Probability and Mathematical Statistics in Data Science

Lecture 19: Section 7.2: Sampling without Replacement
Section 7.3: Law of Averages

Sampling Without Replacement

- The draws in a simple random sample aren't independent of each other.
- This makes calculating variances a little less straightforward than in the case of draws with replacement.
- We will find the variance of a random variable that has a hypergeometric distribution.

Variance of a hypergeometric random variable

- Let $X \sim HG(N, G, n)$, then can write $X = I_1 + I_2 + \cdots + I_n$, where I_k is the indicator of the event that the kth draw is good.

- We can compute the expectation of X using symmetry: $E(X) = \frac{nG}{N}$

- But what about variance?

- Since the indicators are not independent, we can't just add the variances

- We can use the formula: $Var(X) = E(X^2) - \left(\frac{nG}{N}\right)^2$

Variance of a hypergeometric random variable

After a little manipulation this becomes

$$Var(X) = n\frac{G}{N} \cdot \frac{N-G}{N} \cdot \frac{N-n}{N-1}$$

The initial part of this formula is the binomial variance npq. To see this more clearly, write $B = N - G$ for the number of bad elements. Then

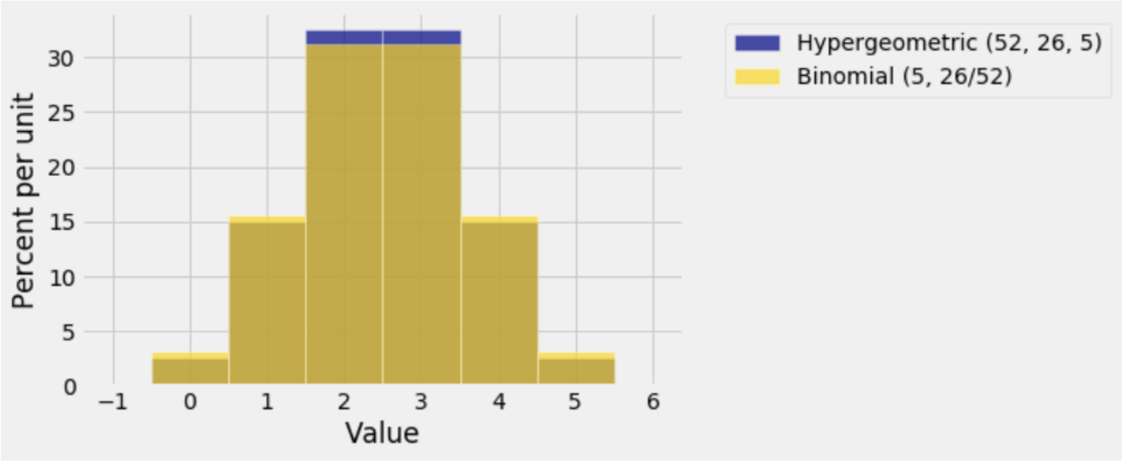
$$Var(X) = (n\frac{G}{N} \cdot \frac{B}{N})\frac{N-n}{N-1}$$

and

$$SD(X) = \sqrt{n\frac{G}{N} \cdot \frac{B}{N}}\sqrt{\frac{N-n}{N-1}} = \sqrt{npq}\sqrt{\frac{N-n}{N-1}}$$

for $p = \frac{G}{N}$.

The finite population correction (fpc) & the accuracy of SRS



Fpc = $\sqrt{\frac{N-n}{N-1}}$

Note that fpc ≤ 1

So SD(HG) $\leq$ SD(Bin)

In general we have that the :

SD of sum of an SRS = SD of sum WITH repl. $\times$ fpc

The finite population correction (fpc) & the accuracy of SRS

- Sampling with and without replacement are essentially the same when the sample size is small relative to the population size. We now have another confirmation of this.

- When the sample size is small relative to the population, the finite population correction is close to 1. That is because

$$\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1} \approx 1 - \frac{n}{N} \approx 1$$

when $\frac{n}{N}$ is small.

The Accuracy of Simple Random Samples

- Suppose a poll is based on a simple random sample drawn from a huge population of voters of whom a proportion p favor a politician. Then the SD of the number of voters who favor the politician is

$$\sqrt{npq}\sqrt{\frac{N-n}{N-1}} \approx \sqrt{npq}$$

- because the fpc is close to 1.

Accuracy of samples

Simple random samples of the same size of 625 people are taken in Berkeley (population: 121,485) and Los Angeles (population: 4 million).

True or false, and explain your choice: The results from the Los Angeles poll will be substantially more accurate than those for Berkeley.

Law of Averages

- Essentially a statement that you are already familiar with: If you toss a fair coin many times, roughly half the tosses will land heads.

- We are going to consider sample sums and sample means of iid random variables $X_1, X_2, \dots, X_n$ where the mean of each X_k is μ and the variance of each X_k is σ^2.

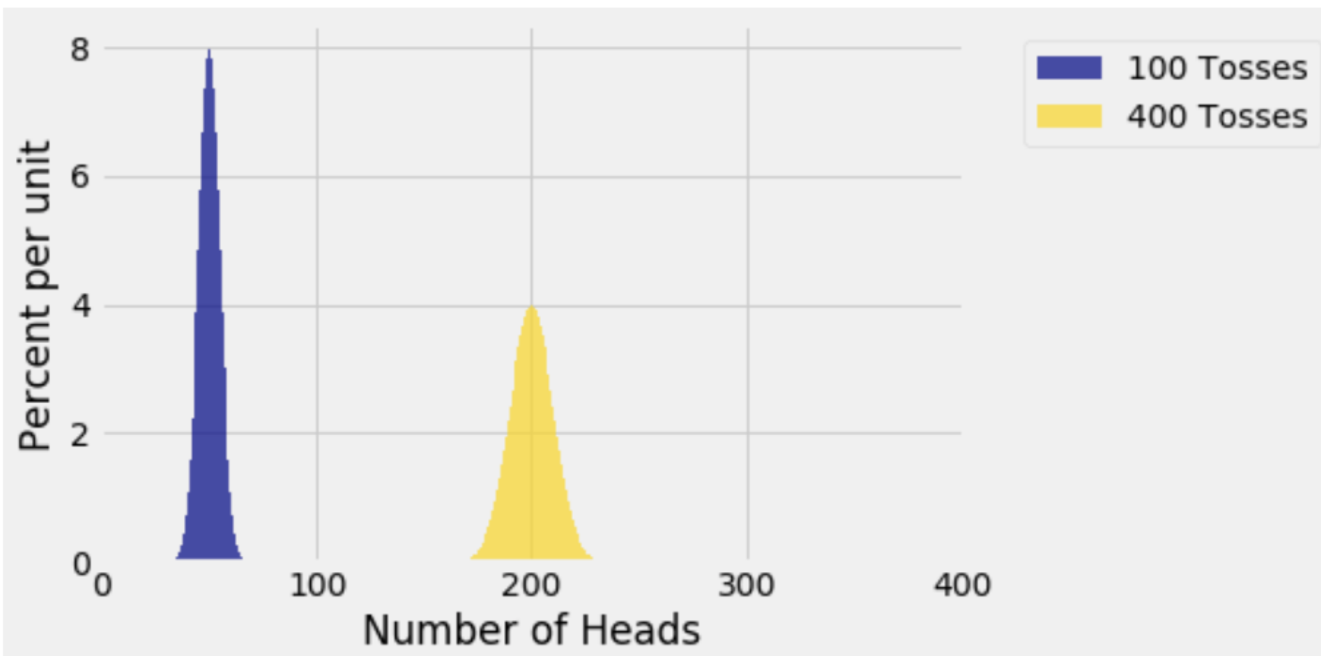
- Define the ***sample sum*** $S_n = X_1 + X_2 + \cdots + X_n$, then $E(S_n) = n\mu, Var(S_n) = n\sigma^2, SD(X_n) = \sqrt{n}\sigma$

- We see here, as we take more and more draws, their sum's variability keeps increasing, which means the values get more and more dispersed around the mean ($n\mu$).

Coin tosses

- Consider a fair coin, toss it 100 times & 400 times, count the number of H
Expect in first case, roughly 50 H, and in second, roughly 200 H.

- So do you think chance of 50 H in 100 tosses and 200 H in 400 tosses should be the same?



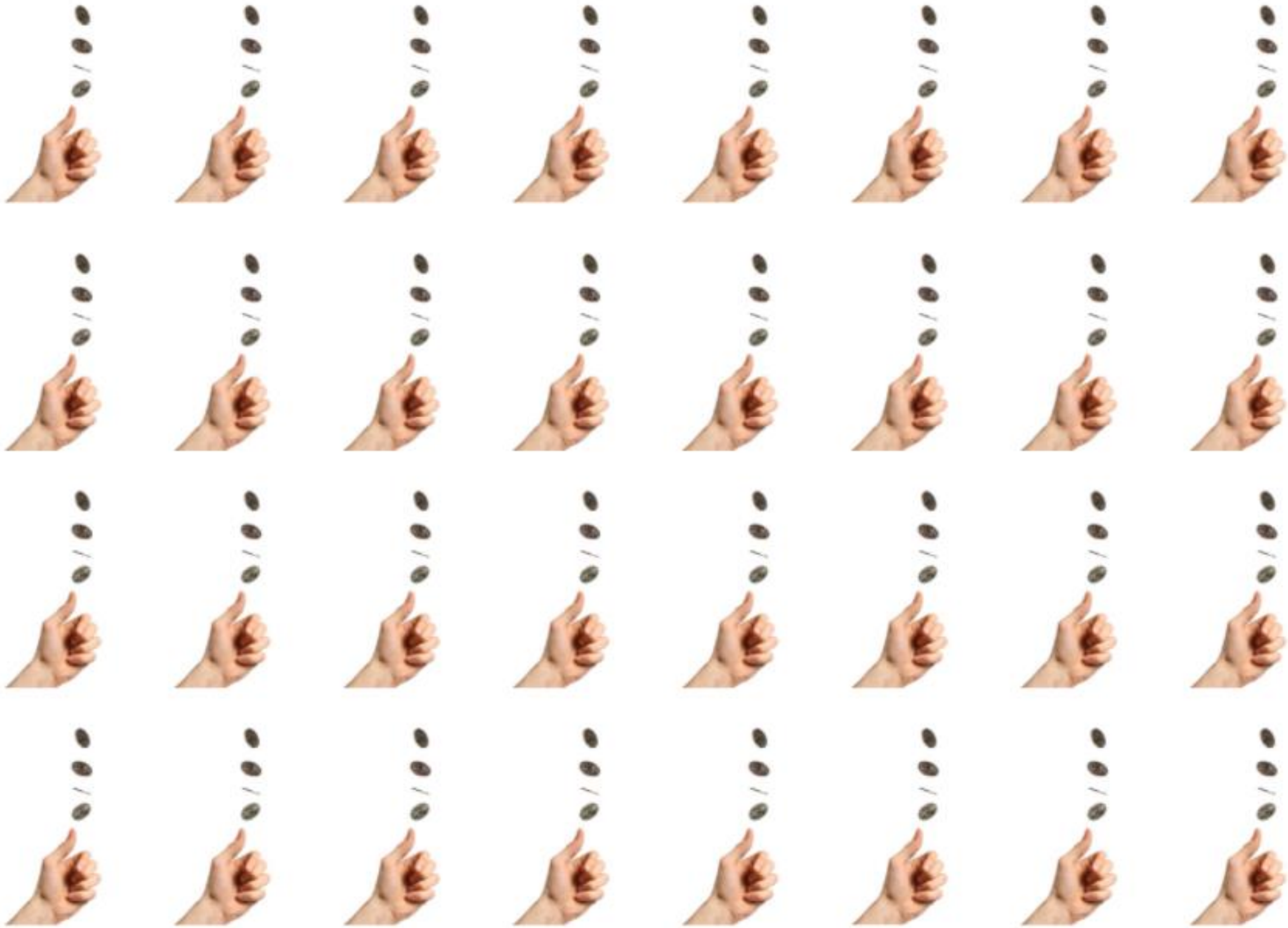
Example: Coin toss

- $SD(S_{100}) =$
- $SD(S_{400}) =$

- P(200 H in 400 tosses)

- P(50 H in 100 tosses)

Law of Averages



Law of Averages

# of heads we should observe ≈ ½ # of tosses + **chance error**

% of heads should be ≈ 50% + **percent error**

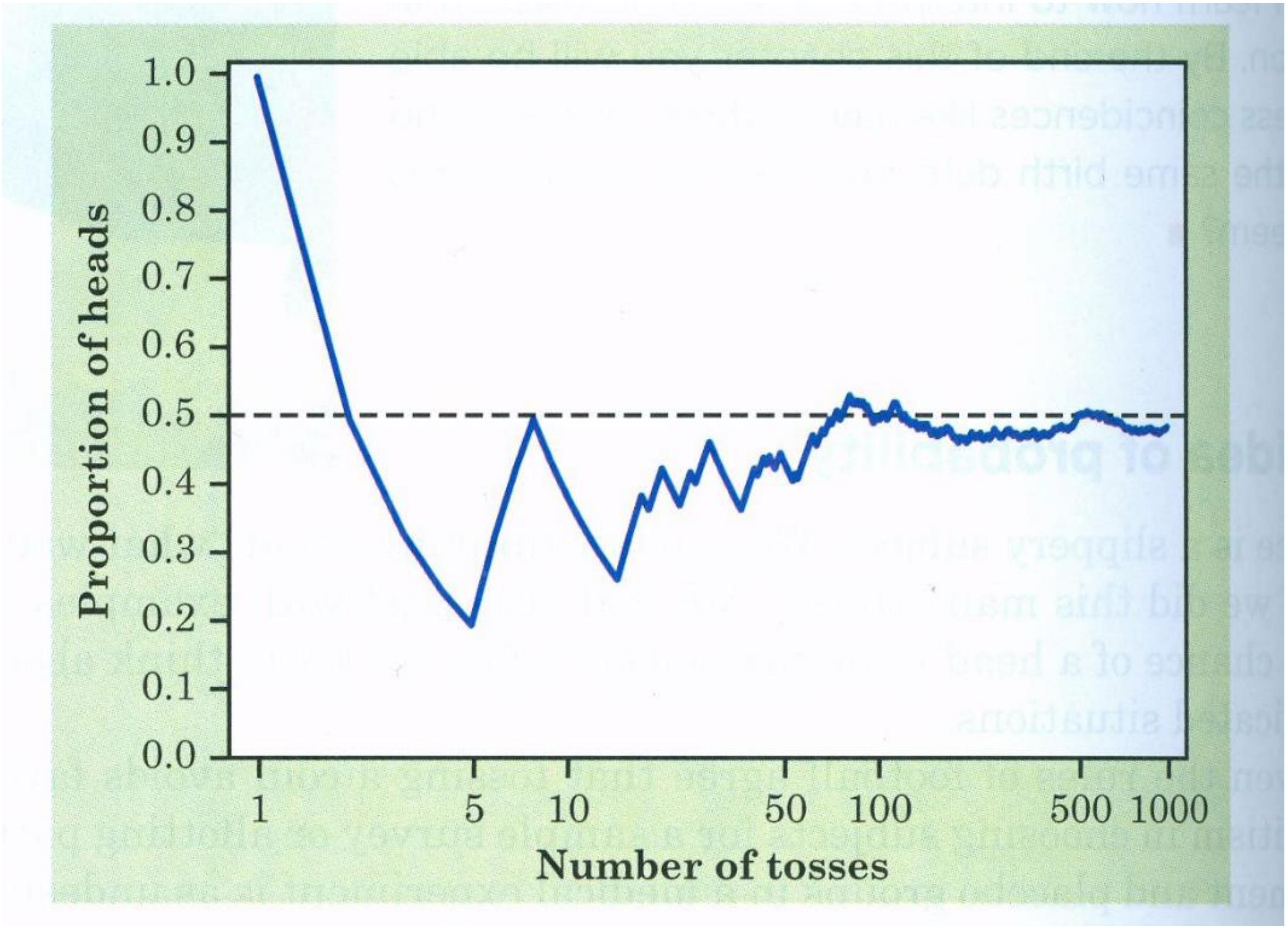
(chance error in percentage terms)

Law of Averages

# tosses	# heads	# expected heads	Chance error	% heads	% expected heads	Percent error
10	4	5	-1	40%	50%	-10%
50	25	25	0	25%	50%	0%
100	44	50	-6	44%	50%	-6%
500	255	250	5	51%	50%	1%
1000	502	500	2	50.2%	50%	0.2%
5000	2533	2500	33	50.66%	50%	0.66%
10000	5067	5000	67	50.67%	50%	0.67%

Note: In a large number of tosses, the percent error will be small

Law of Averages



Law of Averages for a fair coin

- Notice that as the number of tosses of a fair coin increases, the *observed error* (number of heads – half the number of tosses) increases. This is governed by the standard deviation.

- The *percentage* of heads observed comes very close to 50%

- *Law of averages*: The long run *proportion* of heads is very close to 50%.

Sample sum, sample average, and the square root law

- $S_n = X_1 + X_2 + \cdots + X_n$
- Let $A_n = {}^{S_n}/_n$, so $A_n$is the average of the sample (or sample mean).
- If the X_k are indicators, then A_n is a proportion (proportion of successes)
- Note that $E(A_n) = \mu$ and $SD(A_n) = \sigma/\sqrt{n}$
- **The square root law:** the *accuracy* of an estimator is measured by its SD, the ***smaller*** the SD, the ***more accurate*** the estimator, but if you multiply the sample size by a factor, the accuracy only goes up by the **square root** of the factor.
- In our earlier example, we ______ the accuracy by quadrupling the size.

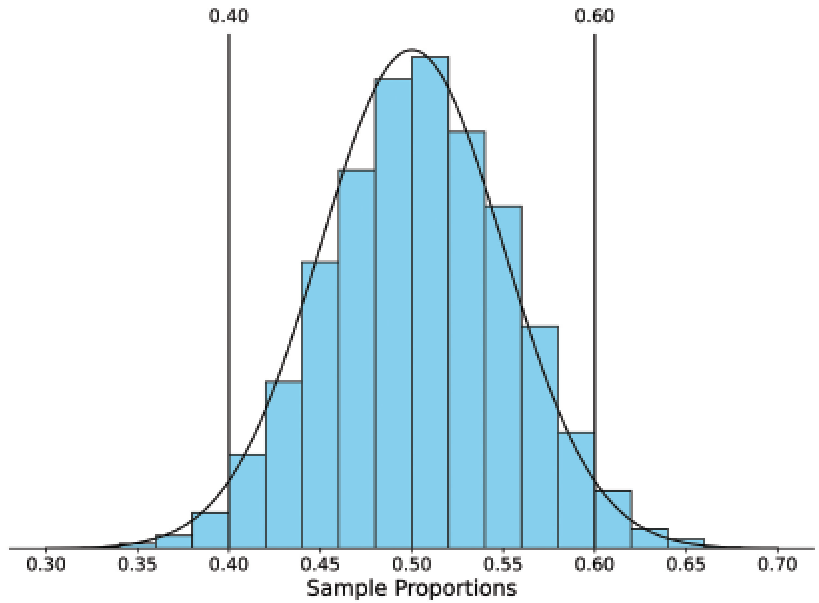
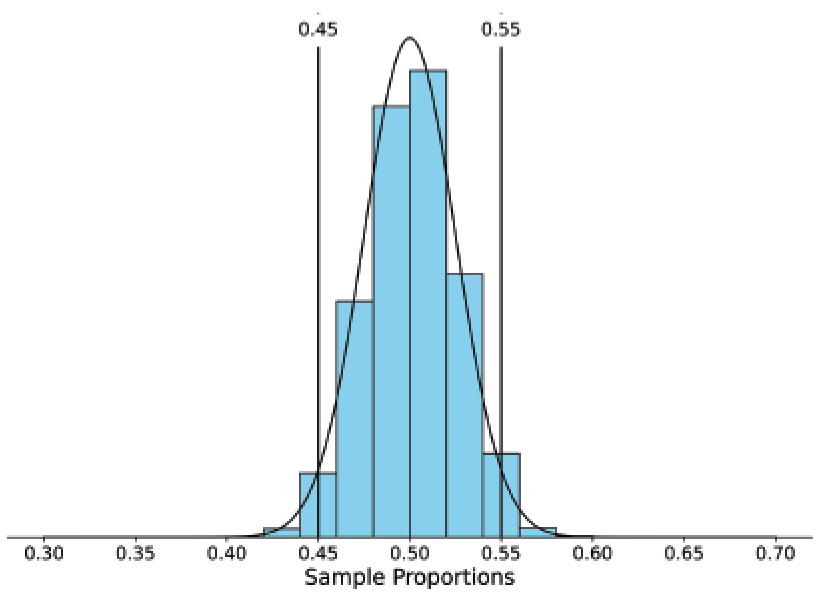


Figure 6.7: Simulation of 10,000 Sample Proportions of Heads (Based on 100 Coin Tosses)



Law of averages

- The law of averages says that if you take enough samples, the proportion of times a particular event occurs is very close to its probability.

- In general, when we repeat a random experiment such as tossing a coin or rolling a die over and over again, the average of the observed values will come the expected value.

- *Law of averages*: The individual outcomes when averaged get very close to the theoretical average (expected value)