STAT 510: Homework 03

David Dalpiaz

Due: Monday, February 14, 11:59 PM

Exercise 1 (Expectation Review)

Let X_1, X_2, and X_3 be independent Uniform$(0, 1)$ random variables. Define $Y = X_1 - 3X_2 + 2X_3$. Provide an upper bound for $P(|Y| \geq 2)$ using Chebyshev's inequality.

Solution

We first note that

$$\mu = \mathbb{E}[Y] = \mathbb{E}[X_1] - 3\mathbb{E}[X_2] + 2\mathbb{E}[X_3] = 0.5 - 3 \cdot 0.5 + 2 \cdot 0.5 = 0.$$

We also note that

$$\mathbb{V}[Y] = \mathbb{V}[X_1] + 9\mathbb{V}[X_2] + 4\mathbb{V}[X_3] = \frac{1}{12} + \frac{9}{12} + \frac{4}{12} = \frac{7}{6}.$$

Then, using Chebyshev's Inequality we have

$$P(|Y - \mu| \geq 2) = P(|Y| \geq 2) \leq \frac{\mathbb{V}[Y]}{4} = \boxed{\frac{7}{24}}.$$

Exercise 2 (Creating a Confidence Interval)

(Based on **LW** 4.4) Let $X_1, X_2, \ldots X_n \sim$ Bernoulli(p). Let $\alpha > 0$ and define

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

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

$$C_n = (\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n).$$

Show that

$$P(C_n \text{ contains } p) \geq 1 - \alpha.$$

Solution

Using Hoeffding's Inequality, we have

$$P(|\hat{p}_n - p| > \epsilon_n) \leq 2e^{-2n\epsilon_n^2}.$$

Given ϵ_n as defined above, note that

$$\alpha = 2e^{-2n\epsilon_n^2}.$$

Thus finally, we have

$$P(C_n \text{ contains } p) = P(|\hat{p}_n - p| < \epsilon_n) \geq 1 - 2e^{-2n\epsilon_n^2} = 1 - \alpha.$$

Exercise 3 (Decreasing Rate Poissons)

(Based on **LW** 5.7) Let $\lambda_n = 1/n$ for $n = 1, 2, \ldots$ and let $X_n \sim \text{Poisson}(\lambda_n)$.

Also define $Y_n = nX_n$. Show that

$$Y_n \xrightarrow{p} 0.$$

Solution

We simply appeal to the definition of convergence in probability.

$$P(|Y_n - 0| > \epsilon) = P(nX_n > \epsilon) = P(X_n > \epsilon/n) \leq P(X_n > 0) = 1 - P(X_n = 0) = 1 - e^{-1/n}$$

Thus we have

$$P(|Y_n - 0| > \epsilon) \to 0$$

as $n \to \infty$. So we conclude that

$$Y_n \xrightarrow{p} 0.$$

Exercise 4 (More Classic Setup)

(**LW** 5.3) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed and $\mu = \mathbb{E}[X_1]$. Give that the variance is finite, show that

$$\bar{X}_n \xrightarrow{qm} \mu.$$

Solution

$$\mathbb{E}\left[(\bar{X}_n - \mu)^2\right] = \mathbb{V}\left[\bar{X}_n\right] = \frac{\mathbb{V}[X_1]}{n} \overset{n\to\infty}{\to} 0$$

The above holds since the variance is finite. Thus

$$\bar{X}_n \xrightarrow{qm} \mu.$$

Exercise 5 (The Sample Variance)

(**LW** 5.3) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed and finite mean $\mu = \mathbb{E}[X_1]$ and finite variance $\sigma^2 = \mathbb{V}[X_1]$. Let $\bar{X}_n$ be the sample mean and let S_n^2 be the sample variance. Show that

$$S_n^2 \xrightarrow{p} \sigma^2.$$

Solution

First note that we can rearrange the sample variance a bit.

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right) \\ &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2\right) \end{aligned}$$

Via the WLLN we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mu^2 + \sigma^2.$$

With the WLLN and the Continuous Mapping theorem, we also have

$$\left(\bar{X}_n\right)^2 \xrightarrow{p} \mu^2.$$

Then, as a result we can say that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\bar{X}_n\right)^2 \xrightarrow{p} \sigma^2.$$

Also note that

$$\frac{n}{n-1} \xrightarrow{p} 1$$

since

$$\lim_{n\to\infty} P\left(\left|\frac{n}{n-1} - 1\right| > \epsilon\right) = 0.$$

Then, finally, we can conclude that

$$S_n^2 \xrightarrow{p} \sigma^2.$$

Exercise 6 (Normal Approximations with the CLT)

(Based on **LW** 2.8) Suppose we have a computer program consisting of $n = 1000$ lines of code. (And somehow, someone wrote it without debugging along the way.) Let X_i be the number of errors on the i-th line of code. Suppose that the X_i are Poisson with mean 0.01 and that they are independent. Let Y be the sum of the X_i, that is, the total errors. Use the CLT to approximate the probability that there are 5 errors or less. Compare this to the exact probability.

Solution

First, note that

$$Y = \sum_{i=1}^{n} X_i \sim \text{Poisson}(\lambda_Y = 10).$$

Via the CLT, we can approximate Y with a Normal distribution, in particular

$$Y \approx \text{Normal}(\mu_Y = 10, \sigma_Y^2 = 10).$$

We then calculate the two probabilities in R.

```
# exact probability
ppois(q = 5, lambda = 10)
```

```
## [1] 0.06708596
```

```
# via approximation
pnorm(q = 5, mean = 10, sd = sqrt(10))
```

```
## [1] 0.05692315
```

```
# with continuity correction
pnorm(q = 5.5, mean = 10, sd = sqrt(10))
```

```
## [1] 0.07736446
```

Note that here the continuity correction doesn't help much. Also note, it is 2020, and we can simply calculate the Poisson probability directly.

Exercise 7 (CLT with Sample Variance)

Assuming the same conditions as the CLT, and knowing that the CLT exists, show that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{D} N(0, 1).$$

where S_n^2 is the sample variance.

Solution

From a previous exercise, we know that

$$S_n^2 \xrightarrow{p} \sigma^2.$$

Thus, we can also claim that

$$1/S_n \xrightarrow{p} 1/\sigma.$$

The CLT gives us

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Then, via Slutky's Theorem, we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{D} N(0, 1).$$

Exercise 8 (Clever Titles are Hard)

(**LW** 2.14) Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. Let $Y_n = \bar{X}_n^2$. Find the limiting distribution of Y_n.

Solution

First note that

$$\mathbb{E}[X_i] = \frac{1}{2} \quad \text{and} \quad \mathbb{V}[X_i] = \frac{1}{12}.$$

The WLLN gives

$$\bar{X}_n \xrightarrow{p} \frac{1}{2}.$$

Then via the Continuous Mapping Theorem, we have

$$Y_n \xrightarrow{p} \frac{1}{4}.$$

It would also be acceptable to note that the CLT give

$$\frac{\sqrt{n}\left(\bar{X}_n - \frac{1}{2}\right)}{\sqrt{\frac{1}{12}}} \xrightarrow{D} N(0, 1).$$

Thus, via the Delta Method we would obtain

$$\frac{\sqrt{n}\left(Y_n - \frac{1}{4}\right)}{\sqrt{\frac{1}{12}}} \xrightarrow{D} N(0, 1).$$

Exercise 9 (Coverage)

(Based on **LW** 4.4) Return to the results from Exercise 2. Set $\alpha = 0.2$ and $p = 0.4$. Use a simulation study to see how often this interval contains p. We call this quantity the interval's *coverage*. Do this for various sample sizes, n, between 1 and 10,000. Plot the coverage versus n. Note, for each n you will need to perform multiple simulations. Use enough values of n, and enough simulations for each, to create a reasonable looking plot.

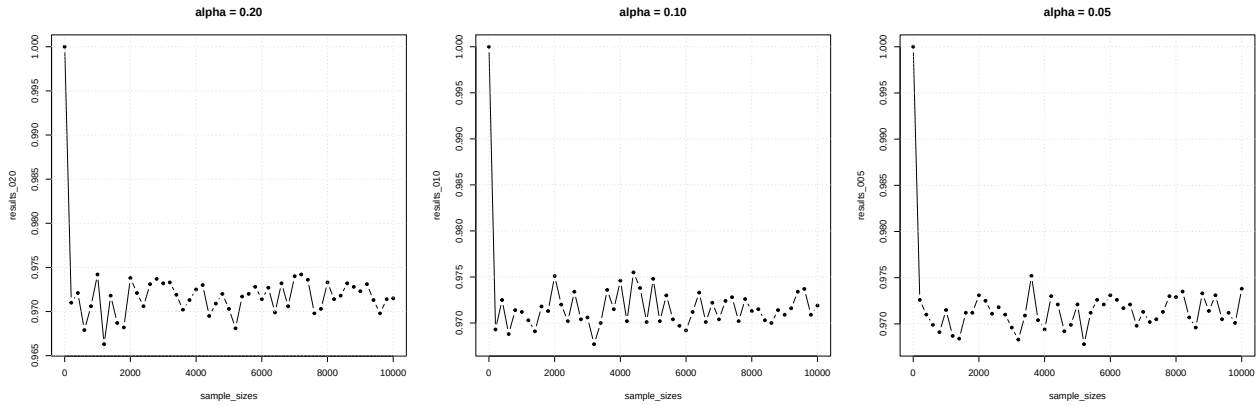
Solution

```
# simple data, create interval, check for true value
check_in_interval = function(sample_size, alpha) {
  data = rbinom(n = sample_size, size = 1, prob = 0.4)
  phat = mean(data)
  margin = sqrt((1 / (2 * sample_size)) * log(2 / 0.2))
  interval = phat + c(-1, 1) * margin
  interval[1] < 0.4 & 0.4 < interval[2]
}

# calculate coverage
calc_coverage = function(sample_size, alpha) {
  mean(replicate(n = 10000, check_in_interval(sample_size, alpha)))
}
```

```
# define sample sizes used
sample_sizes = seq(from = 1, to = 10001, by = 200)

# set seed
set.seed(42)
```

```
# perform simulations
results_020 = sapply(sample_sizes, calc_coverage, alpha = 0.20)
results_010 = sapply(sample_sizes, calc_coverage, alpha = 0.10)
results_005 = sapply(sample_sizes, calc_coverage, alpha = 0.05)
```

```
# plot results
par(mfrow = c(1, 3))
plot(sample_sizes, results_020, type = "b", pch = 20, main = "alpha = 0.20")
grid()
plot(sample_sizes, results_010, type = "b", pch = 20, main = "alpha = 0.10")
grid()
plot(sample_sizes, results_005, type = "b", pch = 20, main = "alpha = 0.05")
grid()
```



We have performed the simulations for two additional α values. Note that for each α, and any sample size, this interval appears to be valid. However note, that this interval appears to be over-covering, and α seems to have little effect.

Exercise 10 (Rate of Convergence)

So far, we have only been concerned with **if** a random variable converges, and to an extent, **how** a random variable converges, but we have not looked at the **rate** of convergence. To investigate this idea, consider random samples from two different distributions.

1. A Bernoulli like distribution with $P(X = -0.2) = P(X = 0.2) = 0.5$.
2. A t distribution with 2 degrees of freedom.

Note that both of these distributions have mean 0.

Generate a sample of size 10,000 from both and plot the sample mean against the sample size. Repeat this process three times and arrange the plots side-by-side. Comment on which distribute you believe converges faster.

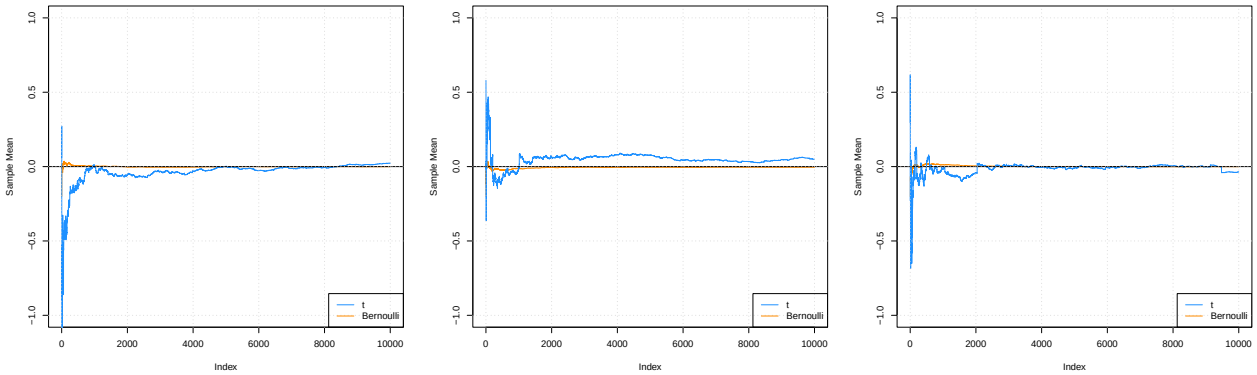
Solution

```
# define helper function
cummean = function(x) {
  cumsum(x) / 1:length(x)
}
```

```
# function to perform simulations and plot
sim_and_plot = function() {
  plot(cummean(sample(c(-0.2, 0.2), replace = TRUE, size = sample_size)),
       type = "l", ylim = c(-1, 1), ylab = "Sample Mean", col = "darkorange")
  lines(cummean(rt(n = sample_size, df = 2)), type = "l", col = "dodgerblue")
  abline(h = 0)
  grid()
  legend("bottomright", legend = c("t", "Bernoulli"),
         col = c("dodgerblue", "darkorange"), lty = 1)
}
```

```
# setup simulation
sample_size = 10000
set.seed(42)
```

```
# run simulations and plot results
par(mfrow = c(1, 3))
sim_and_plot()
sim_and_plot()
sim_and_plot()
```



We see that the "Bernoulli" distribution converges faster.

Exercise 11 (Hodges' Estimator)

Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Define

$$\hat{\theta}_n = \begin{cases} 0 & |\bar{X}_n| \leq n^{-1/4} \\ \bar{X}_n & |\bar{X}_n| > n^{-1/4} \end{cases}$$

Prove that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \begin{cases} 0 & \theta = 0 \\ N(0,1) & \theta \neq 0 \end{cases}$$

Solution

Recall,

$$\bar{X}_n \sim N\left(\theta, \frac{1}{n}\right).$$

First, note that

$$\begin{aligned} P\left(|\bar{X}_n| \leq n^{-1/4}\right) &= P\left(\sqrt{n}|\bar{X}_n| \leq n^{1/4}\right) \\ &= P\left(|Z + \sqrt{n}\theta| \leq n^{1/4}\right) \\ &= \Phi\left(n^{1/4} - \sqrt{n}\theta\right) - \Phi\left(-n^{1/4} - \sqrt{n}\theta\right). \end{aligned}$$

Here we use Z for a standard normal random variable.

Then, if $\theta > 0$

$$P\left(|\bar{X}_n| \leq n^{-1/4}\right) \to 0 - 0 = 0.$$

Also, if $\theta < 0$

$$P\left(|\bar{X}_n| \leq n^{-1/4}\right) \to 1 - 1 = 0.$$

Lastly, if $\theta = 0$

$$P\left(|\bar{X}_n| \leq n^{-1/4}\right) \to 1 - 0 = 1.$$

Now, we consider the two cases. First, when $\theta = 0$. For $\epsilon > 0$,

$$\begin{aligned} P\left(|\sqrt{n}\left(\hat{\theta}_n - 0\right) - 0| > \epsilon\right) &= P\left(|0 - 0| > \epsilon\right) P\left(\sqrt{n}|\bar{X}_n| \leq n^{1/4}\right) \\ &\quad + P\left(|\sqrt{n}\left(\hat{X}_n - 0\right) - 0| > \epsilon\right) P\left(\sqrt{n}|\bar{X}_n| > n^{1/4}\right) \\ &\quad \to 0. \end{aligned}$$

This hold because

$$P\left(|0-0|>\epsilon\right)=0$$
$$P\left(\sqrt{n}|\bar{X}_n|\leq n^{1/4}\right)\to 1$$
$$P\left(\sqrt{n}|\bar{X}_n|>n^{1/4}\right)\to 0$$

Thus we have $\sqrt{n}\left(\hat{\theta}_n-0\right)\xrightarrow{p}0$, and as a result $\sqrt{n}\left(\hat{\theta}_n-0\right)\xrightarrow{D}0$.

Now onto the case where $\theta\neq 0$.

Again, recall that

$$\sqrt{n}\left(\bar{X}_n-\theta\right)\xrightarrow{D}N\left(0,1\right).$$

Now, note that we can write

$$\hat{\theta}_n=\bar{X}_n\mathbf{1}_{[|\bar{X}_n|>n^{-1/4}]}$$

For simplicity of notation, define

$$Y=\mathbf{1}_{[|\bar{X}_n|>n^{-1/4}]}$$

Now, we will show that $Y\xrightarrow{p}1$.

$$P\left(|Y-1|\leq\epsilon\right)\geq P\left(|Y-1|=0\right)=P(Y=1)=P\left(|\bar{X}_n|>n^{-1/4}\right)\to 1.$$

Then, using Slutsky's theorem, we arrive at the desired result.

$$\sqrt{n}\left(\hat{\theta}_n-\theta\right)=\sqrt{n}\left(\bar{X}_n\mathbf{1}_{[|\bar{X}_n|>n^{-1/4}]}-\theta\right)\xrightarrow{D}N\left(0,1\right).$$

Thanks to Dave Zhao for the exercise suggestion. Note that this exercise was meant to be challenging.