Experiments, Central Limit Theorem, Confindence Intervals

Your Name and id

Contents

Simulating Experiments

In this problem we will simulate two important experiments: coin tosses and die rolls.

We will explore empirical probability of events using random samples. The two important functions to understand are "sample" and "replicate".

We use the sample function to set up our experiment, and replicate function to repeat it as many times as we want.

Die Rolls

Setup the experiment where we roll two fair six sided die independently, and calculate the sum of the numbers that appear for each of them.

Use the replicate function to repeat this experiment 10000 times, store the output of the experiment in a variable "sum_die".

Use "sum_die" to calculate the empirical probability of getting the sums 2, 3, 4, ..., 12. Compare your answers with the theoretical values (you calculated these in HW2).

Plot the following: relative frequency of getting sum equal to 4 as a function of number of trials, for this experiment.

Balls and Cells

If n indistinguishable balls are placed randomly into n cells, the probability that exactly one cell remains empty is calculated to be

$$\binom{n}{2}\frac{n!}{n^n}.$$

\

Write a function "exactly_one_cell'' takes input n, and performs one run of this experiment. Hint: the"unique" function applied to a vector gives the unique entries in a vector. You can sample a set of size n (with replacement) from $1, 2, 3, 4, 5, ..., n$, and check if the number of unique entries in this set is $(n-1)$.

```
#define the function 'exactly_one_cell'sds
#takes input number of cells,
#output: true if exactly one cell is empty, else false.
```

```
n <- 3
#balls_to_cell_exp <- replicate(100000, exactly_one_cell(n))
#mean(balls_to_cell_exp)
theoretical_value <- choose(n, 2)*factorial(n)/(n^n)
theoretical_value
```

```
## [1] 0.6666667
```

Now use the replicate function to calculate the empirical probabilities when for each of the following values of $n : 3, 6, 9, 15$ (note any problems that you might run into). Verify that your empirical probabilities indeed line up with what we expect from our theoretical calculations.

Central Limit Theorem

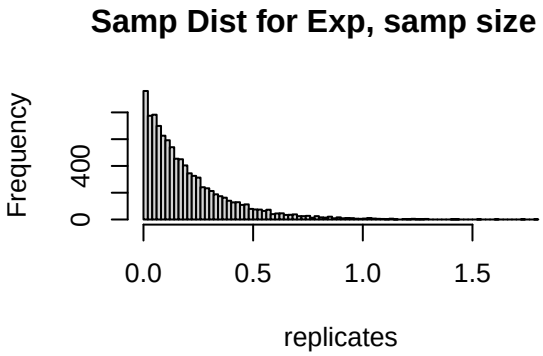
Exponential Distribution

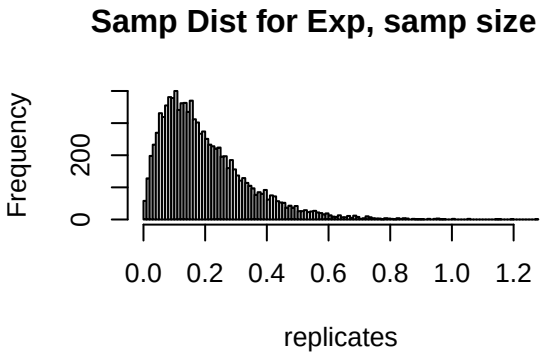
Suppose we are working with a population that has the exponential distibution with $\lambda = 2$.

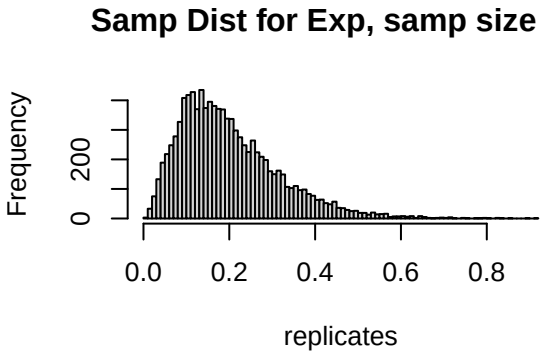
Use the replicate function to get the histograms for the sampling distribution of the sample mean when working with sample sizes $n = 1, 2, 3, 4, 15, 500$. Be sure to have appropriate titles for your histograms.

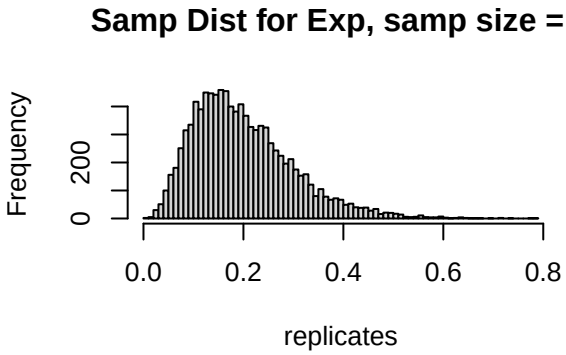
```
  samp_sizes  <- c(1, 2, 3, 4, 15, 500, 1000)

  par(mfrow=c(2,2))
  for(size in samp_sizes){
    replicates <- replicate(10000, {
      mean(rexp(size, rate = 5))
    })
```

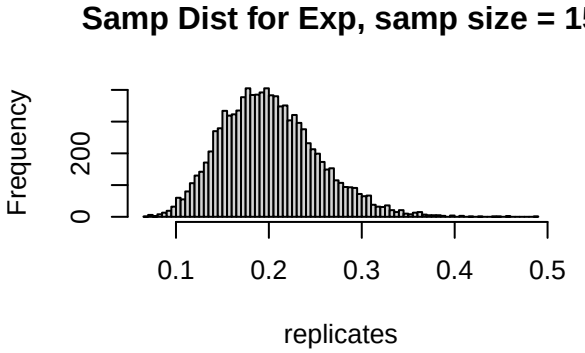
```
    hist(replicates, breaks = 100,
         main = paste("Samp Dist for Exp, samp size =", size ))
}
```

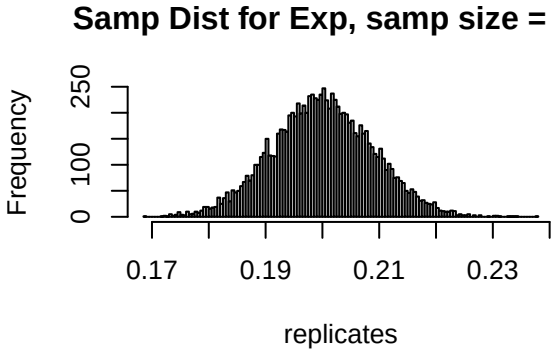


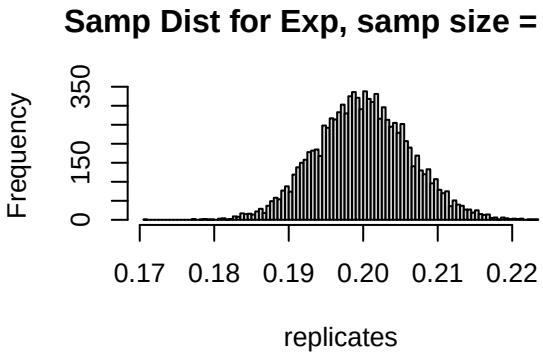












What do you notice?

Discrete Uniform distibution

Suppose we are working with the discrete uniform random variable taking values $\{1, 2, 3, 4, 5, 6\}$.

Define a function "disc_samp" that takes input "n" and returns a random sample of size "n" from this distribution.

Use the "disc_samp" function and the replicate function to to get the histograms for the sampling distribution of the sample mean when working with sample sizes $n = 1, 2, 3, 4, 15, 500$. Be sure to have appropriate titles for your histograms.

What do you notice?

Continuous Uniform distibution

Suppose we are working with the Continuous uniform random variable taking values on $(0, 1)$.

Define a function "cont_uni_samp" that takes input "n" and returns a random sample of size "n" from this distribution.

Use the "cont_uni_samp" function and the replicate function to to get the histograms for the sampling distribution of the sample mean when working with sample sizes $n = 1, 2, 3, 4, 15, 500$. Be sure to have appropriate titles for your histograms.

What do you notice?

Bias, Variance, and MSE for estimator for μ^2

Suppose we have a random sample of size n coming from the normal distribution $N(\mu = 5, \sigma = 1.5)$. Recall from class that $\hat{\theta} = \overline{X}^2$ is a biased estimator for μ^2, and we calculated the bias to be

$$\text{Bias}(\hat{\theta}) = \frac{\sigma^2}{n}.$$

Write function that takes input a random sample and outputs the square of the sample mean (this is the estimator $\hat{\theta}$.

Use the replicate function to calculate the empirical variance and bias of the estimator $\hat{\theta}$ when sampling a sample of size 15 from $N(\mu = 5, \sigma = 1.5)$. Store these numbers in variables "var_emp" and "bias_emp".

Use the replicate function to calculate the Mean Squared Error for $\hat{\theta}$ when sampling a sample of size 15 from $N(\mu = 5, \sigma = 1.5)$. Store this number in "MSE_emp".

If there is justice in this world, you must get

$$\text{mse}_{\text{emp}} \approx \text{var}_{\text{emp}} + \text{bias}_{\text{emp}}.$$

Check if this is true for the estimator $\hat{\theta}$.

Estimator for μ from two samples

Recall from lecture that we constructed an unbiased estimator for μ as a weighted average of sample means coming from two independent random samples of sizes m and n coming from populations with mean μ and variances σ^2 and $k * \sigma^2$ for some choice of k.

$$\hat{\theta} = \delta\overline{X} + (1-\delta)\overline{Y}$$

We calculated the variance of this estimator as a function of σ^2, m, n. We will run a simulation to verify this.

In the following code, we define a function that takes input δ, and run an experiment where: two samples are taken from $N(2,9)$ and $N(2,36)$ respectively, and returning the value of the estimator $\hat{\mu}$.

```
mean_estimate_two_samps <- function(delta){
  samp1 <- rnorm(10, mean = 2, sd = 3)
samp2 <- rnorm(20, mean = 2, sd = 6)
  return(delta*mean(samp1) + (1-delta)*mean(samp2))
}
```

Here we run a simulation to collect 100000 different estimates for μ using the estimator $\hat{\mu}$ with $\delta = \frac{3}{4}$.

```
B <- 100000
means <- replicate(B,mean_estimate_two_samps(3/4))
```

The mean and the variance of the estimates calculated above is the emprical expected value and the empirical variance of the estimator $\hat{\mu}$.

```
mean(means)
```

```
## [1] 1.999021
```

```
var(means)
```

```
## [1] 0.6213265
```

You can verify that the empirical variance is pretty close to the theoretical variance.

```
delta <- 3/4

var <- delta^2*(3^2)/10 + (1-delta)^2*(8^2)/20
var
```

```
## [1] 0.70625
```

Run the same experiment as above with the value of δ that minimizes the variance of $\hat{\mu}$.

Visualizing Distributions

Densities of Standard normal and Students' t-distribution

Construct plot of the pdf of the standard normal distribution, and the plot of the pdf of the Students' t-distribution with df=2, 3, 4, 6, 9, 30. All the graphs must be in a single plot, each graph should be of a different color.

For the plot above, for each pdf of a given color, draw a vertical line of the same color at the 0.05 critical value for that pdf.

What do you notice?

Densities of Chi-Squared Distribution

Construct a single plot with the graphs of the pdf of the Chi-squared distribution with df=2, 3, 4, 6, 9, 30, 40. Each graph must have a different color.

What do you notice?