# Large Sample Theory Project

Jieqi Tu

11/5/2020

The algorithm of this project:

- Draw n (n = 10, 20 and 30) samples from $Gamma(k = 1.67, \theta = 49.98)$.
- Compute $\bar{x}$ and $s$.
- Repeat 1000 times. Then we will get $(\bar{x_1}, s_1), (\bar{x_2}, s_2), \cdots, (x_{\bar{1}000}, s_{1000})$.
- Compute $\bar{\bar{x}} = \frac{\sum \bar{x_i}}{1000}$, $\bar{s} = [\frac{\sum(\bar{x_i} - \bar{\bar{x}})^2}{1000}]^{\frac{1}{2}}$
- Compute $z = \frac{\bar{x_i} - \bar{\bar{x}}}{\bar{s}/\sqrt{n}}$.
- Draw density function based on $z_i, i = 1, 2, \cdots, 1000$.
- Compute $\bar{\bar{s}}_1^2 = \frac{\sum(\bar{x_i} - \bar{\bar{x}})^2}{1000} \times n$ and $\bar{\bar{s}}_2^2 = \frac{\sum s_i^2}{1000}$

In this project, we want to compare $\bar{\bar{s}}_1^2$, $\bar{\bar{s}}_2^2$ and $k\theta^2$.

### Scenario: n = 10

Then let's firstly try n = 10.

```
set.seed(26)
k = 1.67
n = 10
theta = 49.98
x_bar = numeric(1000)
s = numeric(1000)
for (i in 1:1000) {
  sample = rgamma(n, shape = k, scale = theta)
  x_bar[i] = mean(sample)
  s[i] = sd(sample)
}
mean(x_bar)
```

```
## [1] 83.31535
```

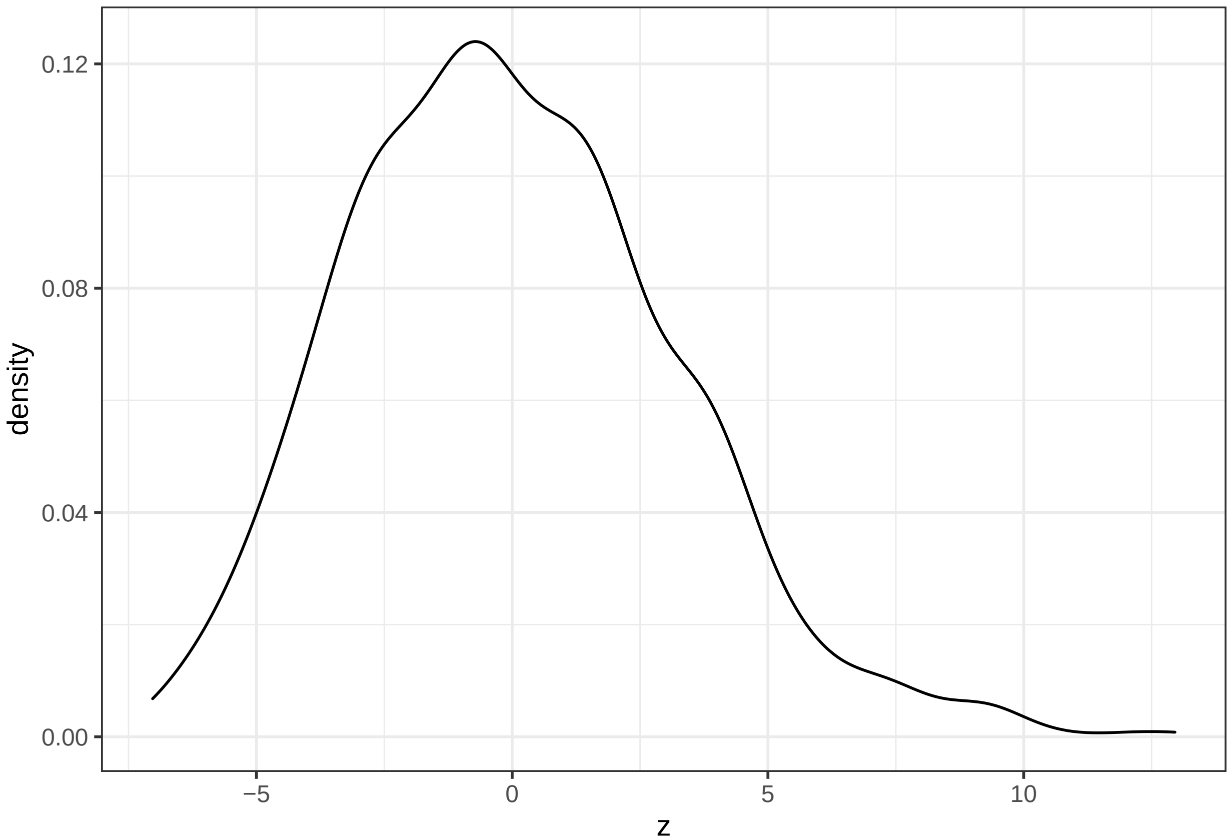
```
sd(x_bar)
```

```
## [1] 18.94633
```

```
z = (x_bar - mean(x_bar))/(sd(x_bar)/sqrt(n))
df = data.frame(z = z)
```

When the sample size n = 10, of 1000 samples, the mean of sample mean is 83.3153535, and the standard deviation of the sample mean is 18.9463298.

```
density1 = ggplot(df, aes(x = z)) + geom_density() + theme_bw()
density1
```




Here we plotted the density of $z$. We can see that the distribution is a little bit right skewed and goes down more rapidly after the peak. The peak is approximately a slightly left to the origin.

```
# Calculate s1 and s2 when n = 10
s1_1 = sum((x_bar-mean(x_bar))^2)/1000 * n; s1_1
```

```
## [1] 3586.044
```

```
s2_1 = sum(s^2)/1000; s2_1
```

```
## [1] 4042.179
```

The values for $\bar{\bar{s}}_1^2$ and $\bar{\bar{s}}_2^2$ are calculated when n = 10. Here, the $\bar{\bar{s}}_2^2$ is slightly larger than $\bar{\bar{s}}_1^2$, and is more close to the theoretical value 4171.660668.

### Scenario: n = 20

Then let us increase the sample size. This time, n = 20.

```
set.seed(26)
n = 20
x_bar2 = numeric(1000)
s2 = numeric(1000)
for (i in 1:1000) {
  sample = rgamma(n, shape = k, scale = theta)
  x_bar2[i] = mean(sample)
  s2[i] = sd(sample)
}
mean(x_bar2)
```

```
## [1] 83.38704
```

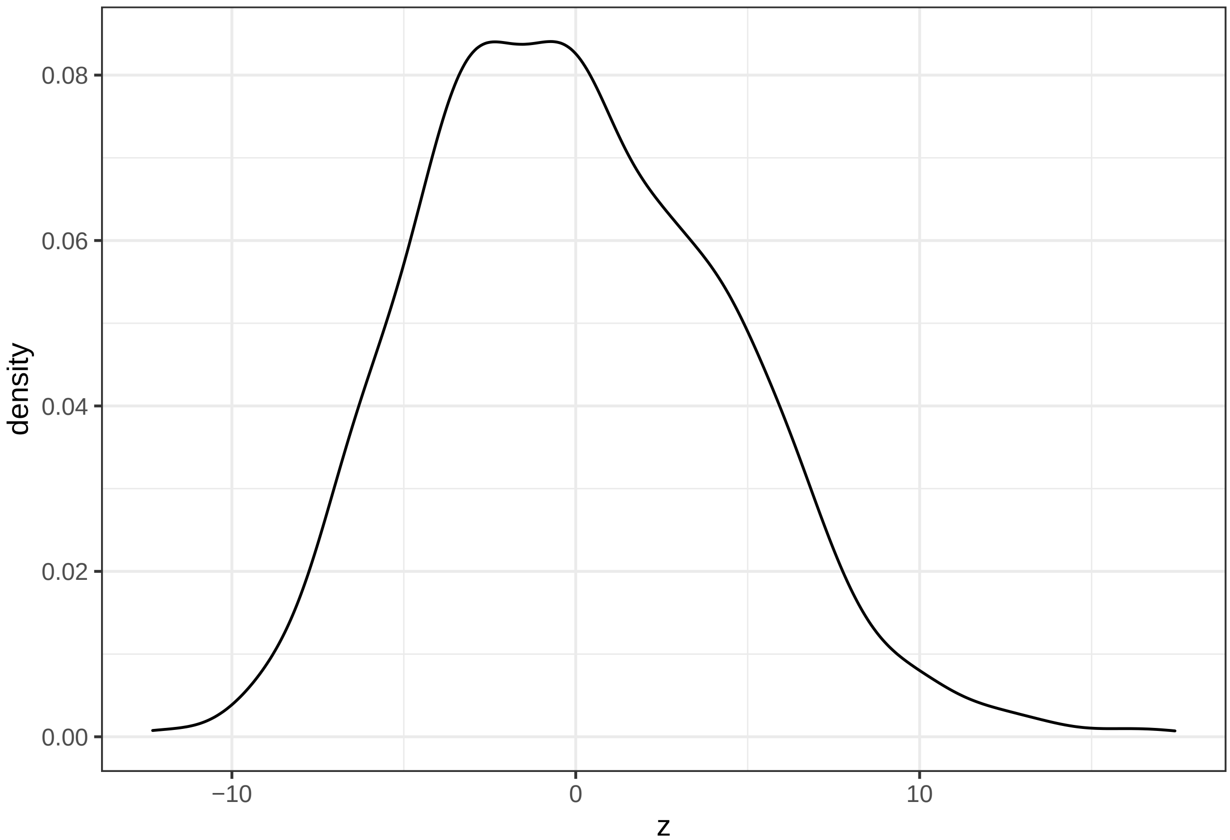
```r
sd(x_bar2)
```

```
## [1] 13.3689
```

```r
z2 = (x_bar2 - mean(x_bar2))/(sd(x_bar2)/sqrt(n))
df2 = data.frame(z = z2)
```

When the sample size n = 20, of 1000 samples, the mean of sample mean is 83.387044, and the standard deviation of the sample mean is 13.3689014. We could notice that, although the mean remains the same as it of n = 10, the standard deviation becomes lower than before.

```r
density2 = ggplot(df2, aes(x = z)) + geom_density() + theme_bw()
density2
```




This is the distribution of $z$ when sample size n = 20. Here we could see that the distribution becomes more symmetric than that of n = 10. This time, the peak also arrives left to the origin. The distribution is still slightly right skewed.

```r
s1_2 = sum((x_bar2-mean(x_bar2))^2)/1000 * n; s1_2
```

```
## [1] 3570.976
```

```r
s2_2 = sum(s2^2)/1000; s2_2
```

```
## [1] 4119.761
```

Here we could see that, $\bar{\bar{s}}_2^2$ is still more close to the theoretical value, compared to $\bar{\bar{s}}_1^2$.

### Scenario: n = 30

Then we want to increase the sample size from 20 to 30, and see what will happen.

```r
set.seed(26)
n = 30
x_bar3 = numeric(1000)
s3 = numeric(1000)
for (i in 1:1000) {
  sample = rgamma(n, shape = k, scale = theta)
  x_bar3[i] = mean(sample)
  s3[i] = sd(sample)
}
mean(x_bar3);
```

```
## [1] 83.47845
```

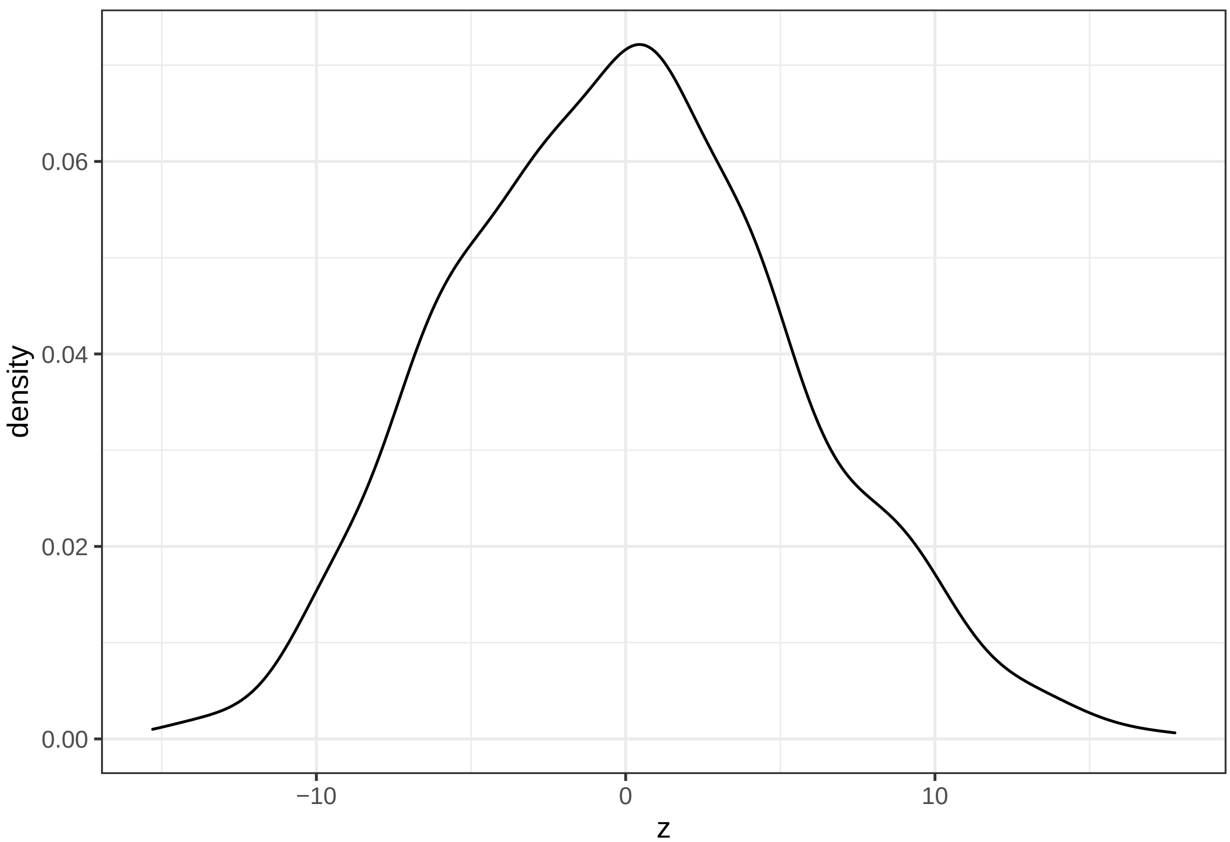
```r
sd(x_bar3)
```

```
## [1] 11.24138
```

```r
z3 = (x_bar3 - mean(x_bar3))/(sd(x_bar3)/sqrt(n))
df3 = data.frame(z = z3)
```

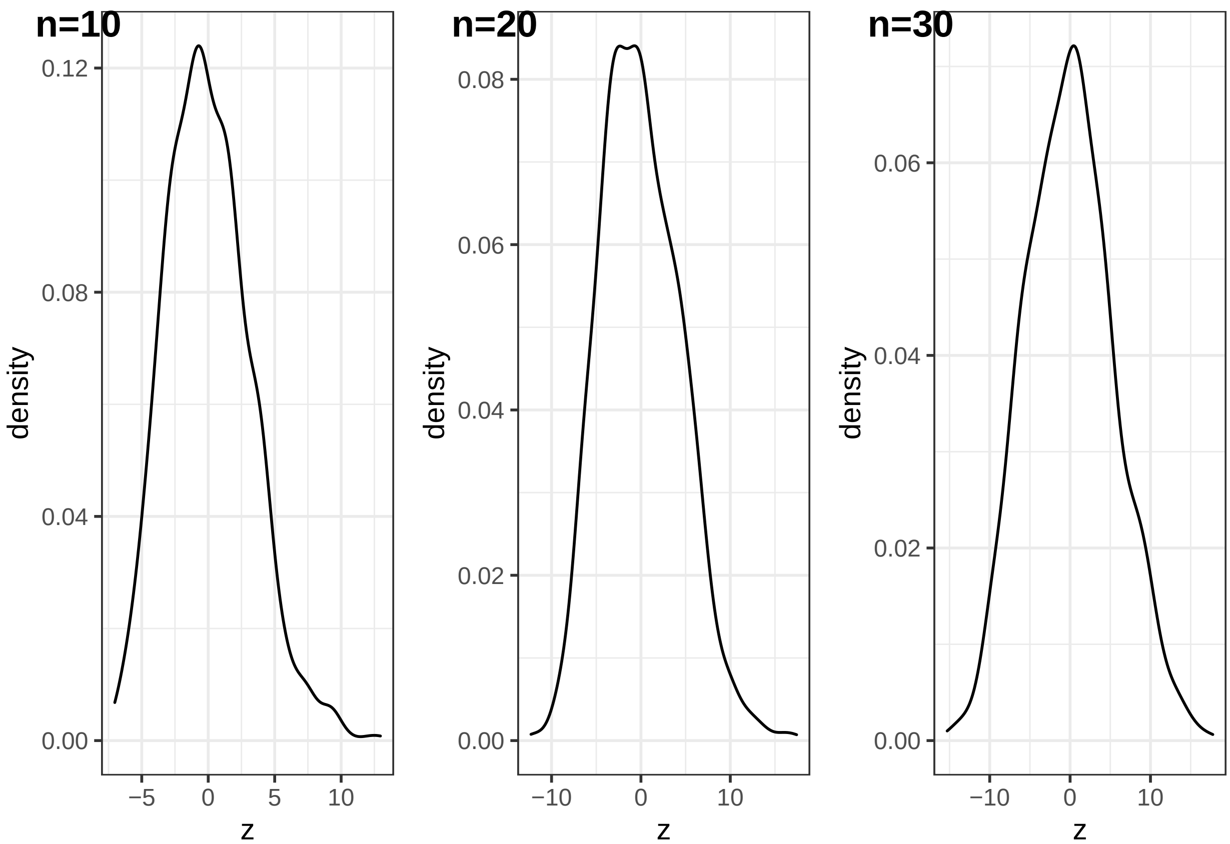
When the sample size n = 30, of 1000 samples, the mean of sample mean is 83.47845, and the standard deviation of the sample mean is 11.2413753. We could see that, although the mean remains close to that of n = 10 and 20, the standard deviation becomes more lower than n = 10 and 20.

```r
density3 = ggplot(df3, aes(x = z)) + geom_density() + theme_bw()
density3
```




The density of $z$ when sample size is 30 looks more symmetric, and almost not skewed. the peak is arrived when $z \approx 0$.

```
ggarrange(density1, density2, density3,
          labels = c("n=10", "n=20", "n=30"),
          ncol = 3, nrow = 1)
```




From the plot panel, we could see that, although they are all bell-shaped and symmetric distributed, the peak density became lower and lower as the sample size increases. Meanwhile, the shape of the distribution becomes more symmetric with one peak at $z = 0$ and looks more normally distributed.

```
s1_3 = sum((x_bar3-mean(x_bar3))^2)/1000 * n; s1_3
```

```
## [1] 3787.265
```

```
s2_3 = sum(s3^2)/1000; s2_3
```

```
## [1] 4127.873
```

Here we could see that, $\bar{\bar{s}}_2^2$ is still more close to the theoretical value, compared to $\bar{\bar{s}}_1^2$.

## Compare 3 scenarios

```
s1 = c(s1_1, s1_2, s1_3)
s2 = c(s2_1, s2_2, s2_3)
compare = data.frame(s1, s2)
theoretical = k*theta^2
theoretical
```

```
## [1] 4171.661
```

```r
compare$'s1-theoretical' = s1 - theoretical
compare$'s2-theoretical' = s2 - theoretical
compare %>% knitr::kable()
```

| s1 | s2 | s1-theoretical | s2-theoretical |
|---:|---:|---:|---:|
| 3586.044 | 4042.179 | -585.6162 | -129.48170 |
| 3570.976 | 4119.761 | -600.6847 | -51.89986 |
| 3787.265 | 4127.873 | -384.3961 | -43.78796 |

We want to compare the $\bar{\bar{s}}_1^2$, $\bar{\bar{s}}_2^2$ and the theoretical value (true) $k\theta^2 = 4171.661$. From the result, we have several findings:

- As the sample size increases, $\bar{\bar{s}}_1^2$ and $\bar{\bar{s}}_2^2$ are getting more closer to the theoretical value.
- No matter what sample size we choose, $\bar{\bar{s}}_2^2$ is always closer to the theoretical value than $\bar{\bar{s}}_1^2$.
- We still need to take it into consideration that $n = 10, 20, 30$ are still not large sample size, so the sample mean and sample variance might not accurately relect the population parameters. Therefore, even though we repeatedly draw samples 1000 times, we still have considerable differences between our estimates and the theoretical value.

Therefore, we want to show that $\bar{\bar{s}}_1^2$ is a biased estimator of the population variance while $\bar{\bar{s}}_2^2$ is an unbiased estimator of the population variance.

### Proof 1: $\bar{\bar{s}}_1^2$ is a biased estimator of the population variance

Since each element in the sample follows $Gamma(k, \theta)$, we have large enough repeated times (1000), according to Lindeberg-Levy CLT, we have:

$$\sqrt{n}(\bar{X}_i - k\theta) \xrightarrow{\text{d}} N(0, k\theta^2)$$

Since $\bar{\bar{s}}_1^2 = \frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{1000} \times n$, let us consider $\bar{x}_i$ as a random variable with $E(\bar{X}_i) = k\theta, Var(\bar{X}_i) = k\theta^2/n$, and $n = 10, 20, 30$. Hence we have $Var(\bar{X}_i) = k\theta^2/n$. However, in order to estimate the variance parameter, the unbiased estimator would be $\frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{1000-1}$ (This result we used here was proved in Large Sample Theory Homework 2). It can be shown that $E(\frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{1000}) = (1 - \frac{1}{1000})k\theta^2/n$. Hence the unbiased estimator for the population variance $k\theta^2$ would be $\frac{\sum(\bar{x}_i - \bar{\bar{x}})^2}{1000-1} \times n$, and $n = 10, 20, 30$. Therefore, the estimates would be $1/1000$ smaller than the theoretical value of population variance $k\theta^2$.

### Proof 2: $\bar{\bar{s}}_2^2$ is an unbiased estimator of the population variance

Since $\bar{\bar{s}}_2^2 = \frac{\sum s_i^2}{1000}$, let us consider $s_i$, $i = 1, 2, \cdots, 1000$ as a random variable. In Large Sample Theory Homework 2, we have shown that $s_i^2 \to \sigma^2$. So for each $s_i$, we have $E(s_i) = \sigma^2 = k\theta^2$. Then, apply the Weak Law of Large Numbers here, we can have: $\frac{\sum s_i^2}{N} \xrightarrow{\text{p}} E(s^2) = k\theta^2$ as $N \to \infty$. Therefore, $\bar{\bar{s}}_2^2$ is an unbiased estimator for $k\theta^2$ if we have a large N.