

8

The Bootstrap

The **bootstrap** is a method for estimating standard errors and computing confidence intervals. Let $T_n = g(X_1, \dots, X_n)$ be a **statistic**, that is, T_n is any function of the data. Suppose we want to know $\mathbb{V}_F(T_n)$, the variance of T_n . We have written \mathbb{V}_F to emphasize that the variance usually depends on the unknown distribution function F . For example, if $T_n = \bar{X}_n$ then $\mathbb{V}_F(T_n) = \sigma^2/n$ where $\sigma^2 = \int (x - \mu)^2 dF(x)$ and $\mu = \int x dF(x)$. Thus the variance of T_n is a function of F . The bootstrap idea has two steps:

Step 1: Estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\hat{F}_n}(T_n)$.

Step 2: Approximate $\mathbb{V}_{\hat{F}_n}(T_n)$ using simulation.

For $T_n = \bar{X}_n$, we have for Step 1 that $\mathbb{V}_{\hat{F}_n}(T_n) = \hat{\sigma}^2/n$ where $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. In this case, Step 1 is enough. However, in more complicated cases we cannot write down a simple formula for $\mathbb{V}_{\hat{F}_n}(T_n)$ which is why we need Step 2. Before proceeding, let us discuss the idea of simulation.

8.1 Simulation

Suppose we draw an IID sample Y_1, \dots, Y_B from a distribution G . By the law of large numbers,

$$\bar{Y}_n = \frac{1}{B} \sum_{j=1}^B Y_j \xrightarrow{P} \int y dG(y) = \mathbb{E}(Y)$$

as $B \rightarrow \infty$. So if we draw a large sample from G , we can use the sample mean \bar{Y}_n to approximate $\mathbb{E}(Y)$. In a simulation, we can make B as large as we like, in which case, the difference between \bar{Y}_n and $\mathbb{E}(Y)$ is negligible. More generally, if h is any function with finite mean then

$$\frac{1}{B} \sum_{j=1}^B h(Y_j) \xrightarrow{P} \int h(y) dG(y) = \mathbb{E}(h(Y))$$

as $B \rightarrow \infty$. In particular,

$$\begin{aligned} \frac{1}{B} \sum_{j=1}^B (Y_j - \bar{Y})^2 &= \frac{1}{B} \sum_{j=1}^B Y_j^2 - \left(\frac{1}{B} \sum_{j=1}^B Y_j \right)^2 \\ &\xrightarrow{P} \int y^2 dF(y) - \left(\int y dF(y) \right)^2 = \mathbb{V}(Y). \end{aligned}$$

Hence, we can use the sample variance of the simulated values to approximate $\mathbb{V}(Y)$.

8.2 Bootstrap Variance Estimation

According to what we just learned, we can approximate $\mathbb{V}_{\hat{F}_n}(T_n)$ by simulation. Now $\mathbb{V}_{\hat{F}_n}(T_n)$ means “the variance of T_n if the distribution of the data is \hat{F}_n .” How can we simulate from the distribution of T_n when the data are assumed to have distribution \hat{F}_n ? The answer is to simulate X_1^*, \dots, X_n^* from \hat{F}_n and then compute $T_n^* = g(X_1^*, \dots, X_n^*)$. This constitutes one draw from the distribution of T_n . The idea is illustrated in the following diagram:

$$\begin{array}{llll} \text{Real world} & F & \implies & X_1, \dots, X_n \implies T_n = g(X_1, \dots, X_n) \\ \text{Bootstrap world} & \hat{F}_n & \implies & X_1^*, \dots, X_n^* \implies T_n^* = g(X_1^*, \dots, X_n^*) \end{array}$$

How do we simulate X_1^*, \dots, X_n^* from \hat{F}_n ? Notice that \hat{F}_n puts mass $1/n$ at each data point X_1, \dots, X_n . Therefore,

drawing an observation from \hat{F}_n is equivalent to drawing one point at random from the original data set.

Thus, to simulate $X_1^*, \dots, X_n^* \sim \hat{F}_n$ it suffices to draw n observations with replacement from X_1, \dots, X_n . Here is a summary:

Bootstrap Variance Estimation

1. Draw $X_1^*, \dots, X_n^* \sim \hat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \dots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \dots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2. \quad (8.1)$$

8.1 Example. The following pseudocode shows how to use the bootstrap to estimate the standard error of the median.

Bootstrap for The Median

Given data $X = (X(1), \dots, X(n))$:

```
T <- median(X)
Tboot <- vector of length B
for(i in 1:B){
  Xstar <- sample of size n from X (with replacement)
  Tboot[i] <- median(Xstar)
}
se <- sqrt(variance(Tboot))
```

The following schematic diagram will remind you that we are using two approximations:

$$\mathbb{V}_F(T_n) \overset{\text{not so small}}{\approx} \mathbb{V}_{\hat{F}_n}(T_n) \overset{\text{small}}{\approx} v_{\text{boot}}.$$

8.2 Example. Consider the nerve data. Let $\theta = T(F) = \int (x - \mu)^3 dF(x) / \sigma^3$ be the skewness. The skewness is a measure of asymmetry. A Normal distribution,

for example, has skewness 0. The plug-in estimate of the skewness is

$$\hat{\theta} = T(\hat{F}_n) = \frac{\int (x - \mu)^3 d\hat{F}_n(x)}{\hat{\sigma}^3} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^3}{\hat{\sigma}^3} = 1.76.$$

To estimate the standard error with the bootstrap we follow the same steps as with the median example except we compute the skewness from each bootstrap sample. When applied to the nerve data, the bootstrap, based on $B = 1,000$ replications, yields a standard error for the estimated skewness of .16. ■

8.3 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. Here we discuss three methods.

Method 1: The Normal Interval. The simplest method is the Normal interval

$$T_n \pm z_{\alpha/2} \hat{\mathbf{se}}_{\text{boot}} \quad (8.2)$$

where $\hat{\mathbf{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ is the bootstrap estimate of the standard error. This interval is not accurate unless the distribution of T_n is close to Normal.

Method 2: Pivotal Intervals. Let $\theta = T(F)$ and $\hat{\theta}_n = T(\hat{F}_n)$ and define the **pivot** $R_n = \hat{\theta}_n - \theta$. Let $\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*$ denote bootstrap replications of $\hat{\theta}_n$. Let $H(r)$ denote the CDF of the pivot:

$$H(r) = \mathbb{P}_F(R_n \leq r). \quad (8.3)$$

Define $C_n^* = (a, b)$ where

$$a = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{and} \quad b = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right). \quad (8.4)$$

It follows that

$$\begin{aligned} \mathbb{P}(a \leq \theta \leq b) &= \mathbb{P}(a - \hat{\theta}_n \leq \theta - \hat{\theta}_n \leq b - \hat{\theta}_n) \\ &= \mathbb{P}(\hat{\theta}_n - b \leq \hat{\theta}_n - \theta \leq \hat{\theta}_n - a) \\ &= \mathbb{P}(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a) \\ &= H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\ &= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

Hence, C_n^* is an exact $1 - \alpha$ confidence interval for θ . Unfortunately, a and b depend on the unknown distribution H but we can form a bootstrap estimate of H :

$$\hat{H}(r) = \frac{1}{B} \sum_{b=1}^B I(R_{n,b}^* \leq r) \quad (8.5)$$

where $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Let r_β^* denote the β sample quantile of $(R_{n,1}^*, \dots, R_{n,B}^*)$ and let θ_β^* denote the β sample quantile of $(\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*)$. Note that $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$. It follows that an approximate $1 - \alpha$ confidence interval is $C_n = (\hat{a}, \hat{b})$ where

$$\begin{aligned} \hat{a} &= \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^* \\ \hat{b} &= \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^*. \end{aligned}$$

In summary, the $1 - \alpha$ **bootstrap pivotal confidence interval** is

$$C_n = \left(2\hat{\theta}_n - \hat{\theta}_{1-\alpha/2}^*, 2\hat{\theta}_n - \hat{\theta}_{\alpha/2}^*\right). \quad (8.6)$$

8.3 Theorem. *Under weak conditions on $T(F)$,*

$$\mathbb{P}_F(T(F) \in C_n) \rightarrow 1 - \alpha$$

as $n \rightarrow \infty$, where C_n is given in (8.6).

Method 3: Percentile Intervals. The **bootstrap percentile interval** is defined by

$$C_n = \left(\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*\right).$$

The justification for this interval is given in the appendix.

8.4 Example. For estimating the skewness of the nerve data, here are the various confidence intervals.

Method	95% Interval
Normal	(1.44, 2.09)
Pivotal	(1.48, 2.11)
Percentile	(1.42, 2.03)

All these confidence intervals are approximate. The probability that $T(F)$ is in the interval is not exactly $1 - \alpha$. All three intervals have the same level of accuracy. There are more accurate bootstrap confidence intervals but they are more complicated and we will not discuss them here.

8.5 Example (The Plasma Cholesterol Data). Let us return to the cholesterol data. Suppose we are interested in the difference of the medians. Pseudocode for the bootstrap analysis is as follows:

```
x1 <- first sample
x2 <- second sample
n1 <- length(x1)
n2 <- length(x2)
th.hat <- median(x2) - median(x1)
B <- 1000
Tboot <- vector of length B
for(i in 1:B){
  xx1 <- sample of size n1 with replacement from x1
  xx2 <- sample of size n2 with replacement from x2
  Tboot[i] <- median(xx2) - median(xx1)
}
se <- sqrt(variance(Tboot))
Normal      <- (th.hat - 2*se, th.hat + 2*se)
percentile <- (quantile(Tboot,.025), quantile(Tboot,.975))
pivotal     <- ( 2*th.hat-quantile(Tboot,.975),
               2*th.hat-quantile(Tboot,.025) )
```

The point estimate is 18.5, the bootstrap standard error is 7.42 and the resulting approximate 95 percent confidence intervals are as follows:

Method	95% Interval
Normal	(3.7, 33.3)
Pivotal	(5.0, 34.0)
Percentile	(5.0, 33.3)

Since these intervals exclude 0, it appears that the second group has higher cholesterol although there is considerable uncertainty about how much higher as reflected in the width of the intervals. ■

The next two examples are based on small sample sizes. In practice, statistical methods based on very small sample sizes might not be reliable. We include the examples for their pedagogical value but we do want to sound a note of caution about interpreting the results with some skepticism.

8.6 Example. Here is an example that was one of the first used to illustrate the bootstrap by Bradley Efron, the inventor of the bootstrap. The data are LSAT scores (for entrance to law school) and GPA.

LSAT	576	635	558	578	666	580	555	661
	651	605	653	575	545	572	594	
GPA	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43
	3.36	3.13	3.12	2.74	2.76	2.88	3.96	

Each data point is of the form $X_i = (Y_i, Z_i)$ where $Y_i = \text{LSAT}_i$ and $Z_i = \text{GPA}_i$. The law school is interested in the correlation

$$\theta = \frac{\int \int (y - \mu_Y)(z - \mu_Z) dF(y, z)}{\sqrt{\int (y - \mu_Y)^2 dF(y) \int (z - \mu_Z)^2 dF(z)}}.$$

The plug-in estimate is the sample correlation

$$\hat{\theta} = \frac{\sum_i (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (Z_i - \bar{Z})^2}}.$$

The estimated correlation is $\hat{\theta} = .776$. The bootstrap based on $B = 1000$ gives $\hat{\mathbf{s}}\hat{\mathbf{e}} = .137$. Figure 8.1 shows the data and a histogram of the bootstrap replications $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. This histogram is an approximation to the sampling distribution of $\hat{\theta}$. The Normal-based 95 percent confidence interval is $.78 \pm 2\hat{\mathbf{s}}\hat{\mathbf{e}} = (.51, 1.00)$ while the percentile interval is $(.46, .96)$. In large samples, the two methods will show closer agreement. ■

8.7 Example. This example is from Efron and Tibshirani (1993). When drug companies introduce new medications, they are sometimes required to show bioequivalence. This means that the new drug is not substantially different than the current treatment. Here are data on eight subjects who used medical patches to infuse a hormone into the blood. Each subject received three treatments: placebo, old-patch, new-patch.

subject	placebo	old	new	old - placebo	new - old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

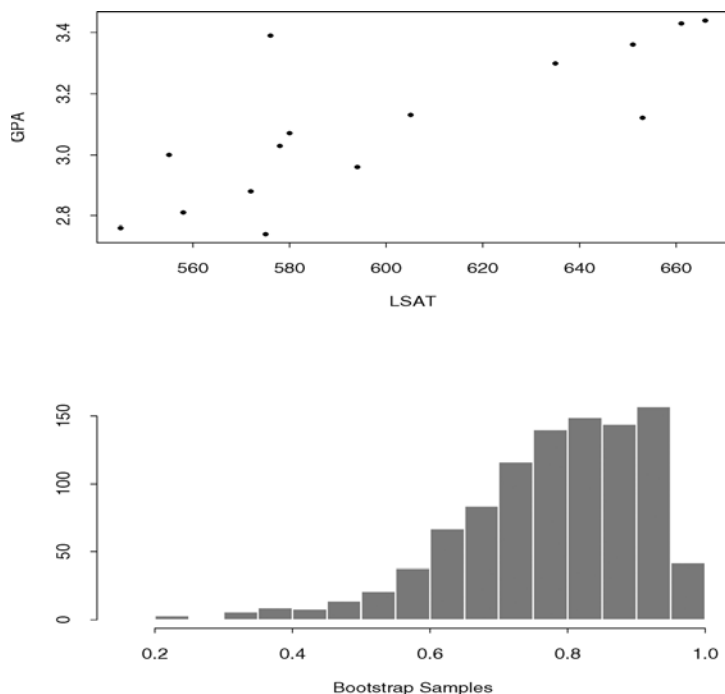


FIGURE 8.1. Law school data. The top panel shows the raw data. The bottom panel is a histogram of the correlations computed from each bootstrap sample.

Let $Z = \text{old} - \text{placebo}$ and $Y = \text{new} - \text{old}$. The Food and Drug Administration (FDA) requirement for bioequivalence is that $|\theta| \leq .20$ where

$$\theta = \frac{\mathbb{E}_F(Y)}{\mathbb{E}_F(Z)}.$$

The plug-in estimate of θ is

$$\hat{\theta} = \frac{\bar{Y}}{\bar{Z}} = \frac{-452.3}{6342} = -0.0713.$$

The bootstrap standard error is $\hat{\text{se}} = 0.105$. To answer the bioequivalence question, we compute a confidence interval. From $B = 1000$ bootstrap replications we get the 95 percent interval $(-0.24, 0.15)$. This is not quite contained

in $(-0.20, 0.20)$ so at the 95 percent level we have not demonstrated bioequivalence. Figure 8.2 shows the histogram of the bootstrap values. ■

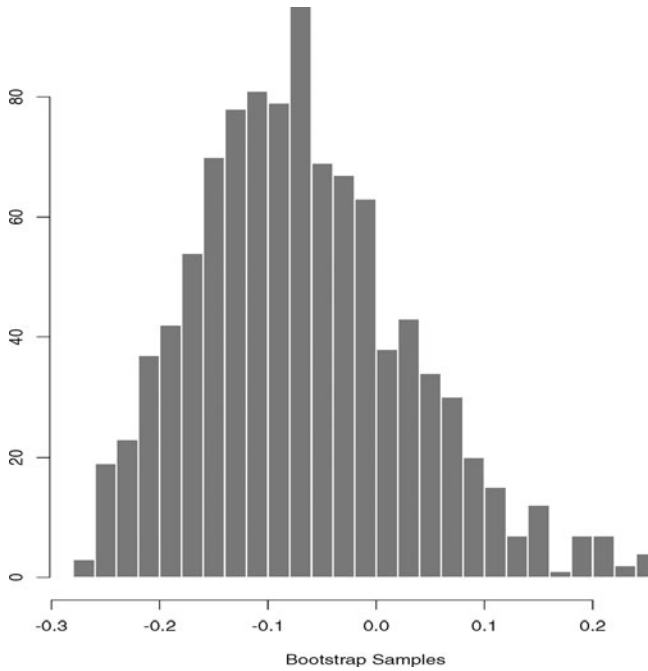


FIGURE 8.2. Patch data.

8.4 Bibliographic Remarks

The bootstrap was invented by Efron (1979). There are several books on these topics including Efron and Tibshirani (1993), Davison and Hinkley (1997), Hall (1992) and Shao and Tu (1995). Also, see section 3.6 of van der Vaart and Wellner (1996).

8.5 Appendix

8.5.1 *The Jackknife*

There is another method for computing standard errors called the **jackknife**, due to Quenouille (1949). It is less computationally expensive than the boot-

strap but is less general. Let $T_n = T(X_1, \dots, X_n)$ be a statistic and $T_{(-i)}$ denote the statistic with the i^{th} observation removed. Let $\bar{T}_n = n^{-1} \sum_{i=1}^n T_{(-i)}$. The jackknife estimate of $\text{var}(T_n)$ is

$$v_{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^n (T_{(-i)} - \bar{T}_n)^2$$

and the jackknife estimate of the standard error is $\hat{\text{se}}_{\text{jack}} = \sqrt{v_{\text{jack}}}$. Under suitable conditions on T , it can be shown that v_{jack} consistently estimates $\text{var}(T_n)$ in the sense that $v_{\text{jack}}/\text{var}(T_n) \xrightarrow{P} 1$. However, unlike the bootstrap, the jackknife does not produce consistent estimates of the standard error of sample quantiles.

8.5.2 Justification For The Percentile Interval

Suppose there exists a monotone transformation $U = m(T)$ such that $U \sim N(\phi, c^2)$ where $\phi = m(\theta)$. We do not suppose we know the transformation, only that one exists. Let $U_b^* = m(\theta_{n,b}^*)$. Let u_β^* be the β sample quantile of the U_b^* 's. Since a monotone transformation preserves quantiles, we have that $u_{\alpha/2}^* = m(\theta_{\alpha/2}^*)$. Also, since $U \sim N(\phi, c^2)$, the $\alpha/2$ quantile of U is $\phi - z_{\alpha/2}c$. Hence $u_{\alpha/2}^* = \phi - z_{\alpha/2}c$. Similarly, $u_{1-\alpha/2}^* = \phi + z_{\alpha/2}c$. Therefore,

$$\begin{aligned} \mathbb{P}(\theta_{\alpha/2}^* \leq \theta \leq \theta_{1-\alpha/2}^*) &= \mathbb{P}(m(\theta_{\alpha/2}^*) \leq m(\theta) \leq m(\theta_{1-\alpha/2}^*)) \\ &= \mathbb{P}(u_{\alpha/2}^* \leq \phi \leq u_{1-\alpha/2}^*) \\ &= \mathbb{P}(U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}) \\ &= \mathbb{P}(-z_{\alpha/2} \leq \frac{U - \phi}{c} \leq z_{\alpha/2}) \\ &= 1 - \alpha. \end{aligned}$$

An exact normalizing transformation will rarely exist but there may exist approximate normalizing transformations.

8.6 Exercises

1. Consider the data in Example 8.6. Find the plug-in estimate of the correlation coefficient. Estimate the standard error using the bootstrap. Find a 95 percent confidence interval using the Normal, pivotal, and percentile methods.

2. (Computer Experiment.) Conduct a simulation to compare the various bootstrap confidence interval methods. Let $n = 50$ and let $T(F) = \int (x - \mu)^3 dF(x) / \sigma^3$ be the skewness. Draw $Y_1, \dots, Y_n \sim N(0, 1)$ and set $X_i = e^{Y_i}$, $i = 1, \dots, n$. Construct the three types of bootstrap 95 percent intervals for $T(F)$ from the data X_1, \dots, X_n . Repeat this whole thing many times and estimate the true coverage of the three intervals.

3. Let

$$X_1, \dots, X_n \sim t_3$$

where $n = 25$. Let $\theta = T(F) = (q_{.75} - q_{.25})/1.34$ where q_p denotes the p^{th} quantile. Do a simulation to compare the coverage and length of the following confidence intervals for θ : (i) Normal interval with standard error from the bootstrap, (ii) bootstrap percentile interval, and (iii) pivotal bootstrap interval.

4. Let X_1, \dots, X_n be distinct observations (no ties). Show that there are

$$\binom{2n-1}{n}$$

distinct bootstrap samples.

Hint: Imagine putting n balls into n buckets.

5. Let X_1, \dots, X_n be distinct observations (no ties). Let X_1^*, \dots, X_n^* denote a bootstrap sample and let $\bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$. Find: $\mathbb{E}(\bar{X}_n^* | X_1, \dots, X_n)$, $\mathbb{V}(\bar{X}_n^* | X_1, \dots, X_n)$, $\mathbb{E}(\bar{X}_n^*)$ and $\mathbb{V}(\bar{X}_n^*)$.
6. (Computer Experiment.) Let X_1, \dots, X_n Normal($\mu, 1$). Let $\theta = e^\mu$ and let $\hat{\theta} = e^{\bar{X}}$. Create a data set (using $\mu = 5$) consisting of $n=100$ observations.
- (a) Use the bootstrap to get the se and 95 percent confidence interval for θ .
- (b) Plot a histogram of the bootstrap replications. This is an estimate of the distribution of $\hat{\theta}$. Compare this to the true sampling distribution of $\hat{\theta}$.
7. Let $X_1, \dots, X_n \sim \text{Uniform}(0, \theta)$. Let $\hat{\theta} = X_{\max} = \max\{X_1, \dots, X_n\}$. Generate a data set of size 50 with $\theta = 1$.
- (a) Find the distribution of $\hat{\theta}$. Compare the true distribution of $\hat{\theta}$ to the histograms from the bootstrap.

(b) This is a case where the bootstrap does very poorly. In fact, we can prove that this is the case. Show that $P(\hat{\theta} = \hat{\theta}) = 0$ and yet $P(\hat{\theta}^* = \hat{\theta}) \approx .632$. Hint: show that, $P(\hat{\theta}^* = \hat{\theta}) = 1 - (1 - (1/n))^n$ then take the limit as n gets large.

8. Let $T_n = \overline{X}_n^2$, $\mu = \mathbb{E}(X_1)$, $\alpha_k = \int |x - \mu|^k dF(x)$ and $\hat{\alpha}_k = n^{-1} \sum_{i=1}^n |X_i - \overline{X}_n|^k$. Show that

$$v_{\text{boot}} = \frac{4\overline{X}_n^2 \hat{\alpha}_2}{n} + \frac{4\overline{X}_n \hat{\alpha}_3}{n^2} + \frac{\hat{\alpha}_4}{n^3}.$$