DATA 1010
Problem Set 12
Due 17 December 2018 at 12 PM

Problem 1

Label each of the following four estimators as either (i) biased and consistent, (ii) biased and inconsistent, (iii) unbiased and consistent, or (iv) unbiased and inconsistent. The matching will be one-to-one.

(a) $X_1, X_2, \ldots$ are i.i.d. Bernoulli random variables with unknown p and estimator

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(b) $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2 and estimator

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}.$$

(c) $X_1, X_2, \ldots$ are i.i.d. uniform random variables on an unknown bounded interval. For $n \geq 100$ we estimate the mean using

$$\widehat{\mu} = \frac{\sum_{i=1}^{100} X_i}{100}.$$

(d) $X_1, X_2, \ldots$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2. For $n \geq 100$ we estimate the standard deviation using

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{100}(X_i - \overline{X})^2}{99}}.$$

Solution

(a) **Unbiased and consistent**. The expectation of $\widehat{p}$ is $(1/n)(np) = p$, and the variance converges to 0 since $\widehat{p}$ is an average of i.i.d., finite-variance random variables. Therefore, the mean squared error converges to 0 as $n \to \infty$

(b) **Biased and consistent**. The estimator is biased because its value is always slightly smaller than the unbiased estimator (which has $n-1$ instead of n in the denominator). The estimator is nevertheless consistent, since the bias and the variance both converge to 0 as $n \to \infty$.

(c) **Unbiased and inconsistent**. The mean of $\widehat{\mu}$ is $(1/100)(100\mu) = \mu$, so the estimator is unbiased. The variance isn't zero and doesn't depend on n, so it cannot converge to 0 as $n \to \infty$. Therefore, the estimator is inconsistent.

(d) **Biased and inconsistent**. This estimator is inconsistent for the same reason as (c). The bias is trickier. Since the variance of $\widehat{\sigma}$ is positive, then we have $\mathbb{E}[\hat{\sigma}^2] - \mathbb{E}[\hat{\sigma}]^2 > 0$, which implies that

$$\mathbb{E}[\hat{\sigma}]^2 < \mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{1}{99}\sum_{i=1}^{100}(X_i - \overline{X})^2\right] = \sigma^2.$$

Thus the bias of $\widehat{\sigma}$ is negative.

Problem 2

Suppose that $X_1, \ldots, X_n$ are independent Unif$[0, \theta]$ random variables, where θ is an unknown parameter, and consider

the following estimators for θ:

$$\widehat{\theta}_1 = \max(X_1, \ldots, X_n), \qquad \widehat{\theta}_2 = 2 \cdot \frac{X_1 + \cdots + X_n}{n}.$$

(a) Find the CDF of $\widehat{\theta}_1$.

(b) Recall that if $F_{\widehat{\theta}_1}(x)$ and $f_{\widehat{\theta}_1}(x)$ are the CDF and PDF of $\widehat{\theta}_1$ respectively, then $\frac{d}{dx} F_{\widehat{\theta}_1}(x) = f_{\widehat{\theta}_1}(x)$.

Differentiate your answer to (a) to find the PDF of $\widehat{\theta}_1$.

(c) Show that $\widehat{\theta}_1$ is consistent.

(d) Find $\mathbb{E}\left[\widehat{\theta}_1\right]$ and $\mathbb{E}\left[\widehat{\theta}_2\right]$. Which estimator is biased?

(e) Find $\text{Var}\left(\widehat{\theta}_1\right)$ and $\text{Var}\left(\widehat{\theta}_2\right)$. Which estimator has lower variance?

(f) Show that the mean squared error of $\widehat{\theta}_1$ is less than the mean squared error of $\widehat{\theta}_2$ whenever $n \geq 3$.

Solution

(a) The probability that $\widehat{\theta}_1$ exceeds $t \in [0, \theta]$ is the probability that all of the X_i's are less than or equal to t. By independence, this probability is $(t/\theta)^n$. Therefore,

$$F_{\widehat{\theta}_1}(t) = \begin{cases} 0 & t \leq 0 \\ (t/\theta)^n & 0 \leq t \leq \theta \\ 1 & \theta \leq t \end{cases}$$

(b) Differentiating $F_{\widehat{\theta}_1}(t)$ gives nt^{n-1}/θ^n.

(c) We did this one in class: the probability that $\widehat{\theta}_1$ is less than $\theta - \epsilon$ is

$$\left(\frac{\theta - \epsilon}{\theta}\right)^n,$$

which converges to 0 as $n \to \infty$.

(d) We have

$$\mathbb{E}[\widehat{\theta}_1] = \int_0^\theta t(nt^{n-1}/\theta^n)\, \mathrm{d}\theta = \frac{n}{n+1}\theta,$$

and

$$\mathbb{E}[\widehat{\theta}_2] = 2\mathbb{E}[X_1 + \cdots + X_n]/n = 2(n\theta/2)/n = \theta.$$

So $\widehat{\theta}_1$ is biased and $\widehat{\theta}_2$ is unbiased.

(e) We have

$$\mathbb{E}[\widehat{\theta}_1] = \int_0^\theta t^2(nt^{n-1}/\theta^n)\, \mathrm{d}\theta = \frac{n}{n+2}\theta^2,$$

so the variance of $\widehat{\theta}_1$ is

$$\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

The variance of $\widehat{\theta}_2$ is

$$\text{Var}(\widehat{\theta}_2) = \frac{4}{n^2}\,\text{Var}(X_1 + X_2 + \cdots + X_n) = \frac{4\sigma^2}{n} = \frac{\theta^2}{3n}.$$

(f) The mean squared error of $\widehat{\theta}_2$ is its variance $\frac{\theta^2}{3n}$, since it is unbiased. The mean squared error of $\widehat{\theta}_1$ is

$$\frac{n\theta^2}{(n+1)^2(n+2)} + \left(\frac{\theta}{n+1}\right)^2 = \frac{2\theta^2}{(n+1)(n+2)}.$$

These expressions are equal when $n = 1$ and when $n = 2$, but the former is larger for all $n \geq 3$.

(Some SymPy code for checking the calculations above:)

Julia

```julia
using SymPy
@vars θ t
@vars n integer=true positive=true
F = t^n/θ^n
f = diff(F,t)
μ = simplify(integrate(t*f,(t,0,θ)))
simplify(integrate(t^2*f,(t,0,θ)))
σ² = factor(integrate(t^2*f,(t,0,θ)) - μ^2)
(μ-θ)^2 + σ² |> simplify |> factor # returns 2θ²/((n+1)(n+2))
```

Problem 3

(a) **Hoeffding's inequality** says that if $Y_1, Y_2, \ldots$ are independent random variables with the property that $\mathbb{E}[Y_i] = 0$ and $a_i \le Y_i \le b_i$ for all i, then for all $\epsilon > 0$ and $t > 0$, we have

$$\mathbb{P}\left(Y_1 + Y_2 + \cdots + Y_n \ge \epsilon\right) \le \mathrm{e}^{-t\epsilon} \prod_{i=1}^{n} \mathrm{e}^{t^2(b_i-a_i)^2/8}.$$

Use Hoeffding's inequality to show that if $X_1, X_2, X_3, \ldots$ is a sequence of independent Bernoulli(p) random variables, then for all $\alpha > 0$, the interval $\left(\overline{X}_n - \sqrt{\frac{1}{2n}\log(2/\alpha)}, \overline{X}_n + \sqrt{\frac{1}{2n}\log(2/\alpha)}\right)$ is a confidence interval for p with confidence level $1-\alpha$. Explain what happens to the width of this confidence interval if n gets large, and also what happens to the width if α is made very small.

(b) As above, consider n independent Bernoulli(p)'s. Find the normal-approximation confidence interval for p

(c) As above, consider n independent Bernoulli(p)'s. Find the Chebyshev confidence interval for p.

(d) Find the numerical values of the half-widths for each of the above confidence intervals when $p = \frac{1}{2}$, $n = 1000$, and $\alpha = 0.05$ (approximating $\overline{X}$ as p).

Solution

(a) Let's define $Y_i = (X_i - p)/n$. Then $\mathbb{E}[Y_i] = 0$, and the the tightest interval $[a_i, b_i]$ that contains the range of Y_i is $[-p/n, (1-p)/n]$. So Hoeffding's inequality says that

$$\mathbb{P}(\overline{X}_n - p \ge \epsilon) \le \mathrm{e}^{-t\epsilon + nt^2/(8n^2)}.$$

Since this inequality holds for all t, we achieve the best upper bound by choosing the value of t which minimizes the exponent on the right-hand side. Since the graph of that expression is a convex parabola, we can find the minimum of the expression by differentiating and finding the unique critical point. We find that the minimizing value of t is $4\epsilon/n$, which means that

$$\mathbb{P}(\overline{X}_n - p \ge \epsilon) \le \mathrm{e}^{-2n\epsilon^2}.$$

Substituting $\epsilon_n = \sqrt{\frac{1}{2n}\log(2/\alpha)}$, we get $\mathbb{P}(\overline{X}_n - p \ge \epsilon_n) \le \alpha/2$. Likewise, we can repeat all of the above for $Y_i = -(X_i - p)/n$ and find that $\mathbb{P}(\overline{X}_n - p \le -\epsilon_n) \le \alpha/2$. So the probability that $|\overline{X}_n - p| \ge \epsilon_n$ is no more than $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$ (by the subadditivity property of probability measures).

As $n \to \infty$, the confidence interval shrinks, and if α is very small, then the confidence interval grows. Both of these are consistent with what you would expect: more data permits a tighter confidence interval, and a higher confidence level requires a wider confidence interval.

(b) The normal-approximation confidence interval is $(\overline{X}_n - z_{\alpha/2}\sqrt{\overline{X}_n(1-\overline{X}_n)/n}, \overline{X}_n + z_{\alpha/2}\sqrt{\overline{X}_n(1-\overline{X}_n)/n})$, where $z_{\alpha/2}$ is the value such that the standard normal distribution assigns mass $1-\alpha$ to $[-z_{\alpha/2}, z_{\alpha/2}]$.

(c) The Chebyshev confidence interval is $(\overline{X}_n - \frac{1}{\sqrt{\alpha}}\sqrt{\overline{X}_n(1-\overline{X}_n)/n}, \overline{X}_n + \frac{1}{\sqrt{\alpha}}\sqrt{\overline{X}_n(1-\overline{X}_n)/n})$, where the expression $\frac{1}{\sqrt{\alpha}}$ is obtained by solving the equation $1/k^2 = \alpha$ for k.

(d) We approximate $\overline{X}_n \approx p$ to find the values

$$\sqrt{\log(2/0.02)/(2\cdot 1000)} \approx 0.048 \quad 1.96\sqrt{(1/2)(1-1/2)/1000} \approx 0.031 \quad \frac{1}{\sqrt{0.05}}\sqrt{(1/2)(1-1/2)/1000} \approx 0.071$$

So we can see that the normal approximation provides the tightest confidence interval, while Hoeffding does better than Chebyshev.

Problem 4

I drew 6 samples from an undisclosed distribution and obtained the following results:

```
c(6.19,7.048,6.143,5.459,4.603,4.335)
```

I also drew 8 samples from another undisclosed distribution and got

```
c(8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309)
```

Determine whether the Wald hypothesis test (with significance $\alpha = 0.05$) rejects the null hypothesis that the mean of the two distributions are equal.

Solution

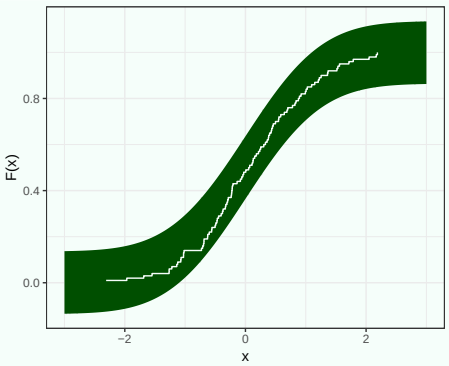
If we store the two given vectors as `u` and `v`, we estimate the standard error of the difference of sample means as

```
se <- sqrt(var(u)/length(u) + var(v)/length(v))
```

which returns 0.98. Thus the observed difference between sample means is `(mean(u) - mean(v))/se` = 1.03 standard deviations from the mean. Since $1.03 < 1.96$, we retain the null hypothesis.

Problem 5

One hundred samples were drawn from the standard normal distribution, and the resulting empirical CDF was plotted in the figure shown. Also plotted are the DKW bounds for $\alpha = 0.05$.



(a) Replicate this figure in `ggplot`. Some tips: do `set.seed(123)` and draw samples with `rnorm`. Evaluate the CDF of the normal distribution with `pnorm`. You can make the CDF graph using the step geom. You'll want to make some tibbles to contain the generated data, and bear in mind that you can specify the data frame on a geom-by-geom basis in ggplot.

(b) Run the code block repeatedly (though comment out the `set.seed` line first!) until you see the empirical CDF graph fall outside the ribbon. How many times did it take? Does your result seem consistent with the claim that the graph of the empirical CDF lies in the ribbon with probability at least $1 - \alpha$?

Solution

(a) We use the ribbon geom for the green band, and the step geom as suggested. We calculate the band half-width in terms of the given α according to the DKW inequality.

```
library(tidyverse)
set.seed(123)
n <- 100
U <- rnorm(n)
xs <- seq(-3,3,length=250)
alpha <- 0.05
eps <- sqrt((1/(2*n))*log(2/alpha))
ggplot() +
  geom_ribbon(aes(x = xs,
                  ymin = pnorm(xs) - eps,
                  ymax = pnorm(xs) + eps),
              fill='darkgreen') +
  geom_step(aes(x = sort(U), y = (1:n)/n), direction='vh')
```

(b) It took me 13 times to see the stairstep curve fall outside the band. This does seem consistent with a 5% confidence interval, since an event whose probability is 95% fails one time in 20 on average.

Problem 6

Consider a distribution ν which is known only via a dozen samples therefrom, the values of which are

```
c(8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309)
```

(a) Obtain a bootstrap estimate of the standard error of the plug-in estimator for the median of five samples from ν.

(b) The actual standard deviation of the median of 5 samples from ν is approximately 2.14. How close is the value you found? Could you have gotten as close as desired to this value by choosing sufficiently many bootstrap re-samplings?

Solution

(a) We calculate

```
var(sapply(1:10^5, function(n) {median(sample(u,5,replace=TRUE))}))
```

which returns (approximately) 2.07.

(b) The value 2.07 is already very close to $T(\widehat{\nu})$ (where T is the "variance of the median of five independent samples from" statistical functional, and ν is the empirical measure), and it could be made even closer by using more than 10^5 bootstrap runs. But it is not necessarily particularly close to $T(\nu) = 2.14$, and it could be made closer only by taking more samples from ν.

Problem 7

Consider the family of densities

$$\left\{ \frac{1}{2(\beta-\alpha)} \mathbf{1}_{\{\alpha \le x \le \beta\}} + \frac{1}{2(\delta-\gamma)} \mathbf{1}_{\{\gamma \le x \le \delta\}} \,:\, \alpha < \beta < \gamma < \delta \right\}.$$

Show that there is no maximum likelihood estimator for this family of distributions. In other words, consider a set of samples drawn from one of these distributions, and show that arbitrarily large likelihoods may be obtained for those samples by choosing suitable values for the parameters.

Solution

Consider a set of samples $x_1, \ldots, x_n$. We define $[\alpha, \beta]$ to be some interval which includes all of the points, and we define $[\delta, \gamma]$ to be a very small interval which is centered at one of the samples. Then the likelihood is positive, and we can make it arbitrarily large from here by shrinking $[\delta, \gamma]$ down around the point at its center. Therefore, there is no choice of parameters which maximizes the likelihood.