Week 3| Probability distributions

pp. 275 - 300, 306-314, 320-322

Frequentist definitions of probability

Long-run (frequentist) probability

- Based on the 'true' (as opposed to 'subjective') probability of an event
- Values which are fixed and can be replicated by others: what you see is what what you get.
- The more data we have, the closer we can approximate the 'true' values of a population.
- Frequentist statements provide binary outcomes: something *is* the case, or it *is not*.

- Assumes probabilities reflect objective realities 'out there' and should be *fixed*. An alternative approach (Bayesian statistics) assumes probabilities are fundamentally 'subjective' and may be altered in light of incoming information.

We will focus on frequentist statistics over this course for the following reasons:
1. Frequentist statistics are more commonly used and understood.
2. Justifications for subjective probablities do not have to be provided (e.g., "*I believe...*" vs "*It was shown...*").
3. Bayesian methods provide variable outcome parameters whereas frequentist methods provide fixed outcomes.

- In practice, we never have 'infinite' trials and thus can never fully approximate what is the 'true' population parameter.

- Frequentist probabilities reflect global (not local/contingent) probabilities. For example, claiming that a coin will display Heads (`H`) 40% of the time does *not* mean that the next flip has a 40% chance of yielding Heads. Instead, if the coin is flipped an 'infinite' amount of times (or at least many, many, many times), than 40% of the flips will be Heads.

Binomial data

We discuss how 'objective' probabilities are contingent on a 'sufficiently' large sample using simulated coin-flip data. Coin flips provide binary outcomes (either `H` or `T`) which can be readily interpreted in terms of probabilities. Flipping coins can produce Heads *or* Tails, meaning the responses are nominal, *not* continuous (A coin-flip cannot result in `.8` Heads or `.2` tails). Data with two nominal categories are called *binomial*.

> Suppose out of five coin flips we view the sequence **HHTTH** - in this case, the objective probability of Heads out of five coin flips is $\frac{3}{5}$ or 0.6. Our demonstration is not restricted to coin flips only - *any* binary outcome parameter may be described in this way (e.g., given the distribution of males and females in a classroom to be **MMFFMFMMFFFM**, we can report the probability of females out of 12 students is $\frac{6}{12}$, or 0.5).

R has a built-in function for generating random samples of values from various *distributions* - the ones we will generally use across the majority of analyses are binomial (categorical), as well as normal, t-statistic, χ^2 and F-statistic distributions (which are all continuous). We will illustrate the various distributions later on.

First, let's generate random samples of increasingly larger sizes to observe how estimated probabilities converge towards their 'true' state (recall that frequentist probabilities deal with objective, or 'true', events)

Simulating binomial experiments

We will simulate 10 coin flips by repeatedly (and randomly) sampling from two outcome possibilities (H|T). We may assume there is an equal chance of selecting either possibility from the space given (H = .5|T = .5). It will be demonstrated how the 'true' probabilities become increasingly realized as sample size increases.

```
set.seed(22)        # Enables replication

tosses <- sample(   # Using R's built-in 'sample' function
  x = c(1,0),       # Represents Heads and Tails as `1` and `0` (sample space)
  size = 10,        # Represents how many tosses are made (sample size)
  replace = TRUE,   # Replaces value taken from sampling space (when appropriate)
  prob = c(.5,.5)   # Describe the probability of selecting each sample (50% for Heads, 50% for Tails)
)

numflips  <- 1:length(tosses)               # Total number of flips
numheads  <- cumsum(tosses)                 # Cumulative summation of all heads
propheads <- round(numheads/numflips,2)     # Proportion rounded to 2 decimals
```

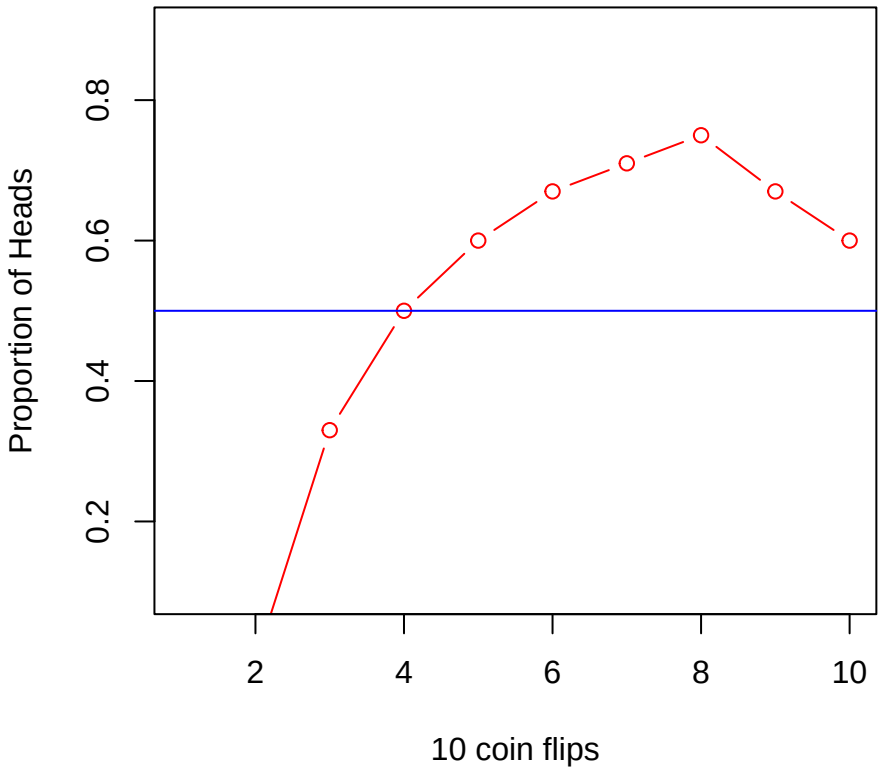
Assuming '1' and '0' represent Head and Tails respectively, our sequence of coin-flips can be represented as 0, 0, 1, 1, 1, 1, 1, 1, 0, 0. Out of 10 coin flips in total, we viewed 6 Heads.

Note how the 'final' proportion of Heads $P(H) = \frac{6}{10} = 0.6$ does not match the 'true' probability of $P(H) = .5$. According to frequentist theory, the 'sample' of coin tosses ($N = 10$) is not representative of the 'population' of coin tosses (essentially infinite). Although we cannot toss a coin an infinite number of times, we can *increase* the sampling space which should reduce the gap between the sample and population estimates.

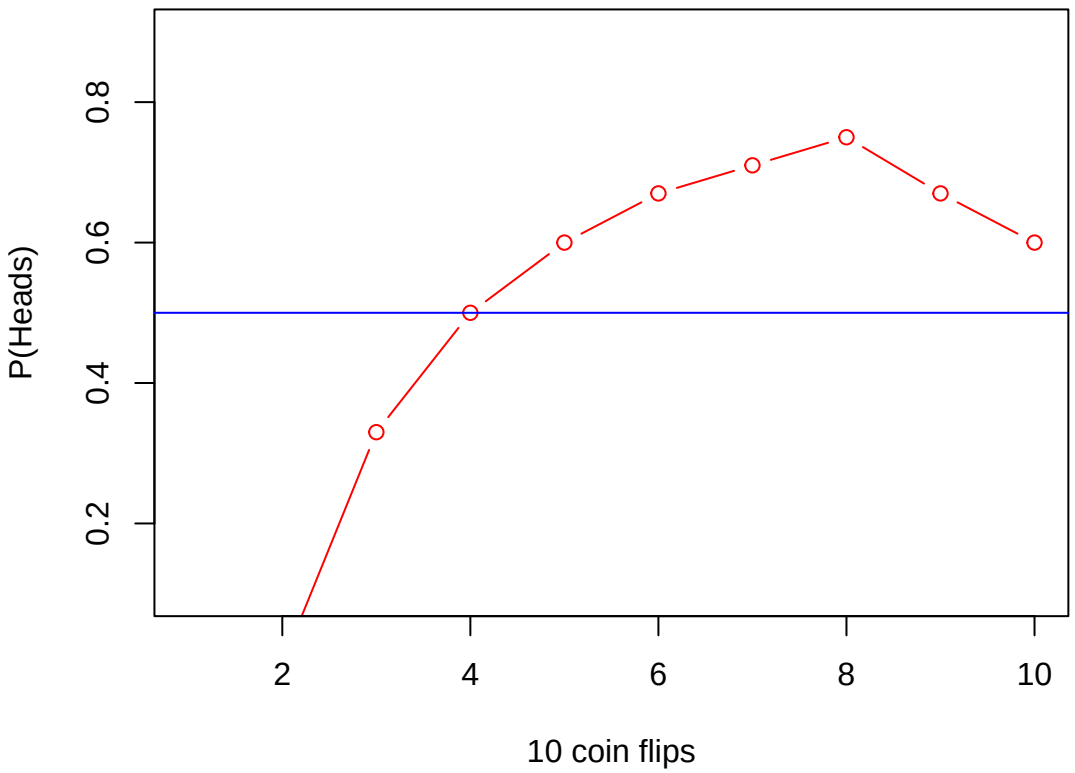
Converging towards the population mean

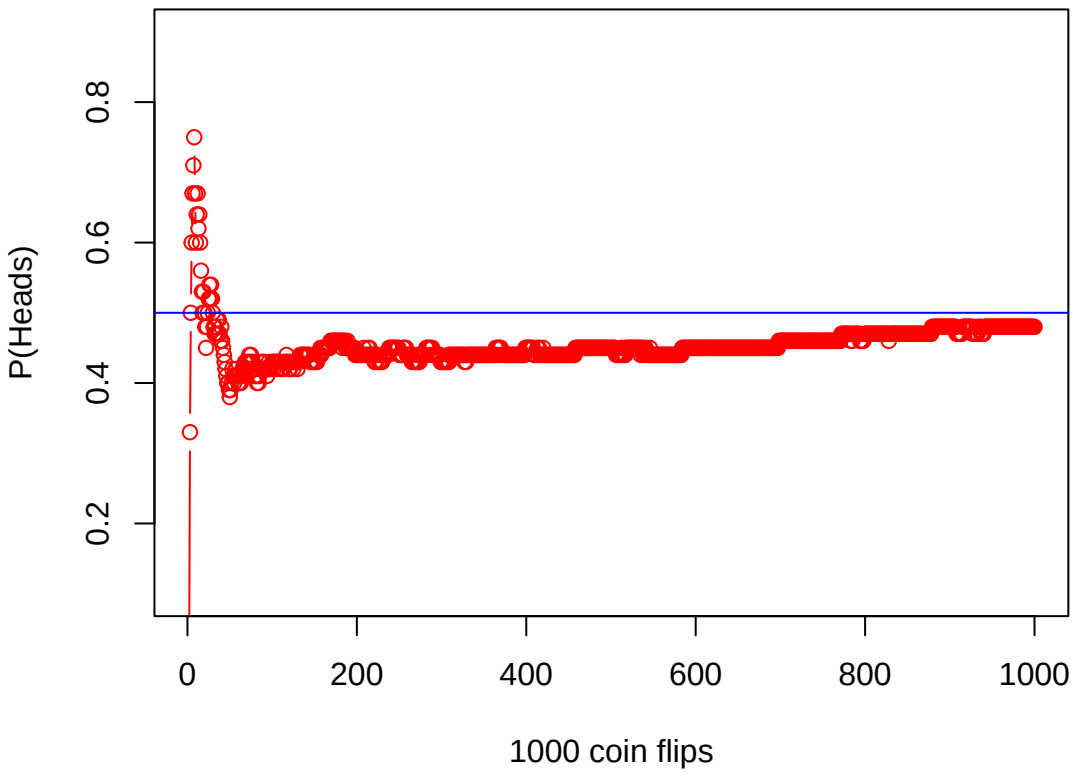
To test the assertion that larger N's shift sample parameters towards population parameters, let's first estimate the probability of heads ($P(H)$) with each flip. This would generate a sequence like the following, 0, 0, 0.33, 0.5, 0.6, 0.67, 0.71, 0.75, 0.67, 0.6 (each probability represents the cumulative number of heads divided by the cumulative number of trials). We can plot the probabilities using a line graph (the true population parameter is represented by the horizontal blue line).

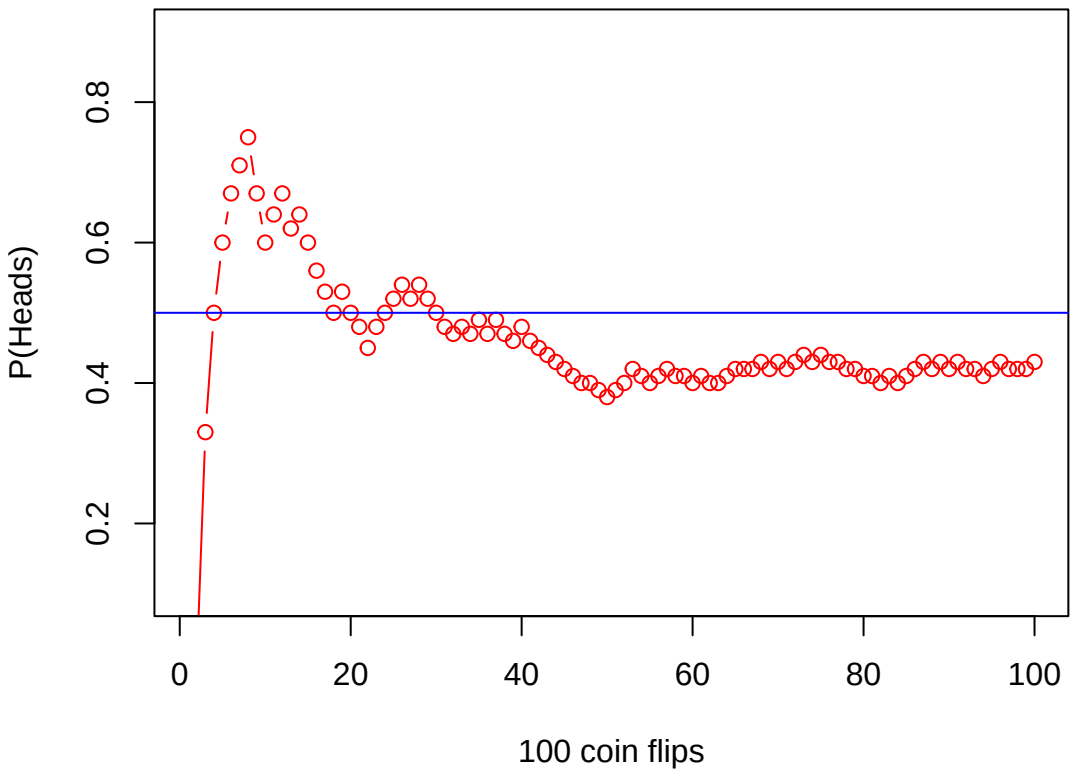
```
plot(propheads,                             # Name of vector created above
     ylim = c(.1,.9),                       # Limits of the y-axis
     type = "b",                            # We want 'both' lines and scattered values
     col = "red",                           # That is red
     xlab = "10 coin flips",                # The label of (horizontal) x-axes
     ylab = "Proportion of Heads")          # The label of (vertical) y-axes
     abline(h = .5,col = "blue")            # Draw a y-intercept of (.5) to illustrate the 'true' param
```

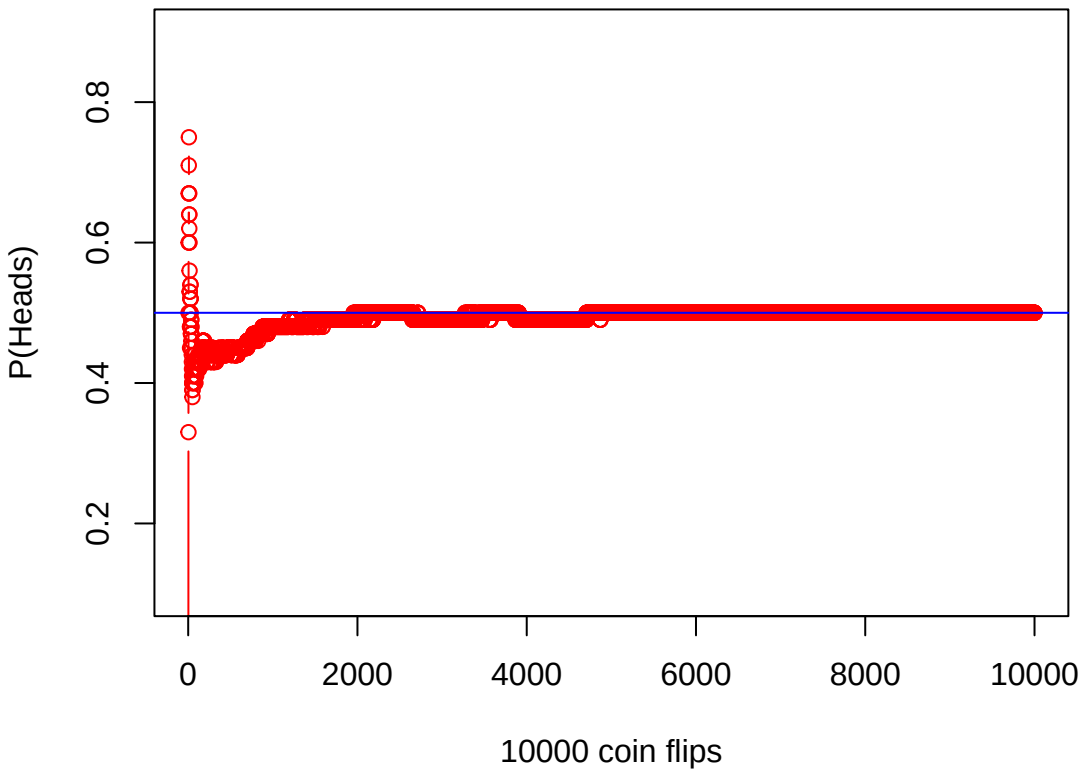


Let's keep flipping the coin and see what happens...


P(Heads)
0.2
0.4
0.6
0.8
2
4
6
8
10
10 coin flips


P(Heads)
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
1000 coin flips


P(Heads)
0.2
0.4
0.6
0.8
0
20
40
60
80
100
100 coin flips



Note the sample probability of observing a Heads approximates the 'true' probability, $P(H) = .5$ after a large number of flips...

Flipping coins can produce Heads *or* Tails, meaning the responses are categorical, not continuous (a coin-flip cannot result in `.8` Heads or `.2` tails). To recall our earlier point, data with two nominal categories ('successes' and 'fails') are described as *binomial*.

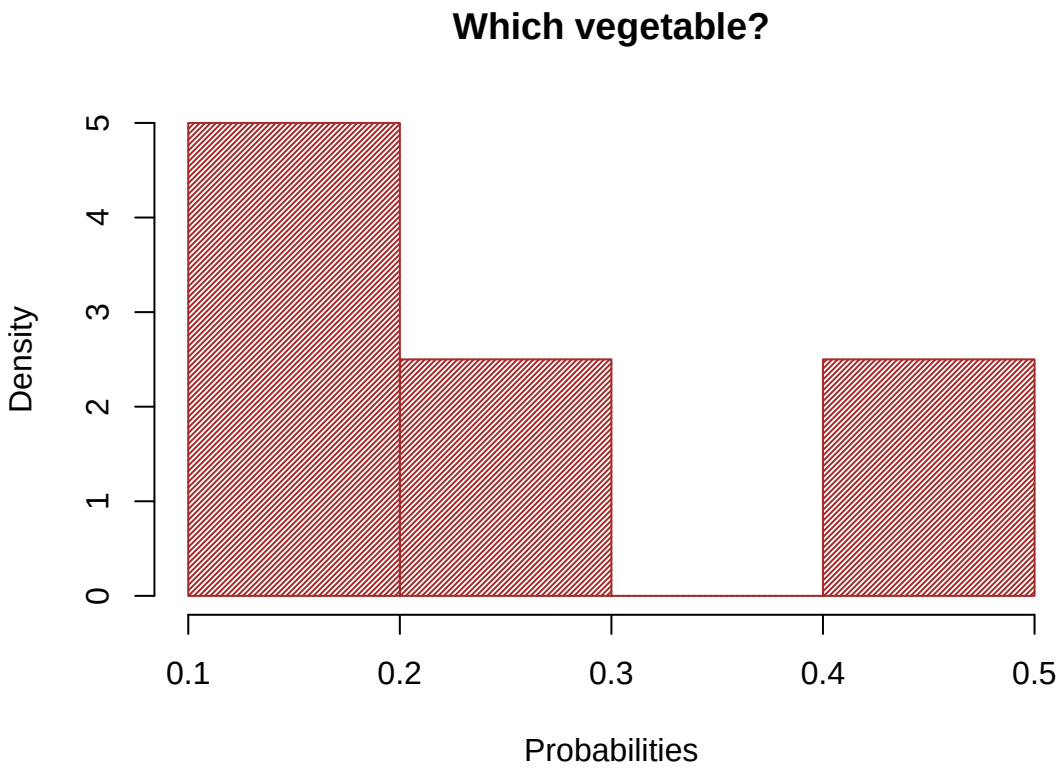
Not all nominal data is binomial

Nominal data can contain multiple (>2) categories and not restricted to binomial interpretations. To illustrate, suppose we recorded which vegetables were purchased over the course of a pre-determined time period (say 100 days). We could describe the probabilities of vegetabl purchase as follows:

Vegetables	Vegetable Label	Purchase probability
Broccoli	V_1	$P(V_1) = .12$
Cucumber	V_2	$P(V_2) = .28$
Tomato	V_3	$P(V_3) = .16$
Carrot	V_4	$P(V_4) = .44$

We could report that 12% of all vegetables purchased will be Broccoli (which is technically *not* the same as

thinking there's as 12% chance of purchasing Brocolli). Also note that $P(V_1) + P(V_2) + P(V_3) + P(V_4) = 1$, satisfying the constraint that all probabilities sum up to 1.



We can visually plot our probabilities to note which vegetable appears likely to be purchased on a given weekday.

Trending towards normal (distributions)

We had previously assumed that our coin was 'fair', meaning that after enough iterations, both Heads and Tails can be expected at probabilities of `.5` and `.5` respectively. We summarized this as $P(H) = .5; P(T) = .5$, which gives a *total probability* of 1.

Now suppose we had a trick coin that is expected to produce Heads 65% of the time. Although the total probability would not change (it must always be 1), we can update the likelihood of viewing Heads and Tails (as 1-Heads). So if $P(H) = .65$, $P(T) = 1 - P(H) = .35$.

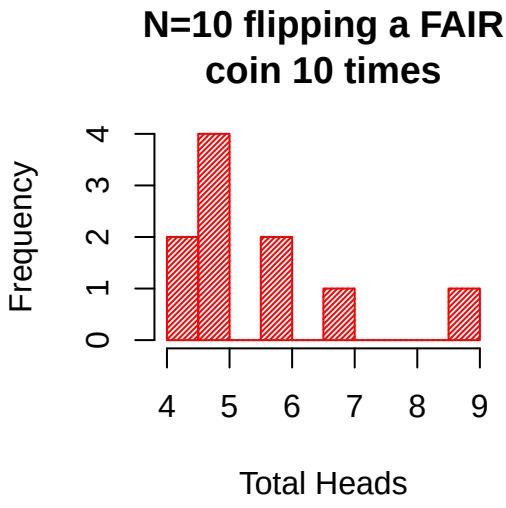
We may summarize the true probabilities of our two coins below:

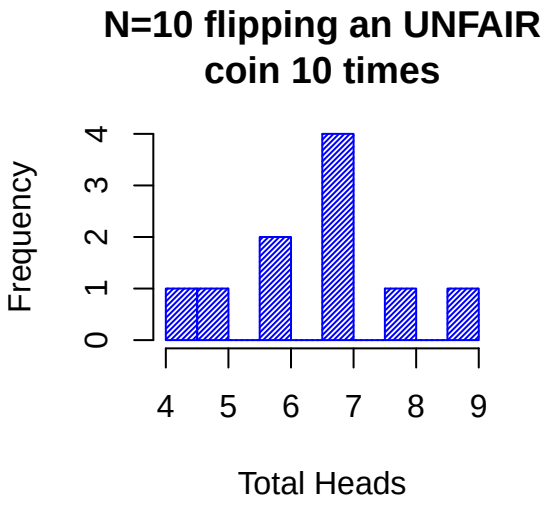
Coin Type	Heads	Tails
Fair (50/50)	P(H) = .5	P (H) =.5
Unfair (65/35)	P(H) = .65	P (T) =.35

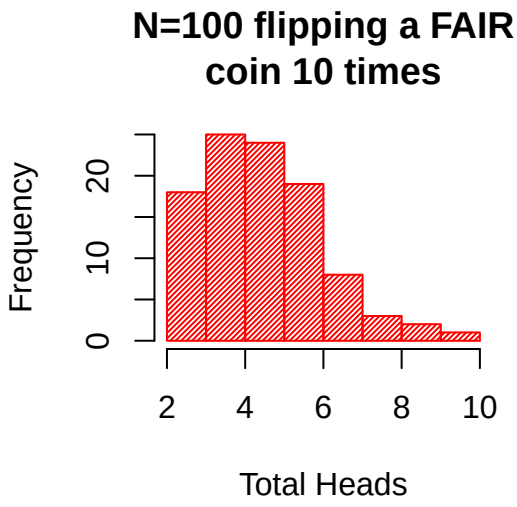
From our example earlier, we know that the true/population-level probabilities will become evident with increasing sample sizes. If we flip an unfair coin many times, the sample probability should become $P(H) = .65$. Let us now examine how probabilities are *distributed* across a sample space.

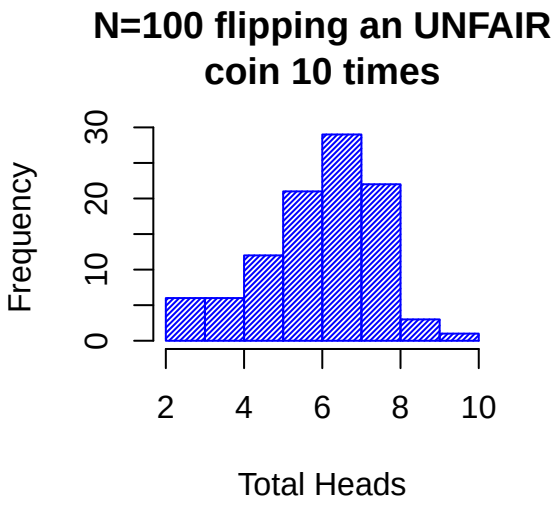
More coin flips!

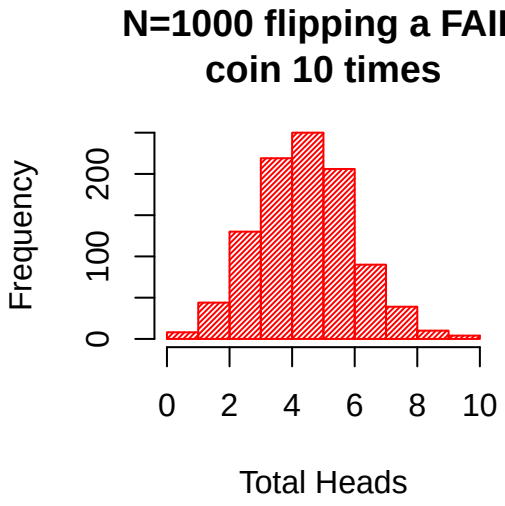
In the example below, we present simulated data of participants, each flipping the same coin across 10 trials, to ensure we have a reliable estimate from each sample.

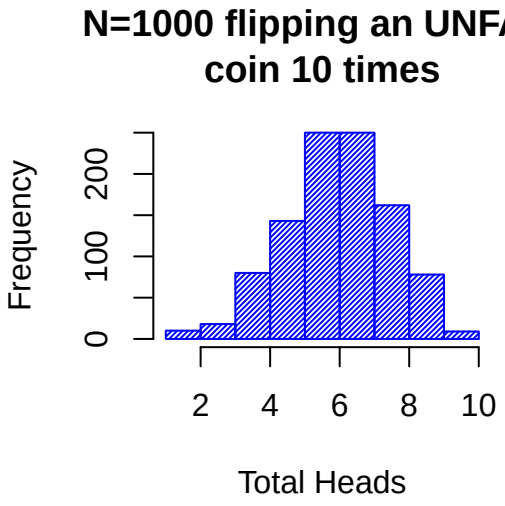


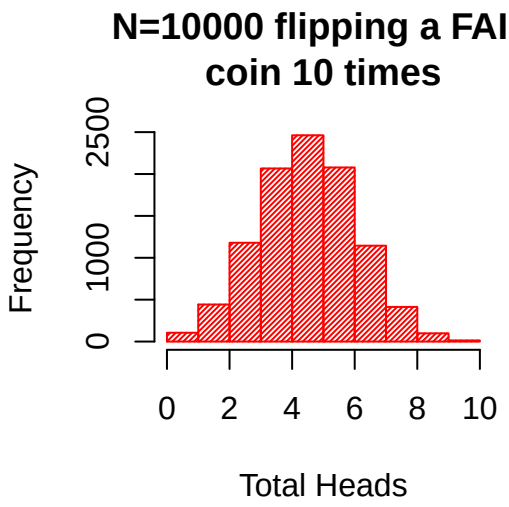


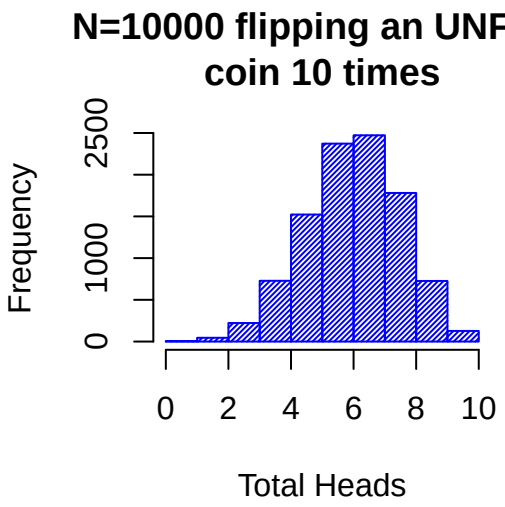








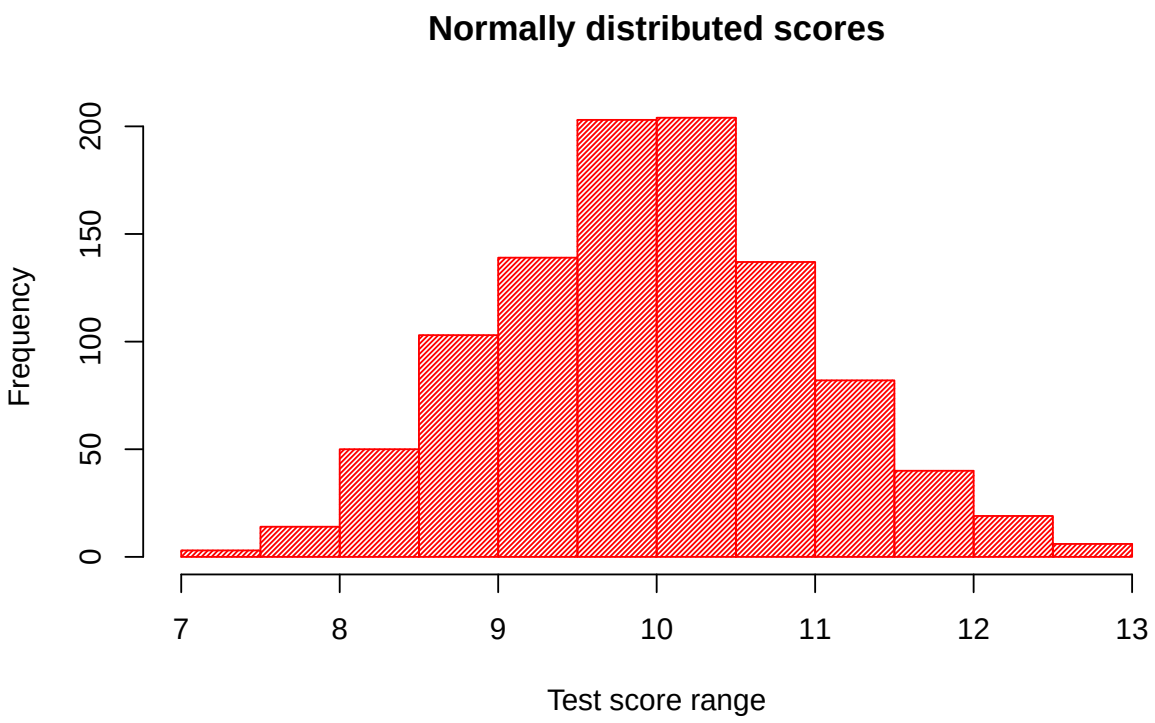




As you increase the size of your sample, the shape of the distribution tends towards **normality** and the **mean** estimate of the distribution approaches the true population mean - this constitutes the essence of the **Central Limit Theorem** (pp. 312-314).

The normal distribution

A core assumption of many frequentist tests is that the data being analyzed is *normally* distributed. Also called a *Gaussian* distribution, this resembles a bell-shaped curve that demonstrates most values measured fall in the middle of the total value range.



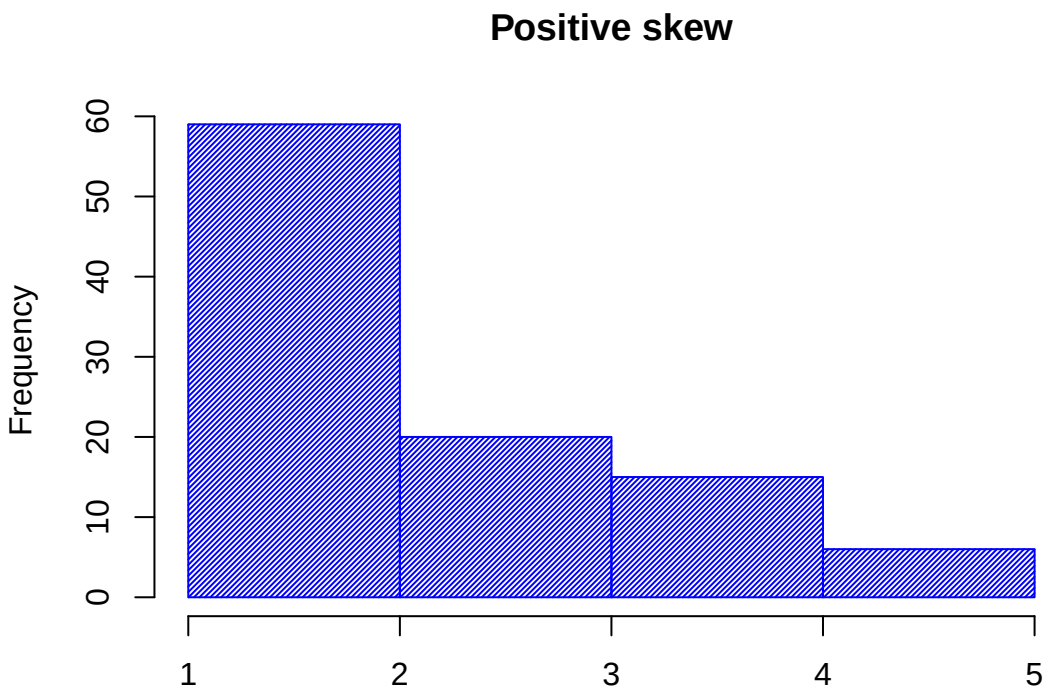
A more technical definition is that 68.3% of the area under the normal curve falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and 99.7% of the distribution is within 3 standard deviations.

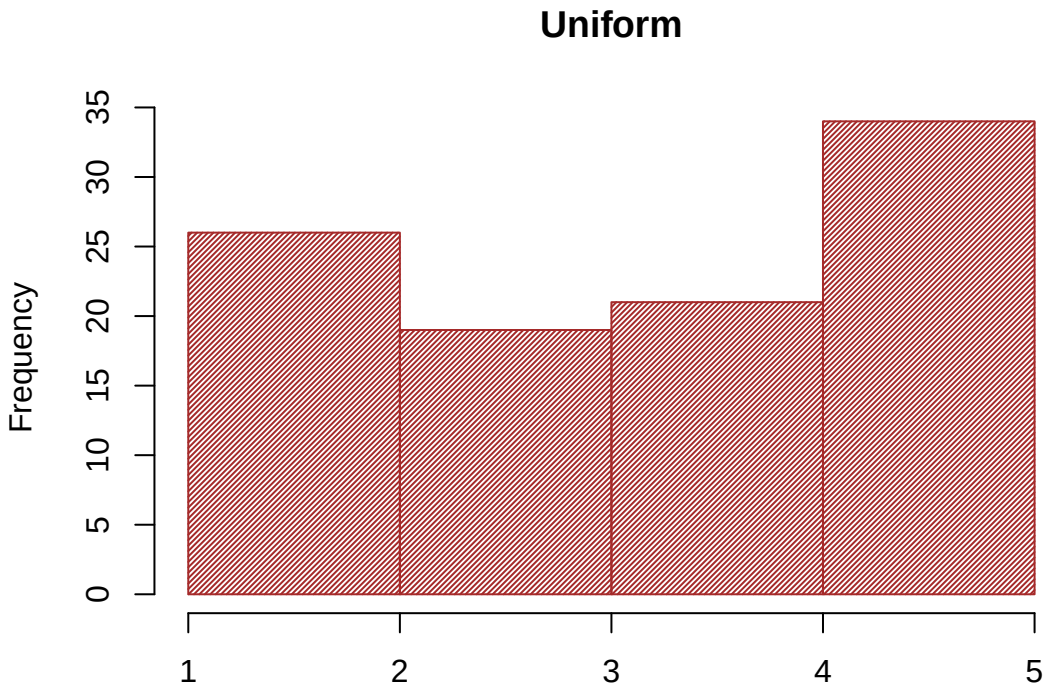
```
hist(rnorm(n=1000, mean = 10, sd = 1),       # See below
     main = "Normally distributed scores",   # Main title
     xlab = "Test score range",              # x-axis label
     breaks = 10,                            # Number of bars
     density = 50,                           # Shading of bars
     col = "red")                            # Bar color
```

`rnorm()` creates a normally distributed vector with three arguments: `n=` sample size, `mean=` desired mean estimate, and `sd=` desired standard deviation. The example above created a 1000-item vector, so `n=1000` with `mean=10` and `sd=1`.

Non-normal distributions

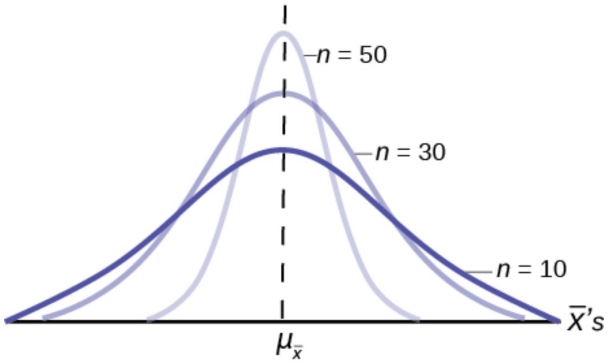
There are various other distributions that are *not* normal. This does not mean there is anything wrong with your data (in fact, many distributions should ideally *not* be normal, such as suicide rates). Having non-normally distributed data means you have to think more carefully about the type of analysis to run, and/or whether you may need to transform the data.


Positive skew
Frequency
0
10
20
30
40
50
60
1
2
3
4
5



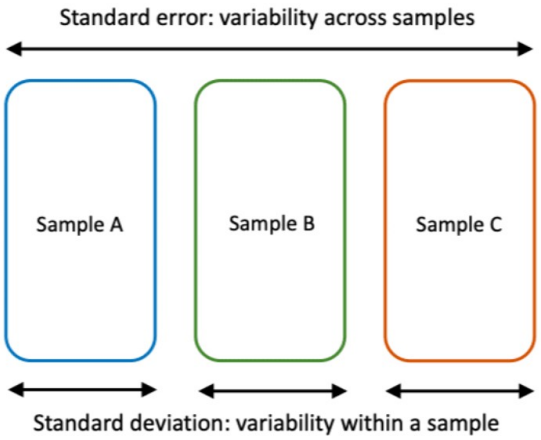
Standard error

One consequence of increasing the sample size (N) is that standard deviation (σ) decreases.



However, a core assumption of frequentist reasoning is that 'true population' estimates are *fixed*. Therefore, a varying σ estimate cannot be correct (it is not fixed). A primary reason we are interested in computing *range* estimates is to find a continuum of values within which the population parameter is likely to be found.

One workaround for this issue is either to compute the *sample* standard deviation (discussed last week). You can next estimate the *Standard Error* (SE) of the mean, which can be computed as $SE = \frac{\sigma}{\sqrt{N}}$. SE constitutes a more reliable estimate of sampling variance relative to standard deviations - the former describes mean variance across *all* tested samples (N) whereas σ describes variation across each particular sample (n1, n2, n3 - see below).



Ranges estimates (continued)

Knowing SE allows estimating another commonly used range estimate, the 95% confidence interval (CI). This describes the limits within which we can be 95% confident that the true population parameter resides.

```
# We want to identify the range where the sample mean will be found 95% of the time across a NORMAL dis
qnorm(p=c(.025,.975)) # 97.5% - 2.5% = 95%
```

```
## [1] -1.959964  1.959964
```

Watch from 1:09 onwards.

In practical circumstances, we generally know sample (not population) parameters. In this case, we should estimate sample standard deviation (called standard error) and adjust our confidence intervals by sampling from a t-distribution instead of a normal distribution.

Suppose a sample mean of $\hat{X} = 6.3$ with a standard deviation of $\sigma = .8$. Assuming our sample contained $N = 30$ students from a **normal** distribution, we can calculate standard error as $SE = \frac{\sigma}{\sqrt{N}} = \frac{.8}{\sqrt{30-1}}$ which gives $SE = 1.46$. Now we can derive the lower $(\hat{X} - (t^{LIM} \times SEM))$ and upper $(\hat{X} + (t^{LIM} \times SEM))$ limits of a confidence interval, where t^{LIM} represents the lower and upper quantiles of a t-statistic distribution. We use the t-statistic distribution whenever we're interested in inferring population parameters but only have sample data.

Let's review how to find t^{LIM} estimates from R.

```
# Suppose we have a sample of N=30
N=30

# And we want to identify the range where the sample mean will be found 95% of the time
qt(p=c(.025,.975),df=N-1) # 97.5% - 2.5% = 95%
```

```
## [1] -2.04523  2.04523
```

We can plug these values into our earlier formula to conclude that the lower and upper limits of a 95% confidence interval are $6.3 - (2.05 \times 1.46) = 3.04$ and $6.3 + (2.05 \times 1.46) = 9.02$. We can roughly conclude that the true population mean is likely to be found within a range of 3.04 and 9.02 about 95% of the time.

Note that CIs can be of any range you desire (e.g., 90%, 99%). 95% is the convention within Psychology. Standard Error (SE) and Confidence Intervals (CIs) are essential range estimates when we don't know the population (which is frequently the case), so make sure you grasp this concept clearly!

Lab activity

1. Create 1000 random values using `(mean=1, sd=10)` as parameters and store these in a variable (*hint* use the `rnorm()` function).
2. Report the mean, standard deviation and *standard error* of your custom variable. Show how you calculated the latter manually.
3. Summarize your variable using a histogram with modified plot title, x-axis and y-axis labels
4. Create a *second* normally distributed variable using `(mean=1, sd=10)` as parameters but this time with 100000 units, then assign these values to a second variable. Similar to earlier, (re)estimate the mean, SD and SEM for the second variable.
5. Provide a second histogram with customized titles and axes labels
6. Does the width of sampling distributions vary between the two variables? Explain briefly why (*hint* remember the central limit theorem).