Master Degree in Statistics and Economics - University of Milano-Bicocca

Lecture 1: Hypothesis Testing: a review

Lecturer: Aldo Solari

1 Darwin data

Charles Darwin collected data over a period of years on the heights of Zea mays plants. The plants were descended from the same parents and planted at the same time. Half of the plants were *self-fertilized*, and half were *cross-fertilized*, and the purpose of the experiment was to compare their *heights*. To this end Darwin planted them in pairs in different pots.

	Pot	Cross	Self
1	I	23.500	17.375
2	I	12.000	20.375
3	I	21.000	20.000
4	II	22.000	20.000
5	II	19.125	18.375
6	II	21.500	18.625
7	III	22.125	18.625
8	III	20.375	15.250
9	III	18.250	16.500
10	III	21.625	18.000
11	III	23.250	16.250
12	IV	21.000	18.000
13	IV	22.125	12.750
14	IV	23.000	15.500
15	IV	12.000	18.000

Cross-fertilized plants seem higher than self-fertilized ones, with averages:

	type	height
1	Cross	20.19
2	Self	17.57

Questions arise:

1. Is the difference in heights too large to have occurred by chance?
2. Can we estimate the height increase, and assess the uncertainty of our estimate?

1.1 Stochastic proof by contradiction

The simplest way to address the first question is by hypothesis testing. Hypothesis testing is a type of *stochastic proof by contradiction*. To begin we first review the paradigm of *deterministic proof by contradiction*:

1. Assume a proposition, the opposite of what you think about, i.e. the opposite conclusion of your theorem
2. Write down a sequence of logical steps/math
3. Derive a contradiction
4. Conclude that the proposition is false (which implies that the theorem is true)

Example 1.1. *Prove the following statement by Contradiction:* There is no greatest even integer.

Proof. Suppose not. [We take the negation of the theorem and suppose it to be true.] Suppose there is greatest even integer N. [We must deduce a contradiction.] Then
for every even integer n, $N \geq n$.
Now suppose $M = N+2$. Then, M is an even integer. [Because it is a sum of even integers.] Also, $M > N$ [since $M = N + 2$]. Therefore, M is an integer that is greater than the greatest integer. This contradicts the supposition that $N \geq n$ for every even integer n. [Hence, the supposition is false and the statement is true.] And this completes the proof. □

In Hypothesis Testing, the *null hypothesis* (H_0) is the proposition. For the Darwin data:
H_0 : There is no difference in height between cross-fertilized and self-fertilized plants.
Choosing an appropriate H_0 requires some care, as we can design many types of tests. Importantly we cannot specify an alternative after looking at our data; the hypothesis must be generated independent on the data.
The paradigm is as follows:

1. Set H_0
2. Collect data (which is noisy)
3. Derive an apparent contradiction (ie if H_0 is true, then this data is very weird)
4. Hence we reject H_0; this is called a "discovery"

Hypothesis testing is stochastic because we might make errors:

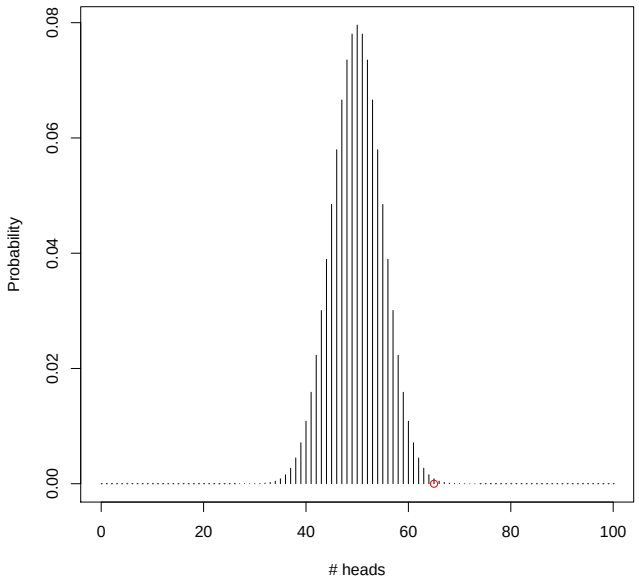
1. *Type I Error*: False Discoveries
2. *Type II Error*: Missed Discoveries

To illustrate this procedure, assume we have a coin and want to determine whether or not it is fair. In this case we have

$$H_0 : \text{Coin is fair } (\theta = 1/2)$$
$$H_1 : \text{Coin is biased } (\theta \neq 1/2)$$

The probability distribution of X = "the number of heads in 100 trials" under H_0 is Binomial($n = 100$, $\theta = 1/2$). After tossing the coin $n = 100$ times we then get $x = 65$ heads and $n - x = 35$ tails.

Figure 1: Tossing a coin 100 times



Is this then enough to reject H_0? To determine this we calculate a p-value associated with our observed data assuming the null hypothesis. A p-value is the probability of seeing what you saw - or something more extreme - given that H_0 is true.

Typically, small p-values imply an unexpected outcome, given that the null (H_0) is true. So if $p = 0.00178993$ then either H_0 isn't true or we are really unlucky and saw this data.

1.2 Galton model

Let's go back do the Darwin data.
Galton considered a model where the height of a self-fertilized plant is

$$Y = \mu + \sigma\varepsilon$$

and of a cross-fertilized plant is

$$X = \mu + \theta + \sigma\epsilon$$

where μ, θ and σ are unknown parameters, and ε and ϵ are independent random variables with mean zero and unit variance.

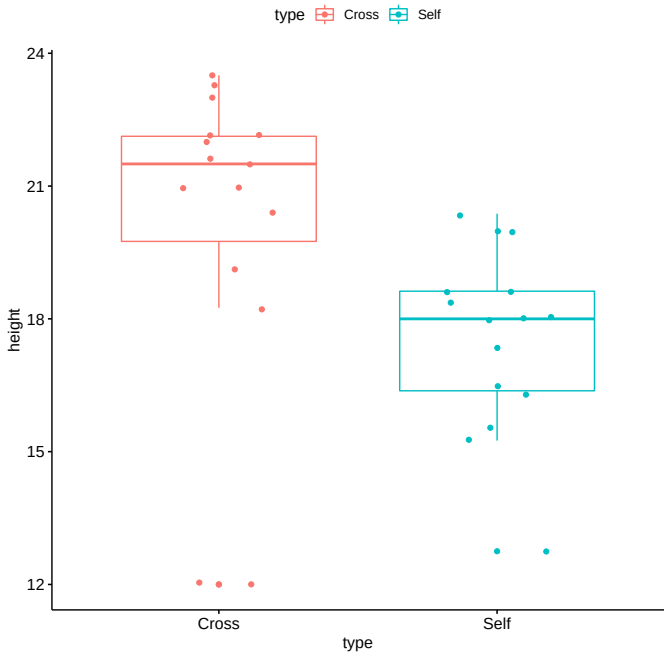
Observations from self-fertilized plants $Y_1, \ldots, Y_{15}$ are i.i.d. as Y, and observations from cross-fertilized plants $X_1, \ldots, X_{15}$ are i.i.d. as X.

Under this model the questions translate to:

1. Is the average height increase $\theta \neq 0$?
2. Can we estimate θ, and assess the uncertainty of our estimate?

The comparison between groups can be visualized by a boxplot:

Figure 2: Darwin data: box-plots



If we assume that ε and ϵ have a $N(0,1)$ distribution, we can use a *two-sample t test*

```
Two Sample t-test

data:  height by type
t = 2.4371, df = 28, p-value = 0.02141
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4173433 4.8159900
```

```
sample estimates:
mean in group Cross  mean in group Self
          20.19167            17.57500
```

1.3 Fisher model

In order to minimize differences in humidity, growing conditions, and lighting, Darwin had taken the trouble to plant the seeds in pairs in the same pots.

Comparison of different pairs would therefore involve these differences, which are not of interest, whereas *comparison within pairs* would depend only on the type of fertilization.

Fisher considered the model

$$Y_i = \mu_i + \sigma\varepsilon_i, \quad X_i = \mu_i + \theta + \sigma\epsilon_i, \quad i = 1, \ldots, n$$

The parameter μ_i represents the effects of the planting conditions for the ith pair, and ε_i and ϵ_i are independent random variables with mean zero and unit variance.

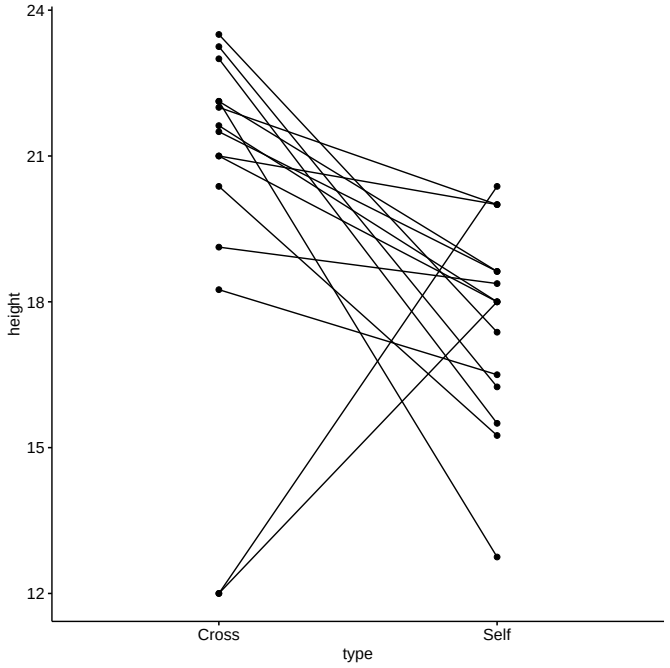
The μ_i could be eliminated by using the differences

$$D_i = X_i - Y_i$$

which have mean θ and variance $2\sigma^2$.

The comparison within pairs can be visualized by the following plot:

Figure 3: Darwin data: difference in height for each pair



If we assume that ε and ϵ have a $N(0, 1)$ distribution, we can use a *paired t test*, or *one-sample t test* for the difference:

```
One Sample t-test

data:  differences
t = 2.148, df = 14, p-value = 0.0497
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.003899165 5.229434169
sample estimates:
mean of x
 2.616667
```

2 P-values

A scientific theory leads to assertions that are testable using empirical data. One way to investigate the extent to which a theory is supported by the data Y is to choose a test statistic, $T = t(Y)$, large values of which cast doubts on the theory.

This theory, the null hypothesis H_0, places restrictions on the distribution of Y and is used to calculate a p-value

$$p_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}})$$

where t_{obs} is the value of T actually observed, i.e. $t_{\text{obs}} = t(y_{obs})$ when we observe the realization y_{obs} of Y.

A distribution computed under the assumption that H_0 is true is called *null distribution*, and then we use $\mathrm{P}_0, \mathrm{E}_0, \ldots$ to indicate probability, expectation and so forth.

The p-value may be written as $p_{\text{obs}} = 1 - F_0(t_{\text{obs}})$, where F_0 is the null distribution function of T, supposed to be continuous and invertible.

One interpretation of p_{obs} stems from the corresponding random variable $P = 1 - F_0(T)$. For $0 \leq u \leq 1$, its null distribution is *Uniform* on the unit interval:

$$\mathrm{P}_0(P \leq u) = \mathrm{P}_0(F_0^{-1}(1-u) \leq T) = 1 - F_0(F_0^{-1}(1-u)) = u$$

If we consider t_{obs} as just decisive evidence against H_0, then we would reject the hypothesis when true a long-run proportion p_{obs} of times.

A more formal definition of p-value is

Definition 2.1. A p-value is the smallest significance level α at which the null hypothesis H_0 would be rejected for the given observation, i.e.

$$P(Y) = \inf\{\alpha \in (0,1) : Y \in R_\alpha\}$$

where R_α is the *rejection region* of the test, assumed to satisfy $R_\alpha \subseteq R_{\alpha'}$ if $\alpha < \alpha'$.

2.1 Valid p-values

Definition 2.2. We say that we have a *valid test* if its p-value is uniformly distributed under H_0, i.e.

$$\mathrm{P}_0(P \leq u) = u \quad \forall u \in (0,1) \tag{1}$$

or more generally if the p-value is *stochastically dominated* by the uniform distribution under H_0, i.e.

$$\mathrm{P}_0(P \leq u) \leq u \quad \forall u \in (0,1) \tag{2}$$

2.2 One- and two-sided tests

Suppose that we have a test statistic T, small and large values of which indicate a departure from H_0. The simplest procedure is then often to contemplate two tests, one for each tail.

With test statistic T, consider two p-values, namely

$$p^-_{\mathrm{obs}} = \mathrm{P}_0(T \leq t_{\mathrm{obs}}), \quad p^+_{\mathrm{obs}} = \mathrm{P}_0(T \geq t_{\mathrm{obs}})$$

When the distribution of T is continuous, the p-value is the smallest between p^-_{obs} and p^+_{obs} multiplied by 2:

$$p_{\mathrm{obs}} = 2\min(p^-_{\mathrm{obs}}, p^+_{\mathrm{obs}})$$

This follows because the null distribution of $Q = \min(P^-, P^+)$ is

$$Q = \min(1 - U(0,1), U(0,1)) = U(0, 1/2)$$

thus the null distribution of $2Q$ is $U(0,1)$.

When the distribution of T is discrete, p_{obs} is q_{obs} plus the achievable p-value from the other tail of the distribution nearest to but not exceeding q_{obs}.

For example, suppose we are testing that a coin is fair by tossing the coin 100 times and getting 65 heads and 35 tails. Then $p^-_{\mathrm{obs}} = \mathrm{P}(\mathrm{Binomial}(100, 1/2) \leq 65) = 0.999105$, $p^+_{\mathrm{obs}} = \mathrm{P}(\mathrm{Binomial}(100, 1/2) \geq 65) = 0.0008949652$, and the p-value is $q_{\mathrm{obs}} + \mathrm{P}(\mathrm{Binomial}(100, 1/2) \leq 34) = 0.0008949652 + 0.0008949652 = 0.00178993$.

3 Nonparametric tests

Tests where the null hypotheses itself is formulated in terms of arbitrary distributions, so-called *nonparametric* or *distribution-free tests*.

Example 3.1. *Assume we have n data points $(X_i, Y_i) \in \mathbb{R}^2$. A parametric null hypothesis would be that X and Y are uncorrelated. A nonparametric test would be that X and Y are* stochastically independent*, i.e. $X \perp\!\!\!\perp Y$. Note that these tests are equivalent if and only if X and Y are normally distributed. In this example if the true relationship between X and Y is linear, the non-parametric test has less power. However, if the relationship between X and Y is non-linear (e.g. $X^2 + Y^2 = 1$) then the parametric test will have less power.*

Consider a more general matched pair model for the Darwin data:

$$Y_i = \mu_i + \sigma_i \varepsilon_i, \quad X_i = \mu_i + \theta + \tau_i \epsilon_i, \quad i = 1, \ldots, n$$

with ε_i and ϵ_i are independent random variables with continuous distribution and mean zero and unit variance.

The height differences may be written as

$$D_i = \theta + (\tau_i \epsilon_i - \sigma_i \varepsilon_i)$$

If we assume that ε_i and ϵ_i are independent and symmetrically distributed around 0, then D_i is symmetrically distributed around θ, i.e.

$$D_i - \theta \stackrel{d}{=} \theta - D_i \qquad i = 1, \ldots, n$$

The null hypothesis is $H_0 : \theta = 0$. If H_0 is true, then the probability that D_i falls on either side of 0 is $1/2$. This suggests that we can base a test on

$$T = \sum_{i=1}^{n} \mathbb{1}\{D_i > 0\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Note that using $>$ or $\geq$ in the indicator function is equivalent because the probability that $D_i = 0$ is 0 since we have assumed that the distribution of D_i is continuous. This is known as the *sign test*.

Under H_0, T has a binomial distribution with size n and probability $1/2$, so its mean and variance are $n/2$ and $n/4$. The one-sided p-values are

$$p^{+}_{\text{obs}} = \mathrm{P}_0(T \geq t_{\text{obs}}) = \sum_{k=t_{\text{obs}}}^{n} \binom{n}{k} \frac{1}{2^n}, \quad p^{-}_{\text{obs}} = \mathrm{P}_0(T \leq t_{\text{obs}}) = \sum_{k=0}^{t_{\text{obs}}} \binom{n}{k} \frac{1}{2^n}$$

By testing $H_0 : \theta = 0$, we have $t_{\text{obs}} = 13$ and a two-sided p-value of

```
Exact binomial test

data:  13 and 15
number of successes = 13, number of trials = 15, p-value = 0.007385
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5953973 0.9834241
sample estimates:
probability of success
             0.8666667
```

4 Goodness-of-fit

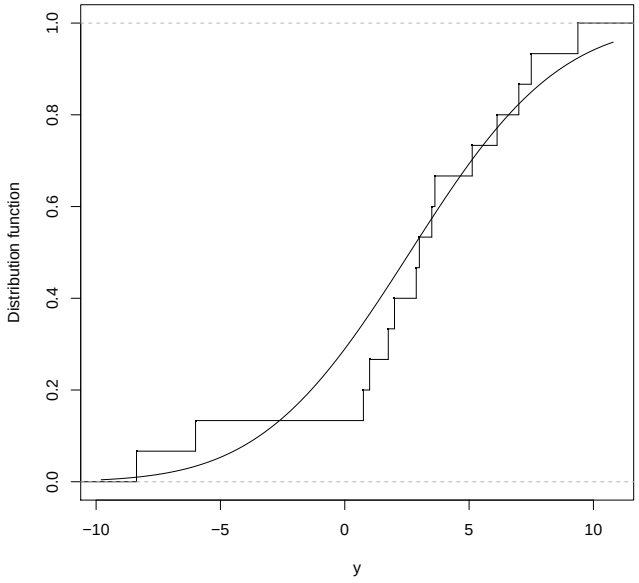
Suppose that the null hypothesis is that the random sample $Y_1, \ldots, Y_n$ is from a known continuous distribution $F(y)$. Then we can compare $F(y)$ with the empirical distribution function

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I\{Y_i \leq y\}$$

whose mean and variance are $F(y)$ and $[F(y)(1 - F(y))]/n$ under H_0.

Under the matched pair model for the Darwin data, we assumed that $D_1, \ldots, D_{15}$ are i.i.d. as $N(\theta, \tau^2)$ where $\tau^2 = 2\sigma^2$. Because θ and τ^2 are unknown, we can by replace them with the sample mean and variance $\hat{\theta}$ and $\hat{\tau}^2$. However, this is not formally correct: the parameters specified in must be pre-specified and not estimated from the data. The problem is that we are using the same data twice: once for estimation, and then for testing. The resulting inference is only approximate in finite samples.

Figure 4: Darwin data: comparison between distribution and empirical distribution



A standard test for H_0 is based on the Kolmogorov-Smirnov statistic

$$T = \sup_y |\hat{F}(y) - F(y)|$$

```
One-sample Kolmogorov-Smirnov test

data:  differences
```

```
D = 0.21285, p-value = 0.4442
alternative hypothesis: two-sided
```

To compute the p-value we can generate B independent sets of data from the null distribution $N(\theta, \tau^2)$, calculating the corresponding statistics T^b and

$$p_{\text{obs}} = \frac{1 + \sum_{b=1}^{B} I\{T^b \geq t_{\text{obs}}\}}{1 + B}$$

If we generate data from $N(\hat{\theta}, \hat{\tau}^2)$ instead of $N(\theta, \tau^2)$, the resulting inference is only approximate. There is some more refined distribution theory for the Kolmogorov-Smirnov test with estimated parameters, but that is not implemented in the R function 'ks.test'.

Figure 5: Darwin data: histogram of the null distribution for the Kolgomorov-Smirnov statistic