Homework 7

STAT 984

Emily Robinson

November 7, 2019

Exercise 5.3

Suppose $X_n \sim \text{binomial}(n, p)$, where $0 < p < 1$.

(a) Find the asymptotic distribution of $g(X_n/n) - g(p)$, where $g(x) = \min\{x, 1-x\}$.

By the CLT, we know

$$\sqrt{n}\left(\frac{X_n}{n} - p\right) \xrightarrow{d} N\left(0, p(1-p)\right).$$

Then

$$g(x) = \begin{cases} p & \text{when } p \in (0, 1/2) \\ 1-p & \text{when } p \in (1/2, 1). \end{cases}$$

Therefore, $g'(p) = 1 \implies [g''(p)]^2 = 1$ when $p \in \{(0, 1/2) \cup (1/2, 1)\}$. Then by the delta method,

$$\sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(p)\right) \xrightarrow{d} N(0, p(1-p).$$

However, consider $p = 1/2$. Then $g'(p)$ does not exist, thus, the delta method does not apply. Therefore, note

$$\sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(1/2)\right) = -\sqrt{n}\left|\frac{X_n}{n} - 1/2\right|.$$

Thus, since the absolute value is a continuous function, by the CLT, for $p = 1/2$, we obtain

$$\sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(p)\right) \xrightarrow{d} -\sqrt{p(1-p)}|Z|$$

where $Z \sim N(0, 1)$.

(b) Show that $h(x) = \sin^{-1}(\sqrt{x})$ is a variance-stabilizing transformation for X_n/n. This is called the *arcsine transformation* of a sample proportion.

Hint: $(d/du)\sin^{-1}(u) = 1/\sqrt{1-u^2}$.

Consider $h(x) = \sin^{-1}(\sqrt{x})$. Then by the chain rule, $h'(x) = \frac{1}{\sqrt{1-x}}\frac{1}{2\sqrt{x}} = \frac{1}{2\sqrt{x(1-x)}}$. Then by the CLT and delta method,

$$\sqrt{n}\left[h\left(\frac{X_n}{n}\right) - h(p)\right] \xrightarrow{d} N\left(0, [h'(p)]^2 p(1-p)\right) = N\left(0, \frac{p(1-p)}{4p(1-p)}\right) = N\left(0, \frac{1}{4}\right).$$

Exercise 5.4

Let $X_1, X_2, ...$ be independent from $N(\mu, \sigma^2)$ where $\mu \neq 0$. Let

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

Find the asymptotic distribution of the coefficient of variation $S_n/\bar{X}_n$.

Note $s_n \xrightarrow{p} \sigma$ and $\bar{X}_n \xrightarrow{P} \mu$ implies $s_n/\bar{X}_n \xrightarrow{P} \sigma/\mu$. Therefore, $s_n/\bar{X}_n$ is a consistent estimator for σ/μ. Then

$$\sqrt{n}\left(s_n/\bar{X}_n - \sigma/\mu\right) = \frac{1}{\bar{X}_n}\sqrt{n}\left(s_n - \frac{\sigma}{\mu}\bar{X}_n\right) = \frac{1}{\bar{X}_n}\left(\sqrt{n}(s_n - \sigma) - \sqrt{n}\left(\frac{\sigma}{\mu}\bar{X}_n\right)\right).$$

Let $g(x) = \frac{\sigma}{\mu}x$. Then $g'(x) = \frac{\sigma}{\mu}$. Therefore, by the CLT and delta method,

$$\begin{aligned} & \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \\ \implies & \sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, [g'(\mu)]^2\sigma^2) \\ \implies & \sqrt{n}(\frac{\sigma}{\mu}\bar{X}_n - \sigma) \xrightarrow{d} N(0, \sigma^4/\mu^2). \end{aligned}$$

Then consider

$$\sqrt{n}(s_n - \sigma) = \sqrt{n}\left(\frac{s_n^2 - \sigma^2}{s_n + \sigma}\right).$$

Note, $s_n + \sigma \xrightarrow{P} 2\sigma$ and $\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$. Then by Slutsky's Theorem,

$$\sqrt{n}\left(\frac{s_n^2 - \sigma^2}{s_n + \sigma}\right) \xrightarrow{d} \frac{1}{2\sigma}N(0, 2\sigma^4) = N\left(0, \frac{2\sigma^4}{(2\sigma)^2}\right) = N\left(0, \frac{\sigma^2}{2}\right)$$

. Then, since $\bar{X}_n$ and s_n are independent,

$$\sqrt{n}(s_n - \sigma) - \sqrt{n}\left(\frac{\sigma}{\mu}\bar{X}_n\right) \xrightarrow{d} N\left(0, \frac{\sigma^4}{\mu^2} + \frac{\sigma^2}{2}\right).$$

Recall, $\bar{X}_n \xrightarrow{P} \mu$. Then by Slutksy's Theorem,

$$\frac{1}{\bar{X}_n}\left(\sqrt{n}(s_n - \sigma) - \sqrt{n}\left(\frac{\sigma}{\mu}\bar{X}_n\right)\right) \xrightarrow{d} \frac{1}{\mu}N\left(0, \frac{\sigma^4}{\mu^2} + \frac{\sigma^2}{2}\right) = N\left(0, \frac{\sigma^4}{\mu^4} + \frac{\sigma^2}{2\mu^2}\right).$$

Exercise 5.5

Let $X_n \sim \text{binomial}(n, p)$, where $p \in (0, 1)$ is unknown. Obtain confidence intervals for p in two different ways:

(a) Since $\sqrt{n}(X_n/n - p) \xrightarrow{d} N[0, p(1-p)]$, the variance of the limiting distribution depends only on p. Use the fact that $X_n/n \xrightarrow{P} p$ to find a consistent estimator of the variance and use it to derive a 95% confidence interval for p.

Since $\sqrt{n}\left(\frac{\bar{X}_n}{n} - p\right) \xrightarrow{d} N(0, p(1-p))$, the variance of the limiting distribution depends only on p. Note that $\frac{X_n}{n} \xrightarrow{P} p$ implies $\frac{X_n(n-X_n)}{n^2} \xrightarrow{P} p(1-p)$. Then by the CLT and Slutksy's Theorem,

$$\sqrt{n}(X_n/n - p)\sqrt{\frac{n^2}{X_n(n-X_n)}} \xrightarrow{d} N(0,1).$$

Therefore,

$$\begin{aligned}
& P\left[-1.96 < \sqrt{n}(X_n/n - p)\sqrt{\frac{n^2}{X_n(n-X_n)}} < 1.96\right] \approx 0.95 \\
\implies \quad & P\left[-1.96\frac{\sqrt{X_n(n-X_n)}}{n^{3/2}} < X_n/n - p < 1.96\frac{\sqrt{X_n(n-X_n)}}{n^{3/2}}\right] \approx 0.95 \\
\implies \quad & P\left[X_n/n - 1.96\frac{\sqrt{X_n(n-X_n)}}{n^{3/2}} < p < X_n/n + 1.96\frac{\sqrt{X_n(n-X_n)}}{n^{3/2}}\right] \approx 0.95.
\end{aligned}$$

(b) Use the result of problem 5.3(b) to derive a 95% confidence interval for p.

From 5.3(b), we know

$$2\sqrt{n}\left[\sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) - \sin^{-1}(\sqrt{p})\right] \xrightarrow{d} N(0,1).$$

Therefore,

$$\begin{aligned}
& P\left[-1.96 < 2\sqrt{n}\left[\sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) - \sin^{-1}(\sqrt{p})\right] < 1.96\right] \approx 0.95 \\
\implies \quad & P\left[-\frac{1.96}{2\sqrt{n}} < \sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) - \sin^{-1}(\sqrt{p}) < \frac{1.96}{2\sqrt{n}}\right] \approx 0.95 \\
\implies \quad & P\left[\sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) - \frac{1.96}{2\sqrt{n}} < \sin^{-1}(\sqrt{p}) < \sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) + \frac{1.96}{2\sqrt{n}}\right] \approx 0.95.
\end{aligned}$$

Then consider

$$f(x) = \begin{cases} 0 & \text{if } x \le 0 \\ [\sin(x)]^2 & \text{if } 0 < x < \pi/2 \\ 1 & \text{if } x \ge \pi/2. \end{cases}$$

Thus,

$$P\left[f\left(\sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) - \frac{1.96}{2\sqrt{n}}\right) < p < f\left(\sin^{-1}\left(\sqrt{\frac{X_n}{n}}\right) + \frac{1.96}{2\sqrt{n}}\right)\right] \approx 0.95.$$

(c) Evaluate the two confidence intervals in parts (a) and (b) numerically for all combinations of $n \in \{10, 100, 1000\}$ and $p \in \{.1, .3, .5\}$ as follows: For 1000 realizations of $X \sim \text{bin}(n, p)$, construct both 95% confidence intervals and keep track of how many times (out of 1000) that the confidence intervals contain p. Report the observed proportion of successes for each (n, p) combination. Does your study reveal any differences in the performance of these two competing methods?

The two methods appear to perfrom close to equally.

```r
sim <- function(n, p) {
    x <- rbinom(1000, n, p)
    ci1 <- sum(abs(sqrt(n) * (x/n - p) * n/sqrt(x *
        (n - x))) < 1.96)
    ci2 <- sum(abs(sqrt(n) * (asin(sqrt(x/n)) - asin(sqrt(p)))) <
        0.98)
    c(ci1, ci2)/1000
}

nseq <- c(10, 100, 1000)
pseq <- c(0.1, 0.3, 0.5)
results <- matrix(NA, 9, 5)
colnames(results) <- c("n", "p", "coverageA", "coverageB",
    "ratio")
it = 0
for (i in 1:3) {
    for (j in 1:3) {
        it = it + 1
        coverage <- sim(nseq[i], pseq[j])
        results[it, 1] <- nseq[i]
        results[it, 2] <- pseq[j]
        results[it, 3] <- coverage[1]
        results[it, 4] <- coverage[2]
        results[it, 5] <- coverage[1]/coverage[2]
    }
}
kable(results)
```

n	p	coverageA	coverageB	ratio
10	0.1	0.671	0.666	1.0075075
10	0.3	0.845	0.970	0.8711340
10	0.5	0.871	0.871	1.0000000
100	0.1	0.927	0.955	0.9706806
100	0.3	0.946	0.946	1.0000000
100	0.5	0.943	0.943	1.0000000
1000	0.1	0.952	0.948	1.0042194

n	p	coverageA	coverageB	ratio
1000	0.3	0.954	0.955	0.9989529
1000	0.5	0.955	0.955	1.0000000

Exercise 5.6

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed Normal $(0, \sigma^2)$ random variables.

(a) Based on the result of Example 5.7, Give an approximate test at $\alpha = .05$ for $H_0 : \sigma^2 = \sigma_0^2$ vs $H_a : \sigma^2 \neq \sigma_0^2$.

From Example 5.7, we know

$$\sqrt{n}\left[\log\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \log(\sigma^2)\right] \xrightarrow{d} N(0,2).$$

Then,

$$T_n = \frac{\sqrt{n}\left[\log\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \log(\sigma_0^2)\right]}{\sqrt{2}} \xrightarrow{d} N(0,1).$$

For a size $\alpha = 0.05$ test, we reject if $|T_n| > 1.96$.

(b) For $n = 25$, estimate the true level of the test in part (a) for σ_0^2=1 by simulating 5000 samples of size $n = 25$ from the null distribution. Report the proportion of cases in which you reject the null hypothesis according to your test (ideally, this proportion will be about .05).

```
sim2 <- function(n, sigsq, samps) {
    reject <- rep(NA, samps)
    for (i in 1:samps) {
        Xn <- rnorm(n = n, mean = 0, sd = sqrt(sigsq))
        Tn <- (sqrt(n) * (log((1/n) * sum(Xn^2)) -
            log(sigsq)))/sqrt(2)
        if (Tn < 0) {
            pvalue <- 2 * pnorm(q = Tn, lower.tail = T)
        } else {
            pvalue <- 2 * pnorm(q = Tn, lower.tail = F)
        }
        reject[i] <- (pvalue < 0.05)
    }
    sum(reject)/samps
}
sim2(25, 1, 5000)
```

```
## [1] 0.0626
```

Exercise 5.8

Assume $(X_1, Y_1), ..., (X_n, Y_n)$ are independent and identically distributed from some bivariate normal distribution. Let ρ denote the population correlation coefficient and r the sample correlation coefficient.

(a) Describe a test of $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ based on the fact that

$$\sqrt{n}[f(r) - f(\rho)] \xrightarrow{d} N(0, 1),$$

where $f(x)$ is a Fisher's transformation $f(x) = (1/2)\log[(1 + x)/(1 - x)]$. Use $\alpha = .05$.

Consider

$$T_n = \sqrt{n}[f(r) - f(\rho_0)] \xrightarrow{d} N(0, 1).$$

For a size $\alpha = 0.05$ test, we reject if $|T_n| > 1.96$.

(b) Based on 5000 repetitions each, estimate the actual level for this test in the case when $\mathrm{E}(X_i) = \mathrm{E}(Y_i) = 0, \mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 1$, and $n \in \{3, 5, 10, 20\}$.

The level of the test improves as the sample increases.

```
fisher <- function(x) {
    log((1 + x)/(1 - x))/2
}
rtest <- function(n) {
    z <- array(rnorm(10000 * n), c(5000, n, 2))
    r <- apply(z, 1, function(x) cor(x[, 1], x[, 2]))
    Tn <- sqrt(n) * (fisher(r) - fisher(0))
    sum(abs(Tn) > 1.96)
}
sapply(c(3, 5, 10, 20), rtest)/5000
```

```
## [1] 0.3872 0.1858 0.1060 0.0724
```

Exercise 5.9

Suppose that X and Y are jointly distributed such that X and Y are Bernoulli (1/2) random variables with $P(XY = 1) = \theta$ for $\theta \in (0, 1/2)$. Let $(X_1, Y_1), (X_2, Y_2), ...$ be independent and identically distributed with (X_i, Y_i) distributed as (X, Y).

(a) Find the asymptotic distribution of $\sqrt{n}[(\bar{X}_n, \bar{Y}_n) - (1/2, 1/2)]$.

Since $X \sim \mathrm{Bern}(1/2)$ and $Y \sim \mathrm{Bern}(1/2)$, we know $\sigma_x^2 = \sigma_y^2 = 1/4$. Since $P(XY = 1) = \theta$, we know $\sigma_x y = E[XY] - E[X]E[Y] = \theta - 1/4$ and $\rho = 4\theta - 1$. Then

$$\sqrt{n}\left[\begin{pmatrix}\bar{X}_n \\ \bar{Y}_n\end{pmatrix} - \begin{pmatrix}1/2 \\ 1/2\end{pmatrix}\right] \xrightarrow{d} N_2\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1/4 & \theta - 1/4 \\ \theta - 1/4 & 1/4\end{pmatrix}\right).$$

(b) If r_n is the sample correlation coefficient for a sample of size n, find the asymptotic distribution of $\sqrt{n}(r_n - \rho)$.

We know, $\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0, A\Sigma^* A^T)$ where

$$A = \left(\frac{-\sigma_{xy}}{2\sigma_x^3 \sigma_y}, \frac{-\sigma_{xy}}{2\sigma_x \sigma_y^3}, \frac{1}{\sigma_x \sigma_y} \right) = (2(1-4\theta), 2(1-4\theta), 4)$$

and

$$\Sigma^* = \begin{pmatrix} cov(X,X) & cov(X,Y) & cov(X,XY) \\ cov(Y,X) & cov(Y,Y) & cov(Y,XY) \\ cov(XY,X) & cov(XY,Y) & cov(XY,XY) \end{pmatrix} = \begin{pmatrix} 1/4 & \theta - 1/4 & \theta/2 \\ \theta - 1/4 & 1/4 & \theta/2 \\ \theta/2 & \theta/2 & \theta(1-\theta) \end{pmatrix}$$

since $X^2 = X$ and $Y^2 = Y$. Therefore,

$$A\Sigma^* A^T = 128\theta^3 - 144\theta^2 + 40\theta.$$

Thus, since $\theta = \frac{\rho+1}{r}$,

$$\sqrt{n}(r_n - \rho) \xrightarrow{d} N(0, 128\theta^3 - 144\theta^2 + 40\theta) \overset{d}{=} N(0, 2\rho^3 - 3\rho^2 - 2\rho + 3).$$

(c) Find a variance stabilizing transformation for r_n.

(d) Based on your answer to part (c), construct a 95% confidence interval for θ.

(e) For each combination of $n \in \{5, 20\}$ and $\theta \in \{.05, .25, .45\}$, estimate the true coverage probability of the confidence interval in part (d) by simulating 5000 samples and the corresponding confidence intervals. One problem you will face is that in some samples, the sample correlation coefficient is undefined because with positive probability each of the X_i or Y_i will be the same. In such cases, consider the confidence interval to be undefined and the true parameter therefore not contained therein.

Hint: To generate a sample of (X, Y), first simulate the X's from their marginal distribution, then simulate the Y's according to the conditional distribution of Y given X. TO obtain this conditional distribution, find $P(Y = 1|X = 1)$ and $P(Y = 1|X = 0)$.