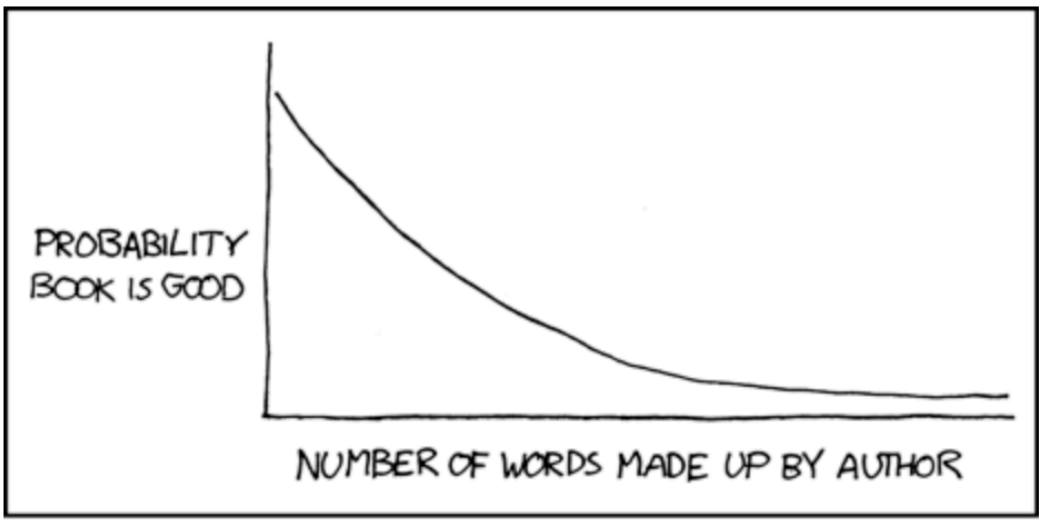
Stat 88: Probability & Mathematical Statistics in Data Science



xkcd.com/483

Lecture 24: 3/17/2021

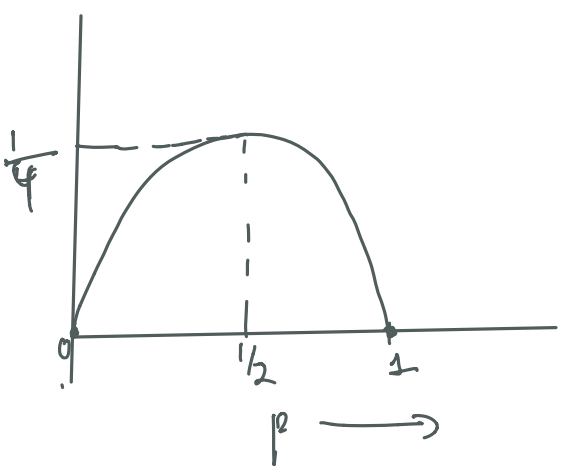
Sections 7.2, 7.3

Sampling without replacement and the Law of Averages

Last time:

- $X \sim Bin(n,p),\ Var(X) = np(1-p) = npq,\ SD(X) = \sqrt{npq}$
- $X \sim Pois(\mu),\ E(X) = Var(X) = \mu,\ SD(X) = \sqrt{\mu}$
- $X \sim Geom(p),\ E(X) = \frac{1}{p}, Var(X) = \frac{1-p}{p^2},\ SD(X) = \frac{\sqrt{1-p}}{p}$
- Consider $X \sim Bernoulli(p),\ Var(X) = p(1-p)$. For what p is the variance highest?

$$SD(X) = \sqrt{pq} = \sqrt{p(1-p)}$$
$$p - p^2$$



Upper bound on variance of a Bernoulli r.v. is $\frac{1}{4}$

(upper for SD is $\frac{1}{2}$)

Variance of a hypergeometric random variable

sum of draws from a 0-1 population

- 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
- Let's just use the formula: $Var(X) = E(X^2) - \left(\frac{nG}{N}\right)^2$
- $X^2 = (I_1 + I_2 + \cdots + I_n)^2 = \sum_{k=1}^{n} I_k^2 + \sum_j \sum_{k \neq j} I_j I_k$ # of pairs = $n(n-1)$

$$E(X^2) = nE\left(I_k^2\right) + n(n-1)E\left(I_j I_k\right) = n\frac{G}{N} + n(n-1)P\left(I_j = 1\right)P\left(I_k = 1 \mid I_j = 1\right)$$

$$E(X^2) = n\frac{G}{N} + n(n-1)\frac{G}{N}\cdot\frac{G-1}{N-1}$$

$$Var(X) = E(X^2) - \left(E(X)\right)^2$$

$\sum_{j=1}^{n} \sum_{k=1, k \neq j}^{n} I_j \cdot I_k$

$(I_1 + I_2)(I_1 + I_2)$

$(I_1 + I_2 + I_3)^2 = I_1^2 + I_2^2 + I_1 I_2 + I_2 I_1 + I_3^2 + I_1 I_3 + I_2 I_3 + I_3 I_1 + I_3 I_2$

$I_1(I_2+I_3) + I_2(I_1+I_3) + I_3(I_1+I_2)$

Variance of a hypergeometric random variable

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

$$= \frac{nG}{N} + n(n-1)\frac{G}{N}\cdot\frac{G-1}{N-1} - \left(\frac{nG}{N}\right)^2$$

$$= \frac{nG}{N}\left[1 + (n-1)\frac{(G-1)}{N-1} - \frac{nG}{N}\right]$$

$$= \frac{nG}{N}\left[\frac{N(N-1) + N(n-1)(G-1) - nG(N-1)}{N(N-1)}\right]$$

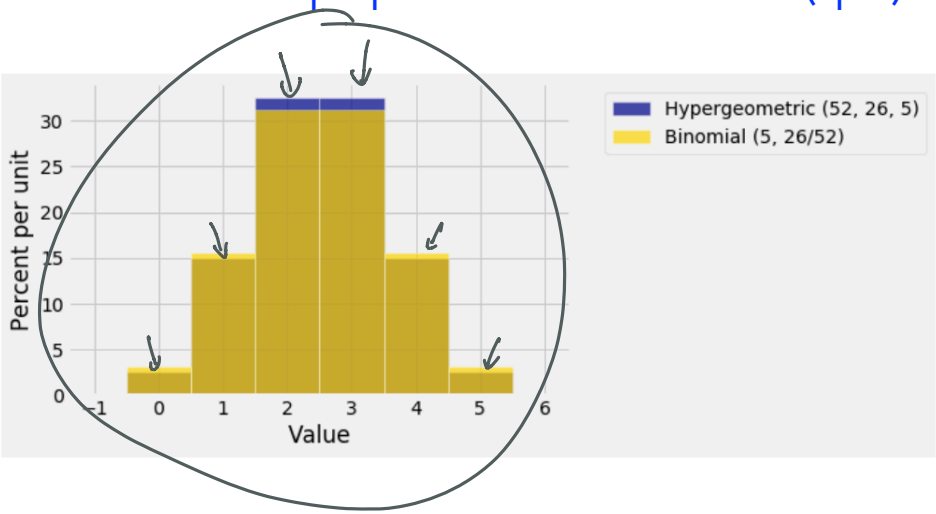
$$= \frac{nG}{N}\left[\frac{N^2 - N + nNG - NG - nN + N - nGN + nG}{N(N-1)}\right]$$

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

$$\text{Var}(X) = \underbrace{n}_{\text{Samp size}}\cdot\underbrace{\frac{G}{N}}_{P(S)}\cdot\underbrace{\frac{N-G}{N}}_{P(F)}\cdot\underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{square of Finite popn correction}}$$

Binomial (n, p) $n \cdot p \cdot (1-p) \cdot fpc$

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



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

Note that fpc ≤ 1

So SD(HG) ≤ SD(Bin)

In general we have that the :

bigger than SD (w/o repl)

SD of sum of an SRS = SD of sum WITH repl. × fpc

Exercise. Plug in values of N, n in your calculator & see what $\sqrt{\frac{N-n}{N-1}}$ will be. , $N = 10^6$, $n = 1000$

$\sqrt{\frac{10^6 - 10^3}{10^6 - 1}} = 0.999 \approx 1$

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.

Accuracy is governed by the SD.

$\sqrt{\frac{N-n}{N-1}}$ ← fpc

$n = 625$

$N_1 = 121485$ ← fp $\sqrt{\frac{121485-625}{121484}} \approx 0.9974$

$N_2 = 4 \times 10^6$

$N_2 \approx 1$

Accuracy depends on sample size

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

A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2021 Grammys. 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.

← you can pretend that sampling is w/ replacement.

- Don't know p. so

Want $SD(X) \leq 0.01$

$SD(X) = \frac{\sqrt{npq}}{n}$

X = percentage of 1's in sample

X = sum of draws

Avg of draws = $\frac{\text{sum of draws}}{n}$.

Use the upper bound on variance to solve this

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.

- 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 ($n\mu$).

Example: Coin toss

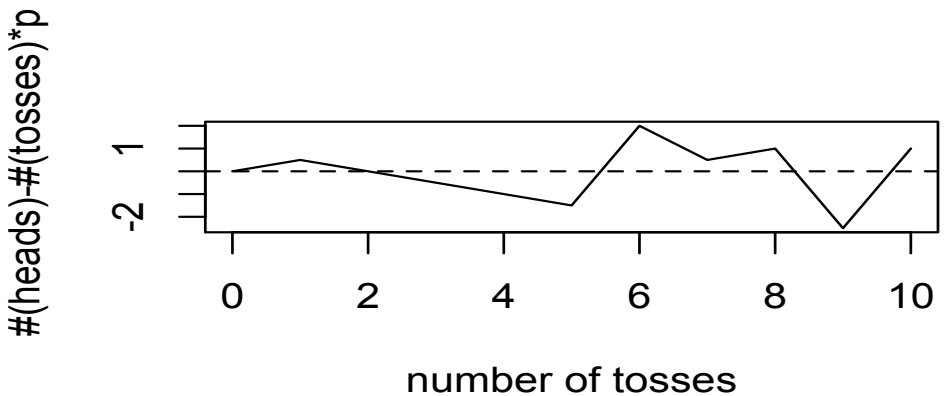
- Consider a fair coin, toss it 100 times & 400 times, count the number of H
- Expect in first case, roughly 50 H, and in second, roughly 200 H. So do you think chance of 50 H in 100 tosses and 200 H in 400 tosses should be the same?

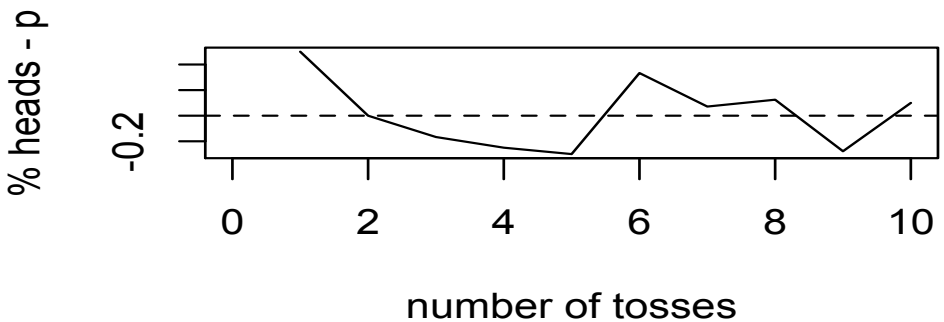
- $SD(S_{100}) =$
- $SD(S_{400}) =$

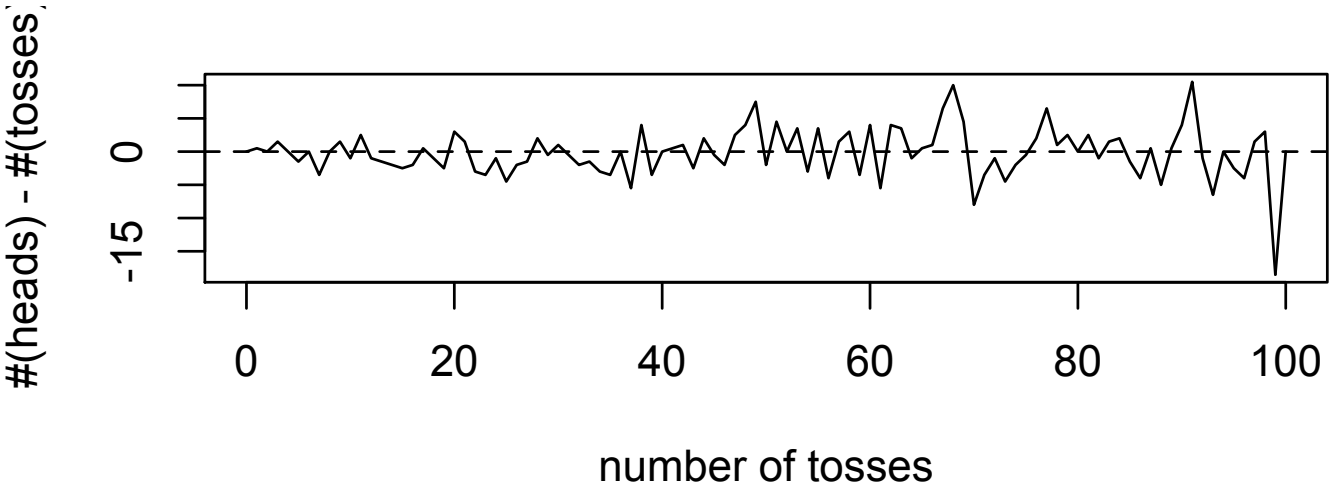
- Picture:

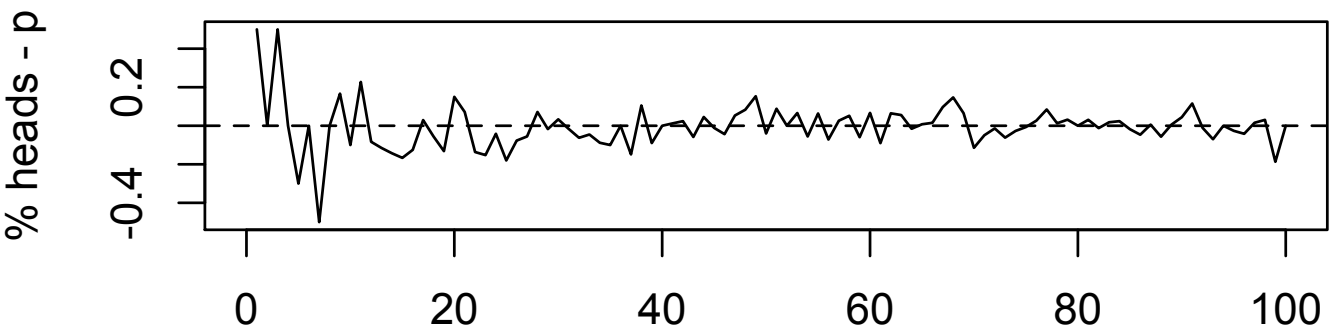
- P(200 H in 400 tosses)
- P(50 H in 100 tosses)

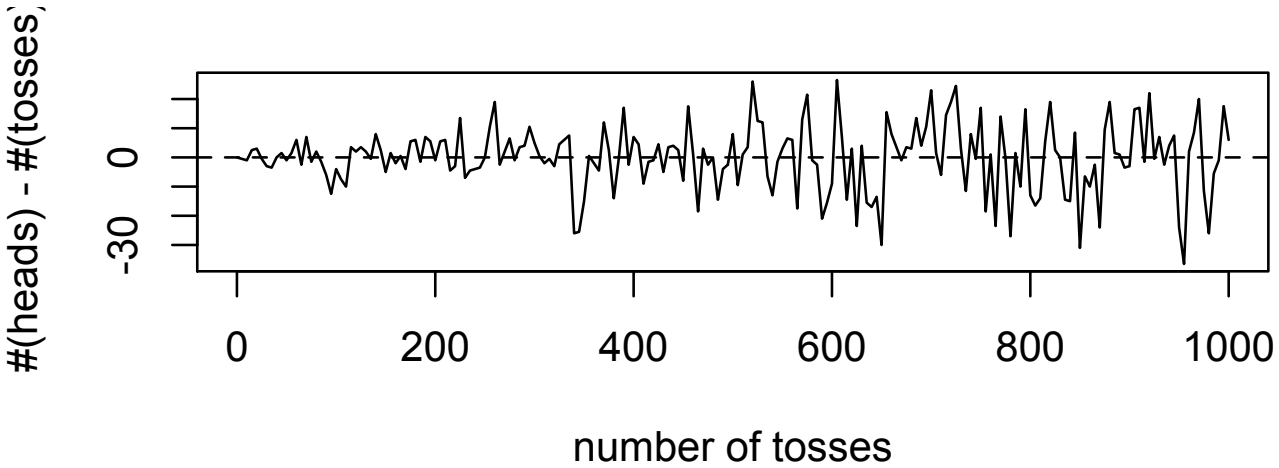
Simulating coin tosses: 10 tosses

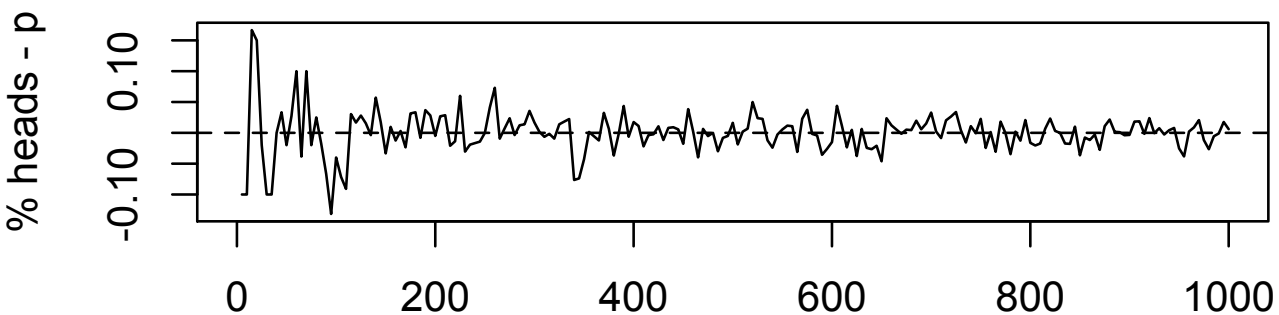


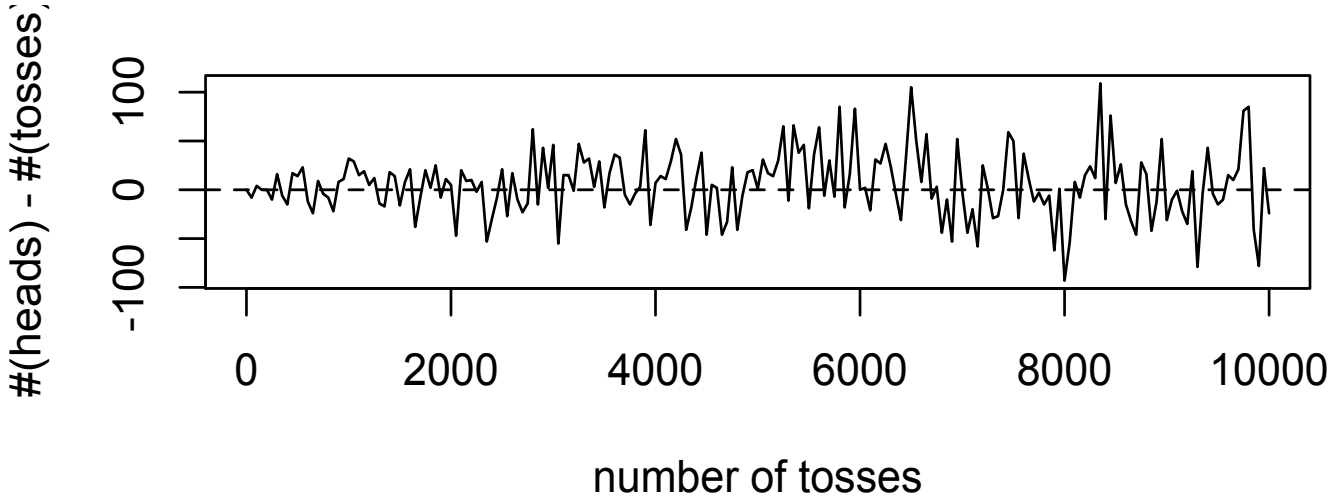


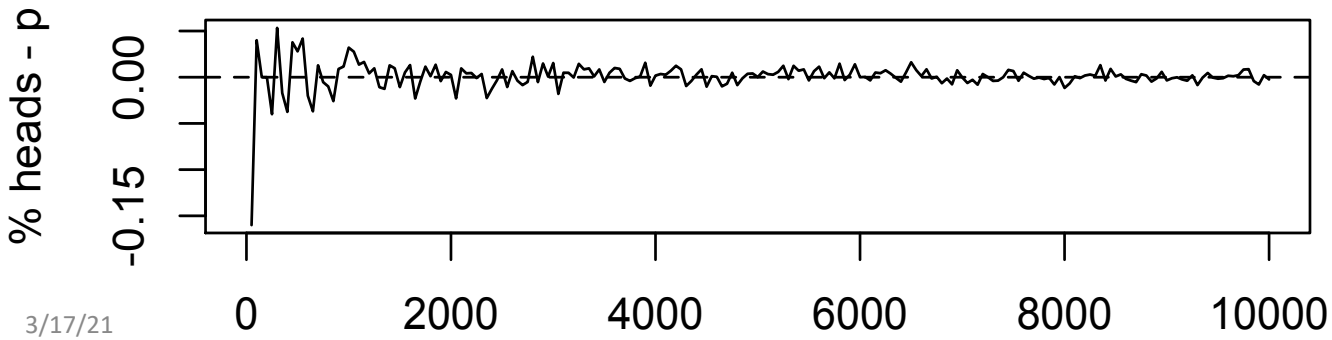

#(heads) - #(tosses)
-15
0
0
20
40
60
80
100
number of tosses


% heads - p
-0.4
0.2
0
20
40
60
80
100


#(heads) - #(tosses)
-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)
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%.

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 $SD(A_n) = ??$

- **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.

Concentration of probability

- This is when the SD decreases, so the probability mass accumulates around the mean, therefore, the larger the sample size, the more likely the values of the sample average $\bar{X}$ fall very close to the mean.

- **Weak Law of Large numbers:**

$For\ c > 0, P(|\ A_n - \mu\ | < c) \to 1\ as\ n \to \infty$

$|\ A_n - \mu\ |$ is the distance between the sample mean and its expectation. So when your sample size is large, then the chance that the sample mean is VERY CLOSE to its expected value is super high.

How close? As close as you like. Just take a large enough sample. BUT the chance that it is *exactly* equal to the expected value is tiny.

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.

- The *percentage* of sixes, when rolling a fair die over and over, is very close to 1/6. True for any of the faces, so the *empirical* histogram of the results of rolling a die over and over again looks more and more like the *theoretical* probability histogram.

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