LANGEBIO - BIOSTATISTICS

OCTAVIO MARTÍNEZ DE LA VEGA

11. Statistical Inference: Hypothesis Testing

Now we know, from the previous section, how to estimate parameters. However, in many cases of interest we want also to test an hypothesis about the parameters that determine the distribution. Here we are going to learn how to do so. There are hundreds, possible thousands of different "Statistical Tests" however all these recipes are based in a few and simple common sense concepts. I do not want you to memorize the "recipes" (formulas or algorithms), what I want is that you understand the concepts behind them.

We begin with an interesting real dataset that we have seen before, and then we will try to generalize to almost any possible test that you could want.

11.1. **Example: 9:7 or 3:1.** Consider again the data of the experiment with *Capsicum*, Garcia-Neria and Rivera-Bustamante (2011; MPMI, Vol. 24, pp. 172-182). They crossed plants resistant to a virus with susceptible ones, obtaining in the F_1 only resistant plants and in the F_2 156 resistant and 130 susceptible descendants (a total of 286 plants in that generation).

The fact that all F_1 plants are resistant indicate that there are at least one locus with a dominant allele for resistance. The simplest model is to think that there is a locus $\mathbb{R}$ with alleles $\{R_1, R_2\}$ such that the phenotype of the individuals are $Phe(R_1R_1) = Resistant$, $Phe(R_1R_2) = Resistant$ and $Phe(R_2R_2) = Susceptible$.

Under that assumption the segregation of the F_2 will give a Mendelian segregation

$$1 : R_1R_1, 2 : R_1R_2, 1 : R_2R_2$$

and thus the phenotypes will be present in a proportion

$$3 : Resistant, 1 : Susceptible$$

Now, let X be a random variable defined as "the number of resistant descendants" in the F_2. We have seen that this variable has the Binomial distribution of parameters $k = 286$ and an unknown parameter p, which we have estimated by maximum likelihood as

$$\hat{p} = \frac{156}{286} = 0.5559441$$

Date: April 2012.

When we did that we did not have any hypothesis, but now we have a genetic hypothesis that *induces* the following Statistical Hypothesis:

$H_0 : p = 3/4 = 0.75$

Let's call this our "null" hypothesis.

What is the **expected** value of X under this hypothesis? -That is, the value of X that we "expect" X to have if H_0 is true. Well We know that if this hypothesis is true then

$$X \sim \mathcal{B}(k = 286, p = 3/4)$$

and thus

$$E[X|H_0] = pk = (3/4)286 = 214.5 \approx 215$$

Now, the observed value of resistants is 156, and it appears to be too low. Lets assume that H_0 is true and let's compare the probabilities of $X = 215$ (the expected value "rounded") with the probability of $X = 156$, the observed value.

```
> dbinom(x=215, size=286, prob=3/4) # Prob. of Expected under H0.
[1] 0.05441006
> dbinom(x=156, size=286, prob=3/4) # Prob. of Observed  under H0.
[1] 3.144458e-14
> # Let's calculate the ratio of the probabilities
> dbinom(x=215, size=286, prob=3/4)/dbinom(x=156, size=286, prob=3/4)
[1] 1.730348e+12
```

That is, we find that it is incredibly more likely, say $1.730348e{+}12 = 1,730,348,000,000$ times more likely to have the expected than the observed value. We can suspect that the null hypothesis is false.

Note that no value of X is *impossible*, but there are some much more likely than others. Evidently, the hypothesis that has more support from the data is the hypothesis that $p = \hat{p} = 156/2860.5559441$. However, that hypothesis does not have "Genetic Meaning". Can we find a Genetic Hypothesis that fit better the data founded?. Let's try.

After some thought, the authors came with a model including two loci, say $\mathbb{A}$ and $\mathbb{B}$, each one containing to alleles, say $\{A, a\}$ and $\{B, b\}$ and they assumed a genetic model of "double recessive epistasis" where the genotypes homocigous for a or b result into a Susceptibel genotype and the upper case letter, A or B, give resistance. Representing by $Ph(\bullet)$ the function "Phenotype" and with R and S denoting the Resistant and Susceptible phenotypes we have

$$Ph(AABB) = Ph(AABb) = Ph(AaBB) = Ph(AaBb) = R$$

and

$$Ph(aabb) = Ph(aaBb) = Ph(aaBB) = Ph(AAbb) = Ph(Aabb) = S$$

Now, this genetic hypothesis imply that the segregation in the F_2 will be

$$9 : R, 7 : S$$

and thus this imply an *alternative* statistical hypothesis, say

$H_a : p = 9/16 = 0.5625$

Note that if this hypothesis is true then

$$X \sim \mathcal{B}(k = 286, p = 9/16)$$

and thus

$$E[X|H_a] = pk = (9/16)286 = 160.875 \approx 161$$

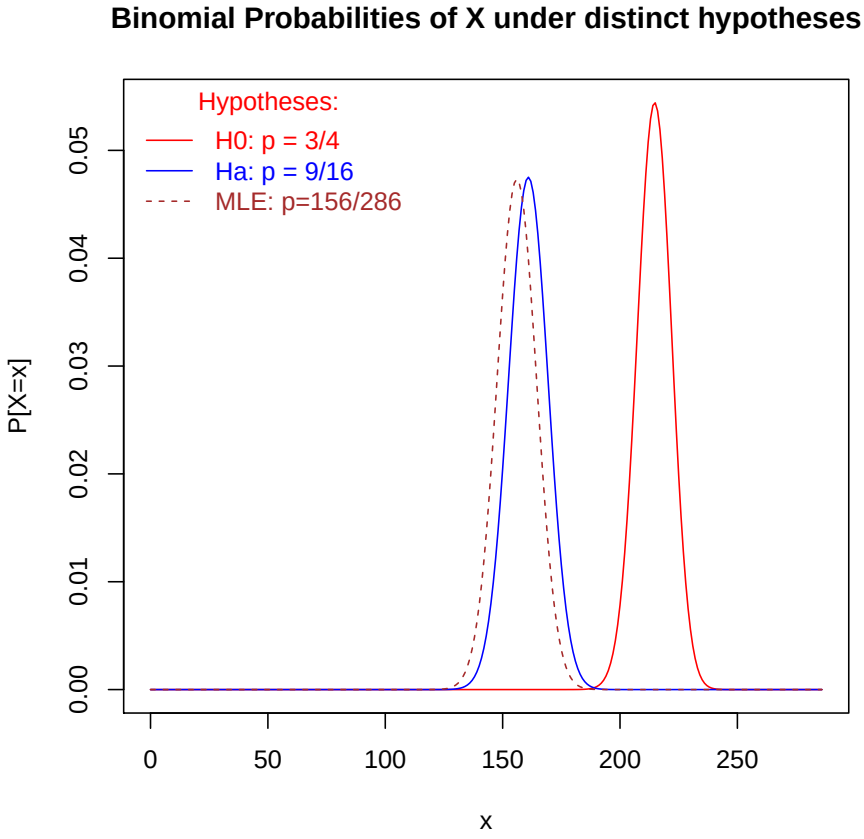
Let's, again, compare the probabilities of this expected value with the observed one.

```
> dbinom(x=161, size=286, prob=9/16) # Prob. of Expected under Ha
[1] 0.04751045
> dbinom(x=156, size=286, prob=9/16) # Prob. of Observed under Ha
[1] 0.03999901
> # Let's calculate the ratio of the probabilities
> dbinom(x=161, size=286, prob=9/16)/dbinom(x=156, size=286, prob=9/16)
[1] 1.187791
```

We see that the probabilities are of comparable size, there is only slightly more likely (1.187791 times) to obtain the "Expected" than the "Observed" value. We can happily conclude that this hypothesis (H_a) fits the data well, and thus we cannot reject it with the evidence that we have.

Let's compare the probability distribution under the "null" (H_0), "alternative" (H_a) and the one with our maximum likelihood estimation of p.

```
> plot(c(0:286), dbinom(c(0:286), 286, 3/4), type="l",
+ col="red", xlab="x", ylab="P[X=x]",
+ main="Binomial Probabilities of X under distinct hypotheses") # Under H0
> points(c(0:286), dbinom(c(0:286), 286, 9/16),
+ type="l", col="blue") # Under Ha
> points(c(0:286), dbinom(c(0:286), 286, 156/286),
+ type="l", col="brown", lty=2) # MLE
> legend("topleft", title="Hypotheses:",
+ legend=c("H0: p = 3/4", "Ha: p = 9/16", "MLE: p=156/286"),
+ bty="n",
+ text.col=c("red", "blue", "brown"),
+ col=c("red", "blue", "brown"),
+ lty=c(1,1,2))
```



Note that the distribution under the two hypothesis are very different, and also that the MLE is very close to the hypothesis H_a, thus this small deviation can be very well explained by the random nature of the phenomenon (error). In fact, in their paper the authors concluded that the inheritance of the character is given by two loci with recessive epistatic effects, as the model shown here as Ha.

11.2. **Definition. Test of a Statistical Hypothesis.** A *test* of a statistical hypothesis H is a rule or procedure to decide whether to reject or not H.

Examples:

- Throw a coin, if it falls "heads" then reject H
- Calculate the statistic

$$z_0 = \frac{|\overline{x} - \mu_0|}{ee(\overline{x})}$$

reject H if $z_0 > 1.96$

11.3. **Definition. Types of Errors.** Assume that we have an statistical hypothesis, H_0, then either H_0 is **true** or H_0 is **false**. Those *states of nature* are unknown to us. Using our test (rule) we can **reject** H_0 or **accept** H_0. The situation is presented in the following table

	States of Nature	
	H_0 is **true**	H_0 is **false**
Reject H_0	Error Type I, $P = \alpha$	Right, $P = 1 - \beta$
Accept H_0	Right, $P = 1 - \alpha$	Error Type II, $P = \beta$

In general we will want α and β to be small, and we can compare distinct tests (rules to reject H_0) comparing the associated errors.

11.4. **Example with *Phycomyces* data.** To illustrate Statistical Tests we will be doing an intensive analysis of the data of the fungus *Phycomyces* (Alfredo Herrera-Estrella's Lab) that you inputed into R in the first homework. These data are

Gene	$S.R_1$	$S.R_2$	$M.R_1$	$M.R_2$
17	22 **(40)**	6 **(10)**	30 **(46)**	18 **(19)**
1661	56 **(101)**	43 **(71)**	1,440 **(2,205)**	2,163 **(2,317)**
7433	40 **(72)**	24 **(39)**	141 **(216)**	64 **(69)**
Others	551,957	608,085	651,479	931,356
Total	552,075	608,158	653,090	933,604

Note: The data are in normal text and the numbers of transcripts per million (TPM) are in **boldface**

As a first question, say that you are interested in knowing if the transcription rate of the gene 17 was the same in the two replicates of the libraries of Sprorangiopohores, S.R1 and S.R2. We can obtain the relevant data in R

```
> phyco # This matrix contain the data

        S.R1   S.R2   M.R1   M.R2
17        22      6     30     18
1661      56     43   1440   2163
7433      40     24    141     64
Other 551957 608085 651479 931359

> apply(phyco[, 1:2], 2, sum) # Number of tags per library

  S.R1   S.R2
552075 608158
```

```
> # Now, the estimated probabilities of expression
> phyco17.S <- phyco[1, 1:2]/apply(phyco[, 1:2], 2, sum)
> phyco17.S # 22/552075 and 6/608158
       S.R1         S.R2
3.984966e-05 9.865857e-06
> phyco17.S[1]/phyco17.S[2] # Ratio of the probabilities =  (22/552075)/(6/60815
    S.R1
4.039148
```

We see that the rate of expression of the gene in the two replicates is around 4; that is, apparently there was "much more" expression in replicate 1 than in replicate 2. The question is: that difference is "significant"? To evaluate this difference we need to compare it with the natural variation that the phenomenon presents, that is with the "error" or noise that is part of the phenomenon.

Formally, we need to write down our hypotheses and to find a rule to test them. Let't assume that we know only the result from replicate 1 (22 tags in a total of 552075 sequenced), and we can assume a Binomial distribution for X, thus only the parameter p needs to be specified. Then our hypotheses are:

- $H_0 : p = 22/552075$
- $H_a : p \neq 22/552075$

Now we need to find a rule (statistical test) to know when to reject H_0. In particular we will want a rule with small errors, say we want small values of

- $\alpha = P[\text{reject } H_0 \text{ when } H_0 \text{ is true}]$
- $\beta = P[\text{accept } H_0 \text{ when } H_0 \text{ is false}]$

It is rational to think that we will reject the hypothesis H_0 if the number of tags is too small or too large to consider that the probability is as stated by the hypothesis. Thus we can design a test of the kind

> Reject H_0 if $X \leq L$ or $X \geq U$ where L and U are some constants (natural numbers) such that $L < U$ and
>
> $$P[X \in [L, U]|H_0] = P[(L \leq X) \cup (X \geq U)|H_0] = \alpha$$

We have a difficulty; because we are dealing with a **discrete** distribution, then it is possible that we cannot obtain the exact α value. We van ask then that

$$P[X \in [L, U]|H_0] = P[(L \leq X) \cup (X \geq U)|H_0] \leq \alpha$$

Now, because we have that $(L \leq X) \cup (X \geq U)$ are disjoint events, we can also put

$$P[(L \leq X) \cup (X \geq U)|H_0] = P[(L \leq X)|H_0] + P[(X \geq U)|H_0] \leq \alpha$$

and thus it is sensible to ask that

$$P[(L \leq X)|H_0] \leq \alpha/2$$

and

$$P[(X \geq U)|H_0] \leq \alpha/2$$

Let's calculate some values for our limits in R.

```
> # Understanding the functions for Binomial
> # The accumulative probability
> pbinom(c(0:5), size=5, prob=.5)
[1] 0.03125 0.18750 0.50000 0.81250 0.96875 1.00000
> # The quantile function
> qbinom(c(0:5)/5, size=5, prob=.5)
[1] 0 2 2 3 3 5
> # Now let's do the calculations for p = 22/552075
> # with size=608158 (of the second library S.R2)
> my.limits <- data.frame(c(1:10), c(608157:608148),
+ rep(NA, 10), rep(NA, 10), rep(NA, 10))
> names(my.limits) <- c("L", "U", "aL", "aU", "a")
> for(i in 1:10){
+ my.limits$aL[i] <- pbinom(my.limits$L[i], 608158, prob=22/552075)
+ my.limits$aU[i] <- pbinom(my.limits$U[i], 608158, prob=22/552075, lower.tail=F)
+ my.limits$a[i] <- my.limits$aL[i]+my.limits$aU[i]
+ }
> my.limits
    L      U           aL aU            a
1   1 608157 7.528874e-10  0 7.528874e-10
2   2 608156 9.514780e-09  0 9.514780e-09
3   3 608155 8.029853e-08  0 8.029853e-08
4   4 608154 5.091726e-07  0 5.091726e-07
5   5 608153 2.587985e-06  0 2.587985e-06
6   6 608152 1.098488e-05  0 1.098488e-05
7   7 608151 4.005687e-05  0 4.005687e-05
8   8 608150 1.281289e-04  0 1.281289e-04
9   9 608149 3.652926e-04  0 3.652926e-04
10 10 608148 9.400706e-04  0 9.400706e-04
```

As we see from the previous calculation, the probability of a "large" number of tags (in the right hand side of the distribution) are so small that can be safely ignored. Thus we can say in this case that

$$P[(X \geq U)|H_0] \approx 0$$

and thus work only with the lower limit asking just that

$$P[(L \leq X)|H_0] \leq \alpha$$

Using our previous calculations we can give rules to reject H_0 and at the same time state the associated probability of Error Type one, α. This rules have the form:

"Reject H_0 if $X \leq L$".

And the associated α probability is given by the table that we obtained in R. For example if we select

"Reject H_0 if $X \leq 6$"

we get an associated error $\alpha = 1.098488e-05 \approx 0.00001$ which is very small indeed. How large needs L to be to obtain $\alpha \approx 0.01$? That is easily computed in R:

```
> qbinom(0.01, 608158, prob=22/552075)
[1] 14
> pbinom(14, 608158, prob=22/552075)
[1] 0.01778581
```

Thus if we set the rule

"Reject H_0 if $X \leq 14$"

we get $\alpha = 0.01778581$

Note that **WE** need to take the (informed) decision about H_0, Statistics only give us the information to do so. In this case we see that there are enough evidence to think that the rate of transcription of the gene changed in the two replicates. Because these libraries were replicates, the change must be caused for an unknown factor that influenced the gene or, possibly, by the own "random" variation of the gene. That is variation "unexplained" by factors in the design of the experiment or "error" variation.

And, what about

$\beta = P[\text{accept } H_0 \text{ when } H_0 \text{ is false}]$?

Here we stated $H_a : p \neq 22/552075$, thus there are *infinite* number of values that fulfill this statment. Lets make a definition.

11.5. **Definition. Power Function.** For a given statistical test, we define its *Power Function* as

$1 - \beta = P[\text{reject } H_0 \text{ when } H_0 \text{ is false}]$

and in our case, given that the test for H_0 is "Reject H_0 if $X \leq L$" we have that

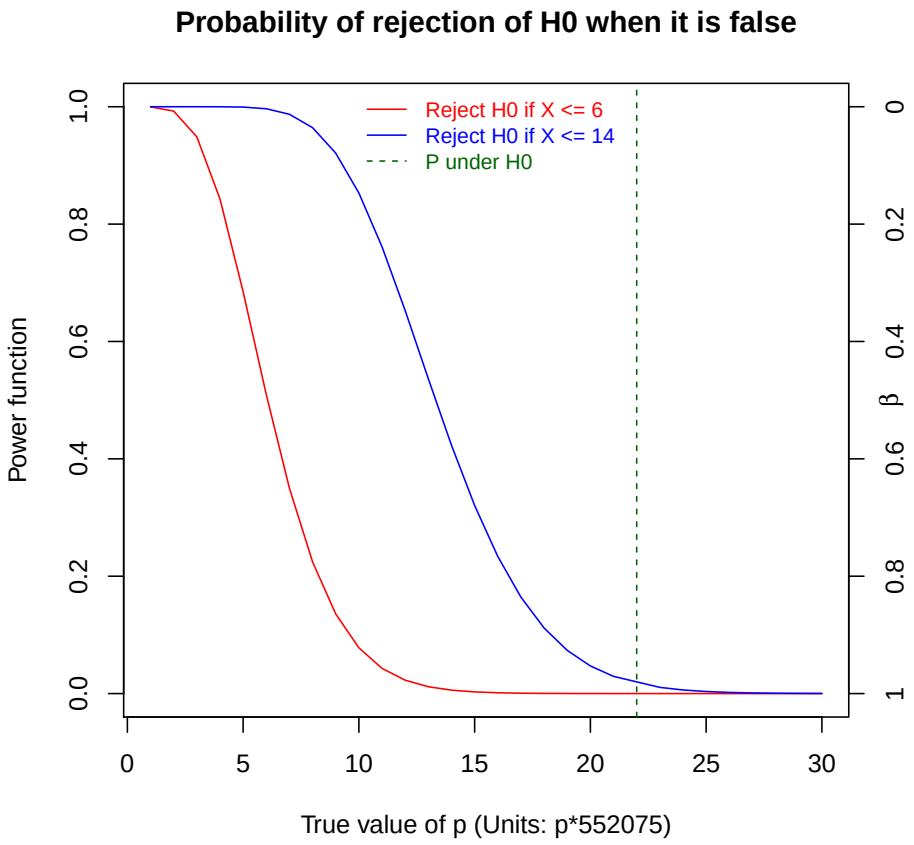
$1 - \beta = P[\text{reject } H_0 \text{ when } H_0 \text{ is false}] = P[X \leq L|H_a] = P[X \leq L|p \neq 22/552075]$

In this case this is a function of p and thus we can evaluate

$$Pow(p) = P[X \leq L | p \neq 22/552075]$$

for our putative tests (say $L = 6$ with $\alpha \approx 0.00001$ and $L = 14$ with $\alpha \approx 0.01$) for a range of p close to $22/552075$.

```
> Pow6 <- pbinom(6, 608158, prob=c(c(1:21)/552075, c(23:30)/552075))
> Pow14 <- pbinom(14, 608158, prob=c(c(1:21)/552075, c(23:30)/552075))
> plot(c(c(1:21), c(23:30)), Pow6, xlab="True value of p (Units: p*552075)",
+ ylab="Power function", type="l", col="red",
+ main="Probability of rejection of H0 when it is false")
> points(c(c(1:21), c(23:30)), Pow14, type="l", col="blue")
> abline(v=22, lty=2, col="darkgreen")
> axis(4, at=c(0:5)*2/10,labels=c(5:0)*2/10)
> mtext(expression(beta), side=4, line=1)
> legend("top", legend=c("Reject H0 if X <= 6",
+ "Reject H0 if X <= 14", "P under H0"),
+ text.col=c("red", "blue", "darkgreen"),
+ bty="n", lty=c(1,1,2),
+ col=c("red", "blue", "darkgreen"), cex=0.85)
```



Note that the power $(1-\beta)$ is larger near the probability fixed by H_0 for the test that accepts a larger probability of Type I error, i.e., $L = 14, \alpha \approx 0.01$, than for the test that accept only a very small probability of Type I error $(L = 14, \alpha \approx 0.00001)$. The fact is that when we reduce one of the errors, the other is increased, thus the researcher must decide how to balance them.

11.6. **Contingency Tables and Hypothesis Testing.** A *Contingency Table* is a matrix (table) presenting the number of cases for two criteria of classification, say **rows** and **columns**. We are going to introduce this in the context of *Transcriptomic*. The data from RNA-Seq experiments consist of a matrix containing the counts of the number of gene tags, where the rows are the genes and the columns the distinct libraries sequenced. The libraries correspond to distinct treatments or replicates, depending on the design of the experiment. In statistical terms, these data represent a Contingency Table from which we can estimate the relative frequency of expression of the genes as well as the dependence of such frequencies on the treatments.

To exemplify, consider the simplest case, where we have two treatments, say T_1 and T_2 and we observe the number of tags for gene A. All the tags that do not correspond to gene A are accumulated in the *Not-A* category. This results in a two by two contingency table shown in the following Table, where X_{ij} represents a random variable of the counts in the *i-th* row and *j-th* column and the total number in each row is represented by $X_{i.}$, while the total for the column is $X_{.j}$ and the full total is represented by $X_{..}$.

Gene	T_1	T_2	Total
A	X_{11}	X_{12}	$X_{1.}$
$Not - A$	X_{21}	X_{22}	$X_{2.}$
Total	$X_{.1}$	$X_{.2}$	$X_{..}$

This table has $(2-1)(2-1) = 1$ degree of freedom (df). In general for a table with r rows and c columns we have $df = (r-1)(c-1)$ (see problem 2 in Homework).

To work on a specific numeric example with real data, consider the following table for the expression of gene 17. By adding rows and columns from our example of *Phycomyces* (see p. 5) I obtain the following 2 (rows) by 2 (columns) table:

Gene	S	M
G17	28	48
Not17	1160205	1586646

This table give us the *frequency* of occurrence of events in a composite sample space. Assume that we are looking, one by one, to the tags sequenced from the libraries S (Sporangiophores) and M (Mycelia), and the tags can be from the gene 17 ($G17$) or from any other gene ($Not17$), that is our sampling space is

$$\Omega = \{(S, G17), (S, Not17), (M, G17), (M, Not17)\}$$

Let's see the table with its totals included

Gene	S	M	Total
G17	28	48	76
Not17	1160205	1586646	2746851
Total	1160233	1586694	2746927

Thus, we realized and impressive number $(2,746,927)$ of realization of the experiment "sample one tag". Here, what is "random" and what is "fixed" by the researcher? The column's criteria (S or M) was fixed by the researcher, however the total number of tags in each library was unknown at the begin of the experiment, and thus can be considered random. Also the other numbers can be considered "random". Which probability functions is reasonable to propose for this phenomenon?.

First let's calculate the relative frequencies for each one of the cells

```
> phyco17 <- matrix(c(28,1160205,48,1586646),
+ nrow=2, ncol=2, dimnames=list(
+ c("G17", "Not17"), c("S", "M"))) # The data
> prob.obs <- phyco17/sum(phyco17)
> prob.obs
                  S            M
G17   1.019321e-05 1.747407e-05
Not17 4.223647e-01 5.776076e-01
```

Thus for each one of the points in Ω we have an estimated value of probability, say approximately $P[(S, G17)] = 1.019e-05$, $P[(S, Not17)] = 4.224e-01$, $P[(M, G17)] = 1.747e-05$ and $P[(M, Not17)] = 5.776e-01$

The important question here is: Are the criteria of classification (treatments and genes) independent?

Let's *assume*, as an **hypothesis** (H_0), that the criteria are independent. Then, by the definition of independence, the following equality must be fulfilled

$$P[(S, G17)] = P[S \cap G17] = P[S]P[G17]$$

If that is fulfilled, also the probabilities for the other three cells or composite events must be, that is

$$P[(S, Not17)] = P[S \cap Not17] = P[S]P[Not17]$$
$$P[(M, G17)] = P[M \cap G17] = P[M]P[G17]$$
$$P[(M, Not17)] = P[M \cap Not17] = P[M]P[Not17]$$

Note that the estimated probabilities $P[S]$, $P[M]$, $P[G17]$ and $P[Not17]$ can be obtained from the totals of the rows and column, that is we can calculate in R

```
> apply(prob.obs, 1, sum) # P[G17] and P[Not17]
         G17        Not17
2.766728e-05 9.999723e-01
> apply(prob.obs, 2, sum) # P[S] and P[M]
        S         M
0.4223749 0.5776251
```

Now we can obtain the probabilities under the hypothesis of independence (H_0). In R this can be done with

```
> prob.H0 <- apply(prob.obs, 1, sum)%*%t(apply(prob.obs, 2, sum))
> attributes(prob.H0) <- attributes(prob.obs)
> prob.H0
```

```
                S            M
G17   1.168596e-05 1.598132e-05
Not17 4.223632e-01 5.776091e-01
```

(please check this by hand to make sure you understand)

Now, if the null hypothesis is true, then we **expect** that the numbers of each one of the cells will be the probabilities obtained under H_0 (obtained only assuming independence), say

```
> expect17 <- prob.H0*sum(phyco17) # Expected values
> expect17
                S            M
G17   3.210049e+01 4.389951e+01
Not17 1.160201e+06 1.586650e+06
> round(expect17) # Expected values (rounded)
            S       M
G17        32      44
Not17 1160201 1586650
> # This can be obtained directly with a function that I programed:
> round(expected(phyco17)) # See problem 3 of homework
            S       M
G17        32      44
Not17 1160201 1586650
```

Thus, now we have two tables: The data Observed and the one Expected under H_0. We must find a statistical test to prove if H_0 is reasonable or likely. Lets have a look at the tables again:

```
> phyco17 # Observed
            S       M
G17        28      48
Not17 1160205 1586646
> round(expect17) # Expected
            S       M
G17        32      44
Not17 1160201 1586650
> round(phyco17 - expect17) # Differences (approximated)
       S  M
G17   -4  4
Not17  4 -4
```

The differences seem quite small (around 4 in each cell) thus it appears that they could be given just by random effects and not by the treatments (different origin of the libraries), but in any case we must find a way to statistically test them. I will give you a general recipe for designing statistical tests.

11.7. **Algorithm? for Designing (classical) Statistical Tests.** In almost all statistical tests, we are interested in one (or more) parameters of a distribution, and the hypothesis of interest (H_0) can be written down by restricting one (ore more) parameters to have specific values. In general lines the procedure to obtain a test is as follows:

- 1. Write down the probability (or density) of the random variable(s) involved.
- 2. Write the hypothesis of interest (H_0), restricting the value of the parameter to be tested.
- 3. Find a statistic (function of the observations) that under H_0 has a **KNOWN** distribution. That is, a statistic that if H_0 is true, then do not have unknown quantities in its distribution. This statistic is call the "test statistic". This of course is the tricky bit.
- 4. Fix a probability of "Error Type I" that you can accept, say α.
- 5. Evaluate statistic with your data. Also find the probability to obtain a value as the one you obtained with your data or "more extreme" (unlikely) given that H_0 is true. You can do that because the distribution is known. If that probability is smaller than α, then reject H_0, otherwise accept it.

We are going to need very soon a distribution of a continuous variable that is called the "Chi-squared" distribution. Let's define it.

11.8. **Chi-squared distribution.** Let $\{Z_1, Z_2, \ldots, Z_k\}$ be a set of independent random variables with Standard Normal Distribution, that is

$$Z_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

Define

$$Q = \sum_{i=1}^{k} Z_i^2$$

Q will have a Chi-squared distribution with k degrees of freedom, we write

$$Q \sim \chi^2(k)$$

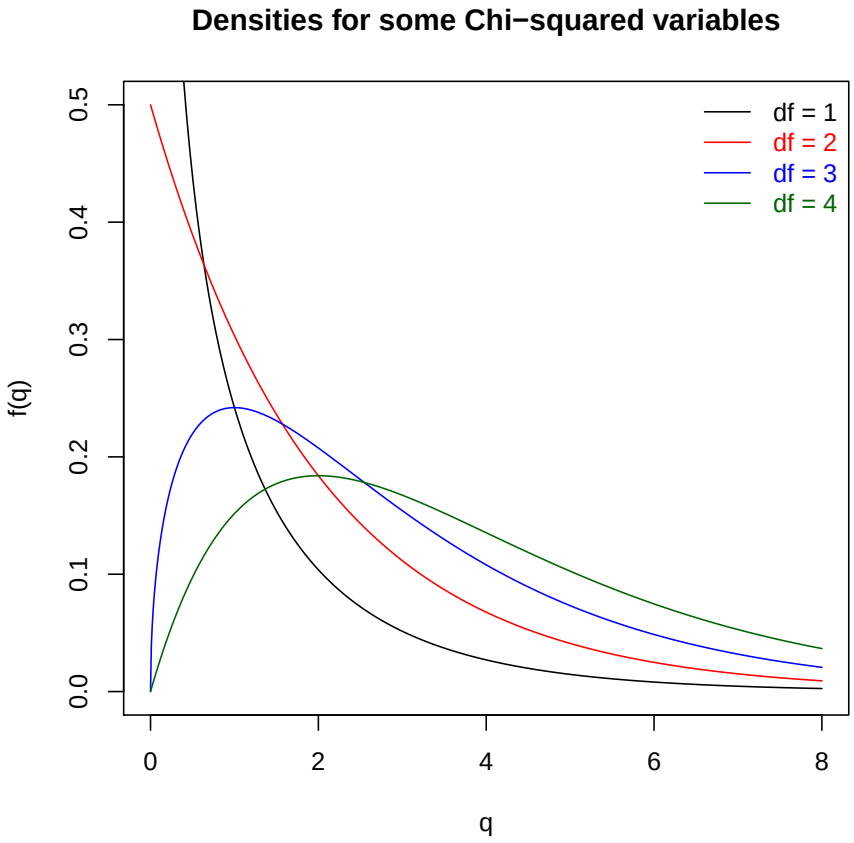
The density of Q is given by

$$f_Q(q) = \frac{1}{2^{k/2}\Gamma(k/2)} q^{\frac{k}{2}-1} e^{-\frac{q}{2}}$$

This is defined for $q \geq 0$ (see $http://en.wikipedia.org/wiki/Chi-squared\_distribution$ for details).

The nice thing about the χ^2 distribution is that because many things are Normally distributed, and the mean and variances of these can be easily estimated, many statistics that are sum of squares of (standard) normals can follow (approximately or exactly) the χ^2 distribution. And it depends only in one parameter: $k =$ the number of degrees of freedom. Let's see how this distribution looks using R (ask for help typing "? pchisq").

```
> plot(c(0:800)/100, dchisq(c(0:800)/100, df=1),
+ type="l", xlab="q", ylab="f(q)", ylim=c(0, 0.5),
+ main="Densities for some Chi-squared variables")
> points(c(0:800)/100, dchisq(c(0:800)/100, df=2),
+ type="l", col="red")
> points(c(0:800)/100, dchisq(c(0:800)/100, df=3),
+ type="l", col="blue")
> points(c(0:800)/100, dchisq(c(0:800)/100, df=4),
+ type="l", col="darkgreen")
> legend("topright", bty="n",
+ legend=paste("df =", c(1:4)),
+ text.col=c("black", "red", "blue", "darkgreen"),
+ col=c("black", "red", "blue", "darkgreen"),
+ lty=c(1,1,1,1))
```



Now let's define a very general test that is based in the Maximum Likelihood Estimation.

11.9. **Maximum Likelihood Ratio Test.** Let $L(\theta|\{x_i\}, H_0)$ be the maximum of the likelihood function of the data, but with parameters restricted by the null hypothesis H_0 and $L(\theta|\{x_i\})$ the maximum of the unrestricted likelihood function of the data. Then the statistic

$$LRT = -2log(\frac{L(\theta|\{x_i\}, H_0)}{L(\theta|\{x_i\})})$$

is called the "Likelihood Ratio Statistics" and it has an approximate $\chi^2(k)$ distribution, where k are the degrees of freedom of the data under the null hypothesis.

In the case of the 2 by 2 contingency tables we have (following the "dot" notation) that the unrestricted maximum likelihood of the data is given by

$$L(\theta|\{x_{ij}\}) = \prod_{ij} \hat{p}_{ij}^{x_{ij}}$$

for $i = 1, 2$, $j = 1, 2$ where ij denote the cells of the table (i=rows, j=columns), and $\hat{p}_{ij}$ are the Maximum Likelihood Estimates of the probabilities, that is

$$\hat{p}_{ij} = \frac{x_{ij}}{x_{..}}$$

Remembering that our hypothesis H_0 restrict our parameters p_{ij} to be the product of the marginal probabilities, i.e., to be independent. Based on that we set:

$$p_{ij} = p_i p_j$$

(note that only one restriction is needed for that) and also noting that the MLE are in that case

$$\hat{p}_i = \frac{x_{i.}}{x_{..}}$$

and

$$\hat{p}_j = \frac{x_{j.}}{x_{..}}$$

we can write the restricted likelihood as

$$L(\theta|\{x_{ij}\}, H_0) = \prod_{ij} \hat{p}_i \hat{p}_j^{x_{ij}}$$

On the other hand we can rewrite

$$LRT = -2log(\frac{L(\theta|\{x_i\}, H_0)}{L(\theta|\{x_i\})}) = 2(l(\theta|\{x_i\}) - l(\theta|\{x_i\}, H_0))$$

where $l(\theta|\{x_i\})$ and $l(\theta|\{x_i\}, H_0)$ are the log likelihood functions of the data. That is, in our case

$$l(\theta|\{x_i\}) = \log(\prod_{ij} \hat{p}_{ij}^{x_{ij}}) = \sum_{ij} x_{ij} \log(\hat{p}_{ij})$$

and

$$l(\theta|\{x_i\}, H_0) = \log(\prod_{ij} \hat{p}_i \hat{p}_j^{x_{ij}}) = \sum_{ij} x_{ij} \log(\hat{p}_i \hat{p}_j)$$

Thus the LRT can be re-written in this case as

$$LRT = 2((\sum_{ij} x_{ij} \log(\hat{p}_{ij})) - (\sum_{ij} x_{ij} \log(\hat{p}_i \hat{p}_j)))$$

or equivalently as

$$LRT = 2\sum_{ij} x_{ij}(\log(\hat{p}_{ij}) - \log(\hat{p}_i\hat{p}_j)) = 2\sum_{ij} x_{ij}\log(\frac{\hat{p}_{ij}}{\hat{p}_i\hat{p}_j})$$

It is easy to see (substituting the values $\hat{p}_{ij}$, $\hat{p}_i$ and $\hat{p}_i$ by their values as functions of the x in the original table, that the LRT statistics can be written as

$$LRT = 2\sum_i O_i \log\frac{O_i}{E_i}$$

where O_i are the observed values (x_ij), E_i are the corresponding expected values and the sum is performed over all the elements (cells) of the table. This statistics is called G by Sokal and Rholf, and that nomenclature is the one that I am going to follow here. Thus we will use:

$$G = 2\sum_i O_i \log\frac{O_i}{E_i}$$

as statistic to test independence in contingency tables. This formula is completely general, that is, it is valid for contingency tables of any size r x c.

> Note that the formula for the G test make sense: If the values O_i and E_i are close then its ratio O_i/E_i will be close to 1, and $\log(1) = 0$, on the contrary if O_i/E_i is far from 1, then $\log(O_i/E_i)$ will be "large". The values of $\log(O_i/E_i)$ are weighted by O_i. Other nice property of G, that we will see later, is that it is **additive**.

By evaluating the statistic G with the data of the example of gene 17 (above) we obtain: $G = 0.9192799$. The probability of having a value as this or larger is calculated, by using the $\chi^2(1)$ (Chi-squared with df=1) as:

$$P[G \geq 0.92|H_0] \approx 0.38$$

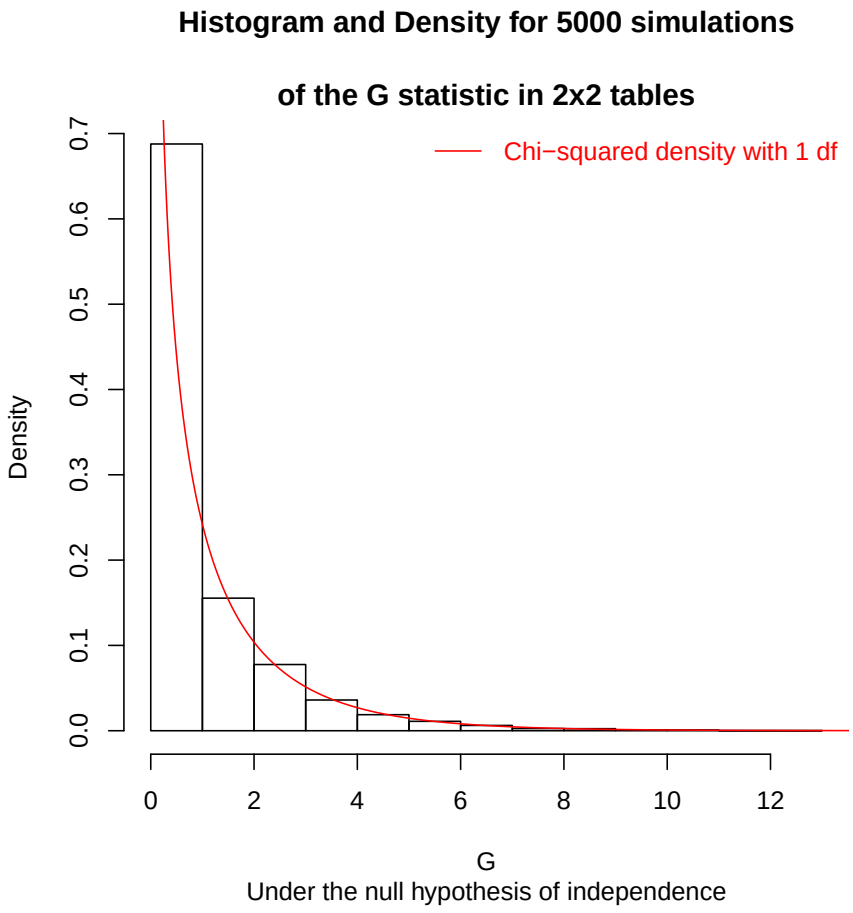
Thus we DO NOT reject the null hypothesis H_0; there is not enough evidence against the independence of the criteria of classification (treatments vs. gene 17). That implies that the rate of transcription of gene 17 is the same for "S" and "M". In contrast, if the value of G is large, that means that it is "unlikely" that the null hypothesis of independence is true, and then we will reject it.

11.10. **Distribution of G under H_0 obtained by simulation in R.** Let generate contingency tables under the null hypothesis of independence and estimate the distribution of the test statistic, "G".

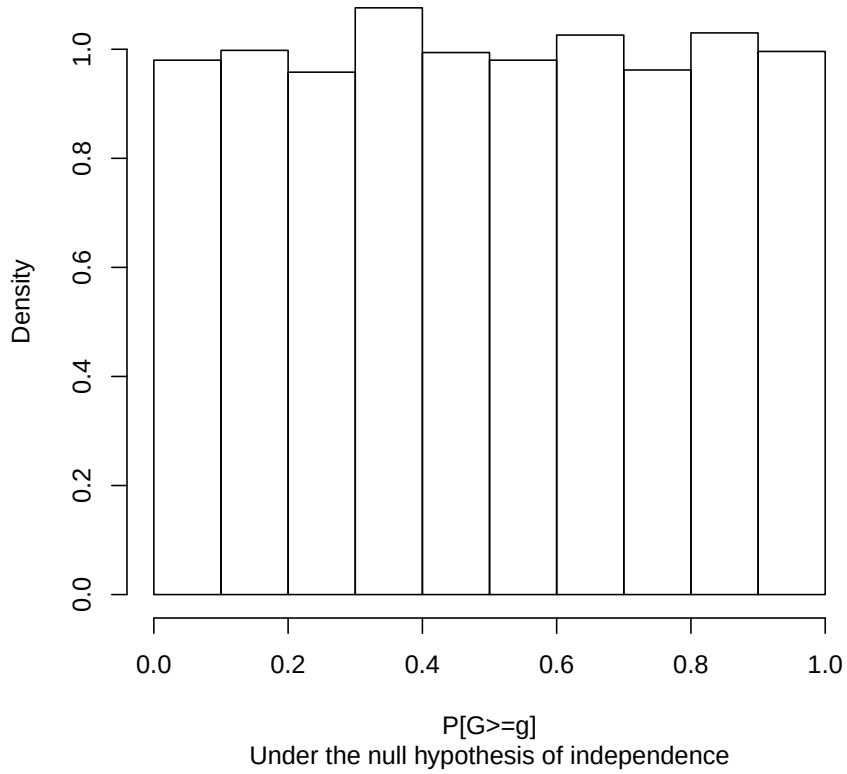
```
> # I programed a function: G.test that perform
> # the G.test As Example see:
> phyco17
```

```
            S       M
G17        28      48
Not17 1160205 1586646
> G.test(phyco17)
        G        df   p.value
0.9192799 1.0000000 0.3376641
> # A function to generate a random 2 by 2 matrix
> genera <- function(lambda=100, n=1000){
+ matrix(c(rpois(2, lambda), rpois(2, n-lambda)),
+ nrow=2, ncol=2, byrow=TRUE)}
> genera() # Example of the output
     [,1] [,2]
[1,]   94  100
[2,]  859  915
> # Let's obtain a substantial number of simulations
> # of the distribution of G (and its probability)
> # when the null hypothesis is true.
> my.sim2 <- data.frame(rep(NA, 5000), rep(NA, 5000))
> for(i in 1:5000){
+ my.sim2[i,] <- G.test(genera())[c(1,3)]
+ }
> names(my.sim2) <- c("G", "p.value")
> # Which proportion of the test are erroneously
> # rejected at alpha=0.05, 0.01 and 0.001?
> length(my.sim2$p.value[my.sim2$p.value <= 0.05])/5000 # Proportion rejected at 0.05
[1] 0.0496
> length(my.sim2$p.value[my.sim2$p.value <= 0.01])/5000 # Proportion rejected at 0.01
[1] 0.0084
> length(my.sim2$p.value[my.sim2$p.value <= 0.001])/5000 # Proportion rejected 0.001
[1] 4e-04
> # Figures
> hist(my.sim2$G, freq=FALSE, xlab="G",
+ main="Histogram and Density for 5000 simulations\n
+ of the G statistic in 2x2 tables",
+ sub="Under the null hypothesis of independence")
> points(c(0:1400)/100, dchisq(c(0:1400)/100,1), type="l",col="red")
> legend("topright", bty="n", lty=1, col="red", text.col="red",
+ legend="Chi-squared density with 1 df")
```

```
> # Now, lets do an histogram of the probabilities
> hist(my.sim2$p.value, freq=FALSE, xlab="P[G>=g]",
+ main="Histogram of probabilities for 5000 simulations\n
+ of the G statistic in 2x2 tables",
+ sub="Under the null hypothesis of independence")
```



How good is the agreement between the observed and expected distribution?



You can see that, when the null hypothesis is true, then the distribution of the probabilities will be approximately uniform. Does it make's sense?

11.11. **Other Statistical Tests for Contingency Tables.** Of course, the $LRT = G$ is not the only test available. There are alternatives, for example the very well known and used Pearson's Chi-square test defined as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

this test is also approximately distributed as $\chi^2(df)$, but the approximation can be poor if the sample size is small. Let see it in R:

```
> # My favorite test:
> G.test(phyco17)
        G        df   p.value
0.9192799 1.0000000 0.3376641
```

```
> # The Pearson's Chi-square test
> chisq.test(phyco17)
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  phyco17
X-squared = 0.6992, df = 1, p-value = 0.4031
> # Let's put it in an object and see its components
> temp <- chisq.test(phyco17)
> names(temp)
[1] "statistic" "parameter" "p.value"   "method"
[5] "data.name" "observed"  "expected"  "residuals"
[9] "stdres"
> # Look at each one of the components, for example
> temp$expected
                  S            M
G17   3.210049e+01 4.389951e+01
Not17 1.160201e+06 1.586650e+06
> # Comparing with our function:
> temp$expected == expected(phyco17)
         S    M
G17   TRUE TRUE
Not17 TRUE TRUE
> # (note you need to program that function, see homework)
```

Other important test is **Fisher's exact test**. It assumes the hypergeometric distribution (thus it assume that the totals are fixed and not random), and then calculate the **exact** probability of obtaining tables "more extreme" that the one analyzed, always assuming the null hypothesis of independence, H_0. Let's try it in R.

```
> fisher.test(phyco17)
        Fisher's Exact Test for Count Data

data:  phyco17
p-value = 0.3556
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4819174 1.2974774
sample estimates:
```

```
odds ratio
 0.7977418
> temp <- fisher.test(phyco17)
> names(temp)
[1] "p.value"     "conf.int"    "estimate"
[4] "null.value"  "alternative" "method"
[7] "data.name"
> temp$conf.int # What is this?
[1] 0.4819174 1.2974774
attr(,"conf.level")
[1] 0.95
```

Compare the probabilities of the three tests (G, Pearson's χ^2 and Fisher's exact test) on the "phyco17" dataset. Which are more alike?

11.11.1. *Homework.*

- 1. A bandit, Narcomuseno Trinquetes, has you captive, and will release you only if you succeed in a Test. The conditions are these: There is a box with two coins, one is normal (fair) and the other had two "heads". You must select a coin from the box and throw the coin as many times as you want, but without looking at the other side of it. You will be released only if you correctly say which coin you selected, otherwise, I am afraid, you will be killed.

  Let X be the random variable defined as the number of throws before "tails" is observed.

  Find the probability function of X.

  Consider:

  H_0 : The fair coin was selected.

  H_a : The tricked coin was selected.

  Propose a statistical test for H_0 as function of a finite value of X and calculate α and β for your test.
- 2. Degrees of freedom in any r x c contingency table (solve it by hand). Consider any example of a contingency table (begin with a 2 x 2 table). Assume that all the totals, of rows and columns, are FIXED and erase the content of the cells. Now in the 2 by 2 table put any number in any cell, the only restriction is that the number must be smaller than the totals of the corresponding row and columns. See how in this case (2 by 2) all the other 3 numbers are determined by the totals, from there the term "degrees of freedom". Try it with larger tables until you are convinced that the degrees of freedom are (r-1)(c-1).

- 3. Program the function "expected" to obtain the expected values in any contingency table. Hint: You do not need to trouble with the calculation of all probabilities (as we did in the presentation). In is easy to see that the expected value of each cell is the product of the totals of the corresponding row and column divided by the full total. In the "dot" notation:
$$E[X_{ij}] = X_{i.}X_{.j}/X_{..}$$
.
- 4. Check "by hand" (of course you can use R as electronic calculator) the calculations to obtain the value of G from the table of data for gene 17.
- 5. Program a function that takes as input a "contingency table" and as output three values (in a vector): "G", "df", "p.value".
- 6. I programed the function "genera" assuming a Poisson distribution, where the cells of the first row are distributed $\mathcal{P}(\lambda)$ and the TOTAL of each column are distributed $\mathcal{P}(n - \lambda)$. This means that the total number of sampled tags in the experiment is a random variable. Program a function wich assumes the binomial distribution with a fixed total for the table, for example a function in which by default the first row has $\mathcal{B}(k = 1000, p = 0.1)$ and the second row is calculated as $k - x$ (the tags which belongs to "other genes"). Test, by simulation, if the values of G obtained from tables using your function also follow the $\chi^2(1)$ distribution.
- 7. Modify the "genera" function to generate tables in which the null hypothesis is false. In particular, assume that the two columns are treatments and then simulate with distinct data of genes with distinct rate of transcription. Which is the rate of rejection in a simulation? Exactly on which quantities it depends? Hint: Take into account the difference in transcription rate and also the sample size.

Have fun!