pop^n median ; $\tilde{X}$ = sample median.

Question: What if θ is the population median (denoted by $\tilde{\mu}$)?

Answer: In this case, we are required to estimate Var$(\tilde{X})$, the variance of the sample median (denoted by $\tilde{X}$) of m observations. It is known that $\boxed{\text{Var}(\tilde{X}) \approx \frac{1}{4m[f(\tilde{\mu})]^2}}$ for a large m, where $f(\tilde{\mu})$ is the pdf of the random variable evaluated at $\tilde{\mu}$.

↑ asymptotic variance of $\tilde{X}$.

Question: Is it possible to estimate Var$(\tilde{X})$ when the pdf $f(x)$ is unknown?

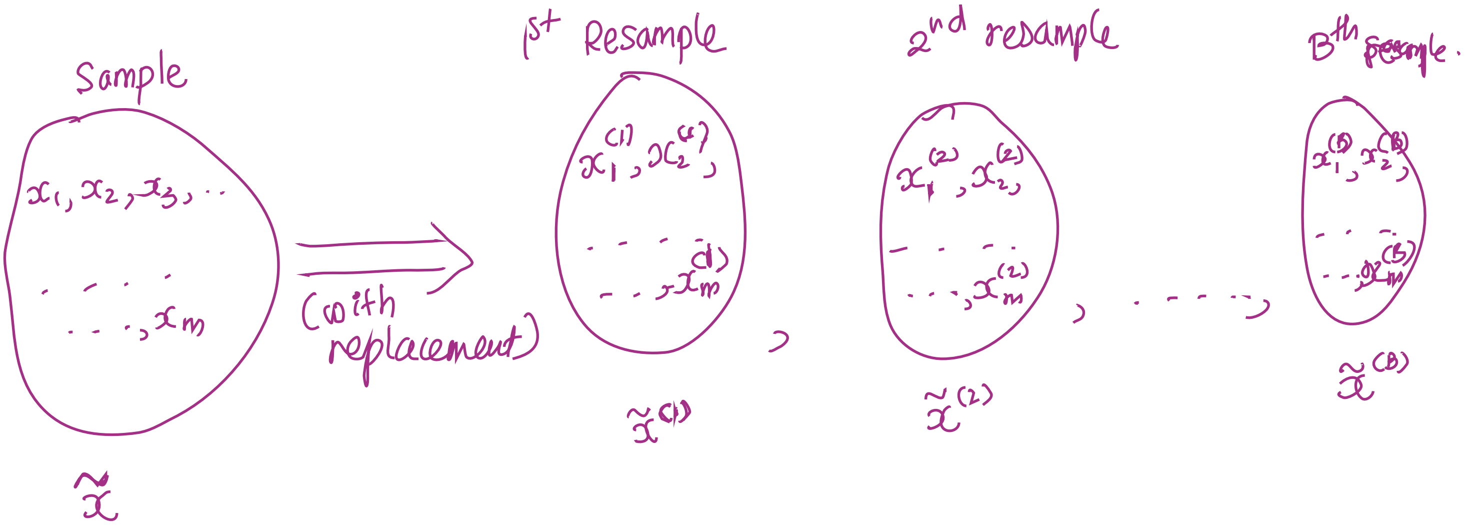
Yes. Using a resampling method.

Rationale (Heuristics): By using a resampling method, it is possible to create replicates of $\tilde{x}$ from each of many resamples. That is, for each of B resamples (where B is some large number, see Notes below), it is possible to calculate its sample median (denoted by $\tilde{x}^{(b)}$ for the b-th sample, $b = 1, \ldots, B$).

The set $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(B)}\}$ constitutes an approximate distribution of $\tilde{X}$, enabling us to estimate Var$(\tilde{X})$ as well as percentiles of $\tilde{X}$.

Notes: The choice for B depends on whether it is for the ① variance estimation or ② percentile estimation (including confidence intervals and hypothesis testing). For the former ①, a relatively small number of resamples, say, $B = 200$ to 1000, suffices. For the latter ②, typically a much larger number is desired, such as $B = 10000$. This is because confidence intervals and hypothesis testing require accurate estimation of extreme percentiles (such as 1% or 5%), and the rate of convergence is much slower at such extreme percentiles.

* Resampling assumes that the sample approximates the pop^{n} well.



* Based on the sample medians from resamples: $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(B)}\}$, we can estimate the var$(\tilde{x})$ using the sample variance formula.
(B= 1000)

* construct 95% C.I. for median $\tilde{\mu}$: (percentile bootstrap)
2.5th and 97.5th percentile of $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots \tilde{x}^{(B)}\}$ (B=10000).

Recall that cdf is one way of uniquely characterizing a distribution. Using mathematical notation, the resampling procedure is equivalent to generating samples from the distribution with the cdf

$$\text{Estimated cdf} \rightarrow \hat{F}_m(t) = P(\widehat{X_i \le t}) = \frac{1}{m}\sum_{i=1}^{m} \underset{1 \text{ or } 0}{I(x_i \le t)}, = \frac{\#\,\text{obs}^{\text{ns}} \le t}{m}$$

which is the total number of x_i not exceeding t divided by m at each $t \in (-\infty, \infty)$.

Note that this is the simple Monte Carlo estimator of

$$F(t) \overset{\text{Def}^{n}}{=} P(X_i \le t) = \underline{\underline{E[I(X_i \le t)]}}.$$

If m is sufficiently large enough, $\hat{F}_m$ is a good approximation of the true cdf $F(t)$ from which x_i are generated.

sample median

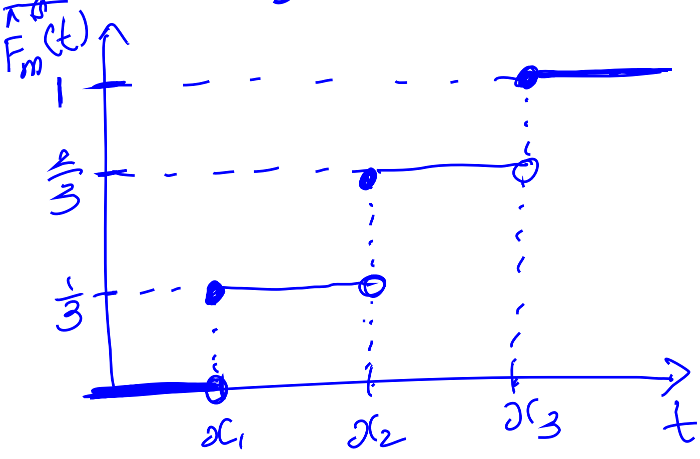
Using the distribution whose cdf $\hat{F}_m$, cdf of $\tilde{X}$, denoted by $F_{\underline{\tilde{X},m}}(t)$, $-\infty < t < \infty$, is estimated by

$$\hat{F}_{\tilde{X},m,B}(t) = \frac{1}{B}\sum_{b=1}^{B} I(\tilde{x}^{(b)} \le t), = \frac{\#\,\tilde{x}^{(b)} \le t}{B}$$

the total number of $\tilde{x}^{(b)}$ not exceeding t divided by B at each $t \in (-\infty, \infty)$. As $B \to \infty$, $\hat{F}_{\tilde{X},m,B}(t)$ converges to $F_{\tilde{X},m}(t)$.

Typical case: $\{x_1, x_2, \ldots, x_m\}$ gives an estimate of $F(t)$, denoted by $\hat{F}_m(t)$ for $t \in (-\infty, \infty)$.

Eg: x_1, x_2, x_3 $(x_1 < x_2 < x_3)$



Note: SLLN shows that the empirical cdf ($\hat{F}_m$) converges pointwise to the actual cdf.

$\hat{F}_m(t) \longrightarrow F(t)$ as $m \to \infty$.

* $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(B)}\}$ gives an estimate of $F_{\tilde{X},m}(t)$, denoted by $\hat{F}_{\tilde{X},m,B}(t)$ for $t \in \mathbb{R}$.

As $B \to \infty$, $\hat{F}_{\tilde{X},m,B}(t) \longrightarrow F_{\tilde{X},m}(t)$.

Several resampling methods are available. We will discuss the following methods to estimate $\mathrm{Var}(\hat{\theta})$ and the $100p$-th percentile of $\hat{\theta}$, denoted by $\tilde{\pi}_p(\hat{\theta})$ (instead of just focusing on $\tilde{x}$).

1. **Nonparametric bootstrap** (creates resamples by sampling observations with replacement from the original data).

2. **Parametric bootstrap** (creates resamples by sampling observations from some assumed distribution). Eg: Normal

3. **Jackknife** (creates resamples by taking one observation out of the original sample).

Bootstrap (nonparametric and parametric) was introduced by Efron (1979), and jackknife was introduced by Quenouille (1949) and Tukey (1958).

* Jackknife: $\{x_1, x_2, \ldots, x_m\}$ — original sample

Create m resamples of size $m-1$ by removing one obs$^{\underline{n}}$ at-a-time.

1st resample: $\{x_2, x_3, \ldots x_m\}$

2nd resample: $\{x_1, x_3, \ldots x_m\}$

$\vdots$

ith " : $\{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots x_m\}$

$\vdots$

mth " : $\{x_1, x_2, \ldots, x_{m-1}\}$

1. Nonparametric Bootstrap

The idea behind nonparametric bootstrap is to create resamples directly from the observations. Let $\{x_1, \dots, x_m\}$ be the set of observations. The following algorithm gives estimated $\text{Var}(\hat{\theta})$ and $\tilde{\pi}_p(\hat{\theta})$.

Algorithm for Estimating Var$(\hat{\theta})$ and $\tilde{\pi}_p(\hat{\theta})$:

1. Create B resamples of $\{x_1, \dots, x_m\}$ by sampling observations with replacement. Denote the b-th resample by $\{x_1^{(b)}, \dots, x_m^{(b)}\}$, $b = 1, \dots, B$.

2. For each of B resamples, calculate $\hat{\theta}$. Let $\hat{\theta}^{(b)}$ be $\hat{\theta}$ from the b-th resample.

   $\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(B)}\}$

3. $\text{Var}(\hat{\theta})$ is estimated by

$$\boxed{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2,}$$

   where $\bar{\hat{\theta}}$ is the sample mean of $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}\}$.

4. Similarly, $\tilde{\pi}_p(\hat{\theta})$ is estimated by calculating the sample $100p$-th percentile of $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)}\}$. In particular, by estimating $\tilde{\pi}_{\alpha/2}(\hat{\theta})$ and $\tilde{\pi}_{1-\alpha/2}(\hat{\theta})$, we obtain the $100(1-\alpha)\%$ nonparametric percentile bootstrap-based confidence interval (CI) of θ, $[\tilde{\pi}_{\alpha/2}(\hat{\theta}), \tilde{\pi}_{1-\alpha/2}(\hat{\theta})]$.

$\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(B)}\}$ approximates the distn of $\hat{\theta}$.
For 95% percentile bootstrap-based C.I. of θ, take 2.5th and 97.5th percentiles of $\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(B)}\}$.

To illustrate, let us consider how the nonparametric bootstrap works for a case with $m = 3$, namely, $\{x_1, x_2, x_3\}$. In such "small-sample" cases, there is no need to consider a large B. Indeed, there are only $3^3 =$ 27 possible unique resamples (when the order matters), which are $\{x_{i_1}, x_{i_2}, x_{i_3}\}$, with $i_1 \in \{1, 2, 3\}$, $i_2 \in \{1, 2, 3\}$, and $i_3 \in \{1, 2, 3\}$.

In general, there are m^m unique resamples of the same sample size when the order matters. As m gets large, m^m gets large rapidly. Hence, only sufficiently large B resamples are created, often allowing duplicate resamples. Lastly, when the order does not matter, there are $\binom{2m-1}{m}$ unique resamples, assuming that each observation is distinct.

Remarks: There are many variations of nonparametric bootstrap. Even though m pseudo-observations are created in each resample, sometimes it is better to have a different number of observations (say, r observations) to improve the accuracy. This type of method is typically referred to as "m out of n bootstrap".

Another kind of nonparametric bootstrap assigns different weights (probabilities of being drawn) for different observations. This type of method is typically referred to as "weighted bootstrap".

Implementation in R with $\{0.55, 0.72, -1.27, 0.03, -0.95, -0.49\}$ assuming $\hat{\theta} = \tilde{X}$:

```
sample <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
# sample size.
m <- length(sample)
theta <- median(sample)
# create B = 1000 resamples.
B <- 1000
# a vector that stores θ's from B resamples.
thetas <- double(B)
for (i in 1:B){
# we do this one by one (but less efficient this way).
resample <- sample(sample, size=m, replace=TRUE)
thetas[i] <- median(resample)
}
# variance estimation
mean((thetas-mean(thetas))^2)
# percentile estimation at 2.5% and 97.5%.
# this is the same as a 95% nonparametric percentile bootstrap-based CI.
quantile(thetas, probs=c(0.025, 0.975))
# bias estimation
mean(thetas)-theta
```

The implementation above uses a `for`-loop, which is quite slow. A better implementation using the `boot` function in the `boot` package can be found below:

```
library(boot)
sample <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
theta <- median(sample)
B <- 1000
# here, i is the random vector of indices.
# this i is necessary to randomly select observations.
med.fun <- function(x, i){return(median(x[i]))}
results <- boot(data=sample, statistic=med.fun, R=B)
thetas <- results$t
# variance estimation
mean((thetas-mean(thetas))^2)
# percentile estimation at 2.5% and 97.5%.
# this is the same as a 95% nonparametric percentile bootstrap-based CI.
quantile(thetas, probs=c(0.025, 0.975))
# bias estimation
mean(thetas)-theta
```

The last part estimates the bias of an estimator $\hat{\theta}$, where the bias is given by

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

We estimate bias$(\hat{\theta})$ by

$$\widehat{\text{bias}}(\hat{\theta}) = \bar{\hat{\theta}} - \hat{\theta},$$

where $\bar{\hat{\theta}}$ is the sample mean of $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$.

Unbiasedness (i.e., $\hat{\theta}$ such that bias$(\hat{\theta}) = 0$, or equivalently, $E[\hat{\theta}] = \theta$) is a popular criterion for judging the quality of $\hat{\theta}$. Hence, a small $\widehat{\text{bias}}(\hat{\theta})$ would indicate that $\hat{\theta}$ is approximately unbiased. A popular threshold for judging unbiasedness is $\widehat{\text{bias}}(\hat{\theta})/\sqrt{\widehat{\text{Var}}(\hat{\theta})} \leq 0.25$ where $\widehat{\text{Var}}(\hat{\theta})$ is estimated variance of $\hat{\theta}$.

Examples of Unbiased Estimators: Sample mean ($\bar{X}$), sample variance (S^2).

2. (Semi-)Parametric Bootstrap

The idea behind parametric bootstrap is to create resamples from some assumed (or known) distribution. Let $\{x_1, \ldots, x_m\}$ be the set of observations, which we believe comes from a particular distribution (such as normal, Student's t, uniform, etc., with all the parameter values specified). The following algorithm gives estimated Var$(\hat{\theta})$ and $\tilde{\pi}_p(\hat{\theta})$.

Algorithm for Estimating Var$(\hat{\theta})$ and $\tilde{\pi}_p(\hat{\theta})$:

1. Create B resamples of $\{x_1, \ldots, x_m\}$ by generating resamples from the assumed distribution. Denote the b-th resample by $\{x_1^{(b)}, \ldots, x_m^{(b)}\}$, $b = 1, \ldots, B$.

2. For each of B resamples, calculate $\hat{\theta}$. Let $\hat{\theta}^{(b)}$ be $\hat{\theta}$ from the b-th resample.

3. Var$(\hat{\theta})$ is estimated by

$$\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2,$$

   where $\bar{\hat{\theta}}$ is the sample mean of $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$.

4. Similarly, $\tilde{\pi}_p(\hat{\theta})$ is estimated by calculating the sample $100p$-th percentile of $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$. In particular, by estimating $\tilde{\pi}_{\alpha/2}(\hat{\theta})$ and $\tilde{\pi}_{1-\alpha/2}(\hat{\theta})$, we obtain the $100(1-\alpha)\%$ parametric percentile bootstrap-based confidence interval (CI) of θ, $[\tilde{\pi}_{\alpha/2}(\hat{\theta}), \tilde{\pi}_{1-\alpha/2}(\hat{\theta})]$.

In the first step, we may use functions in `R` to easily generate resamples. For example, if we believe that the observations are coming from the normal distribution, `rnorm()` can be used. Unless the parameter values are known, they need to be estimated from the data. For the normal distribution, mean and variance (or standard deviation) needs to be estimated first. If parameter estimation is necessary, then we have *semiparametric bootstrap*.

Implementation in R with $\{0.55, 0.72, -1.27, 0.03, -0.95, -0.49\}$ assuming $\hat{\theta} = \tilde{X}$ and that the observations are coming from the normal distribution with unknown parameter values:

```
sample <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
# sample size.
m <- length(sample)
theta <- median(sample)
# create B = 1000 resamples.
B <- 1000
# a vector that stores θ's from B resamples.
thetas <- double(B)
for (i in 1:B) {
# we do this one by one (but less efficient this way).
resample <- rnorm(m, mean=mean(sample), sd=sd(sample))
thetas[i] <- median(resample)
}
# variance estimation
mean((thetas-mean(thetas))^2)
# percentile estimation at 2.5% and 97.5%.
# same as a 95% semiparametric percentile bootstrap-based CI.
quantile(thetas, probs=c(0.025, 0.975))

# bias estimation
mean(thetas)-theta
```

A better implementation using the `replicate` function:

```
sample <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
m <- length(sample)
theta <- median(sample)
mu <- mean(sample)
sigma <- sd(sample)
B <- 1000
thetas <- replicate(B, expr = {
x <- rnorm(m, mean = mu, sd = sigma)
median(x)
})
# variance estimation
mean((thetas-mean(thetas))^2)
# percentile estimation at 2.5% and 97.5%.
# same as a 95% semiparametric percentile bootstrap-based CI.
quantile(thetas, probs=c(0.025, 0.975))
# bias estimation
mean(thetas)-theta
```

Remark: Bootstrap may fail in certain cases, especially when the extremes (maximum or minimum) are considered. The most famous example is variance or percentile estimation of θ where the random variables are uniformly distributed between 0 and some unknown $\theta > 0$. A remedy is to use a variation of bootstrap, such as m out of n bootstrap.

(Semi-)Parametric Bootstrap vs. Nonparametric Bootstrap: Intuitively, it appears that nonparametric bootstrap is better because essentially no distributional assumption is necessary. However, this is not always the case. In particular, if the large-sample distribution (a.k.a. the *asymptotic distribution*) of the estimator $\hat{\theta}$ is known, it may be better to use the (semi-)parametric bootstrap assuming that large-sample distribution. This is because the (semi-)parametric bootstrap-based confidence interval or hypothesis test tends to be more reliable/robust than the nonparametric bootstrap-based ones depending on the situation. Another consideration is to consider a weighted bootstrap method.

Nevertheless, nonparametric bootstrap is still widely used due to its applicability.

3. Jackknife (Section 7.2 in textbook)

Jackknife is another resampling method which utilizes the idea of "leave-one-out" (similar to the one used in Cook's distance in Math 342, and also similar to the idea of cross-validation for model selection). Jackknife is mainly used for bias and variance estimation. Let $\{x_1, \ldots, x_m\}$ be a set of observations. The following algorithm gives estimated bias$(\hat{\theta})$ and Var$(\hat{\theta})$.

Algorithm for Estimating bias$(\hat{\theta})$ and Var$(\hat{\theta})$:

1. Create m resamples of $\{x_1, \ldots, x_m\}$ by removing one observation at a time. Denote the i-th resample by $x_{(i)} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m\}$, $i = 1, \ldots, m$, which omits the i-th observation.

2. For each of m resamples, calculate $\hat{\theta}$. Let $\hat{\theta}_{(i)}$ be $\hat{\theta}$ from the i-th resample.

3. bias$(\hat{\theta})$ is estimated by

$$(m-1)(\bar{\hat{\theta}} - \hat{\theta})$$

   where $\bar{\hat{\theta}}$ is the sample mean of $\{\hat{\theta}_{(1)}, \ldots, \hat{\theta}_{(m)}\}$.

4. Similarly, Var$(\hat{\theta})$ is estimated by

$$\frac{m-1}{m} \sum_{i=1}^{m} \left(\hat{\theta}_{(i)} - \bar{\hat{\theta}} \right)^2 .$$

Some mathematical justification:

Remark: Unlike (semi-)parametric and nonparametric bootstrap, the jackknife method may fail when $\hat{\theta}$ is not "smooth". We call $\hat{\theta}$ smooth when small changes in the data correspond to small changes in the value of $\hat{\theta}$. Although the sample mean $\bar{x}$ is considered smooth, the sample median $\tilde{x}$ is not.

The implementation in R below illustrates the failure.

```
sample <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
theta <- median(sample)
# Jackknife estimation
m <- length(sample)
thetas <- double(m)
# Leave-one-out algorithm
for(i in 1:m){
thetas[i] <- median(sample[-i])
}
# variance estimation
((m-1)/m)*sum((thetas-mean(thetas))^2)
# bias estimation
(m-1)*(mean(thetas)-theta)
```

Because median is not smooth, the jackknife estimate of Var(θ) is quite different from the bootstrap estimate. Similar to the bootstrap case, a variation of jackknife method, such as the delete-d jackknife (that omits d observations in each resample) can be applied to avoid such problem.

Remark: Even though we have discussed resampling methods based on $\hat{\theta}$ estimated from a single sample, it is possible to apply similar methods for the multivariate parameter/sample cases. For example, we may think about estimating variance or constructing a 95% confidence interval of the ratio of two means by looking at $\hat{\theta} = \bar{X}/\bar{Y}$, where $\bar{x}/\bar{y}$ is the ratio of sample means from two independent or paired samples.

Section 7.4.4: The Bootstrap t Interval

In the last section, we learned how to construct (semi-)parametric and nonparametric percentile bootstrap confidence intervals. Instead, we may think about constructing a t-type table based on resamples. Using the t-type table, we may construct confidence intervals by checking appropriate percentiles of the constructed table. Specifically, a $100(1-\alpha)\%$ bootstrap t confidence interval is given by

$$\left[\hat{\theta} - t^{\star}_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\theta})}, \hat{\theta} - t^{\star}_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\theta})}\right],$$

where $t^{\star}_{1-\alpha/2}$ and $t^{\star}_{\alpha/2}$ are the $100(1-\alpha/2)$-th and $100(\alpha/2)$-th percentiles of the constructed t-type table, respectively.

Remark: Note that the definition of $t^{\star}_{\alpha/2}$ is the $100(\alpha/2)$-th percentiles, whereas typically $t_{\alpha/2,m-1}$ is defined as the $100(1-\alpha/2)$-th percentile of the t-distribution with $m-1$ degrees of freedom.

Derivation for the case $\hat{\theta} = \bar{X}$: Let $X_1, X_2, \ldots, X_m$ be i.i.d. random variables with $\mu = E[X_i]$ and $\sigma^2 = \text{Var}(X_i)$. Recall that the t-statistic is given by

$$T = \frac{\bar{X} - \mu}{S/\sqrt{m}},$$

where $\bar{X} = \sum_{i=1}^m X_i/m$ is the sample mean and $S^2 = m\widehat{\text{Var}}(\bar{X}) = \sum_{i=1}^m (X_i - \bar{X})^2/(m-1)$ is the sample variance. When $X_i \sim N(\mu, \sigma^2)$, $T \sim t_{m-1}$ (t-distributed with $m-1$ degrees of freedom) so that by taking $t^\star_{\alpha/2}$ and $t^\star_{1-\alpha/2}$, which are the $100(\alpha/2)$-th and $100(1-\alpha/2)$-th percentiles of the t distribution with $m-1$ degrees of freedom, respectively,

$$\begin{aligned}
1 - \alpha &= P\left(t^\star_{\alpha/2} \leq T \leq t^\star_{1-\alpha/2}\right) \\
&= P\left(t^\star_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{m}} \leq t^\star_{1-\alpha/2}\right) \\
&= P\left(t^\star_{\alpha/2}\frac{S}{\sqrt{m}} \leq \bar{X} - \mu \leq t^\star_{1-\alpha/2}\frac{S}{\sqrt{m}}\right) \\
&= P\left(t^\star_{\alpha/2}\frac{S}{\sqrt{m}} - \bar{X} \leq -\mu \leq t^\star_{1-\alpha/2}\frac{S}{\sqrt{m}} - \bar{X}\right) \\
&= P\left(-t^\star_{\alpha/2}\frac{S}{\sqrt{m}} + \bar{X} \geq \mu \geq -t^\star_{1-\alpha/2}\frac{S}{\sqrt{m}} + \bar{X}\right) \\
&= P\left(\bar{X} - t^\star_{1-\alpha/2}\frac{S}{\sqrt{m}} \leq \mu \leq \bar{X} - t^\star_{\alpha/2}\frac{S}{\sqrt{m}}\right),
\end{aligned}$$

implying that

$$\left[\bar{x} - t^\star_{1-\alpha/2}\frac{s}{\sqrt{m}}, \bar{x} - t^\star_{\alpha/2}\frac{s}{\sqrt{m}}\right]$$

is a valid $100(1-\alpha)\%$ confidence interval. Noting that $t^\star_{1-\alpha/2} = t_{\alpha/2,m-1}$ and $t^\star_{\alpha/2} = -t_{\alpha/2,m-1}$, it also implies that the usual $100(1-\alpha)\%$ confidence interval

$$\left[\bar{x} - t_{\alpha/2,m-1}\frac{s}{\sqrt{m}}, \bar{x} + t_{\alpha/2,m-1}\frac{s}{\sqrt{m}}\right]$$

is equivalent to the one above.

For the general case $\hat{\theta}$, the distribution of

$$T_{\hat{\theta}} = \frac{\hat{\theta} - \theta}{\sqrt{\widehat{\text{Var}}(\hat{\theta})}}$$

is unknown, thus requiring a resampling method to get $t_{\hat{\theta},1-\alpha/2}$ and $t_{\hat{\theta},\alpha/2}$, the $100(1-\alpha/2)$-th and $100(\alpha/2)$-th percentiles of the distribution for $T_{\hat{\theta}}$.

This may be achieved by the following bootstrap, often referred to as the studentized bootstrap.

Algorithm for Studentized bootstrap:

1. Calculate $\hat{\theta}$ from the original sample $\{x_1, \ldots, x_m\}$.
2. Create B resamples of $\{x_1, \ldots, x_m\}$ (by parametric or nonparametric bootstrap). Denote the b-th resample by $\{x_1^{(b)}, \ldots, x_m^{(b)}\}$, $b = 1, \ldots, B$.
3. For each of B resamples, calculate $\hat{\theta}$. Let $\hat{\theta}^{(b)}$ be $\hat{\theta}$ from the b-th resample.
   (a) Now, estimate $\text{Var}(\hat{\theta}^{(b)})$. To achieve this, a second set of bootstrap resamples from each of the B resamples is necessary. Specifically, let $\{x_1^{(b,c)}, \ldots, x_m^{(b,c)}\}$, $c = 1, \ldots, C$ be the c-th resample of the b-th resample, and let $\hat{\theta}^{(b,c)}$ be the estimate of θ from that sample. Then, the estimated $\text{Var}(\hat{\theta}^{(b)})$ is given by

   $$\widehat{\text{Var}}(\hat{\theta}^{(b)}) = \frac{1}{C}\sum_{c=1}^{C}\left(\hat{\theta}^{(b,c)} - \bar{\hat{\theta}}^{(b)}\right)^2,$$

   where $\bar{\hat{\theta}}^{(b)}$ is the sample mean of $\{\hat{\theta}^{(b,1)}, \ldots, \hat{\theta}^{(b,C)}\}$.

   (b) Then, calculate the t-type statistic given by $t^{(b)} = \frac{\hat{\theta}^{(b)} - \hat{\theta}}{\sqrt{\widehat{\text{Var}}(\hat{\theta}^{(b)})}}$.
4. Find $t^{\star}_{1-\alpha/2}$ and $t^{\star}_{\alpha/2}$, the $100(1-\alpha/2)$-th and $100(\alpha/2)$-th percentiles of $\{t^{(1)}, t^{(2)}, \ldots, t^{(B)}\}$.
5. A $100(1-\alpha)\%$ confidence interval is given by

   $$\left[\hat{\theta} - t^{\star}_{1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\theta})}, \hat{\theta} - t^{\star}_{\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{\theta})}\right],$$

   where

   $$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \bar{\hat{\theta}}\right)^2.$$

As long as B and C are relatively large, $t^{\star}_{1-\alpha/2} \approx t_{\hat{\theta},1-\alpha/2}$ and $t^{\star}_{\alpha/2} \approx t_{\hat{\theta},\alpha/2}$.

However, because $B \times C$ number of resamples are required, bootstrap t method is much more computationally expensive. A practical choice may be $B = 1000$ and $C = 200$. See Example 7.12 of `http://personal.bgsu.edu/~mrizzo/SCR/SCRch7.R` for an implementation.

7.5: Better Bootstrap Confidence Intervals

- Discusses how bias corrected accelerated (BCa) confidence interval works.
- Discusses its implementation in R.

We have learned that the empirical Type I error rate of the hypothesis test/coverage rate of the confidence interval is mainly affected by two things, namely:

- the sample size, and
- skewness.

The usual percentile bootstrap does not address any of the problems. Therefore, despite its simplicity and easy interpretation, it typically does not have a good coverage rate. Therefore, it is desirable to have a more sophisticated version of the percentile bootstrap to overcome the problem with skewness. (There is nothing we can do about the sample size, unfortunately.)

The idea behind the bias corrected accelerated (BCa) confidence interval addresses the second problem, skewness (acceleration), in addition to bias correction (and hence the name BCa). Because we are trying to extend the idea of percentile bootstrap, it will have a form $[\hat{\theta}^\star_{\alpha_1}, \hat{\theta}^\star_{\alpha_2}]$, where $\hat{\theta}^\star_{\alpha_i}$ is the (lower) α_i-th percentile of the distribution of $\hat{\theta}^\star$, $i = 1, 2$.

Graphical representation.

Although the details are omitted, after bias correction and skewness adjustment are correctly taken into account, better lower and upper percentiles of the $100(1-\alpha)\%$ BCa confidence interval, denoted by α_1 and α_2, respectively, are given by

$$\begin{aligned}
\alpha_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z_{\alpha/2}}{1 + \hat{a}(\hat{z}_0 - z_{\alpha/2})}\right), \\
\alpha_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 + \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right),
\end{aligned}$$

where $\Phi(x)$ is the cdf of the standard normal distribution evaluated at x.

Moreover,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}^{(b)} < \hat{\theta})\right),$$

is an estimated bias (i.e., estimate of median bias of $\hat{\theta}^{(b)}$ for $\hat{\theta}$), and

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\hat{\theta}} - \theta_{(i)})^3}{6\sum_{i=1}^{n}[(\bar{\hat{\theta}} - \theta_{(i)})^2]^{3/2}}$$

estimates skewness.

The BCa bootstrap confidence interval method is implemented in many `R` packages, such as the `boot.ci()` function in the `boot` package. Detailed implementations can be found in Examples 7.14–7.16 of

`http://personal.bgsu.edu/~mrizzo/SCR/SCRch7.R`

Also, see

http://www.statmethods.net/advstats/bootstrapping.html

for an example in which confidence intervals for the coefficient of determination (R^2), which measures the strength of linear relationship between the two variables.

7.Misc: Resampling-Based Hypothesis Tests

- Discusses resampling-based hypothesis tests for one-, two- and paired-sample cases.
- Shows why the paired-sample test is necessary and better for paired observations.
- Introduces the Kolmogorov-Smirnov test for equality of distributions.
- Discusses how to estimate Type I error rate, power, and coverage rate for those resampling-based tests.

In the statistics course(s) that you have taken, you have seen a number of tests, including, one- and (independent) two-sample t-tests, and possibly χ^2 tests, ANOVA F-tests, and a few others. Even though the critical values and p-values of these tests are often calculated based on some distribution (such as normal, t, χ^2, and F), that distribution is derived based on some assumptions of the data. The most common set of assumptions is that the we have i.i.d. observations from the normal distribution with common variance.

It is not clear how robust these tests are when these assumptions are violated. Thus, we may wish to obtain more robust tests using resampling methods.

This section illustrates how parametric bootstrap, nonparametric bootstrap, and/or permutation can be used for various hypothesis tests, including one, two, and paired sample cases, and equality of distributions. Furthermore, the same kind of idea can be applied to a wide range of hypothesis tests and confidence intervals.

One-Sample Case

Let $X_1, X_2, \ldots, X_m \overset{i.i.d.}{\sim} F$ be a random sample satisfying $E[X_i^4] < \infty$. Here, F denotes some distribution of X. Let $\mu = E[X_i]$ and $\sigma^2 = \text{Var}(X_i)$. To test

$$H_0 \colon \mu = \mu_0 \text{ vs. } H_1 \colon \mu \neq \mu_0,$$

where μ_0 is some constant, we consider the t-test statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{m}},$$

where $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$.

Assuming that H_0 is true, by the Central Limit Theorem (CLT), as $m \to \infty$, $T \sim N(0,1)$ *asymptotically*. In other words, when m is sufficiently large, T is approximately standard normal under H_0.

The idea behind resampling-based hypothesis tests is that a number of test statistics (denoted by $T^{(1)}, T^{(2)}, \ldots, T^{(B)}$ are calculated under the assumption that H_0 is true, from which the approximate distribution of T is obtained. Then, based on the approximate distribution of T, resampling-based p-values can be calculated.

Note: Because the set of resamples may differ, the resampling-based p-value may vary even if you calculate that using the same dataset. However, the variation will be small as long as B is sufficiently large ($B = 10,000$ or more). To completely replicate the results, `set.seed()` can be used.

Note 2: *Permutation* is a resampling method where resamples are created by sampling observations **without** replacement. For the one-sample case, permutation is meaningless.

Algorithm for the Resampling-Based One-Sample t-Test:

1. Calculate the t-test statistic T from the original observations.
2. Create B resamples.
   (a) In the case of semiparametric bootstrap, generate resamples from $N(\mu_0, s^2)$.
   (b) In the case of nonparametric bootstrap, generate resamples from the modified observations
   $$\{x_1^\star, x_2^\star, \ldots, x_m^\star\},$$
   where $x_i^\star = x_i - \bar{x} + \mu_0$, by randomly choosing modified observations with replacement.

   Denote the b-th resample by $\boldsymbol{x}^{(b)} = \{x_1^{(b)}, \ldots, x_m^{(b)}\}$.
3. For each of B resamples, calculate the t-test statistic T. Let $T^{(b)}$ be the t-test statistic calculated using the b-th resample $\boldsymbol{x}^{(b)}$.
4. The p-value is given by
   $$p = \frac{1}{B}\sum_{b=1}^{B} I(|T^{(b)}| \geq |T|).$$

Remark: The algorithm above works for the two-tailed alternative case $H_1\colon \mu_X \neq \mu_0$. In the one-tailed alternative cases, proper adjustments are necessary for the calculations of p-values.

Remark 2: A different way of calculating the p-value is given by $p = \min\{2p_1, 2(1-p_1)\}$, where

$$p_1 = \frac{1}{B}\sum_{b=1}^{B} I(T^{(b)} \geq T).$$

Explanation:

An implementation of the nonparametric bootstrap one-sample t-test using the `replicate()` function is given below:

```
x <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
m <- length(x)
mu0 <- 0
B <- 1000
abs.tstat <- abs(t.test(x, mu=mu0)$statistic)
# Student's t-distribution based p-value
t.test(x, mu=mu0)$p.value
# create resamples
x.star <- x - mean(x) + mu0
xb <- sample(x.star, (m*B), replace=TRUE)
# store the b-th random sample in the b-th row of the matrix.
xbmat <- matrix(xb, nrow=B, ncol=m)
# apply() function calculates the test statistic value for each row.
abs.tb <- abs(apply(xbmat, 1, function(x){t.test(x, mu=mu0)$statistic}))
# p-value calculation
p.value <- mean(abs.tb >= abs.tstat)
p.value
```

Note: There is no simple answer to the question "which resampling method works best in hypothesis testing (or other things, such as confidence intervals)?" In research articles, performances of various resampling methods are compared in terms of Type I error rate and power of the test under various combinations of distributions, sample sizes, and alternatives. The choices are somewhat arbitrary, but they need to be realistic. Typically, one can find a method that works well under many different circumstances by running a number of Monte Carlo simulations for the empirical Type I error rate and power calculations.

Independent Two-Sample Case

Let

$$X_1, X_2, \ldots, X_{m_X} \overset{i.i.d.}{\sim} F_X \text{ and } Y_1, Y_2, \ldots, Y_{m_Y} \overset{i.i.d.}{\sim} F_Y$$

be two independent samples satisfying $E[X_i^4] < \infty$ and $E[Y_i^4] < \infty$. Here, F_X and F_Y denote distribution of X and Y, respectively. Let $\mu_X = E[X_i]$, $\mu_Y = E[Y_i]$, $\sigma_X^2 = \text{Var}(X_i)$, and $\sigma_Y^2 = \text{Var}(Y_i)$. To test

$$H_0\colon \mu_X = \mu_Y \text{ vs. } H_1\colon \mu_X \neq \mu_Y,$$

we consider the t-test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m_X} + \frac{S_Y^2}{m_Y}}},$$

where $\bar{X} = \frac{1}{m_X}\sum_{i=1}^{m_X} X_i$, $\bar{Y} = \frac{1}{m_Y}\sum_{i=1}^{m_Y} Y_i$, $S_X^2 = \frac{1}{m_X-1}\sum_{i=1}^{m_X}(X_i - \bar{X})^2$, and $S_Y^2 = \frac{1}{m_Y-1}\sum_{i=1}^{m_Y}(Y_i - \bar{Y})^2$.

Assuming that H_0 is true, by the Central Limit Theorem (CLT), as $\min\{m_X, m_Y\} \to \infty$ (same as $m_X \to \infty$ and $m_Y \to \infty$) with $m_X/m_Y \to \lambda \in (0, \infty)$ (a technical condition necessary in CLT), T is asymptotically standard normal ($N(0, 1)$). In other words, when m_X and m_Y are sufficiently large, T is approximately standard normally distributed under H_0.

Some remarks on the two-sample t-test:

Permutation may yield a better test (in terms of Type I error rate and power) than the conventional or bootstrap-based tests sometimes in two- or multi-sample setting.

Assumption (for Permutation): Equality of distributions ($F_X = F_Y$) under H_0, where F_X and F_Y are the cdfs of X and Y, respectively.

Rationale: Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_{m_X}\}$ and $\boldsymbol{y} = \{y_1, y_2, \ldots, y_{m_Y}\}$ be observations from the two independent samples. Then, under the assumption $F_X = F_Y$,

$$(\boldsymbol{x}, \boldsymbol{y}) = \{x_1, x_2, \ldots, x_{m_X}, y_1, y_2, \ldots, y_{m_Y}\}$$

forms an i.i.d. sample. For convenience, let $x_{m_X+1} = y_1, x_{m_X+2} = y_2, \ldots, x_{m_X+m_Y} = y_{m_Y}$. That is,

$$(\boldsymbol{x}, \boldsymbol{y}) = \{x_1, x_2, \ldots, x_{m_X}, x_{m_X+1}, x_{m_X+2}, \ldots, x_{m_X+m_Y}\}.$$

To generate permutation-based resamples, consider the permutation operation $\pi(\cdot)$ where $\pi(i) = j$ for some integers $i, j \in \{1, 2, \ldots, m_X + m_Y\}$ such that $\pi(i_1) = \pi(i_2)$ if and only if $i_1 = i_2$. An example is given below for the case $m_X = 2$ and $m_Y = 3$. That is, we have $\boldsymbol{x} = \{x_1, x_2\}$ and $\boldsymbol{y} = \{y_1, y_2, y_3\} = \{x_3, x_4, x_5\}$ as the original samples.

An Example:

Let

$$\{x_{\pi(1)}, x_{\pi(2)}, x_{\pi(3)}, x_{\pi(4)}, x_{\pi(5)}\} = \{x_3, x_2, x_4, x_5, x_1\},$$

where $\pi(1) = 3$, $\pi(2) = 2$, $\pi(3) = 4$, $\pi(4) = 5$, $\pi(5) = 1$, which is a possible permutation.

(i) Consider the resulting resample $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$, and

(ii) Calculate the number of possible permutations when there are $m_X + m_Y$ observations available:

The number of possible permutations grow rapidly as m_X and/or m_Y increases. In practice, only B resamples are used (just like bootstrap), often allowing duplicate resamples.

Note: The assumption $F_X = F_Y$ implies $H_0\colon \mu_X = \mu_Y$ as long as the expectation exists. However, the converse is not true. The most popular example is when $\mu_X = \mu_Y$ with $\sigma_X^2 \neq \sigma_Y^2$, which is commonly known as the Behrens-Fisher situation.

Note 2: There are cases where the use of permutation can be justified (asymptotically) without the assumption $F_X = F_Y$, but not for the small-sample cases. The details are beyond the scope of this course.

The Behrens-Fisher situation:

Algorithm for the Resampling-Based Two-Sample t-Test:

1. Calculate the t-test statistic T from the original observations.

2. Create B resamples.

   (a) In the case of permutation, use the original observations $\boldsymbol{x} = \{x_1, \ldots, x_{m_X}\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_{m_Y}\}$ to generate resamples.

   (b) In the case of semiparametric bootstrap, generate resamples using $N(0, s_X^2)$ (for $\boldsymbol{x}^{(b)}$'s) and $N(0, s_Y^2)$ (for $\boldsymbol{y}^{(b)}$'s).

   (c) In the case of nonparametric bootstrap, use $\boldsymbol{x}_0 = \{x_1 - \bar{x}, \ldots, x_{m_X} - \bar{x}\}$ and $\boldsymbol{y}_0 = \{y_1 - \bar{y}, \ldots, y_{m_Y} - \bar{y}\}$, where $\bar{x}$ and $\bar{y}$ denote the sample means, to generate resamples.

   Denote the b-th resample by $\boldsymbol{x}^{(b)} = \{x_1^{(b)}, \ldots, x_{m_X}^{(b)}\}$ and $\boldsymbol{y}^{(b)} = \{y_1^{(b)}, \ldots, y_{m_Y}^{(b)}\}$.

3. For each of B resamples, calculate the t-test statistic T. Let $T^{(b)}$ be the t-test statistic calculated using the b-th resample $\boldsymbol{x}^{(b)}$ and $\boldsymbol{y}^{(b)}$.

4. The p-value is given by

$$p = \frac{1}{B} \sum_{b=1}^{B} I(|T^{(b)}| \geq |T|).$$

   or $p = \min\{2p_1, 2(1 - p_1)\}$, where

$$p_1 = \frac{1}{B} \sum_{b=1}^{B} I(T^{(b)} \geq T).$$

Remark 1: Many textbooks recommend

$$p = \frac{1}{B+1}[1 + \sum_{b=1}^{B} I(|T^{(b)}| \geq |T|)]$$

instead for the calculation of p-value for the permutation method, although the difference diminishes for a large B.

Remark 2: It is very important to keep in mind that the bootstrap samples must be generated under the assumption of $H_0\colon \mu_X = \mu_Y$. This is often done by assuming that $\mu_X = \mu_Y = 0$. That is the reason why the sample means are subtracted from each sample to comply with the assumption. There is no need to do this for permutation because use of the permutation implies $F_X = F_Y$, which implies $\mu_X = \mu_Y$ (if they exist).

Remark 3: For the parametric bootstrap, if $F_X = F_Y$ is plausible, resamples may be generated from $N(0, s_p^2)$, where

$$s_p^2 = \frac{(m_X - 1)s_X^2 + (m_Y - 1)s_Y^2}{m_X + m_Y - 2}.$$

Otherwise, resamples must be generated separately as described in the algorithm above.

Remark 4: Similarly, for the nonparametric bootstrap, if $F_X = F_Y$ is plausible, resamples may be generated from $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Otherwise (such as $\sigma_X^2 \neq \sigma_Y^2$), resamples must be generated separately for $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$.

An implementation of the permutation method using the `replicate` function:

```
x <- c(0.55, 0.72, -1.27, 0.03, -0.95, -0.49)
y <- c(2.95, 1.79, 2.56)
m.x <- length(x)
m.y <- length(y)
abs.tstat <- abs(t.test(x, y)$statistic)
# Student's t-distribution based p-value
t.test(x,y)$p.value
xy<-c(x,y)
B <- 1000
tstats.permutation <- replicate(B, expr = {
xy.b <- sample(xy, (m.x + m.y), replace=FALSE)
x.b <- xy.b[1:m.x]
y.b <- xy.b[(m.x + 1):(m.x + m.y)]
abs(t.test(x.b, y.b)$statistic)
})
# Permutation p-value calculation
p.value <- mean(tstats.permutation >= abs.tstat)
p.value
```

Note: There is no simple answer to the question "which resampling method works best in hypothesis testing (or other things, such as confidence intervals)?" In research articles, performances of various resampling methods are compared in terms of Type I error rate and power of the test under (somewhat arbitrary) various combinations of distributions, sample sizes, and alternatives. Typically, one can find a method that works well under many different circumstances by running a number of Monte Carlo simulations for the empirical Type I error rate and power calculations.

Paired Case

A paired sample refers to a sample in which there are matched pairs of similar units or pairs observations coming from the same subjects under two different conditions. Simply put, if the two samples are dependent, it is likely that we are talking about the paired sample case.

To formulate a t-test, it is simplest to consider the difference of each pair and discuss in the context of one-sample t-test. To be more specific, let (X_i, Y_i), $i = 1, 2, \ldots, m$, be m pairs of observations, where different pairs are considered independent. Define $D_i = X_i - Y_i$, and let $\mu_D = E[D_i]$. To test

$$H_0\colon \mu_D = \mu_0 \text{ vs. } H_1\colon \mu_D \neq \mu_0,$$

where μ_0 is some constant, we consider the t-test statistic

$$T_D = \frac{\bar{D} - \mu_0}{S_D/\sqrt{m}},$$

where $\bar{D} = \frac{1}{m}\sum_{i=1}^{m} D_i$ and $S_D^2 = \frac{1}{m-1}\sum_{i=1}^{m}(D_i - \bar{D})^2$. If D_i are assumed to be normally distributed i.i.d. random variables, then under H_0, $T_D \sim t_{m-1}$.

Remark: The reason why it is not appropriate to think of this type of problem in the context of independent two-sample t-test is because of the variance of $\bar{D}$.

Some Explanations on the Effect of Dependence Between Two Samples:

1. If X_i and Y_i are independent:

$$\text{Var}(\bar{D}) = \text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_X^2 + \sigma_Y^2}{m}.$$

2. If X_i and Y_i are dependent:

$$\text{Var}(\bar{D}) = \text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2\text{Cov}(\bar{X}, \bar{Y}) = \frac{\sigma_X^2 + \sigma_Y^2}{m} - 2\text{Cov}(\bar{X}, \bar{Y}).$$

Because Cov$(\bar{X}, \bar{Y})$ is typically positive in real-life applications (i.e., positive correlation), Var$(\bar{D})$ is lower assuming dependence, increasing the power of the test.

Example: Taken from http://www.biostathandbook.com/pairedttest.html.

Wiebe and Bortolotti (2002) examined color in the tail feathers of northern flickers. Some of the birds had one "odd" feather that was different in color or length from the rest of the tail feathers, presumably because it was regrown after being lost. They measured the yellowness of one odd feather on each of 16 birds and compared it with the yellowness of one typical feather from the same bird.

There are two nominal variables, type of feather (typical or odd) and the individual bird, and one measurement variable, yellowness. Because these birds were from a hybrid zone between red-shafted flickers and yellow-shafted flickers, there was a lot of variation among birds in color, making a paired analysis more appropriate. They used the two-tailed hypotheses

$$H_0\colon \mu_D = 0 \text{ vs. } H_1\colon \mu_D \neq 0,$$

The data and an implementation is given below.

```
# x is for "typical feather", and y is for "odd feather".
x<-c(-0.255,-0.213,-0.19,-0.185,-0.045,-0.025,-0.015,
0.003,0.015,0.02,0.023,0.04,0.04,0.05,0.055,0.058)
y<-c(-0.324,-0.185,-0.299,-0.144,-0.027,-0.039,-0.264,
-0.077,-0.017,-0.169,-0.096,-0.33,-0.346,-0.191,-0.128,-0.182)
t.test(x,y,mu=0,paired=TRUE)
```

The results indicate that there is strong evidence against H_0 (p-value $\approx$ 0.001). Thus, they concluded that there is a significant difference in the yellowness. Note that we could also do the following instead, using the idea of one-sample t-test.

```
x<-c(-0.255,-0.213,-0.19,-0.185,-0.045,-0.025,-0.015,
0.003,0.015,0.02,0.023,0.04,0.04,0.05,0.055,0.058)
y<-c(-0.324,-0.185,-0.299,-0.144,-0.027,-0.039,-0.264,
-0.077,-0.017,-0.169,-0.096,-0.33,-0.346,-0.191,-0.128,-0.182)
d<-x-y
t.test(d,mu=0)
```

Question: How can we use resampling methods to perform a paired-sample t-test?

Answer: If we use the idea of permutation, we can create resamples by permuting observations within each pair.

Now, let us assume a permuted pair of observations, denoted by $(x_i^\star, y_i^\star)$. There are two possibilities: $(x_i^\star, y_i^\star) = (x_i, y_i)$ (order unchanged) or $(x_i^\star, y_i^\star) = (y_i, x_i)$ (order changed). If we let $d_i = x_i - y_i$, then $d_i^\star = x_i^\star - y_i^\star$ is either d_i or $-d_i$ with an equal chance. That is, essentially, the permutation paired-sample t-test randomly appends either a plus sign or minus sign to the differences, d_i, $i = 1, 2, \ldots, m$.

Keeping that in mind, the data and an implementation for the permutation test is given below using the `replicate()` function.

```
x<-c(-0.255,-0.213,-0.19,-0.185,-0.045,-0.025,-0.015,
0.003,0.015,0.02,0.023,0.04,0.04,0.05,0.055,0.058)
y<-c(-0.324,-0.185,-0.299,-0.144,-0.027,-0.039,-0.264,
-0.077,-0.017,-0.169,-0.096,-0.33,-0.346,-0.191,-0.128,-0.182)
d<-x-y
m<-length(d)
abs.tstat <- abs(t.test(d,mu=0)$statistic)
B <- 10000
tstats.permutation <- replicate(B, expr = {
sign <- sample(c(1,-1), m, replace=TRUE)
d.star <- sign*d
abs(t.test(d.star,mu=0)$statistic)
})
# Permutation p-value calculation
p.value <- mean(tstats.permutation >= abs.tstat)
p.value
```

The permutaion-based p-value is fairly close to the one based on the traditional approach in this case. However, if the normality of d_i cannot be assumed, the permutation approach may yield a more reliable result.

Remark: Of course, it is possible to apply nonparametric or parametric bootstrap to the paired sample case. However, because we know that the paired-sample t-test is equivalent to the one-sample t-test on the differences (d_i), bootstrap p-values can be done by generating resamples from $\{d_1, d_2, \ldots, d_m\}$ either parametrically or nonparametrically, making sure to respect the assumption made in H_0.

Remark 2: The idea of permutation within each pair can be justified in the context of equality of distributions, which states that the cdfs of X_i and Y_i are equal to each other. Clearly, this assumption implies $\mu_D = 0$ as long as the expectation exists.

Even though we have discussed many different ways of obtaining statistical results, “cherry-picking” p-values, known as “p-hacking” could lead to a number of problems. This will be discussed in class.

Testing Equality of Distributions

A popular test for checking equality of two distributions is the Kolmogorov-Smirnov test. Unlike the two-sample t-test, the Kolmogorov-Smirnov test has much wider applicability because they look at the maximum absolute difference between the empirical cdfs of the two samples.

Note that, if we have two samples (of possibly unequal sample sizes) $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$, we may define their empirical cdfs as

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le t) \qquad and \qquad G_m(t) = \frac{1}{m}\sum_{i=1}^{m} I(y_i \le t),$$

respectively, for $-\infty < t < \infty$.

The Kolmogorov-Smirnov test measures the maximum absolute difference between the two empirical cdfs. Specifically, its statistic can be expressed as

$$D = \sup_t |F_n(t) - G_m(t)|.$$

In practice, the maximum distance can be found by looking at the difference between F_n and G_m where the jump occurs. That is, by noting that the jumps occur at $t_1, t_2, \ldots, t_N$, $N \le m + n$,

$$D = \sup_{1 \le i \le N} |F_n(t_i) - G_m(t_i)|.$$

Illustration of the Kolmogorov-Smirnov test:

Even though it is known that, under $H_0\colon F = G$, the test statistic D converges to what is known at the Kolmogorov distribution as $\min\{n, m\} \to \infty$, in practice, we only have finite sample sizes. Therefore, a permutation test idea can be employed to come up with a more accurate (resampling-based) p-value.

How permutation works for the Kolmogorov-Smirnov test:

In R, we may use `ks.test()` to calculate the test statistic D for each of B resamples, from which the p-value can be calculated. Specifically, let D_0 be the test statistic from the original samples and $D^{(b)}$ be the test statistic from the b-th resamples. Then, the (approximate) p-value can be obtained by

$$p = \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} I(D^{(b)} \geq D_0)\right] \approx \frac{1}{B}\sum_{b=1}^{B} I(D^{(b)} > D_0).$$

These two equations will give an almost identical answer as long as B is sufficiently large.

An implementation is given below using `chickwts` data that contain information about weights (in grams) of six groups of newly hatched chicks. Those six groups have different feed supplements.

```
# prepare dataset
attach(chickwts)
x<-sort(as.vector(weight[feed=="soybean"]))
y<-sort(as.vector(weight[feed=="linseed"]))
detach(chickwts)

# exploratory analysis
par(mfrow=c(2,2))
hist(x,main="Soybean fed",freq=FALSE,breaks="scott",xlab="Weight")
hist(y,main="Linseed fed",freq=FALSE,breaks="scott",xlab="Weight")
plot(ecdf(x),main="Empirical CDF of Soybean fed",xlim=c(130,350),xlab="Weigh
plot(ecdf(y),main="Empirical CDF of Linseed fed",xlim=c(130,350),xlab="Weigh

# permutation part
B<-10000
# concatenate samples
z<-c(x,y)
x.length<-length(x)
y.length<-length(y)
z.length<-x.length + y.length
# Db is a vector that stores the $D$ statistic for each resample.
Db<-double(B)
options(warn = -1)
D0<-ks.test(x,y,exact=FALSE)$statistic

# create resamples and calculate $D$
for (i in 1:B){
zb<-sample(z,replace=FALSE)
xb<-zb[1:x.length]
yb<-zb[(x.length+1):z.length]
Db[i]<-ks.test(xb,yb,exact=FALSE)$statistic
}

# $p$-value calculation
p.value<-mean(Db>D0)
options(warn=0)
p.value
```

More on the Size and Power of the Test

In the last chapter, we discussed ways of estimating the Type I error rate and power of the test. Those estimates, called empirical Type I error rate (or empirical size of the test) and empirical power of the test, respectively, are calculated using the test statistics arising from the original sample and resamples. Specifically, if we let $T^{(b)}$ be the test statistic from the b-th resample consisting of m observations, $b = 1, \ldots, B$, then,

$$\frac{1}{B} \sum_{b=1}^{B} I(|T^{(b)}| > t_{\alpha/2,m-1})$$

approximates the true Type I error rate well if B is sufficiently large.

In the context of resampling method, we typically use p-values to calculate the empirical Type I error rate for resampling-based test.

Explanation:

Note: Even though the "test statistic" and "p-value" may be taught as two different things in the introductory level statistics course, because they have a one-to-one relationship, the p-value can actually be viewed as a "test statistic". It is a convenient test statistic because we know that any p-value is in $[0, 1]$, and moreover, under H_0,

$$p\text{-value} \sim \text{Uniform}(0, 1)$$

if the test statistic is continuous.

That is, if we calculate a number of p-values from the distribution under the assumption of H_0, then **the histogram of those p-values should look flat between 0 and 1**, as they are uniformly distributed.

Proof of p-value $\sim$ Uniform$(0, 1)$under H_0:

Exercise: Obtain $B = 10,000$ p-values of the one-sample t-test from B samples with $m = 10$ observations for testing $H_0\colon \mu = 0$ vs. $H_a\colon \mu \neq 0$. Generate these B samples from the standard normal distribution using `rnorm()`. Plot them in a histogram to empirically confirm the statement above.

The exercise above should make it clear that, under H_0, the probability that a p-value is less than α is exactly equal to α when the assumptions on the data are met and that the test statistics are continuous.

To illustrate it, consider the one-sample t-test case above where $H_0\colon \mu = \mu_0$ and $H_a\colon \mu \neq \mu_0$. Assuming that the data are coming from the normal distribution with mean μ_0, we have

$$\alpha = P(|T| > t_{\alpha/2,m-1}) = P(p_T < \alpha)$$

where

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{m}}$$

is the usual test statistic and p_T is the p-value that is calculated from that test statistic.

For the empirical Type I error rate calculation for resampling methods, we treat p-values as test statistics and perform a similar calculation. To do this, we need to think about a couple of things, namely:

- B, the number of resamples to calculate a p-value, and
- MC, the number of Monte Carlo simulations.

A sufficiently large B is required for an accurate p-value calculation, and a sufficiently large MC is required for an accurate estimation of Type I error rate.

Below, a general algorithm for the empirical Type I error rate calculation is described.

Algorithm for Estimating the Type I Error Rate for Resampling-Based Methods:

1. Generate MC random samples from a distribution under H_0. Let $T^{(mc)}$ be the original test statistic calculated from the mc-th set of random samples, $mc = 1, \ldots, MC$.

2. Let $mc = 1$.

3. Generate B resamples (by either (semi-)parametric bootstrap, nonparametric bootstrap, or by permutation) under H_0 using information about the mc-th set of random samples, and calculate the test statistic. Denote the b-th test statistic by $T^{(b,mc)}$, $b = 1, \ldots, B$.

4. Using those B test statistics $(T^{(1,mc)}, T^{(2,mc)}, \ldots, T^{(B,mc)})$ and the original test statistic for the mc-th set ($T^{(mc)}$), calculate the p-value. Let $p^{(mc)}$ be the p-value from the mc-th random sample.

5. If $mc < MC$, increment mc by 1 and go to Step 3. Otherwise, go to Step 6.

6. The empirical Type I error rate at the nominal significance level α is given by

$$\frac{1}{MC} \sum_{mc=1}^{MC} I(p^{mc} < \alpha).$$

Note 1: Empirical power calculations can be done in a similar way. The main difference appears in (1), where you generate MC random samples from a distribution under H_1. As before, we say that the test is *powerful* if the probability of correctly rejecting the null hypothesis is higher than the other tests being considered.

Note 2: Coverage rate calculations can also be done in a similar way. As before, we say that the confidence interval is *reliable* if the empirical coverage rate is close enough to the nominal confidence level $100(1 - \alpha)\%$.

Note 3: Even though the algorithm above was written for the one-sample case, a slight modification makes it applicable to two- and multi-sample cases with suitable test statistics.

Exercise: Write `R` code to check the empirical Type I error rate or empirical power of the permutation-based two-sample Kolmogorov-Smirnov test. Consider generating samples from the following distributions.

1. $X_i \sim N(0,1), i = 1, \ldots, 20, Y_i \sim N(0,1), i = 1, \ldots, 10$

2. $X_i \sim N(0,1), i = 1, \ldots, 10, Y_i \sim N(0,1), i = 1, \ldots, 10$

3. $X_i \sim N(0,1), i = 1, \ldots, 10, Y_i \sim N(0,2), i = 1, \ldots, 20$

4. $X_i \sim N(0,1), i = 1, \ldots, 10, Y_i \sim N(1,1), i = 1, \ldots, 20$