

MSMB-Chapter2-Statistical Modelling

Aleeza Gerstein

2019-09-16

Chapter 2: Statistical Modeling

A simple example of statistical modelling

```
load(url("http://bios221.stanford.edu/data/e100.RData"))
e99 = e100[~which.max(e100)]
barplot(table(e99), space = 0.8, col = "chartreuse4")

library("vcd")

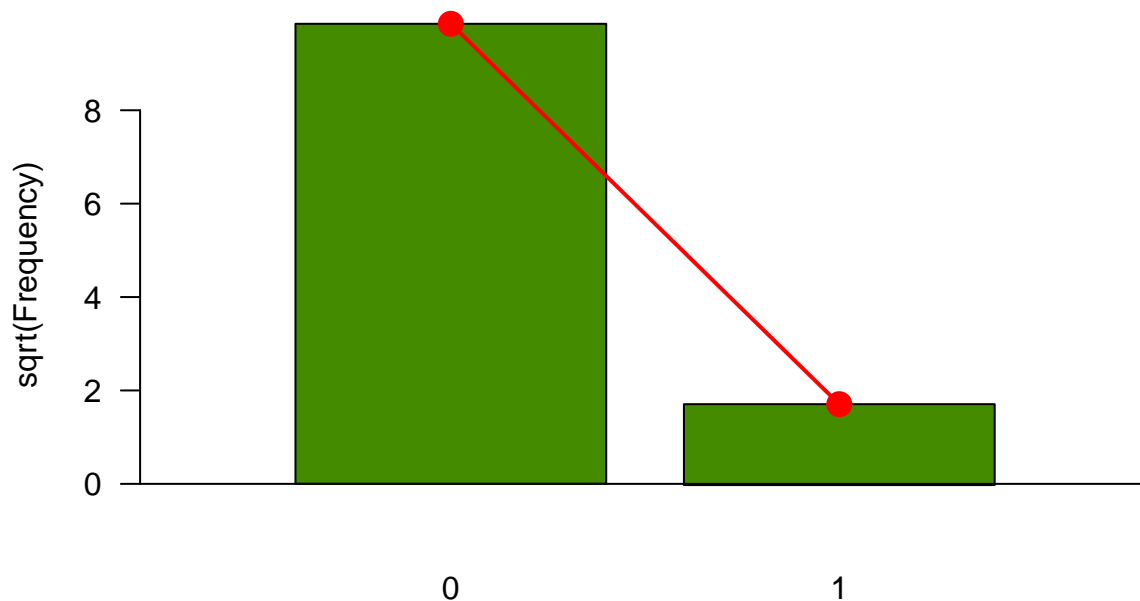
## Loading required package: grid

gf1 = goodfit(e99, "poisson")
rootogram(gf1, xlab = "", rect_gp = gpar(fill = "chartreuse4"))
```

We will learn later what this is?

Question 1: To calibrate what such a plot looks like a known poisson variable, use 'rpois' and $\lambda = 0.05$ to generate 100 Poisson distributed numbers and draw their rootogram

```
pv <- rpois(100, 0.05)
gf_pv = goodfit(pv, "poisson")
rootogram(gf_pv, xlab = "", rect_gp = gpar(fill = "chartreuse4"))
```



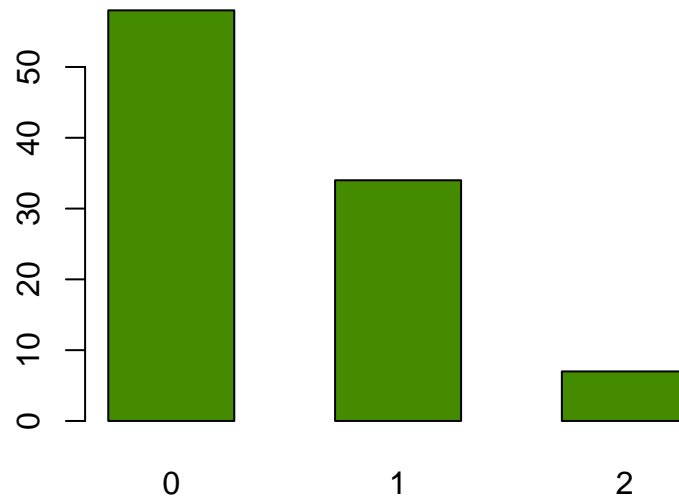


Figure 1: The observed distribution of the epitope data without the outlier.

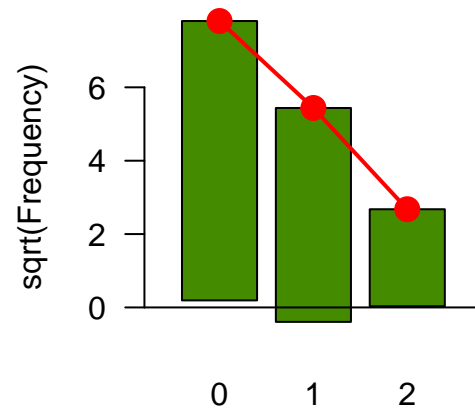


Figure 2: Rootogram showing the square root of the theoretical values as red dots and the square root of the observed frequencies as drop down rectangles. (We'll see a bit below how the `goodfit` function decided which λ to use.)

Question 2: Repeat the simulation with different values of λ . Can you find one that gives count close to the observed count by trial and error?

```
table(e100)

## e100
##  0  1  2  7
## 58 34  7  1

pv0.5 <- rpois(100, 0.5)
table(pv0.5)

## pv0.5
##  0  1  2  3
## 66 22 11  1
```

```
loglikelihood = function(lambda, data = e100) {
  sum(log(dpois(data, lambda)))
}

lambdas = seq(0.05, 0.95, length = 100)
loglik = vapply(lambdas, loglikelihood, numeric(1))
#this is the same as
#loglik = sapply(lambdas, loglikelihood)
plot(lambdas, loglik, type = "l", col = "red", ylab = "", lwd = 2,
     xlab = expression(lambda))
m0 = mean(e100)
abline(v = m0, col = "blue", lwd = 2)
abline(h = loglikelihood(m0), col = "purple", lwd = 2)
```

```
m0

## [1] 0.55

gf = goodfit(e100, "poisson")
names(gf)

## [1] "observed" "count"      "fitted"    "type"      "method"    "df"
## [7] "par"

gf$par

## $lambda
## [1] 0.55

cb = c(rep(0, 110), rep(1, 10))
table(cb)

## cb
##  0  1
## 110 10

probs = seq(0, 0.3, by = 0.005)
likelihood = dbinom(sum(cb), prob = probs, size = length(cb))
plot(probs, likelihood, pch = 16, xlab = "probability of success",
     ylab = "likelihood", cex=0.6)

probs[which.max(likelihood)]
```

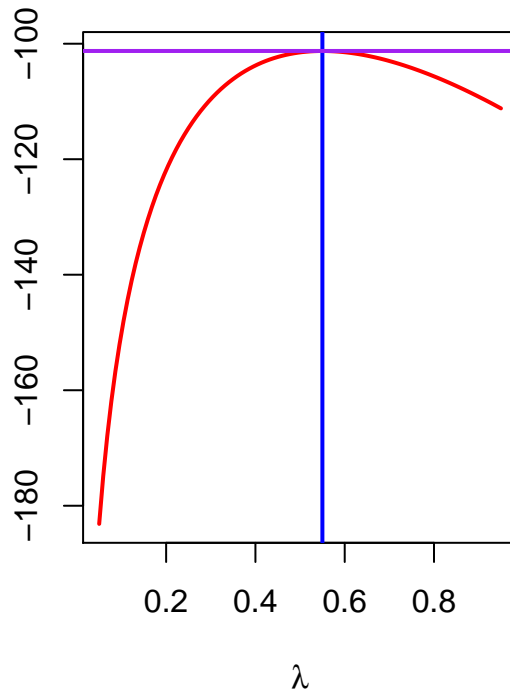


Figure 3: The red curve is the log-likelihood function. The vertical line shows the value of \mathfrak{m} (the mean) and the horizontal line the log-likelihood of \mathfrak{m} . It looks like \mathfrak{m} maximizes the likelihood.

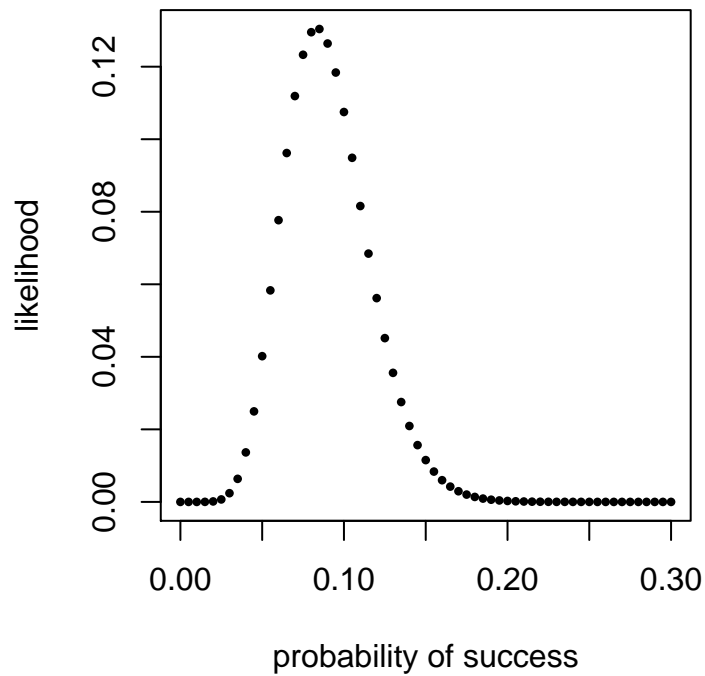


Figure 4: Plot of the likelihood as a function of the probabilities. The likelihood is a function on $[0, 1]$; here we have zoomed into the range of $[(ref : likely1 - 1), (ref : likely1 - 2)]$, as the likelihood is practically zero for larger values of p .

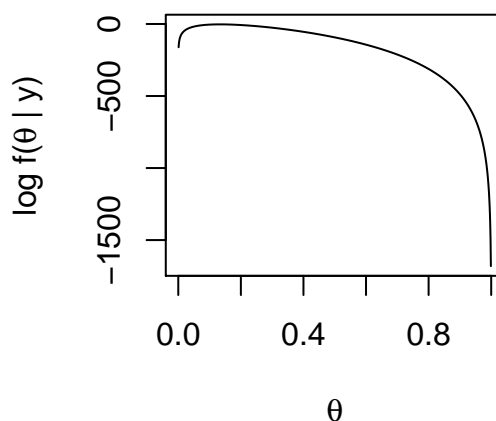


Figure 5: Plot of the log likelihood function for $n = 300$ and $y = 40$.

```
## [1] 0.085
stopifnot(abs(probs[which.max(likelihood)]-1/12) < diff(probs[1:2]))

loglikelihood = function(theta, n = 300, k = 40) {
  115 + k * log(theta) + (n - k) * log(1 - theta)
}

thetas = seq(0, 1, by = 0.001)
plot(thetas, loglikelihood(thetas), xlab = expression(theta),
      ylab = expression(paste("log f(", theta, " | y)")),type = "l")

library("Biostrings")

## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##   clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##   clusterExport, clusterMap, parApply, parCapply, parLapply,
##   parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##   IQR, mad, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##   anyDuplicated, append, as.data.frame, basename, cbind,
##   colnames, dirname, do.call, duplicated, eval, evalq, Filter,
##   Find, get, grep, grepl, intersect, is.unsorted, lapply, Map,
##   mapply, match, mget, order, paste, pmax, pmax.int, pmin,
##   pmin.int, Position, rank, rbind, Reduce, rownames, sapply,
##   setdiff, sort, table, tapply, union, unique, unsplit, which,
##   which.max, which.min
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'
## The following object is masked from 'package:base':
##
##     expand.grid
## Loading required package: IRanges
##
## Attaching package: 'IRanges'
## The following object is masked from 'package:vcd':
##
##     tile
## Loading required package: XVector
##
## Attaching package: 'Biostrings'
## The following object is masked from 'package:base':
##
##     strsplit
staph = readDNAStringSet("http://bios221.stanford.edu/data/staphsequence.ffn.txt", "fasta")
staph[1]

## A DNAStringSet instance of length 1
##      width seq                                     names
## [1] 1362 ATGTCGGAAAAAGAAATTTGG...AAGAAATAAGAAATGTATAA lc1|NC_002952.2_c...
letterFrequency(staph[[1]], letters = "ACGT", OR = 0)

##      A      C      G      T
## 522 219 229 392
```

Question 2.9: Following a similar procedure to Exercise 1.8, test whether the nucleotides are equally distributed across the four possibilities for this first gene.

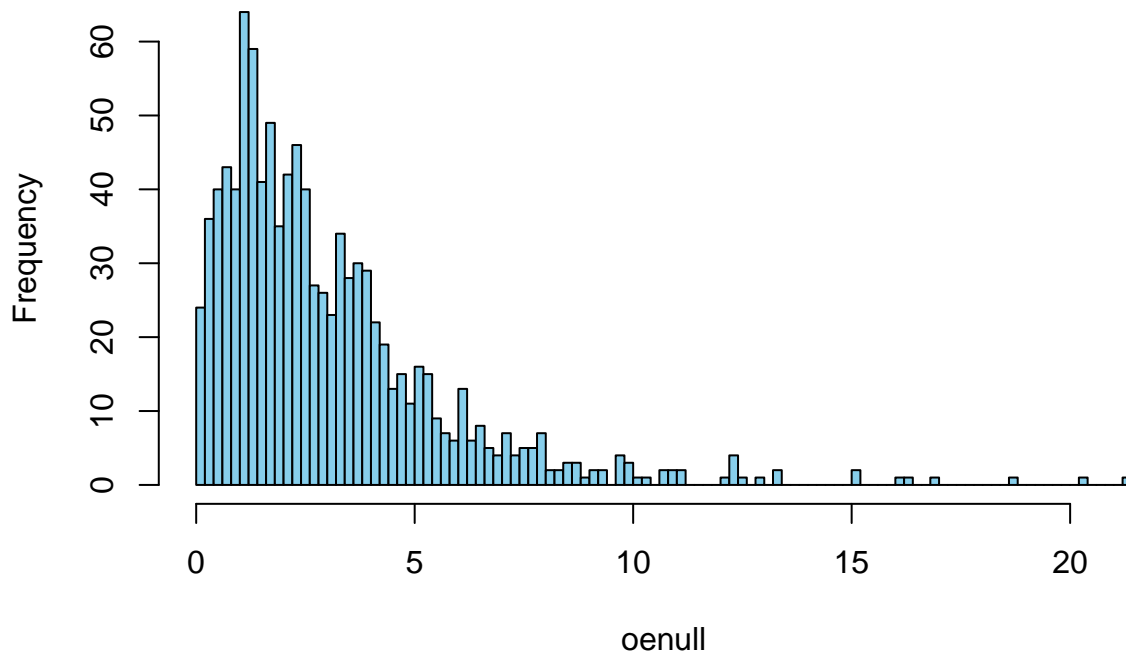
```
lf <- letterFrequency(staph[[1]], letters = "ACGT", OR = 0)
t(rmultinom(1, sum(lf), p = rep(1/4, 4)))

##      [,1] [,2] [,3] [,4]
## [1,] 345 340 340 337

oestat <- function(o, e) {
  sum((e-o)^2 / e)
}

B = 1000
n = sum(lf)
expected = rep(n / 4, 4)

oenull <- replicate(B,
  oestat(e = expected, o = rmultinom(1, n, p = rep(1/4, 4))))
hist(oenull, breaks = 100, col = "skyblue", main="")
```



```
oeostat(e = expected, o = t(lf))
```

```
## [1] 184.4023
```

```
dmultinom(lf, prob = rep(0.25, 4))
```

```
## [1] 9.026662e-45
```

```
letterFrq = vapply(staph, letterFrequency, FUN.VALUE = numeric(4),
                    letters = "ACGT", OR = 0)
colnames(letterFrq) = paste0("gene", seq(along = staph))
tab10 = letterFrq[, 1:10]
computeProportions = function(x) { x/sum(x) }
prop10 = apply(tab10, 2, computeProportions)
round(prop10, digits = 2)
```

```
##   gene1 gene2 gene3 gene4 gene5 gene6 gene7 gene8 gene9 gene10
## A   0.38  0.36  0.35  0.37  0.35  0.33  0.33  0.34  0.38  0.27
## C   0.16  0.16  0.13  0.15  0.15  0.15  0.16  0.16  0.14  0.16
## G   0.17  0.17  0.23  0.19  0.22  0.22  0.20  0.21  0.20  0.20
## T   0.29  0.31  0.30  0.29  0.27  0.30  0.30  0.29  0.28  0.36
```

```
p0 = rowMeans(prop10)
p0
```

```
##           A           C           G           T
## 0.3470531 0.1518313 0.2011442 0.2999714
```

Outer probaability

```
cs = colSums(tab10)
cs
```

```
##   gene1 gene2 gene3 gene4 gene5 gene6 gene7 gene8 gene9 gene10
##   1362  1134   246  1113  1932  2661   831  1515  1287   696
```

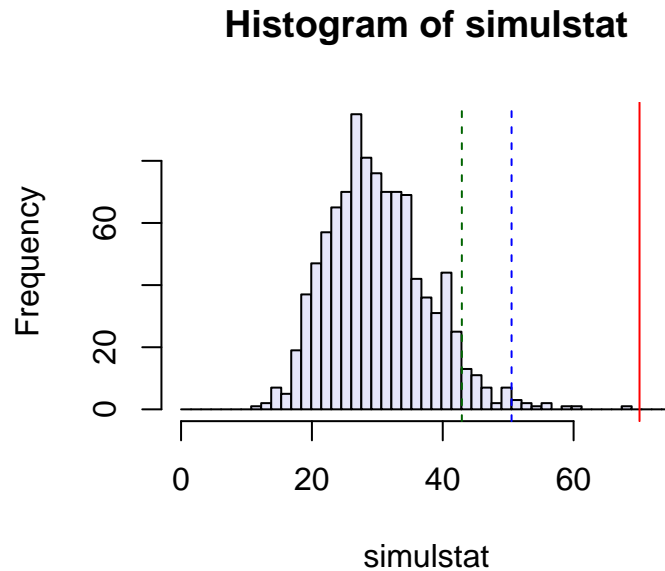


Figure 6: Histogram of `simulstat`. The value of `S1` is marked by the vertical red line, those of the 0.95 and 0.99 quantiles (see next section) by the dotted lines.

```

expectedtab10 = outer(p0, cs, FUN = "*")
round(expectedtab10)

##   gene1 gene2 gene3 gene4 gene5 gene6 gene7 gene8 gene9 gene10
## A   473   394    85   386   671   924   288   526   447   242
## C   207   172    37   169   293   404   126   230   195   106
## G   274   228    49   224   389   535   167   305   259   140
## T   409   340    74   334   580   798   249   454   386   209

randomtab10 = sapply(cs, function(s) { rmultinom(1, s, p0) } )
all(colSums(randomtab10) == cs)

## [1] TRUE

stat = function(obsvd, exptd = 20 * pvec) {
  sum((obsvd - exptd)^2 / exptd)
}
B = 1000

simulstat = replicate(B, {
  randomtab10 = sapply(cs, function(s) { rmultinom(1, s, p0) })
  stat(randomtab10, expectedtab10)
})

S1 = stat(tab10, expectedtab10)
sum(simulstat >= S1)

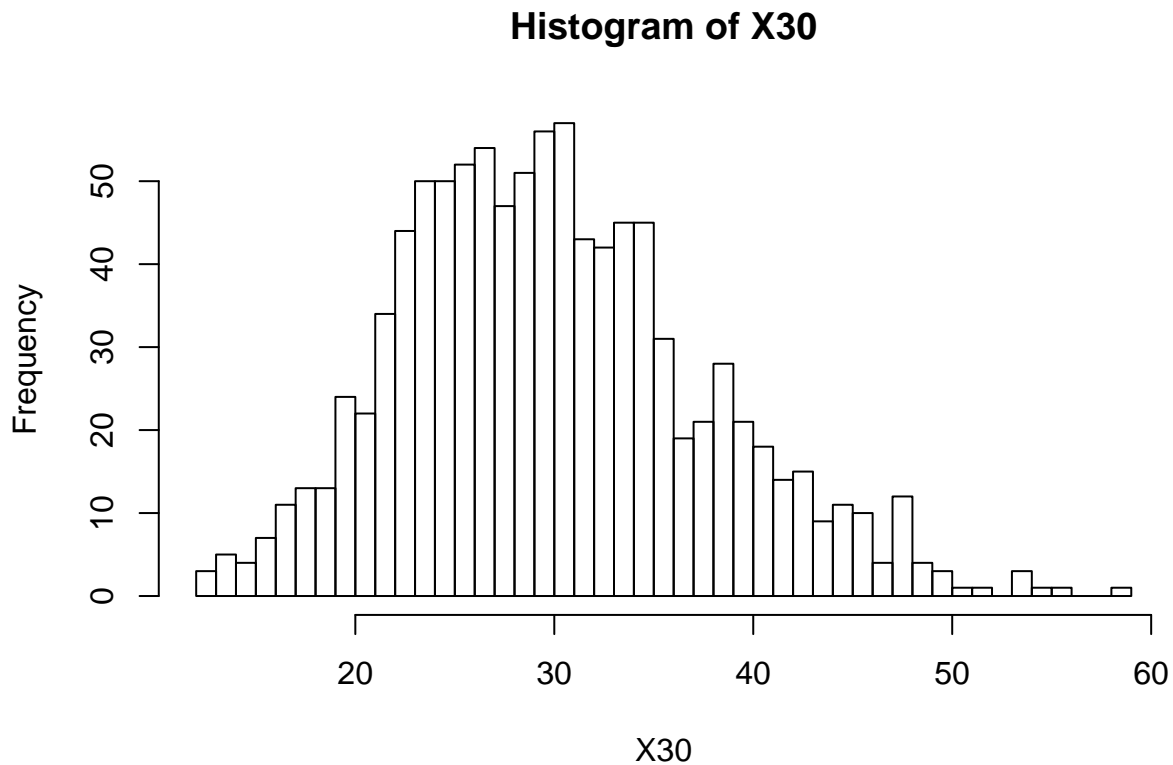
## [1] 0

hist(simulstat, col = "lavender", breaks = seq(0, 75, length.out=50))
abline(v = S1, col = "red")
abline(v = quantile(simulstat, probs = c(0.95, 0.99)),
       col = c("darkgreen", "blue"), lty = 2)

```

Question 2.10 a) Compare the 'simulstat' values and 1000 randomly generated χ^2_{30} random numbers by displaying them in histograms with 50 bins each.

```
X30 <- rchisq(1000, 30)
hist(X30, breaks = 50)
```



b) Compute the quantiles of the simulstat values and compare them to those of the χ^2_{30} distribution.}

```
qs = ppoints(100)
squant <- quantile(simulstat, qs)
quant <- quantile(qchisq(qs, df = 30), qs)
plot(squant, quant)
abline(0, 1)
```

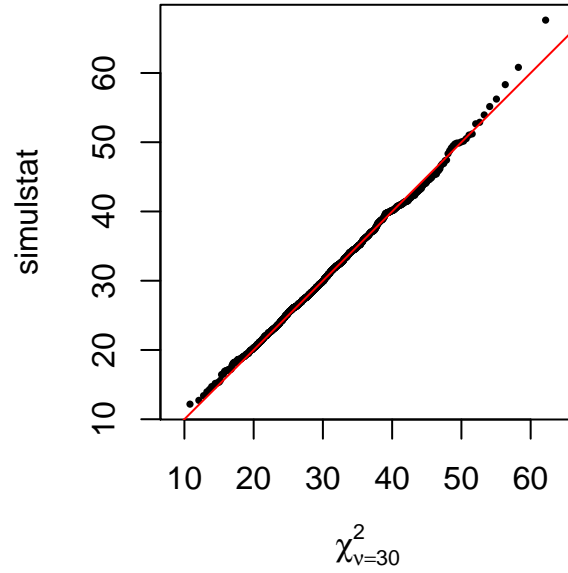
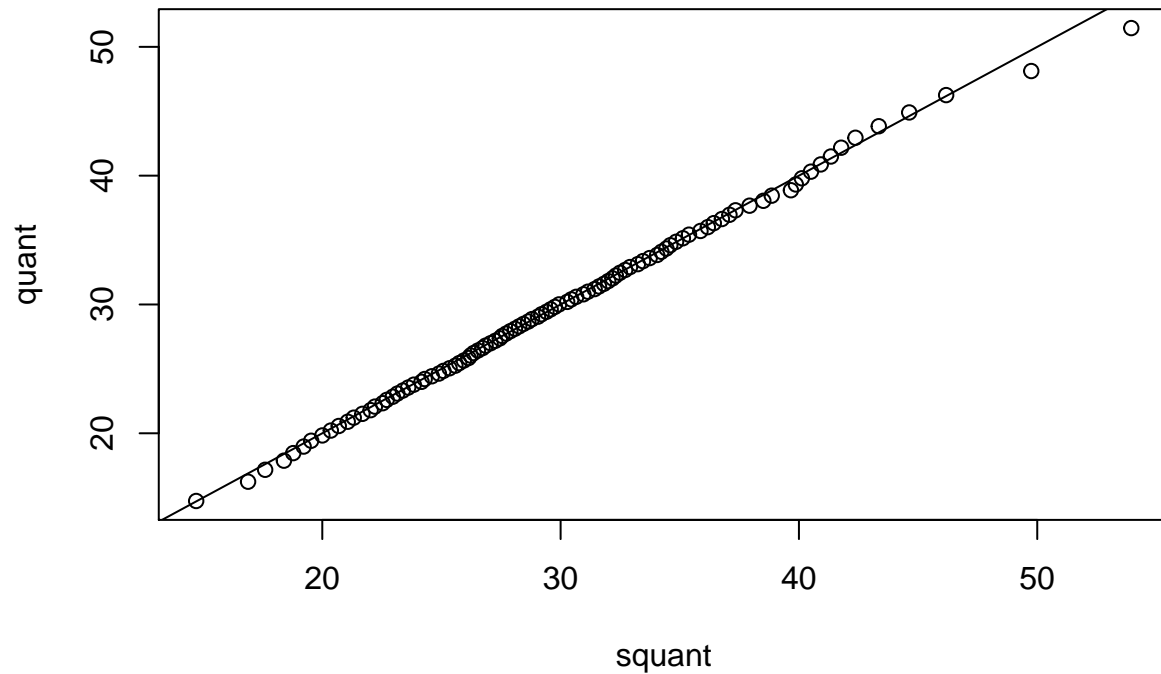


Figure 7: Our simulated statistic's distribution compared to χ^2_{30} using a QQ-plot, which shows the theoretical **quantiles** for the χ^2_{30} distribution on the horizontal axis and the sampled ones on the vertical axis.



```
qqplot(qchisq(ppoints(B), df = 30), simulstat, main = "",
       xlab = expression(chi[nu==30]^2), asp = 1, cex = 0.5, pch = 16)
abline(a = 0, b = 1, col = "red")
```

Compute the p-value that the counts are distributed with multinomial probability $p_A = 0.35$, $p_C = 0.15$, $p_G = 0.2$ $p_T = 0.3$ we observe a value as high as 70.1:

```
1 - pchisq(S1, df = 30)
```

```
## [1] 4.74342e-05
```

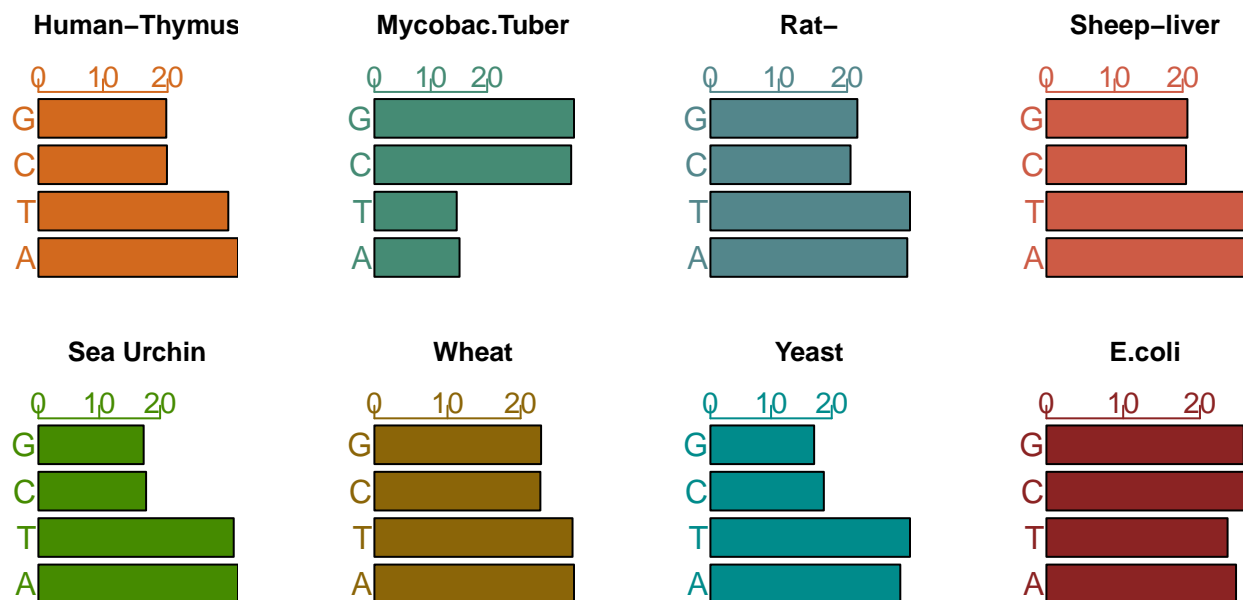


Figure 8: Barplots for the different rows in ChargaffTable. Can you spot the pattern?

2.7 Chagaff's Rule

Chagaff published percentages of the masses of different organisms for nucleotides:

```
load(url("http://bios221.stanford.edu/data/ChargaffTable.RData"))
ChargaffTable
```

```
##           A      T      C      G
## Human-Thymus 30.9 29.4 19.9 19.8
## Mycobac.Tuber 15.1 14.6 34.9 35.4
## Rat-         28.8 29.2 20.5 21.5
## Sheep-liver  29.3 29.3 20.5 20.7
## Sea Urchin   32.8 32.1 17.7 17.3
## Wheat        27.3 27.1 22.7 22.8
## Yeast        31.3 32.9 18.7 17.1
## E.coli       24.7 23.6 26.0 25.7
```

Question 2.13: Do these data seem to come from equally likely multinomial categories? Can you suggest an alternative pattern? Can you do a quantitative analysis of the pattern?

No. $G = C = 0.2$; $A = T = 0.3$

Explain this why 'statChf'

```
statChf = function(x){
  sum((x[, "C"] - x[, "G"])^2 + (x[, "A"] - x[, "T"])^2)
}
chfstat = statChf(ChargaffTable)
permstat = replicate(100000, {
  permuted = t(apply(ChargaffTable, 1, sample))
  colnames(permuted) = colnames(ChargaffTable)
  statChf(permuted)
})
```

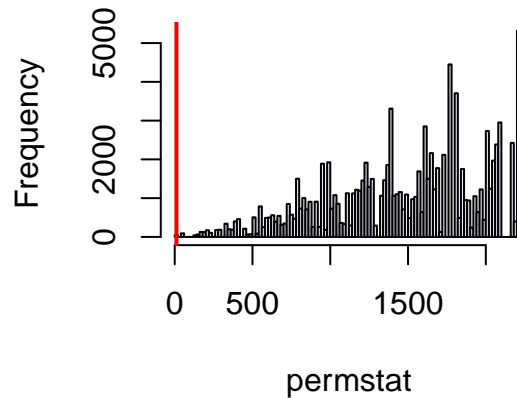


Figure 9: Histogram of our statistic `statChf` computed from simulations using per-row permutations of the columns. The value it yields for the observed data is shown by the red line.

```
pChf = mean(permstat <= chfstat)
pChf

## [1] 0.00018
hist(permstat, breaks = 100, main = "", col = "lavender")
abline(v = chfstat, lwd = 2, col = "red")
```

Question 2.14: When comparing ‘pChf’ we only looked at the values in the null distribution smaller than the observed value. Why did we do this in a one-sided way here?

2.7.1 Two categorical variables

```
load(url("http://bios221.stanford.edu/data/Deuteranopia.RData"))
Deuteranopia

##           Men Women
## Deute      19     2
## NonDeute 1981  1998
chisq.test(Deuteranopia)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Deuteranopia
## X-squared = 12.255, df = 1, p-value = 0.0004641
```

2.7.2 A special multinomial: Hardy-Weinberg equilibrium

```
library("HardyWeinberg")
```

```
## Loading required package: mice
## Loading required package: lattice
##
## Attaching package: 'mice'
## The following objects are masked from 'package:IRanges':
##
##      cbind, rbind
## The following objects are masked from 'package:S4Vectors':
##
##      cbind, rbind
## The following objects are masked from 'package:BiocGenerics':
##
##      cbind, rbind
## The following objects are masked from 'package:base':
##
##      cbind, rbind
```

```
## Loading required package: Rsolnp
```

```
data("Mourant")
Mourant[214:216,]
```

```
##      Population      Country Total  MM  MN  NN
## 214      Oceania Micronesia   962 228 436 298
## 215      Oceania Micronesia   678  36 229 413
## 216      Oceania      Tahiti   580 188 296  96
```

```
nMM = Mourant$MM[216]
nMN = Mourant$MN[216]
nNN = Mourant$NN[216]
loglik = function(p, q = 1 - p) {
  2 * nMM * log(p) + nMN * log(2*p*q) + 2 * nNN * log(q)
}
xv = seq(0.01, 0.99, by = 0.01)
yv = loglik(xv)
plot(x = xv, y = yv, type = "l", lwd = 2,
      xlab = "p", ylab = "log-likelihood")
imax = which.max(yv)
abline(v = xv[imax], h = yv[imax], lwd = 1.5, col = "blue")
abline(h = yv[imax], lwd = 1.5, col = "purple")
```

```
phat = af(c(nMM, nMN, nNN))
phat
```

```
## [1] 0.5793103
```

```
pMM = phat^2
qhat = 1 - phat
```

```
pHW = c(MM = phat^2, MN = 2*phat*qhat, NN = qhat^2)
sum(c(nMM, nMN, nNN)) * pHW
```

```
##      MM      MN      NN
## 194.6483 282.7034 102.6483
```

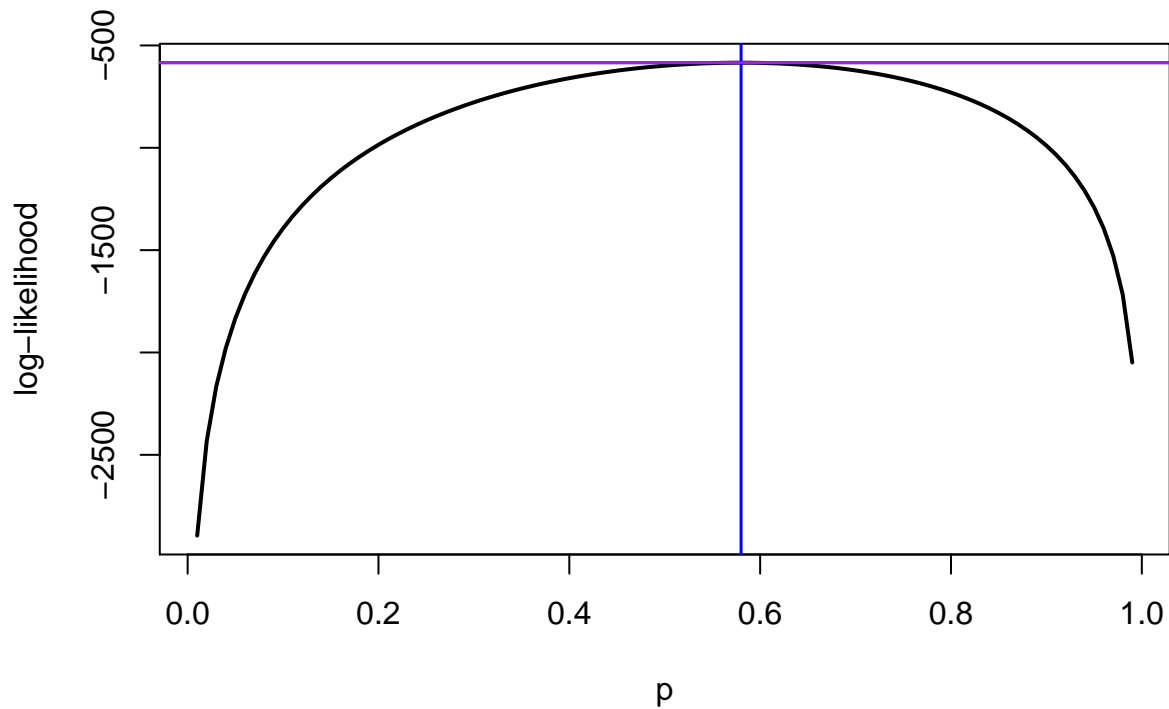


Figure 10: Plot of the log-likelihood for the (ref:chap2-r-HardyWeinberg-1-1) data.

```
pops = c(1, 69, 128, 148, 192)
genotypeFrequencies = as.matrix(Mourant[, c("MM", "MN", "NN")])
HWTernaryPlot(genotypeFrequencies[pops, ],
               markerlab = Mourant$Country[pops],
               alpha = 0.0001, curvecols = c("red", rep("purple", 4)),
               mcex = 0.75, vertex.cex = 1)
```

Question 2.16: Make the ternary plot as in the code above then add the other data points to it. What do you notice? You can back up your discussion with the 'HWChisq' function

```
HWTernaryPlot(genotypeFrequencies[-pops, ],
               markerlab = Mourant$Country[-pops],
               alpha = 0.0001, curvecols = c("red", rep("purple", 4)),
               mcex = 0.75, vertex.cex = 1)
```

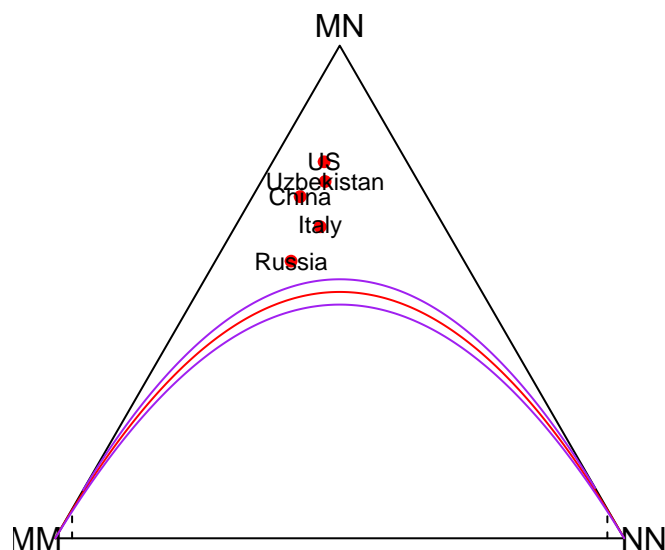
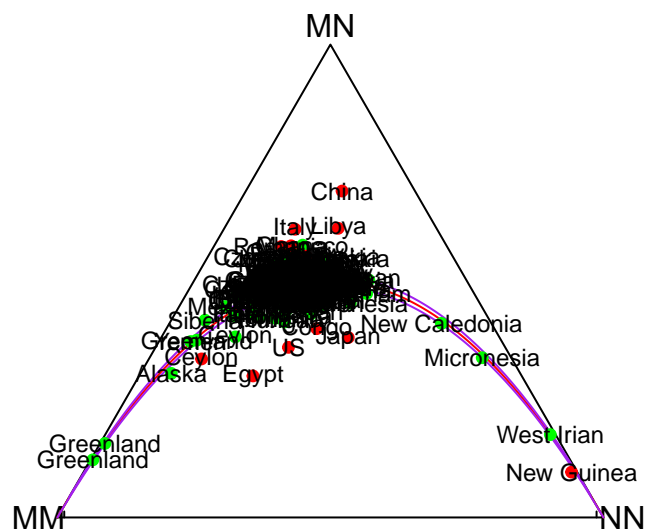


Figure 11: This **de Finetti plot** shows the points as barycenters of the three genotypes using the frequencies as weights on each of the corners of the triangle. The Hardy-Weinberg model is the red curve, the acceptance region is between the two purple lines. We see that the US is the furthest from being in HW equilibrium.



```
unique(Mourant$Population)
```

```
## [1] USSR    Eskimos Jews    Europe Turkey Asia    Africa America Oceania
## Levels: Africa America Asia Eskimos Europe Jews Oceania Turkey USSR
```

```
t <- apply(genotypeFrequencies, 1, function(x) HWChisq(x)$pval)
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 144.1748 DF = 1 p-value = 3.25363e-33 D = 220.9169 f = -0.1589235
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.235995 DF = 1 p-value = 0.07203652 D = -7.715232 f = 0.07737265
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3464892 DF = 1 p-value = 0.5561073 D = -5.506968 f = 0.01392382

## Warning in HWChisq(x): Expected counts below 5: chi-square approximation
## may be incorrect
```

```

## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01751215 DF = 1 p-value = 0.8947205 D = -0.693761 f = 0.01535081

## Warning in HWChisq(x): Expected counts below 5: chi-square approximation
## may be incorrect

## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.04787744 DF = 1 p-value = 0.8267987 D = 0.1177267 f = -0.00262301
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.008482587 DF = 1 p-value = 0.9266178 D = 0.768 f = -0.008326828
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.030071 DF = 1 p-value = 0.1542134 D = 16.0805 f = -0.03262297
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 34.79255 DF = 1 p-value = 3.667752e-09 D = -117.9812 f = 0.07307685
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 6.900338 DF = 1 p-value = 0.008617946 D = 35.78133 f = -0.04846983
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 13.36419 DF = 1 p-value = 0.0002564751 D = -90.71677 f = 0.03670936
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0004297542 DF = 1 p-value = 0.9834606 D = 0.5661652 f = -0.001165108
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5446964 DF = 1 p-value = 0.4604929 D = 10.58073 f = -0.01374264
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6087732 DF = 1 p-value = 0.4352501 D = 6.47025 f = -0.02619219
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.355125 DF = 1 p-value = 0.06699611 D = 12.00037 f = -0.07295075
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1464166 DF = 1 p-value = 0.7019836 D = 6.423477 f = -0.006394214
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 5.153911 DF = 1 p-value = 0.02319407 D = 12.968 f = -0.1045537
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1590772 DF = 1 p-value = 0.6900075 D = 3.65521 f = -0.01315558
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 75.97485 DF = 1 p-value = 2.87301e-18 D = 153.7947 f = -0.1222334
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01592902 DF = 1 p-value = 0.8995655 D = -1.781018 f = 0.003542334
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0002437838 DF = 1 p-value = 0.9875427 D = -0.6262961 f = 0.0004839528
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.003667017 DF = 1 p-value = 0.9517129 D = -1.029919 f = 0.00211036
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.005589011 DF = 1 p-value = 0.9404059 D = 1.588852 f = -0.001482261
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3807199 DF = 1 p-value = 0.5372182 D = 4.707071 f = -0.02249035
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.04590264 DF = 1 p-value = 0.8303529 D = 1.893122 f = -0.008719768
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.871296 DF = 1 p-value = 0.1713267 D = -8 f = 0.05925926
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1104575 DF = 1 p-value = 0.7396239 D = -2.946429 f = 0.01176471
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.571377 DF = 1 p-value = 0.2100071 D = 11.33655 f = -0.03535343
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.941419 DF = 1 p-value = 0.04711102 D = -51.06031 f = 0.01934059
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```



```

## Chi2 = 1.982894 DF = 1 p-value = 0.1590859 D = 11.401 f = -0.04604623
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5305828 DF = 1 p-value = 0.4663619 D = -6.252544 f = 0.02357204
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01190939 DF = 1 p-value = 0.9130992 D = -1.174676 f = 0.005228342
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3854094 DF = 1 p-value = 0.5347223 D = -5.138425 f = 0.02162287
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 6.180854 DF = 1 p-value = 0.01291399 D = -20.439 f = 0.07700075
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6971411 DF = 1 p-value = 0.403746 D = -26.49561 f = 0.006695891
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01093373 DF = 1 p-value = 0.9167214 D = 0.96 f = -0.007326007
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0007832292 DF = 1 p-value = 0.9776731 D = -0.6210543 f = 0.001591953
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0004957161 DF = 1 p-value = 0.9822368 D = -0.0572123 f = 6.037873e-05
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 4.619097 DF = 1 p-value = 0.03161787 D = 37.45367 f = -0.03111942
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.8322854 DF = 1 p-value = 0.3616125 D = 23.0096 f = -0.009274361
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 5.551651 DF = 1 p-value = 0.01846327 D = 19.49032 f = -0.07382427
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.78886 DF = 1 p-value = 0.05159477 D = 15.69025 f = -0.06306123
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.028393 DF = 1 p-value = 0.1543837 D = 10.35281 f = -0.05224809
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.4135673 DF = 1 p-value = 0.5201644 D = -6.708544 f = 0.01710453
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.2455073 DF = 1 p-value = 0.6202564 D = -5.8795 f = 0.01183285
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.707777 DF = 1 p-value = 0.05415947 D = 39.04728 f = -0.02415615
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5879373 DF = 1 p-value = 0.4432179 D = 6.13385 f = -0.02715701
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.000826124 DF = 1 p-value = 0.9770701 D = 0.1765782 f = -0.000652005
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1451798 DF = 1 p-value = 0.7031849 D = -3.932394 f = 0.01117372
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.07615999 DF = 1 p-value = 0.7825703 D = -7.799917 f = 0.002646911
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.174822 DF = 1 p-value = 0.6758614 D = -3.05279 f = 0.01849584
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.92292 DF = 1 p-value = 0.1655351 D = -10.52933 f = 0.04807269
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.139793 DF = 1 p-value = 0.1435213 D = -21.06248 f = 0.02606541
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6048145 DF = 1 p-value = 0.4367469 D = 12.35813 f = -0.0128244
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.471217 DF = 1 p-value = 0.2251537 D = 16.13448 f = -0.02354142
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 8.692228 DF = 1 p-value = 0.003195697 D = 17.30815 f = -0.1289806
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 0.70793 DF = 1 p-value = 0.4001319 D = 11.41815 f = -0.01640346
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9691839 DF = 1 p-value = 0.3248838 D = 11.28345 f = -0.02267043
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.328076 DF = 1 p-value = 0.2491478 D = 34.83068 f = -0.009553988
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.03908429 DF = 1 p-value = 0.8432818 D = -6.318333 f = 0.001716324
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1897645 DF = 1 p-value = 0.6631127 D = -3.034856 f = 0.01989615
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 11.89061 DF = 1 p-value = 0.0005641774 D = -20.09702 f = 0.1493125
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.088042 DF = 1 p-value = 0.07886967 D = -10.24705 f = 0.08084648
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.801502 DF = 1 p-value = 0.179531 D = 9.369813 f = -0.05116476
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.711511 DF = 1 p-value = 0.05403845 D = 21.50579 f = -0.0435345
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 7.355671 DF = 1 p-value = 0.006685156 D = -19.34719 f = 0.09778854
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.002667343 DF = 1 p-value = 0.9588105 D = -0.9848485 f = 0.001680672
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9070016 DF = 1 p-value = 0.3409112 D = 7.372165 f = -0.03224526
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.438673 DF = 1 p-value = 0.1183763 D = 9.881667 f = -0.06626728
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 84.42257 DF = 1 p-value = 3.995607e-20 D = 78.37457 f = -0.2706067
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.05827417 DF = 1 p-value = 0.8092448 D = 2.271593 f = -0.008952853
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 48.05815 DF = 1 p-value = 4.137647e-12 D = 50.05882 f = -0.2395833
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.8464939 DF = 1 p-value = 0.3575461 D = 7.53225 f = -0.03062292
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.904765 DF = 1 p-value = 0.08831811 D = 13.704 f = -0.05541537
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 8.079757 DF = 1 p-value = 0.004476244 D = -21.359 f = 0.0915285
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.002767156 DF = 1 p-value = 0.9580476 D = 0.04435484 f = -0.000241117
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3425901 DF = 1 p-value = 0.5583379 D = -3.392727 f = 0.0281805
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.080873 DF = 1 p-value = 0.079218 D = 10.26079 f = -0.07730038
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 8.437928 DF = 1 p-value = 0.003674746 D = -64.553 f = 0.03211507
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 12.05476 DF = 1 p-value = 0.0005166024 D = 27.20025 f = -0.1113397
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1921472 DF = 1 p-value = 0.6611355 D = -3.04 f = 0.02039721
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.134716 DF = 1 p-value = 0.2867717 D = -28.86422 f = 0.01000193
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5114028 DF = 1 p-value = 0.4745322 D = -12.77292 f = 0.01052765
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 0.1299846 DF = 1 p-value = 0.718448 D = -2.573333 f = 0.01720449
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01194549 DF = 1 p-value = 0.9129682 D = 1.839032 f = -0.002464659
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9405559 DF = 1 p-value = 0.3321349 D = 9.084203 f = -0.02761863
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.4920306 DF = 1 p-value = 0.4830229 D = 30.39793 f = -0.003998096
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.02429297 DF = 1 p-value = 0.8761418 D = -1.62109 f = 0.006057408
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1323965 DF = 1 p-value = 0.7159601 D = 7.571257 f = -0.004695561
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.72444 DF = 1 p-value = 0.0536216 D = 25.74437 f = -0.03645192
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.072349 DF = 1 p-value = 0.3004155 D = 11.37037 f = -0.02517632
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.000815 DF = 1 p-value = 0.3171134 D = 6.964183 f = -0.04001581
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0002798192 DF = 1 p-value = 0.9866538 D = -0.4261472 f = 0.00163321
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01918895 DF = 1 p-value = 0.8898261 D = -1.228296 f = 0.007912835
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 5.257029 DF = 1 p-value = 0.02185831 D = -18.24903 f = 0.07277808
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9132209 DF = 1 p-value = 0.3392613 D = 7.061869 f = -0.03585831
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 10.14062 DF = 1 p-value = 0.001450376 D = 21.04179 f = -0.1230815
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 5.436096 DF = 1 p-value = 0.0197247 D = 14.68253 f = -0.097353
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.880493 DF = 1 p-value = 0.08965854 D = 15.91541 f = -0.04659288
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 9.159442 DF = 1 p-value = 0.00247438 D = -18.75975 f = 0.1220069
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.684379 DF = 1 p-value = 0.1943436 D = -7.7 f = 0.05982906
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.378741 DF = 1 p-value = 0.1229966 D = -8.780271 f = 0.07361042
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 7.584971 DF = 1 p-value = 0.005885692 D = -15.74953 f = 0.1228038
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 6.681023 DF = 1 p-value = 0.009744467 D = 17.8819 f = -0.09685885
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 22.23425 DF = 1 p-value = 2.413303e-06 D = 29.9468 f = -0.1842277
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 8.860607 DF = 1 p-value = 0.002913906 D = 28.94278 f = -0.07779121
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 4.911152 DF = 1 p-value = 0.02668384 D = -13.18271 f = 0.09861193
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6260884 DF = 1 p-value = 0.4287938 D = -6.979167 f = 0.02415111
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5364934 DF = 1 p-value = 0.4638896 D = -11.65944 f = 0.01196872
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1351309 DF = 1 p-value = 0.7131703 D = 4.077072 f = -0.009885658
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 0.9296296 DF = 1 p-value = 0.3349592 D = 10.30127 f = -0.02408535
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.00109149 DF = 1 p-value = 0.9736445 D = 0.6410604 f = -0.002016808
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.02948787 DF = 1 p-value = 0.8636574 D = 1.591934 f = -0.007712555
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9855184 DF = 1 p-value = 0.3208402 D = 9.449013 f = -0.02795144
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.2436828 DF = 1 p-value = 0.6215588 D = 3.197664 f = -0.02328922
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9748977 DF = 1 p-value = 0.3234617 D = -13.93652 f = 0.01812441
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.285431 DF = 1 p-value = 0.1305938 D = 14.42624 f = -0.04109178
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.527046 DF = 1 p-value = 0.4678514 D = 14.56285 f = -0.009475243
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.107158 DF = 1 p-value = 0.2927005 D = -17.3787 f = 0.01654517
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.316154 DF = 1 p-value = 0.5739282 D = -4.545556 f = 0.02042528
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6995281 DF = 1 p-value = 0.4029423 D = 5.49805 f = -0.03547084
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1034337 DF = 1 p-value = 0.7477473 D = -2.062 f = 0.0176146
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.027621 DF = 1 p-value = 0.1544622 D = -9.690086 f = 0.05595058
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.05657749 DF = 1 p-value = 0.8119895 D = 6.210429 f = -0.002483884
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.003919973 DF = 1 p-value = 0.9500773 D = 0.7036329 f = -0.005527517
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6517352 DF = 1 p-value = 0.419493 D = 6.558255 f = -0.02767877
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.321528 DF = 1 p-value = 0.06837824 D = -11.94385 f = 0.07330043
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5794936 DF = 1 p-value = 0.4465109 D = -5.153846 f = 0.03101852
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 257.1487 DF = 1 p-value = 7.178839e-58 D = 142.585 f = -0.4520551
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.411656 DF = 1 p-value = 0.2347816 D = 5.647206 f = -0.05712742
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.004548373 DF = 1 p-value = 0.9462301 D = -0.9524092 f = 0.003067811
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3865915 DF = 1 p-value = 0.5340965 D = 7.775519 f = -0.01344723
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.34518 DF = 1 p-value = 0.06740212 D = -15.0755 f = 0.05807753
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.580895 DF = 1 p-value = 0.2086318 D = -9.68551 f = 0.0426326
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5194688 DF = 1 p-value = 0.4710684 D = 4.394135 f = -0.03512333
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.035923 DF = 1 p-value = 0.08144021 D = 12.79703 f = -0.05905331
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 19.62991 DF = 1 p-value = 9.398615e-06 D = -28.53933 f = 0.1375129
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 10.8335 DF = 1 p-value = 0.0009967997 D = -18.96212 f = 0.1308074
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 2.669591 DF = 1 p-value = 0.1022823 D = -9.506616 f = 0.07397764
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.139921 DF = 1 p-value = 0.07639749 D = 11.23667 f = -0.06619021
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.430547 DF = 1 p-value = 0.0640009 D = 15.5984 f = -0.05663873
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.01769947 DF = 1 p-value = 0.8941623 D = -1.413022 f = 0.005620323
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5206021 DF = 1 p-value = 0.4705849 D = 4.988433 f = -0.03013948
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.000713165 DF = 1 p-value = 0.9786949 D = -0.7998602 f = 0.0007146076
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 146.1379 DF = 1 p-value = 1.211144e-33 D = 95.96041 f = -0.3830149
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.354865 DF = 1 p-value = 0.2444294 D = 8.493766 f = -0.04300505
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1601897 DF = 1 p-value = 0.6889819 D = 3.000362 f = -0.01739344
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.288761 DF = 1 p-value = 0.06975563 D = 24.70451 f = -0.03371269
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 137.959 DF = 1 p-value = 7.439819e-32 D = 82.43134 f = -0.414085
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.1390548 DF = 1 p-value = 0.7092225 D = 2.996328 f = -0.01490683
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.2278697 DF = 1 p-value = 0.6331079 D = 3.000496 f = -0.02429561
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.04569838 DF = 1 p-value = 0.830725 D = 2.423752 f = -0.006558191
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.02150404 DF = 1 p-value = 0.8834142 D = -1.72093 f = 0.005036069
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.002658075 DF = 1 p-value = 0.9588821 D = 0.1014235 f = -0.0007301982
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 46.02595 DF = 1 p-value = 1.166967e-11 D = -49.35077 f = 0.2357325
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.7054962 DF = 1 p-value = 0.400943 D = 6.95025 f = -0.02807617
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.183676 DF = 1 p-value = 0.2766089 D = 12.408 f = -0.02508734
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 4.44235 DF = 1 p-value = 0.03505795 D = 12.058 f = -0.09728744
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.009584193 DF = 1 p-value = 0.9220126 D = 1.434472 f = -0.002978768
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 6.225188 DF = 1 p-value = 0.01259455 D = 24.83833 f = -0.06269747
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.2914313 DF = 1 p-value = 0.5893048 D = -5.659687 f = 0.01461848
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0012818 DF = 1 p-value = 0.97144 D = 0.1296004 f = -0.0005459839
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.7388536 DF = 1 p-value = 0.3900284 D = 6.3196 f = -0.03229552
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.6764863 DF = 1 p-value = 0.4107993 D = 9.200211 f = -0.01970912
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 3.351774 DF = 1 p-value = 0.06713262 D = -26.24257 f = 0.0326708
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 4.35653 DF = 1 p-value = 0.03686733 D = 14.57335 f = -0.07859362
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.425034 DF = 1 p-value = 0.2325766 D = -13.57976 f = 0.02726422
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 11.35796 DF = 1 p-value = 0.0007512515 D = -21.68111 f = 0.1349326
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 12.87362 DF = 1 p-value = 0.0003332464 D = -33.87901 f = 0.09479855
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.84693 DF = 1 p-value = 0.1741409 D = -11.47977 f = 0.04236395
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.297344 DF = 1 p-value = 0.254699 D = 9.412698 f = -0.03626024
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9992242 DF = 1 p-value = 0.3174983 D = 6.027616 f = -0.0469176
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 8.3686 DF = 1 p-value = 0.003817595 D = 33.50919 f = -0.06352565
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5251952 DF = 1 p-value = 0.4686339 D = -8.294584 f = 0.01703919
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5671085 DF = 1 p-value = 0.4514102 D = -11.23508 f = 0.01349178
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.415354 DF = 1 p-value = 0.06459257 D = 13.25731 f = -0.06506889
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.253838 DF = 1 p-value = 0.2628208 D = -15.4113 f = 0.0213627
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 34.4094 DF = 1 p-value = 4.465612e-09 D = 39.14286 f = -0.2238562
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 13.46418 DF = 1 p-value = 0.0002431617 D = 21.09789 f = -0.1636737
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.03480967 DF = 1 p-value = 0.8519951 D = -1.443139 f = 0.01085531
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 60.8135 DF = 1 p-value = 6.274629e-15 D = -40.79528 f = 0.3492887
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.570515 DF = 1 p-value = 0.2101323 D = -7.346899 f = 0.05814863
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.17896 DF = 1 p-value = 0.6722685 D = -2.736554 f = 0.02185108
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 25.08128 DF = 1 p-value = 5.496383e-07 D = 36.18406 f = -0.1732821
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.334906 DF = 1 p-value = 0.5627845 D = -3.924959 f = 0.02583485
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3479463 DF = 1 p-value = 0.555278 D = 5.036 f = -0.0201469
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 19.3649 DF = 1 p-value = 1.079736e-05 D = -24.93812 f = 0.1996038
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.331001 DF = 1 p-value = 0.0679855 D = -16.08 f = 0.05394525
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.4230739 DF = 1 p-value = 0.5154072 D = -5.885496 f = 0.02023303
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.4939089 DF = 1 p-value = 0.4821888 D = -4.435289 f = 0.03311516
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 24.46085 DF = 1 p-value = 7.583506e-07 D = 85.09313 f = -0.06561237
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```

## Chi2 = 3.467243 DF = 1 p-value = 0.06259555 D = -12.97927 f = 0.0684997
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 174.1815 DF = 1 p-value = 9.035356e-40 D = 82.21688 f = -0.5328962
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.08108758 DF = 1 p-value = 0.7758289 D = -2.497478 f = 0.01119958
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 28.15002 DF = 1 p-value = 1.122668e-07 D = 61.89341 f = -0.114172
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.4623857 DF = 1 p-value = 0.4965111 D = 4.48375 f = -0.03029228
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.000456616 DF = 1 p-value = 0.9829516 D = -0.3466495 f = 0.002393217
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3410853 DF = 1 p-value = 0.5592034 D = 8.579024 f = -0.01068477
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3169737 DF = 1 p-value = 0.5734321 D = 4.673218 f = -0.01981632
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.05104855 DF = 1 p-value = 0.8212487 D = 1.6125 f = -0.01312176
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 3.173079 DF = 1 p-value = 0.07486119 D = 16.8789 f = -0.0484857
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 64.13849 DF = 1 p-value = 1.159733e-15 D = -60.21745 f = 0.2632831
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.2555046 DF = 1 p-value = 0.6132255 D = -3.493443 f = 0.02290881
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.5097411 DF = 1 p-value = 0.4752509 D = -4.662069 f = 0.03222731
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3770417 DF = 1 p-value = 0.5391907 D = -3.8 f = 0.03044872
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.7512408 DF = 1 p-value = 0.3860837 D = 5.28932 f = -0.04109465
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.9896148 DF = 1 p-value = 0.3198365 D = 5.1245 f = -0.04817369
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 6.441667 DF = 1 p-value = 0.0111474 D = 14 f = -0.1166667
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.008080521 DF = 1 p-value = 0.9283733 D = 1.323327 f = -0.002738806
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.070909 DF = 1 p-value = 0.3007402 D = -8.017699 f = 0.03611288
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.3002523 DF = 1 p-value = 0.5837243 D = -3.422 f = 0.02750317
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.09548535 DF = 1 p-value = 0.7573162 D = -1.91353 f = 0.01643735

## Warning in HWChisq(x): Expected counts below 5: chi-square approximation
## may be incorrect

## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 16.73636 DF = 1 p-value = 4.294986e-05 D = -8.089721 f = 0.1282257

## Warning in HWChisq(x): Expected counts below 5: chi-square approximation
## may be incorrect

## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 0.0281032 DF = 1 p-value = 0.8668664 D = -0.1361868 f = 0.003017241
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 7.308418 DF = 1 p-value = 0.006863233 D = -21.22661 f = 0.08873014
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)

```

```
## Chi2 = 0.2354712 DF = 1 p-value = 0.6274964 D = -2.592552 f = 0.02214105
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 = 1.141699 DF = 1 p-value = 0.2852937 D = 6.648276 f = -0.04703357
```

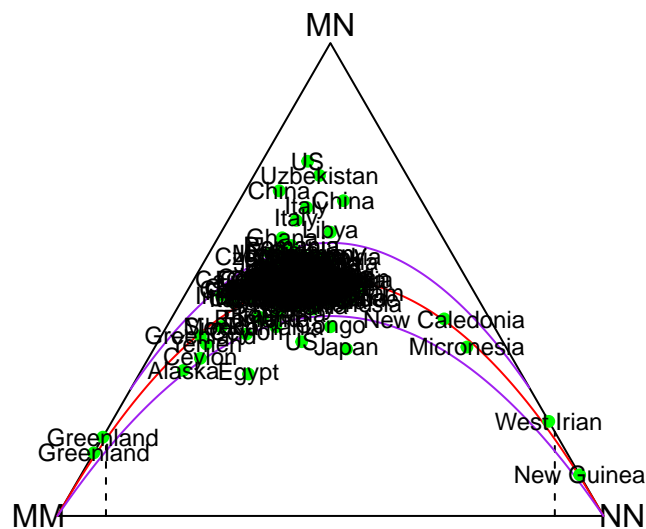
```
Mourant$p.val <- t
subset(Mourant, p.val < 0.05)
```

##	Population	Country	Total	MM	MN	NN	p.val
## 1	USSR	Russia	5728	1743	3222	763	3.253630e-33
## 8	Europe	Austria	6557	2185	2993	1379	3.667752e-09
## 9	Europe	Austria	3000	914	1548	538	8.617946e-03
## 10	Europe	Austria	10000	3156	4761	2083	2.564751e-04
## 16	Europe	Czechoslovakia	500	135	274	91	2.319407e-02
## 18	Europe	Czechoslovakia	5110	1457	2824	829	2.873010e-18
## 28	Europe	France	10694	3356	5178	2160	4.711102e-02
## 33	Europe	France	1082	370	490	222	1.291399e-02
## 38	Europe	Germany	4867	1446	2482	939	3.161787e-02
## 40	Europe	Germany	1059	274	567	218	1.846327e-02
## 55	Europe	Germany	546	157	303	86	3.195697e-03
## 61	Europe	Hungary	554	209	229	116	5.641774e-04
## 65	Europe	Italy	800	263	357	180	6.685156e-03
## 69	Europe	Italy	1164	254	736	174	3.995607e-20
## 71	Europe	Italy	850	221	518	111	4.137647e-12
## 74	Europe	Italy	1000	417	424	159	4.476244e-03
## 78	Europe	Jugoslavia	8278	2895	3891	1492	3.674746e-03
## 79	Europe	Jugoslavia	1000	304	543	153	5.166024e-04
## 94	Europe	Romania	1035	376	465	194	2.185831e-02
## 96	Europe	Romania	694	197	384	113	1.450376e-03
## 97	Europe	Romania	604	147	331	126	1.972470e-02
## 99	Europe	Romania	641	250	270	121	2.474380e-03
## 102	Europe	Romania	528	196	225	107	5.885692e-03
## 103	Europe	Romania	743	198	405	140	9.744467e-03
## 104	Europe	Romania	672	204	385	83	2.413303e-06
## 105	Europe	Romania	1503	425	802	276	2.913906e-03
## 106	Europe	Spain	535	153	241	141	2.668384e-02
## 128	USSR	Uzbekistan	1265	207	916	142	7.178839e-58
## 136	Asia	Ceylon	1068	607	358	103	9.398615e-06
## 137	Asia	Ceylon	660	319	252	89	9.967997e-04
## 144	Asia	China	1004	134	693	177	1.211144e-33
## 148	Asia	China	812	181	563	68	7.439819e-32
## 154	Asia	Japan	841	233	320	288	1.166967e-11
## 157	Asia	Japan	500	137	272	91	3.505795e-02
## 159	Asia	Japan	1633	536	842	255	1.259455e-02
## 165	Asia	Japan	743	187	400	156	3.686733e-02
## 167	Asia	Japan	646	207	278	161	7.512515e-04
## 168	Asia	Japan	1465	523	647	295	3.332464e-04
## 172	Asia	Japan	2121	576	1122	423	3.817595e-03
## 177	Africa	Libya	700	126	428	146	4.465612e-09
## 178	Africa	Marocco	521	137	300	84	2.431617e-04
## 180	Africa	Egypt	508	250	152	106	6.274629e-15
## 183	Africa	Ghana	853	243	490	120	5.496383e-07
## 186	Africa	Congo	501	163	200	138	1.079736e-05
## 190	America	Canada	5734	2370	2764	600	7.583506e-07
## 192	America	US	619	90	473	56	9.035356e-40
## 194	America	US	2186	587	1208	391	1.122668e-07


```
## 201    America          US    937  372  337  228 1.159733e-15
## 207    America          Mexico  500  166  268   66 1.114740e-02
## 212    Oceania    New Guinea 1148   12  110 1026 4.294986e-05
## 214    Oceania    Micronesia  962  228  436  298 6.863233e-03
```

Question 2.17: Divide all total frequencies by 50, keeping the same proportions for each of the genotypes, and recreate the ternary plot. a) what happens to the points? b) what happens to the confidence regions?

```
newgf = round(genotypeFrequencies / 50)
HWTernaryPlot(newgf,
  markerlab = Mourant$Country,
  alpha = 0.0001, curvecols = c("red", rep("purple", 4)),
  mcex = 0.75, vertex.cex = 1)
```



Confidence regions get bigger

2.7.3 Concatenating several multinomials: sequence motifs and logos

```
library("seqLogo")
load(url("http://bios221.stanford.edu/data/kozak.RData"))
kozak
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## A 0.33 0.25 0.4 0.15 0.20 1 0 0 0.05
## C 0.12 0.25 0.1 0.40 0.40 0 0 0 0.05
## G 0.33 0.25 0.4 0.20 0.25 0 0 1 0.90
## T 0.22 0.25 0.1 0.25 0.15 0 1 0 0.00
```

```
pwm = makePWM(kozak)
seqLogo(pwm, ic.scale = FALSE)
```

2.8 Modelling sequence dependencies: Markov chains

```
## Package: markovchain
```

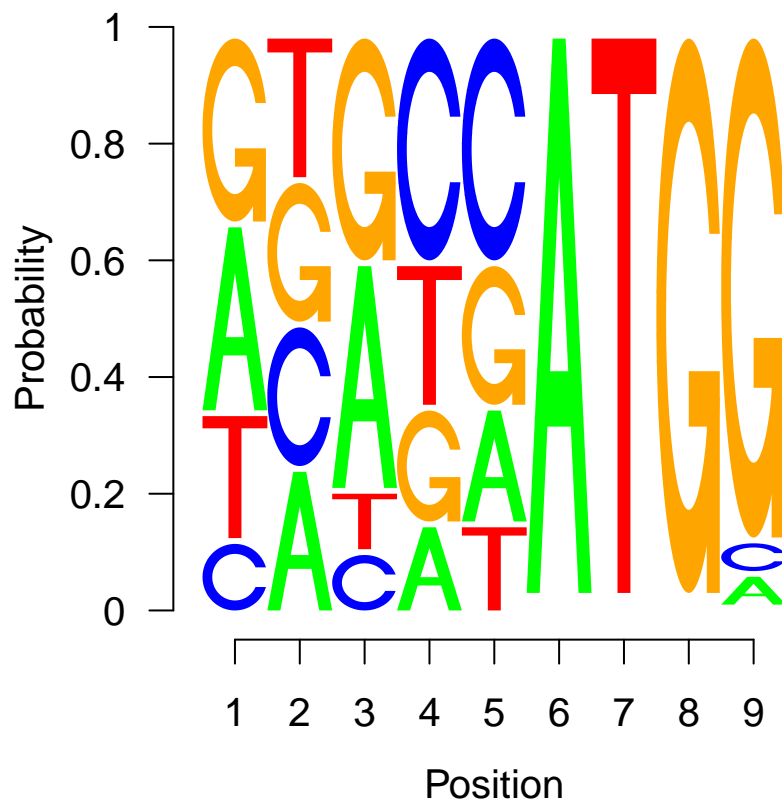


Figure 12: Here is a diagram called a sequence logo for the position dependent multinomial used to model the Kozak motif. It codifies the amount of variation in each of the positions on a log scale. The large letters represent positions where there is no uncertainty about which nucleotide occurs.

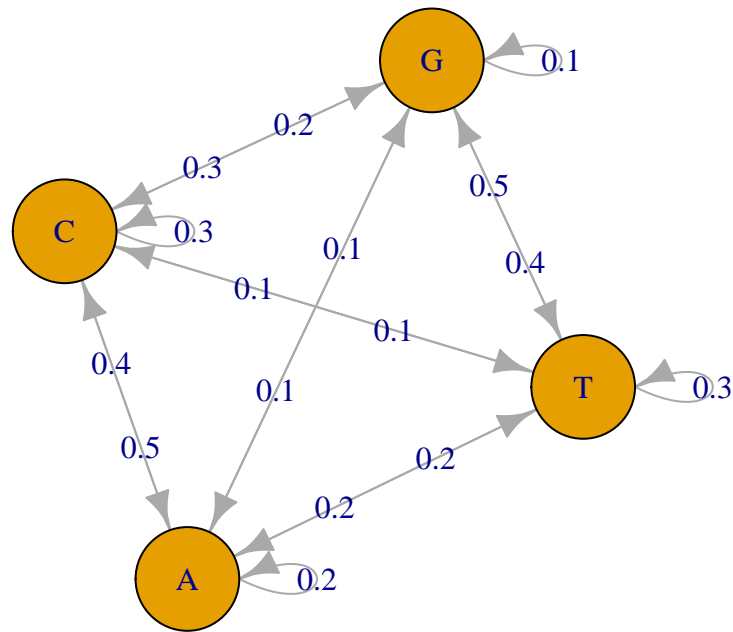


Figure 13: Visualisation of a 4-state Markov chain. The probability of each possible digram (e.g., CA) is given by the weight of the edge between the corresponding nodes. So for instance, the probability of CA is given by the edge $C \rightarrow A$. We'll see in Chapter @ref(Chap:Images) how to use **R** packages to draw these type of network graphs.

```

## Version: 0.8.0
## Date: 2019-09-13
## BugReport: http://github.com/spedygiorgio/markovchain/issues
##
## Attaching package: 'igraph'
##
## The following object is masked from 'package:Biostrings':
##
## union
##
## The following object is masked from 'package:IRanges':
##
## union
##
## The following object is masked from 'package:S4Vectors':
##
## union
##
## The following objects are masked from 'package:BiocGenerics':
##
## normalize, path, union
##
## The following objects are masked from 'package:stats':
##
## decompose, spectrum
##
## The following object is masked from 'package:base':
##
## union

```

2.9 Bayesian thinking

The *baysian paradigm* is a practical approach in which *prior* and *posterior* distributions are used as models of our knowledge **before** and **after** collecting some data and making an observation. We formalize our prior probability as $P(H)$. After seeing the data we have the posterior probability, $P(H|D)$.

2.9.1 Example: Haplotype frequencies

```
haplo6 <- read.table("../data/haplotype6.txt",header = TRUE)
haplo6
```

##	Individual	DYS19	DXYS156Y	DYS389m	DYS389n	DYS389p
## 1	H1	14	12	4	12	3
## 2	H3	15	13	4	13	3
## 3	H4	15	11	5	11	3
## 4	H5	17	13	4	11	3
## 5	H7	13	12	5	12	3
## 6	H8	16	11	5	12	3

2.9.2 Simulation study of the Bayesian paradigm for the binomial

Instead of assuming our parameter Θ has one single value, the bayesian world view allows us to see it as a draw from a statistical distribution. When we are looking at a parameter that expresses a proportion or a probability, and that takes its values between 0 and 1, it is convenient to use the *beta distribution*

Why 50 & 350 for alpha and beta?

```
## Loading required package: reshape2
##
## Attaching package: 'reshape2'
## The following object is masked from 'package:tidyr':
##
## smiths
```

The distribution of γ

What is the distribution of Y if Θ itself also varies according to some distribution. We call this the *marginal* distribution of Y.

```
rtheta = rbeta(100000, 50, 350)
y = vapply(rtheta, function(th) {
  rbinom(1, prob = th, size = 300)
}, numeric(1))
hist(y, breaks = 50, col = "orange", main = "", xlab = "")
```

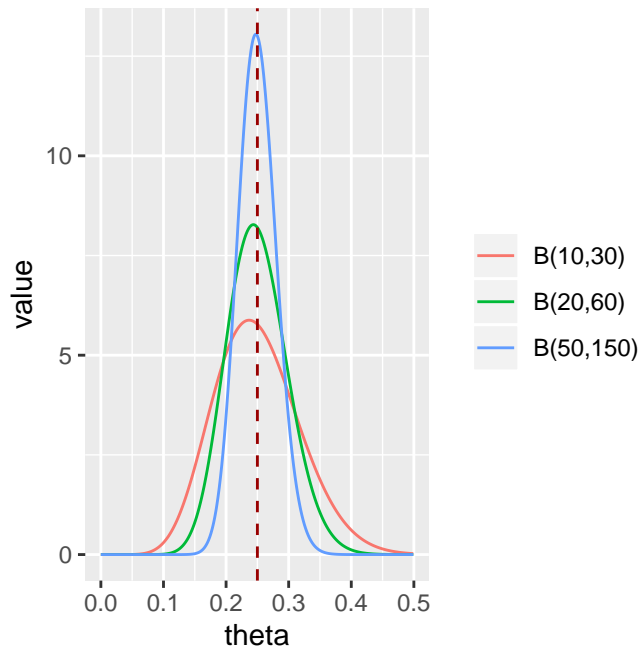
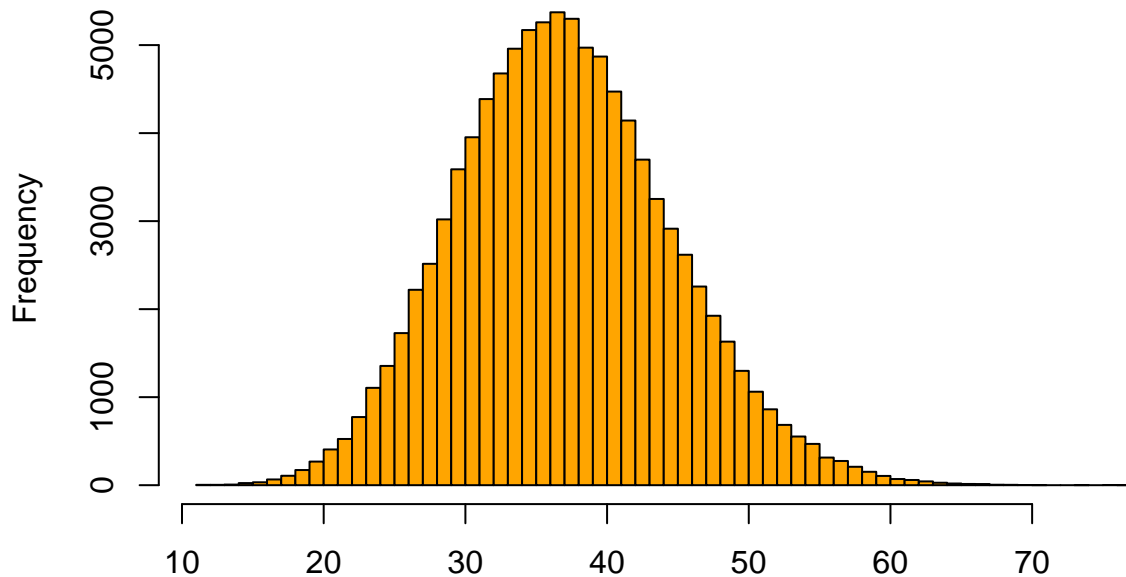


Figure 14: Beta distributions with $\alpha = 10, 20, 50$ and $\beta = 30, 60, 150$ used as a {prior} for a probability of success. These three distributions have the same mean ($\frac{\alpha}{\alpha+\beta}$), but different concentrations around the mean.



```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

```
Mode(y)
```

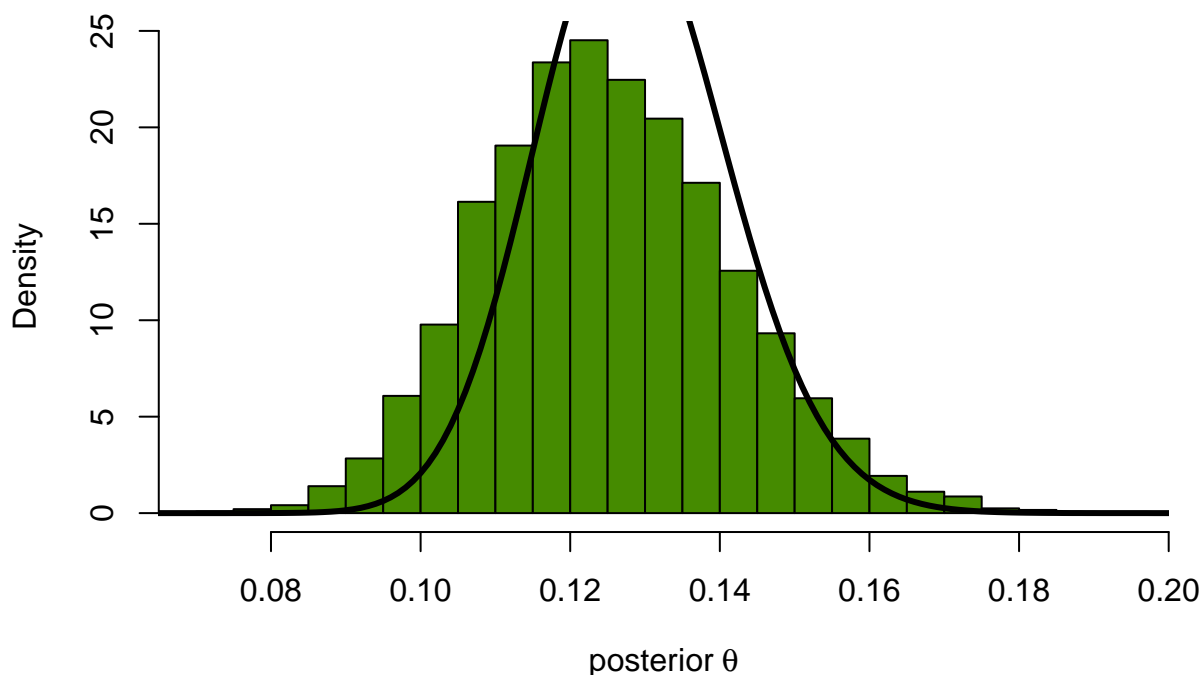
```
## [1] 37
```

Question 2.18: Verify that we can get the same result as in the above code chunk by using R's vectorization capabilities and writing 'rbinom(length(rtheta), rtheta, size = 300)'.

Compute the posterior distribution of Θ by conditioning on outcomes where Y is 40. Compare it to the theoretical posterior `densPostTheory`. [HELP!](#)

```
rtheta = rbeta(100000, 50, 350) #there are 10,000 observations, pulled from a beta distribution with sh
#y = vapply(rtheta, function(th) {rbinom(1, prob = th, size = 300)}, numeric(1))
#y is a vector of length rtheta where a single observation was drawn from the binomial distribution 300
# y = 40 is asking for the rtheta values that quasi-by-chance gave 40/300 successes from the binomial d

thetaPostEmp = rtheta[y == 40] #~ 45% of all y values are equal to 40; 40 is a most common result
hist(thetaPostEmp, breaks = 40, col = "chartreuse4", main = "",
     probability = TRUE, xlab = expression("posterior"~theta))
densPostTheory = dbeta(thetas, 90, 610)
lines(thetas, densPostTheory, type="l", lwd = 3)
```



Check the means of both distributions computed above and see that they are close to four significant digits:

Calculate the integral under the curve

```
mean(thetaPostEmp)
```

```
## [1] 0.125026
```

```
mean(rtheta)
```

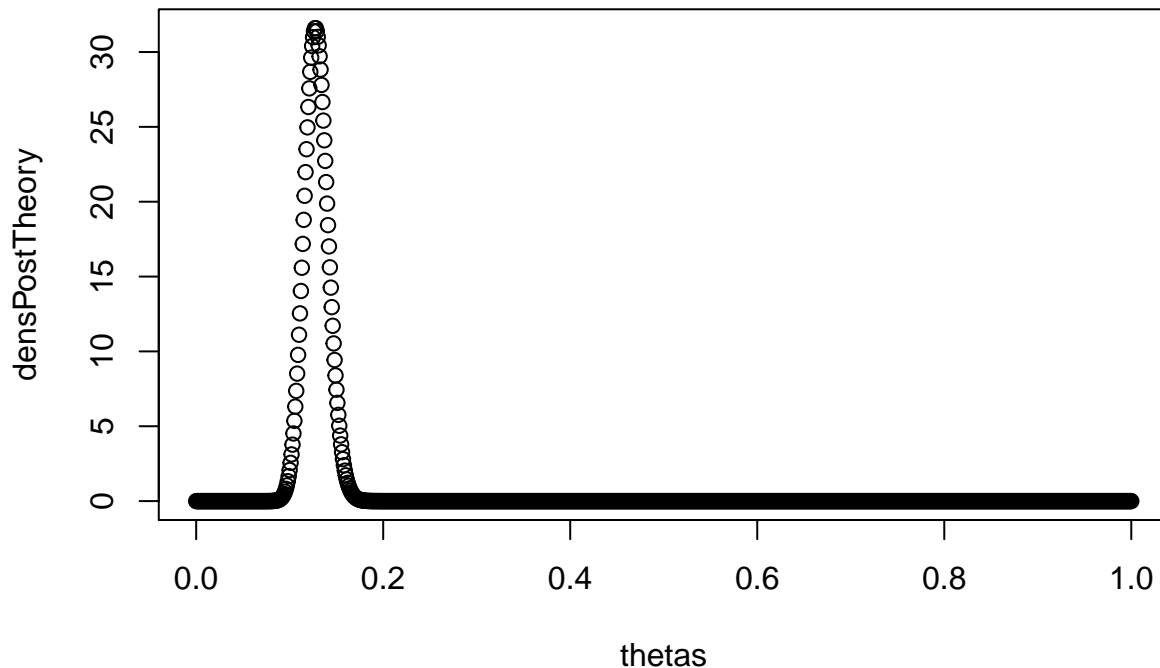
```
## [1] 0.1249781
```

```
dtheta = thetas[2]-thetas[1]
```

```
sum(thetas * densPostTheory * dtheta)
```

```
## [1] 0.1285714
```

```
plot(thetas, densPostTheory)
```



To approximate the mean of the theoretical distribution `densPostTheory` we computed the integral $\int_0^1 \Theta f(\Theta) d\Theta$ using numerical integration, i.e., the sum over the integrand. Use *monte carlo integration* instead. Again, where do the shape numbers 90, 610 come from?? Why is this equivalent??

```
thetaPostMC = rbeta(n = 1e6, 90, 610)
mean(thetaPostMC)
```

```
## [1] 0.1285737
```

Check the concordance between Monte Carlo simulation sample `thetaPostMC` and our sample `thetaPostEmp` using a Q-Q plot

```
qqplot(thetaPostMC, thetaPostEmp, type = "l", asp = 1)
abline(a = 0, b = 1, col = "blue")
```

Question 2.19 What is the difference between the simulation that results in ‘`thetaPostEmp`’ and the Monte Carlo simulation that leads to ‘`thetaPostMC`’?

Posterior distribution is also a beta

The parameters $\alpha = 90$ and $\beta = 610$ were obtained by summing the prior parameters $\alpha = 50$ and $\beta = 350$ with the observed successes $y = 40$ and the observed failures $n-y = 260$, obtaining the posterior

$$\text{beta}(90, 610) = \text{beta}(\alpha + y, \beta + (n - y))$$

We can use this to give the best estimate we can for Θ with its uncertainty given by the posterior distribution. This is called the MAP estimate, $\frac{\alpha-1}{\alpha+\beta-2} = \frac{89}{698} = 0.1275$

Suppose we have a second series of data After seeing our previous data we now have a new prior, $\text{beta}(90, 610)$. Now we collect a new data set with $n = 150$ observations and only $y = 25$ successes

The new posterior will be $\text{beta}(90+25 = 115, 610+125 = 735)$; the mean is $\frac{115}{115+735} = 0.135$, thus one estimate of Θ is 0.135.

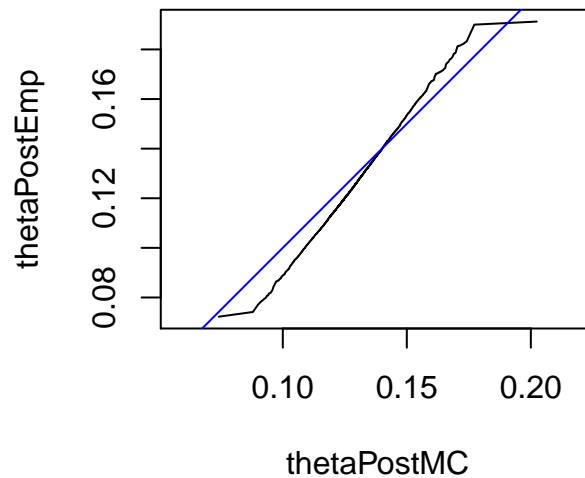


Figure 15: QQ-plot of our Monte Carlo sample `thetaPostMC` from the theoretical distribution and our simulation sample `thetaPostEmp`. We could also similarly compare either of these two distributions to the theoretical distribution function `pbeta(., 90, 610)`. If the curve lies on the line $y = x$ this indicates a good agreement. There are some random differences at the tails.

The theoretical *maximum a posteriori (MAP) estimate* would be the mode of $\text{beta}(115, 735) = 0.134$.

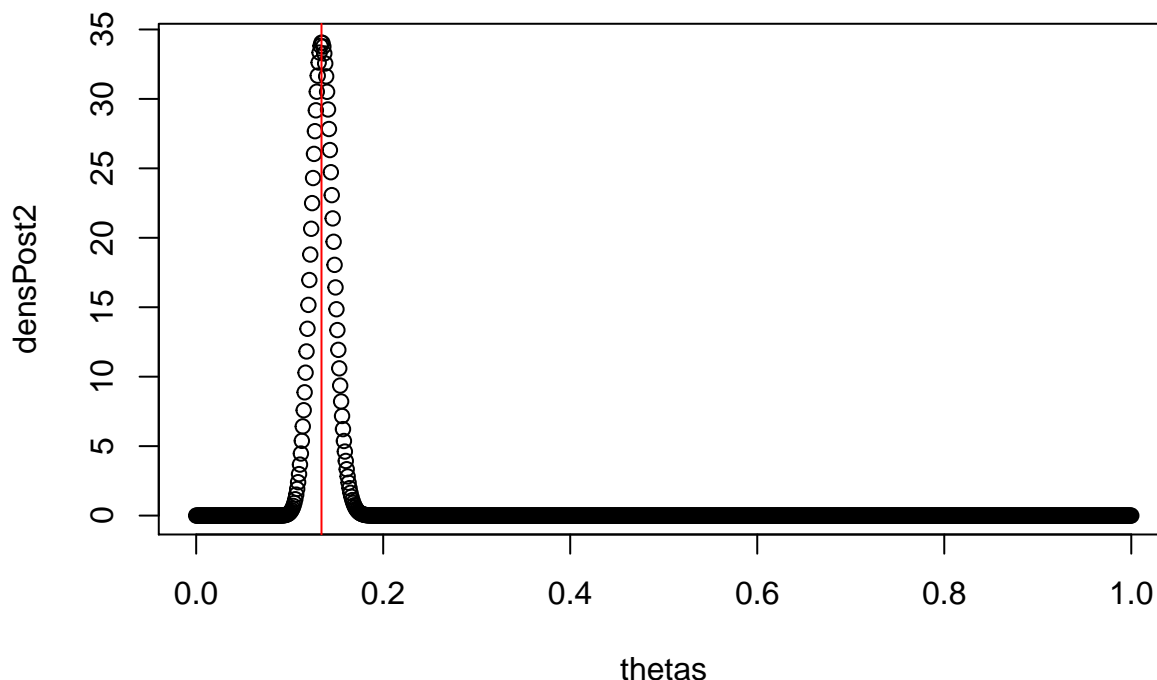
```
densPost2 = dbeta(thetas, 115, 735)
mcPost2 = rbeta(1e6, 115, 735)

sum(thetas*densPost2*dtheta) #mean, by numeric integration

## [1] 0.1352941
mean(mcPost2) #mean, by MC

## [1] 0.1352882
thetas[which.max(densPost2)]

## [1] 0.134
plot(thetas, densPost2)
abline(v=thetas[which.max(densPost2)], col="red")
```

Question 2.20 Redo all the computations replacing our original prior with a softer prior (less peaked), meaning that we use less prior information. How much does that change the final result?

```
rtheta = rbeta(100000, 50, 100) #there are 10,000 observations, pulled from a beta distribution with sh
densPostTheory = dbeta(thetas, 90, 610)
```

Confidence statements for the proportion parameter

Reach a conclusion about where the proportion actually lies given the data. The *posterior credibility interval* is a bayesian analog of a confidence interval. We can take the 2.5th and 97.5th percentiles of the posterior distribution $P(L \leq \Theta \leq U) = 0.95$ as

```
quantile(mcPost2, c(0.025, 0.975))
```

```
##      2.5%      97.5%
## 0.1131169 0.1591032
```

2.10 Example: occurrence of a nucleotide pattern in a genome

```
library("Biostrings")
```

```
## GENETIC_CODE
## IUPAC_CODE_MAP
## vignette(package = "Biostrings")
## vignette("BiostringsQuickOverview", package = "Biostrings")
```

```
library("BSgenome")
```

```
## Loading required package: GenomeInfoDb
## Loading required package: GenomicRanges
## Loading required package: rtracklayer
```

```
##
## Attaching package: 'rtracklayer'
## The following object is masked from 'package:igraph':
##
##      blocks
ag = available.genomes()
length(ag)

## [1] 93
ag[1:2]

## [1] "BSgenome.Alyrata.JGI.v1"
## [2] "BSgenome.Amelliifera.BeeBase.assembly4"
library("BSgenome.Ecoli.NCBI.20080805")
Ecoli
shineDalgarno <- "AGGAGGT"
ecoli <- Ecoli$NC_010473

window = 50000
starts = seq(1, length(ecoli) - window, by = window)
ends   = starts + window - 1
numMatches = vapply(seq_along(starts), function(i) {
  countPattern(shineDalgarno, ecoli[starts[i]:ends[i]],
    max.mismatch = 0)
}, numeric(1))
table(numMatches)

## numMatches
##  0  1  2  3  4
## 48 32  8  3  2
```

Question 2.22: What distribution might this fit? Poisson

What does the poisson plot below mean?

```
library("vcd")
gf = goodfit(numMatches, "poisson")
summary(gf)
```

```
##
##      Goodness-of-fit test for poisson distribution
##
##              X^2 df  P(> X^2)
## Likelihood Ratio 4.134932  3 0.2472577
distplot(numMatches, type = "poisson")
```

Inspect the matches using the matchPattern function

```
sdMatches <- matchPattern(shineDalgarno, ecoli, max.mismatch = 0)
sdMatches
```

```
##      Views on a 4686137-letter DNAString subject
```

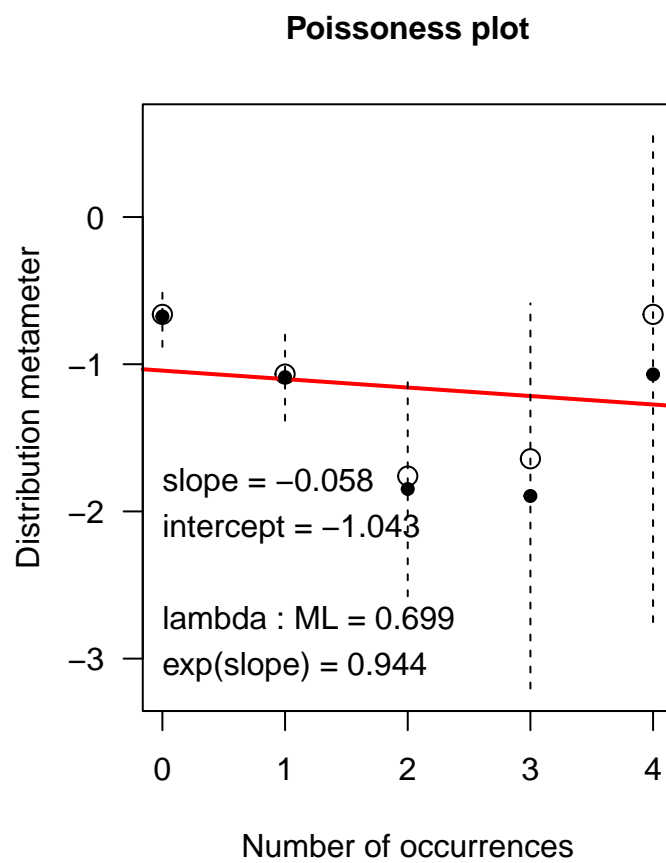


Figure 16: Evaluation of a Poisson model for motif counts along the sequence.

```
## subject: AGCTTTTCATTCTGACTGCAACGGGCAATATG...CAAATAAAAAACGCCTTAGTAAGTATTTTC
## views:
##      start      end width
## [1]   56593   56599     7 [AGGAGGT]
## [2]  199644  199650     7 [AGGAGGT]
## [3]  202176  202182     7 [AGGAGGT]
## [4]  214433  214439     7 [AGGAGGT]
## [5]  217429  217435     7 [AGGAGGT]
## ...      ...      ...    ...
## [61] 4438786 4438792     7 [AGGAGGT]
## [62] 4498085 4498091     7 [AGGAGGT]
## [63] 4536658 4536664     7 [AGGAGGT]
## [64] 4546821 4546827     7 [AGGAGGT]
## [65] 4611626 4611632     7 [AGGAGGT]
```

What are the distances between them?

```
betweenmotifs <- gaps(sdMatches)
```

Find a model for the distribution of gap sizes between motifs. If they occur at random locations we expect them to follow an *exponential distribution*.

```
library("Renext")
```

```
## Loading required package: evd
```

```
##
```

```
## Attaching package: 'evd'
```

```
## The following object is masked from 'package:igraph':
```

```
##
```

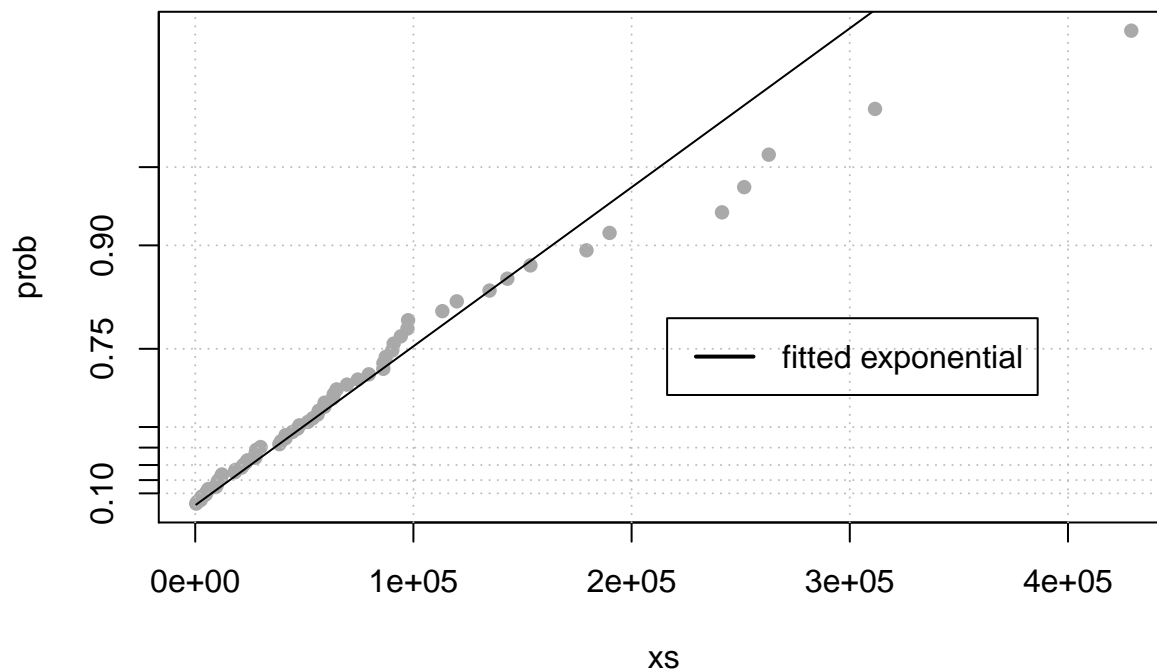
```
##      clusters
```

```
## The following object is masked from 'package:lattice':
```

```
##
```

```
##      qq
```

```
expplot(width(betweenmotifs), rate = 1/mean(width(betweenmotifs)),
        labels = "fitted exponential")
```



Question 2.23: There appears to be a slight deviation from the fitted line in Figure 2.23 at the right tail of the distribution, i.e., for the largest values. What could be the reason?

Bias in where genes are located?

2.10.1 Modeling in the case of dependencies

Dependency modelling using a **markov chain**. Discover differences between regions called CpG islands and the rest of the genome.

```
library("BSgenome.Hsapiens.UCSC.hg19")
chr8 = Hsapiens$chr8
CpGtab = read.table("../data/model-based-cpg-islands-hg19.txt",
                    header = TRUE)
nrow(CpGtab)
```

```
## [1] 65699
```

```
head(CpGtab)
```

```
##      chr  start    end length CpGcount GCcontent pctGC obsExp
## 1 chr10  93098  93818    721      32      403 0.559 0.572
## 2 chr10  94002  94165    164      12       97 0.591 0.841
## 3 chr10  94527  95302    776      65      538 0.693 0.702
## 4 chr10 119652 120193    542      53      369 0.681 0.866
## 5 chr10 122133 122621    489      51      339 0.693 0.880
## 6 chr10 180265 180720    456      32      256 0.561 0.893
```

```
irCpG = with(dplyr::filter(CpGtab, chr == "chr8"), #filter to only include chromosome 8
             IRanges(start = start, end = end))    #define IRanges object with start and end positions
```

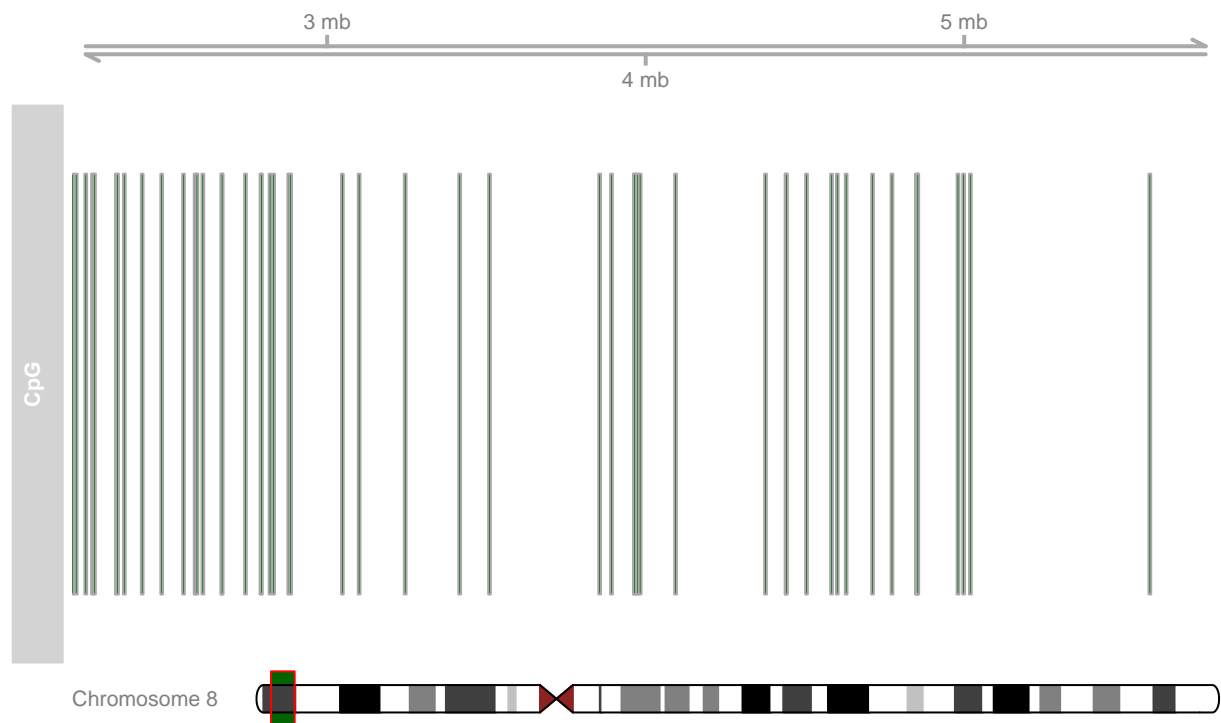


Figure 17: **Gviz** plot of CpG locations in a selected region of chromosome 8.

IRanges is a general “container” for mathematical intervals. **GRanges** is a container for the genomic locations and their associated annotations.

```
grCpG = GRanges(ranges = irCpG, seqnames = "chr8", strand = "+")
genome(grCpG) = "hg19"
```

```
library("Gviz")
ideo = IdeogramTrack(genome = "hg19", chromosome = "chr8")
plotTracks(
  list(GenomeAxisTrack(),
       AnnotationTrack(grCpG, name = "CpG"), ideo),
  from = 2200000, to = 5800000,
  shape = "box", fill = "#006400", stacking = "dense")
```

```
CGIview = Views(unmasked(Hsapiens$chr8), irCpG)
NonCGIview = Views(unmasked(Hsapiens$chr8), gaps(irCpG))
```

Compute the transition counts in CpG islands and non-islands using the data:

```
seqCGI = as(CGIview, "DNAStringSet") #there are 2855 different sequences
seqNonCGI = as(NonCGIview, "DNAStringSet")
dinucCpG = sapply(seqCGI, dinucleotideFrequency) #calculates all dinucleotides for a set of sequences
dinucNonCpG = sapply(seqNonCGI, dinucleotideFrequency)
dinucNonCpG[, 1]
```

```
## AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT
## 389 351 400 436 498 560 112 603 359 336 403 336 330 527 519 485
```

```
NonICounts = rowSums(dinucNonCpG) #there are 16 different dinucleotides -> sum across all sequences
```

```
IslCounts = rowSums(dinucCpG)
```

For a four state Markov chain we define the transition matrix as a matrix where the rows are the “From” state and the columns are the “to” state.

```
TI = matrix( IslCounts, ncol = 4, byrow = TRUE)
TnI = matrix(NonICounts, ncol = 4, byrow = TRUE)
dimnames(TI) = dimnames(TnI) =
  list(c("A", "C", "G", "T"), c("A", "C", "G", "T"))
```

Use the counts of numbers of transitions of each type to compute frequencies and put them in matrices:

```
MI = TI /rowSums(TI)
MI
```

```
##           A           C           G           T
## A 0.20457773 0.2652333 0.3897678 0.1404212
## C 0.20128250 0.3442381 0.2371595 0.2173200
## G 0.18657245 0.3145299 0.3450223 0.1538754
## T 0.09802105 0.3352314 0.3598984 0.2068492
```

```
MN = TnI / rowSums(TnI)
MN
```

```
##           A           C           G           T
## A 0.3351380 0.1680007 0.23080886 0.2660524
## C 0.3641054 0.2464366 0.04177094 0.3476871
## G 0.2976696 0.2029017 0.24655406 0.2528746
## T 0.2265813 0.1972407 0.24117528 0.3350027
```

Question 2.24: Are the transitions different in the different rows? i.e., $P(A|C) \neq P(A|T)$ No.

Question 2.25: Are the relative frequencies of the different nucleotides different in CpG islands compared to elsewhere?

```
freqIsl = alphabetFrequency(seqCGI, baseOnly = TRUE, collapse = TRUE)[1:4]
freqIsl / sum(freqIsl)
```

```
##           A           C           G           T
## 0.1781693 0.3201109 0.3206298 0.1810901
```

```
freqNon = alphabetFrequency(seqNonCGI, baseOnly = TRUE, collapse = TRUE)[1:4]
freqNon / sum(freqNon)
```

```
##           A           C           G           T
## 0.3008292 0.1993832 0.1993737 0.3004139
```

Question 2.26: Use a χ^2 statistic to compare the frequencies between the observed and ‘freqIsl’ and ‘freqNon’ frequencies

```
t <- data.frame(freqIsl, freqNon)
chisq.test(t)
```

```
##
## Pearson's Chi-squared test
##
## data:  t
## X-squared = 423898, df = 3, p-value < 2.2e-16
```

Given a sequence for which it's unknown if it's a CpG island or not, ask the probability if it is. Compute a score based on what is called the odds ratio -> use probabilities for each dinucleotide for islands and compare to probability for non-islands. Then take their ratio and see if it's larger or smaller than one.

Probabilities will be the products of many small terms and become small, so work around this by taking the logarithm. This is the *log-likelihood score*.

```
alpha = log((freqIsl/sum(freqIsl)) / (freqNon/sum(freqNon)))
beta  = log(MI / MN)
```

```
x <- "ACGTTATACTACG"
scorefun = function(x) {
  s <- unlist(strsplit(x, ""))
  score <- alpha[s[1]]
  if (length(s) >= 2)
    for (j in 2:length(s))
      score = score + beta[s[j-1], s[j]]
  score
}

testX <- scorefun(x)
```

Then pick sequences of length `len = 100` out of the 2855 sequences in the `seqCGI` object and then out of the 2854 `seqNonCGI` object. Drop sequences that contain any letter other than A, c, T, G (i.e. “.”). Sample with probabilities proportional to their length minus `len` and pick subsequences of length `len` out of them.

```
generateRandomScores = function(s, len = 100, B = 1000) {
  alphFreq = alphabetFrequency(s)
  isGoodSeq = rowSums(alphFreq[, 5:ncol(alphFreq)]) == 0
  s = s[isGoodSeq]
  slen = sapply(s, length)
  prob = pmax(slen - len, 0)
  prob = prob / sum(prob)
  idx = sample(length(s), B, replace = TRUE, prob = prob)
  ssmpl = s[idx]
  start = sapply(ssmpl, function(x) sample(length(x) - len, 1))
  scores = sapply(seq_len(B), function(i)
    scorefun(as.character(ssmpl[[i]][start[i]+(1:len)])))
  )
  scores / len
}

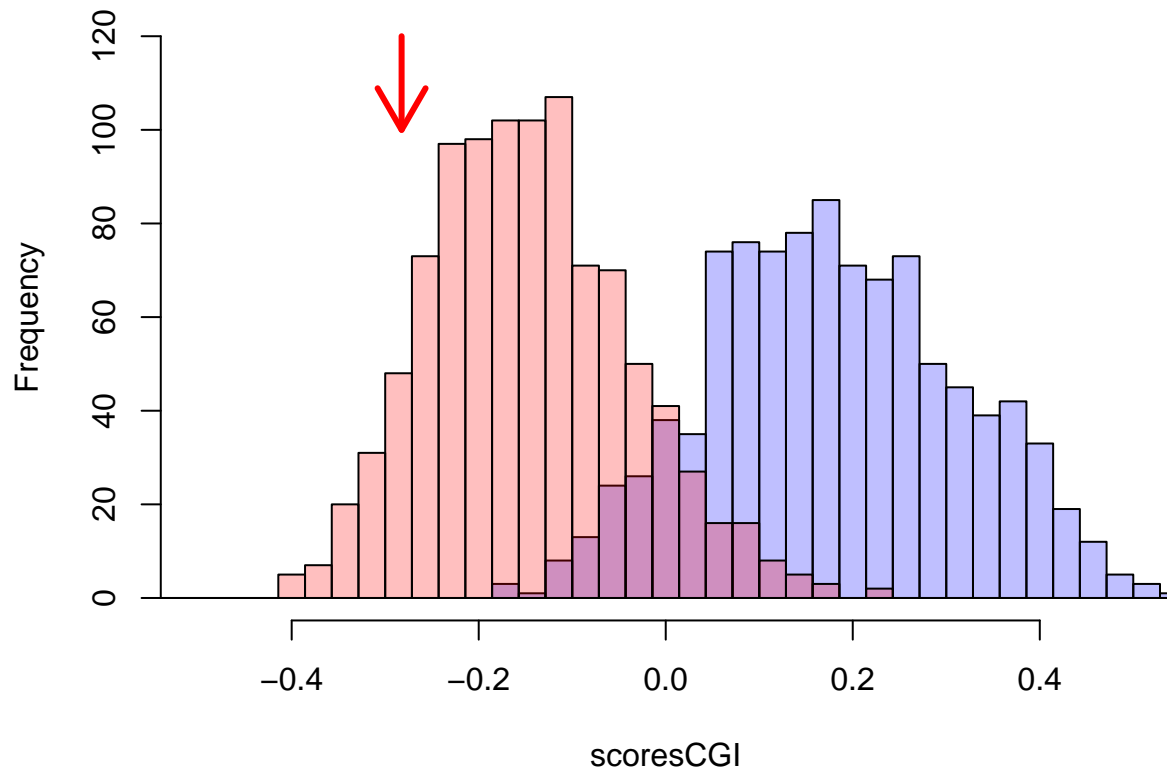
scoresCGI = generateRandomScores(seqCGI)
scoresNonCGI = generateRandomScores(seqNonCGI)
```

```
br = seq(-0.7, 0.7, length.out = 50) #need to change this to -0.7 or else the breaks did not span the range
h1 = hist(scoresCGI, breaks = br, plot = FALSE)
h2 = hist(scoresNonCGI, breaks = br, plot = FALSE)
plot(h1, col = rgb(0, 0, 1, 1/4), xlim = c(-0.5, 0.5), ylim = c(0, 120))
plot(h2, col = rgb(1, 0, 0, 1/4), add = TRUE)
```



```
arrows(testX, 120, testX, 100, lwd = 3, col="red")
```

Histogram of scoresCGI



consider this training data. Cool.

We can