STAT430 Homework #5: Due Friday, March 8, 2019.

Name:

---

0. We will continue with **Chapter 8** on estimation. You should have read **Sections 8.1-8.4** on point estimation; please continue with **Sections 8.5-8.10** on interval estimation and related topics.

1. Suppose $[\hat{\theta}_L, \hat{\theta}_U]$ is a $(1-\alpha)$ 100% confidence interval for the target parameter θ.

(a) Find a $(1-\alpha)$ 100% confidence interval for the new target parameter $a + b\theta$, where a is a known real number and $b > 0$ is a known real number.

Answer:

$$1 - \alpha = P(\hat{\theta}_L \leq a + b\theta \leq \hat{\theta}_U), b > 0$$
$$= P(\frac{\hat{\theta}_L - a}{b} \leq \theta \leq \frac{\hat{\theta}_U - a}{b})$$
$$[\frac{\hat{\theta}_L - a}{b}, \frac{\hat{\theta}_U - a}{b}]$$

(b) Find a $(1-\alpha)$ 100% confidence interval for the new target parameter $a + b\theta$, where a is a known real number and $b < 0$ is a known real number.

Answer:

$$1 - \alpha = P(\hat{\theta}_L \leq a + b\theta \leq \hat{\theta}_U), b < 0$$
$$= P(\frac{\hat{\theta}_L - a}{b} \geq \theta \geq \frac{\hat{\theta}_U - a}{b})$$
$$[\frac{\hat{\theta}_U - a}{b}, \frac{\hat{\theta}_L - a}{b}]$$

(c). Suppose $Y_1, \ldots, Y_{16}$ are iid $N(\mu, \sigma^2)$ and are measured in degrees Fahrenheit.
Form a 90% confidence interval for the true mean μ if the realized data are as follows:

```
Y <- c(97.91, 98.81, 97.42, 98.01, 98.25, 98.18, 98.90, 98.25, 99.59, 98.50, 98.60, 97.92, 97.48, 99.21
```

Answer:

This is the generic 90% CI for $\bar{Y}$

$$\bar{Y} \pm t_{\alpha/2,n-1}\sqrt{S^2/n}$$

```
Y <- c(97.91, 98.81, 97.42, 98.01, 98.25, 98.18, 98.90, 98.25, 99.59, 98.50, 98.60, 97.92, 97.48, 99.21

Y_bar <- mean(Y)

lower <- Y_bar + qt(.05,length(Y)-1)*sqrt(var(Y)/length(Y))
```

```r
upper <- Y_bar - qt(.05,length(Y)-1)*sqrt(var(Y)/length(Y))


lower
```

```
## [1] 98.15876
```

```r
upper
```

```
## [1] 98.69249
```

So a 90% CI for $\bar{Y}$ is,

$$[98.15985, 98.6914]$$

(d). Suppose we are interested in the true mean in degrees Celsius = (5/9) * (Fahrenheit - 32), instead of Fahrenheit. Use the data in part (c) to compute a 90% confidence interval for the true mean in degrees Celsius.

Answer:

```r
X <- c(97.91, 98.81, 97.42, 98.01, 98.25, 98.18, 98.90, 98.25, 99.59, 98.50, 98.60, 97.92, 97.48, 99.21

Y <- (5/9)*(X -32)

Y_bar <- mean(Y)

lower <- Y_bar + qt(.05,length(Y)-1)*sqrt(var(Y)/length(Y))
upper <- Y_bar - qt(.05,length(Y)-1)*sqrt(var(Y)/length(Y))


lower
```

```
## [1] 36.75487
```

```r
upper
```

```
## [1] 37.05138
```

So a 90% CI for the temperature in Celsius is,

$$[36.75547, 37.05078]$$

---

2. The code in the answer block below computes and plots exact $(1-\alpha)$ 100% two-sided confidence intervals for `nreps` replicates of the sampling experiment in which $Y_1, \ldots, Y_n$ iid $N(\theta, \sigma^2)$, with θ as our target and σ^2 unknown. The pivot used to produce the CI is

$$Q = \frac{\bar{Y} - \theta}{\sqrt{S^2/n}} \sim t_{n-1},$$

where $\bar{Y}$ is the sample mean of the observations and S^2 is the sample variance. The quantiles then come from the t distribution with $n-1$ degrees of freedom, and by isolating θ in the probability statement, we get

$$1 - \alpha = P\left(q_{\alpha/2} \leq Q \leq q_{1-\alpha/2}\right) = P\left(\bar{Y} + q_{\alpha/2}\sqrt{\frac{S^2}{n}} \leq \theta \leq \bar{Y} + q_{1-\alpha/2}\sqrt{\frac{S^2}{n}}\right),$$

so

$$\hat{\theta}_L = \bar{Y} + q_{\alpha/2}\sqrt{\frac{S^2}{n}}, \quad \hat{\theta}_U = \bar{Y} + q_{1-\alpha/2}\sqrt{\frac{S^2}{n}}.$$

(Because the t distribution is symmetric about zero, $q_{\alpha/2} = -q_{1-\alpha/2}$, and we often write $\bar{Y} \pm q_{1-\alpha/2}\sqrt{S^2/n}$ for the CI.)

Run the code below with the given settings, then increase the sample size to $n = 100$ and run again. Comment on the width and the *coverage* (empirical proportion of hits) of the CIs at the two sample sizes. What changes and what does not change?

Answer:

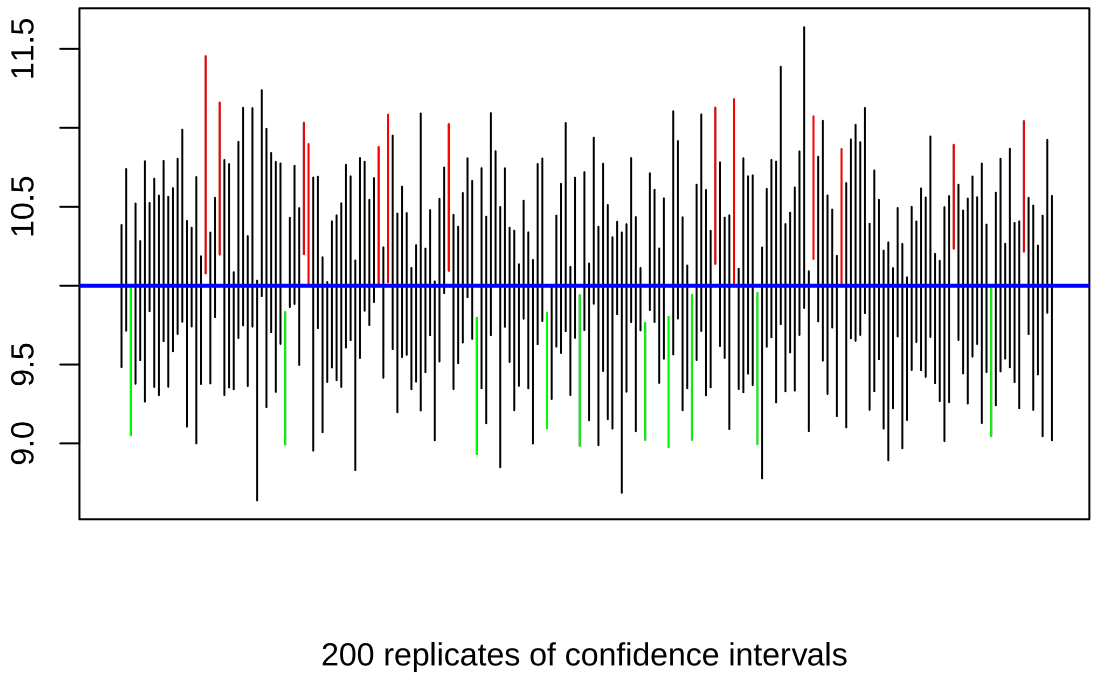
```
nreps <- 200
theta <- 10
n <- 10
alpha <- 0.1
set.seed(4302019)
Y <- rnorm(n * nreps, mean = theta, sd = 1)
YY <- matrix(Y, n, nreps)  # nreps iid samples of size n, one in each column
sample_mean <- apply(YY, MAR = 2, FUN = "mean")  # sample mean, Ybar, for each column
sample_var  <- apply(YY, MAR = 2, FUN = "var")   # sample variance, S^2, for each column
# Compute lower confidence limit, using quantile from t distribution
theta_hat_L <- sample_mean + qt(alpha / 2, df = n - 1) * sqrt(sample_var / n)
# Compute upper confidence limitk using quantile from t distribution
theta_hat_U <- sample_mean + qt(1 - alpha / 2, df = n - 1) * sqrt(sample_var / n)
too_high <- (theta < theta_hat_L)  # CI is too high and missed theta
too_low  <- (theta > theta_hat_U)  # CI is too low and missed theta
mean(too_low + too_high)      # proportion of misses should be close to alpha
```

```
## [1] 0.115
```

```
1 - mean(too_low + too_high)  # proportion of hits should be close to 1 - alpha
```

```
## [1] 0.885
```

```
# Plot the CIs.
plot(c(0,1), c(min(theta_hat_L), max(theta_hat_U)), type = "n",
     xlab = paste(nreps, "replicates of confidence intervals"), ylab = "", xaxt = "n")
x <- (1:nreps) / nreps  # index for each of the nreps CIs
segments(x, theta_hat_L, x, theta_hat_U)
segments(x[too_high], theta_hat_L[too_high], x[too_high], theta_hat_U[too_high], col = "red")
segments(x[too_low], theta_hat_L[too_low], x[too_low], theta_hat_U[too_low], col = "green")
abline(h = theta, col = "blue", lwd = 2)
```



The proportion of coverage is 0.885 when n = 10.

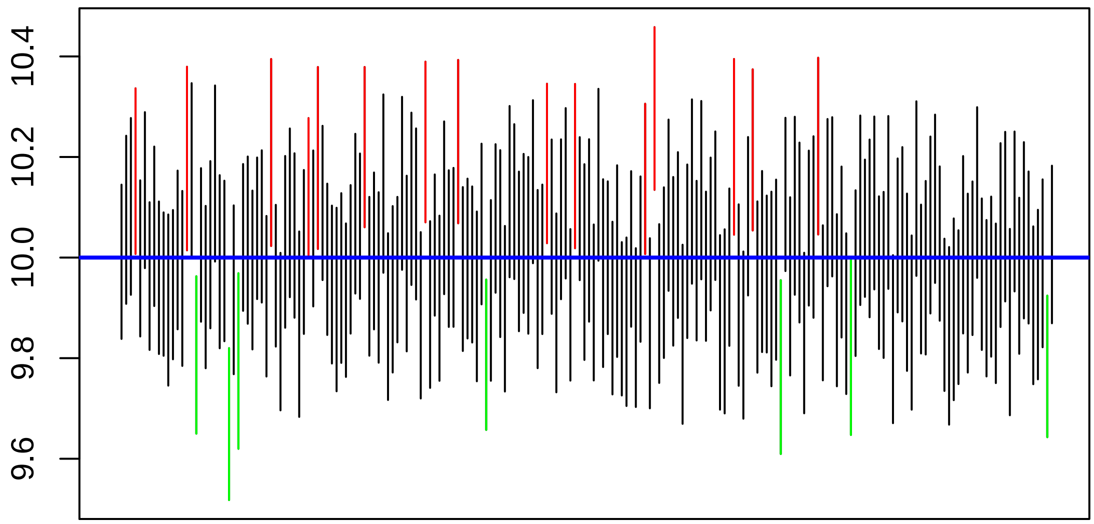
```
nreps <- 200
theta <- 10
n <- 100
alpha <- 0.1
set.seed(4302019)
Y <- rnorm(n * nreps, mean = theta, sd = 1)
YY <- matrix(Y, n, nreps)  # nreps iid samples of size n, one in each column
sample_mean <- apply(YY, MAR = 2, FUN = "mean")  # sample mean, Ybar, for each column
sample_var  <- apply(YY, MAR = 2, FUN = "var")   # sample variance, S^2, for each column
# Compute lower confidence limit, using quantile from t distribution
theta_hat_L <- sample_mean + qt(alpha / 2, df = n - 1) * sqrt(sample_var / n)
# Compute upper confidence limitk using quantile from t distribution
theta_hat_U <- sample_mean + qt(1 - alpha / 2, df = n - 1) * sqrt(sample_var / n)
too_high <- (theta < theta_hat_L)  # CI is too high and missed theta
too_low  <- (theta > theta_hat_U)  # CI is too low and missed theta
mean(too_low + too_high)      # proportion of misses should be close to alpha
```

```
## [1] 0.11
```

```
1 - mean(too_low + too_high)  # proportion of hits should be close to 1 - alpha
```

```
## [1] 0.89
```

```
# Plot the CIs.
plot(c(0,1), c(min(theta_hat_L), max(theta_hat_U)), type = "n",
     xlab = paste(nreps, "replicates of confidence intervals"), ylab = "", xaxt = "n")
x <- (1:nreps) / nreps  # index for each of the nreps CIs
segments(x, theta_hat_L, x, theta_hat_U)
segments(x[too_high], theta_hat_L[too_high], x[too_high], theta_hat_U[too_high], col = "red")
segments(x[too_low], theta_hat_L[too_low], x[too_low], theta_hat_U[too_low], col = "green")
abline(h = theta, col = "blue", lwd = 2)
```



200 replicates of confidence intervals

When n = 100, the proportion of coverage is 0.89. So, as n increases the coverage converges to 0.9 because we want to get 90% CI.

---

3. In class, we considered the sampling experiment in which $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$, with $\theta = \sigma^2$ as our target. The pivot used to produce the CI is

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where S^2 is the sample variance. The quantiles then come from the χ^2 distribution with $n-1$ degrees of freedom, and by isolating θ in the probability statement, we obtained the two-sided CI,

$$\left[\frac{(n-1)S^2}{q_{1-\alpha/2}}, \frac{(n-1)S^2}{q_{\alpha/2}}\right].$$

Modify the code from problem 2 above to compute and plot `nreps=200` 95% confidence intervals for σ^2 when $n = 20$ and the true population parameters are $\mu = 10$, $\sigma^2 = 1$. Report your empirical coverage (proportion of hits).

Answer:

```
nreps <- 200
theta <- 1
n <- 20
alpha <- 0.05
set.seed(4302019)
Y <- rnorm(n * nreps, mean = 10, sd = theta)
YY <- matrix(Y, n, nreps)  # nreps iid samples of size n, one in each column
sample_mean <- apply(YY, MAR = 2, FUN = "mean")  # sample mean, Ybar, for each column
sample_var  <- apply(YY, MAR = 2, FUN = "var")   # sample variance, S^2, for each column
# Compute lower confidence limit, using quantile from t distribution
theta_hat_L <- sample_var*(n-1)/qchisq(1-alpha / 2, df = n - 1)
```

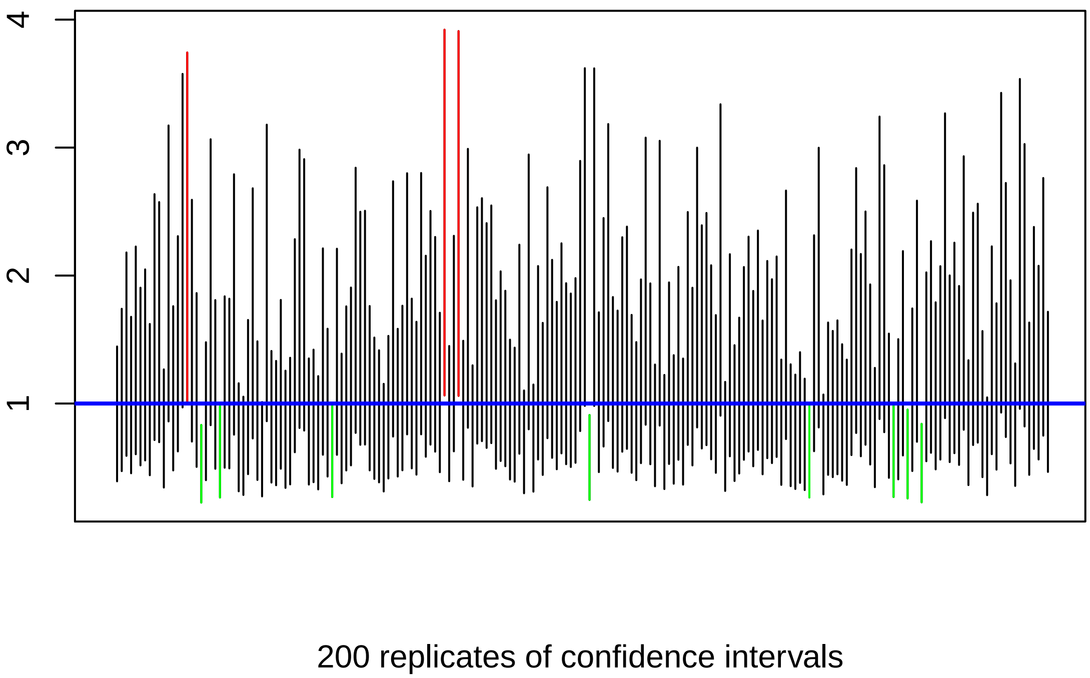
```r
# Compute upper confidence limitk using quantile from t distribution
theta_hat_U <- sample_var*(n-1)/qchisq(alpha / 2, df = n - 1)
too_high <- (theta < theta_hat_L)  # CI is too high and missed theta
too_low  <- (theta > theta_hat_U)  # CI is too low and missed theta
mean(too_low + too_high)      # proportion of misses should be close to alpha
```

```
## [1] 0.055
```

```r
1 - mean(too_low + too_high)  # proportion of hits should be close to 1 - alpha
```

```
## [1] 0.945
```

```r
# Plot the CIs.
plot(c(0,1), c(min(theta_hat_L), max(theta_hat_U)), type = "n",
     xlab = paste(nreps, "replicates of confidence intervals"), ylab = "", xaxt = "n")
x <- (1:nreps) / nreps  # index for each of the nreps CIs
segments(x, theta_hat_L, x, theta_hat_U)
segments(x[too_high], theta_hat_L[too_high], x[too_high], theta_hat_U[too_high], col = "red")
segments(x[too_low], theta_hat_L[too_low], x[too_low], theta_hat_U[too_low], col = "green")
abline(h = theta, col = "blue", lwd = 2)
```



so 0.945 is the proportion of coverage by the simulation.

---

4. Consider the setting of problem 3. The true variance, σ^2, is naturally bounded below by zero. In some cases, we might only be interested in finding an upper confidence limit for σ^2, so that we have the one-sided confidence interval $(-\infty, \hat{\theta}_U]$, but we know this is actually $[0, \hat{\theta}_U]$. For this problem, the target is still σ^2 and we still use the same pivot.

(a). Show that

$$1 - \alpha = P\left(0 \leq \sigma^2 \leq \frac{(n-1)S^2}{q_\alpha}\right),$$

so that

$$\left[0, \frac{(n-1)S^2}{q_\alpha}\right]$$

is a valid $(1-\alpha)$ 100% confidence interval for σ^2.

Answer:

Since, $Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$1-\alpha = P(\frac{(n-1)S^2}{q_{\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S^2}{q_{1-\alpha/2}})$

So,

$1-\alpha = P(q_\alpha \leq Q) = P(\frac{(n-1)S^2}{q_\alpha} \geq \sigma^2)$

Therefore,

$$[0, \frac{(n-1)S^2}{q_\alpha}]$$

is a valid $(1-\alpha)$100% CI.

(b). Using exactly the same setting as problem 3, modify the code to compute and plot the one-sided 95% confidence intervals, and report their empirical coverage. In this case, you can replace the lower endpoint by a vector of 0's: `theta_hat_L <- rep(0, nreps)`.

Answer:

```
nreps <- 200
theta <- 1
n <- 20
alpha <- 0.05
set.seed(4302019)
Y <- rnorm(n * nreps, mean = 10, sd = theta)
YY <- matrix(Y, n, nreps)  # nreps iid samples of size n, one in each column
sample_mean <- apply(YY, MAR = 2, FUN = "mean")  # sample mean, Ybar, for each column
sample_var  <- apply(YY, MAR = 2, FUN = "var")   # sample variance, S^2, for each column
# Compute lower confidence limit, using quantile from t distribution
theta_hat_L <- 0
# Compute upper confidence limitk using quantile from t distribution
theta_hat_U <- sample_var*(n-1)/qchisq(alpha, df = n - 1)
too_high <- (theta < theta_hat_L)  # CI is too high and missed theta
too_low  <- (theta > theta_hat_U)  # CI is too low and missed theta
mean(too_low + too_high)      # proportion of misses should be close to alpha
```

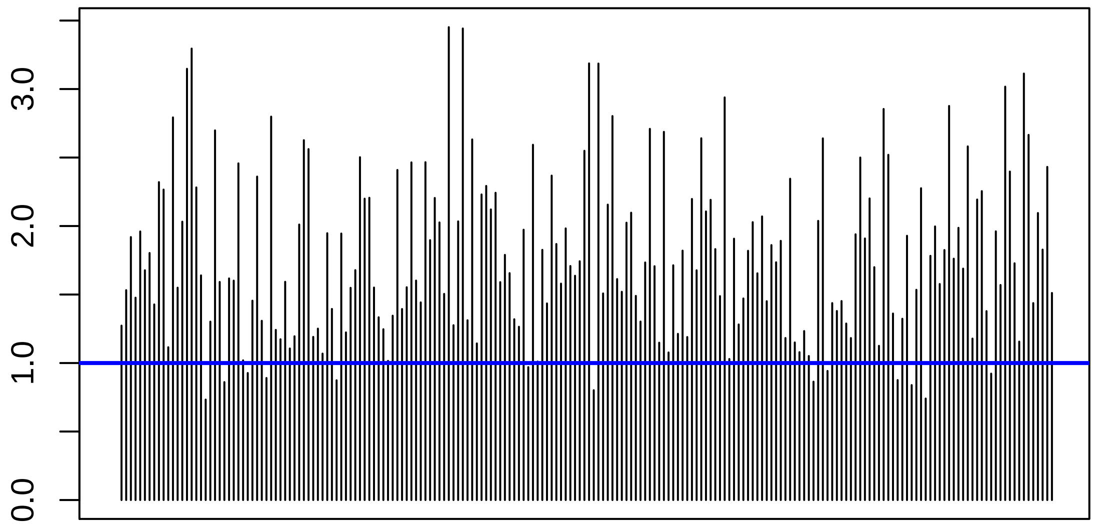
```
## [1] 0.065
```

```
1 - mean(too_low + too_high)  # proportion of hits should be close to 1 - alpha
```

```
## [1] 0.935
```

```
# Plot the CIs.
plot(c(0,1), c(min(theta_hat_L), max(theta_hat_U)), type = "n",
     xlab = paste(nreps, "replicates of confidence intervals"), ylab = "", xaxt = "n")
x <- (1:nreps) / nreps  # index for each of the nreps CIs
```

```r
segments(x, theta_hat_L, x, theta_hat_U)
segments(x[too_high], theta_hat_L[too_high], x[too_high], theta_hat_U[too_high], col = "red")
segments(x[too_low], theta_hat_L[too_low], x[too_low], theta_hat_U[too_low], col = "green")
abline(h = theta, col = "blue", lwd = 2)
```



200 replicates of confidence intervals

---

5. Complete **Exercise 8.40** of the text. In this problem, the variance is *known* to be 1.

Answer:

a)

$$0.95 = P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U)$$

Where the target is $\theta = \mu$

Pivot:

$Q = \frac{Y-\theta}{\sigma} \sim N(0,1)$, where $\sigma = 1$

Quantiles:

$1 - \alpha = P(q_{\alpha/2} \leq Q \leq q_{1-\alpha/2}) = P(Y - q_{\alpha/2} \geq \theta \geq Y - q_{1-\alpha/2})$

Report:

$$[Y - q_{1-\alpha/2}, Y - q_{\alpha/2}]$$

is a 95% CI for μ where $q_{\alpha/2} = -1.96$

So, we get

$$[Y - 1.96, Y + 1.96]$$

b)

$$0.95 = P(Q \leq q_\alpha)$$

$$0.95 = P(\theta \leq 1.645 + Y)$$

So,

$$(-\infty, 1.645 + Y]$$

is a 95% upper confidence limit for μ

c)

$$0.95 = P(Q \geq q_\alpha)$$

$$0.95 = P(\theta \geq -1.645 + Y)$$

So,

$$[-1.645 + Y, \infty)$$

is a 95% lower confidence limit for μ

6, parts 6(a) and 6(b). Complete parts (a) and (b) of **Exercise 8.44** of the text.

Answer:

a)

$$F_Y(y) = \int_0^y \frac{2(\theta - t)}{\theta^2} dt = \frac{2y}{\theta} - \frac{y^2}{\theta^2}$$

b)

$$Q = \frac{Y}{\theta}$$

Using the CDF method,

$$P(\frac{Y}{\theta} \leq y) = P(Y \leq \theta y)$$

Now plug in θy in to the CDF of Y

$$F_Y(\theta y) = \frac{2\theta y}{\theta} - \frac{(\theta y)^2}{\theta^2} = 2y - y^2 = F_{\frac{Y}{\theta}}(y)$$

Which does not depend on θ so Q is a valid pivot.

6(c). Quantiles are defined by the property that

$$F_Q(q_\alpha) = P(Q \leq q_\alpha) = \alpha,$$

where $F_Q(y)$ is the cumulative distribution function of the random variable, Q. Use this fact and the result of 6(b) to find a general expression for the αth quantile of the pivotal distribution, then use `R` to compute the quantiles $q_{0.025}$, $q_{0.05}$, $q_{0.1}$, $q_{0.9}$, $q_{0.95}$, and $q_{0.975}$.

Answer:

So we know,

$\alpha = 2q_\alpha - q_\alpha^2$

```r
q_solve <- function(alpha){

  polyroot(c(1,-2, alpha))

}

q_solve(0.025)
```

```
## [1]  0.5031647+0i 79.4968353-0i
```

```r
q_solve(0.05)
```

```
## [1]  0.5064113-0i 39.4935887+0i
```

```r
q_solve(.1)
```

```
## [1]  0.513167-0i 19.486833+0i
```

```r
q_solve(.9)
```

```
## [1] 0.7597469+0i 1.4624753-0i
```

```r
q_solve(.95)
```

```
## [1] 0.817256+0i 1.288007+0i
```

```r
q_solve(.975)
```

```
## [1] 0.8634729-0i 1.1878091+0i
```

So, here is the quantiles

$q_{0.025} = 0.503$

$q_{0.05} = 0.506$

$q_{0.1} = 0.513$

$q_{0.9} = 0.759$

$q_{0.95} = 0.8172$

$q_{0.975} = 0.8635$

6(d). Use your computed quantiles and pivotal quantity to compute a 90% two-sided confidence interval for θ.

Answer:

$$0.9 = P(q_{0.05} \leq Q \leq q_{0.95}) = P(\frac{Y}{q_{0.95}} \leq \theta \leq \frac{Y}{q_{0.05}})$$

So a 90% CI is

$$[\frac{Y}{0.8172}, \frac{Y}{0.506}]$$

6(e). Use your computed quantiles and pivotal quantity to compute the 90% upper confidence limit for θ (This is **Exercise 8.45(a)** of the book, so you can check your answer in the back.)

Answer:

$$[-\infty, \frac{Y}{0.759}]$$

---

7. (a), (b) and (c): Complete **Exercise 8.47** of the text. (d) Then repeat part (c) with the same setup ($\bar{y} = 4.77$, 95%), but with $n = 21$. Comment on the difference between the results in (c) and (d).

Answer:

7. a, b, c) Complete 8.47

Refer to 8.46. Assume that $Y_1, \ldots Y_n$ is a sample of size n from an exponential dist with mean θ

a) Use the method of mgf to show that $2\sum_{i=1}^{n} Y_i/\theta$ is a pivotal quantity and has a χ^2 dist. with 2n df.

b) Use the pivotal quantity $2\sum_{i=1}^{n} Y_i/\theta$ to derive a 95% CI for θ

c) If a sample of size n = 7 yields $\bar{y} = 4.77$, use the result from part b) to give a 95% CI for θ

d) Repeat part c) with the same setup ($\bar{y} = 4.77$, 95%), but with n = 21. Comment on the difference between the results in c) and d)

a) mgf

$E(e^{tY})$ $\qquad Y_i \sim Exp(\theta)$

$f(y) = \frac{1}{\cancel{\beta}\theta} e^{-y/\theta}$

target $\mu = \theta$

$\mu_{Y_i} = \theta$, $\sigma^2_{Y_i} = \theta^2$

mgf of Exp(θ) = ~~exp~~ $\cancel{e^{\mu t + \frac{t^2\sigma^2}{2}}}$ $= (1-\beta t)^{-1}$

Pivot $= 2\sum_i Y_i/\theta = Q = \frac{2}{\theta}\sum Y_i$

$mgf(Q) = E(e^{tQ})$

$= E(e^{t\,2\sum Y_i/\theta})$?

$= E(e^{t\,2Y_1/\theta}) \cdot E(e^{t\,2Y_2/\theta}) \cdots E(e^{t\,2Y_n/\theta})$

$E(e^{(\frac{2t}{\theta})Y_i})$

$= \left(E(e^{t2Y_i/\theta})\right)^n = ((1-\theta(\frac{2t}{\theta}))^{-1})^n$

$= ((1-2t)^{-1})^n$

$= \left(\frac{1}{1-2t}\right)^n$ is mgf of χ^2_{2n}

$Q \sim \chi^2_{2n}$

Does not depend on θ

b) $2\sum Y_i/\theta \sim \chi^2_{2n}$ | $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, $\frac{2nS^2}{\sigma^2} \sim \chi^2_{2n}$

$0.95 = 1 - 0.05 = \cancel{P(q_{0.025}}$

$= P(q_{0.025} \le Q \le q_{0.975})$

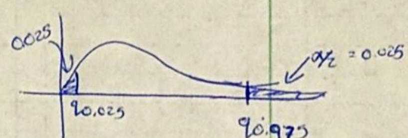
$= P(q_{0.025} \le \frac{2}{\theta}\sum Y_i \le q_{0.975})$

$= P\left(\frac{2\sum Y_i}{q_{0.025}} \ge \theta \ge \frac{2\sum Y_i}{q_{0.975}}\right)$ (general equation)

$\cancel{q_{0.025}}$

c) $\bar{y} = \frac{\sum y_i}{n}$, $n = 7$, $\bar{y} = 4.77$

$\sum y_i = n\bar{y} = 7 \cdot 4.77 = 33.39$



$\left[\frac{2 \cdot 33.39}{26.119}, \frac{2 \cdot 33.39}{5.629}\right]$

$[2.557, 11.864]$ is a 95% CI for θ, $n=7$

d) $n = 21$, $\bar{y} = 4.77$

$\sum y_i = 21 \cdot 4.77 = 100.17$

$\left[\frac{2 \cdot 100.17}{\text{qchisq}(0.975, df=42)}, \frac{2 \cdot 100.17}{\text{qchisq}(0.025, df=42)}\right]$ is a 95% CI for θ, $n=21$