HYPOTHESIS TESTING WITH THE BOOTSTRAP

Ilir PALLA

Department of Mathematics, University of Korca, Albania
e-mail:pallailir@yahoo.com

Abstract: This article contains the bootstrap methods to test the equality of means of two random samples. Such a problem is called a two-sample problem. This method can be used even when both random samples have not a normal distribution or when are not paired. We have given the algorithms for computation of the achieved significance level of the test, are constructed from Bradley Efron, who is the first person to treat bootstrap methods. We have used R program to get results. We have simulated data of two known normal random samples to compare results with results that are given from bootstrap methods. A way to judge the acceptance or reject the null hypothesis we use the p-values. This method can be used even for many other hypothesis tests which we have not treated in this article.

Key words: *bootstrap tests, p-values, two-sample problem, null hypothesis.*

Introduction

We observe two independent random variables X and Y drawn from possibly different probability distributions F and G. $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_m)$ .We wish to test hypothesis H_0 of no difference between F and G, $H_0 : F = G$ **(1)**

The equality $F = G$ means that F and G assign equal probabilities to all sets, $\Pr ob\{A|F\} = \Pr ob\{A|G\}$ for any subset of the common sample space of the $X's$ and $Y's$. If H_0 is true, then there is no difference between the probabilistic behavior of a random X or a random Y. In the first part of the article is given the known cases of statistics. In the second part of the article is given the achieved significance level and bootstrap methods for hypothesis test. Finally I have taken some results using R program.

1. Random variables have normal distribution

We wish to test the null hypothesis $H_0 : \mu_X = \mu_Y$ versus $H_A : \mu_X \neq \mu_Y$.

a) If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, the variances are known. In this case we can use criteria $Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$, when null hypothesis H_0 is true, Z has standard normal distribution. Allowed area is $]-z_{\alpha/2}, z_{\alpha/2}[$, where $\Phi(z_{\alpha/2}) = (1-\alpha)/2$, Φ the Laplace's function, α the level of significance.

b) If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$, the variances are unknown but equal $T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)s_X^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim S(n+m-2)$, where H_0 is true. Allowed area is $]-t_{\alpha/2}, t_{\alpha/2}[$, where $t_{\alpha/2}$ is found in table of student distribution with $\alpha/2$ and $n+m-2$ degree of freedom.

c) If X and Y are paired, then we put random variable $Z = X - Y$, where $Z_i = X_i - Y_i$, $i = 1, \ldots, n$. Now we have $Z \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$ which can treated as in the case "one sample", student criterion.

d) Variances are unknown and unequal. We can use Welch criterion. $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \sim S(n_0)$, where

$$n_0 = \frac{s_X^2 / n + s_Y^2 / m \;^2}{s_X^2 / n \;^2 \;/\; n-1 \;+\; s_Y^2 / m \;^2 \;/\; m-1}$$

2. Hypothesis testing with the bootstrap

In all cases treated with the above assumptions we had about random variables to come from a normal distribution.

In this part we describe how bootstrap methods can be used to produce significance tests. The basic idea to test the hypothesis with the bootstrap method is to test the hypothesis, without mathematical assumptions.

The simplest situation involves a simple null hypothesis H_0 which completely specifies the probability distribution of the data. Thus, we are dealing with a single sample $X = (X_1, X_2, ..., X_n)$ from a population with CDF (cumulative distribution function) F, then H_0 specifies that $F = F_0$, where F_0 contains no unknown parameters, for example "exponential with mean 1". The more usual situation in practice is that H_0 is a composite null hypothesis, which means that some aspects of F are not determined and remain unknown when H_0 is true. An example: "normal with mean 5", the variance of the normal distribution being unspecified.

We observe two independent random samples $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_m)$ drawn from possibly different probability distribution F and G.

An example would be the data: measurements for cholesterol before treatment and after treatment.

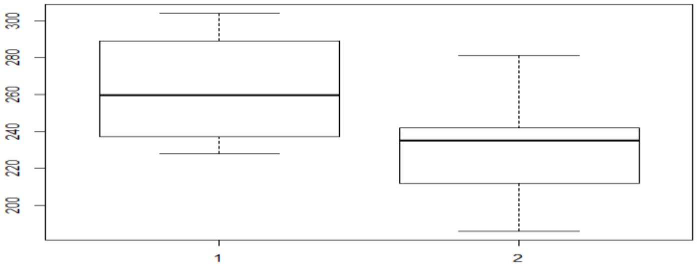

***Fig1**. Box plot for cholesterol data.*

The difference of the means, $\hat{\theta} = \bar{x} - \bar{y}$ =32, encourages us to believe that the treatment distribution F gives effect in lowering cholesterol. If we cannot decisively reject the possibility that H_0 is true, then we have not successfully demonstrated the treatment is effectively.

In our case the test statistic $\hat{\theta}$ is the difference of means. We will assume here that if the null hypothesis H_0 is not true, we expect to observe larger values of

$\hat{\theta}$, than if H_0 is true. All we say that the larger the value of $\hat{\theta}$ we observe, the stronger is the evidence against H_0.

The achieved significance level

The achieved significance level (ASL) of the test is defined to be the probability of observing at least that large a value when the null hypothesis is true (Bradley Efron)

$$ASL = \Pr ob\{\hat{\theta}^{*} \geq \hat{\theta} \mid H_0) \quad \textbf{(2)}$$

The quantity $\hat{\theta}$ is fixed at its observed value. The random variable $\hat{\theta}^{*}$ has the null hypothesis distribution, the distribution of $\hat{\theta}$ if H_0 is true. The star notation differentiates between the actual observation and a hypothetical generated according to null hypothesis. Where ASL is less than 0.1, we have borderline evidence against H_0, if ASL<0.05 reasonably strong evidence against H_0, if ASL<0.025 strong evidence against H_0, if ASL<0.01 very strong evidence against H_0.

If we are in situation the variances are equal and known

$$ASL = \Pr ob\{Z > \frac{\hat{\theta}}{\sqrt{\sigma^2{}_x / n + \sigma^2{}_y / m}}\} = 1 - \Phi(\frac{\hat{\theta}}{\sqrt{\sigma^2{}_x / n + \sigma^2{}_y / m}}). \quad \textbf{(3)}$$

```
### ASL calculation, when we know variances
ASL.a<-1-pnorm((t.obs)/(sqrt(var.x*(1/n.x)+var.y*(1/n.y))))
```

The variances are unknown, but equal:

$$F = N(\mu_X, \sigma^2)\ ,\ G = N(\mu_Y, \sigma^2) \quad \textbf{(4)}$$

$$H_0 : \hat{\theta} \sim N(0, \sigma^2(\frac{1}{n} + \frac{1}{m}))\ , \quad \textbf{(5)}$$

$$ASL = \Pr ob\{t_{df} > \frac{\hat{\theta}}{\bar{\sigma}\sqrt{1/n + 1/m}}\}, \quad \textbf{(6)}$$

Standard estimate for σ is $\bar{\sigma} = \sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2]/[n+m-2]}$

Variances are unknown and unequal. We can use R to find p-value. The Welch test

```
t.test(sim.x,sim.y,alternative="greater").
```

The bootstrap achieved significance level (ASL$_{boot)}$

A bootstrap hypothesis test is based on a test statistic $\hat{\theta}$. To emphasize that a test statistic need not be an estimate of a parameter, we denote it here by $T(Z)$. In our example we have $T(Z) = \bar{X} - \bar{Y}$. The achieved significance level is:

$$ASL = \Pr ob\{T(Z^{*}) \geq t(z) | H_0\} \quad \textbf{(7)}$$

The quantity $t(z)$ is fixed at its observed value and the random variable Z^{*} has a distribution specified by the null hypothesis H_0. Call this distribution F_0. Bootstrap hypothesis testing uses a "plug-in" style estimate for F_0. Denote the combined sample by Z and let its empirical distribution be $\hat{F}_0$, putting probability $1/(n+m)$ on each member of Z**.** Under H_0**,** $\hat{F}_0$ provides a nonparametric estimate of the common population that gave rise to both x and y.

Algorithm to compute ASL

1. Draw R samples of size $n+m$ with replacement from Z. Call the first n observation X^{*} and the remaining m observation Y^{*}.
2. Evaluate $T(Z^{*r}) = \bar{X}^{*} - \bar{Y}^{*}$, $r = 1,..., R$
3. Approximate ASL_{boot} by $A\hat{S}L_{boot} = number\{T(Z^{*r}) \geq t_{obs}\}/R$

Use R program

```
> before <-c(237,289,257,228,303,275,262,304,244,233); mean.bef=mean(before)
> after <- c(194,240,230,186,265,222,242,281,240,212);  mean.aft=mean(after)
> total<-c(before,after); mean.tot=mean(total)
> diff <- function(x) {mean(x[1:10]) - mean(x[11:20])}
> d<-bootstrap(total,theta=diff,nboot=999)
>hist(d$thetastar,freq = FALSE,col = "5",main="BOOTSTRAP REPLICATIONS",
xlab="bootstrap values")
> t.obs <- mean.bef-mean.aft;   abline(v=t.obs,col="red",lwd=3,lty=1)
> asl<-sum(d$thetastar >= t.obs)/nboot;  paste("ASL =",asl)
```

Achieved significance level $A\hat{S}L =$ 0.02<0.05, this mean that, we reject hypothesis with significance level α =0.05.

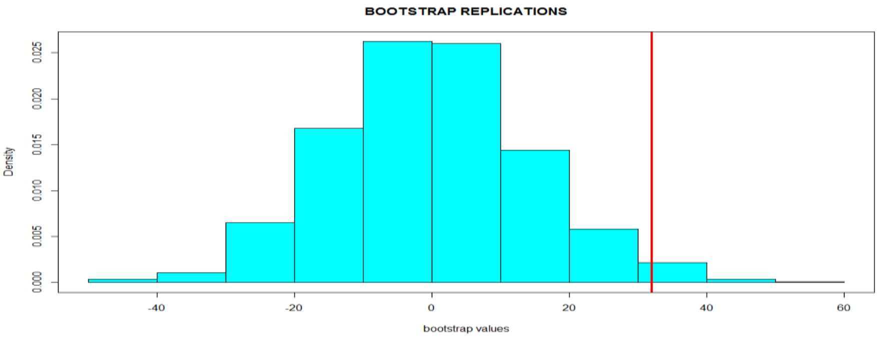

***Fig. 2.** Histogram of bootstrap replications of $\bar{X} - \bar{Y}$. Red line is observation value.*

P-value

P-values (A.C. Davison and D.V. Hinkley) $p_{boot} = Prob(T^{*} \geq t \,|\, \hat{F}_0)$ we can use in nonparametric bootstrap test. p_{boot} to approximate by

$$p = \frac{1 + number\{t^{*r} \geq t\}}{R+1}$$ using $t^{*1}, t^{*2}, ..., t^{*R}$ from R bootstrap samples.

Studentized statistic

More accurate testing can be obtained through the use of a studentized statistic.

Statistic test we take $$T(Z) = \frac{\bar{X} - \bar{Y}}{\bar{\sigma}\sqrt{1/n + 1/m}} \quad \textbf{(8)}$$

$\bar{\sigma}$ is the pooled estimate of standard error. Repeating the above bootstrap algorithm, use (8) and find $A\hat{S}L_{boot}$. Studentization does not affect the answer, it does produce a different value for $A\hat{S}L_{boot}$. However, in this particular approach to bootstrapping the two sample problem, the difference is typically quite small.Under the null hypothesis and assuming normal populations with equal variances, this has a Student's t distribution with $n+m-2$ degrees of freedom. If we are not willing to assume that the variances in the two population is equal, we could base the test on

$$T(Z) = \frac{\bar{X} - \bar{Y}}{\sqrt{\bar{\sigma}_x^2 / n + \bar{\sigma}_y^2 / m}} \quad , \qquad \textbf{(9)}$$

where $\bar{\sigma}_x^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$, $\quad \bar{\sigma}_y^2 = \sum_{i=1}^{m} (Y_i - \bar{Y})^2 / (m-1)$. With normal populations, the quantity (7) no longer has a Student's t distribution.

The equal variance assumption is attractive for the t-test. In considering a bootstrap hypothesis test for comparing the two means, there is no compelling reason to assume equal variances and hence we don't make that assumption. The algorithm I this case is:

1. Let $\hat{F}$ put equal probability on the points $\tilde{X}_i = X_i - \bar{X} + \bar{Z}$, $i = 1,...,n$, and $\hat{G}$ put equal probability on the points $\tilde{Y}_i = Y_i - \bar{Y} + \bar{Z}$, $i = 1,...,m$, where $\bar{X}$ and $\bar{Y}$ are the group means and $\bar{Z}$ is the mean of the combined sample.
2. Form R bootstrap data sets (X^*, Y^*) where X^* is sampled with replacement from $\tilde{X}_1,...,\tilde{X}_n$ and Y^* is sampled with replacement from $\tilde{Y}_1,...,\tilde{Y}_m$.
3. Evaluate $T(.)$ defined by (7) on each data set,

$$T(Z^{*r}) = \frac{\bar{X}^* - \bar{Y}^*}{\sqrt{\bar{\sigma}_x^{2*} / n + \bar{\sigma}_y^{2*} / m}} \text{ , } r = 1,..., R \qquad \textbf{(10)}$$

4. Approximate ASL_{boot} by $\hat{ASL}_{boot} = number\{T(Z^{*r}) \geq t_{obs}\} / R$, where $t_{obs} = t(z)$ is the observed value of the statistic.

In R this algorithm we can write commands:

```
sim.x <-rnorm(50,10,5) ; n.x=length(sim.x); mean.x=mean(sim.x);var.x=var(sim.x)
sim.y <- rnorm(40,10,7); n.y <- length (sim.y); mean.y <- mean(sim.y)
var.y <- var(sim.y);
 t.obs <- (mean.x-mean.y)/sqrt(var.x/n.x + var.y/n.y)
total <- c(sim.x,sim.y); mean.tot <- mean(total)
x.tilde = sim.x - mean.x + mean.tot;  y.tilde = sim.y - mean.y + mean.tot; R=999
mean.x.star <- var.x.star <- numeric(); mean.y.star <- var.y.star <- numeric()
t.star <- numeric()
### Start the bootstrap procedure
for (r in 1:R){
  x.tilde.star <- sample(x.tilde, replace=TRUE)
  mean.x.star[r] <- mean(x.tilde.star)
  var.x.star[r] = var(x.tilde.star)
  y.tilde.star <- sample(y.tilde, replace=TRUE)
  mean.y.star[r] <- mean(y.tilde.star)
  var.y.star[r] <- var(y.tilde.star)
  t.star[r] <- (mean.x.star[r]-mean.y.star[r])/
    sqrt(var.x.star[r]/n.x + var.y.star[r]/n.y)
}                              ### finish the bootstrap procedure
### Calculate the approximate ASL
ASL.star <- sum( t.star >= t.obs)/R
hist(t.star,freq = FALSE,col = "5", main="Bootstrap estimations",  xlab="Bootstrap
values")
abline(v=t.obs,col="red",lwd=3,lty=1);
 paste("ASL.star =",ASL.star)
```

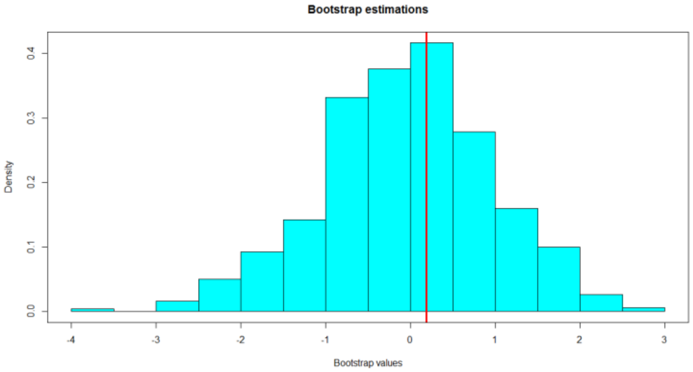

Fig.3 *Histogram of bootstrap replications.*

Achieved significance level ASL =0.41>0.05. We can't reject hypothesis with significance level α =0.05.

Practical simulation

Table 1. ASL values. $X \sim N(10,5)$ and $Y \sim N(10,7)$. Seven simulations.

Sim.	1	2	3	4	5	6	7
Knowvariance	0.4342	0.9157	0.1197	0.8036	0.4349	0.5478	0.5961
Welch test	0.4264	0.9692	0.0893	0.8886	0.4158	0.5551	0.6163

ASL_{boot}	0.4104	0.9659	0.0880	0.8998	0.4164	0.5965	0.6006

Conclusion

The main practical difficulty with hypothesis test comes in calculating the ASL, **(2)**. In most problems the null hypothesis **(1)**, $F = G$, leave us with a family of possible null hypothesis distributions, rather than just one. In normal case **(4),** for instance, the null hypothesis family **(5)** includes all normal distributions with expectation 0. In order to actually calculate the ASL, we had to either approximate the null hypothesis variance as in **(6)**, Student's method, but only applies to the normal situation. In considering a bootstrap hypothesis test for comparing the two means, there is no compelling reason to assume equal variances and hence we don't make that assumption.

References

- Bradley Efron and Robert J Tibshirani (1993): *An Introduction to the Bootstrap*. Chapman and Hall/CRC p: 202-233.

- Bradley Efron (1987): Journal of the American Statistical Association, March 1987, Vol. 82, No. 397, Theory and Methods , p. 171-174.

- Davison, A. C. and Hinkley, D. V. (1997): *Bootstrap Methods and Their Applications*. Cambridge University Press, p: 71-72, 159-160.

- Hall P. (1988): Theoretical Comparison of Bootstrap Confidence Intervals. *Ann. Statist*, p. 927-923.