S04 - 2008
The Over-Reliance on the Central Limit Theorem

Abstract

The objective is to demonstrate the theoretical and practical implication of the central limit theorem. The theorem states that as n approaches infinity, the distribution of the sample mean approaches normality with mean equal to the population mean and variance equal to the population variance divided by n. However, as n approaches infinity, the variance of the mean approaches zero. In practice, the population variance is unknown, and so the sample variance is used to estimate the population distribution. In that case, we assume the format of a t-distribution, which requires the assumption that the population is itself normally distributed. In this presentation, we use data visualization to show some problems that can occur when assuming that n is sufficiently large to assume that the sample mean is normally distributed. In particular, we use PROC SURVEYSELECT to sample data from non-normal distributions to compare the distribution of the sample mean to that of the population mean.

Introduction

Regression requires the assumption that the residuals are normally distributed. However, most healthcare data are exponential or gamma because of the presence of extreme outliers. As shown in the previous section, there is a considerable difference between the mean and the median.

Linear regression requires large samples to be effective. Power analysis tends to assume that the population distribution is sufficiently homogeneous to be normally distributed. However, healthcare outcomes tend to be exponential or gamma distributions, and we must consider just how large n has to be before the Central Limit Theorem is realistic.[65] To examine the issue, we take samples of different sizes to compute the distribution of the sample mean. The following code will compute 100 mean values from sample sizes starting with 5 and increasing to 10,000.

```
PROC SURVEYSELECT DATA=nis.nis_205 OUT=work.samples
METHOD=SRS N=5 rep=100 noprint;
RUN;
proc means data=work.samples noprint;
  by replicate;
  var los;
  output out=out mean=mean;
run;
```

Once we have computed the means, we can graph them using kernel density estimation (a smoothed histogram). We show the difference between the distribution of the population, and the distribution of the sample mean for the differing sample sizes. Figures 1-4 show the distribution of the sample mean compared to the distribution of the population for differing sample sizes. To compute the distribution of the sample mean, we collect 100

different samples using the above code. We compute the mean for the patient length of stay using the National Inpatient Sample.

Figure 1. Sample Mean With Sample=5

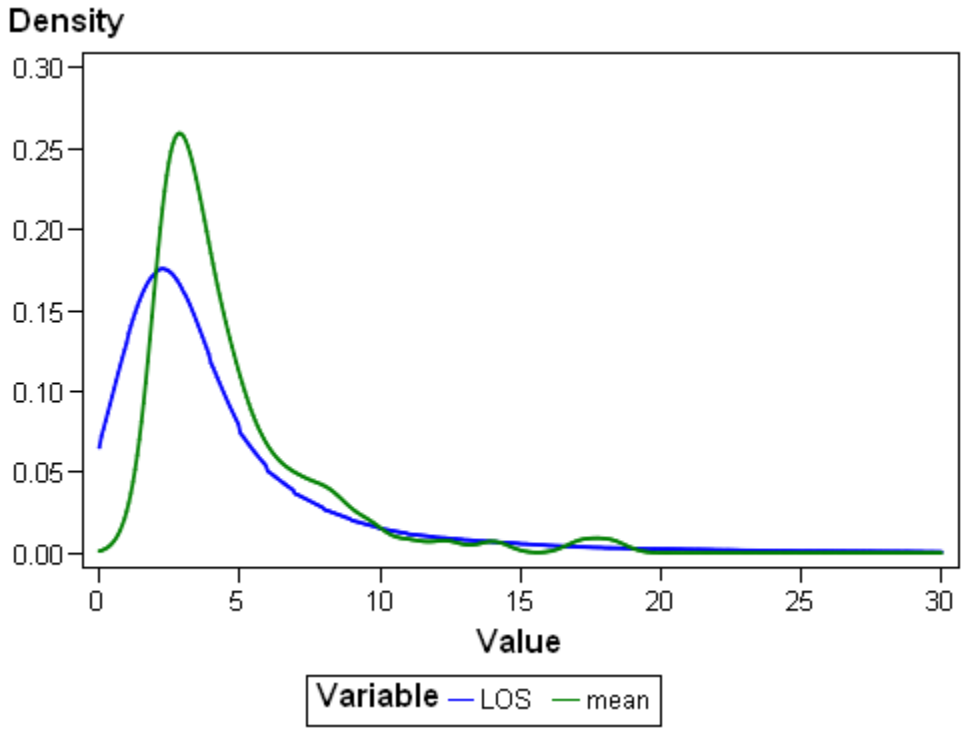


Figure 2. Sample Mean With Sample=30

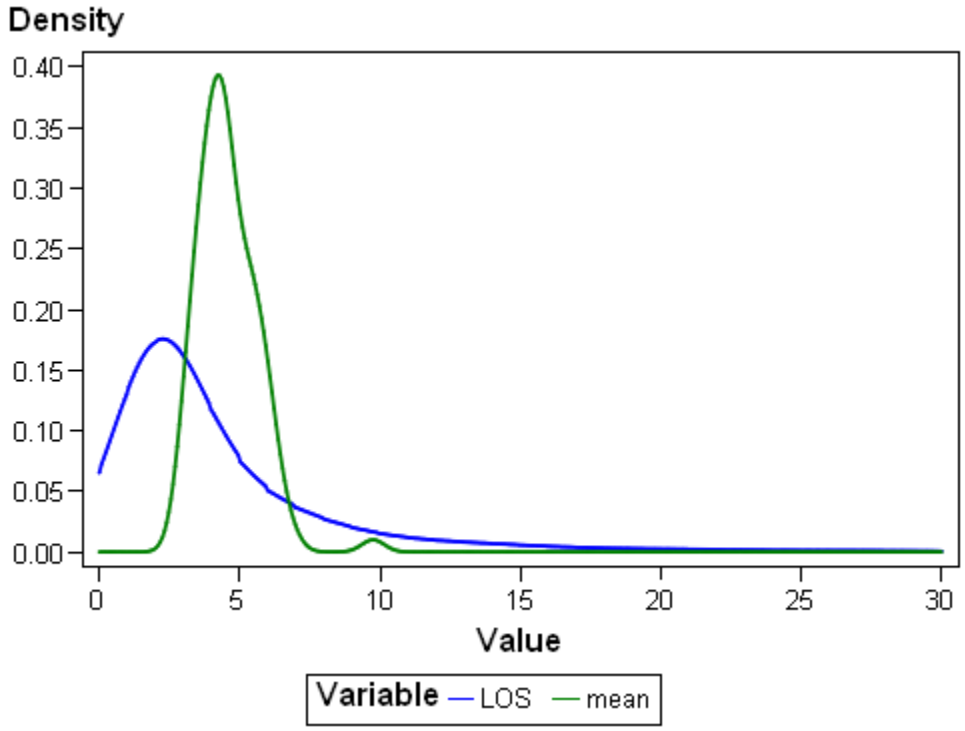


Figure 3. Sample Mean With Sample=100

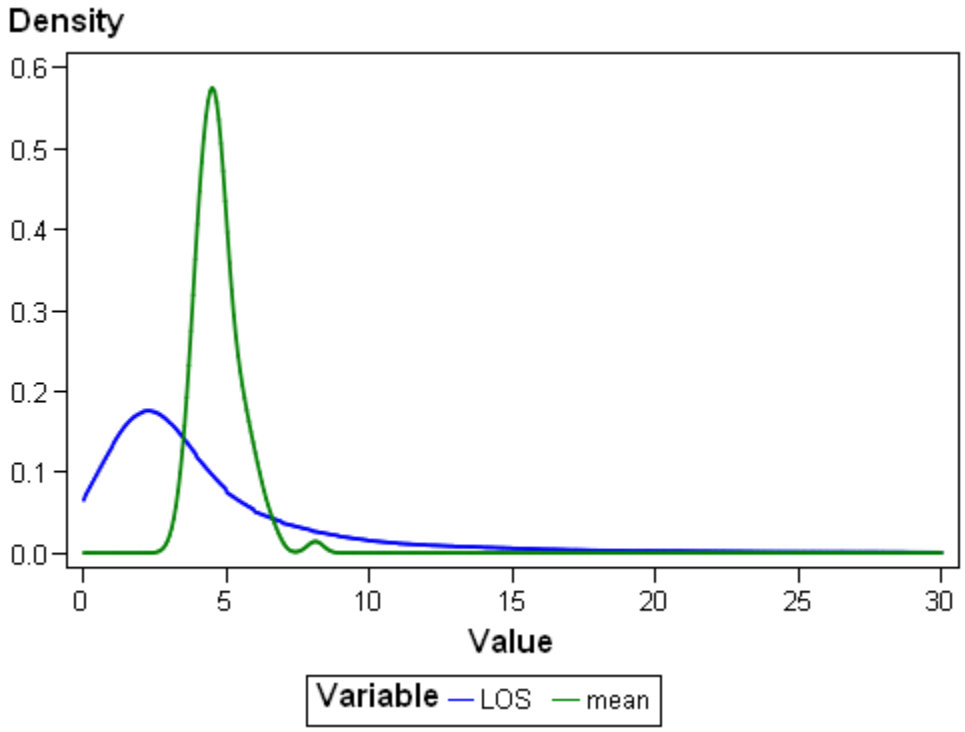
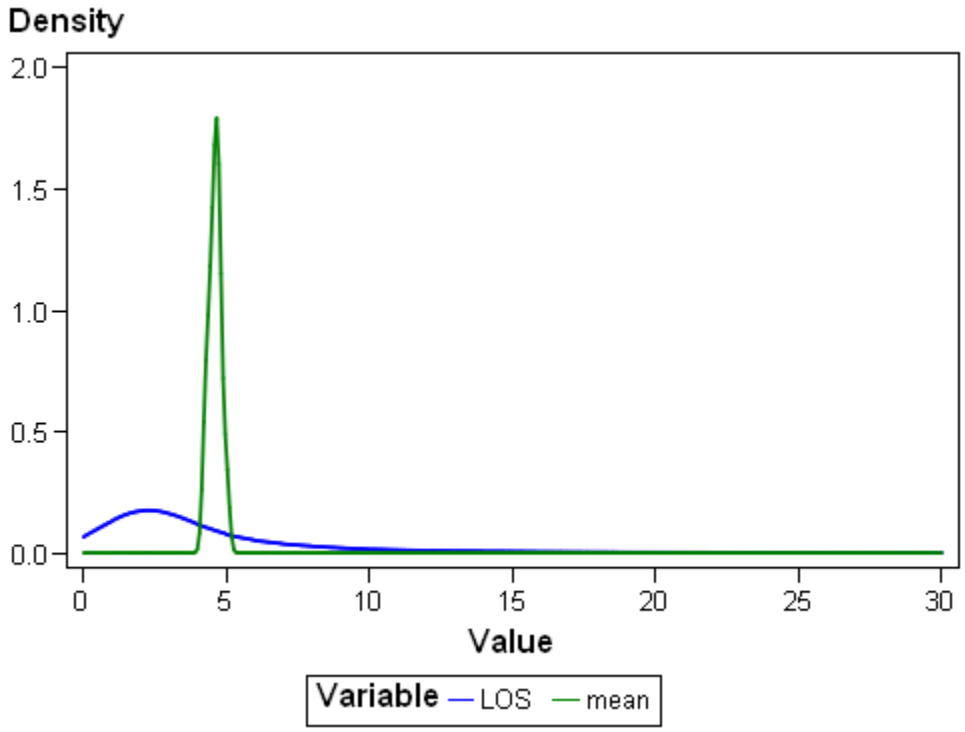


Figure 4. Sample Mean With Sample=1000



In Figure 1, the sample mean peaks slightly to the right of the peak of the population distribution; this peak is much more exaggerated in Figure 2. The reason for this shift in the peak is because the sample mean is susceptible to the influence of outliers, and the population is very skewed. Because it is so skewed, the distribution of the sample mean is not entirely normal. As the sample increases to 100 and then to 1000, this shift from the population peak to the sample peak becomes much more exaggerated. We use the same sample sizes for 1000 replicates (Figures 5-8).

Figure 1. Sample Mean for Sample Size=5 and 1000 Replicates

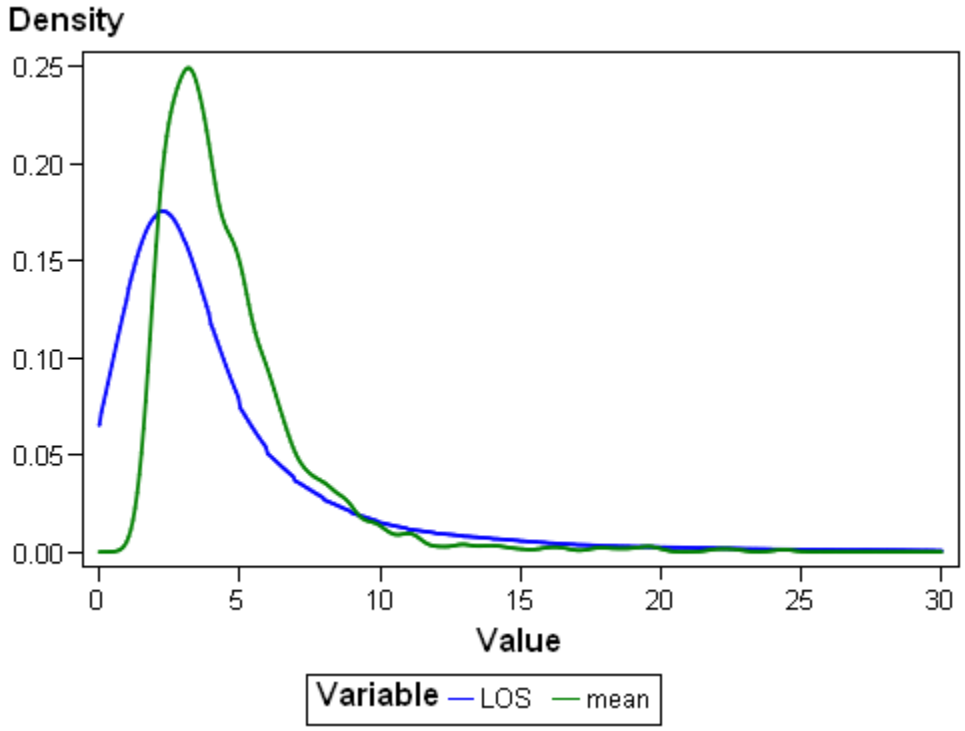


Figure 2. Sample Mean for Sample Size=30 and 1000 Replicates

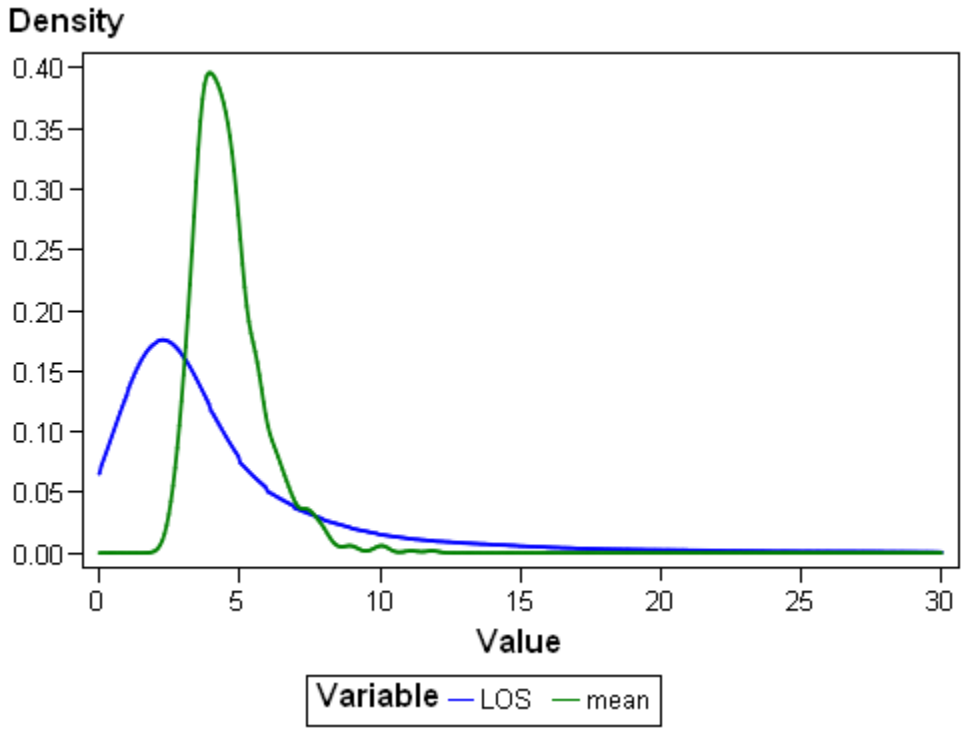


Figure 3. Sample Mean for Sample Size=100 and 1000 Replicates

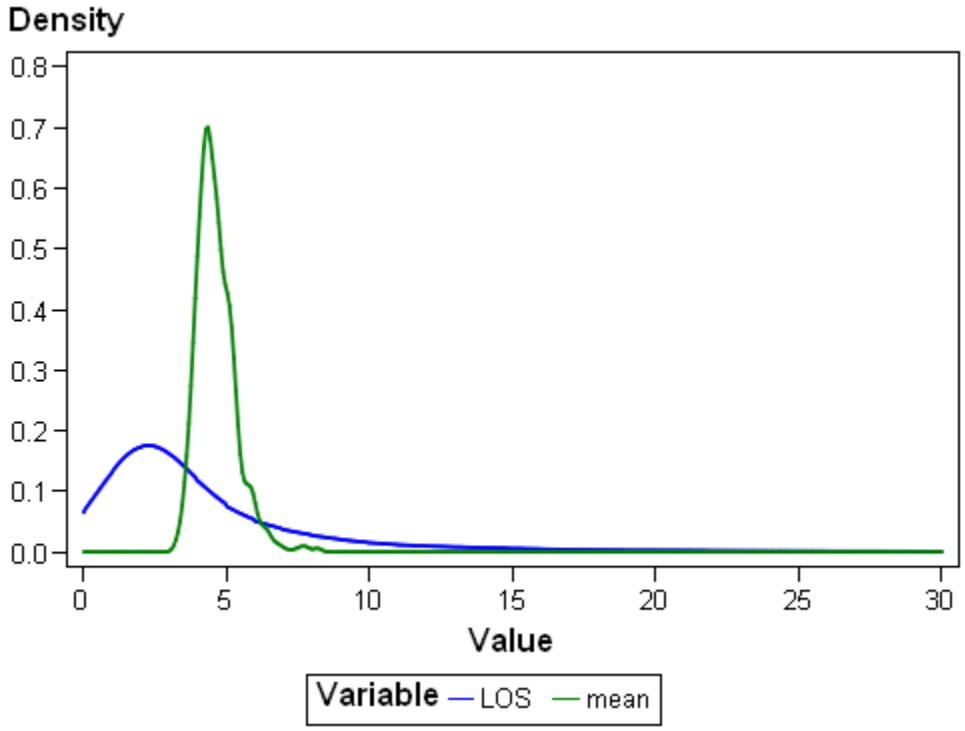
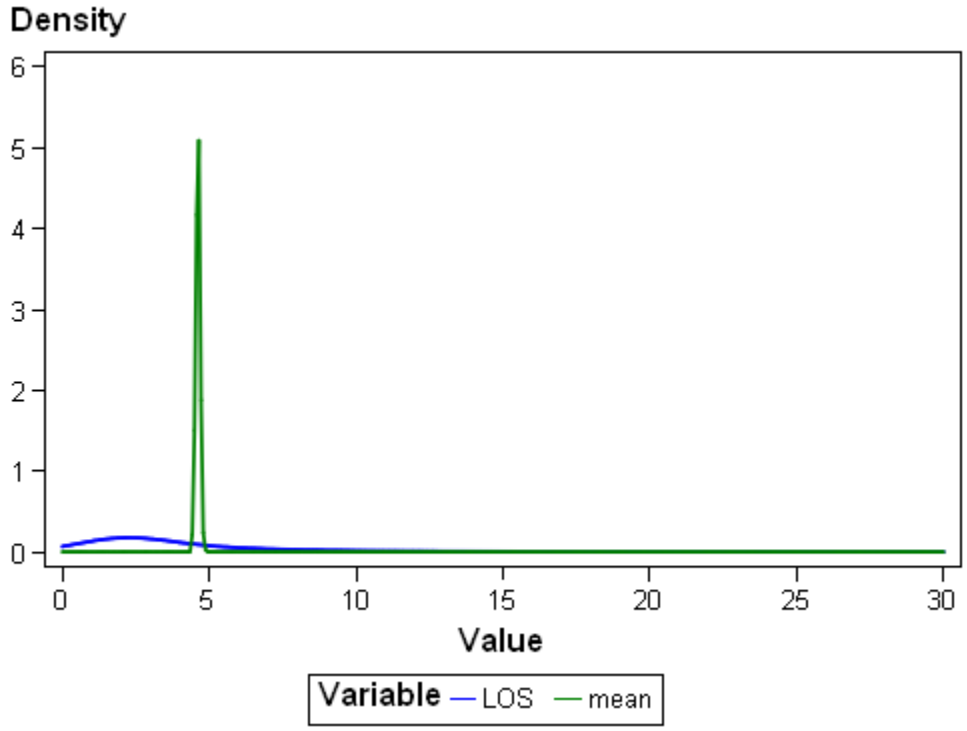


Figure 4. Sample Mean for Sample Size=1000 and 1000 Replicates



It is again noticeable that the sample mean is shifted away from the peak value of the population distribution because of the skewed distribution. However, the distribution of the mean is not normally distributed.