Homework 1

Joshua Ingram

2/17/2021

Question 1 - Two Estimators for σ

For question 1, we will explore the performances of the two estimators s and s_{MLE} by using simulations in R. For a sample $x_1, x_2, \cdot, x_n$, the standard deviation s is given by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad n > 1. \tag{1}$$

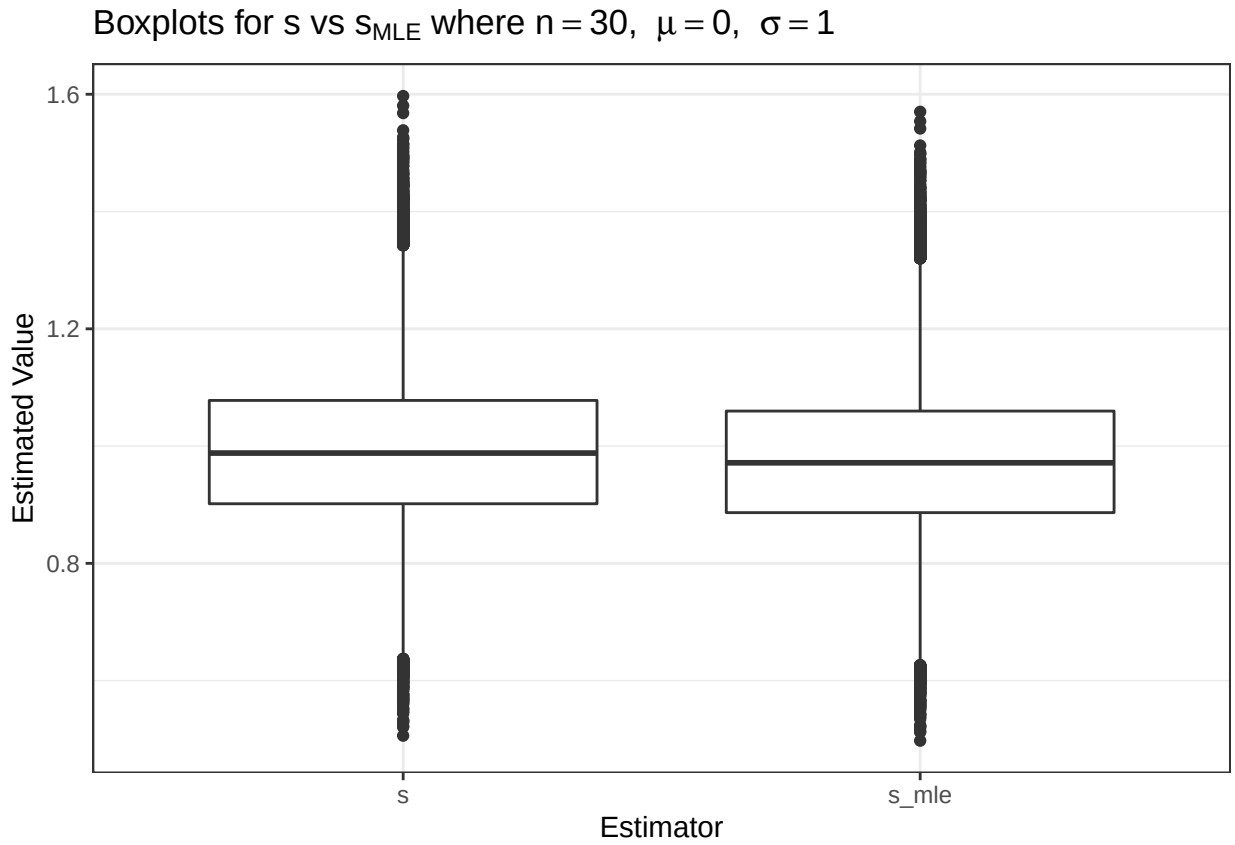
Alternatively, the Maximum Likelihood Estimator σ is given by

$$s_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad n \geq 1. \tag{2}$$

We use the `rnorm` function to create the simulated data, with parameters to determine sample size, μ, and σ. To better understand these estimators and make comprehensive comparisons, we will vary the simulated data by sample size, μ, and σ. We first create a function to compute s_{MLE}.

```
# built-in sd function uses the formula given for s
# Creating function for s_mle
sd_mle <- function(x){
  sqrt(sum((x - mean(x))^2)/length(x))
}
```

Now we compare the estimators for a sample size of 30 when drawing from the standard normal distribution.



Before delving into more simulations, we observe in the side-by-side boxplots that the estimate given by s is closer to the true value $\sigma = 1$ than is s_{MLE}. However, the spread of the estimates given by s_{MLE} is slightly less than that of s. See the following table for summaries.

Table 1: Simulation Results - n = 30, mu = 0, sigma = 1

	mean	sd
s	0.9913094	0.1308707
s_mle	0.9746475	0.1286711

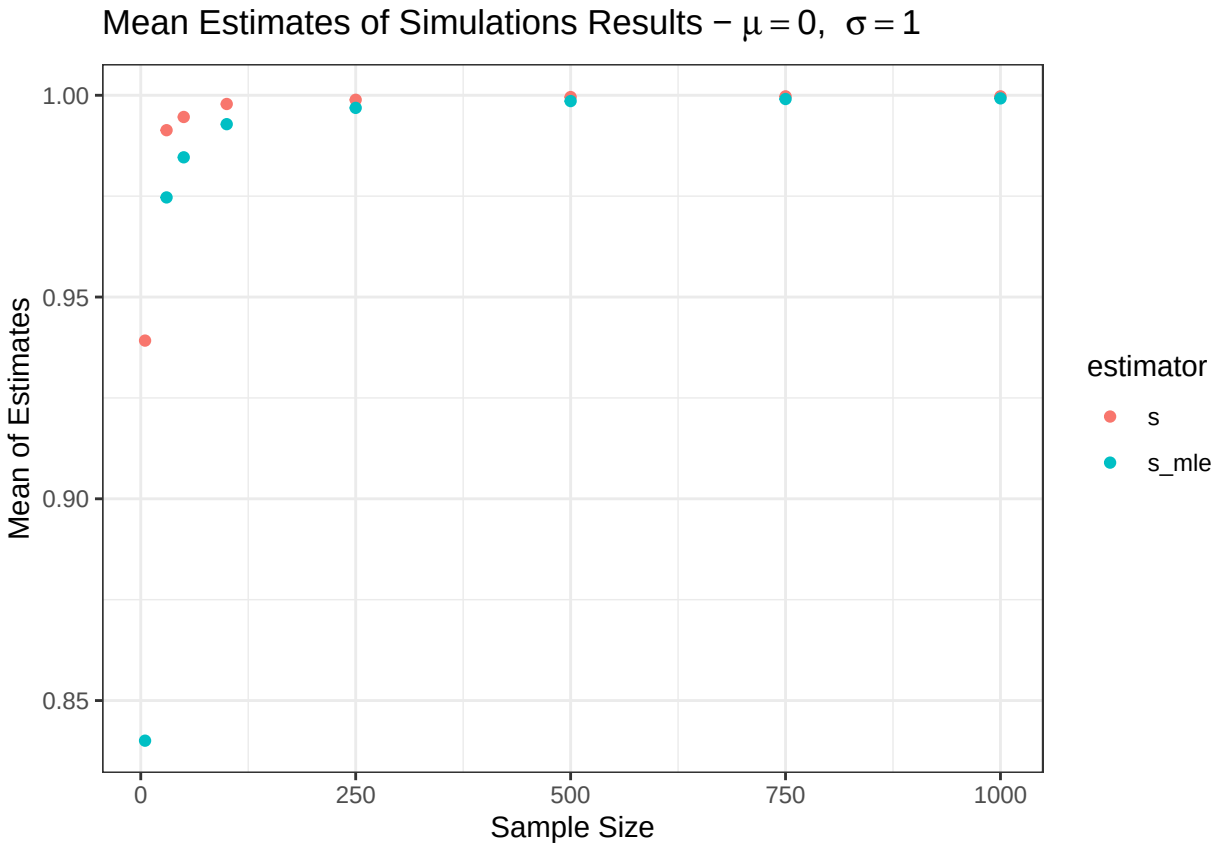
After the initial exploration, we will now vary the sample sizes, as well as the mean and standard deviation of the distribution we draw from to see how our estimates perform. We create a function to get the estimates for the simulated datasets.

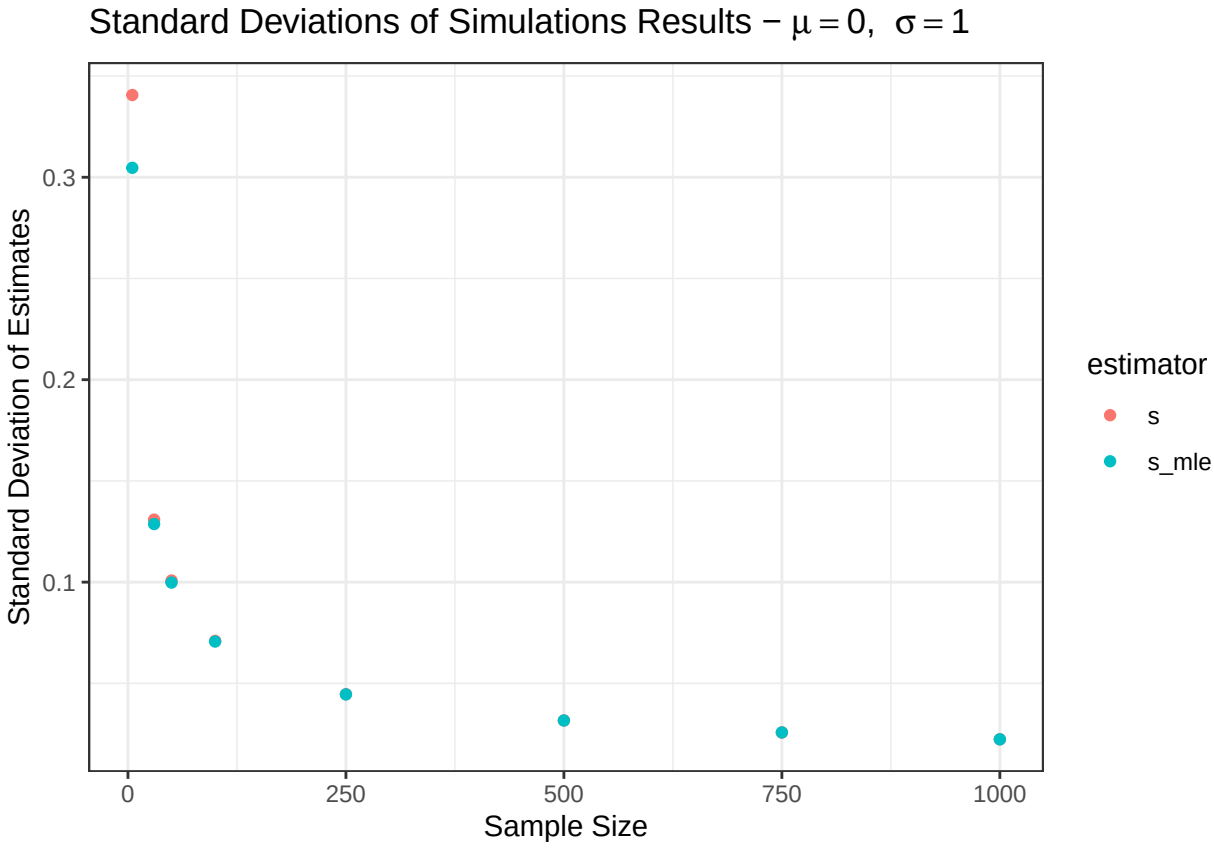
```
sim_sd <- function(sim_count = 100000, s, mu, sigma){

  sim_data <- lapply(1:sim_count, function(x) rnorm(s, mu, sigma))
  data_sd <- unlist(lapply(sim_data, function(x) sd(x)))
  data_sdm <- unlist(lapply(sim_data, function(x) sd_mle(x)))

  df_results <- data.frame(data_sd, data_sdm)
  colnames(df_results) <- c("s", "s_mle")

  return(df_results)
}
```

Simulations Results - $\mu = 0$ and $\sigma = 1$

Table 2: Simulation Results - mu = 0, sigma = 1

estimator	n	mean	sd
s	5	0.9392096	0.3406318
s_mle	5	0.8400546	0.3046704
s	30	0.9913511	0.1308808
s_mle	30	0.9746886	0.1286810
s	50	0.9946263	0.1007899
s_mle	50	0.9846298	0.0997770
s	100	0.9978439	0.0709992
s_mle	100	0.9928421	0.0706433
s	250	0.9988816	0.0446364
s_mle	250	0.9968818	0.0445471
s	500	0.9995624	0.0316783
s_mle	500	0.9985623	0.0316466
s	750	0.9997384	0.0257645
s_mle	750	0.9990716	0.0257474
s	1000	0.9997344	0.0223562
s_mle	1000	0.9992344	0.0223450

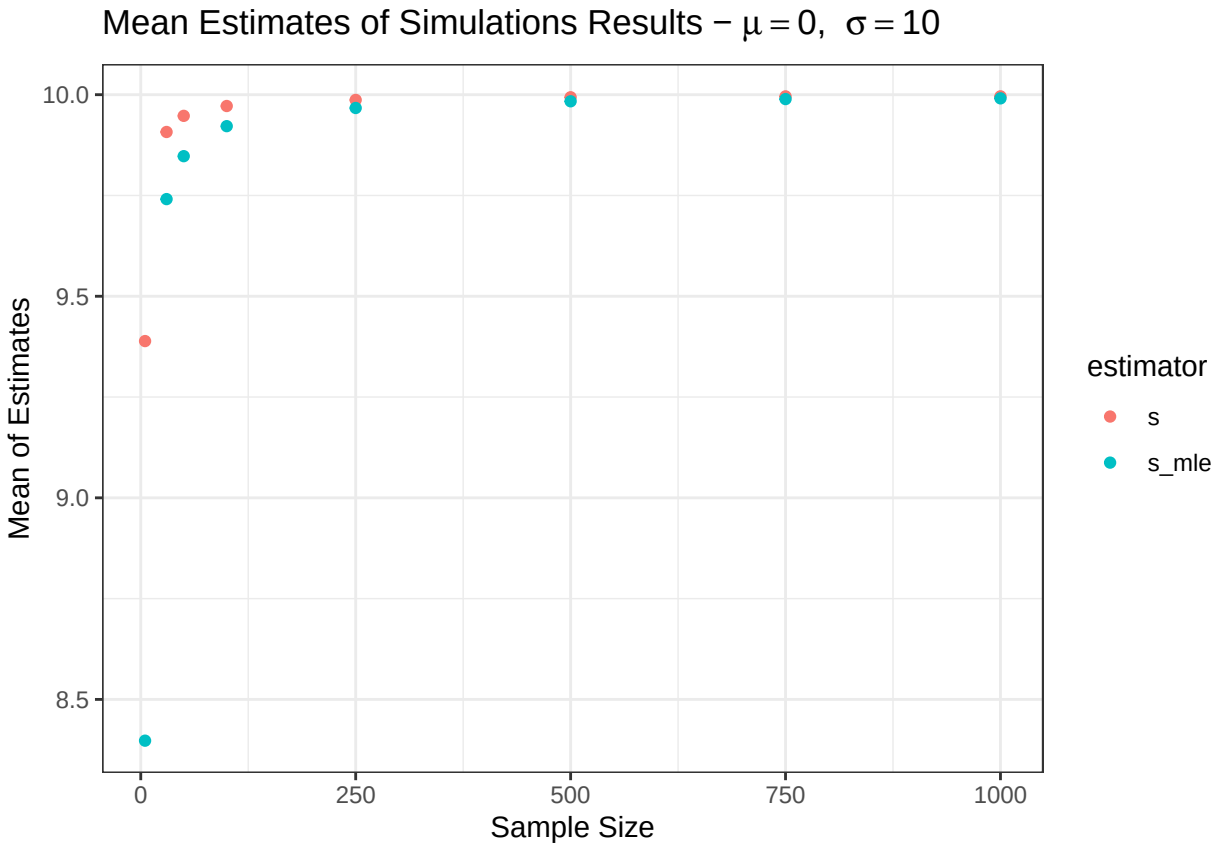

Mean Estimates of Simulations Results − μ = 0, σ = 1
Mean of Estimates
1.00
0.95
0.90
0.85
0
250
500
750
1000
Sample Size
estimator
s
s_mle

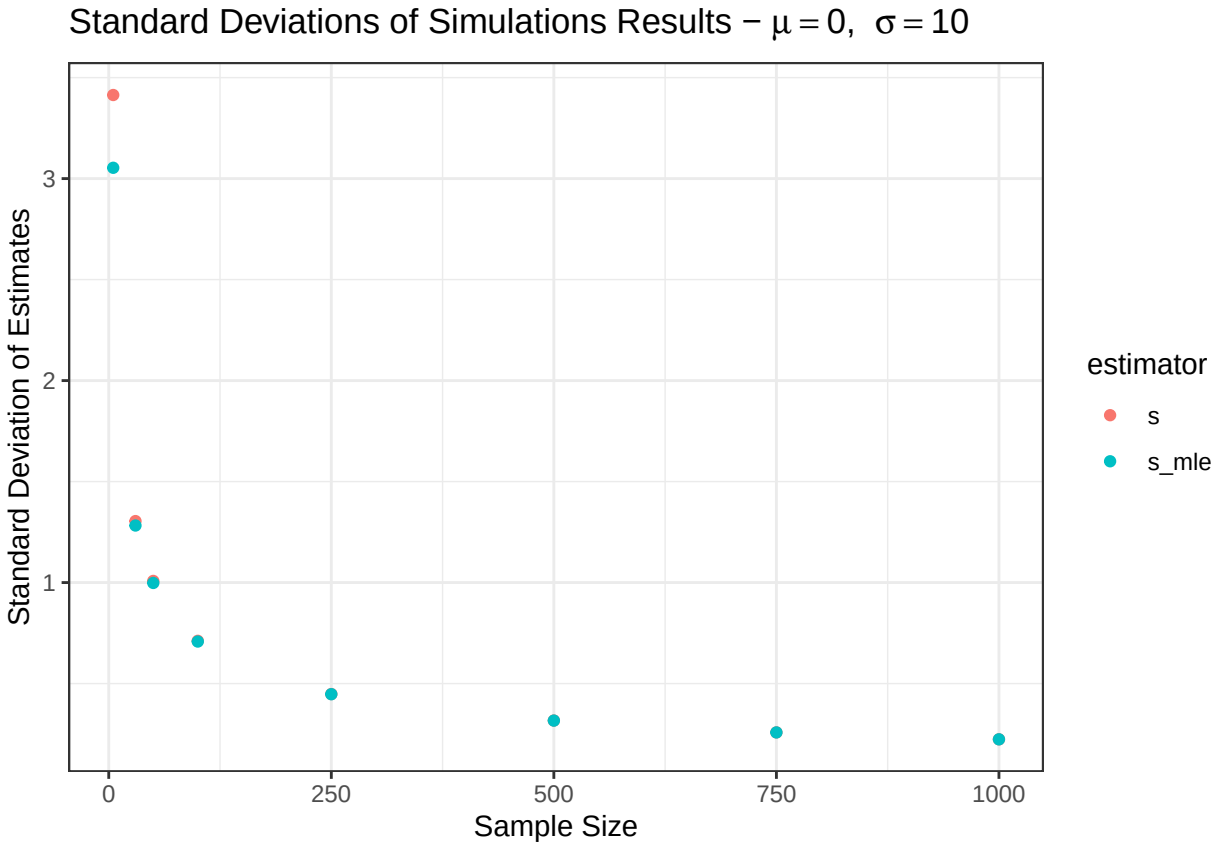

Standard Deviations of Simulations Results − μ = 0, σ = 1
Standard Deviation of Estimates
0.3
0.2
0.1
0
250
500
750
1000
Sample Size
estimator
s
s_mle

Simulations Results - $\mu = 0$ and $\sigma = 10$

Table 3: Simulation Results - mu = 0, sigma = 10

estimator	n	mean	sd
s	5	9.388796	3.4139534
s_mle	5	8.397594	3.0535327
s	30	9.907478	1.3043087
s_mle	30	9.740954	1.2823860
s	50	9.947482	1.0081245
s_mle	50	9.847505	0.9979924
s	100	9.971906	0.7115955
s_mle	100	9.921921	0.7080285
s	250	9.986967	0.4480673
s_mle	250	9.966973	0.4471703
s	500	9.993609	0.3168675
s_mle	500	9.983611	0.3165505
s	750	9.995750	0.2581318
s_mle	750	9.989084	0.2579597
s	1000	9.995871	0.2239063
s_mle	1000	9.990872	0.2237943

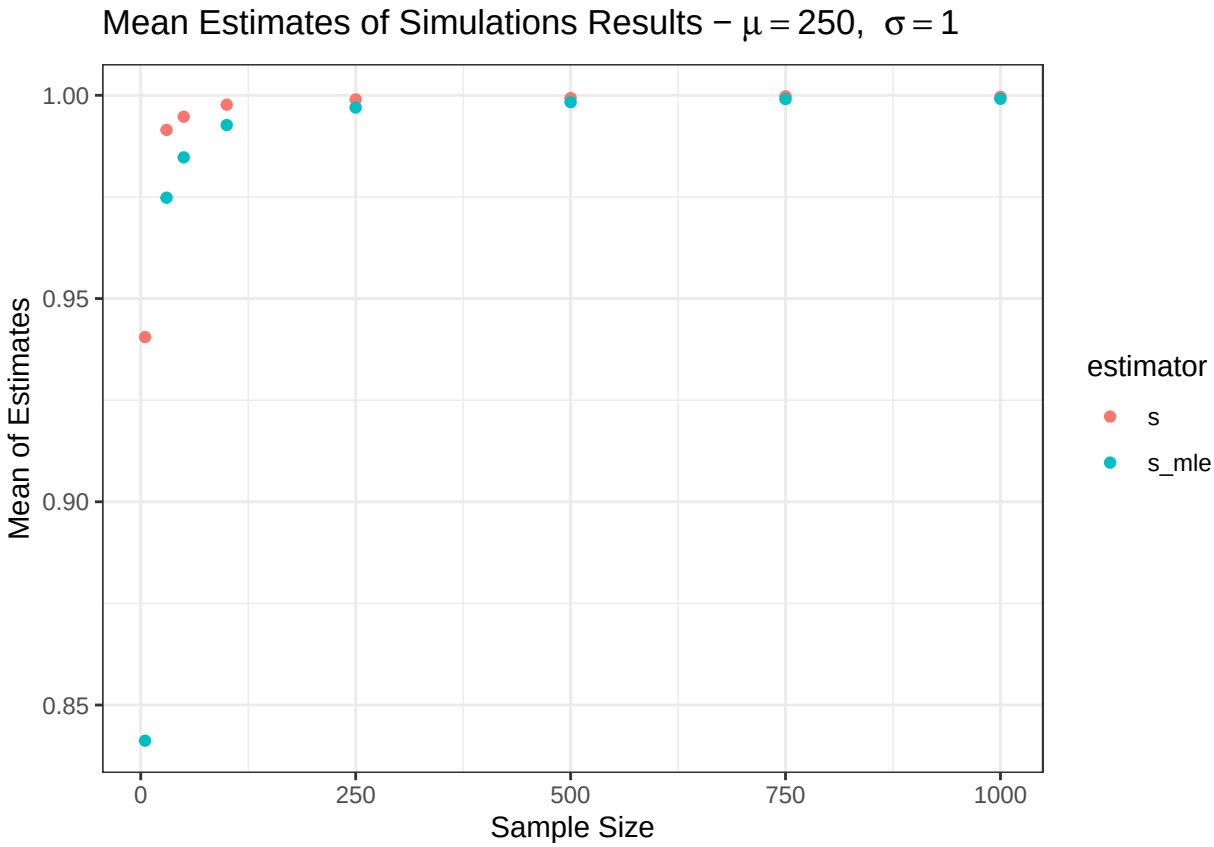

Mean Estimates of Simulations Results − μ = 0, σ = 10
Mean of Estimates
10.0
9.5
9.0
8.5
0
250
500
750
1000
Sample Size
estimator
s
s_mle

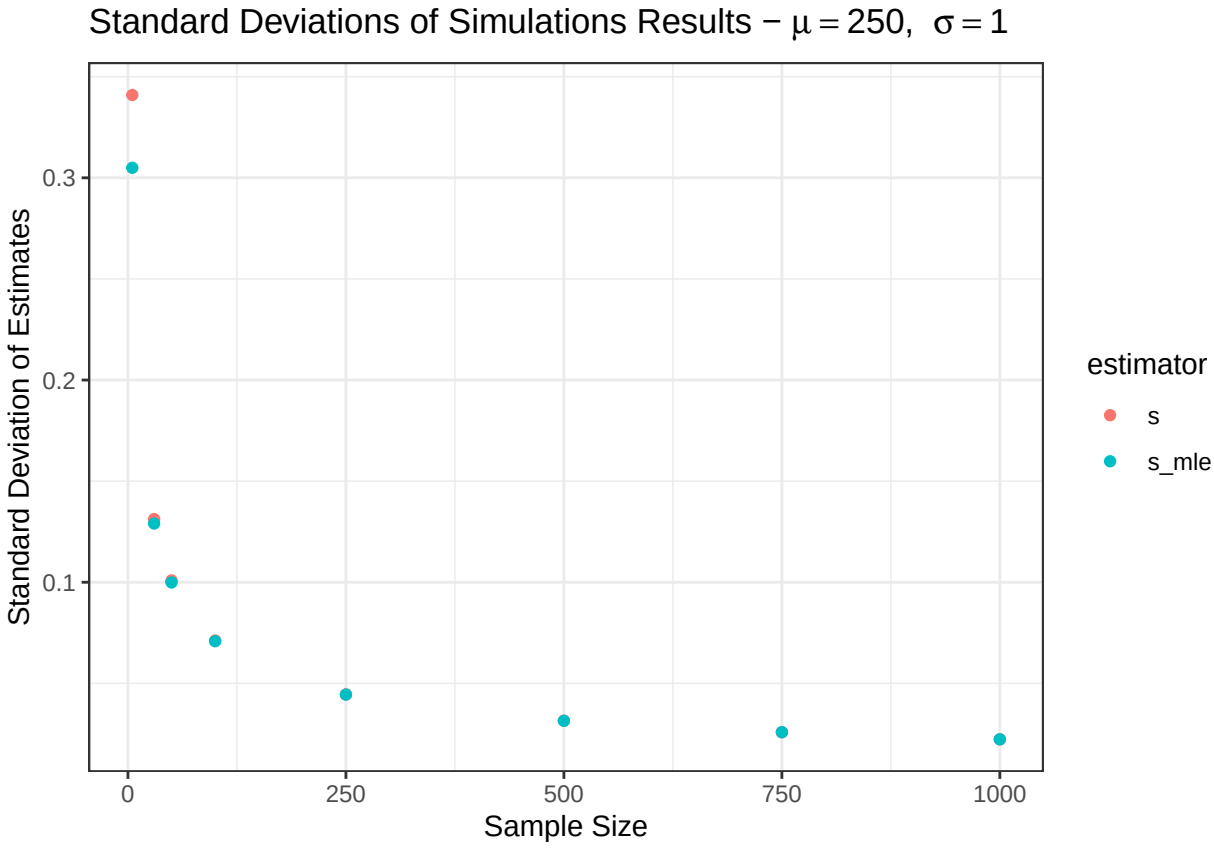

Standard Deviations of Simulations Results − μ = 0, σ = 10
Standard Deviation of Estimates
3
2
1
0
250
500
750
1000
Sample Size
estimator
s
s_mle

Simulations Results - $\mu = 250$ and $\sigma = 1$

Table 4: Simulation Results - mu = 250, sigma = 1

estimator	n	mean	sd
s	5	0.9405234	0.3409381
s_mle	5	0.8412297	0.3049443
s	30	0.9914756	0.1312558
s_mle	30	0.9748109	0.1290497
s	50	0.9947315	0.1009007
s_mle	50	0.9847339	0.0998866
s	100	0.9976871	0.0712028
s_mle	100	0.9926861	0.0708459
s	250	0.9989830	0.0445345
s_mle	250	0.9969831	0.0444453
s	500	0.9993202	0.0315197
s_mle	500	0.9983204	0.0314882
s	750	0.9997461	0.0258722
s_mle	750	0.9990794	0.0258550
s	1000	0.9996243	0.0223406
s_mle	1000	0.9991243	0.0223295

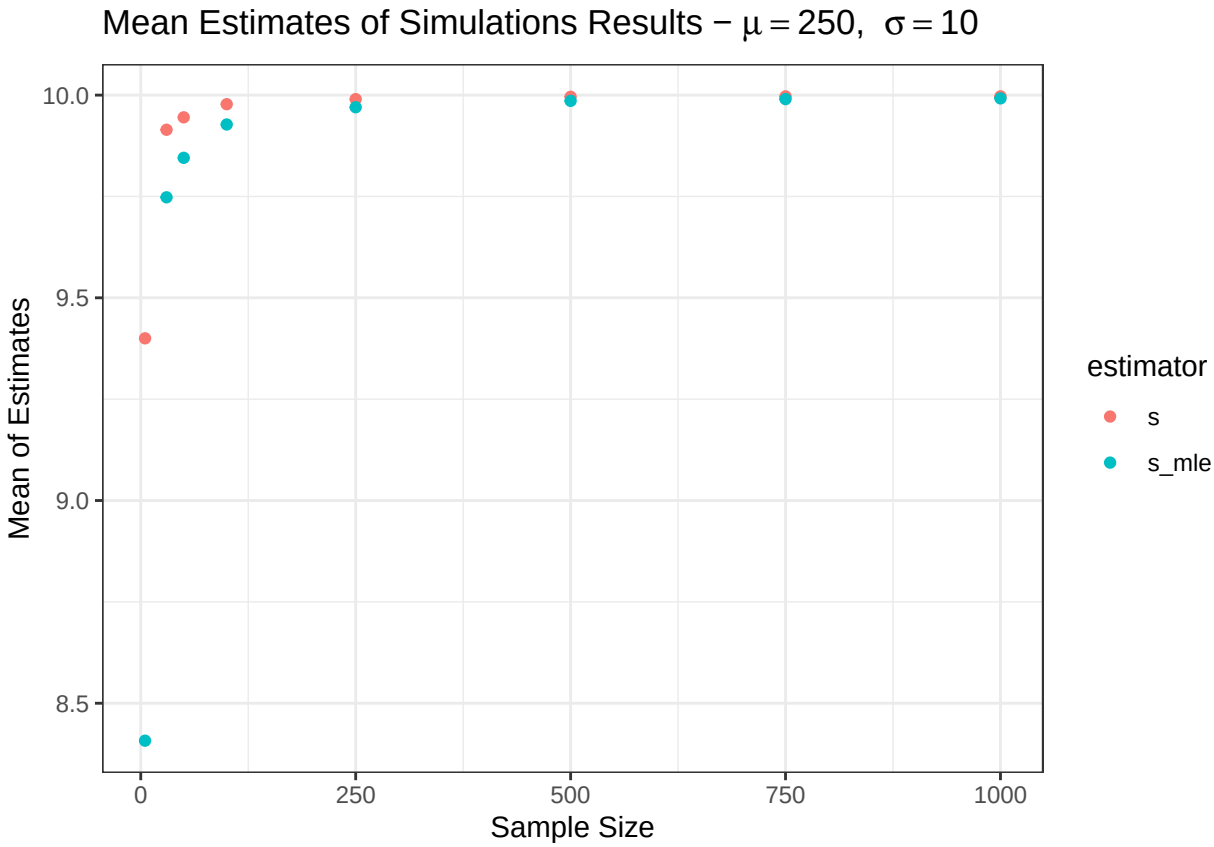

Mean Estimates of Simulations Results − μ = 250, σ = 1
Mean of Estimates
1.00
0.95
0.90
0.85
0
250
500
750
1000
Sample Size
estimator
s
s_mle

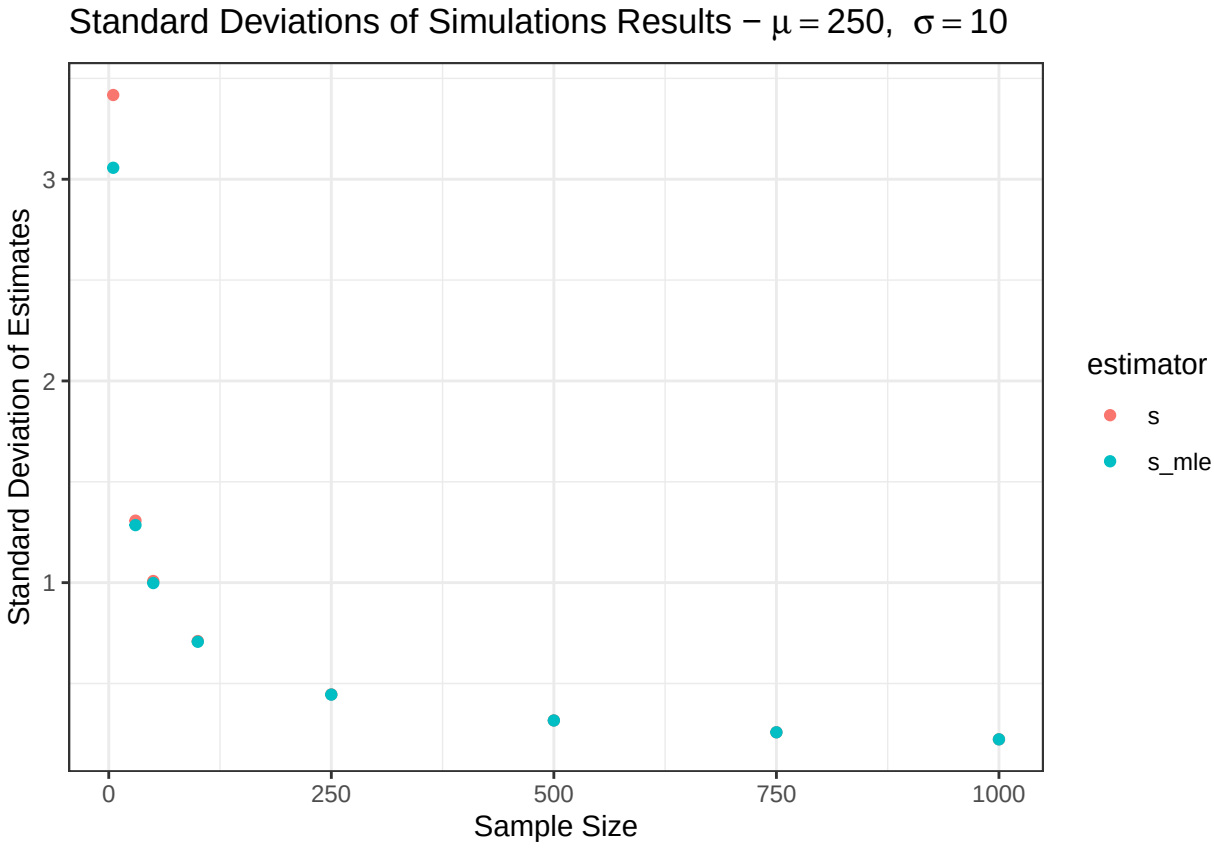

Standard Deviations of Simulations Results − μ = 250, σ = 1
Standard Deviation of Estimates
0.3
0.2
0.1
0
250
500
750
1000
Sample Size
estimator
s
s_mle

Simulations Results - $\mu = 250$ and $\sigma = 10$

Table 5: Simulation Results - mu = 250, sigma = 10

estimator	n	mean	sd
s	5	9.400060	3.4175510
s_mle	5	8.407670	3.0567506
s	30	9.914549	1.3072651
s_mle	30	9.747906	1.2852927
s	50	9.945245	1.0080995
s_mle	50	9.845290	0.9979676
s	100	9.977548	0.7105025
s_mle	100	9.927535	0.7069411
s	250	9.989984	0.4458717
s_mle	250	9.969984	0.4449791
s	500	9.995814	0.3169712
s_mle	500	9.985813	0.3166541
s	750	9.996782	0.2582924
s_mle	750	9.990116	0.2581201
s	1000	9.996947	0.2233575
s_mle	1000	9.991947	0.2232458


Mean Estimates of Simulations Results − μ = 250, σ = 10
Mean of Estimates
10.0
9.5
9.0
8.5
0
250
500
750
1000
Sample Size
estimator
s
s_mle


Standard Deviations of Simulations Results − μ = 250, σ = 10
Standard Deviation of Estimates
3
2
1
0
250
500
750
1000
Sample Size
estimator
s
s_mle

After creating simulated data for several sample sizes and combinations of parameters, we obtained the estimates for the standard deviation given by s and s_{MLE}. For small sample sizes, s clearly peforms better on average for giving the point estimate for σ and approaches the true value as n increases. s_{MLE} does not peform as well, especially for small sample sizes, but converges to the true value as n increases. However, s_{MLE} has the smaller standard deviation in its estimates for small sample sizes, though both s and s_{MLE} have small standard deviations as the sample size gets increasingly large. If we were to create side-by-side boxplots for all the combinations of parameters and sample sizes, the results would be similar to the first example where $n = 30$.

It does appear that changing μ results in a different performance of either estimator. The standard deviations in the estimates are larger when σ is increased to 10, though this seems logical, and both s and s_{MLE} display the same patterns in their performances. To summarize, s performs better on average for the "point estimate" of σ for all sample sizes, as s_{MLE} appears to have a slight bias, and s_{MLe} has a considerably smaller standard deviation for small sample sizes.

Question 2 - Are $\bar{x}$ and s independent?

a.

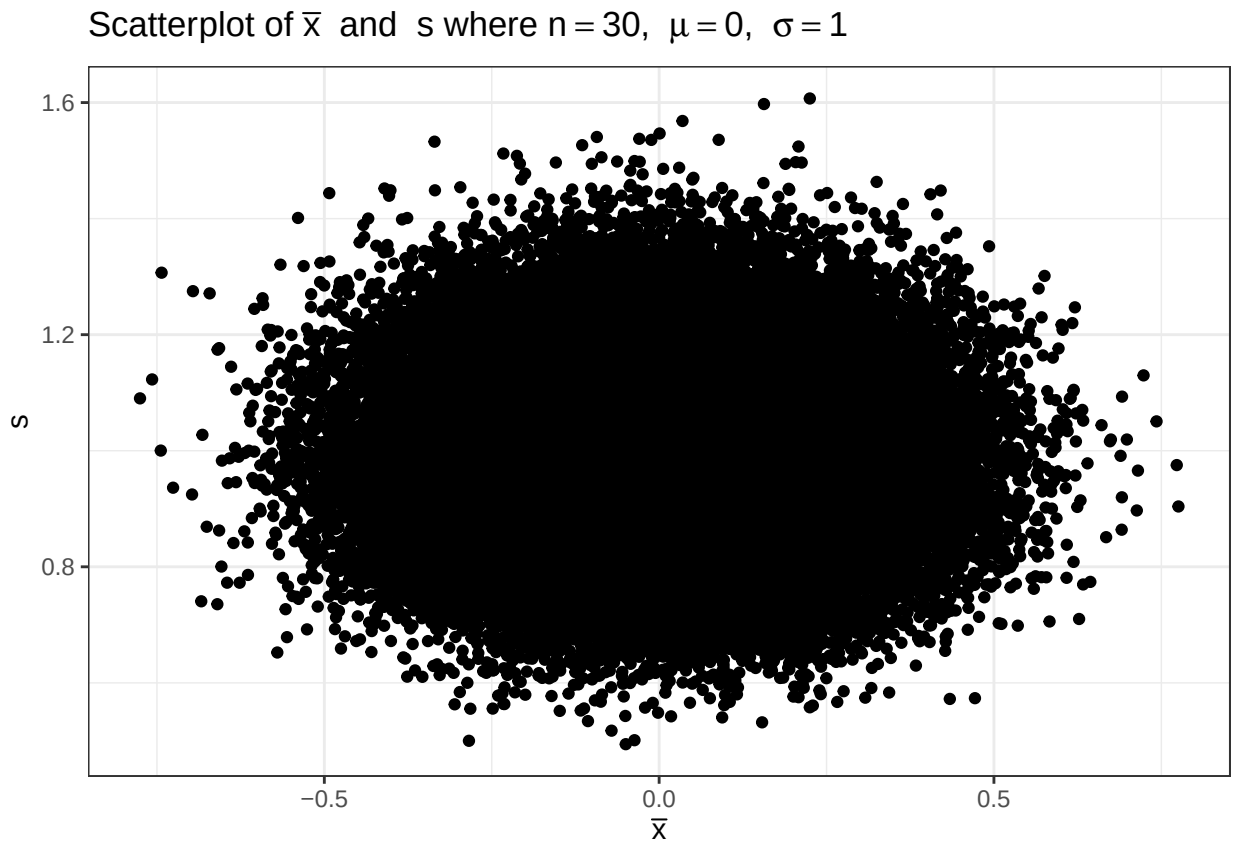


Table 6: Correlation Coefficient

x
-0.005893

There is no relationship between $\bar{x}$ and s, as the scatterplot just shows a dense cloud. The correlation coefficent is also extremely small, being less than 0.01.

b.

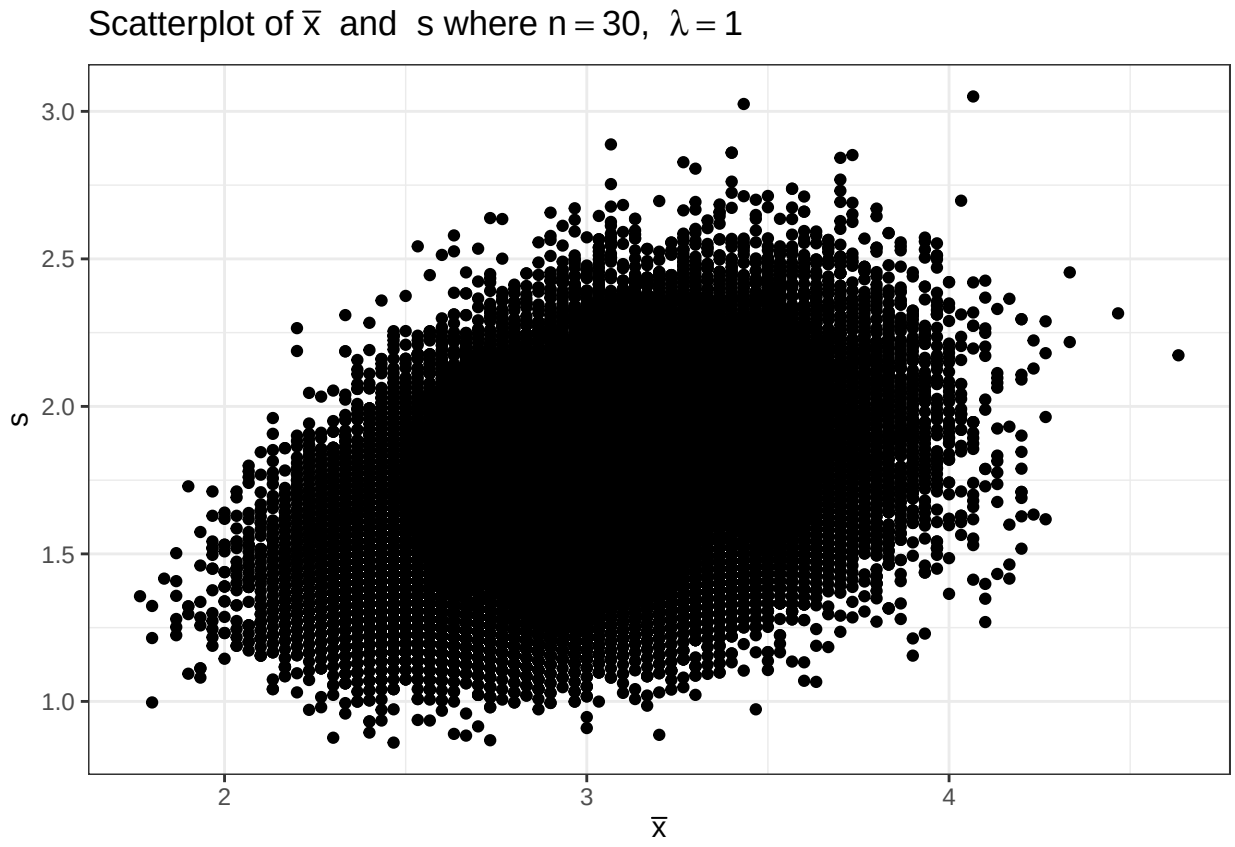


Table 7: Correlation Coefficient

x
0.3720036

There is a clear positive relationship between $\bar{x}$ and s when sampling from a Poisson distribution, as seen in the scatterplot.The correlation coefficient supports this, being about 0.4. This makes sense, as the parameter λ is both the mean and the variance of the distribution, so as one increases so does the other.