Stat 20: Discussion at section

B. M. Bolstad, bolstad@stat.berkeley.edu

Sept 29, 2003

Estimating the sampling distribution of a proportion

The sampling distribution of a proportion tells us how likely it is to see a certian proportion when we take a sample of a certain size from a population. We will run a simulation:

1. Toss a coin 10 times, record $\hat{p}_H$, the estimate of the proportion of heads you will get tossing a coin.
2. Repeat step 1 a lot of times.
3. Toss a coin 20 times, record $\hat{p}_H$, the estimate of the proportion of heads you will get tossing a coin.
4. Repeat step 3 a lot of times.
5. Toss a coin 100 times, record $\hat{p}_H$, the estimate of the proportion of heads you will get tossing a coin.
6. Repeat step 5 a lot of times.

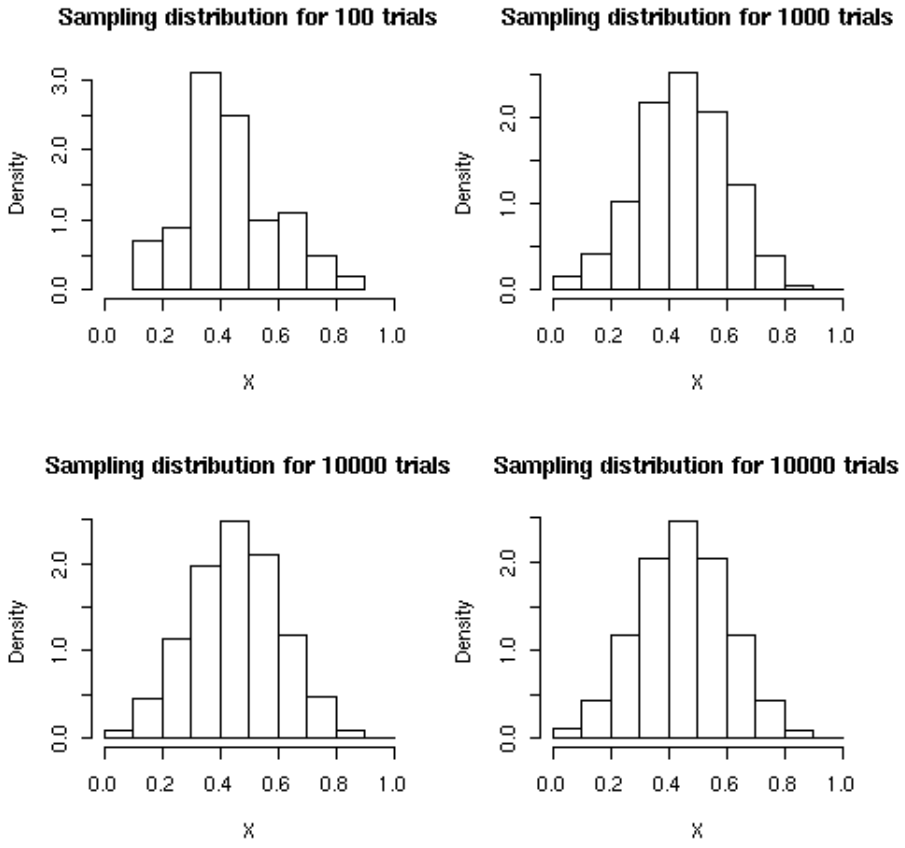
Some things to think about

1. What is p_H? How does it differ from $\hat{p}_H$?
2. How do we estimate $\hat{p}_H$?
3. What happens to the sampling distribution as n, the sample size, gets larger?
4. What happens as the number of trials gets larger?
5. Where is the center of the distribution?

Simulation Results when $n = 10$

Rather than tossing a real coin, I will use a computer to simulate coin tosses with probability of heads 0.5 the same as a real fair coin. In addition I can in seconds simulate 1 million trials.

First we plot histograms as we progressively run a larger number of trials. One thing to note is that as we run more and more trials the sampling distribution seems to stablize and become symetrical.

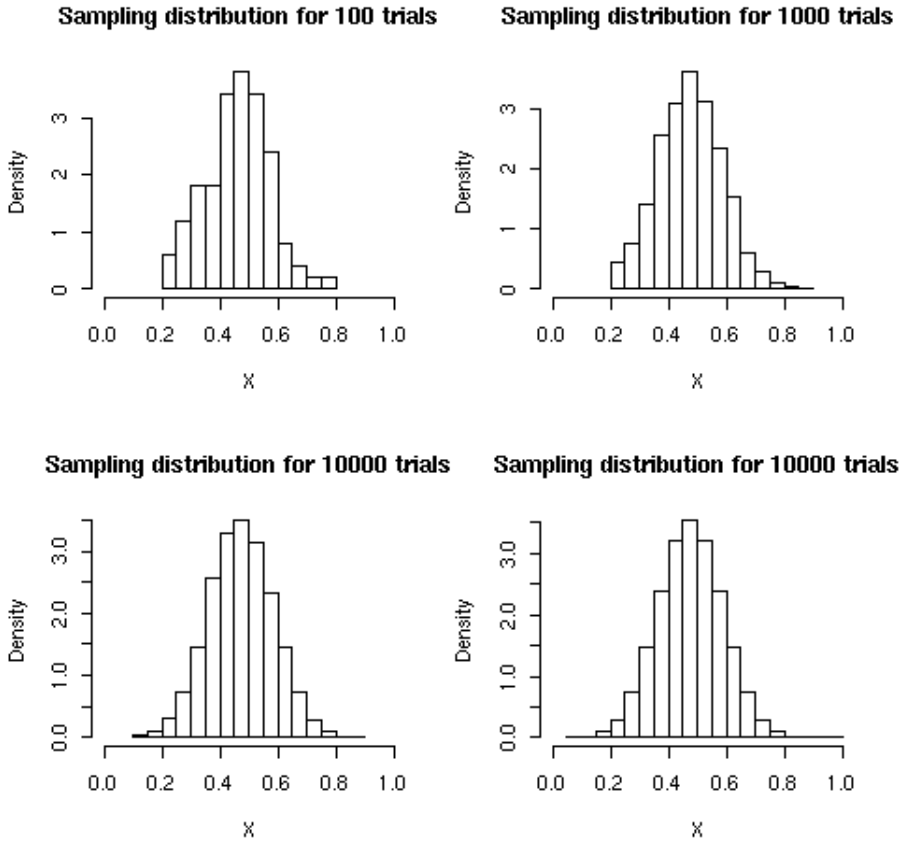


We also look the mean and variance of our observed $\hat{p}_H$ in this table. Note that the mean seems to be stabilizing towards 0.5 and the variability towards 0.0250 as the number of trials increases

Number of trials	Mean $\hat{p}_H$	Variance $\hat{p}_H$
100	0.4840	0.0268
1000	0.4990	0.0243
10000	0.5029	0.0250
1000000	0.4999	0.0250

Simulation Results when $n = 20$

We again plot histograms as we progressively run a larger number of trials. One thing to note is that as we run more and more trials the sampling distribution seems to stablize and become symetrical. The distribuions look a little less speread out than when $n = 10$. The table shows that the mean was again converging to 0.5 and the variance to 0.0125.

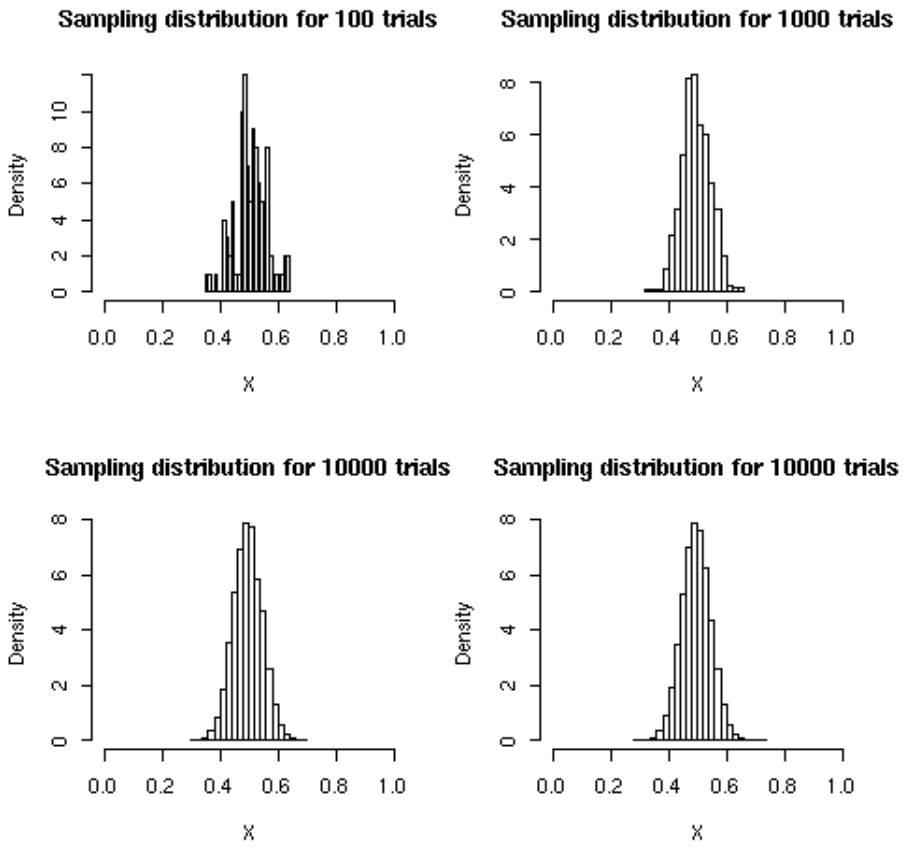


Now we look at the observed means and variances we see that the mean is converging to 0.5 and the variance is converging to 0.0125.

Number of trials	Mean $\hat{p}_H$	Variance $\hat{p}_H$
100	0.4845	0.0129
1000	0.4990	0.0126
10000	0.4980	0.0124
1000000	0.5000	0.0125

Simulation Results when $n = 100$

Finally, lets look at the histograms for $n = 100$. These seem to be much less spread than the earlier cases. Again, a larger number of trials gave us a smoother, more symetric histogram.



Number of trials	Mean $\hat{p}_H$	Variance $\hat{p}_H$
100	0.5097	0.0031
1000	0.4995	0.0024
10000	0.5004	0.0025
1000000	0.5000	0.0025

Summary of results

We saw that a larger number of trials gave smoother histograms. Increasing the sample size reduced the spread of the histogram and we saw the variance was smaller in each case. $\hat{p}_H$ seemed to center around 0.5 which is the theoretical probability (for a fair coin). We will later see that

$$E[\hat{p}_H] = p_H$$

and the standard error

$$SE(\hat{p}_H) = \sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n}}$$