DATA 581

Modeling and Simulation II

Lecture 1: Simulation Studies; Bootstrapping

What We Discuss Today

- Simulation studies in data analysis
- A simulation study to compare sample mean and sample median
- Bootstrapping
- Parametric vs nonparametirc bootstapping

Motivation

Assume we have been given the task of estimating the effectiveness of a brand new drug in a clinical trial;

1. we collected the data
2. we have considered three models that we believe can explain the effect of new drug.

Question: Which model perform better?

Motivation

Note: We do not know the true effect of the treatment;

- We cannot compare how the model estimates it to the actual effect.
- As opposed to a predictive model, in which we know the true labels and can evaluate our predictions against them.

Choosing the right model can greatly influence the conclusion one draws;

- For instance, Type I and Type II errors depend heavily on the methods used and/or assumptions made.

This is where simulation studies come into their own.

Definition

Simulation studies involve creating data by pseudo-random sampling.

A simulation study provides empirical evidence of the performance of statistical methods in different scenarios.

Using simulation studies can help you make better decisions when;

- choosing statistical models.
- evaluating a new or existing model

Data Generating Mechanisms

When an analyst designs a simulation study, they typically spend most of their time on generating dataset.

Data generation mechanis utilize random numbers generators to produce a dataset.

- Data can be generated using parametric draws from a known distribution, or
- by sampling with replacement from an existing dataset.

It is usually necessary to simulate a few scenarios, such as varying the sample size and/or effect size.

Sample Mean vs Sample Median

Example: Design a simulation study to compare sample mean and sample median for the following scenarios;

- It is known (e.g. Hooker, 1907) that the median can sometimes be more appropriate than the mean when measuring central tendency.
- For normal data, it is known that the variance of the mean is less than the variance of the median.
- When there are outliers, the median is often recommended over the mean.

Simulation Study

In our study, we

1. **simulated random data from several different distributions: normal, t distribution with $2, 10, 20$ degrees of freedom at sample sizes $n = 10, 30$ and 100.**

2. **calculated the means and medians for each sample**

3. **created a boxplot of the means and medians for visual comparison**

4. **computed the variances of the means and medians for numerical comparison**

Results: Tabular Variance Comparisons

Sample Size: 10

	variance of means	variance of medians
normal	0.10	0.14
t20	0.12	0.15
t10	0.13	0.15
t2	1.70	0.23

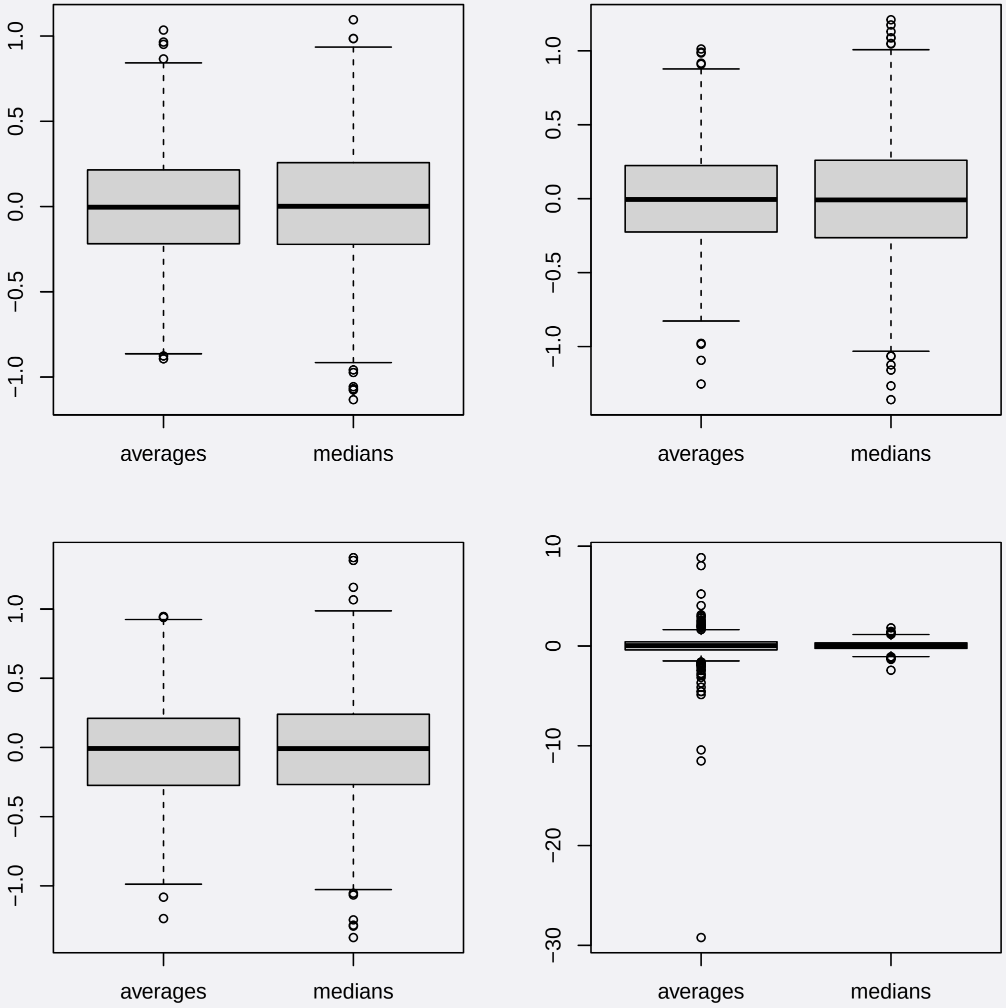
Sample Size: 30

	variance of means	variance of medians
normal	0.04	0.05
t20	0.04	0.05
t10	0.04	0.05
t2	0.70	0.08

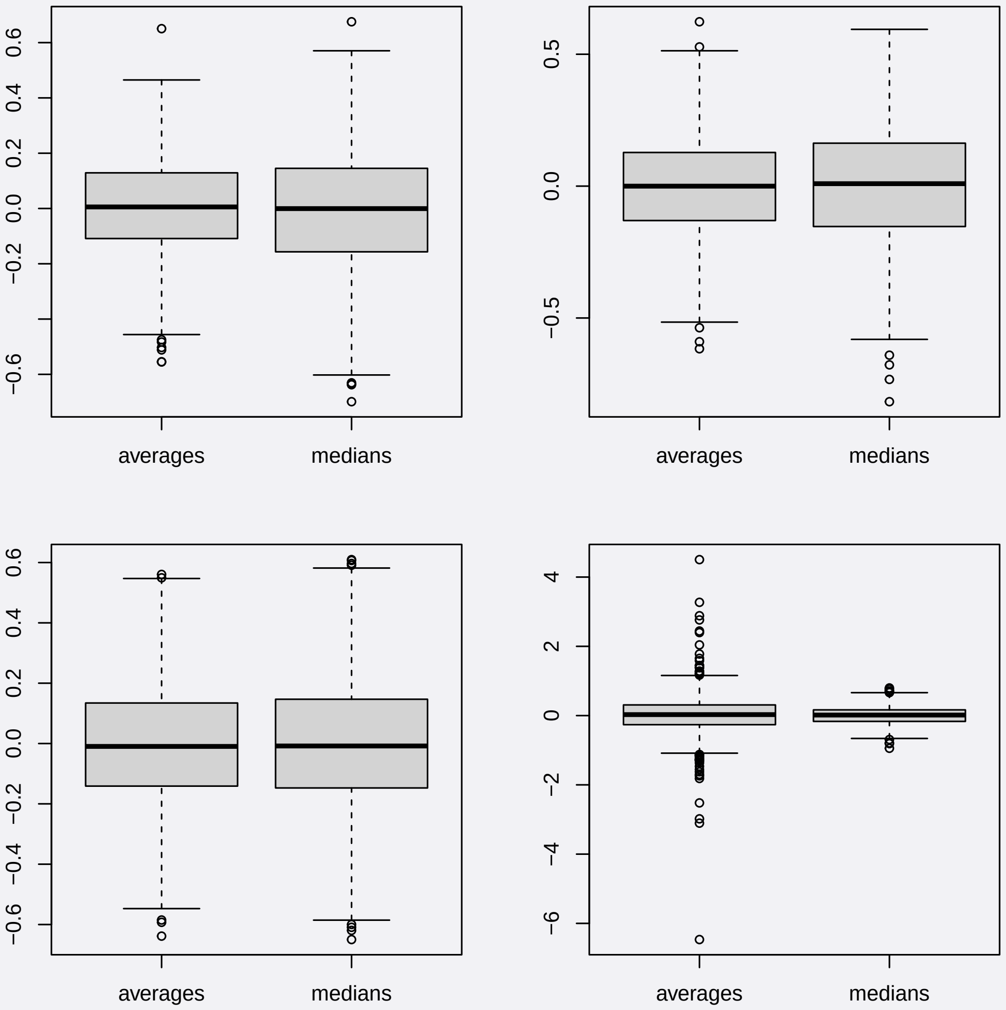
Sample Size: 100

	variance of means	variance of medians
normal	0.01	0.01
t20	0.01	0.02
t10	0.01	0.02
t2	0.11	0.02

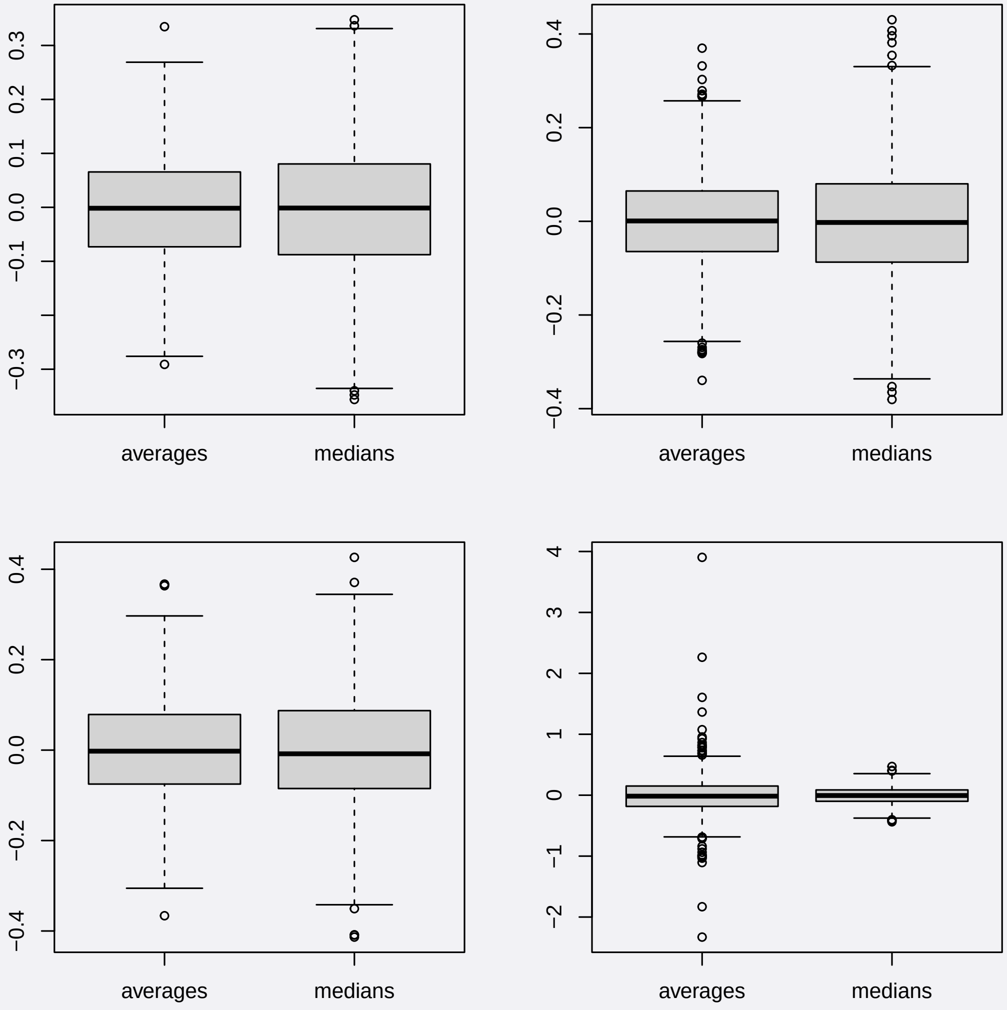
Results: Boxplot Comparisons (Sample Size: 10)



Results: Boxplot Comparisons (Sample Size: 30)



Results: Boxplot Comparisons (Sample Size: 100)



Conclusions

For normal, t_{20}, and t_{10}, the mean has smaller variance than the median, independent of sample size.

For t_2 data, the mean has a much larger variance.

The variability of the mean and median tends to be similar except for the t_2 case, where the variability of the mean can be quite extreme.

Conclusion: For normal/ close to normal distribution (t_{20}, and t_{10}) use mean and when there are outliers (t_2) , the median is often recommended over the mean.

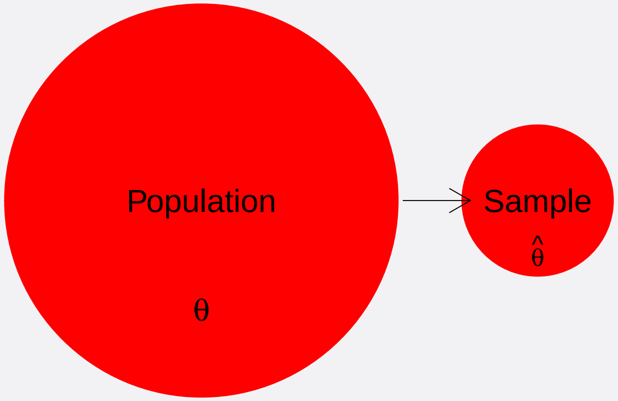
Data Generation for Simulation Study

- In the last example, we simulate data from known distributions (normal, t distribution).

- How about when where data distribution in not known? i.e we only have small dataset.
  - Boostrapping Technique

Review; A typical statistical problem

Estimate the population parameter θ using the sample estimate $\widehat{\theta}$.



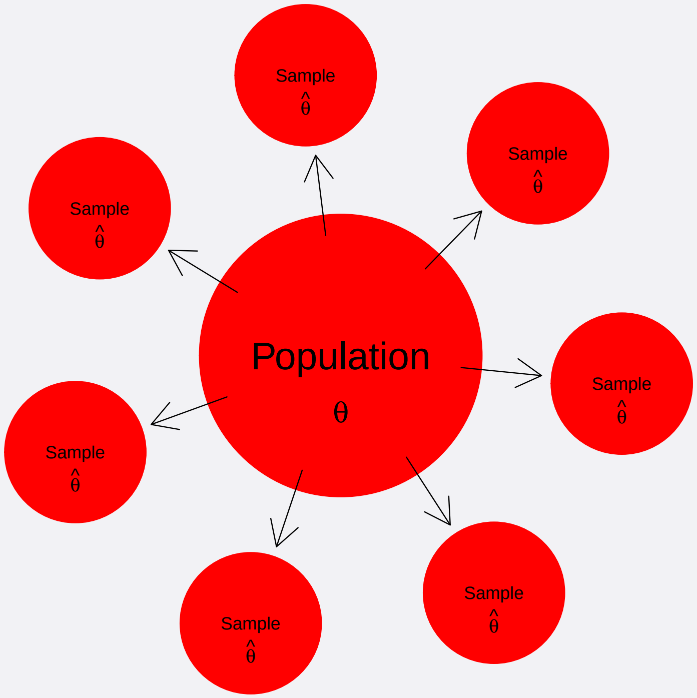
- **Fact:** $\theta \neq \widehat{\theta}$

- **Question:** How wrong is the estimate?

- **Statistical Answer: assess variability of $\widehat{\theta}$**
  - We can use standard errors, confidence intervals, p-values for hypothesis tests about θ.
  - These can be obtained by evaluating sampling distribution of the statistic ($\widehat{\theta}$).

Review

In an ideal world:



Assess variability of the sample estimate $\widehat{\theta}$ by taking additional samples, obtaining new estimates of θ each time.

However, this is not feasible for many scenarios!

Bootstrapping

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement.

By sampling over and over again, bootstrapping approximates the true population data; Law of Large Numbers.

When we are limited to a small dataset (sample), as well as saving time and money, bootstrapped samples can be quite good approximations for population parameters.

Forms of Bootstrapping

There are two major forms of bootstrapping which differ primarily in how the population is estimated.

1. Parametric Bootsrapping: when data distribution is known or assumed.

2. Nonparametric (Emperical) Bootsrapping: when data distribution is not known.

3. Semiparametric

Example - Nonparametric bootstrapping

Example: Suppose we have a the following measurements of cholesterol level before and after taking a drug in our sample.

```
## Before : 197 203 201 202 204 205 203 203 205 201
##  After : 199 198 199 202 198 199 199 198 197 199
```

We want to evaluate whether the new drug is effective.

One way is to compare the average level of chlostrol before and after taking the drug.

```
## Average difference of cholesterol levels is : -3.6
```

We can see that cholesterol level has reduced by 3.6, but is this enough to say that the drug is effective? This variation might be due to chance!

The Nonparametric bootstrap

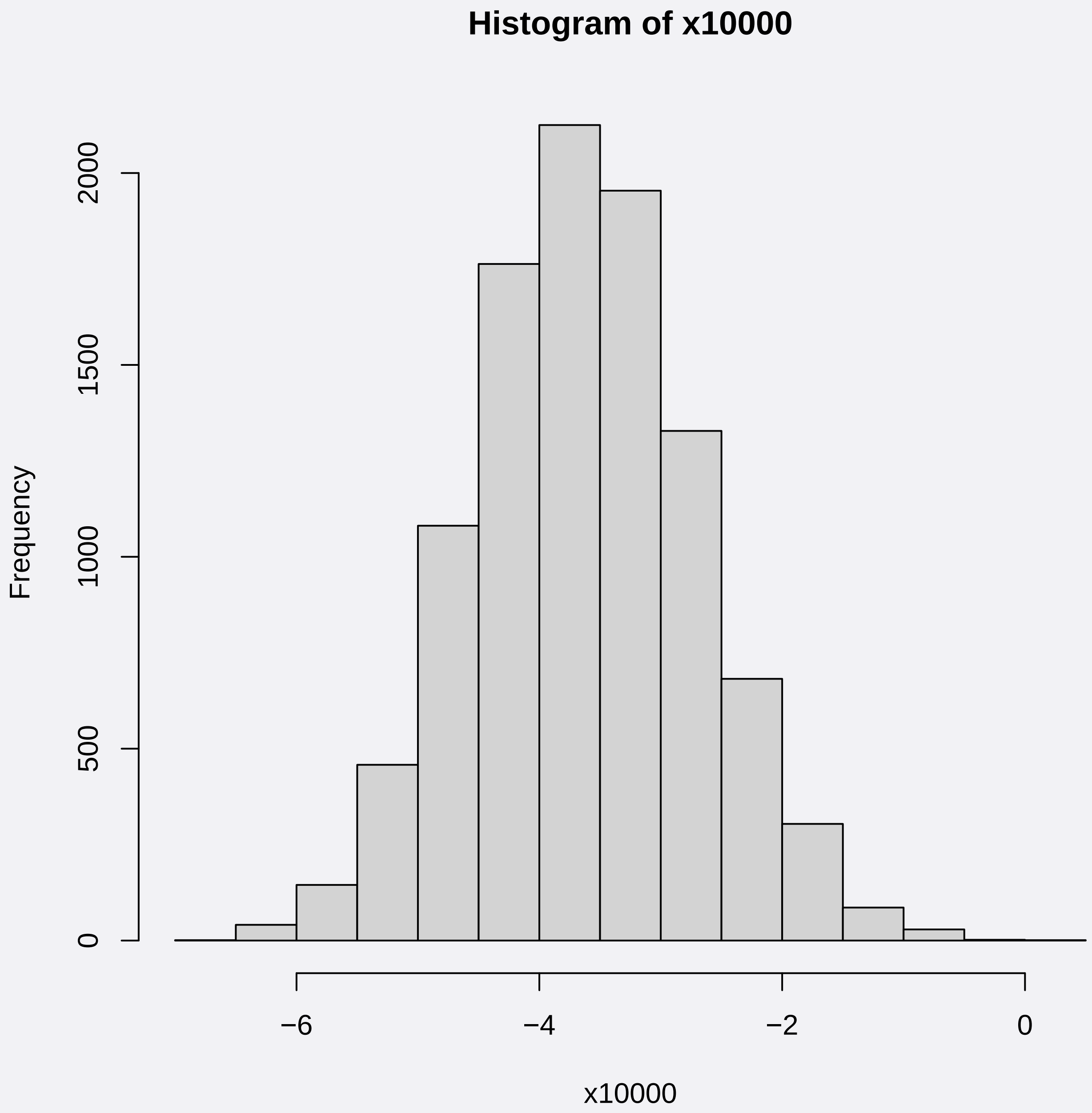
We don’t know the distrubition of the population.

We can use bootstrapping to simulate from this data set.

1. **Draw a resample with replacement from the original sample, with exactly the same size.**
2. **Calculate the statistic (mean of the difference) for the resample and store it.**
3. **Go to 1, repeat hundreds or even thousands of times.**

This can be done in R as below:

```r
x10000<- replicate( 10000,
mean(sample(diff, size = 10, replace= TRUE)))
```


Histogram of x10000
Frequency
0
500
1000
1500
2000
−6
−4
−2
0
x10000

The Nonparametric bootstrap

The sampling distribution seems to be normal!

We can describe the simulated population:

```
## Center of the bootstraped population is : -3.58165
##  Standard devation is : 0.9179175
```

Note: This is not a coincidence! Based on the **central limit theorom**, the sampling distribution for mean of difference follows a normal distribution centered around true population parameter with estimated standard error $SE = \frac{sample std.}{\sqrt{n}}$.

we can check SE with our sample,

```r
cat('estimated SE is:', sd(diff)/sqrt(10))
```

```
## estimated SE is: 0.9683893
```

Pretty close!

The Nonparametric bootstrap

We can use SE (standard error of the sampling dist) to obtain margin of error and confidence interval;

Confidene interval = point estimate $\pm$ critical value $\times SE$

Critical value depends on sampling distribution and confidence level.

For example, in this case since sampling distribution is normal, critical value for 95% confidence level is : 1.96

So, we can say we're 95% confident that cholestrol level is reduced by the drug fall in the range below:

$$-3.6 \pm 1.96 * 0.917 = (-5.39, -1.8)$$

Is drug effective?

The Nonparametric bootstrap

If the sampling distribution for the statistic is not known, we can not estimate the SE directly however, we can still calculate the confidence interval.

1. **calculated the statistic from the bootstrapping samples.**

2. **Sort the bootstrapped sample statistic,**

3. **For** $(1\alpha)100\%$ **percentile bootstrap confidence interval for the population parameter, calculate** $(\alpha/2)100$**th and** $(1-\alpha/2)100$**th percentiles**

    - **for example, 90% confidence interval is obtained by;**

$$CI_{90\%} = 95\%percentile - 5\%percentile$$

What to Take Away from this Lecture

- Some notes on simulation studies
- bootstrapping as a simulation procedure
- Parametric bootstrap vs Nonparametric bootsrapping