Stats 332 Assignment 3

Junqiao Lin

Mar 06, 2019

```
library(survey)
```

Question 1.

```
#importing the dataset
Q1.dat <- read.csv("clusters.csv")
Q1.datnoclust <- Q1.dat$price
```

a)

```
#Wants consistance results
set.seed(1734)
#Design the sample for 1000 samples.
Q1a.acclst <- c()
for(x in 1:1000){
  Q1a.acclst[[x]] <- mean(sample(Q1.datnoclust,40))
}

Q1a.var <- var(Q1a.acclst)
Q1a.var
```

```
## [1] 0.1768737
```

Hence the variance of design A for the given experiment is around 0.1768737.

b)

```
#Assumeing we are mark on correctness
Q1clulst <- c()

for(x in 1:10){
  Q1clulst[[x]] <- mean(Q1.datnoclust[seq(((x-1)*10 + 1), x*10)])
}

print(Q1clulst)
```

```
##  [1] 18.34162 18.48854 12.08536 10.15489 13.14202 10.56011 13.64063
##  [8] 14.87469 17.14089 15.30463
```

Hence the mean of the cluster is the following

1	18.34162	2	18.48854	3	12.08536	4	10.15489	5	13.14202
6	10.56011	7	13.64063	8	14.87469	9	17.14089	10	15.30463

c)

```
#A bit rough, but it works
Q1c.acclst <-c()
set.seed(1717)
```

```r
for(x in 1:1000){
  Q1c.acclst[[x]] <- mean(Q1clulst[sample(seq(1:10),4)])
}


Q1c.var<- var(Q1c.acclst)
Q1c.var
```

```
## [1] 1.329108
```

Hence the variance of design A for the given experiment is around 1.329108.

d)

```r
Q1c.var/Q1a.var
```

```
## [1] 7.514447
```

Hence $\frac{Var\tilde{\mu}_B}{Var\tilde{\mu}_A} = 7.514447$ according to my simulation. e)

```r
Q1e.clustactvar = (sum((Q1clulst - mean(Q1.datnoclust))^2)/9) * (1-4/10)/4
Q1e.actvar = (1 - 40/100) * var(Q1.datnoclust)/40
Q1e.clustactvar/Q1e.actvar
```

```
## [1] 7.470915
```

Hence $\frac{Var\tilde{\mu}_B}{Var\tilde{\mu}_A} = 7.470915$. Which is close to my estimate.

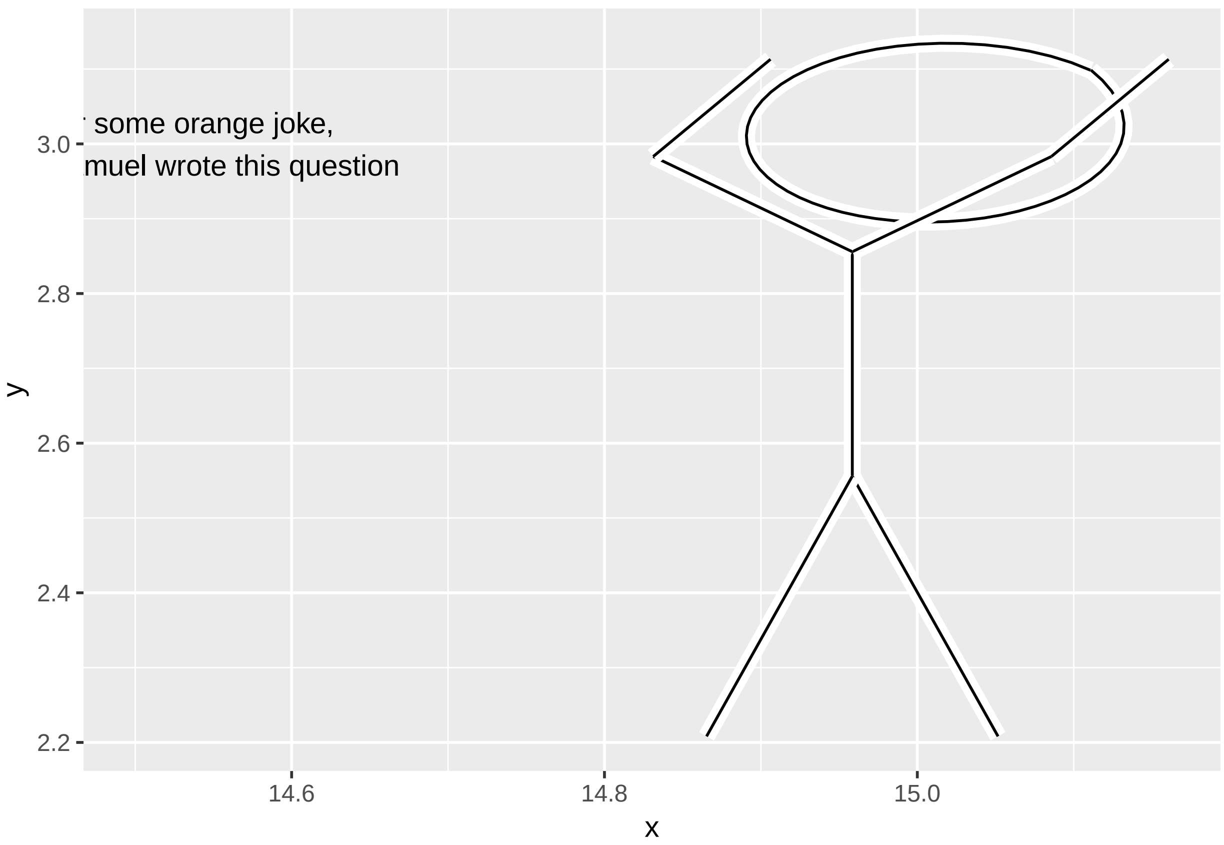
```r
library(xkcd)
```

```r
ratioxy = 1
mapping <- aes(x, y, scale, ratioxy, angleofspine,
+ anglerighthumerus, anglelefthumerus,
+ anglerightradius, angleleftradius,
+ anglerightleg, angleleftleg, angleofneck)
```

```
## Warning: Duplicated aesthetics after name standardisation:
```

```r
dataman <- dataman <- data.frame(x= c(15), y=c(3),
 scale = c(0.3) ,
 ratioxy = ratioxy,
 angleofspine = -pi/2 ,
 anglerighthumerus = c(pi/4),
 anglelefthumerus = c(pi/2 + pi/4),
 anglerightradius = c(pi/3),
 angleleftradius = c(pi/3),anglerightleg = 3*pi/2 - pi / 12,
 angleleftleg = 3*pi/2 + pi / 12 ,
 angleofneck = runif(1, 3*pi/2-pi/10, 3*pi/2+pi/10))

p <- ggplot() +
  xkcdman(mapping, dataman) +
  annotate("text", x=14.5, y = 3, label = "No snake or some orange joke, \n Are you sure samuel wrote t
p
```


' some orange joke,
muel wrote this question
3.0
2.8
2.6
2.4
2.2
y
14.6
14.8
15.0
x

Question 2.

a)

```
Q2.dat <- read.csv("citrus.csv")
Q2.clust <- svydesign(id=~cnum+onum, data=Q2.dat, fpc=~fpc1+fpc2)

svymean(~Q2.dat$quality,Q2.clust )
```

```
##                    mean     SE
## Q2.dat$quality 7.2027 0.2098
```

Hence the estimate of the mean quality of the orange is 7.2 and the standard error is around 0.20

b)

```
var(Q2.dat$quality)
```

```
## [1] 1.549705
```

Hence assuming SRS FPC ≈ 1, $SE(\tilde{\mu}) \approx \sqrt{\frac{1}{200}(1)\tilde{\sigma}} = 0.08802570647$.

c) Since the orange is taken from different crate, and different crate might be coming from different farm. Although your method is easier, and have a lower SE, however since the sample only contains 10 oranges coming from 20 crates, you are effectively only estimating the quality of orange coming from the 20 farm where the crate came from instead of oranges from Florida, which can have a huge source of bias.

Question 3.

a) Response: The burn time for a single flare. Factors: Which design types is used. Level: Flare type 1 or Torch type 2 (2 levels). Treatment: Buring a flare from type 1 or buring a flare from type 2. Experimental unit: A torch

b) Make sure the torch is pick uniformly random form the avaliable torches, make sure the environment used is same.

c) Under the null hypothesis assume there is no difference bewteen the average burn time.

Calculting $\bar{Y}_{1\cdot}$ and $\bar{Y}_{2\cdot}$ we have the following:

```
#I am just going to purposely name this goose and nogoose, also hi Nam
goose <- c(65.2,82.3,81.1,67.4,57.3,59.5,66.1,75.0,82.2,70.0)
nogoose <- c(67.9,59.8,75.0,72.7,86.3,77.8,62.7,85.6,69.0,82.9)

mean(goose)
```

```
## [1] 70.61
```

```
mean(nogoose)
```

```
## [1] 73.97
```

And calculating $\tilde(\sigma)$

```
v.hat <-  (9*var(goose) + 9*var(nogoose))/18
v.hat
```

```
## [1] 85.27167
```

Hence calculating the test statistic:

$$S = \frac{(70.61 - 73.97)}{(\sqrt{85.27167 * \frac{2}{10}})}$$

Which is the following

```
test <- abs((70.61 - 73.97)/(sqrt(85.27167 *  (2/10))))
2 * pt(test, 18, lower.tail=F)
```

```
## [1] 0.4264947
```

Hence we fail to reject the null hypothesis since $0.42 \geq 0.05$

Question 4

a)

I would recommend the use of blocking by the induvidual judge since each person might have different innate opinion towards beer in general regardless of brand. With each judge, I would recommend to select the judge randomly and I would recommend giving each judge both brand of beer and do not inform them the brand of the beer to avoid potential bias.

b)

```
# Thanks Nam... , so samuel said we are getting mark by correctness
shorter <- c(7,3, 9, 7, 6, 8, 2, 6, 8, 4)
longer <- c(5, 4, 7, 7, 5, 6, 2, 7, 6, 3)

# number of units
n <- sapply(list(shorter, longer), length)

# overall mean
mu.hat <- mean(c(shorter, longer))

t.hat <- sapply(list(shorter, longer), mean) - mu.hat

b.hat <- colMeans(rbind(shorter, longer)) - mu.hat

v.hat <- var(shorter - longer)/2
test <- abs(diff(t.hat)/sqrt(v.hat * (2/length(b.hat))))
df <- length(b.hat) - 1
2 * pt(test, df, lower.tail=F)
```

```
## [1] 0.06970738
```

Hence we fail to reject the null hypthesis since $0.06970738 \geq 0.05$

c)

```
v.hat.nb <- ((n[1] - 1)*var(shorter) + (n[2] - 1)*var(longer))/(sum(n) - 2)
test.nb <- abs(diff(t.hat)/sqrt(v.hat.nb * sum(1/n)))
df.nb <- sum(n) - 2
2 * pt(test.nb, df.nb, lower.tail=F)
```

```
## [1] 0.3942286
```

Hence we fail to reject the null hythesis since $0.3942286 \geq 0.05$