STA 032 R Final

Hardy Jones
999397426
Professor Melcon
Winter 2015

1. (a) $P(Z)$

n	probability
10 000	0.3240
100 000	0.3300

(b) $P(T)$

n	probability
10 000	0.3220
100 000	0.3190

(c) $P(P)$

n	probability
10 000	0.3540
100 000	0.3510

(d) $P(M|Z)$

n	probability
10 000	0.7760
100 000	0.7860

(e) $P(M^c|P^c)$

n	probability
10 000	0.3060
100 000	0.2970

(f) $P(Z \cap M^c)$

n	probability
10 000	0.0724
100 000	0.0705

(g) $P(T \cup M)$

n	probability
10 000	0.7680
100 000	0.7740

2. (a) θ_A

estimate	error
0.9810	0.0572

(b) θ_B

estimate	error
0.1480	0.0236

(c) θ_C

estimate	error
4.710	0.417

3. (a) i.

	Lower	Upper
Corrected	0.685	0.871
Uncorrected	0.707	0.893

ii.

	Lower	Upper
Corrected	0.240	0.543
Uncorrected	0.229	0.540

(b) i.

Corrected	Uncorrected
0.9614	0.9329

ii.

Corrected	Uncorrected
0.9475	0.9475

iii.

Corrected	Uncorrected
0.9601	0.9445

iv.

Corrected	Uncorrected
0.9362	0.9362

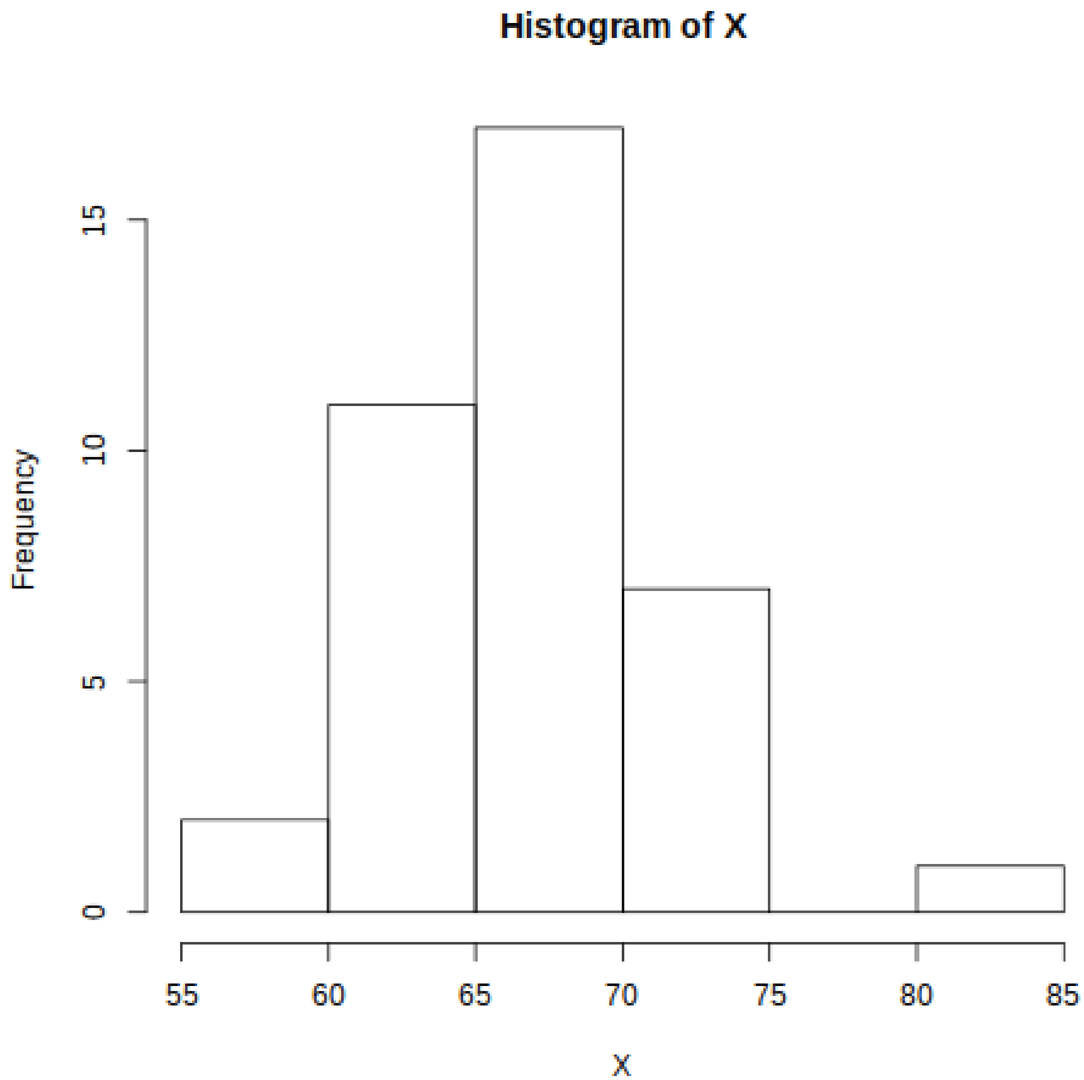
(c) The difference is large enough that using the corrected version makes sense. It does not add more asymptotic complexity, but it provides a greater coverage probability. You basically get better coverage for free.

(d) The true proportion affects the uncorrected version more than the corrected version. The sample size does not really affect either version.

(e) The corrected version can have worse coverage when the true mean is very low and the sample size is very large.

(f) Given the results of (d) and (e), it would depend on the situation. If the sample size is very large and the true mean is very small then it would be better to use the uncorrected version. If not, it's better to use the corrected version.

4. We start by generating a histogram:



The data appears to be distributed approximately normally.

(a) $\mu_0 = 67$

Non-parametric	Parametric	Theoretical
0.637	0.627	0.629

(b) $\mu_0 = 68$

Non-parametric	Parametric	Theoretical
0.146	0.151	0.153

(c) $\mu_0 = 69$

Non-parametric	Parametric	Theoretical
0.00628	0.00849	0.00863

Appendix A R code

Problem 1

```
male <- "Male"
female <- "Female"
sexProbs <- c(0.65, 0.35)

protoss <- "Protoss"
terran <- "Terran"
zerg <- "Zerg"
races <- c(protoss, terran, zerg)
maleProbs <- c(0.3, 0.3, 0.4)
femaleProbs <- c(0.45, 0.35, 0.2)

Simulate <- function(n) signif(Raw(n), 3)

Raw <- function(n) {
  sex <- sample(c(male, female), n, prob = sexProbs, replace = TRUE)
  race <- sapply(c(1:n), function (x) ChooseRace(sex[x]))

  c( "P(Z)"            = sum(race == zerg) / n
   , "P(T)"            = sum(race == terran) / n
   , "P(P)"            = sum(race == protoss) / n
   , "P(M|Z)"          = sum(sex == male & race == zerg) / sum(race == zerg)
   , "P(M^c|P^c)"      = sum(sex != male & race != protoss) / sum(race != protoss)
   , "P(Z \\cap M^c)" = sum(race == zerg & sex != male) / n
   , "P(T \\cup M)"   = sum(race == terran | sex == male) / n
   )
}

ChooseRace <- function(sex)
  sample(races, 1, prob = if (sex == male) maleProbs else femaleProbs)
```

Problem 2

```
library(MASS)

Bootstrap <- function(X) function(f) function(n) {
  bs <- replicate(n, f(sample(X, length(X), replace = TRUE)))
  signif(c(estimate = mean(bs), error = sd(bs)), 3)
}

BootstrapDays <- Bootstrap(quine$Days)

ThetaA <- BootstrapDays(function(x) sd(x) / mean(x))
ThetaB <- BootstrapDays(function(x) median(x) / diff(range(x)))
ThetaC <- BootstrapDays(function(x) diff(range(x)) / sd(x))
```

Problem 3

```r
Corrected <- function(trials, alpha) {
  X <- sum(trials)
  n <- length(trials)
  n.tilde <- n + 4
  p.tilde <- (X + 2) / n.tilde
  z <- qnorm(alpha / 2, lower.tail = FALSE)
  foo <- sqrt(p.tilde * (1 - p.tilde) / n.tilde)
  bar <- z * foo
  signif(c(low = p.tilde - bar, high = p.tilde + bar), 3)
}

Uncorrected <- function(trials, alpha) {
  X <- sum(trials)
  n <- length(trials)
  p.hat <- X / n
  z <- qnorm(alpha / 2, lower.tail = FALSE)
  foo <- sqrt(p.hat * (1 - p.hat) / n)
  bar <- z * foo
  signif(c(low = p.hat - bar, high = p.hat + bar), 3)
}

Confidences <- function(trials, alpha) Conf(alpha)(trials)

Conf <- function(alpha) function(trials)
  matrix( c(Corrected(trials, alpha), Uncorrected(trials, alpha)), 2, 2, TRUE
        , list(c("Corrected", "Uncorrected"), c("Lower", "Upper")))

Proportion <- function(pop, alpha, size, simulations, true.mean) {
  # This is a `size x simulations` matrix.
  sims <- replicate(simulations, sample(pop, size))

  # Construct a confidence interval for each simulation.
  # This is a `4 x simulations` matrix
  confs <- apply(sims, 2, Conf(alpha))
  # Grab row vectors of each respective low and high in the interval.
  # Each is a `1 x simulations` row vector.
  corrected.low    <- confs[1, ]
  corrected.high   <- confs[3, ]
  uncorrected.low  <- confs[2, ]
  uncorrected.high <- confs[4, ]

  covered.mean <- covered(true.mean)

  # Find out how many intervals covered the true.mean.
  # These are both `1 x simulations` row vectors.
  corrected.covered <- mapply(covered.mean, corrected.low, corrected.high)
  uncorrected.covered <- mapply(covered.mean, uncorrected.low, uncorrected.high)

  # Using the duck typing of R, take the mean of both covered vectors.
  c(mean(corrected.covered), mean(uncorrected.covered))
}
```

```r
# Helper function that curries its arguments.
covered <- function(val) function(low, high)
    low <= val && val <= high
```

Problem 4

```r
X <- read.csv("STA-32-Winter-2015-Final-Data.txt")$X64

png("prob4.png")
hist(X)
dev.off()

NonParImpl <- function(b) function(mu0) {
  X.boot <- X - mean(X) + mu0
  bs <- replicate(b, mean(sample(X.boot, length(X.boot), replace = TRUE)))
  signif(mean(bs < mean(X)), 3)
}

NonPar <- NonParImpl(100000)

ParImpl <- function(b) function(mu0) {
  s <- sd(X)
  bs <- replicate(b, mean(rnorm(X, mu0, s)))
  signif(mean(bs < mean(X)), 3)
}

Par <- ParImpl(100000)

Theor <- function(mu0) {
  X.bar <- mean(X)
  s <- sd(X)
  n.root <- sqrt(length(X))
  z <- (X.bar - mu0) / (s / n.root)
  signif(pnorm(z), 3)
}
```