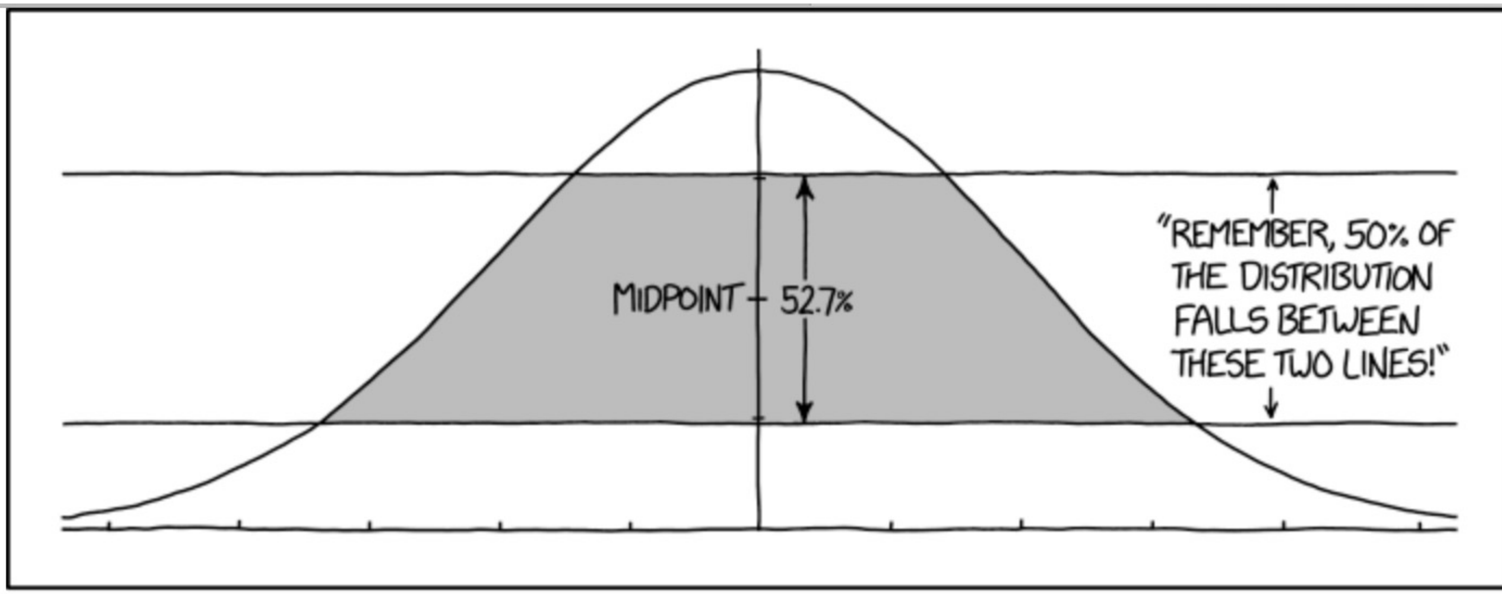
Stat 88: Prob. & Math. Statistics in Data Science



HOW TO ANNOY A STATISTICIAN

xkcd.com/2118

Lecture 28: 4/1/2024

The law of averages, distribution of a sample sum

7.3, 8.1, 8.2

$S = S_n$, $A_n = \bar{X}$

Story so far…

$X_1, X_2, \ldots X_n$ are iid r.v.

$S = X_1 + X_2 + \ldots + X_n$

$E(X_k) = \mu$, $Var(X_k) = \sigma^2$

$E(S) = n\mu$

$S = S_n$, $A_n = \bar{X}$

$A_n = \frac{S}{n}$

- Variance and SD of sums of iid random variables:

$Var(S_n) = n\sigma^2$ $\boxed{SD(S_n) = \sqrt{n}\,\sigma}$

- Variance of a Binomial rv

$\boxed{Var(X) = npq = np(1-p)}$

$X_k \sim Bernoulli(p)$

$X \sim Bin(n, p)$

$E(X) = np$

- SD of a sample ***sum*** increases with *n*, whereas the SD of a sample ***mean*** decreases with *n*.

$\boxed{SD(A_n) = \frac{\sigma}{\sqrt{n}}}$

$SD(S_n) = \sqrt{n} \cdot \sigma$

$SD(A_n) = SD\left(\frac{S_n}{n}\right) = \frac{1}{n} SD(S_n) = \frac{\sqrt{n}\,\sigma}{\sqrt{n}}$

$Var(X) = E(X^2) - E(X)^2$

$X \sim Ber(p)$

$X^2 \sim Ber(p)$

$Var(X) = p - p^2$

- When we have a simple random sample (SRS), the draws are without replacement (like drawing cards from a deck).

- Variance of hypergeometric rv:

$X \sim HG(N, G, n)$

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

- Finite population correction:

$$\boxed{f.p.c = \sqrt{\frac{N-n}{N-1}}}$$

SD of sum of iid r.v. * f.p.c

= SD of sum of SRS

Suppose X is sum of draws of tickets from a box.

SD of sum WITHOUT REPL = SD of SUM w/ Repl * fpc

Accuracy of samples (depend on the SD of the sample mean/sum)

- 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.

Fpc in case of Berkeley: 0.9974285

Fpc in case of LA: 0.999922

Example adapted from Statistics, by FPP

- A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2022 Oscars. To keep costs down, they want to take as small a sample as possible, but their client will only tolerate a random error of 1 percentage point or so in the estimate. Should they use a sample size of 100, 2500, or 10000? The population is very large and the fpc is about 1.

What n to use? Note that the number of people who have watched the Oscars in the sample is a rv with the $HG(N, G, n)$ distiribution.

Note that N is very large so can pretend that we are sampling w/ replacement.

Pretend that we are sampling w/replacement

X = # of people in sample that watched the Oscars

$X \sim \text{Bin}(n, p)$ (fpc ≈ 1)

$\approx$ approximately

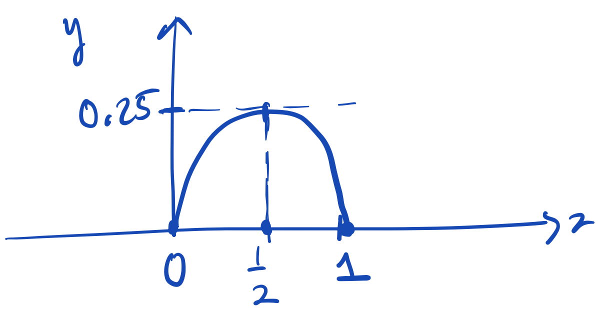
Percent of people in sample that watched Oscars $= \dfrac{X}{n}$

$\mathbb{E}(X) = np$ $\quad$ $\text{Var}(X) = np(1-p)$

$SD(X) = \sqrt{np(1-p)}$

$$SD\left(\frac{X}{n}\right) = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq 0.01$$

$f(x) = x(1-x)$, $0 \leq x \leq 1$



$x(1-x) \leq 0.25$

$\sqrt{x(1-x)} \leq 0.5$

$$SD\left(\frac{X}{n}\right) \leq \frac{0.5}{\sqrt{n}} \leq 0.01$$

Therefore, what can we say about n?

$$\frac{0.5}{0.01} \leq \sqrt{n}$$

$$50 \leq \sqrt{n}$$

$$2500 \leq n$$

Example (adapted from *Statistics*, by Freedman, Pisani, and Purves)

- Note that the number of people who have watched the Oscars in the sample is a rv with the $HG(N, G, n)$ distribution, but we are told that N is very large & $fpc \approx 1$, so we can approximate the prob. using the $Bin(n, p)$ distribution, where p is the percentage of people who watched the Oscars (which is what we are trying to estimate).

- $SD\left(\frac{S_n}{n}\right) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{pq}}{\sqrt{n}} \leq \frac{0.5}{\sqrt{n}} \leq 0.01 \Rightarrow n \geq 2500$

Exercise 7.4.11

Exercise

Each Data 8 student is asked to draw a random sample and estimate a parameter using a method that has chance 95% of resulting in a good estimate.

Suppose there are 1300 students in Data 8. Let X be the number of students who get a good estimate. Assume that all the students' samples are independent of each other.

- a) Find the distribution of X

- b) Find $E(X)$ and $SD(X)$.

- c) Find the chance that more than 1250 students get a good estimate.

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 $X_1, X_2, \ldots, X_n$ where the mean of each X_k is μ and the variance of each X_k is σ^2.

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

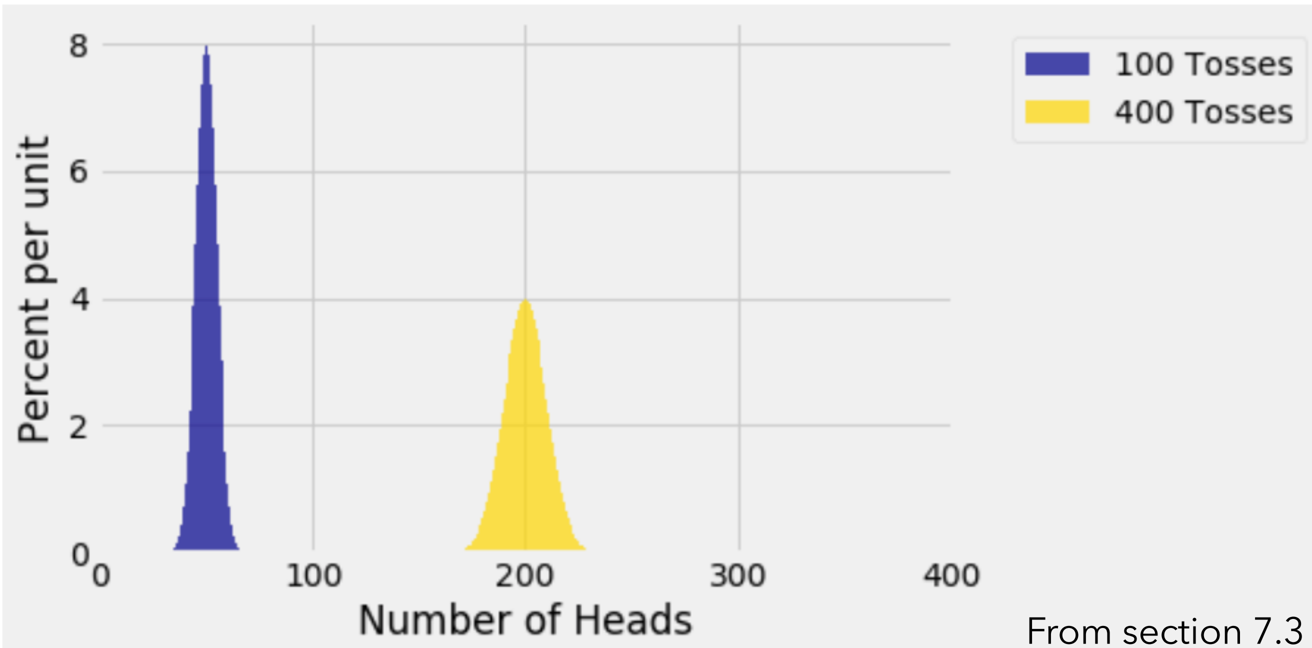
  fixed

- We see here, as we take more and more draws, the variability of the sum 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?



From section 7.3

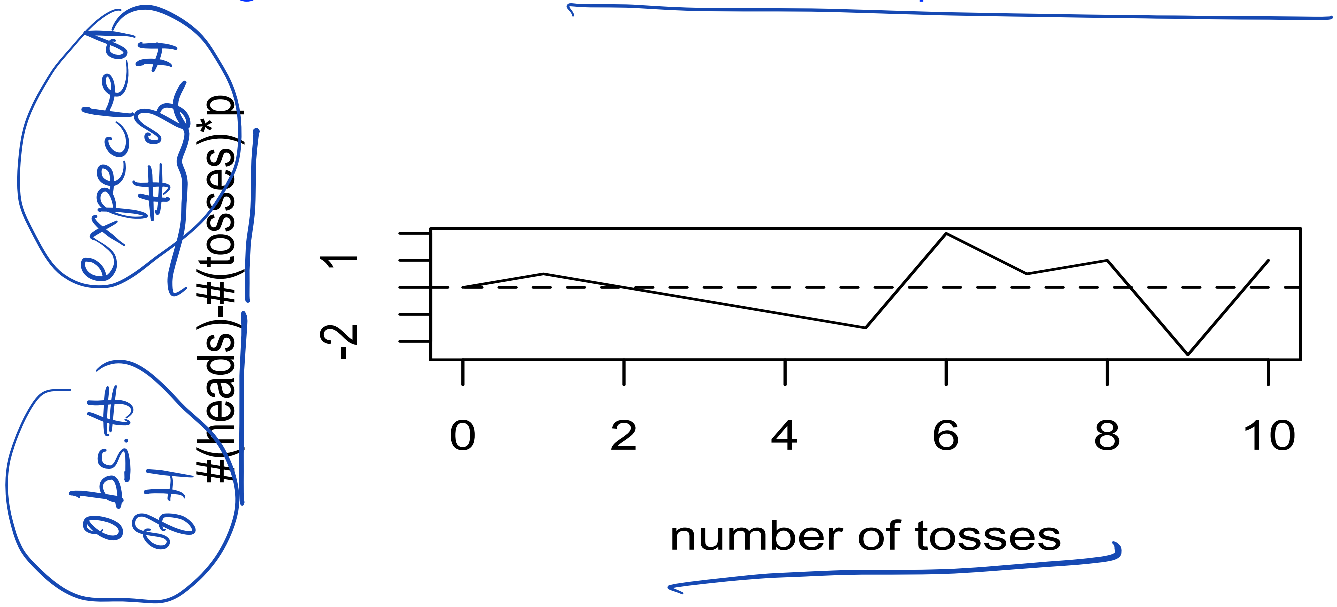
Example: Coin toss

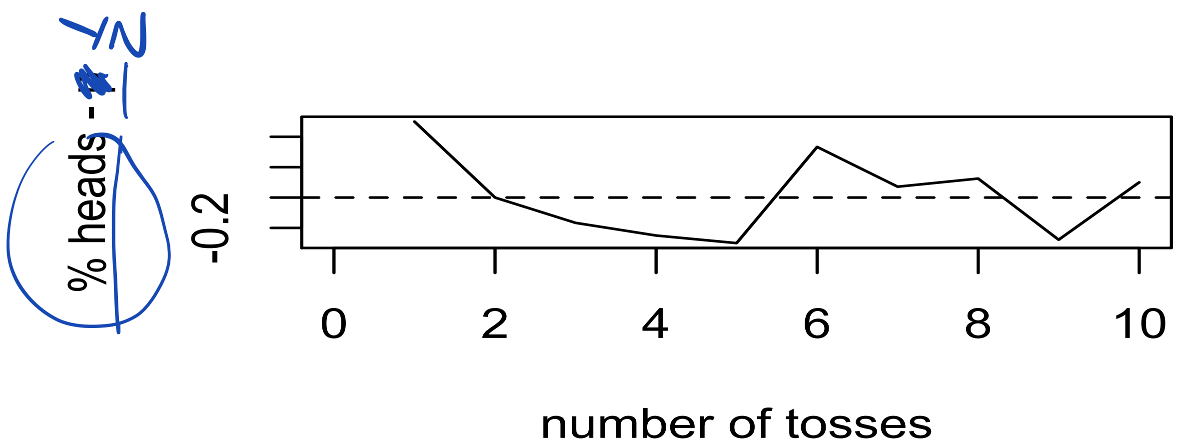
- $SD(S_{100}) =$
- $SD(S_{400}) =$

- P(200 H in 400 tosses)

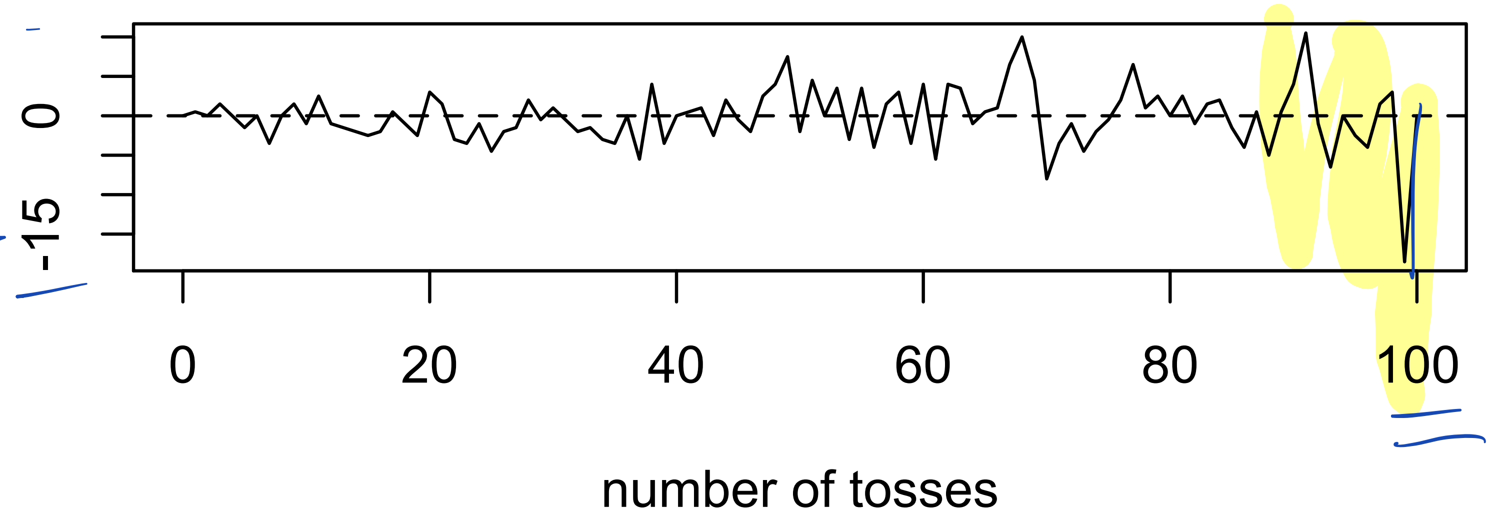
- P(50 H in 100 tosses)

Simulating coin tosses: 10 tosses (adapted from FPP)

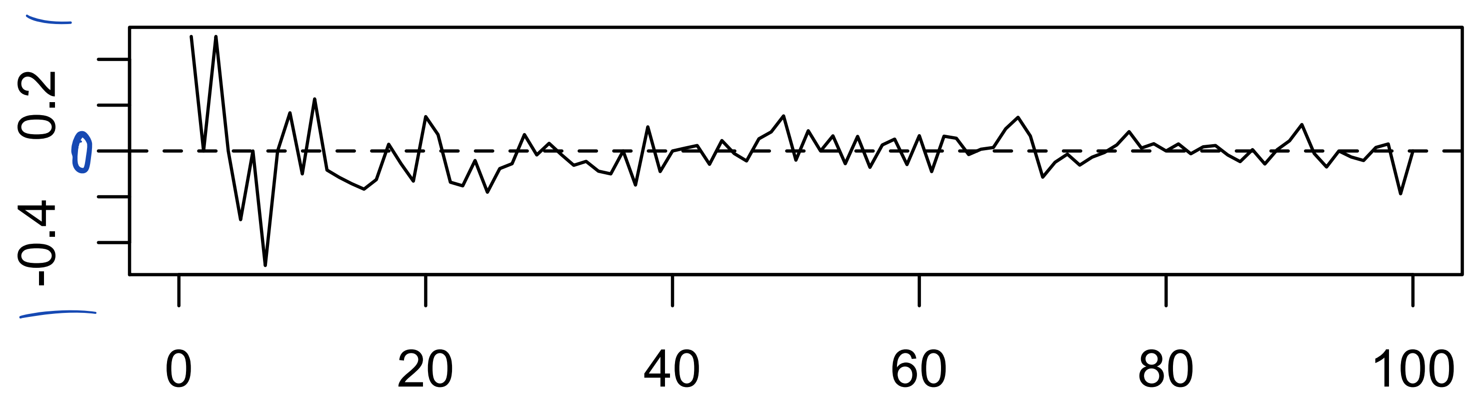


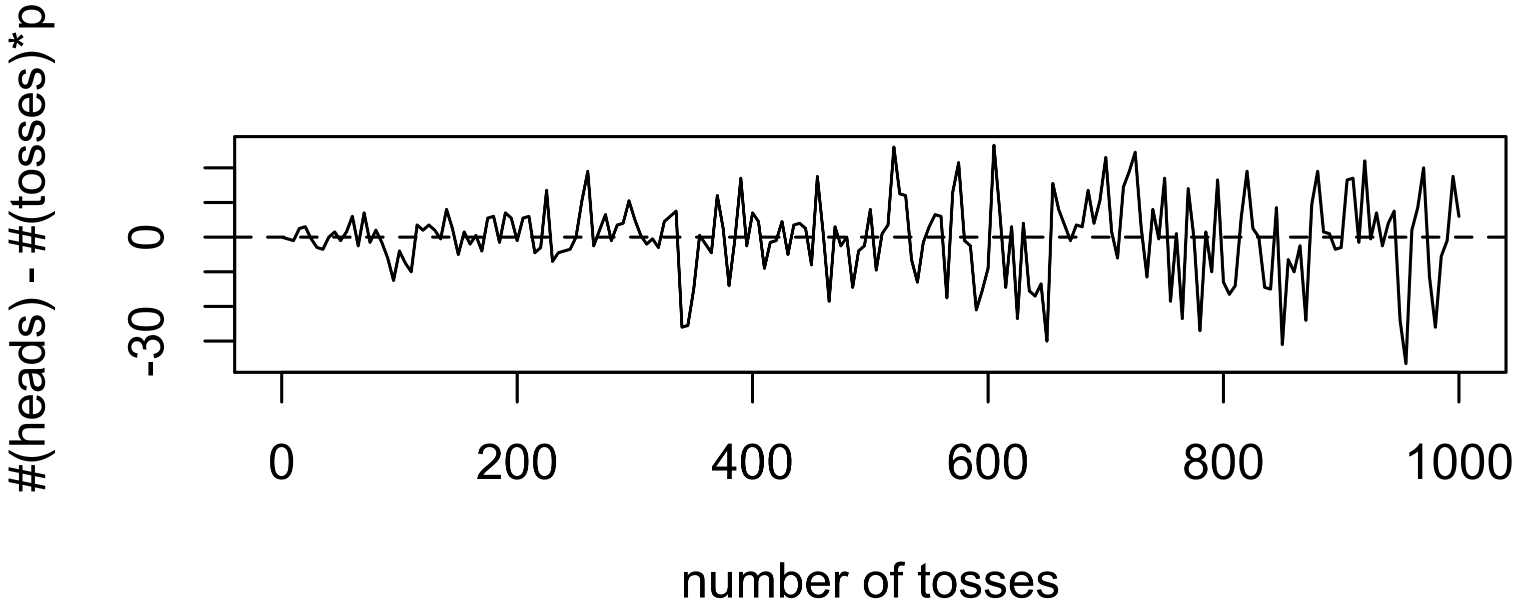


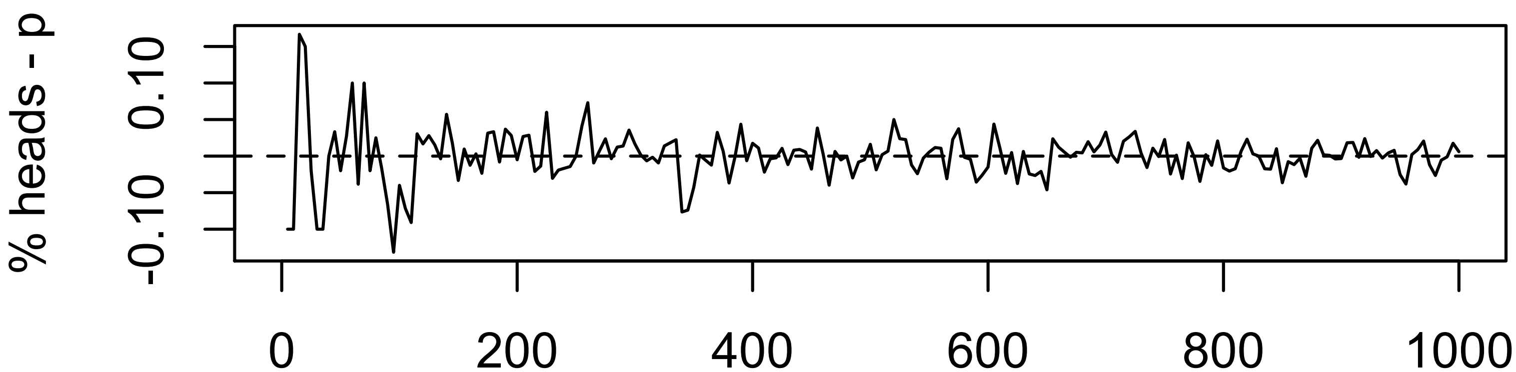
Observed error= #H- #tosses/2

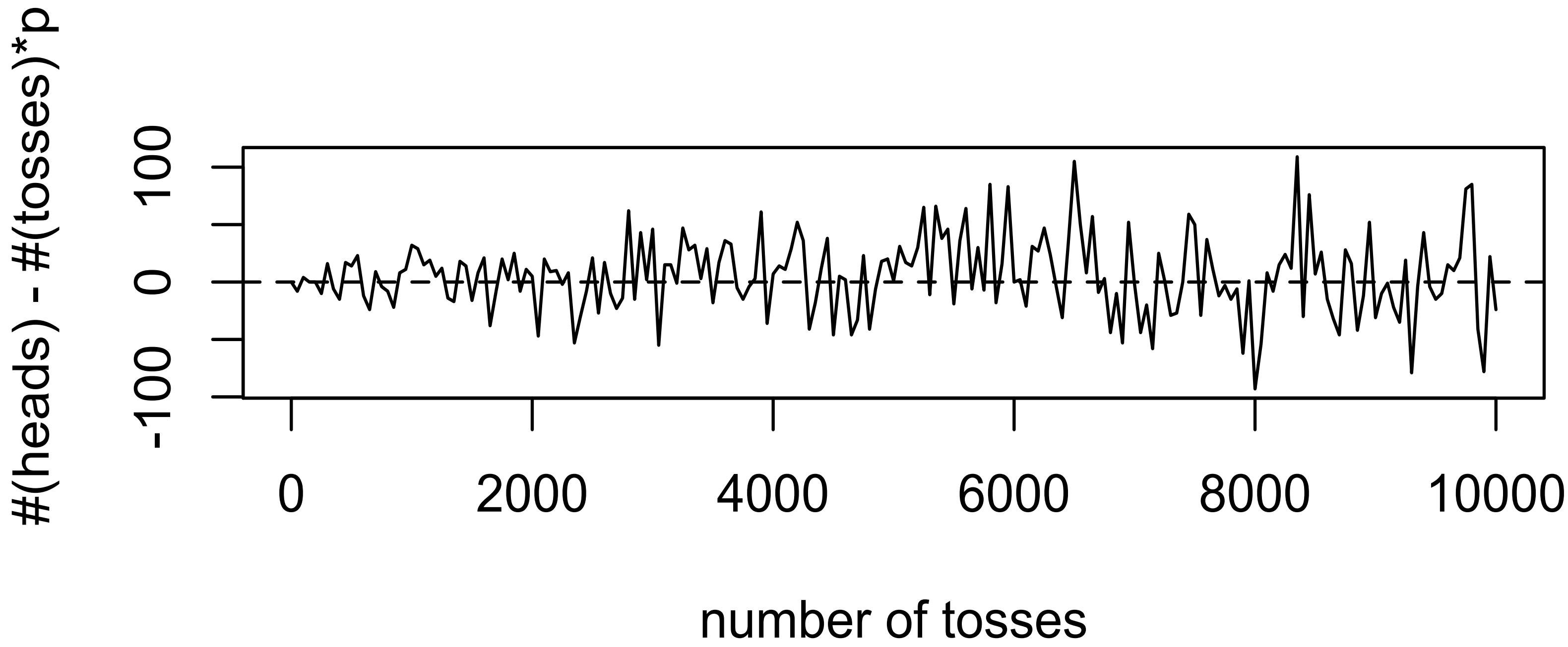


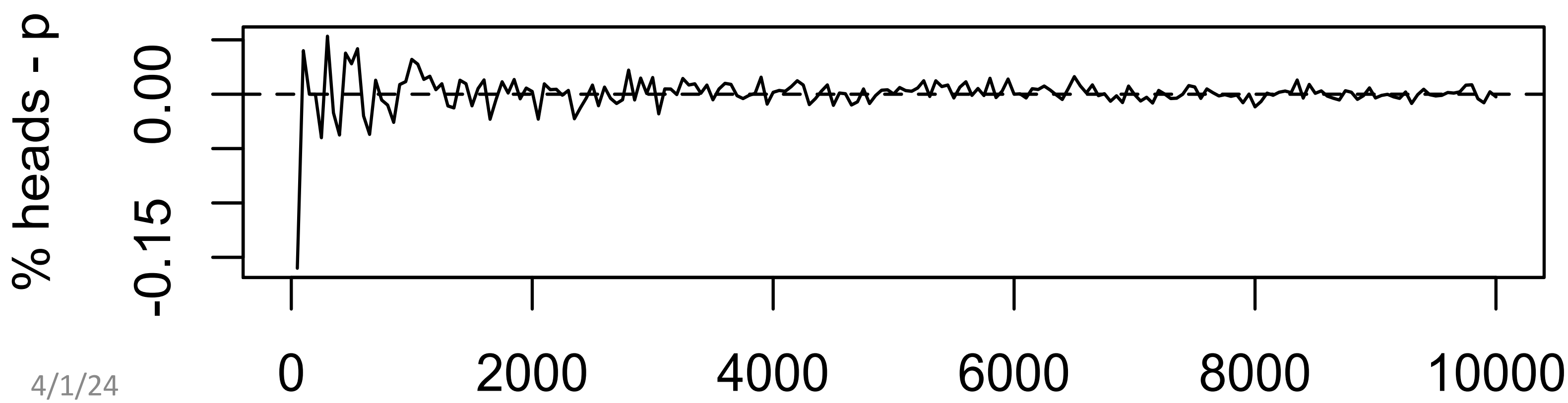
% error= %H-0.5




#(heads) - #(tosses)*p
-30
0
0
200
400
600
800
1000
number of tosses


% heads - p
-0.10
0.10
0
200
400
600
800
1000


#(heads) - #(tosses)*p
100
0
-100
0
2000
4000
6000
8000
10000
number of tosses


% heads - p
0.00
-0.15
0
2000
4000
6000
8000
10000

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 error.

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

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