preamble

Biostatistics: an introduction

Ana M. Pires

Departamento de Matemática – IST
apires@math.ist.utl.pt

21 and 23 May 2013

Outline

1. What is Biostatistics?
2. From Mendel and Fisher to Statistical Genomics
3. A closer look at Mendel's data (using tests of hypotheses and understanding the *p-value*)
4. Statistical tests: further examples
5. Multiple tests and multitudes of tests
6. Some important (bio)statistical models
7. Some ideas of multivariate analysis
8. Conclusions/recommendations

1. What is Biostatistics?

biostatistics = biology + statistics

is the application of statistics to a wide range of topics in biology.

The science of biostatistics includes:

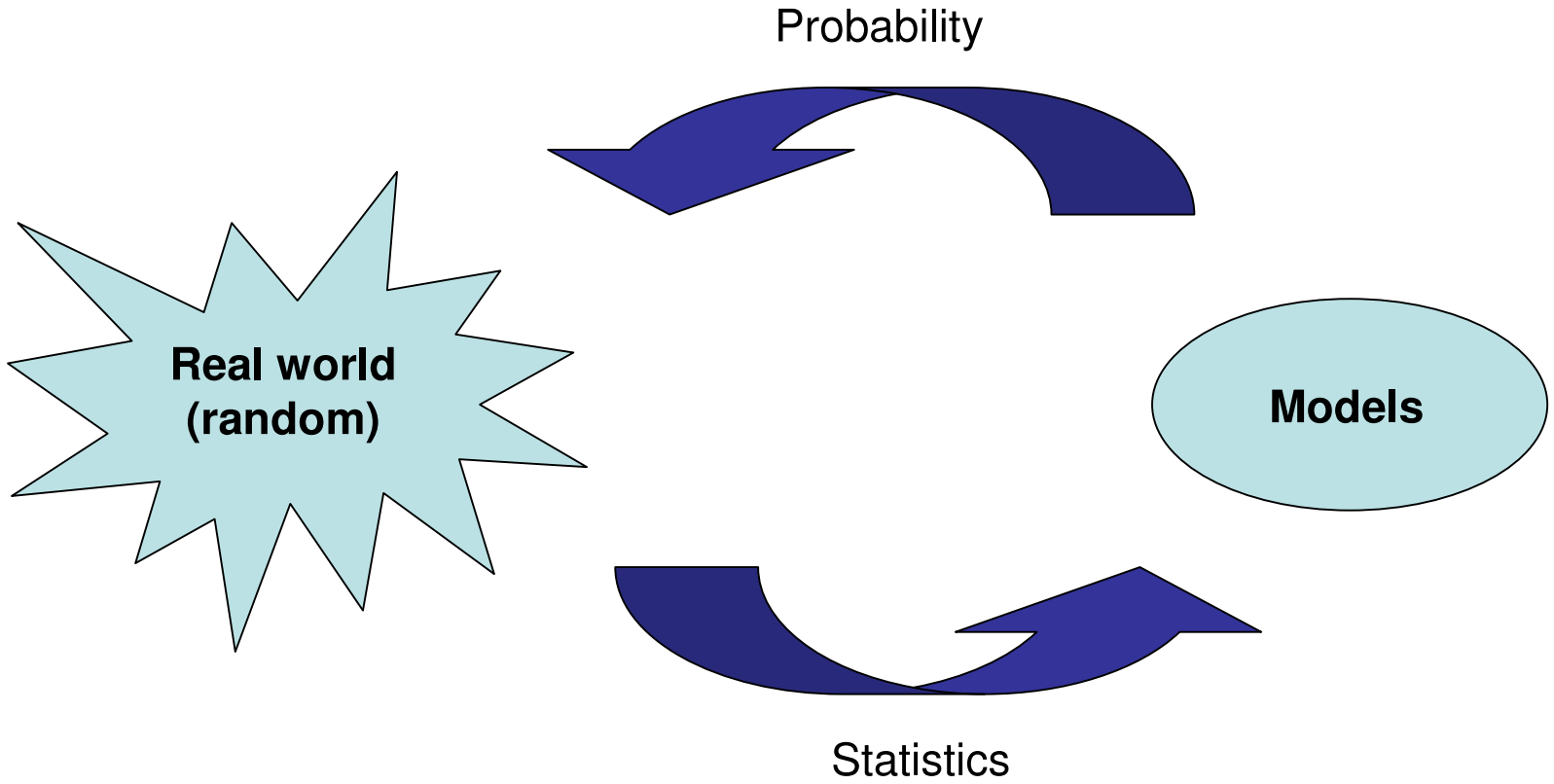
- the design of biological experiments (especially in medicine and agriculture);
- the collection, summarization, and analysis of data from those experiments;
- and the interpretation of, and inference from, the results.

Applications of biostatistics

- Public health, including epidemiology, health services research, nutrition, and environmental health.
- Design and analysis of clinical trials in medicine.
- Genomics, population genetics, and statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals (animal breeding). In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to human diseases.
- Ecology, ecological forecasting.
- Biological sequence analysis.

Statistical methods are also a fundamental component of bioinformatics.

1. What is Biostatistics?



Statistics:

- Parameter estimation
- Confidence regions
- Hypotheses tests

> *all models are wrong,*
> *some models are useful.*
> G.E.P. Box

Statistics and Genetics are linked since the beginning of both sciences. Two examples:

1) Gregor Mendel (1822–1884)

- The founder of modern genetics
- Produced lots of statistical data
- Based on those data formulated the two laws of heredity
- Mendel did not use formal statistical methods (tests) because they were not yet invented
- . . . Mendel was a man far ahead of his time.

2) Sir Ronald Fisher (1890–1962)

- Made fundamental contributions to statistics (described as "a genius who almost single-handedly created the foundations for modern statistical science")
- Is considered the founder of quantitative genetics (in a recent book Richard Dawkins calls him "the greatest of Darwin's successors")

2. From Mendel and Fisher to Statistical Genomics

Some of the most important books of Fisher:

- Statistical Methods for Research Workers (1925)
- The Genetical Theory of Natural Selection (1930)
- The Design of Experiments (1935)
- Statistical tables for biological, agricultural and medical research (1938, with Frank Yates)
- The Theory of Inbreeding (1949)
- Contributions to Mathematical Statistics (1950)
- Statistical Methods and Statistical Inference (1956)

Fisher's famous quotations:

> "Any finite body of data contains only a limited amount of information, which cannot be increased by any amount of ingenuity expended by statisticians."

...the message to biologists is clear: if you want to work with microarrays, you need to find yourself one of these precious experts (a statistician) – and don't wait until after you've collected your data. The following advice, from pioneering British geneticist and statistician Ronald Fisher, rings even more true today than when he uttered it, back in 1938:[1]

> "To call a statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of."

[1] Tilstone, C. (2003). DNA microarrays: Vital statistics, Nature 424, 610-612

2. From Mendel and Fisher to Statistical Genomics

Statistics and Genetics have evolved closely linked:

- Statistical methods are essential to many discoveries in Genetics
- The complexity of the data and the challenges posed by Genetics induce the "creation" of new statistical methods.

This is especially true in the current genomics era!

3. A closer look at Mendel's data

Chronology:

1856–1863 Mendel performed his experiments during this period. He produced around 29,000 garden pea plants from controlled crosses and registered several of their observable characteristics (phenotype), such as, shape and color of the seeds, height, flower color, etc.

1865 Mendel presented the results of his experiments in a communication entitled *Experiments on Plant Hybridization*, in two meetings of the Society of Natural History of Brünn.

1866 The paper with the same title was published in the proceedings of that society. The paper had little impact and would have been cited only three times in the next 35 years.

1900 His work was rediscovered independently by Hugo de Vries, Carl Correns and Erich von Tschermak.

3. A closer look at Mendel's data

1902 The first statistical analysis of Mendel's data is published in the first volume of Biometrika (Weldon, W.R.F., 1902. *Mendel's law of alternative inheritance in peas.* Biometrika 1: 228–254), using the recently invented chi-square test (Pearson, 1900).

1911 Fisher produced a first comment about Mendel's results, in a communication to the Cambridge University Eugenics Society, while he was still an undergraduate:

> *"It is interesting that Mendel's original results all fall within the limits of probable error; if his experiments were repeated the odds against getting such good results is about 16 to one. It may just have been luck; or it may be that the worthy German abbot, in his ignorance of probable error, unconsciously placed doubtful plants on the side which favoured his hypothesis"*

3. A closer look at Mendel's data

1936 Fisher published the paper *Has Mendel's work been rediscovered?* (Annals of Science 1: 115–137), where he expresses the same concern but this time presenting a detailed analysis, both of Mendel's experiments and data. He also attributes the alleged forgery, not to Mendel himself, but to an unknown assistant:

> *"Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected. This possibility is supported by independent evidence that the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations"*

3. A closer look at Mendel's data

1964 Ironically Fisher's paper also remained overlooked until the centennial of Mendel's article: De Beer, G. (1964). *Mendel, Darwin and Fisher (1865-1965)*. Notes and Records of the Royal Society 19: 192–226.

1964–2007 During this period at least 50 papers have been published about the controversy created by Fisher. Some elucidative titles:

Are Mendel's results really too close?

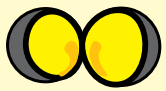
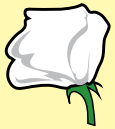
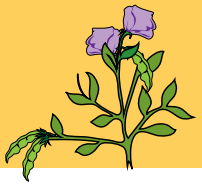
The too-good-to-be-true paradox and Gregor Mendel

Did Mendel cheat?

2008 A group of scientists from different fields published the book *Ending the Mendel-Fisher Controversy* (Franklin, A., Edwards, A.W.F., Fairbanks, D., Hartl, D. and Seidenfeld, T.).

But is it really the end of the controversy?

The seven characters (traits) studied by Mendel:

Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
						
Grey & Round	Yellow	White	Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
						
White & Wrinkled	Green	Violet	Constricted	Green	Terminal pods, Flowers top	Short (¾ -1ft)
1	2	3	4	5	6	7

These traits are all binary, determined by a single gene with a dominant allele. Moreover the seven genes are unlinked.

The model:

F_0		AA	$\times$	aa		
			$\downarrow$			
F_1			Aa	$\times$		Aa
				$\downarrow$		
F_2			AA	Aa		aa
	Phenotype		3		:	1
(F_3)	Genotype	1	:	2	:	1

Note:

F_1	A	a
A	AA	Aa
a	Aa	aa

$\longrightarrow \quad F_2$

(tests of hypotheses and the *p-value*)

Imagine the following very simple experiment:

A coin is tossed 50 times and all the results are "heads" (H).

Can we draw any conclusion from this experiment?

If p denotes the probability of H in one single toss and X is the number of heads in 50 tosses (note that X is a random variable), then

$$P(X = 50) = p^{50}$$

More precisely, X follows a Binomial(n, p) distribution, which implies that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

with $n = 50$.

If $p = 0.5$ we get $P(X = 50) = P(X = 0) = 1/2^{50} = 8.88 \times 10^{-16}$

(the probability of winning the first prize in the Euromillions with a single key is 1.31×10^{-8})

Statistical test to ascertain the validity of the hypothesis "the coin is fair":

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_1 : p \neq \frac{1}{2}$$

Decision: reject H_0 if the number of heads observed in n tosses (x) is much larger than the number expected when H_0 is true ($n/2$). **How do you know how much is "too much"?**

Considerations about errors:

Type I error: reject H_0 when H_0 is true; (or false positive)

Type II error: fail to reject H_0 when H_0 is false; (or false negative)

(tests of hypotheses and the *p-value*)

Level of significance:

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$$

Fixing α at a small value (usually 5%) it is possible to determine the critical points (separation points between the rejection and the acceptance regions).

Another possibility (much better!) is to use the p-value:

> The p-value is the probability of observing, when the null hypothesis is true, a result that is more, or as, unfavorable than the one actually observed.

We may then reject H_0 if the p-value is smaller than a pre-specified threshold, which is precisely the level of significance, α.

In the example

p-value$=P(X = 0|p = 0.5) + P(X = 50|p = 0.5) = 2/2^{50} = 1.78 \times 10^{-15}$

In R:

```
> binom.test(50,50)

        Exact binomial test

data:  50 and 50
number of successes = 50, number of trials = 50, p-value = 1.776e-15
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.9288783 1.0000000
sample estimates:
probability of success 1

> 2/2**50
[1] 1.776357e-15
```

(tests of hypotheses and the *p-value*)

What if instead of 50 "heads" in 50 tosses we had obtained 30 "heads" in 50 tosses?

```
> binom.test(30,50)

        Exact binomial test

data:  30 and 50
number of successes = 30, number of trials = 50, p-value = 0.2026
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4517940 0.7359216
sample estimates:
probability of success 0.6
```

3. A closer look at Mendel's data

In one of his experiments, Mendel obtained after a controlled cross:

in a total of 6887 seeds, 5138 yellow (A) and 1749 green (a).

According to his theory the seeds A and a should appear according to the ratio 3:1

Are these data in accordance with the theory, or not?

The situation is the same as in the coin example. Denote by p the probability of an A seed:

$$H_0 : p = \frac{3}{4} \quad \text{versus} \quad H_1 : p \neq \frac{3}{4}$$

3. A closer look at Mendel's data

Result:

```
> binom.test(5138,6887,p=3/4)

        Exact binomial test

data:  5138 and 6887
number of successes = 5138, number of trials = 6887, p-value = 0.4524
alternative hypothesis: true probability of success is not equal to 0.75
95 percent confidence interval:
 0.7355878 0.7562900
sample estimates:
probability of success 0.7460433
```

3. A closer look at Mendel's data

Edwards (1986) organized the data presented by Mendel (1866) in such a way that they can be analyzed with the test

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0$$

where p is the probability of success and we have a sample of n cases, from which y are successes.

Overall there are 84 independent tests.

3. A closer look at Mendel's data

n	y	p_0
90	43	0.5
43	20	0.5
47	25	0.5
110	57	0.5
57	31	0.5
53	27	0.5
87	44	0.5
44	25	0.5
43	22	0.5
98	49	0.5
49	24	0.5
49	22	0.5
166	87	0.5
87	47	0.5
79	38	0.5

n	y	p_0
565	372	2/3
519	353	2/3
301	198	2/3
102	67	2/3
96	68	2/3
198	138	2/3
103	65	2/3
367	245	2/3
113	76	2/3
122	79	2/3
245	175	2/3
122	78	2/3

3. A closer look at Mendel's data

n	y	p_0
6887	5138	0.75
7545	5667	0.75
57	45	0.75
35	27	0.75
31	24	0.75
29	19	0.75
43	32	0.75
32	26	0.75
112	88	0.75
32	22	0.75
34	28	0.75
32	25	0.75
36	25	0.75
39	32	0.75
19	14	0.75

n	y	p_0
97	70	0.75
37	24	0.75
26	20	0.75
45	32	0.75
53	44	0.75
64	50	0.75
62	44	0.75
929	705	0.75
1181	882	0.75
580	428	0.75
858	651	0.75
1064	787	0.75
556	423	0.75
423	315	0.75
133	101	0.75

n	y	p_0
639	480	0.75
480	367	0.75
159	122	0.75
175	127	0.75
70	52	0.75
78	60	0.75
44	30	0.75
76	60	0.75
37	26	0.75
79	55	0.75
43	33	0.75
37	30	0.75

3. A closer look at Mendel's data

n	y	p_0	p_0^*
100	64	2/3	0.63
100	71	2/3	0.63
100	60	2/3	0.63
100	67	2/3	0.63
100	72	2/3	0.63
100	65	2/3	0.63
127	78	2/3	0.63
52	38	2/3	0.63
60	45	2/3	0.63
30	22	2/3	0.63
60	40	2/3	0.63
26	17	2/3	0.63
55	36	2/3	0.63
33	25	2/3	0.63
30	20	2/3	0.63

$$p_0^* = \left(1 - \left(\frac{3}{4}\right)^{10}\right) \times \frac{2}{3}$$

$\left(\frac{3}{4}\right)^{10}$ = probability of wrongly classifying an Aa plant as AA

3. A closer look at Mendel's data

Computation of the *p-value* of each test:

- Exact method:

$$p\text{-}value = P\left(Y \in \{z : P(Y = z) \leq P(Y = y)\}\right),^{2} \text{ where } Y \sim Bin(n, p_0).$$

- Approximate method (valid for large n):

$$p\text{-}value = 2 \times \min\left\{P\left(Y^* \leq y + \frac{1}{2}\right), P\left(Y^* \geq y - \frac{1}{2}\right)\right\},$$

where $Y^* \sim \mathcal{N}\left(np_0, np_0(1-p_0)\right)$.

- Another approximate method (valid for large n and equivalent to the previous one, without continuity correction):

$$p\text{-}value = P\left(Q \geq \frac{(y - np_0)^2}{np_0(1-p_0)}\right), \quad \text{where} \quad Q \sim \chi^2_1.$$

[2] When $p_0 = 1/2$, $\Leftrightarrow$ $p\text{-}value = 2 \times \min\{P(Y \leq y), P(Y \geq y), 0.5\}$.

3. A closer look at Mendel's data

n	y	p_0	p-$value$
90	43	0.5	0.752
43	20	0.5	0.761
47	25	0.5	0.771
110	57	0.5	0.775
57	31	0.5	0.597
53	27	0.5	1.000
87	44	0.5	1.000
44	25	0.5	0.451
43	22	0.5	1.000
98	49	0.5	1.000
49	24	0.5	1.000
49	22	0.5	0.568
166	87	0.5	0.587
87	47	0.5	0.520
79	38	0.5	0.822

n	y	p_0	p-$value$
565	372	2/3	0.688
519	353	2/3	0.545
301	198	2/3	0.760
102	67	2/3	0.834
96	68	2/3	0.449
198	138	2/3	0.407
103	65	2/3	0.465
367	245	2/3	1.000
113	76	2/3	1.000
122	79	2/3	0.701
245	175	2/3	0.119
122	78	2/3	0.565

3. A closer look at Mendel's data

n	y	p_0	p-val.
6887	5138	0.75	0.452
7545	5667	0.75	0.832
57	45	0.75	0.544
35	27	0.75	1.000
31	24	0.75	1.000
29	19	0.75	0.282
43	32	0.75	1.000
32	26	0.75	0.541
112	88	0.75	0.445
32	22	0.75	0.416
34	28	0.75	0.429
32	25	0.75	0.839
36	25	0.75	0.443
39	32	0.75	0.360
19	14	0.75	1.000

n	y	p_0	p-val.
97	70	0.75	0.557
37	24	0.75	0.182
26	20	0.75	1.000
45	32	0.75	0.605
53	44	0.75	0.206
64	50	0.75	0.666
62	44	0.75	0.464
929	705	0.75	0.545
1181	882	0.75	0.814
580	428	0.75	0.502
858	651	0.75	0.581
1064	787	0.75	0.436
556	423	0.75	0.590
423	315	0.75	0.822
133	101	0.75	0.842

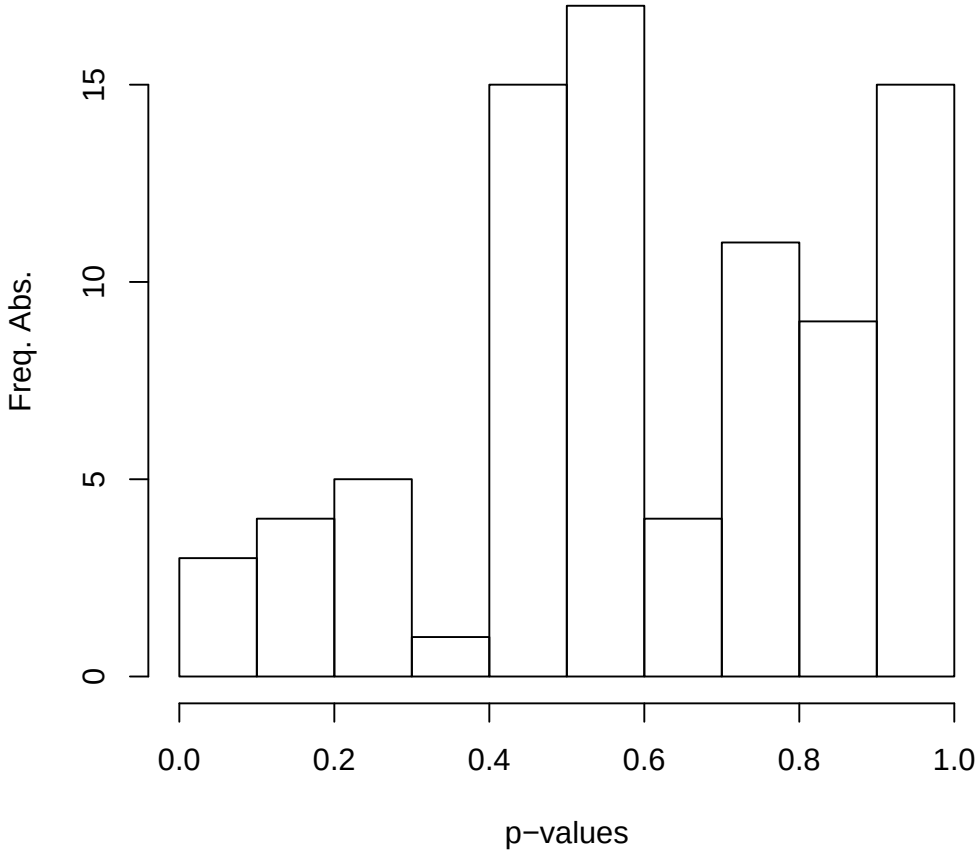
n	y	p_0	p-val.
639	480	0.75	1.000
480	367	0.75	0.493
159	122	0.75	0.648
175	127	0.75	0.485
70	52	0.75	0.890
78	60	0.75	0.794
44	30	0.75	0.298
76	60	0.75	0.508
37	26	0.75	0.568
79	55	0.75	0.298
43	33	0.75	1.000
37	30	0.75	0.453

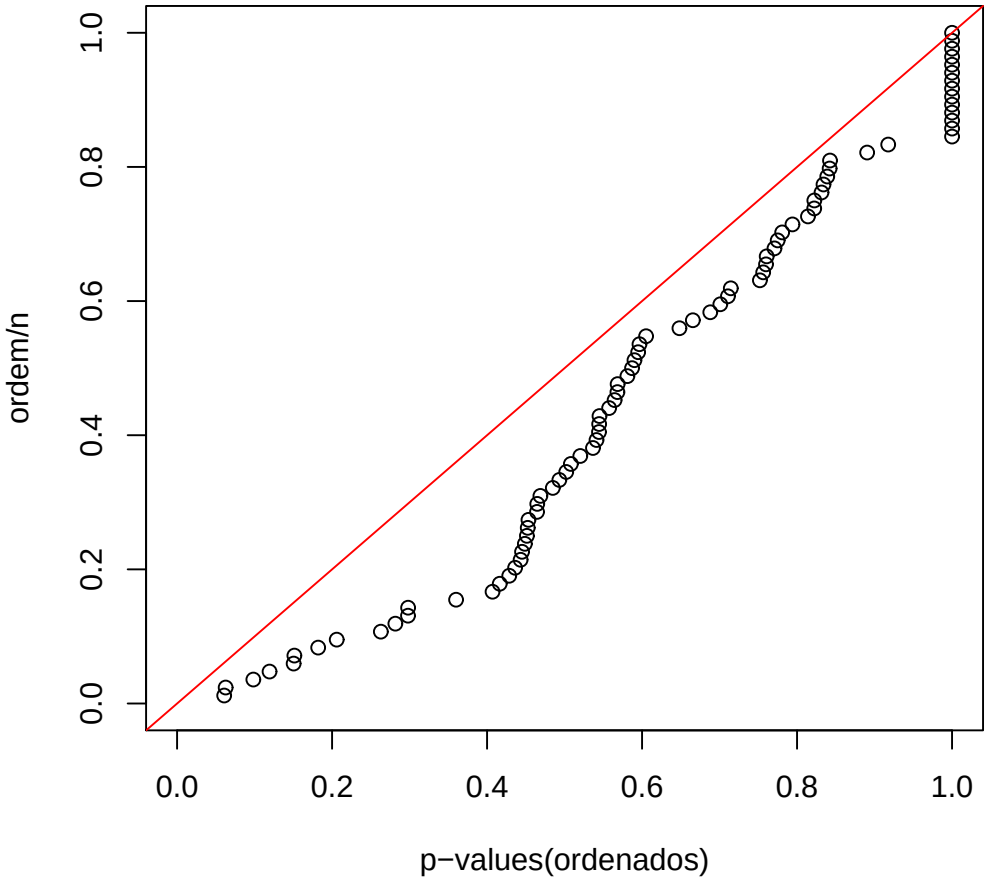
3. A closer look at Mendel's data

n	y	p_0	$p\text{-value}$	p_0^*	$p\text{-value}^*$
100	64	2/3	0.596	0.63	0.918
100	71	2/3	0.397	0.63	**0.098**
100	60	2/3	0.168	0.63	0.537
100	67	2/3	1.000	0.63	0.469
100	72	2/3	0.289	0.63	**0.063**
100	65	2/3	0.751	0.63	0.756
127	78	2/3	0.221	0.63	0.714
52	38	2/3	0.379	0.63	0.151
60	45	2/3	0.217	0.63	**0.061**
30	22	2/3	0.562	0.63	0.263
60	40	2/3	1.000	0.63	0.595
26	17	2/3	1.000	0.63	0.842
55	36	2/3	0.886	0.63	0.781
33	25	2/3	0.356	0.63	0.150
30	20	2/3	1.000	0.63	0.711

3. A closer look at Mendel's data

Analysis of the 84 exact *p-values* (the results are a bit different if the χ^2 p-values are used):





mean = 0.62 **standard deviation = 0.263** **minimum = 0.061**

3. A closer look at Mendel's data

When H_0 is true, the *p-value* (seen as a random variable, function of Y) follows asymptotically a *Uniform* $(0,1)$ distribution.

Proof: As asymptotically the three methods to compute the p-values are all equivalent we can consider the *P-value* given by

$$U = 2\left(1 - \Phi(|Z_0|)\right) \quad \text{com} \quad Z_0 = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}}$$

As, under H_0, $Z_0 \overset{a}{\sim} \mathcal{N}(0,1)$, we have that, for $0 < x < 1$ (and in limit, as $n \to \infty$)

$$\begin{aligned} P(U \leq x|H_0) &= P\left(2\left(1 - \Phi(|Z_0|)\right) \leq x|H_0\right) = P\left(|Z_0| \geq \Phi^{-1}\left(1 - \frac{x}{2}\right)\middle| H_0\right) = \\ &= 2P\left(Z_0 \geq \Phi^{-1}\left(1 - \frac{x}{2}\right)\middle| H_0\right) = 2\left(1 - \left(1 - \frac{x}{2}\right)\right) = x \end{aligned}$$

3. A closer look at Mendel's data

What happens if we test the hypothesis that the *p-values* associated with Mendel's data follow a *Uniform* $(0, 1)$ distribution?

Kolmogorov-Smirnov test:	$D = 0.2523,$	p-value $\simeq 4.537 \times 10^{-5}$
χ^2 goodness-of-fit test:[3]	$X^2 = 36,$	p-value $= 3.965 \times 10^{-5}$

There seems to be a strong evidence that the *p-values* associated with Mendel's data do not follow a *Uniform* $(0, 1)$ distribution.

What does this mean? What has happened?

[3] 10 classes with $E_i = 8.4$

3. A closer look at Mendel's data

Some conclusions:

- It appears that Mendel's results are too close to his theory, than what is expected under random fluctuations (but maybe not as much as Fisher concluded, he mentioned at some point that the chances of obtaining better results were 7 in 100000).
- This is most likely a case of *publication bias*, which means that the results published are a subset of all the results obtained, choosing the ones that better support the theory (there is at least one instance where Mendel mentions that he has repeated the experiment).
- This conclusion does not diminish the extraordinary pioneering work of Mendel.

Chi-square tests

$$H_0\colon\ p_i = p_{i0},\ \forall_{i=1,\ldots,k} \text{ versus } H_1\colon\ \exists_{i=1,\ldots,k}\colon\ p_i \neq p_{i0}$$

Chi-square statistic (Pearson, 1900):

$$\text{under } H_0,\ Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{a}{\sim} \chi^2_{k-\beta-1}$$

where

O_i - observed (absolute) frequency of class i

E_i - expected (absolute) frequency of class i under H_0, $E_i = n\,p_{i0}$ or $E_i = n\,\hat{p}_{i0}$

k - number of classes

β - number of parameters estimated

Example 1: verify the results given in slide 33 for the chi-square test.
Data (observed frequency in the intervals $[0, 0.1[\cdots [0.8, 0.9[, [0.9, 1]$):

3 4 5 1 15 17 4 11 9 15

Example 2: use the chi-square test in lines 1 to 3 of the left table, slide 29

n	y	p_0	*p-val.(1)*	*p-val.(2)*
6887	5138	0.75	0.452	
7545	5667	0.75	0.832	
57	45	0.75	0.544	

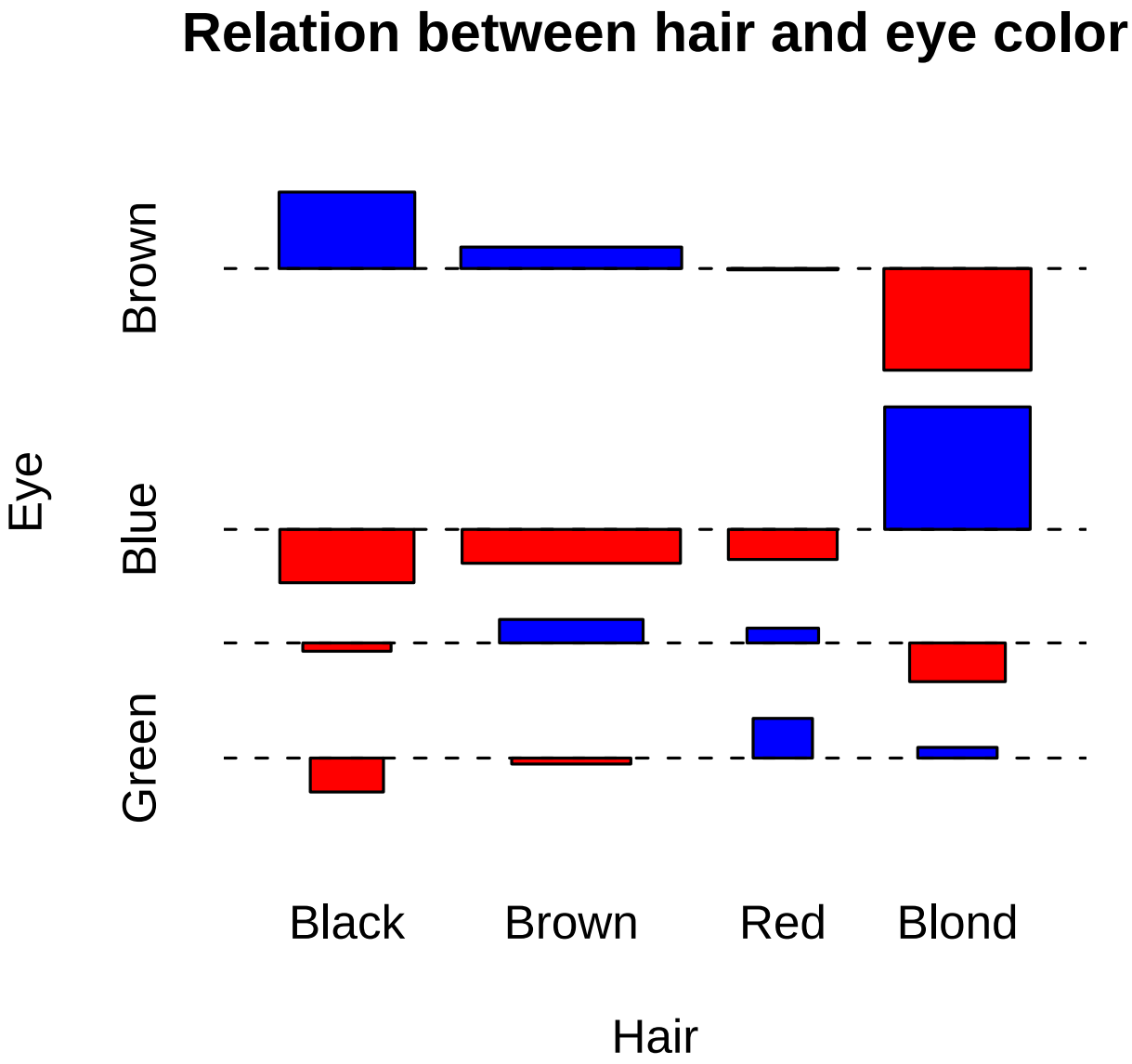
Example 3: Distribution of hair color, eye color and sex in 592 statistics students. (source: HairEyeColor{datasets})

Eye:	Sex=Male				Sex=Female				All			
Hair:	Brown	Blue	Hazel	Green	Brown	Blue	Hazel	Green	Brown	Blue	Hazel	Green
Black	32	11	10	3	36	9	5	2	68	20	15	5
Brown	53	50	25	15	66	34	29	14	119	84	54	29
Red	10	10	7	7	16	7	7	7	26	17	14	14
Blond	3	30	5	8	4	64	5	8	7	94	10	16

Are the characteristics hair and eye color independent (ignore variable sex)?

```
> library(datasets)
> x <- margin.table(HairEyeColor, c(1, 2))
> chisq.test(x)
        Pearson's Chi-squared test
X-squared = 138.2898, df = 9, p-value < 2.2e-16
```

```
residuals:   Hair
Eye      Black Brown   Red Blond
  Brown   4.40  1.23 -0.07 -5.85
  Blue   -3.07 -1.95 -1.73  7.05
  Hazel  -0.48  1.35  0.85 -2.23
  Green  -1.95 -0.35  2.28  0.61
```



4. Statistical tests: further examples

There are thousands of statistical tests...

In the previous slides you have seen examples of chi-square tests, used to test: goodness-of-fit, a single probability, independence (in the practical we use also a chi-square test — for Hardy-Weinberg equilibrium).

In all these situations the data are categorical (classes). There are other tests for the same situations, and also there are tests for testing similar hypothesis with continuous data. And, finally, there are lots and lots of other tests for other situations.

Comparing two groups with respect to a continuous variable

Compare what?

- the means (at the population level, μ_1 and μ_2)
- the medians (idem, m_1 and m_2)
- the true distributions ($P(X_1 \leq x)$ and $P(X_2 \leq x)$)

i. Comparing the means with two independent samples

Assumptions:

Group 1: $X_1 \sim N(\mu_1, \sigma_1^2)$, Group 2: $X_2 \sim N(\mu_2, \sigma_2^2)$

The samples $(X_{11}, \ldots, X_{1n_1})$ and $(X_{21}, \ldots, X_{2n_2})$ are independent,

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown but equal variances)

Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$

Using a well know result (Student, 1908)

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is the pooled variance, we obtain the test statistic

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ (under } H_0)$$

4. Statistical tests: further examples

Ex.: The age (in years) at the first symptoms of a certain disease in a group of men and in an independent group of women were:

Homens	58	52	50	49	56	52	54	48	41	37	67	70
Mulheres	26	41	57	66	36	55	41	61	53	50	52	37	50

Is there a significant difference of the age at the first symptoms of that disease between men and women?

Suppose that (is it reasonable?)

Age for men: $X_1 \sim N(\mu_1, \sigma_1^2)$, Age for women: $X_2 \sim N(\mu_2, \sigma_2^2)$,

the samples are independent

and $\sigma_1^2 = \sigma_2^2 = \sigma^2$. To test

$$H_0 : \ \mu_1 = \mu_2 \text{ versus } H_1 : \ \mu_1 \neq \mu_2$$

at the $100 \times \alpha\%$ significance level, we compute t_0 and reject H_0 if $|t_0| > a$, where $a : P(T_0 > a) = \alpha/2$.

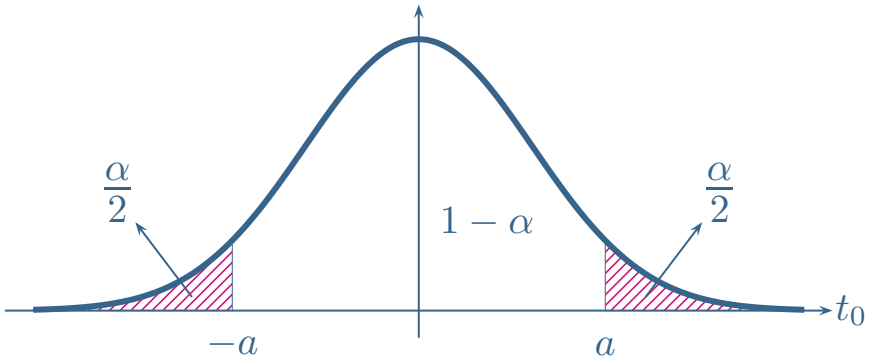
From the given samples we compute

$$\bar{x}_1 = 52.83 \quad s_1^2 = 88.33$$
$$\bar{x}_2 = 48.08 \quad s_2^2 = 126.58$$

$$s_p^2 = \frac{11 \times 88.33 + 12 \times 126.58}{12 + 13 - 2} = 108.29$$

and $t_0 = \dfrac{52.83 - 48.08}{\sqrt{108.29\left(\frac{1}{12} + \frac{1}{13}\right)}} = 1.142$

$\alpha = 5\%$ and $T \sim t_{23} \Rightarrow a = 2.07$
(because $P(T > 2.07) = 0.025$).



Since $t_0 = 1.142 < 2.07$ there is no evidence to reject H_0 (equal ages).

P-value $= 2 \times P(T > 1.142) = 0.2652 > 0.05$.

Obs.: When the assumption of equal variances is not supported ($\sigma_1^2 \neq \sigma_2^2$) it is preferable to use the Welch-(or Satterthwaite) modified test:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu \ \ (\text{sob } H_0)$$

with

$$\nu = \frac{(\nu_1 + \nu_2)^2}{\frac{\nu_1^2}{n_1 - 1} + \frac{\nu_2^2}{n_2 - 1}}, \quad \text{where} \quad \nu_1 = \frac{S_1^2}{n_1} \ \text{ e } \ \nu_2 = \frac{S_2^2}{n_2}$$

4. Statistical tests: further examples

```
> Homens<-c(58,52,50,49,56,52,54,48,41,37,67,70)
> Mulheres<-c(26,41,57,66,36,55,41,61,53,50,52,37,50)
> t.test(Homens,Mulheres)
Standard Two-Sample t-Test
data:  Homens and Mulheres
t = 1.1418, df = 23, p-value = 0.2653
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.861134  13.373954
sample estimates:
 mean of x mean of y
  52.83333  48.07692

> t.test(Homens,Mulheres,var.equal=F)
Welch Modified Two-Sample t-Test
data:  Homens and Mulheres
t = 1.1503, df = 22.792, p-value = 0.2619
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.801706  13.314526
```

4. Statistical tests: further examples

Exploratory plots:

```
> qqplot(Homens,Mulheres)
> abline(0,1)
```

`> boxplot(Homens,Mulheres)` (also assumption of equal variances)

Plots to verify the assumption of normality

```
> qqnorm(Homens)
> qqline(Homens)
```

```
> qqnorm(Mulheres)
> qqline(Mulheres)
```

4. Statistical tests: further examples

Other tests (with less assumptions but also smaller power):

```
> wilcox.test(Homens,Mulheres)

        Wilcoxon rank sum test with continuity correction

data:  Homens and Mulheres
W = 93.5, p-value = 0.4134
alternative hypothesis: true location shift is not equal to 0

> ks.test(Homens,Mulheres)

        Two-sample Kolmogorov-Smirnov test

data:  Homens and Mulheres
D = 0.2179, p-value = 0.9283
alternative hypothesis: two-sided
```

Assumptions of the Wilcoxon test: continuous data, the two distributions are identical except may be in location

Assumptions of the Kolmogorov-Smirnov test: continuous data

4. Statistical tests: further examples

How to determine an appropriate sample size? This sample size depends on

1. H_0 and H_a (value of the alternative hypothesis important to detect)
2. The significance level, α
3. The power wanted for H_a detection
4. The variances of the groups

```
> power.t.test(delta=5,sd=10,power=0.8,type="two.sample",alternative = "two.sided")

     Two-sample t test power calculation
              n = 63.76576
          delta = 5
             sd = 10
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
 NOTE: n is number in *each* group
```

ii. Paired samples (a special case of dependence)

- We say we have a paired sample if by purpose the two variables of interest, X_1 and X_2, are measured on the same subject/object (longitudinal study) or in pairs of subjects/objects chosen on purpose to be similar in several characteristics (case-control study).
- Paired samples are very likely dependent.
- Paired studies tend to be more effective in detecting effects (need less sample units than unpaired studies to detect the same effect).
- These type of data are simple to analyze: consider $D = X_1 - X_2$ and use one-sample methods (two-samples $\longrightarrow$ one sample).
- If it is possible to assume $D \sim N(\mu_D, \sigma_D^2)$ than the test can be based on

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

4. Statistical tests: further examples

Ex.: In a clinical trial to study the effect of hydrochlorothiazide (H) on blood pressure, each patient (suffering from hypertension) receives a placebo (P) and one month later H. The values of the systolic blood pressure at both moments and the corresponding differences were:

P	211	210	210	203	196	190	191	177	173	170	163
H	181	172	196	191	167	161	178	160	149	119	156
P − H	30	38	14	12	29	29	13	17	24	51	7

In order to find if the effect of H is significant consider $D = P - H$ and

$$H_0 : \ \mu_D = 0 \text{ versus } H_1 : \ \mu_D \neq 0$$

4. Statistical tests: further examples

The test statistic is $T_0 = \dfrac{\bar{D}}{S_D/\sqrt{n}} \sim t_{n-1}$ (sob H_0)

Computations: $\bar{d} = 24$, $s_D^2 = 171.4$, $n = 11$ and $t_0 = \dfrac{24}{\sqrt{171.4/11}} = 6.08$

The decision is to reject H_0 for $\alpha = 5\%$ ($P(T > 2.23) = 0.025$).

P-value $= 2 \times P(T > 6.08) = 0.000118$ (very low!)

```
> t.test(DPH)

        One Sample t-test

data:  DPH
t = 6.08, df = 10, p-value = 0.0001188
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 15.20469 32.79531
sample estimates:
mean of x        24
```

5. Multiple tests and multitudes of tests

- In many applications, and genomics is one of them, it is necessary to apply, to the same dataset, hundreds or thousands of tests. An example is the detection of differentially expressed genes in microarray experiments (another example is given in the lab problem).

- When we perform m tests using a significance level of α, we expect to reject H_0 for $m \times \alpha$ tests when H_0 is true for all of them. These wrong rejections are called false positives.

- Multiple test methods are methods designed to limit the number of false positives and work by adjusting either the significance level or the p-values.

Criteria:

- **PCER** (Per Comparison Error Rate): is the usual error rate, α.

  FWER (Family Wise Error Rate): is the probability of at least a false positive.

  FDR (False Discovery Rate): is the expected proportion of type I errors (false positives) among the rejected hypotheses.

Correction methods of the p-values
(the result is called adjusted p-value):

- **Methods that control the FWER:** Bonferroni ($\tilde{p}_j = \min(mp_j, 1)$), Sidák ($\tilde{p}_j = 1 - (1 - p_j)^m$), Holm, Hochberg, Sidák single-step, Sidák step-down

- **Methods that control the FDR:** Benjamini and Hochberg (BH), Benjamini and Yekutieli (BY).

Ex.: Colon cancer data set

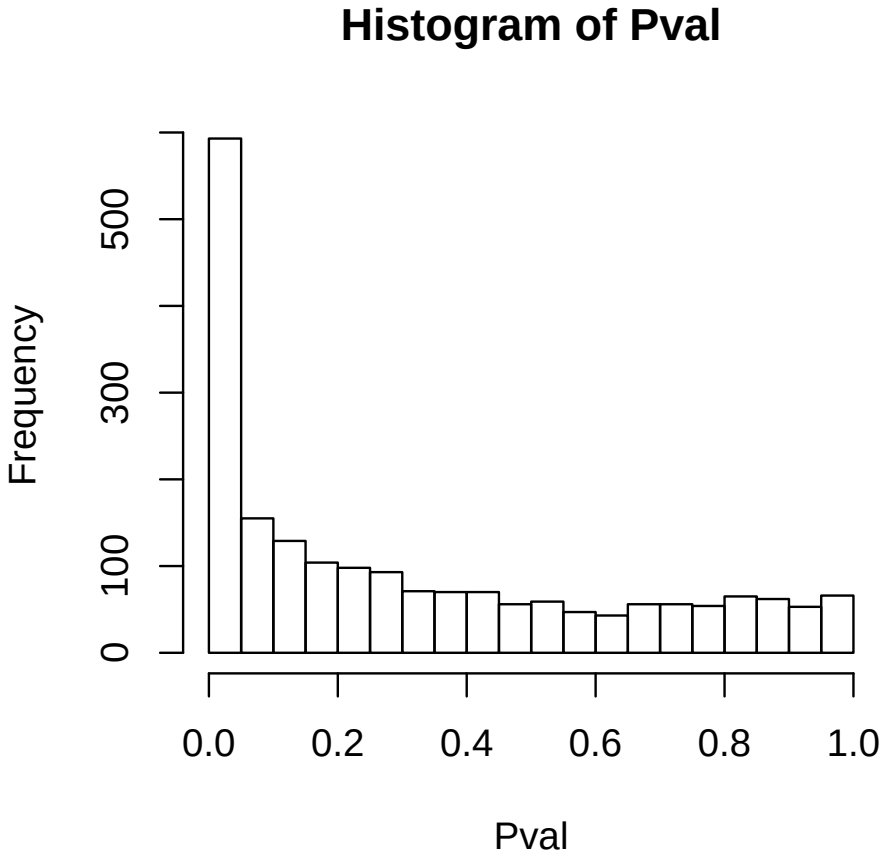
This data set is the result of a microarray experiment (Alon et al., 1999). It contains 62 observations on subjects classified into two groups (G_1: subjects with colon cancer, with 40 observations; G_2: healthy subjects, with 22 observations) and measured on 2000 variables (gene expression levels).

The aim is to find differentially expressed genes between the two conditions (sick and healthy patients).

Solution: consider each gene expression at a time and apply a t-test to find out if there is a significant mean difference between the two groups.

```
> group<-read.table("cancer.txt")   # group codes
> X<-read.table("log.col.nor.txt")   # log normalized data
```

```
> Pval<-rep(NA,2000)
> for (i in 1:2000) Pval[i]<-t.test(X[group==0,i],X[group==1,i])$p.val
> summary(Pval)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
2.770e-10 2.829e-02 2.085e-01 3.163e-01 5.516e-01 9.998e-01
> hist(Pval,breaks=20)
```



An histogram of this form gives indication of a mixture of true and false null hypotheses

5. Multiple tests and multitudes of tests

Number of genes selected without adjustment and with several adjustment methods:

```
> sum(Pval<0.05)
[1] 593
> adj.Holm<-p.adjust(Pval)
> sum(adj.Holm<0.05)
[1] 45
> adj.bonf<-p.adjust(Pval,method="bonf")
> sum(adj.bonf<0.05)
[1] 44
> adj.BH<-p.adjust(Pval,method="BH")
> sum(adj.BH<0.05)
[1] 348
> adj.BY<-p.adjust(Pval,method="BY")
> sum(adj.BY<0.05)
[1] 122
```

6. Some important (bio)statistical models

- **Linear regression**

  Y - dependent variable (continuous)

  X_i - explanatory variables

  ε - random error

$$\text{Model:} \qquad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  β_0, β_1, ...- parameters to be estimated using data information

- **Logistic regression**

  Y - dependent variable (binary: $P(Y = 1) = p$; $P(Y = 0) = 1 - p$)

  X_i - explanatory variables

$$\text{Model:} \qquad \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

  β_0, β_1, ...- parameters to be estimated using data information

7. Some ideas of multivariate analysis

When, for each of n individuals (or objects) several variables are measured/observed (p), we obtain a **multivariate data set**.

Example: the colon cancer data set is a multivariate data set with $n = 62$ objects and $p = 2001$ variables (2000 continuous variables, the genes, plus a binary variable, the group code)

The area of statistics which studies/develops methods for this kind of data is called **Multivariate Analysis**. Examples of multivariate methods

- Principal Component Analysis (reducing the number of variables using linear combinations)
- Cluster analysis (finding groups)
- Discriminant analysis (separating groups) (Fisher, 1936)
- Factor analysis (finding hidden/latent variables) ...

Traditionally $n > p$ (many methods only work in this case).

7. Some ideas of multivariate analysis

In microarray data we have $p >> n$ (p is the number of genes n is the number of arrays):

	a_1	a_2	$\cdots$	a_n
$gene_1$	x_{11}	x_{12}	$\cdots$	x_{1n}
$gene_2$	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$gene_p$	x_{p1}	x_{p2}	$\cdots$	x_{pn}

Consequences:

- Huge number of statistical tests (see previous section)!
- Multivariate analysis turned upside down!

One advice: always try to find out whether the results make sense and validate as much as possible!

8. Conclusions/recommendations

1. **Garbage In, Garbage Out** (GIGO) A bad experiment can not be rescued by the p-value
2. **Be aware of convenience sampling** (randomization is very important).
3. **A good experimental design is crucial**
   - always have a plan before starting the experiment;
   - reduce the sources of variability as much as possible;
   - use controls whenever possible;
   - a designed study is always preferable to an observational study.
4. The hypotheses to test must be defined before the experiment.

8. Conclusions/recommendations

5. **Beware of multiple testing problems.**. Always adjust the p-values (Bonferroni correction or other).
6. A p-value < 0.05 appears once every 20 tests when H_0 is true. **Maintain skepticism** and perform independent validation.
7. Failure to reject the null hypothesis is not equivalent to accept the null hypothesis (**type II errorsI**).
8. Do not confound **statistical significance** with **scientific relevance** or **practical significance**. Always analyze the size of the effects.

Bibliography

Amaratunga D. and Cabrera J. (2003). *Exploration and Analysis of DNA Microarray and Protein Array Data.* Wiley. New York.

Edwards, A.W.F. (1986). Are Mendel's results really too close? *Biol. Rev.*, 61, 295–312

Falconer, D.S. and MacKay, T.F.C. (1996). *Quantitative Genetics.* Prentice Hall, New York.

Franklin, A., Edwards, A.W.F., Fairbanks, D., Hartl, D. and Seidenfeld, T. (2008). *Ending the Mendel-Fisher Controversy.* University of Pittsburgh Press, Pittsburgh

Liu, B. (1997). *Statistical Genomics: Linkage, Mapping and QTL Analysis.* CRC Press.

Novitski, C.E. (1995). Another look at some of Mendel's results. *J. Heredity*, **86**, 62–66.

Piegorsch, W.W. (1986). The Gregor Mendel controversy: Early issues of goodness-of-fit and recent issues of genetic linkage. *Hist. Sci.*, **24**, 173–182.

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W. and Zhao, Y. (2003). *Design and Analysis of DNA Microarray Investigations.* Springer, New York.

Weir, B.S. (1996). *Genetic Data Analysis II.* Sinauer, Sunderland, MA.