Hypothesis testing

Null hypothesis H_0 and alternative hypothesis H_1.
Simple and compound hypotheses.

Simple : the probabilistic model is specified completely.

Compound : the probabilistic model is not specified completely (generally it will contain parameters to be estimated).

Example 1: we want to test whether data are compatible with the assumption that their true mean is μ_0. Then it could be set as:

H_0: $X_1, \ldots, X_n \sim N(\mu_0, \sigma_0^2)$ and independent. [simple if σ^2 known];

H_1: $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ and independent, with $\mu \neq \mu_0$. [compound]

Rejection region

Example 2: we have two groups, and we wish to test whether they can be considered as samples from the same population, or from two populations with different means. General assumption:

$$X_1, \dots, X_n \sim N(\mu_X, \sigma^2),\ Y_1, \dots, Y_m \sim N(\mu_Y, \sigma^2),\ \text{and independent.}$$

H_0: $\mu_X = \mu_Y,\ \sigma^2 > 0.$

H_1: $\mu_X \neq \mu_Y,\ \sigma^2 > 0.$

Both are compound, but H_0 is 'simpler' than H_1.

How does a test work? We select a *rejection region* C: if data fall in C, we reject H_0 (and accept H_1); if data do not fall in C, we accept (do not reject) H_0.

Errors of first and second species

Error of first species: rejecting H_0 if H_0 is true;
Error of second species: accepting H_0 if H_1 is true.

A smaller rejection region C decreases error of first species, but increases those of second species; a larger C vice versa.

A test of hypothesis is a region C: it will have a *level* (the risk I take of errors of 1st species) and a *power* (the probability of not making errors of 2nd species).

Level and power of a test

Level The probability of an error of 1st species, i.e. to reject H_0 when H_0 is true.

Power $1-$ the probability of an error of 2st species, i.e. to reject H_0 when H_0 is false.

If hypotheses were simple, level and power could be computed exactly.

In actual tests, the level can often be computed or bounded from above; the power will depend on exact parameter value.

Level and power of a test

Level The probability of an error of 1st species, i.e. to reject H_0 when H_0 is true.

Power $1-$ the probability of an error of 2st species, i.e. to reject H_0 when H_0 is false.

If hypotheses were simple, level and power could be computed exactly.
In actual tests, the level can often be computed or bounded from above; the power will depend on exact parameter value.
Ideally the level should be close to 0 and the power close to 1. But to decrease the level, we should reject H_0 less often, thus decrease the power.

Solution? Choose the level α [often 5%]. Then among all possible tests of level α (i.e. rejection regions s.t. $\mathbb{P}(X \in C|H_0) \leq \alpha$) choose the one of highest power [*uniformly most powerful test*]; this is not always possible, but it is the rationale for many well known tests.

One-sample test on the mean

Example 1: (with σ^2 unknown).

H_0 : $X_1, \ldots, X_n \sim N(\mu_0, \sigma^2)$ and independent, where $\sigma^2 > 0$.

H_1 : $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and independent, where $\mu \neq \mu_0$, $\sigma^2 > 0$.

The test quantity used is $T = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$ where S^2 is the sample variance.

It is natural (and can be justified rigorously) to reject H_0 when T is far away from 0.

Under H_0, T follows a $t(n-1)$ distribution. Then we find t_α s.t. $\mathbb{P}(|t(n-1)| > t_\alpha) = \alpha$. Reject H_0 if $|T| > t_\alpha$, accept it otherwise.

One-sample test on the mean. II

If *[unilateral alternative]*

H_1 : $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and independent, where $\mu > \mu_0$, $\sigma^2 > 0$,

then the rejection region is for T positive and large.
Hence we find t'_α s.t. $\mathbb{P}(t(n-1) > t\_\alpha) = \alpha$.
Reject H_0 if $T > t'_\alpha$, accept it otherwise.
Vice versa if the alternative hypothesis is $\mu < \mu_0$.

Often programs (e.g. R) return the p-value $= \mathbb{P}(|t(n-1)| > T)$ (for a bilateral test), [or $\mathbb{P}(t(n-1) > T)$ against $\mu > \mu_0$]. If the p-value is less than the level we chose, reject H_0; otherwise accept.

Observations: unilateral alternatives make it easier rejecting the null hypothesis (hence they are seldom used).
In practice, border-line results suggest further research.

Test on equality of the means

Independent samples (e.g. 2 groups with different treatments)

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2),\ Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2),\ \text{and independent.}$$

$$H_0 :\ \mu_X = \mu_Y,\ \sigma^2 > 0. \qquad H_1 :\ \mu_X \neq \mu_Y,\ \sigma^2 > 0.$$

Test on equality of the means

Independent samples (e.g. 2 groups with different treatments)

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2),\ Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2),\ \text{and independent.}$$

$$H_0:\ \mu_X = \mu_Y,\ \sigma^2 > 0. \qquad H_1:\ \mu_X \neq \mu_Y,\ \sigma^2 > 0.$$

Estimate of σ^2 : $\quad S^2_{X,Y} = \dfrac{1}{n+m-2}\left((n-1)S^2_X + (m-1)S^2_Y\right)$

Test on equality of the means

Independent samples (e.g. 2 groups with different treatments)

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2),\ Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2),\ \text{and independent.}$$

$$H_0:\ \mu_X = \mu_Y,\ \sigma^2 > 0. \qquad H_1:\ \mu_X \neq \mu_Y,\ \sigma^2 > 0.$$

Estimate of σ^2 : $\quad S^2_{X,Y} = \dfrac{1}{n+m-2}\left((n-1)S^2_X + (m-1)S^2_Y\right)$

Under H_0 $T = \dfrac{\bar{Y} - \bar{X}}{s_{X,Y}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$ follows a $t(m+n-2)$ distribution.

Test on equality of the means

Independent samples (e.g. 2 groups with different treatments)

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2), \; Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2), \text{ and independent.}$$

$$H_0: \; \mu_X = \mu_Y, \; \sigma^2 > 0. \qquad H_1: \; \mu_X \neq \mu_Y, \; \sigma^2 > 0.$$

Estimate of σ^2 : $\quad S_{X,Y}^2 = \dfrac{1}{n+m-2}\left((n-1)S_X^2 + (m-1)S_Y^2\right)$

Under H_0 $T = \dfrac{\bar{Y} - \bar{X}}{s_{X,Y}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$ follows a $t(m+n-2)$ distribution.

Assumptions: normality, independence, equality of variances (should be checked [sometimes variable transformations help]).

Test on equality of the means

Independent samples (e.g. 2 groups with different treatments)

$$X_1, \ldots, X_n \sim N(\mu_X, \sigma^2),\ Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma^2),\ \text{and independent.}$$

$$H_0:\ \mu_X = \mu_Y,\ \sigma^2 > 0. \qquad H_1:\ \mu_X \neq \mu_Y,\ \sigma^2 > 0.$$

Estimate of σ^2 : $\quad S^2_{X,Y} = \dfrac{1}{n+m-2}\left((n-1)S^2_X + (m-1)S^2_Y\right)$

Under H_0 $T = \dfrac{\bar{Y} - \bar{X}}{s_{X,Y}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$ follows a $t(m+n-2)$ distribution.

Assumptions: normality, independence, equality of variances (should be checked [sometimes variable transformations help]).
A modified version (Welch t-test) works without assuming equal variances.

Test on equality of the means. II

Paired samples (e.g. same individuals before/after treatment)

General assumption

$$D_i = Y_i - X_i \sim N(\mu, \sigma^2), \quad i = 1 \ldots n.$$

No assumption on X_i and Y_i, but only on their differences (the effect of treatment).

$$H_0: \ \mu = 0, \ \sigma^2 > 0. \qquad H_1: \ \mu \neq 0, \ \sigma^2 > 0.$$

This is simply a test that the true mean of $D = Y - X$ is 0.

It will be easier rejecting H_0 because generally s_D is much smaller than $\sqrt{2}s_{X,Y}$.
Basic assumption: the effect of treatment is additive (does not depend on the original value of X_i).

Paires samples. An example

Body and encephalus temperature measured on 6 ostriches kept at hot outside temperature:

Ostrich	Body T	encephalus T
1	38.51	39.32
2	38.45	39.21
3	38.27	39.20
4	38.52	38.68
5	38.62	39.09
6	38.18	38.94

Paires samples. An example

Body and encephalus temperature measured on 6 ostriches kept at hot outside temperature:

Ostrich	Body T	encephalus T
1	38.51	39.32
2	38.45	39.21
3	38.27	39.20
4	38.52	38.68
5	38.62	39.09
6	38.18	38.94

$$\bar{X} = 38.425 \quad \bar{Y} = 39.073 \quad S_D = 0.283 \quad T = \sqrt{n}\frac{\bar{Y} - \bar{X}}{S_D} = 5.6099.$$

$$p\text{-value} = \mathbb{P}(|t(5)| > 5.6099) = 0.00249.$$

Paires samples. An example

Body and encephalus temperature measured on 6 ostriches kept at hot outside temperature:

Ostrich	Body T	encephalus T
1	38.51	39.32
2	38.45	39.21
3	38.27	39.20
4	38.52	38.68
5	38.62	39.09
6	38.18	38.94

$$\bar{X} = 38.425 \quad \bar{Y} = 39.073 \quad S_D = 0.283 \quad T = \sqrt{n}\frac{\bar{Y} - \bar{X}}{S_D} = 5.6099.$$

$$p\text{-value} = \mathbb{P}(|t(5)| > 5.6099) = 0.00249.$$

Reject $\mu_{Y-X} = 0$.

Non-parametric tests

One may question the assumption that differences are normally distributed.
There exists tests that are not based on a specific *parametric* form.

Non-parametric tests

One may question the assumption that differences are normally distributed.

There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. H_1 : median of $D \neq 0$.

Non-parametric tests

One may question the assumption that differences are normally distributed.

There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. H_1 : median of $D \neq 0$.

Under H_0, $\mathbb{P}(D_i > 0) = \mathbb{P}(D_i < 0) = \frac{1}{2}$. Count number of positive n_+ and negative n_- observations: if they are far from $\frac{n}{2}$, reject H_0.

Non-parametric tests

One may question the assumption that differences are normally distributed.

There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. H_1 : median of $D \neq 0$.

Under H_0, $\mathbb{P}(D_i > 0) = \mathbb{P}(D_i < 0) = \frac{1}{2}$. Count number of positive n_+ and negative n_- observations: if they are far from $\frac{n}{2}$, reject H_0.

Example of ostriches: $n_+ = 6$, $n_- = 0$. Which is the probability, under H_0, to have a result as extreme (or more)?

Non-parametric tests

One may question the assumption that differences are normally distributed.
There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. H_1 : median of $D \neq 0$.

Under H_0, $\mathbb{P}(D_i > 0) = \mathbb{P}(D_i < 0) = \frac{1}{2}$. Count number of positive n_+ and negative n_- observations: if they are far from $\frac{n}{2}$, reject H_0.

Example of ostriches: $n_+ = 6$, $n_- = 0$. Which is the probability, under H_0, to have a result as extreme (or more)?

$$\mathbb{P}(n_+ = 6) = \left(\frac{1}{2}\right)^6 = 0.015625 \ = \mathbb{P}(n_+ = 0)$$

Non-parametric tests

One may question the assumption that differences are normally distributed.
There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. $\quad$ H_1 : median of $D \neq 0$.

Under H_0, $\mathbb{P}(D_i > 0) = \mathbb{P}(D_i < 0) = \frac{1}{2}$. Count number of positive n_+ and negative n_- observations: if they are far from $\frac{n}{2}$, reject H_0.

Example of ostriches: $n_+ = 6$, $n_- = 0$. Which is the probability, under H_0, to have a result as extreme (or more)?

$$\mathbb{P}(n_+ = 6) = \left(\frac{1}{2}\right)^6 = 0.015625 = \mathbb{P}(n_+ = 0), \quad p\text{-value} = 3.125\%.$$

Non-parametric tests

One may question the assumption that differences are normally distributed.
There exists tests that are not based on a specific *parametric* form.

Sign test: H_0 : median of $D = 0$. H_1 : median of $D \neq 0$.

Under H_0, $\mathbb{P}(D_i > 0) = \mathbb{P}(D_i < 0) = \frac{1}{2}$. Count number of positive n_+ and negative n_- observations: if they are far from $\frac{n}{2}$, reject H_0.

Example of ostriches: $n_+ = 6$, $n_- = 0$. Which is the probability, under H_0, to have a result as extreme (or more)?

$$\mathbb{P}(n_+ = 6) = \left(\frac{1}{2}\right)^6 = 0.015625 = \mathbb{P}(n_+ = 0), \quad p\text{-value} = 3.125\%.$$

Sign test is very robust, but not very powerful (*rejecting* H_0 is difficult).
There exist intermediate tests such as Wilcoxon's test, that use not only signs, but also **ranks** of observations.

Chi-square test

General chi-square test:
We have k types of events that can occur in each trial, with a priori probabilities for them

$$p_1^0, \ldots, p_k^0 \qquad \text{with} \qquad p_1^0 + \cdots + p_k^0 = 1.$$

After n trials, we observe

$$n_1 \text{ events of type } 1, \; n_2 \text{ of } 2, \ldots, n_k \text{ of } k, \; \text{with } n_1 + \cdots + n_k = n.$$

Are data compatible with expectations? ($k = 2$ is binomial)

Classical test: **chi-square**:
Set $E_i = np_i^0$ (expected number of events of type i under H_0),

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i} \sim \chi^2(k-1) \quad \text{for } n \text{ large.}$$

Find c_α s.t. $\mathbb{P}(\chi^2(k-1) > c_\alpha) = \alpha$. If $X^2 > c_\alpha$, reject H_0; accept it otherwise.

Chi-square for data fit to a distribution

The values $p_1^0, \ldots, p_k^0$ can be those arising from some distribution.
Often the distribution will contain parameters to be estimated (e.g. λ of Poisson).
One can use the chi-square: if m parameters are estimated, $X^2 \sim \chi^2(k - m - 1)$ (of course $m < k - 1$).
Example: data (Von Bortkiewicz, 1898) on Prussians soldiers kicked to death by horses:

i (deaths)	n_i (number of corps/years)
0	109
1	65
2	22
3	3
4	1
Total	200

Chi-square example (continued)

Estimate λ with the sample mean
$\hat{\lambda} = (1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1)/200 = 0.61$. Compute E_i.
Join the classes ≥ 3 (rule of thumb: $E_i \geq 5$) to obtain:

i	n_i	$\hat{E}_i$
0	109	108.67
1	65	66.29
2	22	20.22
≥ 3	4	4.82
Total	200	

Compute $X^2 \approx 0.32$. p-value $= \mathbb{P}(\chi^2(2) > 0.32) = 85.2\%$.

Chi-square test of independence

A classical use of chi-square is when we observe two qualitative variables X and Y.

H_0 : variables are independent; H_1: they are not independent.

X: k levels, Y: l levels (if $k = l = 2$, a 2×2 contingency table).

Data: n_{ij} (# of observ. with $X = i$ and $Y = j$).

Chi-square test of independence

A classical use of chi-square is when we observe two qualitative variables X and Y.

H_0 : variables are independent; H_1: they are not independent.

X: k levels, Y: l levels (if $k = l = 2$, a 2×2 contingency table).

Data: n_{ij} (# of observ. with $X = i$ and $Y = j$).

$$H_0 : \mathbb{P}(X = i,\ Y = j) = p_i q_j \qquad \text{for all } i \text{ and } j$$

$p_i,\ i = 1 \ldots k-1$, $q_j,\ j = 1 \ldots l-1$ to be estimated from data.

$$H_1 :\ \mathbb{P}(X = i,\ Y = j) \neq p_i q_j.$$

Computations in test of independence

Row totals: $\quad n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$

column totals: $\quad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$

grand total: $\quad n_{\bullet\bullet} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}.$

$$\hat{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}} \quad \text{so that} \quad X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

This to be compared with $\chi^2(k \cdot l - k - l + 1) = \chi^2((k-1)(l-1))$.

Computations in test of independence

Row totals: $n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$

column totals: $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$

grand total: $n_{\bullet\bullet} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}.$

$$\hat{E}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}} \quad \text{so that} \quad X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

This to be compared with $\chi^2(k \cdot l - k - l + 1) = \chi^2((k-1)(l-1))$.

Chi-square is only an approximation. One can perform exact tests based on binomial (Fisher's test)

Example of test of independence

In a (hypotethical) study a set of individuals was treated with an antiviral or a placebo, and then experimentally infected with a mild strain of influenza. From following analyses, individuals were classified as "No virus", "Virus but no symptoms", "severe infections" obtaining the table:

	NV	VNS	SI	Total
Antiviral	8	21	4	33
Placebo	6	14	12	32
Total	14	35	16	65

Example: computations

Observations:

NV	VNS	SI	Total
8	21	4	33
6	14	12	32
14	35	16	65

Expected values:

$$\frac{33\cdot 14}{65} = 7.1 \quad \frac{33\cdot 35}{65} = 17.8 \quad \frac{33\cdot 16}{65} = 8.1$$

$$\frac{32\cdot 14}{65} = 6.9 \quad \frac{32\cdot 35}{65} = 17.2 \quad \frac{32\cdot 16}{65} = 7.9$$

$$X^2 = \frac{(8-7.1)^2}{7.1} + \frac{(21-17.8)^2}{17.8} + \frac{(4-8.1)^2}{8.1} + \frac{(6-6.9)^2}{6.9} + \frac{(14-17.2)^2}{17.2} + \frac{(12-7.9)^2}{7.9} = 5.605.$$

p-value $= \mathbb{P}(\chi^2(2) > 5.605) = 6.1\%$.

We cannot reject independence, though it is a borderline case.

Comparison of means of multiple groups

When we have many (more than 2) groups, we may think to perform t-tests for $\mu_1 = \mu_2$, then $\mu_1 = \mu_3$, then $\mu_2 = \mu_3 \ldots$

Comparison of means of multiple groups

When we have many (more than 2) groups, we may think to perform t-tests for $\mu_1 = \mu_2$, then $\mu_1 = \mu_3$, then $\mu_2 = \mu_3 \ldots$ Why is this not appropriate? not optimal?

Comparison of means of multiple groups

When we have many (more than 2) groups, we may think to perform t-tests for $\mu_1 = \mu_2$, then $\mu_1 = \mu_3$, then $\mu_2 = \mu_3 \ldots$
Why is this not appropriate? not optimal?
If we perform many tests, we need to correct probability levels. If we perform 20 tests, each with 5% probability of being positive, we suspect some may become positive just for chance...

Comparison of means of multiple groups

When we have many (more than 2) groups, we may think to perform t-tests for $\mu_1 = \mu_2$, then $\mu_1 = \mu_3$, then $\mu_2 = \mu_3 \ldots$

Why is this not appropriate? not optimal?

If we perform many tests, we need to correct probability levels. If we perform 20 tests, each with 5% probability of being positive, we suspect some may become positive just for chance...

Tests are not independent...

Comparison of means of multiple groups

When we have many (more than 2) groups, we may think to perform t-tests for $\mu_1 = \mu_2$, then $\mu_1 = \mu_3$, then $\mu_2 = \mu_3 \ldots$
Why is this not appropriate? not optimal?
If we perform many tests, we need to correct probability levels. If we perform 20 tests, each with 5% probability of being positive, we suspect some may become positive just for chance...
Tests are not independent...
Proper way to correcting for this: analysis of variance.