Homework 05

Brown University

DATA 1010

Fall 2019

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

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

(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 $\hat{p}$ is $(1/n)(np) = p$, and the variance converges to 0 since $\hat{p}$ 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 $\mathrm{Unif}[0, \theta]$ random variables, where θ is an unknown parameter, and consider the following estimators for θ:

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

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

(b) Recall that if $F_{\hat{\theta}_1}(x)$ and $f_{\hat{\theta}_1}(x)$ are the CDF and PDF of $\hat{\theta}_1$ respectively, then $\frac{d}{dx} F_{\hat{\theta}_1}(x) = f_{\hat{\theta}_1}(x)$. Differentiate your answer to (a) to find the PDF of $\hat{\theta}_1$.

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

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

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

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

Hint: this problem is pretty calculus intensive. SymPy is your friend.

solution

(a) The probability that $\hat{\theta}_1$ 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 $F_{\hat{\theta}_1}(t)$ gives nt^{n-1}/θ^n.

(c) The probability that $\hat{\theta}_1$ 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) \,
```

```
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 $\hat{\theta}_1$ is biased and $\hat{\theta}_2$ is unbiased.

(e) We have

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

so the variance of $\hat{\theta}_1$ 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 $\hat{\theta}_2$ is

```
$$\operatorname{Var}(\widehat{\theta}_2) =
  \frac{4}{n^2}\operatorname{Var}(X_{1}+X_{2}+\cdots+X_{n}) =
  \frac{4\sigma^2}{n} = \frac{\theta^2}{3n}.$$
```

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

```
[ ]
```

```
# solution

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 $a_i \leq Y_i \leq b_i$ for all i, then for all $\epsilon > 0$ and $t > 0$, we have

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

Use Hoeffding's inequality to show that if $X_1, X_2, X_3, \ldots$ is a sequence of independent $\text{Bernoulli}(p)$ 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 $\text{Bernoulli}(p)$'s. Find the normal-approximation confidence interval for p

(c) As above, consider n independent $\text{Bernoulli}(p)$'s. Find the Chebyshev confidence interval for p. (Chebyshev's inequality says that the probability of any random variable deviating from its mean by more than k standard deviations is no more than $1/k^2$.)

(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

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

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

$$\mathbb{P}(\overline{X}_n - p \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

Substituting $\epsilon_n = \sqrt{\frac{1}{2n}\log(2/\alpha)}$, we get $\mathbb{P}(\overline{X}_n - p \geq \epsilon_n) \leq \alpha/2$. Likewise, we can repeat all of the above for $Y_i = -(X_i - p)/n$ and find that $\mathbb{P}(\overline{X}_n - p \leq -\epsilon_n) \leq \alpha/2$. So the probability that $|\overline{X}_n - p| \geq \epsilon_n$ 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.

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

```
$$\sqrt{\log(2/0.02)/(2\cdot1000)} \approx 0.048
  \quad
  1.96\sqrt{(1/2)(1-1/2)/1000} \approx 0.031
  \quad
  \frac{1}{\sqrt{0.05}}\sqrt{(1/2)(1-1/2)/1000} \approx 0.071$$
```

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

Problem 4

I drew 6 observations from an undisclosed distribution and obtained the following results: `u = [6.19,7.048,6.143,5.459,4.603,4.335]`

I also drew 8 observations from another undisclosed distribution and got `v = [8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309]`

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

(b) Repeat with Welch's t-test in place of the the Wald test.

solution

We estimate the standard error of the difference of sample means as

```
[ ]  # solution
     using Statistics
     se = sqrt(std(u)^2/length(u) + std(v)^2/length(v))
```

solution

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

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

```
[8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309]
```

(a) Obtain a bootstrap estimate of the standard deviation of the median of five independent 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

```
[ ]  # solution
     using Random, Statistics, StatsBase
     X = [8.924,4.698,6.095,4.223,3.643,1.624,1.444,6.309]
     std(median(sample(X,5)) for _ in 1:10^6)
```

solution

No, we could not get arbitrarily close to the correct value, because bootstrapping only allows us to approximate the plug-in estimator arbitrarily well. To get a better estimate of the actual value of the statistical functional, we would need more observations from the original distribution.

Problem 6

Consider a population of patients who have had a recent heart attack. A randomized trial is conducted in which each patient is assigned with equal probability (and independently of any attribute of the patient) to either a heart medication regimen or a placebo. Each patient has a unknown genetic attribute X which is uniformly distributed on $[0, 1]$.

Suppose that the probability that a patient will comply with the regimen (that is, take the medication as prescribed) is $(1 + X)/2$. The conditional probability that they will survive the following decade, given X and given that they take the drug, is $3X/4$. The conditional probability that they will survive the following decade is, given X and given that they do not take the drug, is $(X + 1)/4$.

(a) Fill out the following tables, indicating the probability of each outcome. The eight numbers should sum to 1.

Drug condition	survives	does not survive
complaint	_	_
non-compliant	_	_

Placebo condition	survives	does not survive
complaint	_	_
non-compliant	_	_

Hint: you want to work out the conditional probabilities of each event given X, and then find the expected value of the resulting conditional probability by integrating against the density of X.

(b) Does a randomly selected patient have a higher conditional probability of surviving if they take the drug or if they do not? Does comparing the survival probabilities for the drug and placebo conditions give the correct answer to this question?

(c) Suppose you know yourself well enough to be confident that you'd be relatively unlikely to comply with a prescription regimen if you had participated in this clinical trial. Based on the given probability model, should you take the drug? Would you get the right answer or the wrong answer if you just compared the survival rates for compliant and noncompliant patients?

solution

Solution.
(a) We calculate the indicated values using SymPy:

```
[ ]  # solution
     using SymPy
     @vars x
     survival_given_compliance = 3x/4
     survival_given_noncompliance = (x+1)/4
     compliance_prob = (1+x)/2

     drug_table =
     [
       survival_given_compliance * compliance_prob        (1-
     survival_given_compliance) * compliance_prob
       survival_given_noncompliance * (1-compliance_prob) (1-
     survival_given_noncompliance) * (1-compliance_prob)
     ]

     integrate.(drug_table,Ref((x,0,1)))/2 # Ref protects the tuple
     from the dot, so it doesn't try to broadcast
```

```
[ ]  # solution
     placebo_table =
     [
       survival_given_noncompliance * compliance_prob        (1-
     survival_given_noncompliance) * compliance_prob
       survival_given_noncompliance * (1-compliance_prob) (1-
     survival_given_noncompliance) * (1-compliance_prob)
     ]

     integrate.(placebo_table,Ref((x,0,1)))/2
```

solution

(b) Yes, a randomly selected patient has a higher probability of surviving with the drug than without it. Comparing rates does give the correct answer.

(c) No, you should not take the drug. If you know you're not inclined to be compliant, that's an indication that your value of X is small. And the conditional probability of survival is larger without the drug when X is small:

```
[ ]  # solution
     using Plots, LaTeXStrings
     plot(0:0.01:1, [x->3x/4, x->(x+1)/4], label = ["drug", "no
     drug"],
          xlabel=L"X", ylabel="survival probability", leg = :topleft,
     ylims = (0,1))
```

Problem 7

(a) Write a function which a distribution `D`, together with an α value and a positive integer n and returns `true` or `false` according to whether the empirical CDF for a random sample of size n obeys the bound in the DKW inequality.

(b) Run the function many times with `α = 0.05` and check that it returns `true` around 95% of the time.

[ ]
```
function DKW_check(D, α, n)
    sample = sort(rand(D,n)) # solution
    for i in eachindex(sample) # solution
        ∈ = √(log(2/α)/(2n)) # solution
        if max(abs(cdf(D, sample[i]) - i/n), abs(cdf(D,
sample[i]) - (i-1)/n)) > ∈ # solution
            return false # solution
        end # solution
    end # solution
    true # solution
end
```

[ ]
```
D = Uniform(0,1)
α = 0.05
n = 1000
mean(DKW_check(D, α, n) for _ in 1:10000)
```

Problem 8

Consider a family of distributions of the form $\mathrm{Uniform}(\theta, \theta + 1)$, where θ is a parameter. Given a sample $X_1, \ldots, X_n$, show that there isn't a unique maximum likelihood estimator for θ

solution

The likelihood is constant for all θ values for which $(\theta, \theta + 1)$ traps all of the observations. For other values of θ, the likelihood is zero.

Since the sample range is necessarily strictly less than 1, there will be an interval of θ values (from the sample maximum minus 1 up to the sample minimum) which have the same likelihood.

```
[ ]
```