

Symbiosis I is the first of a sequence of three Biology, Mathematics and Statistics integrative courses. Students who pass Symbiosis I get credit for Biology I and Introductory Statistics. The first Module (or Chapter) in Symbiosis I is 'The Scientific Method', a topic that naturally brings to the surface the issue of testing statistical hypotheses. In traditional intro stat courses, hypotheses testing is covered toward the end of the semester because the methods traditionally taught require the knowledge of sampling distributions. In Symbiosis I we were able to teach statistical inference from the beginning of the course by using randomization methods (permutation test and bootstrapping) for inference about means and doing a brief early introduction to the Binomial distribution in order to do test hypotheses about a population proportion. Material used in the teaching of these topics is displayed below.. Toward the end of the semester, the classical topics (t-test, large sample inference for proportions) were also covered; by then the students already had an understanding of concepts such as p-value, error types, and power of a test.

### Randomization to test

$$H_0: \mu_1 = \mu_2$$

In 1876 Darwin was aware that inbreeding does not have good results in humans and wondered if something similar happened among plants. Specifically he wanted to know if plants obtained by cross-fertilization tend to be taller than those produced by self-fertilization.

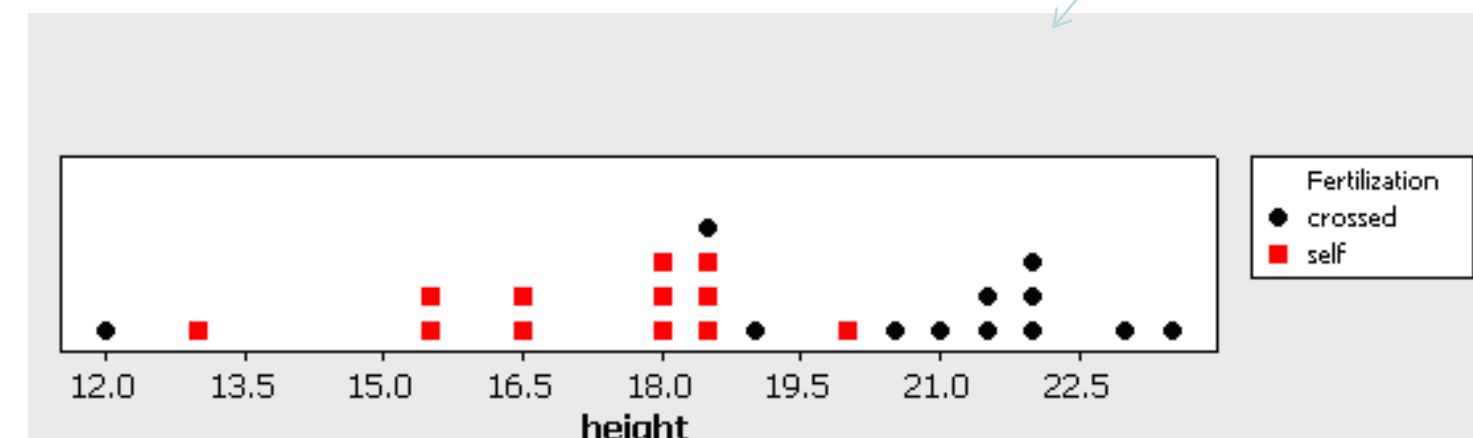
Statistical hypotheses:

(Null)  $H_0$ : the average height of crossed-fertilized and self-fertilized plants are equal

(Alternative)  $H_a$ : on average crossed-fertilized plants are taller than self-fertilized plants

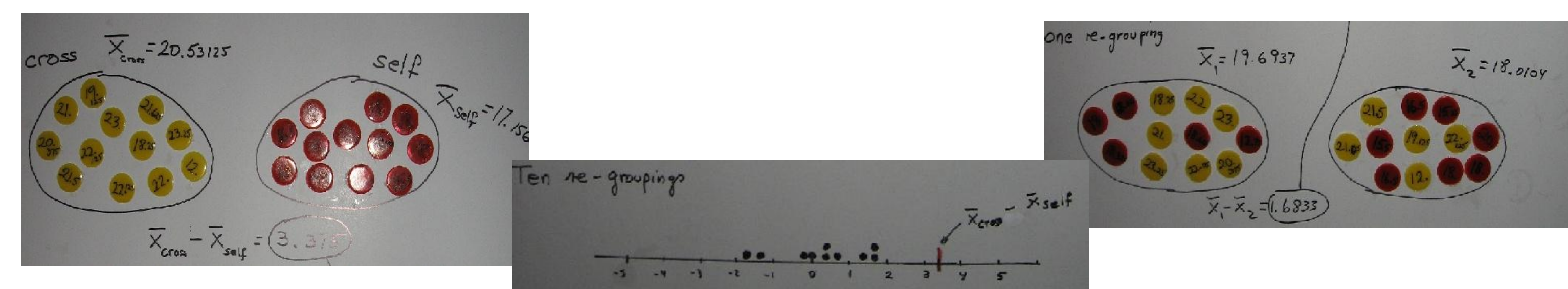
$$H_0: \mu_{cf} = \mu_{sf}$$

$$H_1: \mu_{cf} > \mu_{sf}$$



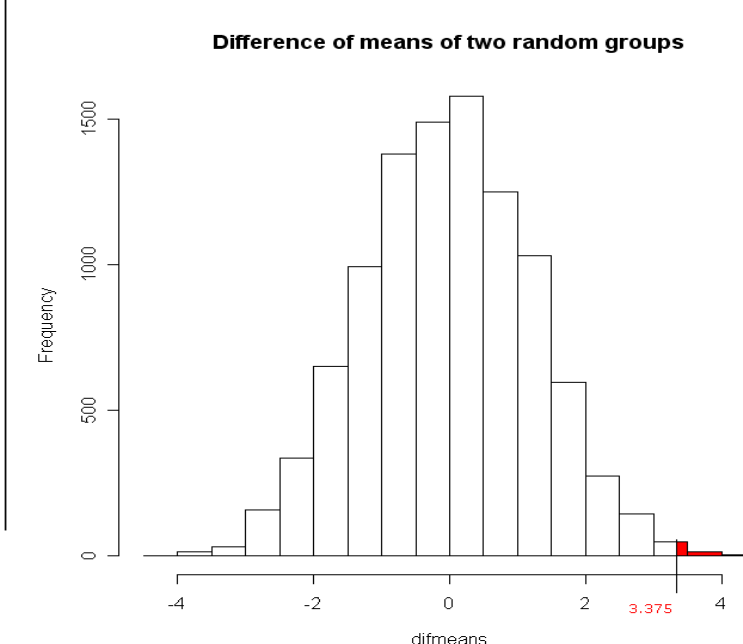
The rationale behind the randomization or permutation test is the following. If there was no difference between the means of the two populations, we can consider the data coming from a single population and the difference in means between the two groups would not be particularly large in absolute value as compared with the difference of means of two groups obtained by randomly re-grouping of the data.

First, the values of Darwin's observations were written over plastic chips, the means of each group were calculated. The 24 chips were mixed and 12 of them were selected at random to form 'group 1', the remaining formed 'group 2'. The difference in the means of the two random groups was calculated. Each group of students did a random separation in two groups and the differences in means were stored.

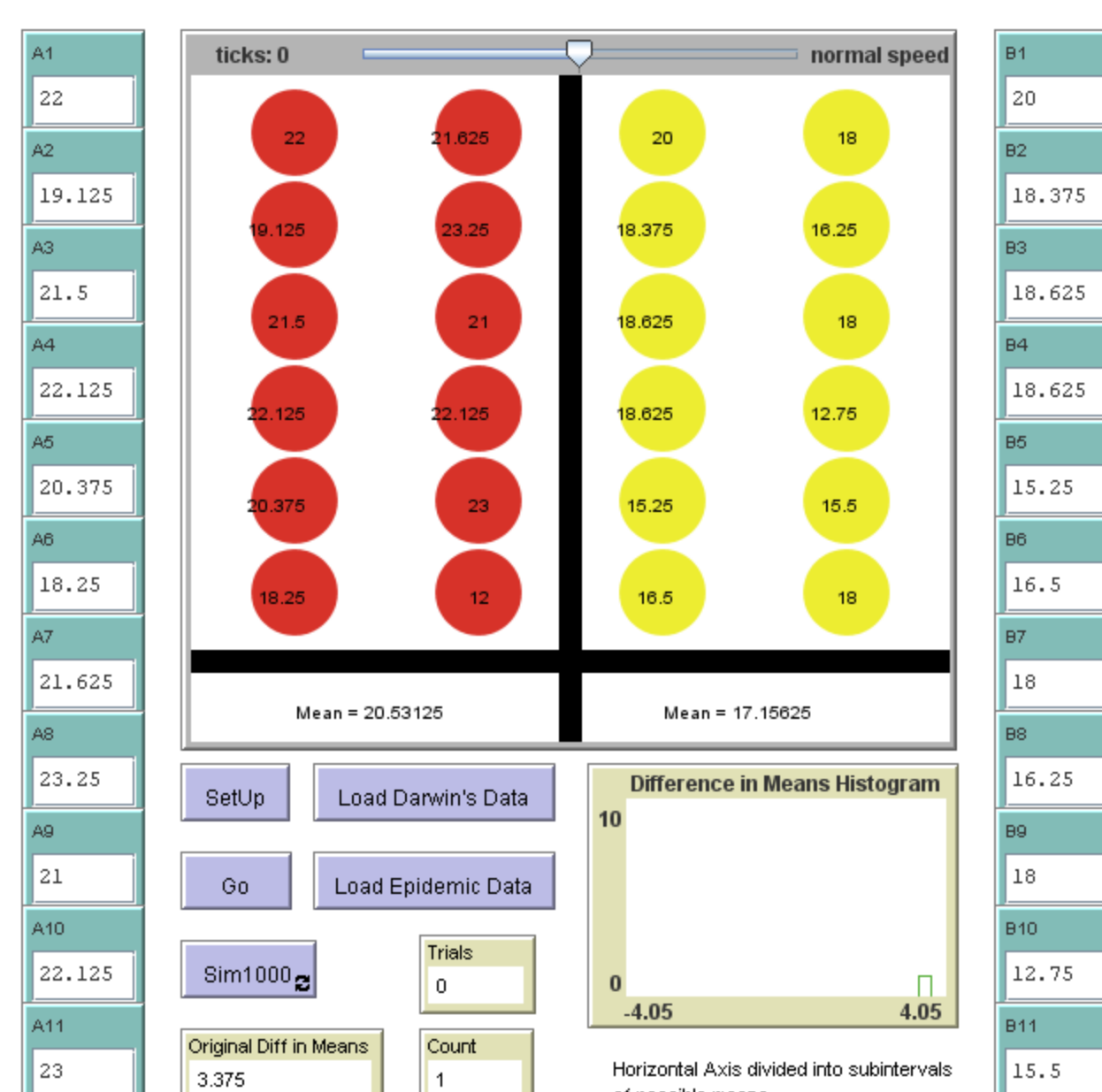


Then a program written in R was used to do 10000 random re-groupings of the 24 observations in two randomly formed groups. The proportion of random re-groupings in which the difference of means was equal or higher than the difference for the two true groups was found (this proportion becomes the approximated p-value); it was very small indeed. Later, an applet in NetLogo was prepared that could be used instead of the program in R. The applet can be found at <http://math.etsu.edu/symbiosis/netlogo/RandomizationTest.h>

```
## Calculating an approximated p-value of the randomization test by simulating 20000 re-groupings
## randomize(10000) # generates 20000 random re-groupings of the 24 observations
## diffmeans(10000) # calculates the difference of means for each of the 20000 re-groupings
## pvalue(10000) # calculates the proportion of re-groupings in which the difference of means is greater or equal than the difference of the two true groups
## zeros will go to the other locations
## zeros will go to the other locations
## zeros will go to the other locations
```



### The Randomization Test



Something nice about the permutation test is that, contrary to the t-test, it can be applied to the comparison of medians, standard deviations, etcetera.

### Binomial distribution to test

$$H_0: p = p_0$$

Placing Probability in the context of the scientific method

Context: Testing statistical hypotheses as part of the scientific method

**Example of research questions**

- Has inbreeding a negative effect in plants?
- Is there an epidemic?
- What makes wasps successful in mating?

**Examples of scientific hypotheses:**

- Plants produced by cross-fertilization tend to be taller than plants produced by self-fertilization.
- The proportion of people with this disease is over the threshold of 2%.
- Larger wasps are more successful at mating.

**Examples of statistical hypotheses:**

$H_0: \mu_{cross} = \mu_{self}$   
 $H_a: \mu_{cross} > \mu_{self}$

$H_0: p \leq 0.02$   
 $H_a: p > 0.02$

**Why do we study probability?**

To be able to answer questions such:

- In regular conditions 2% of the population presents a certain condition, however if the proportion of people with that condition is higher than 2% in the population, then it is considered that an epidemic is taking place. You examine 50 randomly selected people and find that 6 of them present the condition.

**What is the probability of this happening just by chance if there is no epidemic?**

To answer that question will enable to make up our minds with respect to hypotheses such as

$H_0: p \leq 0.02$   
 $H_a: p > 0.02$

What is Probability?

**Classic Definition of Probability**

Probability of an event: 'favorable outcomes' / 'possible outcomes'.

In order to be able to apply this definition two conditions are needed:

- There is a finite number of outcomes
- All the outcomes are equally likely to happen

**Probability as a limit of the relative frequency**

The graph shows the accumulated relative proportion of heads in 10,000 tosses of a fair coin

**Axiomatic definition of Probability**

**AXIOMS:**

- $P(A) \geq 0$
- $P(S) = 1$
- If A and B are mutually exclusive events:  $P(A \cup B) = P(A) + P(B)$

**AND CONSEQUENCES:**

- $P(A^c) = 1 - P(A)$
- If  $A \subset B$  then  $P(A) \leq P(B)$
- $P(A) \leq 1$
- For two events A and B:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

### Binomial Distribution

**Binomial Distribution**

$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$  for  $x=0,1,2,3,\dots,n$

The conditions necessary for the Binomial distribution to be the appropriate way of calculating probabilities are:

- A known number (n) of independent replicates of a Bernoulli experiment
- The probability of 'success' in each replicate is known (p) and it is the same for all the replicates
- We want to know the probability of getting 'x' successes in the 'n' trials.

**Examples:**

- Imagine you are tossing a coin, not 3 times but 10 times, what would be the probability of getting 3 heads?

$P(X=3) = \binom{10}{3} 0.5^3 (0.5)^7$

- Imagine we have a big sack of seeds, we know 85% of the seeds get to germinate, we plan to plant 12 seeds. What is the probability that exactly 10 of them get to germinate? Write an expression to calculate the probability

$P(X=10) = \binom{12}{10} 0.85^{10} (0.15)^2$

Things that help us with calculations:

- Pascal triangle
- Binomial Probability tables (germany)
- Statistical software. Example with R: `dbinom(3,10,0.5)` calculates

$P(X=3) = \binom{10}{3} 0.5^3 (0.5)^7 = 0.1171875$

We can calculate the probability of several values at once

`sum(dbinom(0:12,12,0.85))` calculates  $P(X \leq 12)$

`sum(dbinom(0:12,12,0.85))` calculates  $P(X \leq 12)$

### Testing Hypothesis

**An important application of the Binomial distribution:**

Testing hypotheses about a population proportion.

Example: I have bought two color plastic chips (one side red and one side yellow) to use instead of coins to do simulations in genetics. I want to know if the probability of getting red is equal to 1/2.

$H_0: p = 0.5$   
 $H_a: p \neq 0.5$  (where p is P(Red))

Help me by doing an experiment:

- Toss it 10 times
- Count the number of reds you get
- Calculate the probability of getting that number of reds or a 'more extreme value' (more reds than the alternative hypothesis H<sub>a</sub> would expect when the null hypothesis is true).
- Do your experimental data constitute evidence against the null hypothesis? (Would you reject the null hypothesis?)
- If you want to see a similar example in a biological context read a deck's story: <http://www.etsu.edu/symbiosis/causes/conjecture/example/ducks-main.doc>

**An important application of the Binomial distribution:**

Testing hypotheses about a population proportion.

Example: I have bought two color plastic chips (one side red and one side yellow) to use instead of coins to do simulations in genetics. I want to know if the probability of getting red is equal to 1/2.

$H_0: p = 0.5$   
 $H_a: p \neq 0.5$  (where p is P(Red))

Help me by doing an experiment:

- Toss it 10 times
- Count the number of reds you get
- Calculate the probability of getting that number of reds or a 'more extreme value' (more reds than the alternative hypothesis H<sub>a</sub> would expect when the null hypothesis is true).
- Do your experimental data constitute evidence against the null hypothesis? (Would you reject the null hypothesis?)
- If you want to see a similar example in a biological context read a deck's story: <http://www.etsu.edu/symbiosis/causes/conjecture/example/ducks-main.doc>

**Exercise:**

Imagine that a person knows that seeds of a certain type have a probability 0.85 of germinating. This person has found a sack of seeds in the basement of his house, he had forgotten about the sack for a long time and he is afraid that the seeds are too old and they might not have 0.85 probability of germinating but a lower one. Not to go through all the work and expense of planting the seeds and not getting a good yield, he selects 10 seeds at random from the sack and plants them, only 5 of them got to germinate.

**Write the null and alternative hypothesis.**

**Calculate the p-value**

**arrive at a conclusion.**

Apply what we just learn to answer the research question we started with:

In regular conditions 2% of the population presents a certain condition, however if the proportion of people with that condition is higher than 2% in the population, then it is considered that an epidemic is taking place.

$$H_0: p \leq 0.02$$

$$H_a: p > 0.02$$

You examine 50 randomly selected people and find that 6 of them present the condition

**What is the probability of this happening just by chance when there is no epidemic?**

**What is your conclusion about the statistical hypotheses?**

**What is your answer to the research question?**

### Bootstrapping to obtain confidence intervals and test $H_0: \mu = \mu_0$

**Bootstrapping**

Bradley Efron (Stanford U., circa 1980)

When we do not know the sampling distribution of the statistic a randomization method called 'bootstrap' can help us to do inference about a parameter. The bootstrap method relies in the only thing we have: the sample data, and the common way of producing an empirical distribution is by re-sampling. Important: the distribution obtained by randomization is NOT the sampling distribution of the statistic but something we use instead of it when we don't know it!

Consider as example the length (millimeters) of 40 cells randomly selected from the onion root of

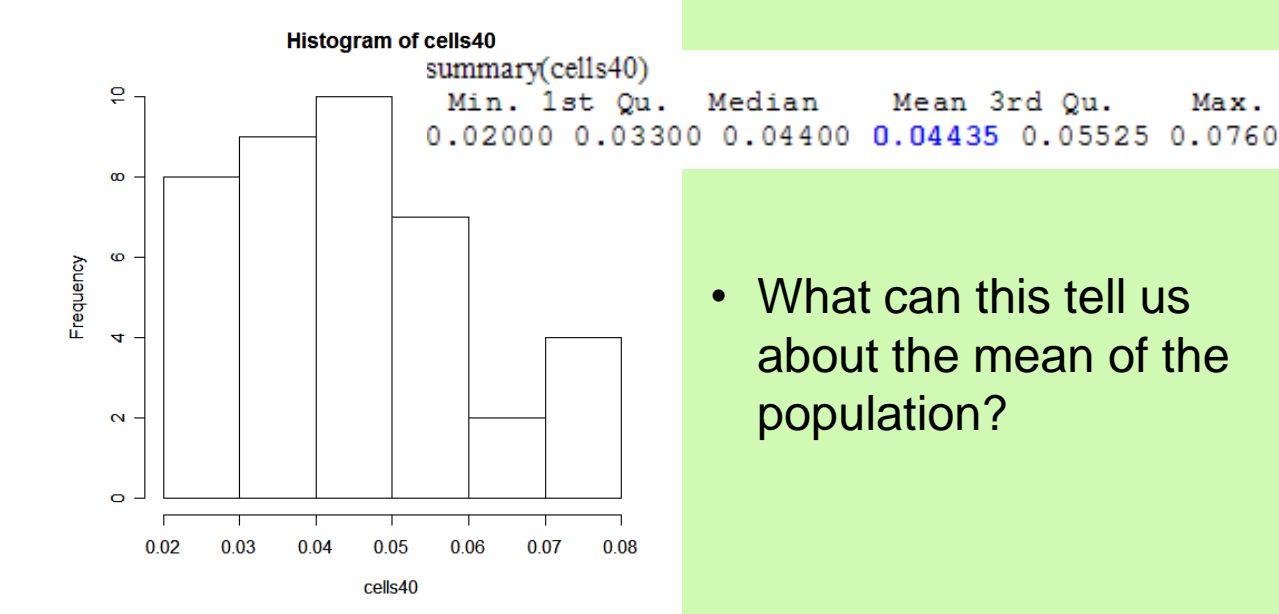
cells roots (Data by K. McDonald & J. Pinner, ETSU)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02000	0.03300	0.04400	0.04435	0.05525	0.07600

length (millimeters)

Cells in the elongation region of an onion root

### Summary of the sample:

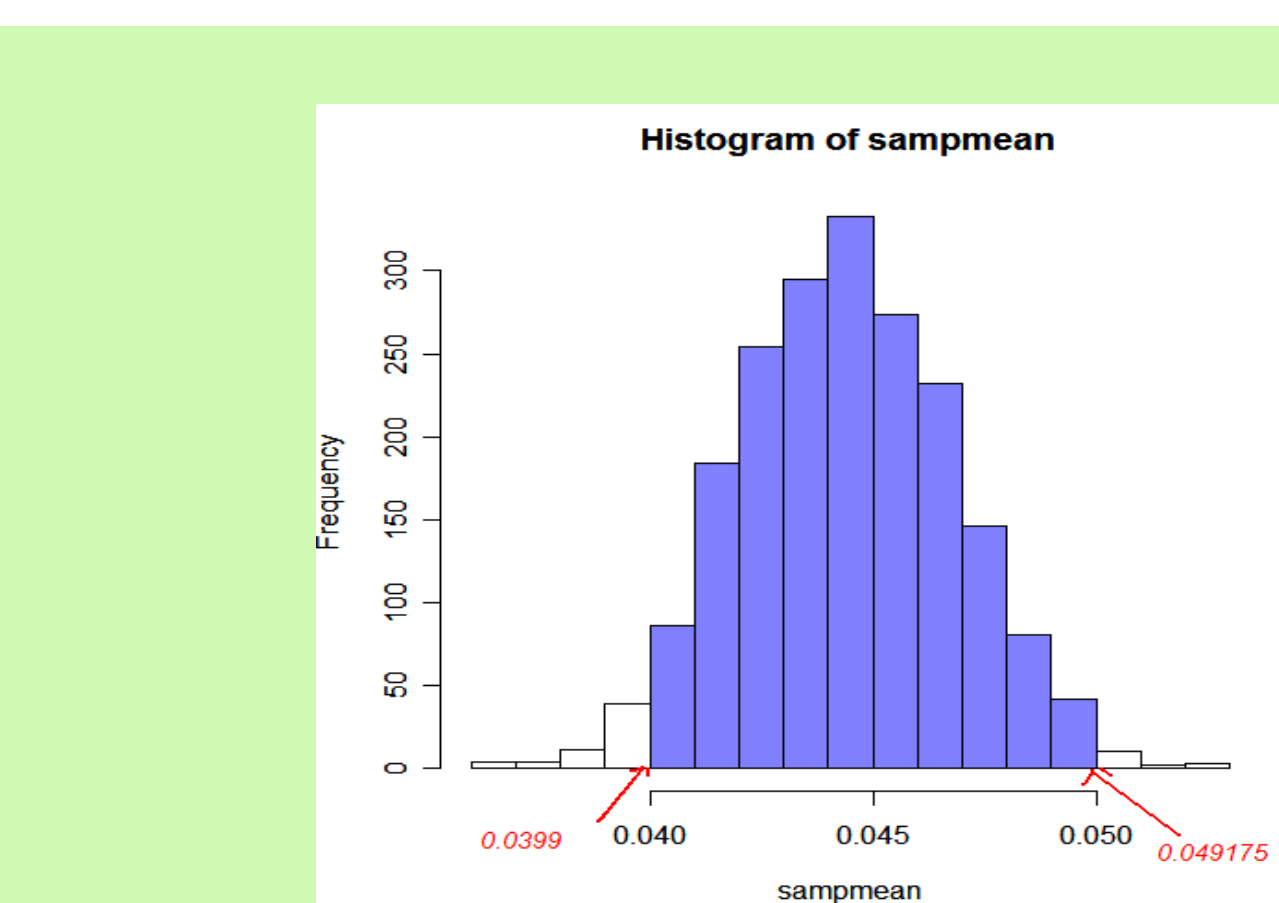
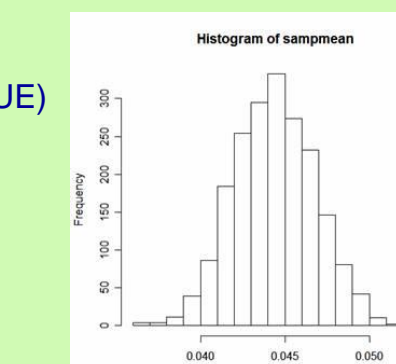


What can this tell us about the mean of the population?

### Let's do bootstrapping

Now we will take a many (2000 is the standard) samples (with replacement) of size 40 from those 40 values. For each one of those samples we will calculate the sample mean and observe the distribution of those values. (Commands in R)

```
mcell<-mean(cells40)
sampsam<-double(2000)
for (i in 1:2000){
  subsam<-sample(cells40,40,replace=TRUE)
  sampsam[i]<-mean(subsam)
}
lowend<-quantile(sampsam,0.025)
highend<-quantile(sampsam,0.975)
lowend
highend
hist(sampsam)
```



95% confidence interval for the population mean (0.0399, 0.049175)

We could use the same method to work with the median, or any other statistic of interest, instead of the mean.

**Note 1:** - The most common % of confidence when working with confidence intervals are 90%, 95% and 99% confidence intervals, the higher the % of confidence the wider the interval because we will be discarding smaller tails at each end.

**Note 2:** - Sometimes people are not interested in 'two-sided confidence intervals' but in 'one-sided ones', in that case we would discard the lower or upper 5% depending of the case.

### Test of hypothesis about a mean using bootstrapping

Somebody has made the statement "the average length of cells in the elongation region of the onion root is 0.045 millimeters". Is his statement reasonable?

We need to decide if we reject or not his statement

$$H_0: \mu = 0.045$$

We select a simple random sample of 40 cells from the elongation region of onion roots

95% confidence interval (0.0399, 0.049175)

0.045 IS IN THAT INTERVAL, thus we consider it a plausible value and do not reject the null hypothesis.

Computer intensive methods such as the randomization test or bootstrapping are NOT the only way performing test of hypotheses or building confidence intervals, but they are the ones that require fewer assumptions and mathematical tools to develop the formulas, and they are quite flexible with regard to what parameters we want to test hypotheses about (means, medians, proportions, variances, etcetera), that is why we study them at the beginning of the course. As we progress in the course and we acquire more knowledge about sampling distributions we will see other more traditional ways of calculating p-values and calculating confidence intervals (among them the 't-tests' and 't-confidence intervals').