Solutions to assignment 1

1.5 Target population: homeless people (in US?)
Sampling unit: homeless person who received medical attention from one of the clinics in the HCH project.
Sampling frame: the medical record of the clinic patients
Observation unit: homeless persons of the clinic.

The argument is not valid because this is a sample of convenience. The sampled population is consisted of homeless people who chose to visit the clinic. People with mental illness are more likely to visit a clinic than normal people. Hence the proportion from the sample may overestimate that of the target population.

1.9 Target population: Entries in Wikipedia
Sampling frame: Wikipedia's science articles
Sampling unit: Wikipedia's science articles
Observation unit: Wikipedia's science articles.

This is a judgement sample, usually not a good representative of the target population. Also the nonresponse rate (16%) is little high.
1.13. Target population: people who are going to attend JSM meeting in 2006.

Sampling unit: person who attended the 2005 JSM.

Sampling frame: registration list of 2005 JSM.

Observation unit: attendees of 2005 JSM.

The sampled population is not the same as the target population because persons who attended 2005 JSM may not attend 2006 JSM. Vise versa. In addition, the response rate may not be very high. People receiving a request through an email is not obliged to submit their opinions online in a short period of given time.

2.6

a. See next page. The shape is right-skewed.

b. The estimated mean is 1.78, with a standard error 0.3674.

c. Yes. Because the sample mean will be approximately normal when the sample size is large, by CLT.

d. The proportion of faculty members with no publications is

\[ \hat{p} = \frac{28}{50} = 0.56 \]

\[ \text{SE}(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{50}{80}\right) \frac{0.56 \cdot 0.44}{49}} \]

A 95% CI of \( \hat{p} \) is given by [\( \hat{p} \pm t_{0.025, n-1} \text{SE}(\hat{p}) \), \( \hat{p} + t_{0.025, n-1} \text{SE}(\hat{p}) \)]

\[ = (0.422, 0.698) \]
29. To show \( n = \frac{Z^2 \frac{s}{\mu_0} S^2}{\epsilon^2 + \frac{Z^2 \frac{s}{\mu_0} S^2}{N}} \) satisfies \( \epsilon = Z^2 \left( 1 - \frac{n}{N} \right) \frac{S}{\mu_0} \) is equivalent to show \( \epsilon^2 = Z^2 \left( 1 - \frac{n}{N} \right) \frac{S^2}{N} = \frac{Z^2 \frac{s}{\mu_0} S^2}{N} - \frac{Z^2 \frac{s}{\mu_0} S^2}{N} \)

Substitute \( n \) in (29) in the right-hand side to get

\[
\frac{Z^2 \frac{s}{\mu_0} S^2}{Z^2 \frac{s}{\mu_0} S^2} \left( \epsilon^2 + \frac{Z^2 \frac{s}{\mu_0} S^2}{N} \right) = \frac{Z^2 \frac{s}{\mu_0} S^2}{N} = \epsilon^2. \quad \text{Hence the proof.}
\]

2. 12

a. For 95\% CI with margin of error 0.10

\[ n_0 = \frac{(1.96)^2}{4(0.10)^2} \approx 96. \]

The sample size required in an SRSWOR is

\[ n = \frac{n_0}{1 + n_0/N} = \frac{96}{1 + \frac{96}{580}} \approx 83. \]

b. Let \( p = \frac{\sum_{i=1}^{N} Y_i}{N} \) where \( Y_i = \begin{cases} 1, & \text{not overdue} \\ 0, & \text{overdue} \end{cases} \)

be the population proportion of not overdue.

In a sample of size \( n \)

\[ \hat{p} = \frac{\sum_{i=1}^{N} Y_i}{n} \] where \( Y_i = \begin{cases} 1, & \text{1st select not overdue} \\ 0, & \text{i\textsuperscript{th} select overdue} \end{cases} \)

\[ \Pr(Y_i = 1) = p, \quad \Pr(Y_i = 0) = 1 - p \]

from an SRSWOR because \( Y_i = 1 \iff \text{independently pick one from } Y_i = 1 \)
Hence $V(p^\hat{}) = \frac{V(\hat{\theta})}{n^2} = \frac{\hat{\theta}(1-\hat{\theta})}{n} = \frac{1}{n} p(1-p)$

$V(p^\hat{}) = \frac{1}{n} \hat{p}(1-\hat{p})$  \[se(p^\hat{}) = \sqrt{\frac{1}{n} \hat{p}(1-\hat{p})}\]

In this case \( \hat{p} = \frac{27}{120} = 0.225 \)  \[se(p^\hat{}) = 0.0381\]

A 95\% CI for \( p \) is given by \( [\hat{p} - t_{\alpha/2, n-1} \cdot se(p^\hat{}), \hat{p} + t_{\alpha/2, n-1} \cdot se(p^\hat{})] \)

\[= (0.225 - 1.96 \times 0.0381, 0.225 + 1.96 \times 0.0381) = (0.150, 0.300)\]

2.14

a. First of all, not all members are listed in the online category. Secondly, the information obtained online may be inaccurate, e.g., the employment category may change but is not updated online.

b. C. did not provide data set.

2.16

a. See attachment. The shape of the data is highly skewed with most of golf courses charging less than $40, but some charging over $100.

b. The average weekday greens fees to play 9 holes is $20.15 and the standard error of this estimator is 1.630

2.17

a. See attachment. The shape is left skewed.

b. The mean back tee yardage is 5309.767 with s.e. 164.532.
\( \hat{p} = 0.708 \quad n = 120 \quad N = 14938 \)

\[
SE(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}} = \sqrt{\left(1 - \frac{120}{14938}\right) 0.708 \times 0.292/120} = 0.0413
\]

A 95% CI for the population proportion is given by

\[
[\hat{p} - 1.96 \times SE(\hat{p}), \hat{p} + 1.96 \times SE(\hat{p})] = (0.627, 0.789)
\]

2.22 a.

\[
CV(\hat{p}) = CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \frac{\sqrt{V(\hat{p})}}{E(\hat{p})} = \frac{\sqrt{\frac{N-n}{n-1} \frac{p(1-p)}{n}}}{p} = \sqrt{\frac{N-n}{n-1} \frac{1-p}{pn}}
\]

For a large \( N \) \( 1 - \frac{1}{n} \approx 1 \)

\( CV(\bar{y}) \) for a sample size \( n \) is \( n = 1 \)

\[
CV = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{S}{\bar{y}}} = \frac{S}{\bar{y}}
\]

by 2.26 \( n = \frac{3 \sigma^2 \bar{y}^2}{(r \bar{y})^2 + \frac{3 \sigma^2 S^2}{N}} = \frac{3 \sigma^2 \left(\frac{S}{\bar{y}}\right)^2}{r^2 + \frac{3 \sigma^2 \bar{y}^2}{N}} = \frac{3 \sigma^2 CV^2}{r^2 + \frac{3 \sigma^2 CV^2}{N}}
\]

for sample size 1

\[
CV(\hat{p}) = \sqrt{\frac{1-p}{p}} \quad \text{then sample size}
\]

\[
n = \frac{3 \sigma^2 (1-p)/p}{r^2 + \frac{3 \sigma^2 (1-p)/p}{N}} = \frac{3 \sigma^2 (1-p)}{r^2 p + \frac{3 \sigma^2 (1-p)}{N}}
\]

b. With large \( N \) \( S^2 \approx p(1-p) \)

\[
n = \frac{3 \sigma^2 p(1-p)}{e^2 + \frac{3 \sigma^2 p(1-p)}{N}} \quad \text{by 2.25 since} \quad fpc = 1 \quad \frac{n \times 0.01}{N} \approx 0
\]

\[
n \approx \frac{3 \sigma^2 p(1-p)}{e^2}
\]
By 2.26 \( n = \frac{3 \sigma^2}{r^2 \delta} \)

Hence 0.001 0.005 0.01 0.05 0.1 0.5 0.7 0.9 0.95 0.99 0.995

By 2.25 5 22 43 203 385 897 1067 897 385 203 43 22 5

By 2.26 4.26 \times 10^2 3.40 \times 10^6 4.23 \times 10^6 3.24 \times 10^7 3.84 \times 10^6 1.11 \times 10^8 1.71 \times 10^7 3.73 \times 10^5 585, 249, 44, 225

When \( p \) is extremely small, \( n \) is extremely large by relative error.

23. \( P(\text{no missing}) = \binom{3078 - 19}{3000} = \frac{3059! 2778!}{3078! 2759!} \)

\[
= \frac{2778 \cdot 2777 \cdot \ldots \cdot 2760}{3078 \cdot 3077 \cdot \ldots \cdot 3060} = 0.1416
\]

25. See attachment.

28:

a. \( Q_i \) take possible values 0, 1, \ldots, \( n \)

\[
\sum Q_i = n \quad Q_i = \sum_{j=1}^{n} Y_{ij} \quad Y_{ij} = 1 \quad \text{if the } i^{th} \text{ unit appears in the } j^{th}
\]

\[
P(Y_{ij} = 1) = p_i = \frac{1}{N}
\]

Hence \( Q_1, Q_2, \ldots, Q_N \) is multinomial with \( n \) trials and \( p_i = \ldots \)

\[
P_N = \frac{1}{N}
\]

b. \( E(\hat{t}) = \frac{N}{n} \sum_{i=1}^{N} E(Q_i) Y_i = \frac{N}{n} \sum_{i=1}^{N} \frac{n}{N} Y_i = \frac{N}{n} \sum_{i=1}^{N} Y_i = t
\]
C. \[ V(t) = \left( \frac{N}{n} \right)^2 \sum_{i=1}^{N} (Q_i) Y_i^2 \]
\[ = \frac{N^2}{n^2} \sum_{i=1}^{N} \frac{n}{N} \frac{N-1}{N} Y_i^2 \]
\[ = \frac{N^2}{n} \sum_{i=1}^{N} Y_i^2 \]

29. Let \( P_{k,n} \) be the probability that any particular sample is chosen at the \( k \)th step of the sampling process \( k=1,2,\ldots \).
Assume \( P_{k,n} = \frac{1}{(n+k)} \) for \( k \geq 1 \). Then when the \((k+1)\)th member of the population is encountered, either:

a) the updated sample does not contain the \((k+1)\)th member
or b) the \( l \)th population member (for some \( l < k+1 \)) which was present in the current sample is replaced by the \((n+k+1)\)th member

When (a) occurs, the probability of the updated sample is

\[ P_{k+1,n} = P_{k,n} \left[ 1 - \frac{n}{(n+k+1)} \right] = \frac{1}{(n+k+1)} \]

when (b) occurs, the prob. that the \( l \)th member was chosen from among the \( n \) sample members is \( 1/n \). The updated sample could also be obtained if the \( l \)th member was replaced by any of the other \( k \) population members not in the current sample. Since each of those possible samples has prob. \( P_{k,n} \), the unconditional prob. of the updated sample is

\[ P_{k+1,n} = P_{k,n} \left[ \frac{1}{(n+k+1)} \right] = \frac{1}{(n+k+1)} \]

Hence the series (29) converges.
Back Tee Yardage

Estimated Sampling Distribution
### 2.6 code
## a.
faculty <- c(rep(0, 28), rep(1, 4), rep(2, 3), rep(3, 4),
             rep(4, 4), rep(5, 2), rep(6, 1), rep(8, 2), 9, 10)
postscript('hist2.6.eps')
hist(faculty, xlab = 'Referred Publications',
     ylab = 'Frequency', breaks = 11, main = '')
dev.off()
## b. mean and se
f.mean <- mean(faculty)
n <- 50
N <- 807
f.se <- sqrt(1/n * (1-n/N)*var(faculty))
## d.
p <- 28/50
p.se <- sqrt((1-n/N) * p * (1-p) / (n-1))
l.limit <- p - qt(0.975, n-1)*p.se
u.limit <- p + qt(0.975, n-1)*p.se

### 2.16 code
## a.
golfsrs <- read.table('golfsrs.txt', header = T, sep = ', ')
wkday9 <- golfsrs$WKDAY9
postscript('hist2.16.eps')
hist(wkday9, xlab = 'Weekday Greens Fees',
     ylab = 'Frequency', main = '')
dev.off()
## b.
wkday9.mean <- mean(wkday9)
n <- 120
N <- 14938
wkday9.se <- sqrt(1/n * (1-n/N)*var(wkday9))

### 2.17 code
## a.
backtee <- golfsrs$BACKTEE
postscript('hist2.17.eps')
hist(backtee, xlab = 'Back Tee Yardage',
     ylab = 'Frequency', main = '')
dev.off()
## b.
backtee.mean<-mean(backtee)
n<-120
N<-14938
backtee.se<-sqrt(1/n*(1-n/N)*var(backtee))

### 2.18 code
hole18<-sum(1*(golfsrs$HOLES==18))
p.18<-hole18/120

### 2.22 code
pvec<-c(0.001,0.005,0.01,0.05,0.1,0.3,0.5,0.7,0.9,0.95,0.99,0.995,0.999)
n2.25<-1.96^2*pvec*(1-pvec)/0.03^2
n2.26<-1.96^2*(1-pvec)/0.03^2/pvec^3

### 2.25 code
acres92<-read.table('agstrat.txt',header=T,sep=',')$ACRES92
samplemeans<-rep(0,500)
for(i in 1:500) {
  s1<-sample(acres92, 100, replace=T)
  samplemeans[i]<-mean(s1)
}
postscript('hist2.25.eps')
hist(samplemeans, main='',xlab='Estimated Sampling Distribution', ylab='Frequency')
dev.off()