class16_assignment

Bruce Mallory

3/3/2021

problem #1

A machine produces rope at a mean rate of mean of 4 feet per minute with standard deviation of 5 inches. Assume that the amounts produced in different minutes are independent and identically distributed, approximate the probability that the machine will produce at least 250 feet in one hour.

```
#mu = 4 ft/min or 48 inch/min
mu <- 48
#sd = 5 inches/min
sd <- 5
#n= 1 hour or 60 minutes
n <- 60
#z_star is 250 ft/hour or 250*12 / 60 in/min
z_star <- (250*12/60)
prob <- pnorm(z_star, mu, sd/sqrt(n), lower.tail=FALSE)
cat("The probability of producing 250 ft or more of rope in an hour is",
    round(100*prob, 3),"%.")
```

```
The probability of producing 250 ft or more of rope in an hour is 0.097 %.
```

problem #2

Assume that the distribution of the number of defects on any given bolt of cloth is the Poisson distribution with mean 5, and the number of defects on each bolt is counted for a random sample of 125 bolts. Determine the probability that the average number of defects per bolt in the sample will be less than 5.5 defects per bolt.

```r
#lamda = 5 defects/bolt
lam <- 5
#z_star = 5.5 defects/bolt
z_star <- 5.5
#using the normal approximation of Poisson (see rationale below)
sd <- sqrt(lam)
n <- 125



prob.pois <- ppois(z_star, lambda = lam, lower.tail = TRUE)
#Why isn't the poisson calculation the same as the normal approximation???
cat("Using ppois(), I get", prob.pois, "\n")
```

```
Using ppois(), I get 0.6159607
```

```r
#I know the normal is correct (saw it on the web)
prob.norm <- pnorm(z_star, lam, sd/sqrt(n), lower.tail=TRUE)
cat("The probability that there will be less than 5.5 defects per bolt is",
    round(100*prob.norm, 1),"%.")
```

```
The probability that there will be less than 5.5 defects per bolt is 99.4 %.
```

I've used Normal approximation to Poisson because "the random variable that follows Poisson distribution can be approximated to the normal distribution when the lambda is large. The thumb rule for this is that the lambda must be greater than approximately 5 and the sample size must be greater than 100."

problem #3

Suppose that the proportion of defective items in a large manufactured lot is 0.1. What is the smallest random sample of items that must be taken from the lot in order for the probability to be at least 0.99 that the proportion of defective items in the sample will be less than 0.13?

```r
p <- 0.1
prob <- 0.99
z_star <- 0.13

#using the binomial
for (n in 1:1000){
  if (pbinom(trunc(n*z_star), n, p, lower.tail=TRUE) > 0.99){
    break
  }
}
cat("The smallest sample size is", n)
```

```
The smallest sample size is 554
```

```r
#using the normal approximation
for (m in 1:1000){
  if (pnorm(z_star, p, sqrt(p*(1-p)/m), lower.tail=TRUE) > 0.99){
    break
  }
}
cat("The smallest sample size is", m)
```

```
The smallest sample size is 542
```

problem #4

Suppose that 16 digits are chosen at random with replacement from the set 0, . . . , 9. What is the probability that their average will lie between 4 and 6?

```
E_X <- 0
for (i in 0:9){
  E_X <- E_X + .1*i
}

E_X2 <- 0
for (i in 0:9){
  E_X2 <- E_X2 + .1*i^2
}

var <- E_X2 -E_X^2

n <- 16

prob <- pnorm(6, E_X, sqrt(var/16)) - pnorm(4, E_X, sqrt(var/16))

cat("The probability that the average of the 16 digits will be between 4 and 6 is",
    round(100*prob, 1),"%.")
```

```
The probability that the average of the 16 digits will be between 4 and 6 is 73.9 %.
```

I've use the first and second moments to generate μ and σ. Then used normal probability calculations.

$E[X] = (\frac{1}{10})(0 + 1 + 2 + ... + 9) = (\frac{1}{10})(\frac{90}{2}) = 4.5 = \mu$

$E[X^2] = (\frac{1}{10})(0^2 + 1^2 + 2^2 + ... + 9^2) = (\frac{1}{10})(1^2 + 2^2 + ... + 9^2) = (\frac{1}{10})(\frac{(9)(9+1)(9+\frac{1}{2})}{3} = 28.5 = \sigma^2 + \mu^2$

So $\sigma^2 = E[X^2] - \mu^2 = \sqrt{28.5 - 4.5^2}$

problem #5

Select a skewed distribution from which to sample. Using R, demonstrate the convergence of the mean value of your samples to the Normal distribution. Assume you are making this demonstration to someone who has little or no statistical training. Your demonstration should take no more than 10 minutes. Along with your code, outline the commentary you would use in your demonstration.

```r
pennydata <- read_excel("pennydata.xls")
ages <- rep(pennydata$Age, pennydata$number)
pop <- as.data.frame(ages)

mu <- round(mean(pop$ages),2)
SD <- round(sd(pop$ages), 2)

gg_pop <- ggplot(data=pop, aes(ages, y=..count../sum(..count..))) +
  geom_histogram(color="white", fill="black", binwidth=1) +
  geom_vline(xintercept=mu, linetype="dashed", color="blue", size=.5) +
  geom_text(
    aes(x=30, label=paste("mu =",mu), y=.24),
    color = "blue") +
   geom_text(
    aes(x=30, label=paste("SD =",SD), y=.20),
    color = "blue") +
  labs(title="Penny popultion (N=752)", x="Ages", y="Proportion") +
  coord_cartesian(ylim=c(0, .30), xlim=c(0,50)) +
  theme(axis.line.y=element_line(color="black", size=.5),
        panel.background = element_blank(),
        panel.grid.major.x = element_line(color="grey", size=0.1),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(color="grey", size=0.1),
        panel.grid.minor.y = element_blank(),
        plot.title=element_text(color="blue")) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))

samples <- data.frame(
  "size_4" = rep(0, 750),
  "size_16" = rep(0, 750),
  "size_64" = rep(0, 750)
  )

for (i in 1:750){
  samples$size_4[i] <- mean(sample(pop$ages, 4, replace=TRUE))
  samples$size_16[i] <- mean(sample(pop$ages, 16, replace=TRUE))
  samples$size_64[i] <- mean(sample(pop$ages, 64, replace=TRUE))
}

x_bar4 <- round(mean(samples$size_4),2)
se4 <- round(sd(samples$size_4), 2)
gg_samples_4 <- ggplot(data=samples, aes(size_4, y=..count../sum(..count..))) +
 geom_histogram(color="white", fill="grey", binwidth=1) +
  geom_vline(xintercept=mu, linetype="dashed", color="darkred", size=.5) +
   geom_text(
    aes(x=29, label=paste("x-bar =",x_bar4), y=.24),
    color = "darkred") +
```

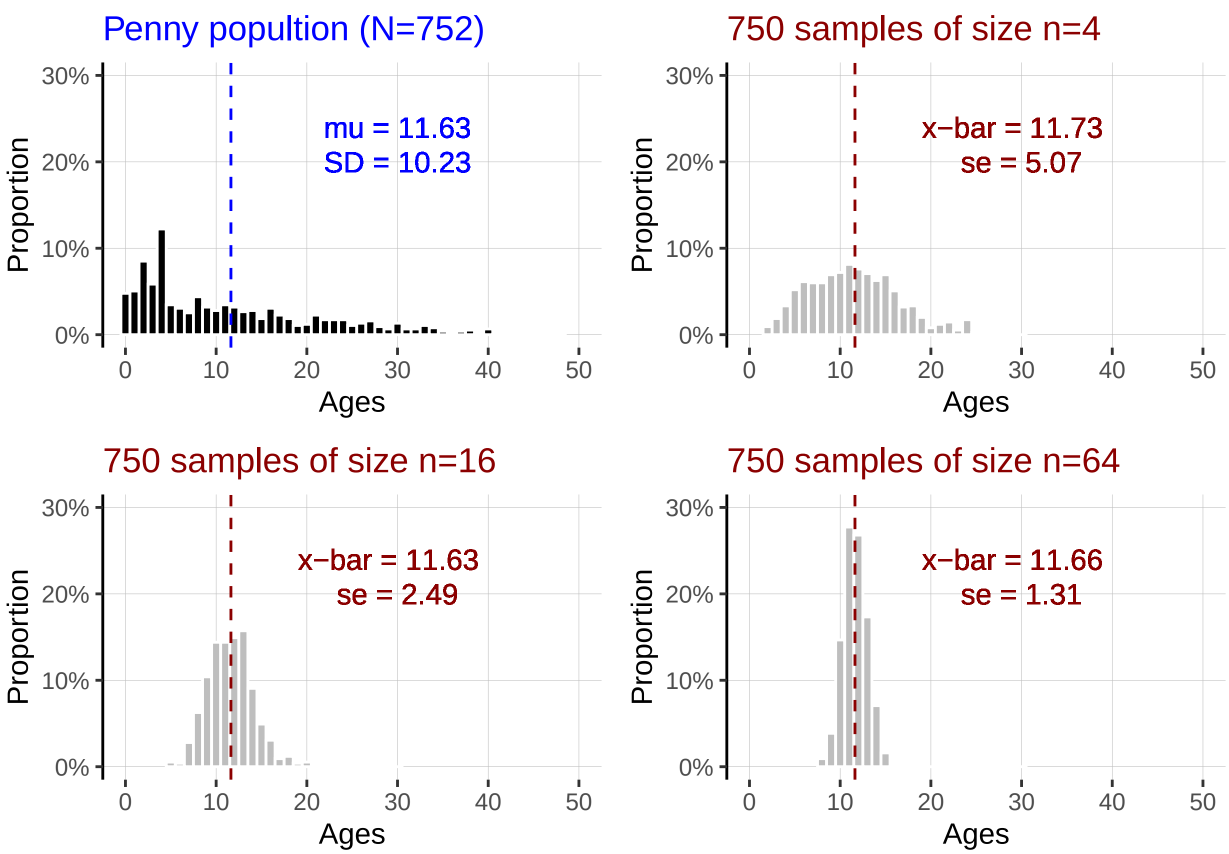
```
    geom_text(
     aes(x=30, label=paste("se =",se4), y=.20),
     color = "darkred") +
  labs(title="750 samples of size n=4", x="Ages", y="Proportion") +
  coord_cartesian(ylim=c(0, .30), xlim=c(0,50)) +
  theme(axis.line.y=element_line(color="black", size=.5),
        panel.background = element_blank(),
        panel.grid.major.x = element_line(color="grey", size=0.1),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(color="grey", size=0.1),
        panel.grid.minor.y = element_blank(),
        plot.title=element_text(color="darkred")) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))


x_bar16 <- round(mean(samples$size_16),2)
se16 <- round(sd(samples$size_16), 2)
gg_samples_16 <- ggplot(data=samples, aes(size_16, y=..count../sum(..count..))) +
 geom_histogram(color="white", fill="grey", binwidth=1) +
  geom_vline(xintercept=mu, linetype="dashed", color="darkred", size=.5) +
   geom_text(
    aes(x=29, label=paste("x-bar =",x_bar16), y=.24),
    color = "darkred") +
   geom_text(
    aes(x=30, label=paste("se =",se16), y=.20),
    color = "darkred") +
  labs(title="750 samples of size n=16", x="Ages", y="Proportion") +
  coord_cartesian(ylim=c(0, .30), xlim=c(0,50)) +
  theme(axis.line.y=element_line(color="black", size=.5),
        panel.background = element_blank(),
        panel.grid.major.x = element_line(color="grey", size=0.1),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(color="grey", size=0.1),
        panel.grid.minor.y = element_blank(),
        plot.title=element_text(color="darkred")) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))

x_bar64 <- round(mean(samples$size_64),2)
se64 <- round(sd(samples$size_64), 2)
gg_samples_64 <- ggplot(data=samples, aes(size_64, y=..count../sum(..count..))) +
 geom_histogram(color="white", fill="grey", binwidth=1) +
  geom_vline(xintercept=mu, linetype="dashed", color="darkred", size=.5) +
   geom_text(
    aes(x=29, label=paste("x-bar =",x_bar64), y=.24),
    color = "darkred") +
   geom_text(
    aes(x=30, label=paste("se =",se64), y=.20),
    color = "darkred") +
  labs(title="750 samples of size n=64", x="Ages", y="Proportion") +
  coord_cartesian(ylim=c(0, .30), xlim=c(0,50)) +
  theme(axis.line.y=element_line(color="black", size=.5),
        panel.background = element_blank(),
        panel.grid.major.x = element_line(color="grey", size=0.1),
```

```r
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(color="grey", size=0.1),
        panel.grid.minor.y = element_blank(),
        plot.title=element_text(color="darkred")) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))

grid.arrange(gg_pop, gg_samples_4, gg_samples_16, gg_samples_64, ncol=2)
```



The sampling distribution is approximately: $N(\mu, \frac{\sigma}{\sqrt{n}})$