Probability

Law of large numbers

let's take an (intuitive) **computational** approach to probability, the following code simulates rolling a die n times and computes the ratio of the number of ones that land up over n

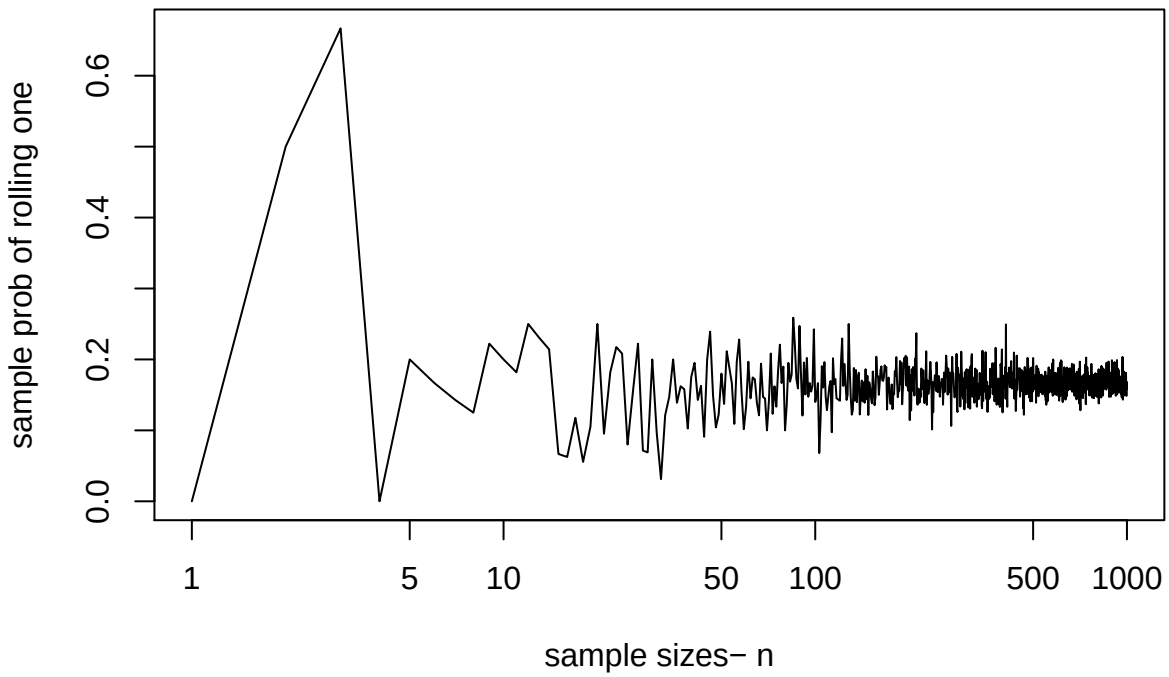
```
sample_prob <- vector()   # Initialize an empty vector

sample_sizes <- 1:1000   # these will be our sample sizes, n

for (N in sample_sizes) {
die <- 1:6
rolls <- replicate(N,sample(die,size=1,replace=TRUE))
sample_prob[N] <- table(
  factor(rolls,levels = c(1,2,3,4,5,6)))[1]/N
}
   # can also do similar code in python!
```

$p = \frac{\text{number of ones}}{n} \to 1/6 \approx 0.167 \qquad as \qquad n \to \infty$

```
plot(sample_sizes,sample_prob,log='x',
  type="l",xlab = 'sample sizes- n', ylab='sample prob of rolling one')
```



Probability of this *or* that

Probability is essentially counting the number of events of one type divide by the total number of all possible events, thus it is related to combinatorics

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

compare to the following expression from set theory

$|A \cup B| = |A| + |B| - |A \cap B|$

let's test this on a dataframe, computing some "probabilities"

We will look at data from the five games the Los Angeles Lakers played against the Orlando Magic in the 2009 NBA finals. Each row represents a shot Kobe Bryant took during these games

---

```
#download.file("http://www.openintro.org/stat/data/kobe.RData",
      #        destfile = "kobe.RData")
load("kobe.RData")
```

vs	game	quarter	time	description	basket
ORL	1	1	9:47	Kobe Bryant makes 4-foot two point shot	H
ORL	1	1	9:07	Kobe Bryant misses jumper	M
ORL	1	1	8:11	Kobe Bryant misses 7-foot jumper	M
ORL	1	1	7:41	Kobe Bryant makes 16-foot jumper (Derek Fisher assists)	H
ORL	1	1	7:03	Kobe Bryant makes driving layup	H
ORL	1	1	6:01	Kobe Bryant misses jumper	M
ORL	1	1	4:07	Kobe Bryant misses 12-foot jumper	M

---

let A = prob that shot was taken in game 1

let B = prob that shot was taken in quarter 1 (of *any* game)

```
total_shots <- nrow(kobe)
  total_shots
```

```
## [1] 133
```

```
A <- nrow(filter(kobe,game==1))
  A
```

```
## [1] 34
```

```
B <- nrow(filter(kobe,quarter==1))
  B
```

```
## [1] 36
```

```
A_and_B <- nrow(filter(kobe,quarter==1 & game==1))
  A_and_B
```

```
## [1] 9
```

```
A_or_B <- nrow(filter(kobe,quarter==1 | game==1))
  A_or_B
```

```
## [1] 61
```

summarizing:

P(A)	P(B)	P(A) + P(B)	P(A and B)	P(A or B)
34/133	36/133	70/133	9/133	61/133

$61/133 = 34/133 + 36/133 - 9/133$ everything checks out!

takeaway - avoid over (under) counting

Think about if your boss asked you what proportion of your clients meet some criterias: A ***or*** B

This is a fairly important and common question in the business world

While probability may seem very simple, it is important to keep fundamental rules in mind to avoid simple counting errors

This issue will be addressed in the lab exercise the "42" problem

Marginal and Joint probabilities

- based on single variable, is called **marginal** probability
- for two or more variables is called a **joint** probability

consider 2 **categorical** variables (columns) from kobe dataframe

1. **basket** which cant take values: 'H' (hit) or 'M' (miss),
2. **quarter** which can take values: 1,2,3,4,1OT

This data can be obtained using `table()` function

```
baskets_marginal <- table(select(kobe,basket))

addmargins(baskets_marginal)  # get totals
```

Basket marginals

	counts
H	58
M	75
TOTAL	133

	proportion
H	0.436
M	0.564
TOTAL	1.000

quarter marginals

	counts
1	36
1OT	7
2	25
3	34
4	31
TOTAL	133

	proportion
1	0.271
1OT	0.053
2	0.188
3	0.256
4	0.233
TOTAL	1.000

0.564 is the prob of kobe missing any given shot in the series

0.256 is the prob that a random shot, from any of the 5 games, taken by kobe was placed in the 3rd quarter

you could argue that these are not probs, just historical %s...

Contingency table from data frame

Now let's create a contingency table which is the joint distribution for baskets and quarters

```
contingency_table <- table(select(kobe,basket,quarter))
addmargins(contingency_table)
```

	1	1OT	2	3	4	TOTAL
H	18	3	11	16	10	58
M	18	4	14	18	21	75
TOTAL	36	7	25	34	31	133

Now we can use this to talk about **conditional probabilities**

Row proportions

	1	1OT	2	3	4	TOTAL
H	0.310	0.052	0.190	0.276	0.172	1
M	0.240	0.053	0.187	0.240	0.280	1
TOTAL	0.271	0.053	0.188	0.256	0.233	1

These are probabilities of shots being taken in various quarters, **conditioned on** whether the shot is a hit (H) or miss (M), notice the bottom row is the marginal (unconditioned) for example

$P(\text{Q3} \mid \text{M}) = 0.240$

Of the missed shots, what probability (or percentage) of them occurred in the 3rd quarter? (Answer: 0.240)

Column proportions

	1	1OT	2	3	4	TOTAL
H	0.5	0.429	0.44	0.471	0.323	0.436
M	0.5	0.571	0.56	0.529	0.677	0.564
TOTAL	1.0	1.000	1.00	1.000	1.000	1.000

These are probabilities of shot being a hit or miss, **conditioned on** what quarter the shot was taken on, notice the rightmost column is the marginal (unconditioned) for example

$P(\text{H} \mid \text{Q2}) = 0.44$

In the 2nd quarter, what probability (or percentage) of shots did kobe hit? (Answer: 0.44)

How to get proportions in R

```
round(prop.table(contingency_table, 1),2) # row percentages
```

```
##       quarter
## basket    1  1OT    2    3    4
##      H 0.31 0.05 0.19 0.28 0.17
##      M 0.24 0.05 0.19 0.24 0.28
```

```
round(prop.table(contingency_table, 2),2) # column percentages
```

```
##       quarter
## basket    1  1OT    2    3    4
##      H 0.50 0.43 0.44 0.47 0.32
##      M 0.50 0.57 0.56 0.53 0.68
```

Cell proportions: $P(A \text{ and } B)$

The relationship between conditional (row or column proportions) and cell proportions is:

$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

```
mytable <- round(prop.table(contingency_table),3) # cell percentages
addmargins(mytable)
```

```
##       quarter
## basket     1   1OT     2     3     4   Sum
##    H   0.135 0.023 0.083 0.120 0.075 0.436
##    M   0.135 0.030 0.105 0.135 0.158 0.563
##    Sum 0.270 0.053 0.188 0.255 0.233 0.999
```

ex: $P(\text{H} \mid \text{Q2}) = \frac{P(H \text{ and } Q2)}{P(Q2)} = \frac{0.083}{0.188} = 0.44$

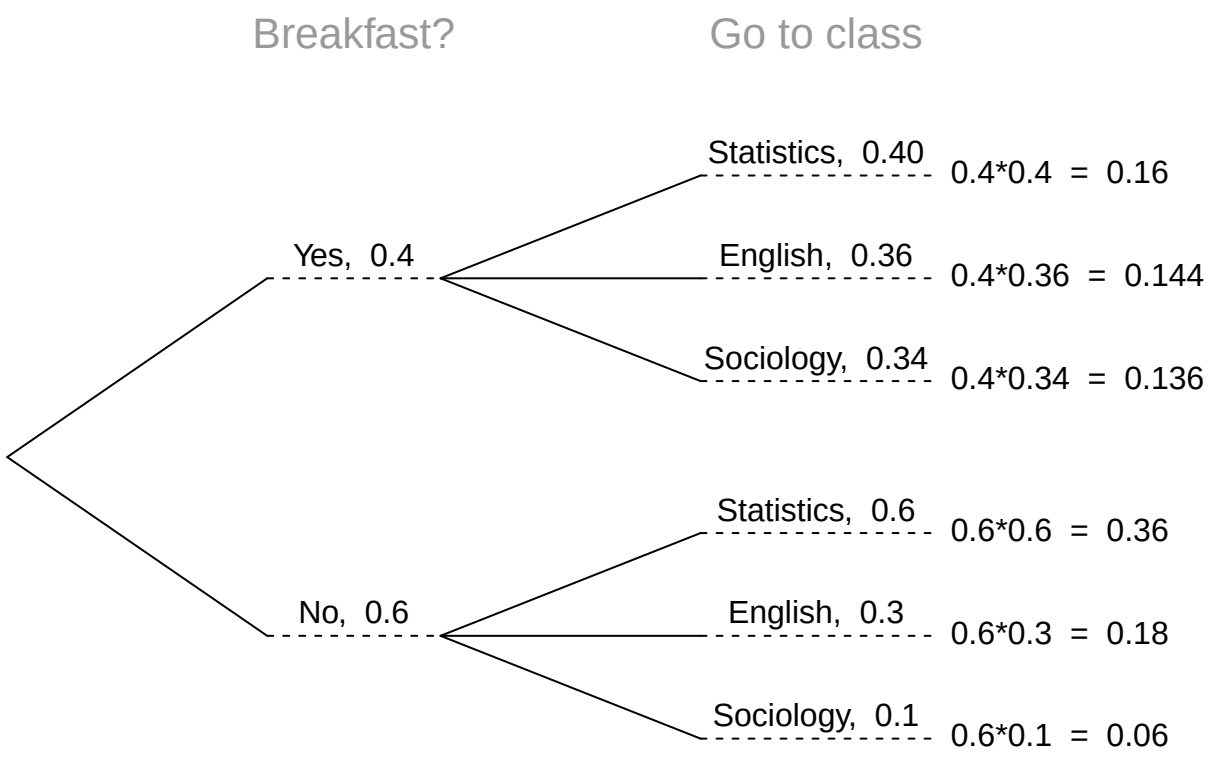
Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors

Tree diagrams can be built with the openintro package

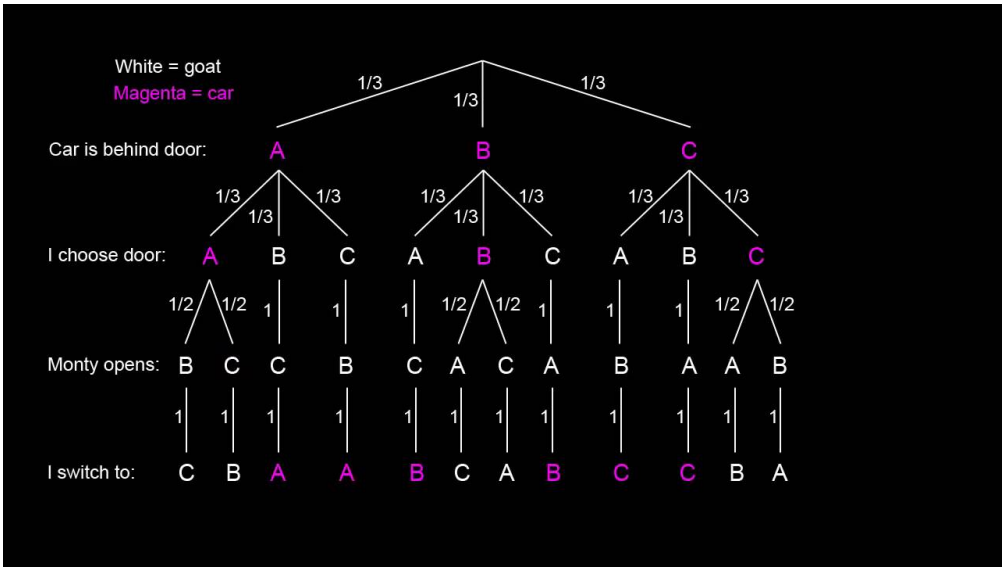
```
install.packages('openintro')
library('openintro')

treeDiag(c('Breakfast?','Go to class'), c(.4,.6),
   list(c(.4,.36,.34),c(.6,.3,.1)), c('Yes','No'),
   c('Statistics','English','Sociology'), showWork=TRUE)
```



Decision Tree

Combine decisions with probabilistic outcomes → decision tree



"Formal" probability

You can approach probability from different angles, you can think in terms of frequency tables, or, you can think about it from a more mathematical perspective. This perspective is important for talking about *distributions*

let's start with **discrete** case:

Say a random var X takes a number of (possibly countably infinite) discrete values $\{x_i\}$ where each x_i is a real number. Then a probability mass function (pmf), $p(x_i)$ must satisfy

1. $p(x_i) \geq 0$ for all outcomes i
2. $\sum_i^\infty p(x_i) = 1$

Continuous pdf

Say a random var X takes values x on a continuum R_X. The probability that x lies in the interval $[a, b]$ is

$P(a \leq x \leq b) = \int_a^b f(x)dx$

where $f(x)$ is called a probability density function satisfying

1. $f(x) \geq 0$ for all x
2. $\int_{R_x} f(x)dx = 1$

Graphically, the probability is the **area** underneath the curve $f(x)$ between $x = a$ and $x = b$.

Cumulative dist functions (CDFs)

Discrete CDF: $F(x) = \sum_{x_i \leq x} p(x_i)$

Continuous CDF: $F(x) = \int_{-\infty}^x f(t)dt$

Properties

1. non-*decreasing*: if $a < b$ then $F(a) \leq F(b)$
2. $F(x)$ takes values on the interval $[0, 1]$

3. $P(a \leq x \leq b) = F(b) - F(a)$

Expectation - mean and variance,SD

discrete

$E(X) = \sum_i x_i p(x_i) = \mu$ - the "mean"

continuous

$E(X) = \int_{-\infty}^{\infty} x f(x) dx = \mu$ - the "mean"

mode & median

most frequent value (non-unique e.g. "bimodal") & middle value

variance

$\sigma^2 = V(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$ $S.D. = \sqrt{\sigma^2} = \sigma$

Moments

$E(X) = \mu$ called the 1st moment

$E(X^n)$ is called the nth moment

μ is a measure of central tendency

σ^2 is a measure of spread about μ

$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$ measure of asymmetry about μ

$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}$ measure of the tail of the dist

Sample mean, Sample variance/SD

The sample mean over n samples*, $X_1, X_2, ..., X_n$, from the distribution (population) is

$\overline{X} = \frac{1}{n} \sum_i X_i$ unbiased estimator of μ

The sample variance over this same sample is

$s^2 = \frac{1}{(n-1)} \sum_i^n (\overline{X_i} - \overline{X})^2$ unbiased estimator of σ^2

$s = \sqrt{s^2}$ biased estimator of σ

We will discuss sample *bias* a bit later in the course

```
*we assume each sample is i.i.d
```

Sampling distributions

It's important to understand that if a random variable, X has some distribution, then statistics over i.i.d samples of size n of X, such as $\overline{X}$, also are random variables themselves

Thus these statistics, have their own *separate* distributions; which is not the same distribution as that of the underlying random variable X necessarily

X and $\overline{X}$ have different distributions

n is a **parameter** of the distribution of $\overline{X}$

Standard error of the mean

The standard error of the mean (SEM) can be expressed as

S.E.M. $= \frac{\sigma}{\sqrt{n}}$

where

σ is the standard deviation of the population

n is the size (number of observations) of the sample.

If the population's (or "distribution's") standard deviation is unkown, the standard error of the mean is usually estimated as the sample standard deviation divided by the square root of the sample size (assuming statistical independence of the values in the sample).

Deriviation

- If $x_1, x_2, \ldots, x_n$ are n independent observations from a population that has a mean μ and standard deviation σ, then the variance of the total $T = (x_1 + x_2 + \cdots + x_n)$ is $n\sigma^2$
- The variance of T/n (the mean $\bar{x}$) must be $n\left(\frac{\sigma^2}{n^2}\right) = \frac{\sigma^2}{n}$. Alternatively, $\text{var}(\frac{T}{n}) = \frac{1}{n^2}\text{var}(T) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$
- And the standard deviation of T/n must be $\sigma/\sqrt{n}$

This has some important consequences as we shall see by example

Rolling a die

A die has a discrete pmf with expectation and variance

$E(X) \equiv \mu = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$

$\sigma^2 = \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 - 3.5^2) \approx 2.9$

$\sigma \approx \sqrt{2.9} = 1.7$

When people say "standard error of the mean", what they are saying is that if you draw n samples and calculate the sample mean, $\overline{X}$, **that mean itself has a distribution** which is normal with standard deviation given by approximately $\sigma/\sqrt{n}$

Next, I will show you how to empirically study this:

S.E.M by Simulation: double averaging

1. Fix a sample size value n
2. Draw n samples, $X_1, X_2, ..., X_n$, from the dist/population
3. Compute the sample mean, $\overline{X} = \frac{1}{n}\sum_i X_i$
4. go back,repeat steps 2 and 3, m times, (e.g. $m = 100$)

5. Estimate the standard error of the mean by calculating the standard deviation of the means, $\overline{X_1}, \overline{X_2}, ..., \overline{X_m}$ Est(S.E.M.) $= \sqrt{\frac{1}{(m-1)} \sum_j^m (\overline{X_j} - \overline{\overline{X}})^2}$ where $\overline{\overline{X}}$ is the average of the averages!! You will need to draw $n \times m$ samples, **for each value of n**
6. change n then, go back to step 1

Example: back to tossing dice

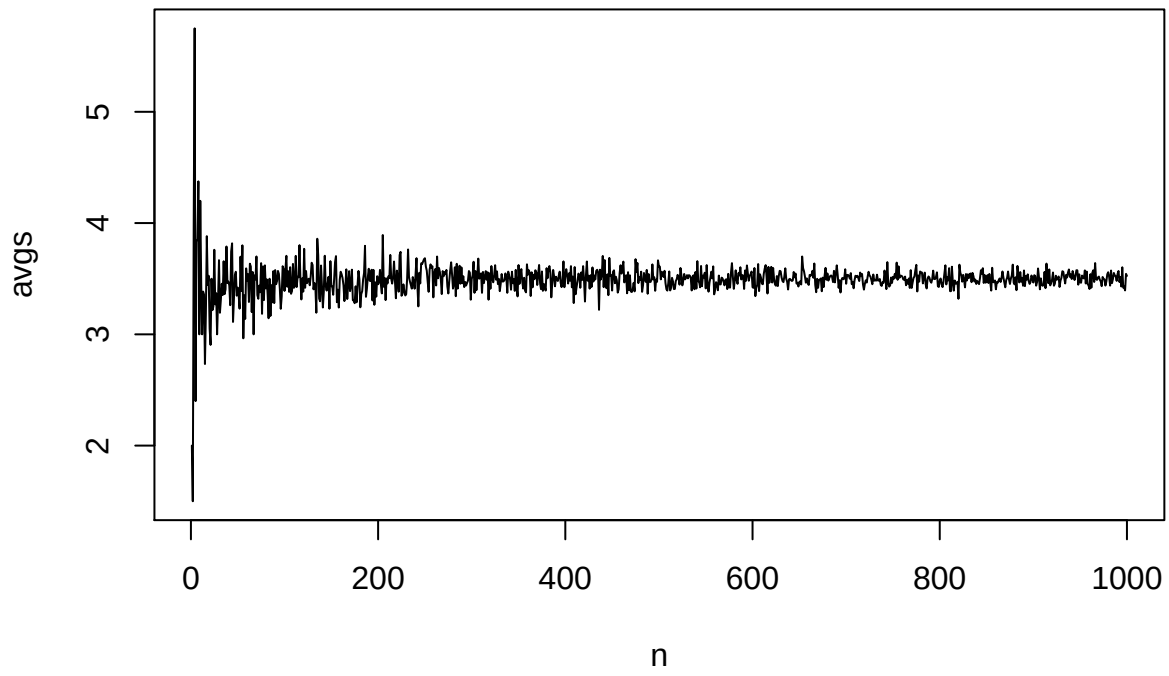
```
sample_sizes <- 1:1000   # note its samples SIZES, not sample SIZE
sds <- vector()
avgs <- vector()
standard_errors <- vector()

for (N in sample_sizes) {
die <- 1:6
rolls <- replicate(N,sample(die,size=1,replace=TRUE))
avgs[N] <- mean(rolls)
sds[N] <- sd(rolls)

standard_errors[N] <- sd(     # Note the two layers of averaging
  replicate(50,mean(
    replicate(N,sample(die,size=1,replace=TRUE)))))
}
```

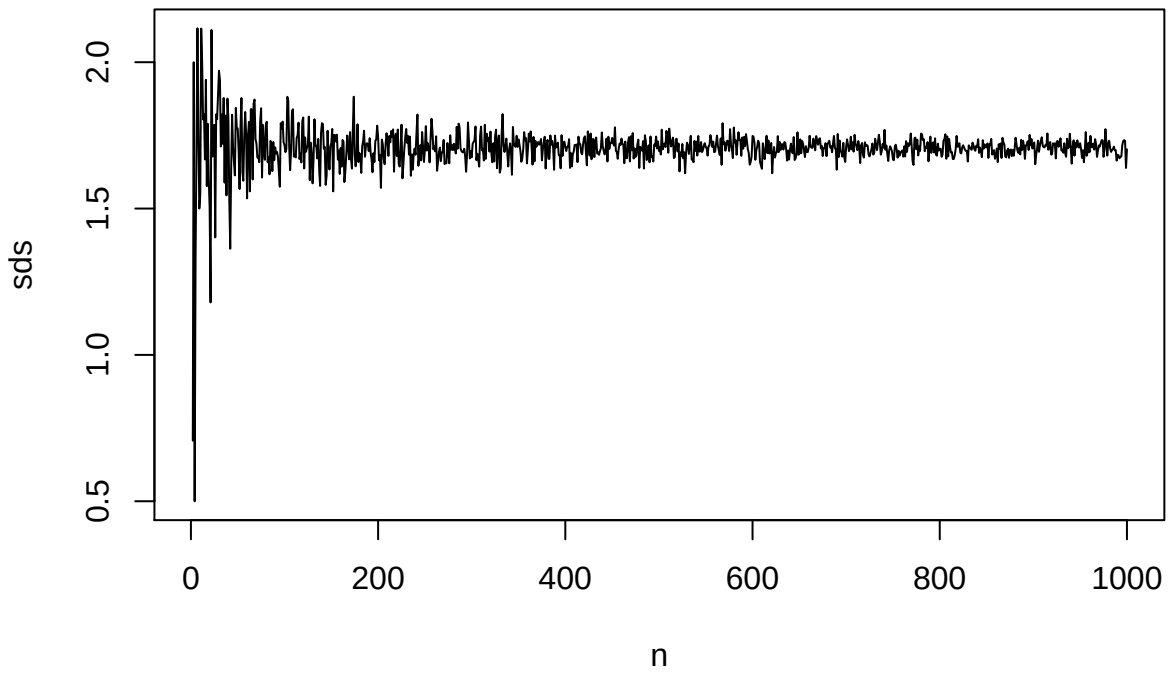
Means of the samples (size n)

Notice it tends towards 3.5



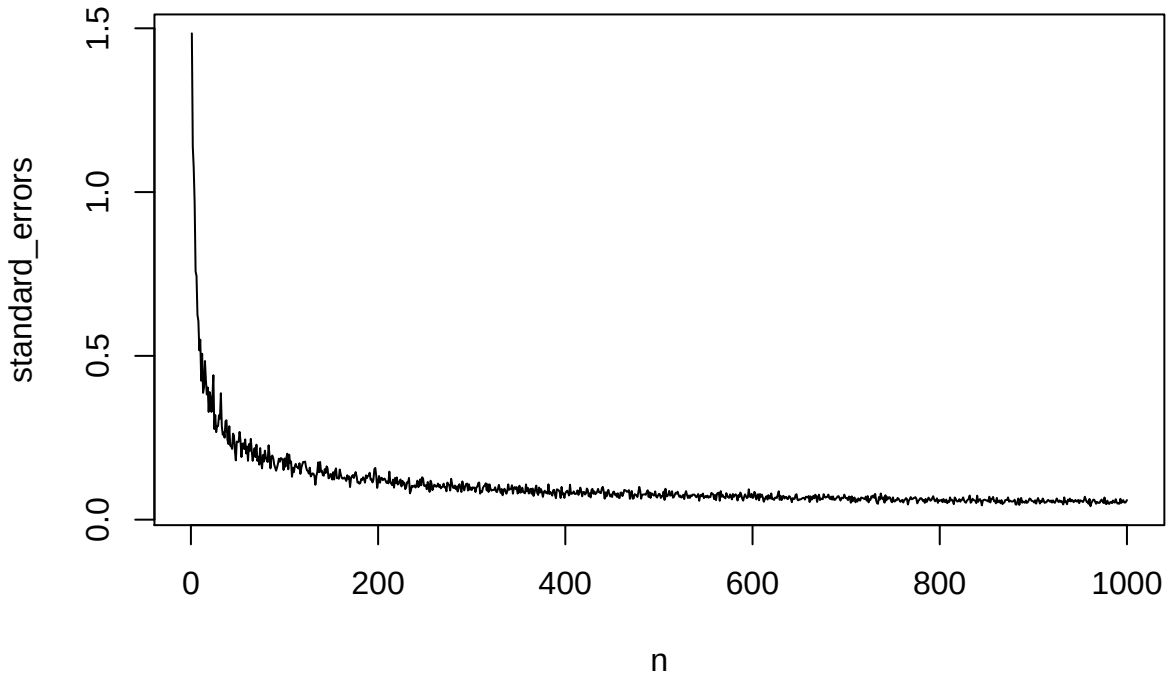
S.D.s of samples (size n)

Notice it tends to 1.7



S.D.s of Means of samples (size n)

the Est(S.E.M.) as function of n, $[m = 50]$ $\sim 1.7/\sqrt{n}$



Summary: sorry to be redundant

Distribution of the *sample statistic*

The means of samples of size n are normally distributed with standard deviation $\sigma/\sqrt{n}$ where sigma is the population (underlying distribution's) standard deviation 1.7. Note that n is a parameter of this so-called "sampling distribution"

Distribution of the sample *itself*

Each of the random variables in a sample of size n are UNIFORMLY (discrete) distributed with standard deviation 1.7 and mean 3.5

Required Conditions

samples must be i.i.d, as usual