Sampling Distribution and Interpretation of Confidence Intervals

Computational Statistics

November 3, 2023

```
library(tidyverse)
```

Sampling distribution of sample proportion

1a. For each pair of the sample sizes $n \in \{10, 25, 50, 100, 250\}$ and population proportions $p \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$, collect 10,000 samples and compute the sample proportions. Plot the histograms of the sample proportions for all pairs of considered sample sizes and population proportions. Summarize the $5 \times 5 = 25$ histograms in **one plot**. Comment on the result.

Solution:

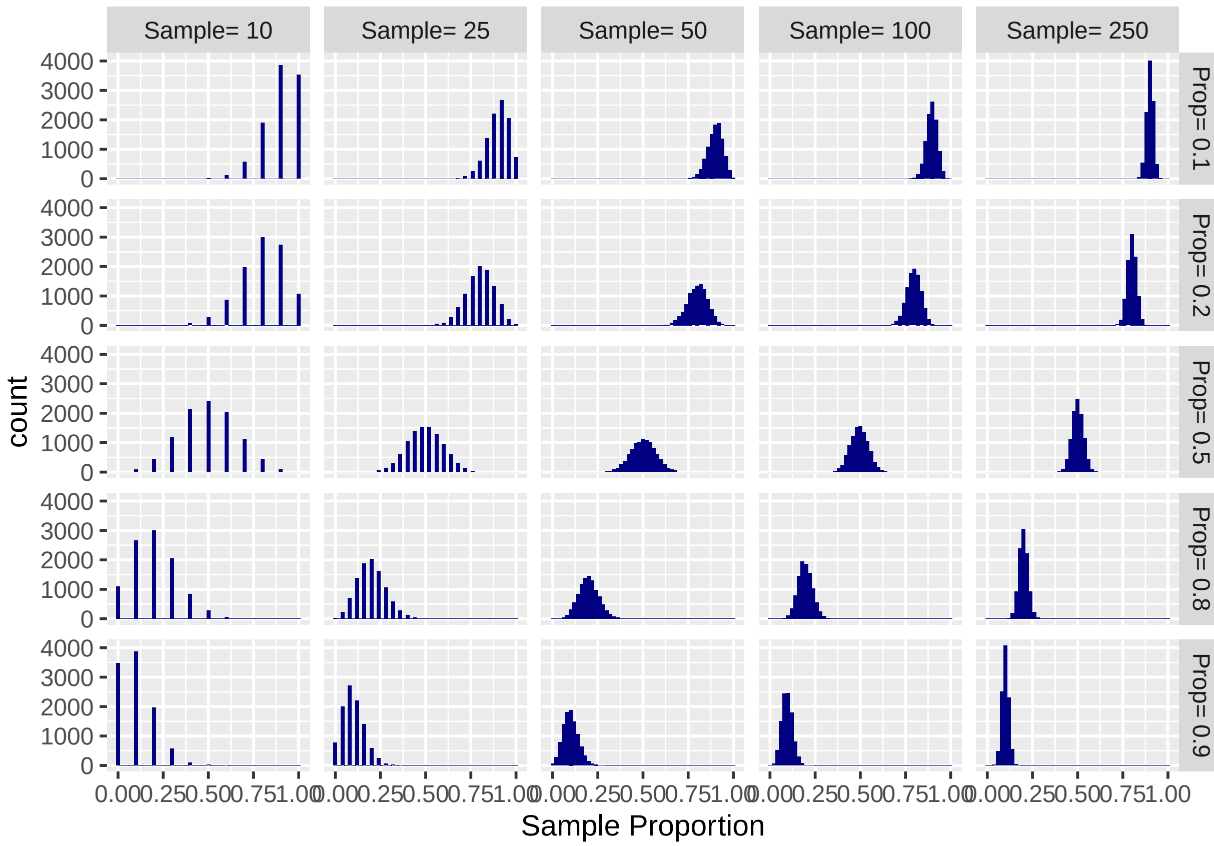
```
library(ggplot2)
library(dplyr)
set.seed(2000)
B <- 10000
p <- c(0.1, 0.2, 0.5, 0.8, 0.9)
size <- c(10, 25, 50, 100, 250)
samples <- vector("list", length = length(p) * length(size))
for (i in 1:length(p)) {
    for (j in 1:length(size)) {
        # Compute intermediate results
        sampled_values <- replicate(B,{s<-sample(c(0,1),size=size[j], replace=TRUE,
                                                 prob=c(p[i],1-p[i]))
          mean(s)})

        samples[[(i - 1) * 5 + j]] <- data.frame(
            p =paste("Prop=", p[i])  ,
            size =paste("Sample=",size[j]) ,
            p_hat = sampled_values
        )
    }
}
```

```r
df_samples <- bind_rows(samples)

df_samples$size <- factor(df_samples$size, levels = paste("Sample=", size))
df_samples$p <- factor(df_samples$p, levels = paste("Prop=", p))

ggplot(df_samples, aes(x = p_hat)) +
    geom_histogram(binwidth = 0.02, fill="navyblue") +
    labs(x = "Sample Proportion") +
    facet_grid(p ~ size)
```




```r
summary(df_samples)
```

```
##          p                     size           p_hat
##  Prop= 0.1:50000   Sample= 10 :50000   Min.   :0.00
##  Prop= 0.2:50000   Sample= 25 :50000   1st Qu.:0.18
##  Prop= 0.5:50000   Sample= 50 :50000   Median :0.50
##  Prop= 0.8:50000   Sample= 100:50000   Mean   :0.50
##  Prop= 0.9:50000   Sample= 250:50000   3rd Qu.:0.82
##                                        Max.   :1.00
```

Inference:

From the graph of Sample Population to the Population Proportion, we can infer the following:

(1) Shapes of all the graphs represent a general "bell curve" pattern indicating that the data is normally distributed.

(2) As the sample size increases, there is less variability in the sample values/data. Which means, the data approaches mean value as the sample size increases.

This justifies the Central Limit Theorem: "If the sample size is large enough, the sampling distribution of the sample mean will be approximately normal, regardless of the shape of the original population distribution."

Therefore, Large sample size resembles the shape of the population distribution.

(3) When Prop is close to 0.5, there is less variability and as the Population proportion deviates from the mid value, there is corresponding deviation from the central value.

(4) The distribution of the data in graphs with sample size 50 is consistent indicating the high proximity to normal curve.

Sampling distribution of sample mean

Consider the following three population distributions: 1) a Normal distribution with mean 4 and standard deviation $\sqrt{8}$, 2) a Uniform distribution between $4 - 2\sqrt{6}$ and $4 + 2\sqrt{6}$, and 3) a Gamma distribution with shape $k = 2$ and scale $\theta = 2$. The parameters are chosen such that the means and standard deviations of the three distributions are the same.

2a. From each of the three distributions, calculate the sample means for 10,000 independent samples, with sample size $n = 2$, 5, and 20. Plot the histograms of the sample means for all pairs of considered sample size ($n = 2$, 5, or 20) and population distribution (Normal, Uniform, or Gamma). Summarize the 9 histograms in **one plot** (similar to the one in Lecture 13, Page 15). Comment on the result.

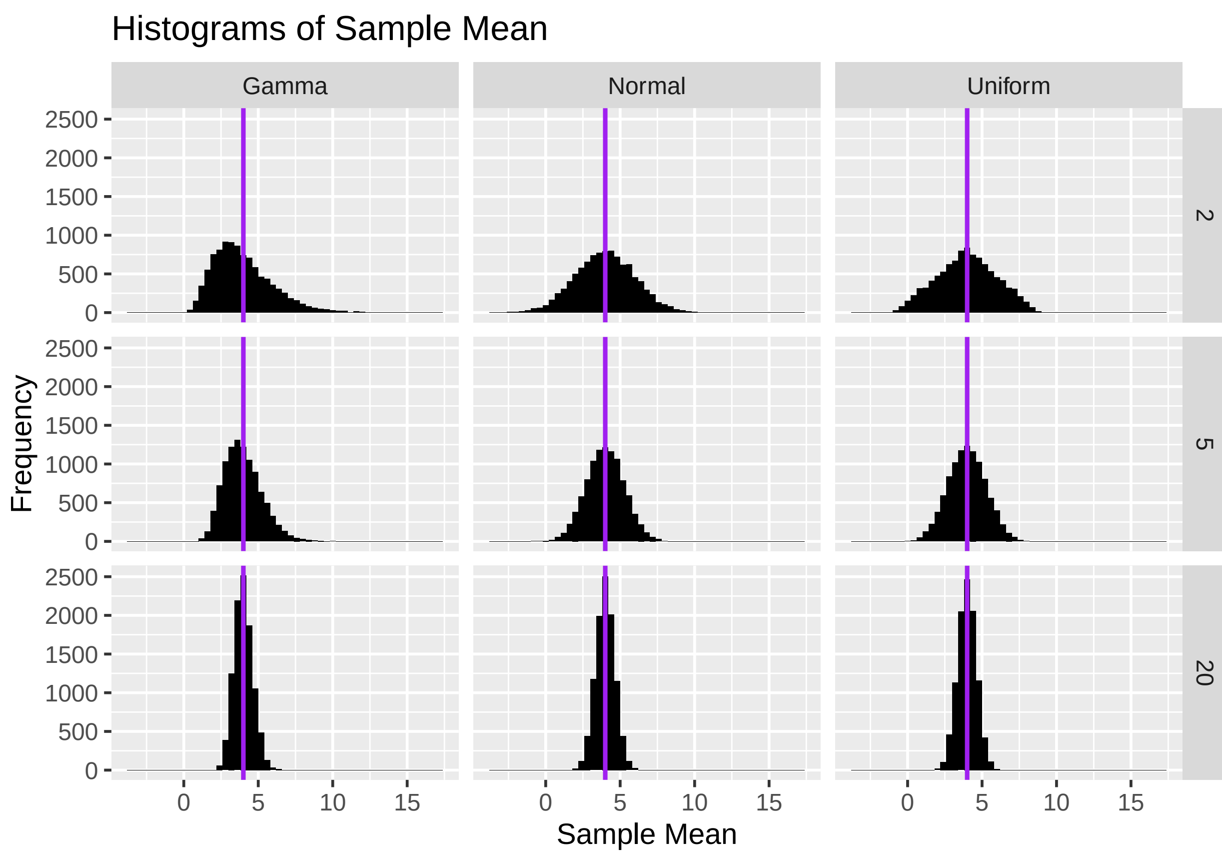
Solution:

```
library(dplyr)
library(tidyverse)
B <- 10000
ns <- c(2, 5, 20)
smean_list <- vector("list", 3)
for (i in seq_along(ns)) {
    # Compute intermediate results
    sample_mean_normal<-replicate(B,{
      mean(rnorm(ns[i],mean=4, sd=sqrt(8)))})
```

```r
    sample_mean_uni<-replicate(B,{
      mean(runif(ns[i],min=4-(2*sqrt(6)), max=4+(2*sqrt(6))))})
    sample_mean_gamma<-replicate(B,{
      mean(rgamma(ns[i],shape=2, scale=2))})
    smean_list[[i]] <- data.frame(
        smean = c(sample_mean_normal,sample_mean_uni,sample_mean_gamma)  ,
        type =rep(c("Normal","Uniform","Gamma"), each=10000)  ,
        n =  ns[i]
    )
}
smean_all <- bind_rows(smean_list)

smean_all |>
    ggplot(aes(x = smean)) +
    geom_histogram(binwidth = 0.4, position="identity",fill="black") +
    geom_vline(xintercept = 4, linetype = "solid", color = "purple" ,size=0.8) +
    labs(title="Histograms of Sample Mean",x = "Sample Mean", y="Frequency") +
    facet_grid(n ~ type)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```r
summary(smean_all)
```

```
##      smean             type                 n
##  Min.   :-3.54   Length:90000       Min.   : 2
##  1st Qu.: 3.20   Class :character   1st Qu.: 2
##  Median : 3.96   Mode  :character   Median : 5
##  Mean   : 3.99                      Mean   : 9
##  3rd Qu.: 4.72                      3rd Qu.:20
##  Max.   :17.25                      Max.   :20
```

Inference:

(1) From the graph it is evident that as the sample size increases, Gamma and Uniform distribution are close to the Normal Distribution and the values are centered around the true mean *(indicated by the purple line)* . This suggests that the sample means are unbiased estimators fo the population mean.

(2) With a shorter sample size, the data in Gamma distribution is right-skewed.

(3) Irrespective of the sample size, the sampling distribution of the mean is approximately normal in normally distributed data.

(4) Lower sample size, indicates more variability therefore the graph is spread horizontally in contrast to large sample size where the graph is peaked.

Interpretation of confidence intervals

Consider drawing samples of size 100 for random variable $X \in \{0, 1\}$ and $\Pr(X = 1) = p = 0.45$:

```
set.seed(1997)
p <- 0.45
N <- 100
x <- sample(c(0, 1), size = N, replace = TRUE, prob = c(1 - p, p))
```

Confidence interval on the mean from one sample is given by:

```
x_bar <- mean(x)  # sample mean
se_hat <- sqrt(x_bar * (1 - x_bar) / N)  # standard error
c(x_bar - 1.96 * se_hat, x_bar + 1.96 * se_hat)
```

```
## [1] 0.266 0.454
```

3a. Take 10,000 samples, construct 10,000 95% confidence intervals based on the samples, and compute the proportion of the times those intervals contain the true parameter $p = 0.45$.

Solution:

```
result<-replicate(10000,{
  x <- sample(c(0, 1), size = N, replace = TRUE, prob = c(1 - p, p))
x_bar <- mean(x)  # sample mean
se_hat <- sqrt(x_bar * (1 - x_bar) / N)  # standard error
ci_lb<-x_bar - 1.96 * se_hat
ci_ub<- x_bar + 1.96 * se_hat
(p>ci_lb) && (p<ci_ub)
}
)

mean_3a<-mean(result)
```

The proportion of times the confidence intervals contains the true parameter p=0.45 is 0.945.

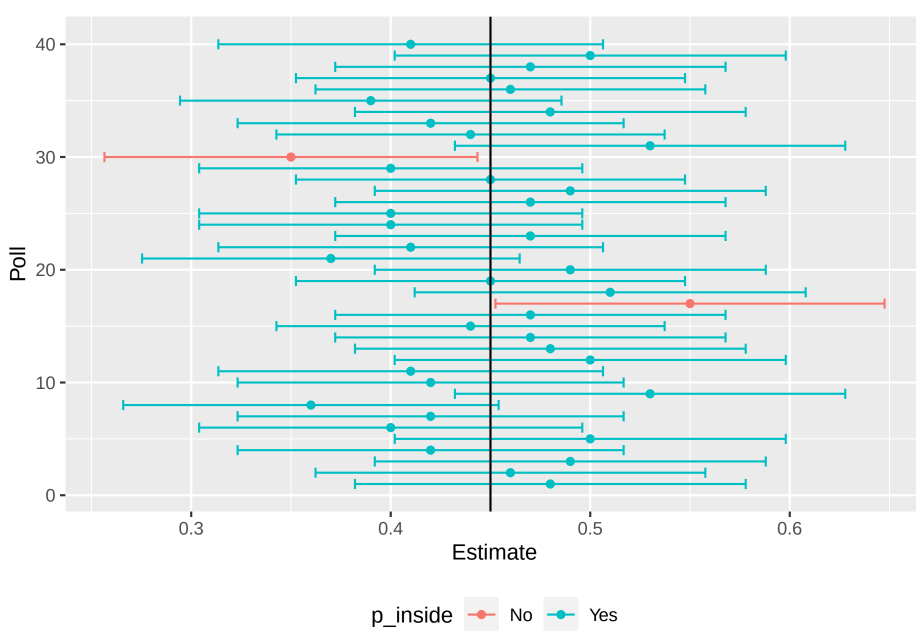
3b. Construct confidence intervals from 40 samples of size 100, and create a plot similar to that in Lecture 13, Page 28 to summarize the result.

Solution:

Confidence Intervals from 40 samples of size 100 each constructed on probabitlity of success as 0.45 and probability of failure as 0.55.

```r
set.seed(1200)
tab <- replicate(40, {
    x2 <- sample(c(0, 1), size = N, replace = TRUE, prob = c(1 - p, p))
    x_bar2 <- mean(x2)
    se_hat2 <- sqrt(x_bar2 * (1 - x_bar2) / N)
    ci_low<-x_bar2 - 1.96 * se_hat2
    ci_hi<- x_bar2 + 1.96 * se_hat2
    hit <- (p>ci_low) && (p<ci_hi)
    c(x_bar2, ci_low, ci_hi, hit)
})

tab_2 <- data.frame(poll = 1:ncol(tab), t(tab))
#t() denotes the transpose of tab
names(tab_2) <- c("Poll", "Estimate", "low", "high", "hit")
tab_3 <- mutate(tab_2, p_inside = ifelse(hit, "Yes", "No") )
ggplot(tab_3, aes(Poll, Estimate, ymin = low, ymax = high, col = p_inside)) +
    geom_point() +
    geom_errorbar() +
    coord_flip() +
    geom_hline(yintercept = p) +
    theme(legend.position = "bottom")
```




```r
summary(tab_3)
```

```
##       Poll          Estimate          low              high             hit
##  Min.   : 1.0   Min.   :0.350   Min.   :0.257   Min.   :0.443   Min.   :0.00
```

```
##  1st Qu.:10.8   1st Qu.:0.410   1st Qu.:0.314   1st Qu.:0.506   1st Qu.:1.00
##  Median :20.5   Median :0.455   Median :0.357   Median :0.553   Median :1.00
##  Mean   :20.5   Mean   :0.450   Mean   :0.353   Mean   :0.547   Mean   :0.95
##  3rd Qu.:30.2   3rd Qu.:0.482   3rd Qu.:0.385   3rd Qu.:0.580   3rd Qu.:1.00
##  Max.   :40.0   Max.   :0.550   Max.   :0.452   Max.   :0.648   Max.   :1.00
##    p_inside
##  Length:40
##  Class :character
##  Mode  :character
##
##
##
```

The above graph represents, Means and 95% CIs of 40 samples (N = 100) drawn from a normal population with mean m and s.d.

Intervals (in red) do not capture the mean within the intervals.