Simple Ain't Easy: Real-World Problems with Basic Summary Statistics

John Myles White and other contributors

Problem 1

The sampling distribution of the median can be multimodal if the source distribution is multimodal.

Example:

Let D be a distribution defined as a 50/50 mixture of two normals. As a specific example, we will assume that D is a mixture of two normals: $\mathcal{N}(-10, 1)$ and $\mathcal{N}(+10, 1)$. The PDF for this distribution is shown in Figure **median/001/001**:

For this bimodal source distribution, D, the sampling distribution is not close to being normally distributed, because the median is most likely to be defined by either two points from the first mixture component or by two points from the second mixture component. This can be seen in the simulation results shown below:

In this case, the contrast between the erratic behavior of the sample mean and the sample median is very stark. As predicted by the Central Limit Theorem, the sampling distribution of the mean is approximately normal:

In the table below, we show the estimated standard deviations of the sampling distributions of the sample median and sample mean for this case. The median is roughly 20x more variable:

Std. Dev. of Sample Median	**Std. Dev. of Sample Mean**
2.918197	0.1612805

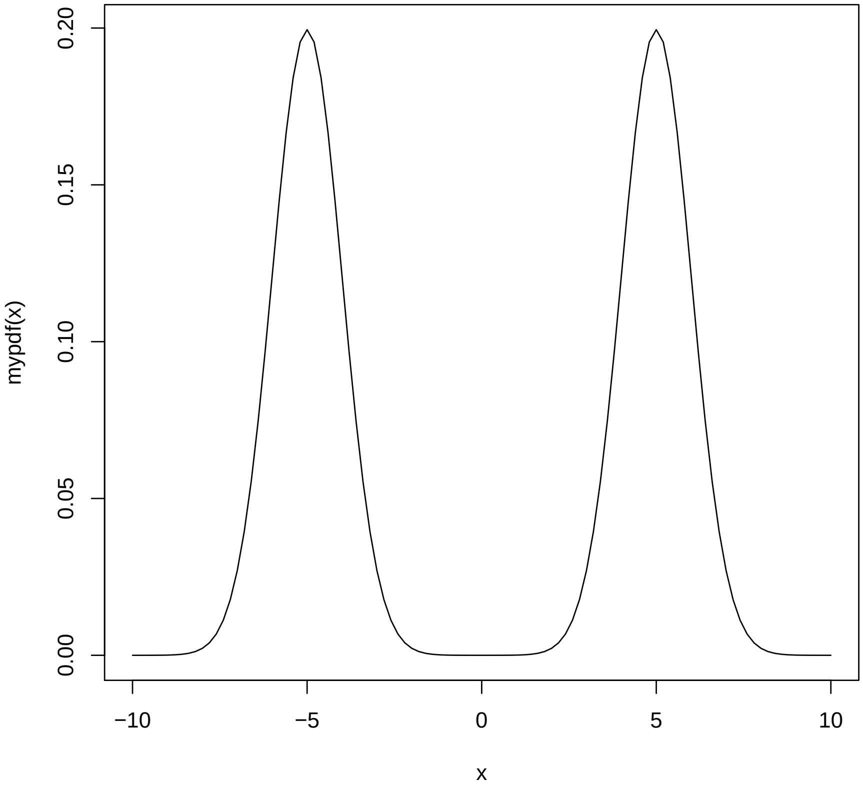

Figure 1: Probability Density Function of Source Distribution

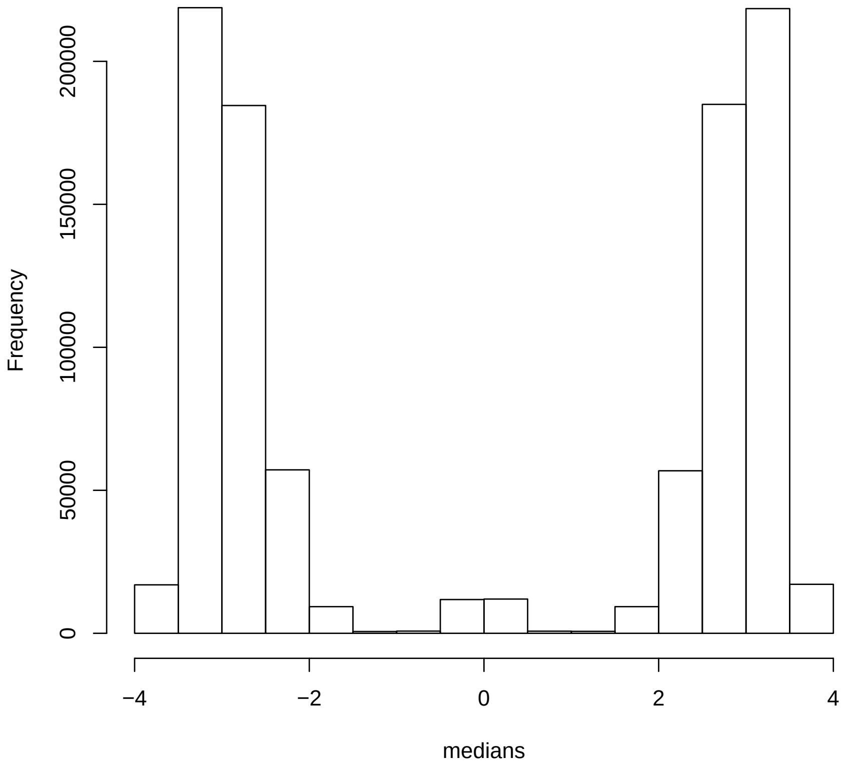

Figure 2: Histogram of the Sampling Distribution of the Median

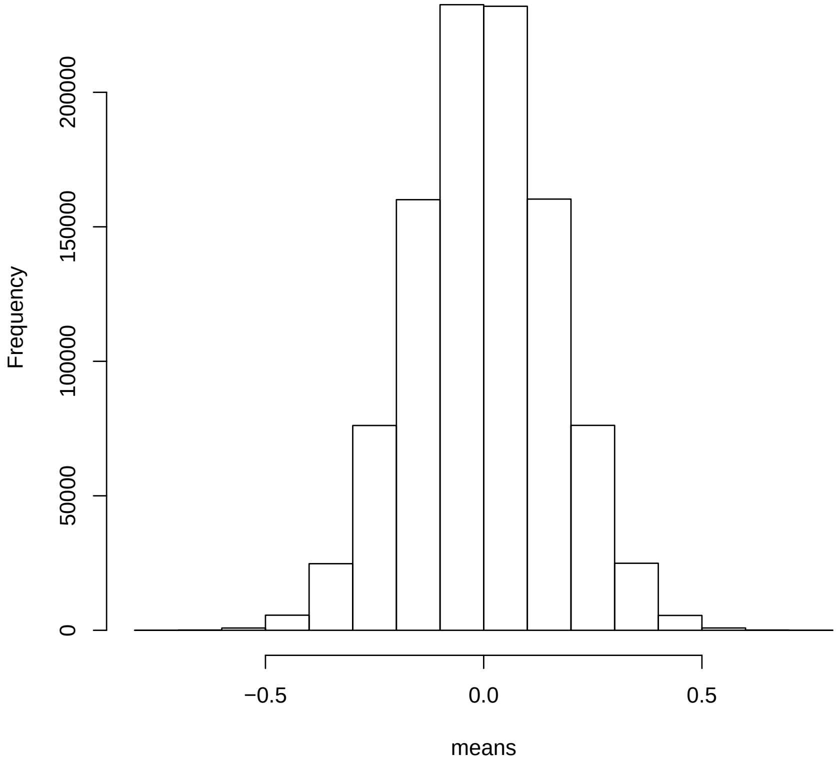

Figure 3: Histogram of the Sampling Distribution of the Mean

Problem 2

For binary outcome data, the median almost certainly equals 0 or 1.

Example:

Let D be a Bernoulli distribution with parameter, p. Consider any finite sample, $(x_1, \ldots, x_n)$, of n IID draws from D. Suppose that we calculate the median of this sample.

If n is odd, the median is always a specific draw from this distribution, which is either 0 or 1.

If n is even, the median is either 0, 0.5, or 1. Importantly, the median is only equal to 0.5 if the number of 0's and 1's in the data is exactly balanced. This is fairly rare event, especially as p gets further from 0.5.

Problem 3

The median is not a function of all parameters of the data generating process.

Example 3:

Consider the following mixture model:

- We draw 99 values from a uniform distribution over the interval $[0, 1]$.
- We draw 1 value from a right-shifted exponential, defined as $y \sim X + 10$, where X is an exponential distribution with mean t.

In this case, the sample distribution of the median is independent of t.

Consider the contrast between the sampling distributions of the median and the mean for two source distributions: one with $t = 1$ and another with $t = 1000$.

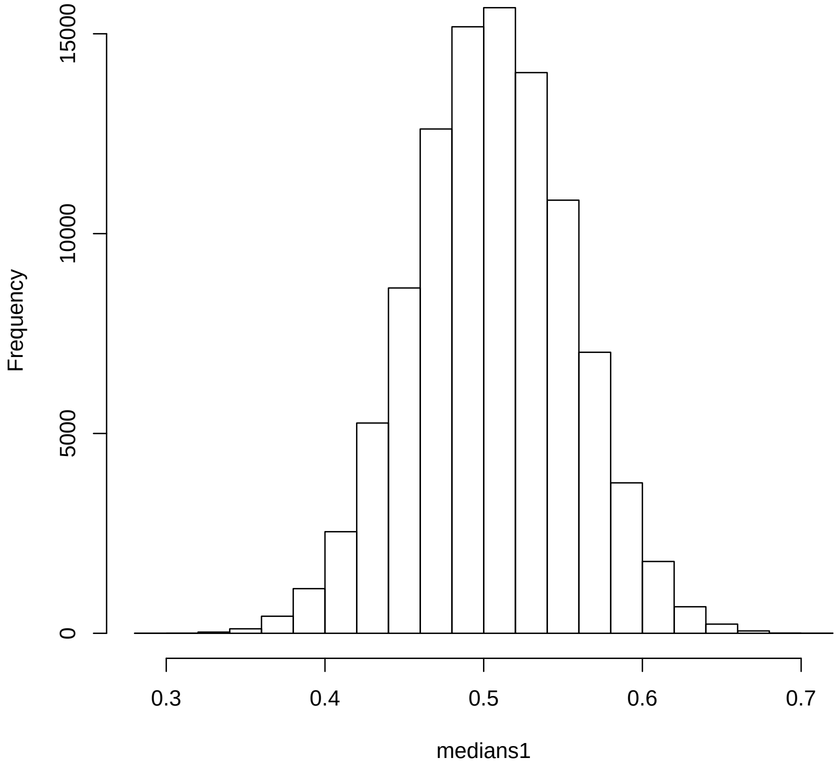

Figure 4: Sampling Distribution of Medians: $t = 1$

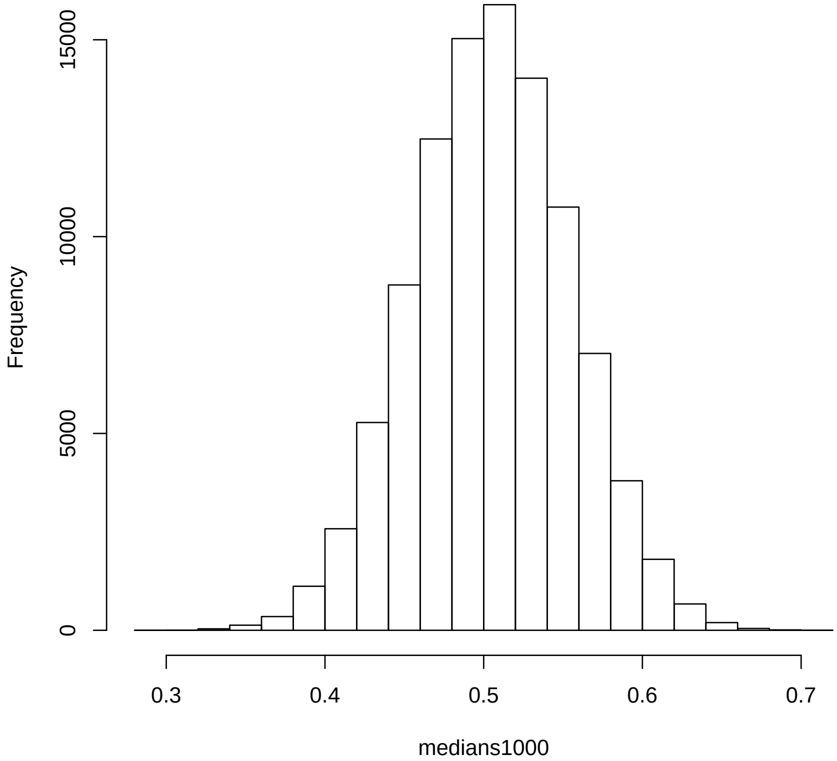

Figure 5: Sampling Distribution of Medians: $t = 1000$

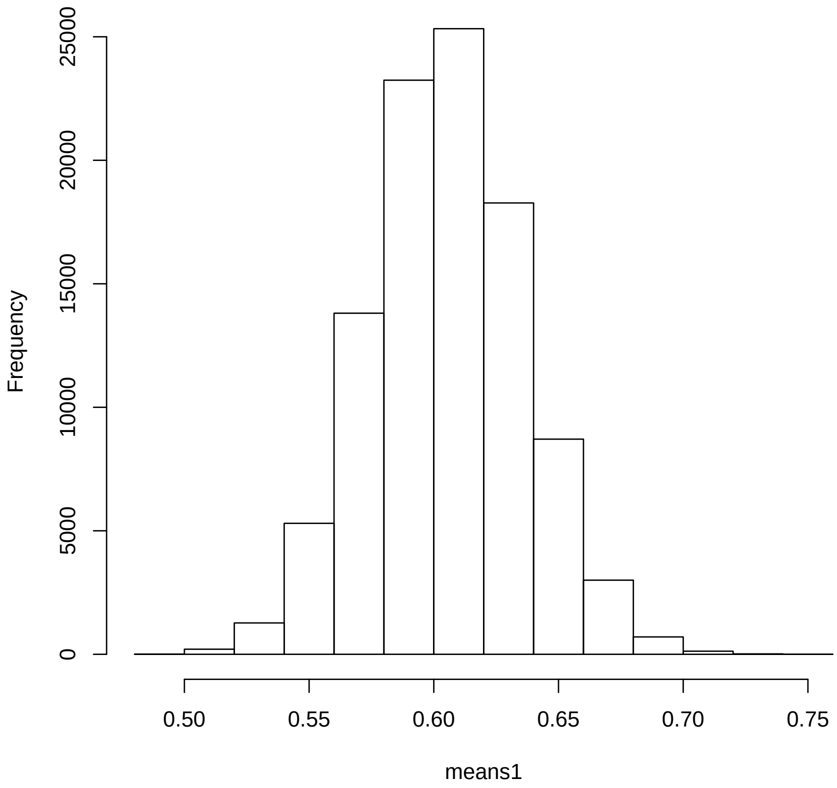

Figure 6: Sampling Distribution of Means: $t = 1$

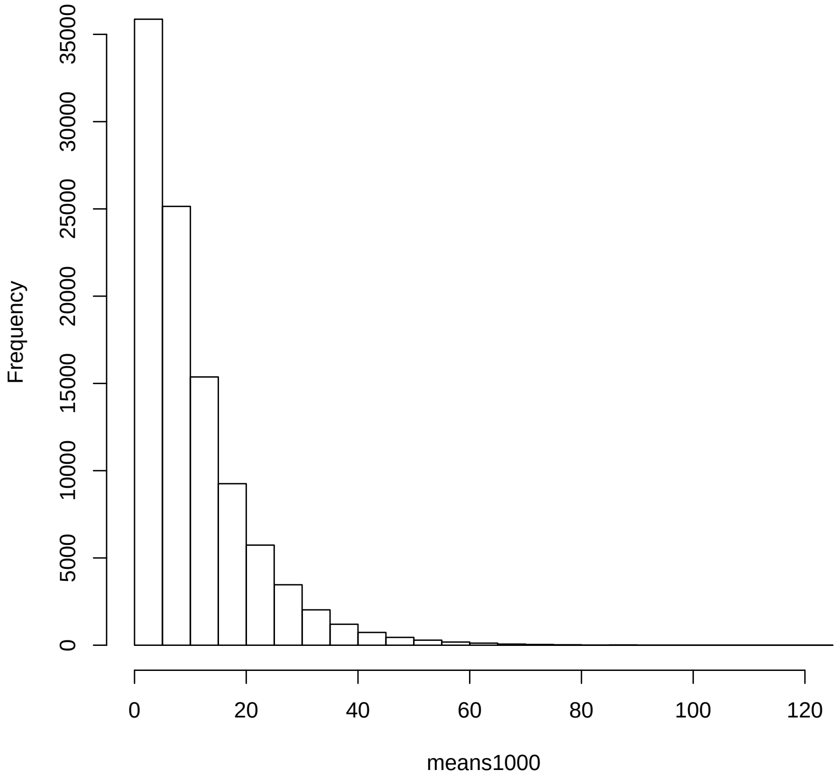

Figure 7: Sampling Distribution of Means: $t = 1000$