Overview of statistics I

Stéphane Aris-Brosou

September 14, 2009

uOttawa
L'Université canadienne
Canada's university

Today's objectives

- gain some statistical vocabulary commonly encountered in bioinformatics
- introduction to hypothesis testing
- intuitive understanding of null distributions and of *p*-values
- brief introduction to R through a few examples and exercises

An snippet of genomic DNA

```
AGTGGTCGCACCACAGCCGTGTGGTAGGACCCATGACAGCCTCTCTCTCCCCACGCCCCACCCCTCCTGCATCTCTGCTCGCCCCCATGCCACGTCTTTCCATCA
CACAGCTCACCATGTCCTGAAGGAGGCAGTGCAGAGCAACAGGTGAAGCGGGCAATGCTCAAGGGTTGGAGGAGGAACAGGAGTCAAGGGAGAGAGGTGCAGGAG
GGTACCCCGGTCCAGTGGGTGACCCTGCCACTTGGTCATTGAACAGAGCTGATGCGTGACGTGAGGTCCAGGCTCAGGAGCCCTCAACTTCACTCTCACCACAGC
AACGCTCCTTGAGGTGGGCAAAGTGGCTCAGGCAGGTGTTAACGAGCCCCTCAGAGCCTCTGGGGCTCAGCAACATGCACCTGGCCAACCAGGGTACAACTTCTC
GAAGGGCCTGACGTACCACCGCATCGTAGAGGCTTTCCGGTTTCGCCAAGAGGACCCTGCTTGGGGACCCCAAGTTTGTGGATGTGGGTAAGGGGCAGGGGCTGG
GACTCGTGGGTGACCCCCAGTCTTGGCTTCTGCCCTGGGGGCACACCACAGACACTGCACTGGTATGTGTCACCCCTTTTCTCCCTGCCACTCTGCACAGCCCC
CCATCATCTACAACCTCTGGTTCGCTATGACGTGAAGAGGGCCGGCCCCGGCTGCACAACAAGCTTCTGCCCAACGTCACGACAGTGCATTGACCAGGTGGGCCG
CCCAGGGCATCTGGGCTGGAGGCCTGGATCATCACAGAGTGGAGGTGTCCTCTCTCTAGTGCCTGGGCCATCTGGAGCCCCTGTGACCAAGCCCCCTGCTCCAGT
TGACCTGGCCTGAAATGGCACCACCTGGGCTGAGGCCTGTGACCGTGTGGTTCAGGTGGCTCTGGAGCCCTGCTCAGGCTTCCCCTCTCACCCCCAGGCAGTGAC
AGTGGTCGCACCACAGCCGTGTGGTAGGACCCATGACAGCCTCTCTCTCCCCACGCCCCACCCCTCCTGCATCTCTGCTCGCCCCCATGCCACGTCTTTCCATCA
CACAGCTCACCATGTCCTGAAGGAGGCAGTGCAGAGCAACAGGTGAAGCGGGCAATGCTCAAGGGTTGGAGGAGGAACAGGAGTCAAGGGAGAGAGGTGCAGGAG
GGTACCCCGGTCCAGTGGGTGACCCTGCCACTTGGTCATTGAACAGAGCTGATGCGTGACGTGAGGTCCAGGCTCAGGAGCCCTCAACTTCACTCTCACCACAGC
AACGCTCCTTGAGGTGGGCAAAGTGGCTCAGGCAGGTGTTAACGAGCCCCTCAGAGCCTCTGGGGCTCAGCAACATGCACCTGGCCAACCAGGGTACAACTTCTC
GAAGGGCCTGACGTACCACCGCATCGTAGAGGCTTTCCGGTTTCGCCAAGAGGACCCTGCTTGGGGACCCCAAGTTTGTGGATGTGGGTAAGGGGCAGGGGCTGG
GACTCGTGGGTGACCCCCAGTCTTGGCTTCTGCCCTGGGGGCACACCACAGACACTGCACTGGTATGTGTCACCCCTTTTCTCCCTGCCACTCTGCACAGCCCC
CCATCATCTACAACCTCTGGTTCGCTATGACGTGAAGAGGGCCGGCCCCGGCTGCACAACAAGCTTCTGCCCAACGTCACGACAGTGCATTGACCAGGTGGGCCG
CCCAGGGCATCTGGGCTGGAGGCCTGGATCATCACAGAGTGGAGGTGTCCTCTCTCTAGTGCCTGGGCCATCTGGAGCCCCTGTGACCAAGCCCCCTGCTCCAGT
TGACCTGGCCTGAAATGGCACCACCTGGGCTGAGGCCTGTGACCGTGTGGTTCAGGTGGCTCTGGAGCCCTGCTCAGGCTTCCCCTCTCACCCCCAGGCAGTGAC
AGTGGTCGCACCACAGCCGTGTGGTAGGACCCATGACAGCCTCTCTCTCCCCACGCCCCACCCCTCCTGCATCTCTGCTCGCCCCCATGCCACGTCTTTCCATCA
CACAGCTCACCATGTCCTGAAGGAGGCAGTGCAGAGCAACAGGTGAAGCGGGCAATGCTCAAGGGTTGGAGGAGGAACAGGAGTCAAGGGAGAGAGGTGCAGGAG
GGTACCCCGGTCCAGTGGGTGACCCTGCCACTTGGTCATTGAACAGAGCTGATGCGTGACGTGAGGTCCAGGCTCAGGAGCCCTCAACTTCACTCTCACCACAGC
AACGCTCCTTGAGGTGGGCAAAGTGGCTCAGGCAGGTGTTAACGAGCCCCTCAGAGCCTCTGGGGCTCAGCAACATGCACCTGGCCAACCAGGGTACAACTTCTC
GAAGGGCCTGACGTACCACCGCATCGTAGAGGCTTTCCGGTTTCGCCAAGAGGACCCTGCTTGGGGACCCCAAGTTTGTGGATGTGGGTAAGGGGCAGGGGCTGG
GACTCGTGGGTGACCCCCAGTCTTGGCTTCTGCCCTGGGGGCACACCACAGACACTGCACTGGTATGTGTCACCCCTTTTCTCCCTGCCACTCTGCACAGCCCC
CCATCATCTACAACCTCTGGTTCGCTATGACGTGAAGAGGGCCGGCCCCGGCTGCACAACAAGCTTCTGCCCAACGTCACGACAGTGCATTGACCAGGTGGGCCG
CCCAGGGCATCTGGGCTGGAGGCCTGGATCATCACAGAGTGGAGGTGTCCTCTCTCTAGTGCCTGGGCCATCTGGAGCCCCTGTGACCAAGCCCCCTGCTCCAGT
TGACCTGGCCTGAAATGGCACCACCTGGGCTGAGGCCTGTGACCGTGTGGTTCAGGTGGCTCTGGAGCCCTGCTCAGGCTTCCCCTCTCACCCCCAGGCAGTGAC
AGTGGTCGCACCACAGCCGTGTGGTAGGACCCATGACAGCCTCTCTCTCCCCACGCCCCACCCCTCCTGCATCTCTGCTCGCCCCCATGCCACGTCTTTCCATCA
CACAGCTCACCATGTCCTGAAGGAGGCAGTGCAGAGCAACAGGTGAAGCGGGCAATGCTCAAGGGTTGGAGGAGGAACAGGAGTCAAGGGAGAGAGGTGCAGGAG
GGTACCCCGGTCCAGTGGGTGACCCTGCCACTTGGTCATTGAACAGAGCTGATGCGTGACGTGAGGTCCAGGCTCAGGAGCCCTCAACTTCACTCTCACCACAGC
AACGCTCCTTGAGGTGGGCAAAGTGGCTCAGGCAGGTGTTAACGAGCCCCTCAGAGCCTCTGGGGCTCAGCAACATGCACCTGGCCAACCAGGGTACAACTTCTC
GAAGGGCCTGACGTACCACCGCATCGTAGAGGCTTTCCGGTTTCGCCAAGAGGACCCTGCTTGGGGACCCCAAGTTTGTGGATGTGGGTAAGGGGCAGGGGCTGG
```

What can we say about this genomic sequence?

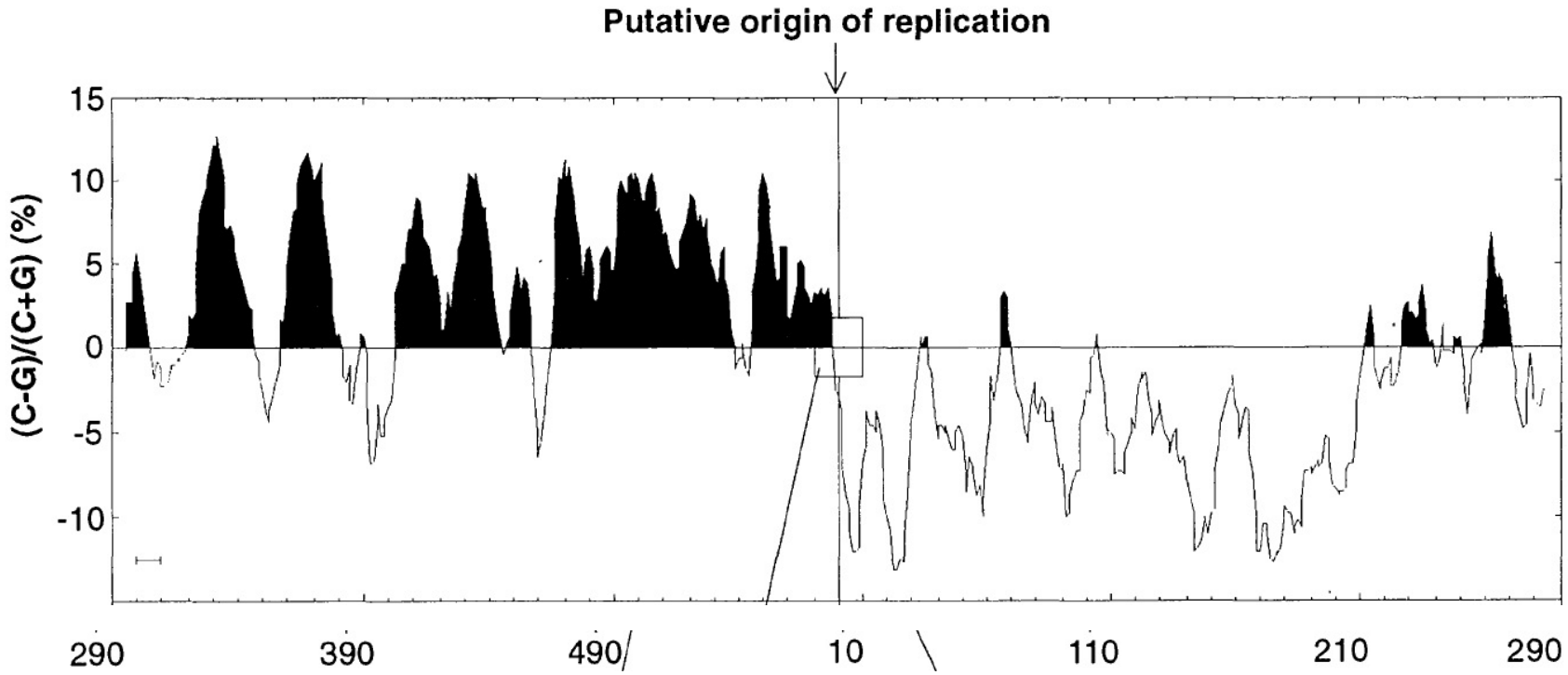
- how to best describe (*summarize*) the information contained in this sequence?
- can we determine what kind of organism it is sampled from?
- does it have unique features that tell it apart from other organisms? (comparative study)
- where are the protein-coding genes? What about noncoding genes? Regulatory elements?

A first idea: simple counts

- count base frequencies
- use GC content
- GC skew = (#G - #C)/(#G + #C)
- calculate GC skew along the sequence using a sliding window of a fixed (and predefined) width

Application: origins of replication with weak consensus patterns

In *Mycoplasma genitalim*:



Distance to putative origin of replication (Lobry, 1996)

Recall: GC skew = (#G - #C)/(#G + #C)

Application: origins of replication with weak consensus patterns

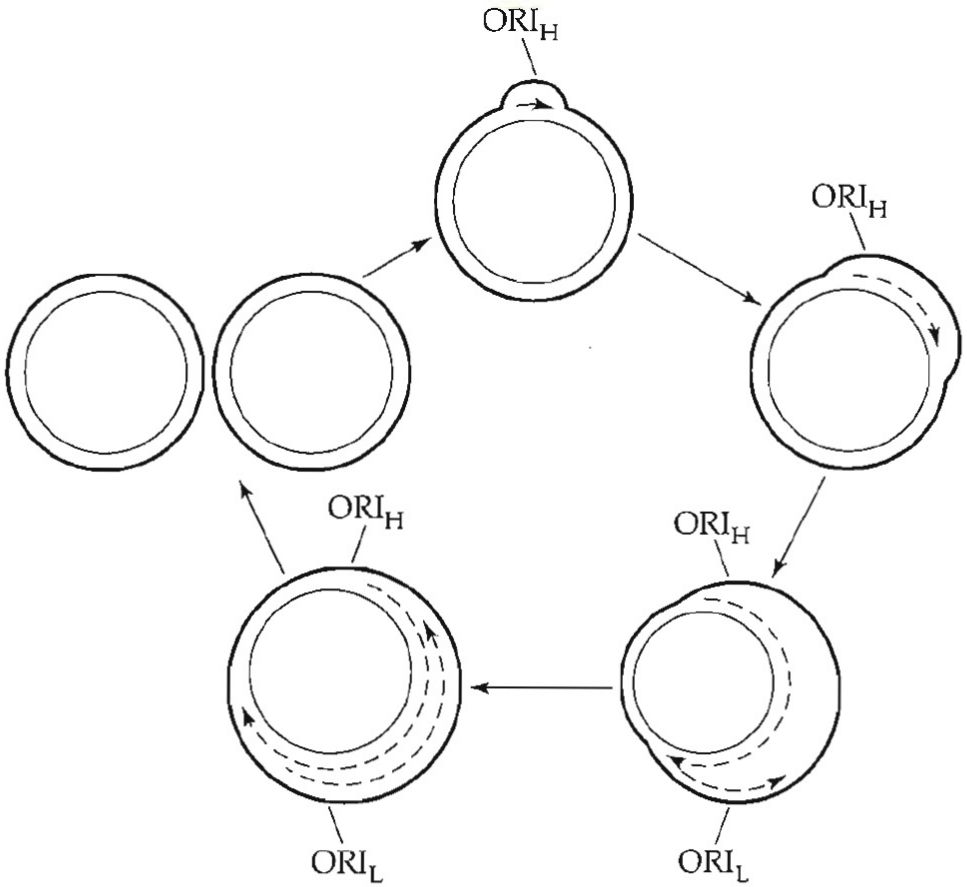


Figure 6.7 Replication of a mitochondrial genome. The inner and outer solid circles denote the light and heavy strands on the parental genome; the dashed lines denote newly synthesized strands. Replication is initiated at ORI$_H$ (*upper right*) and proceeds to build a new heavy strand. Once ORI$_L$ is exposed, synthesis begins in the opposite direction, using the original heavy strand as a template (*lower right*). Note that the parental heavy-strand DNA closest to ORI$_H$ remains single-stranded for the longest time. (Modified from Brown et al. 2005.)

(Lynch, 2007)

What is it that we want? (*desiderata*)

- to find a set of rules that govern observed DNA patterns
- how can we measure our "surprise" to an observed pattern?
- what is the technical jargon?

Probability

- The probability of a particular event occurring is the frequency of that event over a very long series of repetitions.
- Notation: P

Example

- $P(\text{tossing a head}) = 0.5$
- $P(\text{rolling a 6}) = 0.167$
- $P(\text{average age in a population sample} > 21) = 0.25$

Random variable

A random variable is a quantity whose values are random and that cannot be predicted with absolute accuracy.

Example

- X = age of an individual
- Y = length of a gene
- p_{GC} = fraction of nucleotides that are G or C

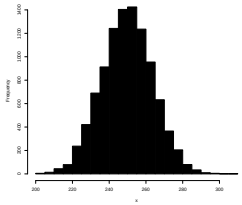
Random variables are described by their probability distribution function (pdf).

Probability distribution

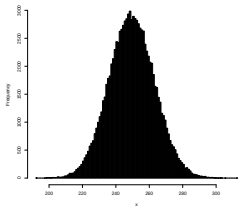
- The **distribution** of a random variable describes the possible values of the variable and the probability of each value.
- For **discrete** random variables, the distribution can be enumerated; for **continuous** ones, we describe the distribution with a (continuous) function.

Examples of probability distributions

- discrete (univariate)



- binomial(n, p): commonly assumed in problems with series of independent success/failure trials where success probability is fixed

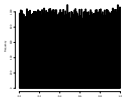


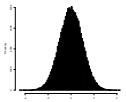
- Poisson(r): commonly assumed in problems with counts of "rare events" occurring in a given time period

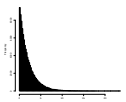
- discrete (multivariate)

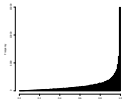
  - multinomial(n, θ): generalization of the binomial distribution

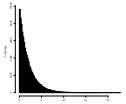
- continuous (univariate)











  - uniform(a, b)
  - normal(μ, σ^2)
  - exponential(r)
  - beta(α, β)
  - Chi-square(df)

- continuous (multivariate)

  - multivariate normal(M, Σ): generalization of the normal
  - Dirichlet(α): generalization of beta distribution

Break! (with R)

- use R to plot the different shapes that can be taken by the beta distribution depending on its two parameters: list and classify these shapes;
- same question with the Chi-square distribution.

Move to Section 1 of the R script posted on your Virtual Campus (VC).

Back to our *desiderata*

- to find a set of rules that govern observed DNA patterns (=probability distributions)
- how can we measure our "surprise" to an observed pattern? (intuitive approach based of a simulation experiment)
- what is a null distribution?

Back to our *desiderata*

- to find a set of rules that govern observed DNA patterns (=probability distributions)
- how can we measure our "surprise" to an observed pattern? (intuitive approach based of a simulation experiment)
- what is a null distribution?

- a probability model describes a method for simulating observations from a model (H_0)
- probability theory helps us derive asymptotic results

Back to our DNA example: the simulation experiment

- alphabet: $\{T, C, A, G\}$ (state space)
- probability: p_T, p_C, p_A, p_G $(\sum p_i = 1)$
- we are interested in the distribution of p_A
- more specifically: for a given sequence, we want to know the probability that: "the frequency of A's is drawn from a particular distribution, *e.g.* uniform".

Scientific questions and null hypotheses (H_0)

- parametric null hypotheses
  - a normal distribution has a specified mean and variance
  - a gene has no GC content bias (GC skew $= 0$)
  - its frequency of A's is random ($p_A - \frac{1}{4} = 0$)

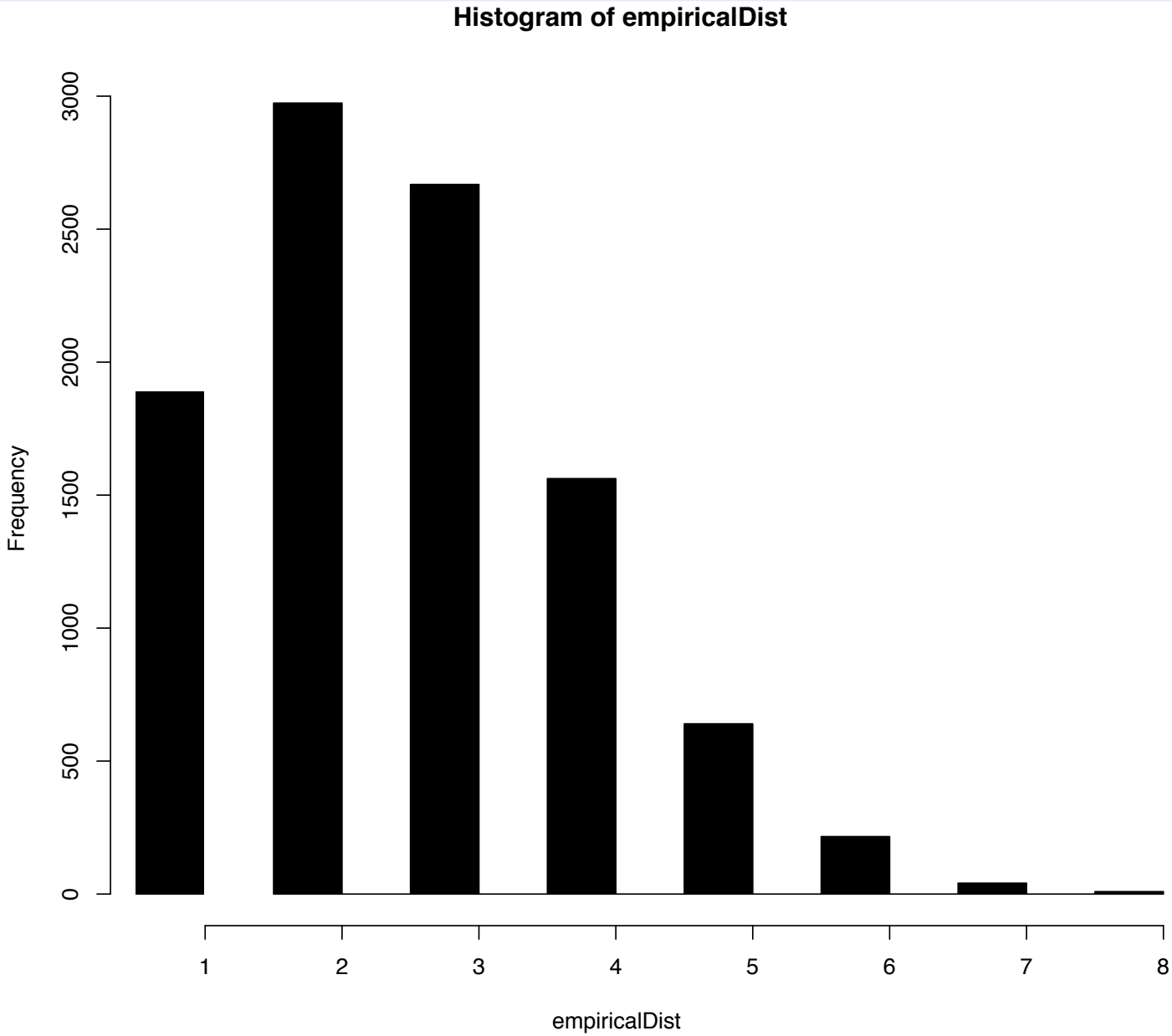
- non-parametric null hypotheses
  - a distribution is of normal form with both mean & variance unspecified
  - two phylogenetic trees are not significantly different

Null hypotheses and the "surprise effect" (with R)

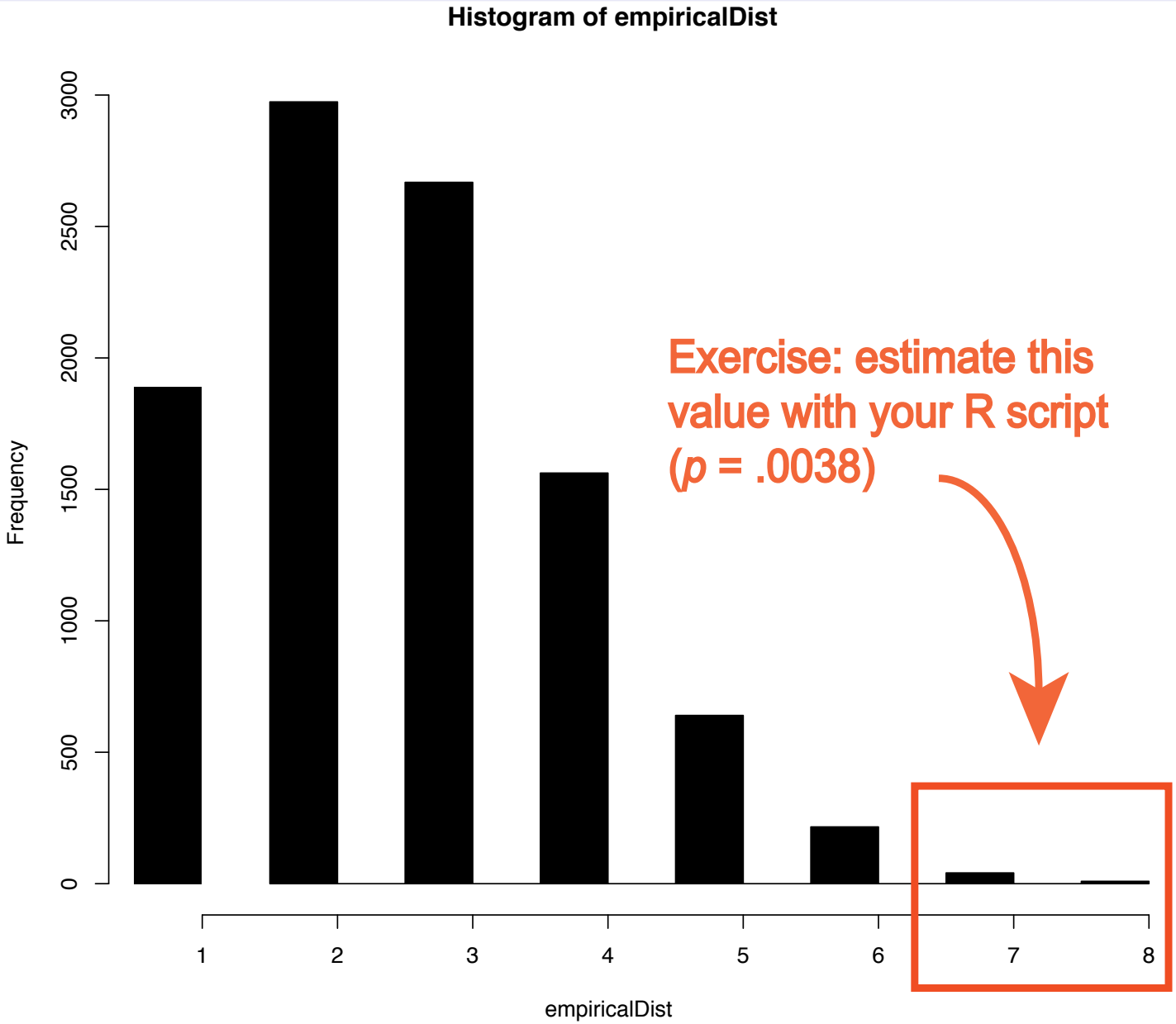
- draw a sequence X with the R script
- eg, $X =$ CAGGCGTAAG, where $p_A = \frac{3}{10}$
- what is the probability that the A's are random (*i.e.*, come from a uniform distribution)?
- H_0: base frequencies in $X \sim U$
- we assume that evolution generated sequence X following a process we fully understand (H_0)
- if we were to rewind time and repeat the evolutionary process H_0 many times, how often would we observe more A's than observed here (3 out of 10)?

Move to Section 2 of the R script posted on your VC.

How often do we draw a value > 6?



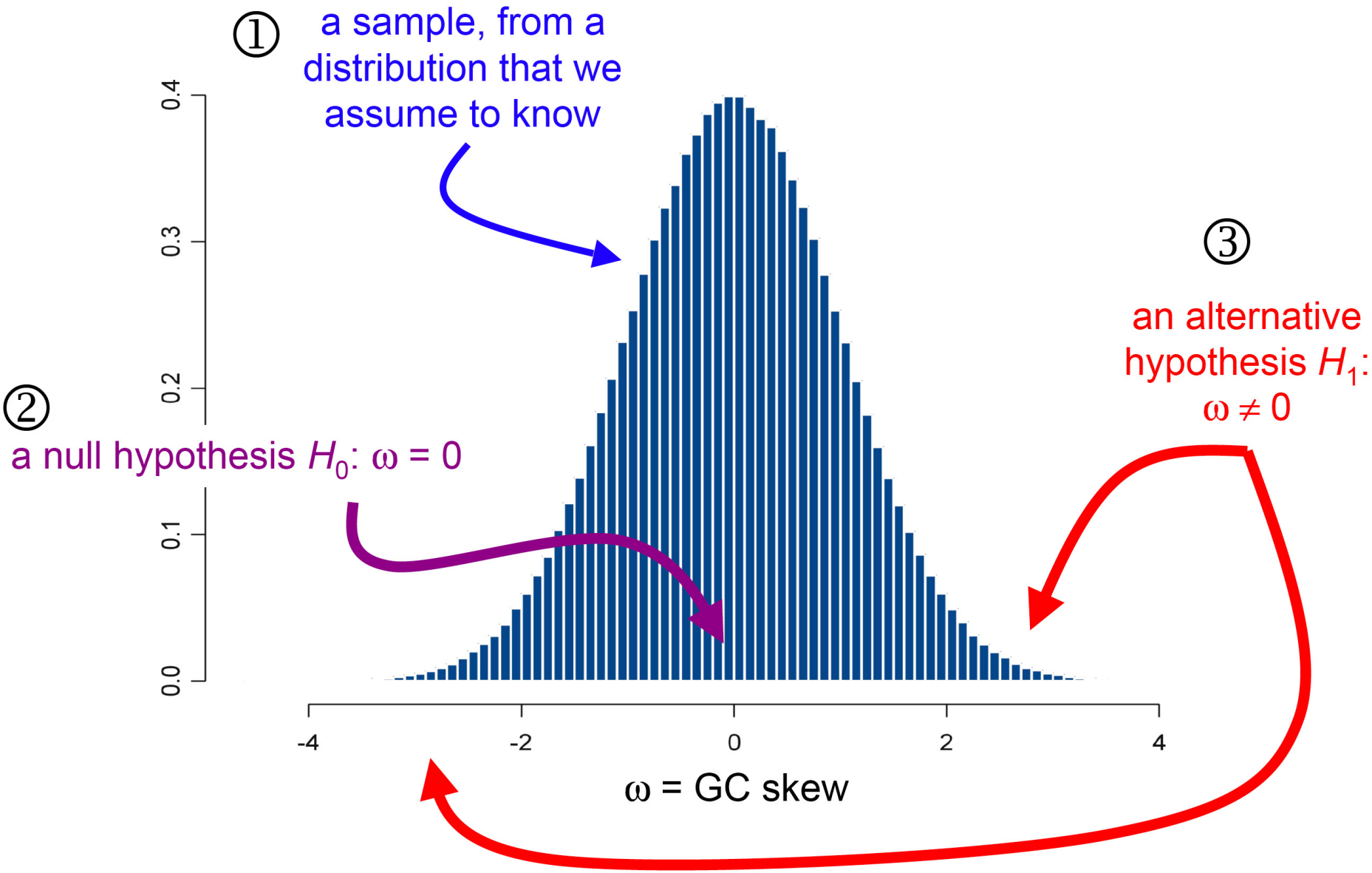
How often do we draw a value > 6?



Do we always need to perform simulations to tests H_0?

- if large sample size AND if asymptotic theory worked out, the answer is NO
- *e.g.*, see lectures on BLAST and Molecular Phylogenies
- when asymptotic theory is used (based on function with parameters), the test is said to be parametric
- otherwise, the test is non-parametric
- non-parametric tests are not limited to permutation (simulation) tests! (rank, signed rank, *etc.*). **Power?**

More formally: What we need to test a null hypothesis?



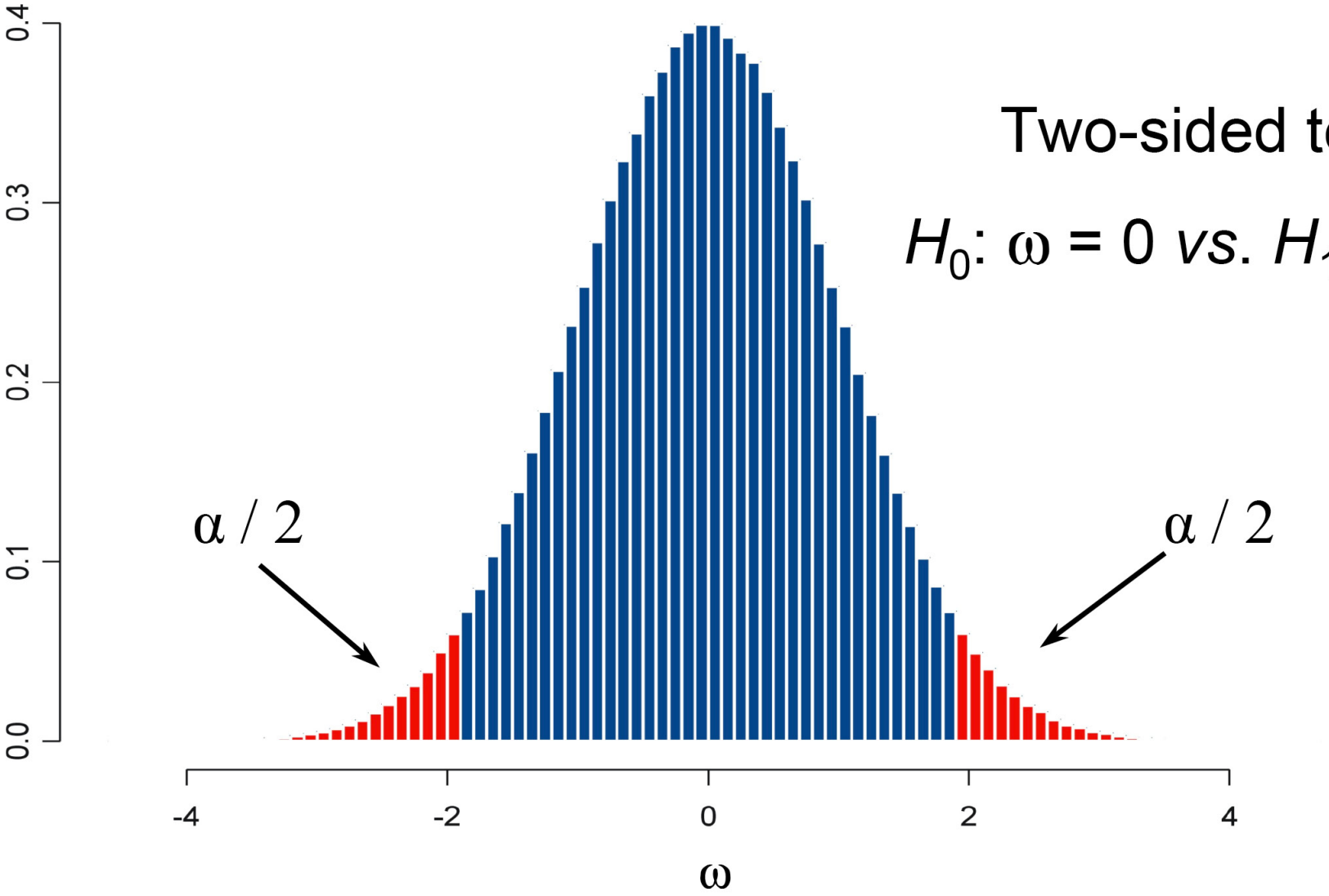
In other words...

- We want to determine the critical region w so that, given H_0, the probability of rejecting H_0 is equal to a pre-assigned value α:

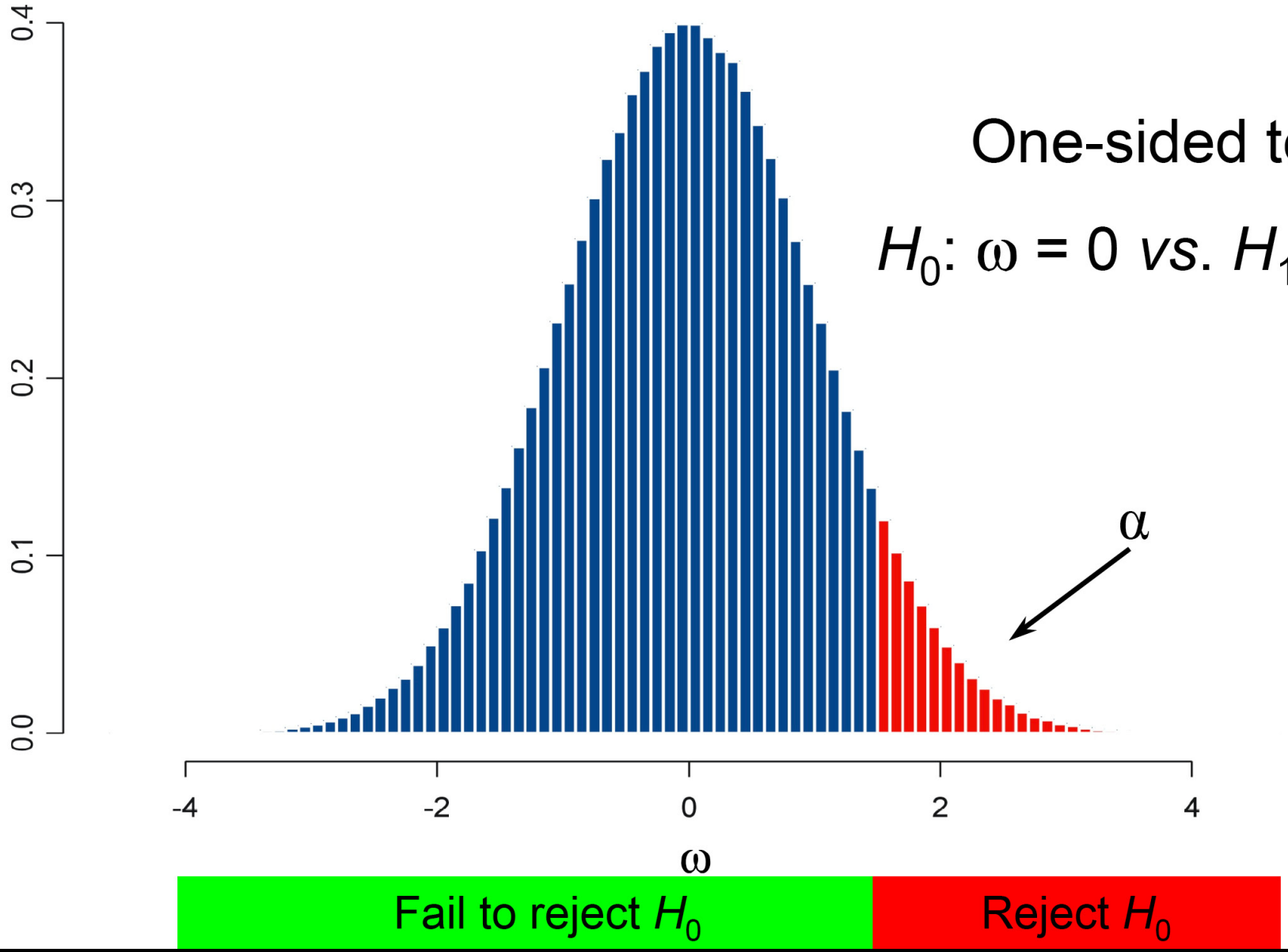
$$P(\omega \in w | H_0) = \alpha$$

- α is called the size of the test.

In practice, we want to determine



The one-sided version of the test



Two types of errors

- Type I: (aka, "error of the first kind", "α error rate", or "false positive") falsely reject H_0
- Type II: (aka, "error of the second kind", "β error rate", or "false negative") falsely fail to reject H_0

	the assigned class is:	
H_0 is:	true	false
true	correct	Type I error
false	Type II error	correct

(See text p.253)

Conclusions

- vocabulary: event, probability, random variables, probability distributions (with examples, equations can be found anywhere. . .);
- tests of statistical hypotheses: null and alternative hypotheses (distributions);
- *p*-value: expression of the level of surprise; if we were able to reproduce how the data were generated under the null, how often would we observe results that are more extreme than those originally observed?
- we will see a related measure in phylogenetics (Bayes factors).

Objectives

- the `freq.HA.xls` and `freq.NA.xls` files contain the nucleotide frequencies (order: ACGT) of the HA and NA genes of influenza viruses
- questions:
  - do these genes have similar GC contents (= freq(G) + freq(C))?
  - propose two solutions, one parametric and one non-parametric

Move to Section 3 of the R script posted on your VC.

1-Read files from R

```
> setwd("~/tmp") # where the files are
>
> HAfreq <- read.table("freq.HA.xls", header =
TRUE)
> NAfreq <- read.table("freq.NA.xls", header =
TRUE)
> HAfreq # to check content of object
> HAfreq[,2] # lists 2nd column
```

2-Compute GC content of each gene

```
> freqGC_HA <- HAfreq[,3] + HAfreq[,4]
> freqGC_NA <- NAfreq[,3] + NAfreq[,4]
```

- we now want to compare these two variables
- it is always a good idea to visualize your data before any statistical analysis

3-Data visualization

```
> # first possibility
> plot(freqGC_HA,pch="+", xlab="GeneID",
      ylab="GC content",ylim=c(.39,.46),
      col="red")
> points(freqGC_NA,pch=3,col="blue")
>
> # second possibility
> plot(density(freqGC_HA), col="red" )
> lines(density(freqGC_NA), col="blue")
```

4-A parametric approach: the two-sided t-test

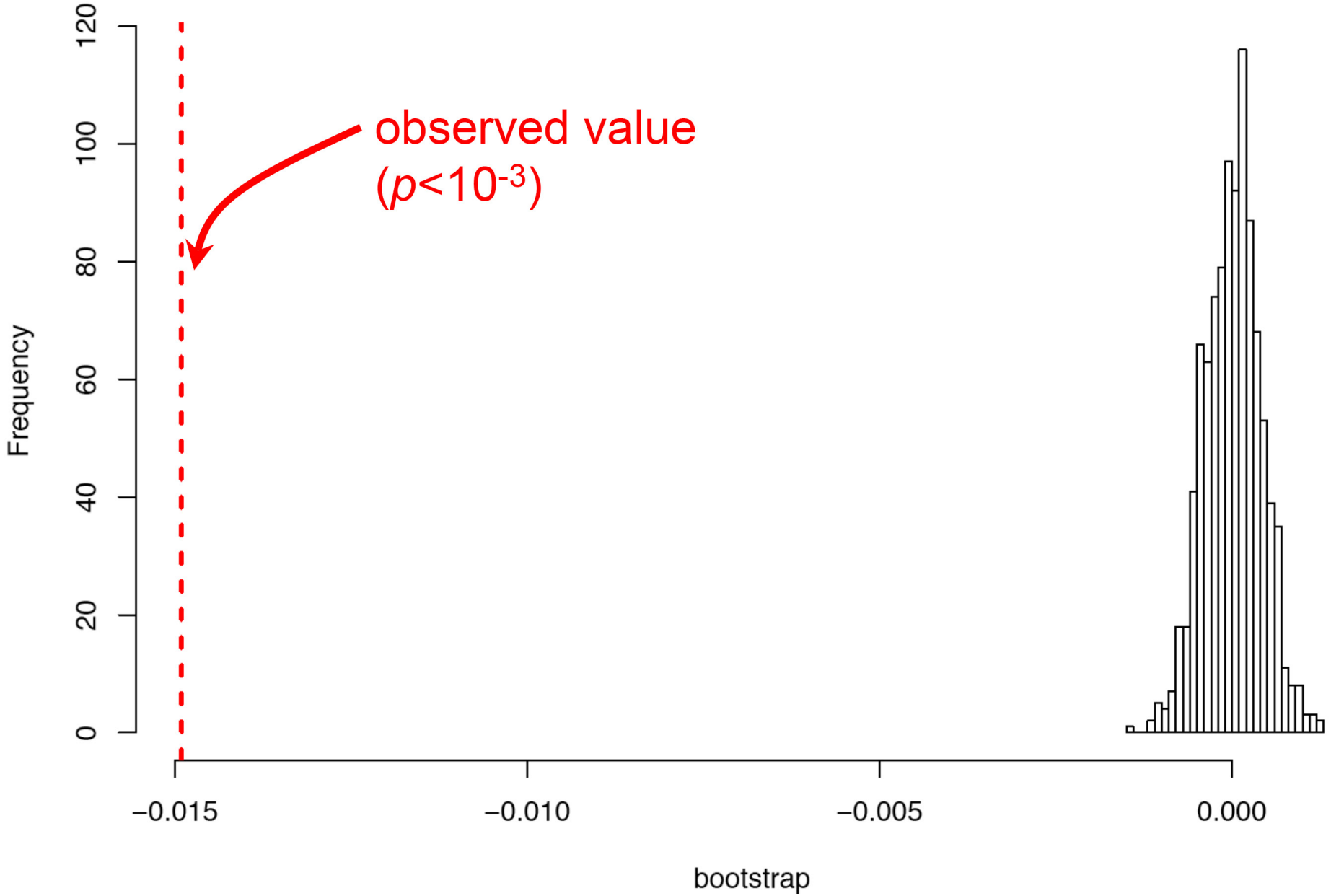
```
> t.test(freqGC_HA, freqGC_NA,
         alternative = c("two.sided"),
         conf.level = 0.95,
         var.equal = FALSE)
>
> # assumption check
> testStat <- freqGC_HA - freqGC_NA
> qqnorm(testStat)
> shapiro.test(testStat)
```

5-A non-parametric approach: the bootstrap

```
> N <- 1000
> bootstrap <- numeric(N)
> for(i in 1:N){
      sampHA <- sample(c(freqGC_HA, freqGC_NA),
                       length(freqGC_HA), replace=T)
      sampNA <- sample(c(freqGC_HA, freqGC_NA),
                       length(freqGC_NA), replace=T)
      bootstrap[i] <- mean(sampHA) - mean(sampNA)
}
```

5-A non-parametric approach: the bootstrap (contd)

```
> hist(bootstrap, nclass=20,
      xlim=c(min(bootstrap,mean(testStat)),
      max(bootstrap,mean(testStat))))
> abline(v=mean(testStat),
      col="red", lwd=2, lty=2)
>
> count <- 0
> for(i in 1:N){
      if( abs(bootstrap[i]) > abs(mean(testStat)))
                        count <- count + 1
}
> Pvalues = count / N
```


observed value
(p<10^{-3})
Frequency
0
20
40
60
80
100
120
−0.015
−0.010
−0.005
0.000
bootstrap

A few words to wrap up

- The way resampling is carried out should directly reflect your null hypothesis: here, this is achieved by sampling from the **empirical joint** distribution of GC contents over the two genes;
- The precision of the *p*-value (# decimal places) depends on the number of replicates (here, 1/1000).

Assignments for next class (Monday)

- text: pp.50-62 & pp.487-491
- exercise 10, p.63-64