Statistical tests:
What does it mean? How to use it?

P. Reynaud-Bouret

CNRS, University of Nice Sophia-Antipolis, France

Les Mini-Cours de "Modélisation, Statistiques et Programmation Pour Tous", Nice, December 4th, 2015
Plan

1. Test ?
2. Tests that you can build by yourself
3. Multiple testing
4. "Goodness-of-fit" tests
5. Independence testing
Basic Rules of Statistics
Basic Rules of Statistics

Observations

= what is recorded = data
It is the realisation of something random and we only see some tracks.

Ex: weight of French people → observations = weight of people in this room
Basic Rules of Statistics

Observations

= what is recorded = data
It is the realisation of something random and we only see some tracks.

Ex : weight of French people → observations = weight of people in this room

THE Unbreakable Rule of Statistics

Data, Nothing else but Data, Everything lies in the Data !!!

→ a statistical procedure takes data in entry and gives an answer, which is just a more understandable transformation of the data....
Basic Rules of Statistics (2)

- **Estimator**: if a quantity is unknown (ex: mean weight in the french population, \( m \)), it can be estimated by a transformation of the data (average of the weights of the people in the room, \( \hat{m} \)).
Basic Rules of Statistics (2)

- **Estimator**: if a quantity is unknown (ex: mean weight in the French population, $m$), it can be estimated by a transformation of the data (average of the weights of the people in the room, $\hat{m}$).

- **Confidence interval**: of course, there is an error in the estimation $\rightarrow$ a *margin of error*
  $\rightarrow m$ is in $[\hat{m} - \epsilon, \hat{m} + \epsilon]$ with more than $1 - \alpha$ (95 %) chance.
**Basic Rules of Statistics (2)**

- **Estimator**: if a quantity is unknown (ex: mean weight in the French population, \( m \)), it can be estimated by a transformation of the data (average of the weights of the people in the room, \( \hat{m} \)).

- **Confidence interval**: of course, there is an error in the estimation \( \rightarrow \) a margin of error
  
  \[ m \text{ in } [\hat{m} - \epsilon, \hat{m} + \epsilon] \text{ with more than } 1 - \alpha \text{ (95 \%)} \text{ chance} \]

Choice of \( \epsilon \):
- depends on \( \alpha \) and a priori on data,
- less a priori = better for statistics \( \rightarrow \) \( \epsilon \) = function of the data.
Basic Rules of Statistics (2)

- **Estimator**: if a quantity is unknown (ex: mean weight in the French population, $m$), it can be estimated by a transformation of the data (average of the weights of the people in the room, $\hat{m}$).

- **Confidence interval**: of course, there is an error in the estimation $\rightarrow$ a margin of error

  $m$ is in $[\hat{m} - \epsilon, \hat{m} + \epsilon]$ with more than $1 - \alpha$ (95 %) chance. Choice of $\epsilon$:

  - depends on $\alpha$ and a priori on data,
  - less a priori = better for statistics $\rightarrow \epsilon = $ function of the data.

$\rightarrow$ the randomness was taken into account in $\epsilon$ and $\alpha$.

When the question is quantitative (weight ?) $\rightarrow$ statistical answer that quantifies the ”almost” in a quite intuitive manner.
Definition of a Statistical Test
Test

→ YES/NO question (Are we heavier than 10 years ago? etc)
Test

→ YES/NO question (Are we heavier than 10 years ago? etc)
The data satisfy an hypothesis:

Hypotheses

Null Hypothesis = "YES is true"
Alternative = "NO is true"
Test

→ YES/NO question (Are we heavier than 10 years ago? etc)
The data satisfy an hypothesis:

Hypotheses

Null Hypothesis = "YES is true"
Alternative = "NO is true"

The statistician does not know which hypothesis holds. His/her answer is given by a test $\Delta$ whose value can only be 0 or 1.

Test

$\Delta = 0$: "The null hypothesis is accepted."
$\Delta = 1$: "The null hypothesis is rejected."

Rk: A statistician will NEVER simply say "YES" or "NO" since he/she knows he/she can be mistaken.
The errors of a test

→ Where is ”almost” in the statistical answer and how is it quantified?
  
  • First kind error : probability to wrongly reject the null hypothesis
    ∼ One says NO when YES holds
  
  • Second kind error : probability to wrongly accept the null hypothesis
    ∼ One says YES when NO holds

Generally, it is not possible in practice to control both errors but only one of them
→ asymmetry
→ a test has a ”dearie” hypothesis (”chérie”) = the null hypothesis
→ for a prescribed level $\alpha$ (ex : 5%), one wants

  probability of first kind error $\leq \alpha$,

→ test of level $\alpha$
The toy example of testing: the penguin problem
The "toy" example

Penguins eggs have usually one chance out of two to hatch out. A factory settles near the colony.
The "toy" example

We observe $n$ penguin eggs of this colony to see if they hatch or not.

$X_i = 1$

$X_i = 0$

The observed empirical frequency is $\bar{X} = 0.48$.

Is the factory responsible for a decrease in birth rate of the penguins?
The choice of the null hypothesis

A test is built such that probability of "wrongly rejecting $H_0" \leq \alpha = 5%.
Let $p$ be the probability that an egg hatches.

The green

The boss
The choice of the null hypothesis

A test is built such that probability of "wrongly rejecting $H_0" \leq \alpha = 5\%.

Let $p$ be the probability that an egg hatches.

**The green**

$H_0 : p < 1/2$

**The boss**

$H_0 : p \geq 1/2$
The choice of the null hypothesis

A test is built such that probability of "wrongly rejecting $H_0$" \( \leq \alpha = 5\% \).

Let \( p \) be the probability that an egg hatches.

\[ H_0 : p < \frac{1}{2} \]
\[ H_1 : p \geq \frac{1}{2} \]

\[ H_0 : p \geq \frac{1}{2} \]
\[ H_1 : p < \frac{1}{2} \]
The test of the green

- The total **bad faith** : $\Delta = 0$. It is of level $\alpha$!
The test of the green

- The total **bad faith** : $\Delta = 0$. It is of level $\alpha$!
  - $\rightarrow$ a powerful test (*fair play*) tries for a fixed level to minimize the probability of second kind error.
The test of the green

- The total **bad faith** : $\Delta = 0$. It is of level $\alpha$!
  - a powerful test (*fair play*) tries for a fixed level to minimize the probability of second kind error.
  - $\neq$ symmetrically minimize the probabilities of both errors.
The test of the green

- The total **bad faith** : $\Delta = 0$. It is of level $\alpha$!
  - a powerful test (*fair play*) tries for a fixed level to minimize the probability of second kind error.
  - ≠ symmetrically minimize the probabilities of both errors.

- One can show that the most powerful test in necessarily of the shape $\Delta = 1_{\bar{X} \geq c_{\alpha}^{\text{green}}}$ avec $c_{\alpha}^{\text{green}}$ tq $P_{p=1/2}(\bar{X} \geq c_{\alpha}^{\text{green}}) \sim \alpha$. 


$\text{CLT, ...}$ one can take (for $n$ very large) $c_{\alpha}^{\text{green}} = 1/2 + z_{1-\alpha} \sqrt{\bar{X}/(1-\bar{X})}$, if $\alpha = 5\%$, $z_{1-\alpha} \approx 1.64$. 

$\frac{13}{56}$
The test of the green

- The total **bad faith** : $\Delta = 0$. It is of level $\alpha$!
  - $\rightarrow$ a powerful test (*fair play*) tries for a fixed level to minimize the probability of second kind error.
  - $\neq$ symmetrically minimize the probabilities of both errors.
- One can show that the most powerful test in necessarily of the shape $\Delta = \mathbf{1}_{\bar{X} \geq c_{\alpha}^{\text{green}}}$ avec $c_{\alpha}^{\text{green}}$ tq $\mathbb{P}_{p=1/2}(\bar{X} \geq c_{\alpha}^{\text{green}}) \simeq \alpha$.
- [...] CLT, [...] one can take (for $n$ very large)

$$c_{\alpha}^{\text{green}} = \frac{1}{2} + z_{1-\alpha} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}},$$

where $z_t$ $t$-quantile of $N(0, 1)$. If $\alpha = 5\%$, $z_{1-\alpha} \simeq 1.64$
The test of the green (fair playing)

Since $c^\text{green}_\alpha > 1/2$, even the most powerful test of this null hypothesis, (i.e. with this committed stance) rejects if

$$\bar{X} > 1/2$$

and

$$\bar{X} < c^\text{green}_\alpha.$$  

In doubt, he prefers to say "Yes" to his dearie hypothesis: "The factory is polluting". Only grudgingly ($\bar{X} > c^\text{green}_\alpha$) the green who is fair playing will admit that the factory has not done anything.
The test of the boss (fair playing)

He rejects $H_0$ if $\bar{X} < c^\text{boss}_\alpha$ avec

$$c^\text{boss}_\alpha = \frac{1}{2} - z_{1-\alpha} \sqrt{\frac{\bar{X}(1 - \bar{X})}{n}}.$$ 

Since $c^\text{boss}_\alpha < 1/2$, in doubt, he prefers to say "Yes" to his dearie hypothesis : "The factory has done nothing", as long as $\bar{X}$ is not small enough. Only grudgingly ($\bar{X} < c^\text{boss}_\alpha$) the boss who is fair playing will admit that the factory has polluted the colony.
Summary of the toy example

NB: The disagreement region $\rightarrow_{n \to \infty} 0$

The boss rejects his $H_0$. He admits the pollution made by the factory.

The green accepts his $H_0$. The factory is harmless.

The green accepts his $H_0$. The factory is polluting.

The boss rejects his $H_0$. He admits the pollution made by the factory.

In doubt, both keep their respective dearie hypothesis.

Agreement

Disagreement
Moral

The test gives a scientifically sensible answer only when it rejects: *NO has more meaning than YES.*
Moral

- The test gives a scientifically sensible answer only when it rejects: *NO has more meaning than YES.*

- When it accepts, it only means "No clear evidence of contradiction between $H_0$ and the data" i.e. "$H_0$ is plausible".
Moral

- The test gives a scientifically sensible answer only when it rejects: *NO has more meaning than YES.*
- When it accepts, it only means "No clear evidence of contradiction between $H_0$ and the data" i.e. "$H_0$ is plausible".
- When it rejects, it means "The probability of seeing this under $H_0$ is very small, smaller than $\alpha$, hence $H_0$ is quite unlikely."
Moral

- The test gives a scientifically sensible answer only when it rejects: *NO has more meaning than YES.*
- When it accepts, it only means ”No clear evidence of contradiction between $H_0$ and the data” i.e. ”$H_0$ is plausible”.
- When it rejects, it means ”The probability of seeing this under $H_0$ is very small, smaller than $\alpha$, hence $H_0$ is quite unlikely.”
Moral

- The test gives a scientifically sensible answer only when it rejects: *NO has more meaning than YES.*
- When it accepts, it only means "No clear evidence of contradiction between $H_0$ and the data" i.e. "$H_0$ is plausible".
- When it rejects, it means "The probability of seeing this under $H_0$ is very small, smaller than $\alpha$, hence $H_0$ is quite unlikely."

If there is a choice, as a scientific person, we should go against our own camp: what we want to show = the alternative $H_1$. Unfortunately most of the time, there is no choice ....
p-values
Levels or p-values

Level $\alpha$ are often difficult to decide. Why always 5% ???
Levels or p-values

Level $\alpha$ are often difficult to decide. Why always 5% ? ? ?

**p-value**

The limit value of $\alpha$ which makes the test pass from acceptation to rejection.
### Levels or p-values

Level $\alpha$ are often difficult to decide. Why always 5% ? ? ?

<table>
<thead>
<tr>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The limit value of $\alpha$ which makes the test pass from acceptation to rejection.</td>
</tr>
</tbody>
</table>

The smaller, the less likely $H_0$ but a large p-value does NOT mean that $H_0$ is "very true".
Levels or p-values

Level $\alpha$ are often difficult to decide. Why always 5% ? ? ?

**p-value**

The limit value of $\alpha$ which makes the test pass from acceptance to rejection.

The smaller, the less likely $H_0$ but a large p-value does NOT mean that $H_0$ is "very true".

The p-values are uniform under $H_0$. 
A recipe for testing
A recipe to build a test

1. Decide for $H_0$ and $H_1$
A recipe to build a test

1. Decide for $H_0$ and $H_1$
2. Find a good estimate $\hat{\theta}$ of the parameter $\theta$ on which the question is based.
A recipe to build a test

1. Decide for $H_0$ and $H_1$
2. Find a good estimate $\hat{\theta}$ of the parameter $\theta$ on which the question is based.
3. Looking at $H_1$, decide qualitatively when you want to reject (i.e. vote for $H_1$).
A recipe to build a test

1. Decide for $H_0$ and $H_1$
2. Find a good estimate $\hat{\theta}$ of the parameter $\theta$ on which the question is based.
3. Looking at $H_1$, decide qualitatively when you want to reject (i.e. vote for $H_1$).
4. Find the distribution of $\hat{\theta}$ under $H_0$.
   - If it is not possible to know it perfectly or approximately, try to estimate everything that you do not know in it, and change accordingly the test statistic until you know the distribution under $H_0$.
   - If it is still not possible, try to exchange $H_0$ and $H_1$. 

5. Thanks to this decide quantitatively when you want to reject at level $\alpha$, for any $\alpha$.
6. Transform the data into p-values thanks to this.
7. Interpret the p-value, i.e. how small it is? 

$10^{-16}$ (definitely not possible to be $H_0$), $0.001$, $0.01$, $0.05$...
A recipe to build a test

1. Decide for $H_0$ and $H_1$.
2. Find a good estimate $\hat{\theta}$ of the parameter $\theta$ on which the question is based.
3. Looking at $H_1$, decide qualitatively when you want to reject (i.e. vote for $H_1$).
4. Find the distribution of $\hat{\theta}$ under $H_0$.
   - If it is not possible to know it perfectly or approximately, try to estimate everything that you do not know in it, and change accordingly the test statistic until you know the distribution under $H_0$.
   - If it is still not possible, try to exchange $H_0$ and $H_1$.
5. Thanks to this decide quantitatively when you want to reject at level $\alpha$, for any $\alpha$. 
A recipe to build a test

1. Decide for \( H_0 \) and \( H_1 \)
2. Find a good estimate \( \hat{\theta} \) of the parameter \( \theta \) on which the question is based.
3. Looking at \( H_1 \), decide qualitatively when you want to reject (i.e. vote for \( H_1 \)).
4. Find the distribution of \( \hat{\theta} \) under \( H_0 \).
   - If it is not possible to know it perfectly or approximately, try to estimate everything that you do not know in it, and change accordingly the test statistic until you know the distribution under \( H_0 \).
   - If it is still not possible, try to exchange \( H_0 \) and \( H_1 \).
5. Thanks to this decide quantitatively when you want to reject at level \( \alpha \), for any \( \alpha \).
6. Transform the data into p-values thanks to this.
A recipe to build a test

1. Decide for $H_0$ and $H_1$.
2. Find a good estimate $\hat{\theta}$ of the parameter $\theta$ on which the question is based.
3. Looking at $H_1$, decide qualitatively when you want to reject (i.e. vote for $H_1$).
4. Find the distribution of $\hat{\theta}$ under $H_0$.
   - If it is not possible to know it perfectly or approximately, try to estimate everything that you do not know in it, and change accordingly the test statistic until you know the distribution under $H_0$.
   - If it is still not possible, try to exchange $H_0$ and $H_1$.
5. Thanks to this decide quantitatively when you want to reject at level $\alpha$, for any $\alpha$.
6. Transform the data into p-values thanks to this.
7. Interpret the p-value, i.e. how small it is? $10^{-16}$ (definitely not possible to be $H_0$), 0.001, 0.01, 0.05...
Test of "the mean = something fixed"
Test on the mean

Observe $X_1, \ldots, X_n$ of unknown mean $m$

**Example**

The $X_i$’s are the weights of people in this room, $m$ is the mean weight of the French population)
Test on the mean

Observe $X_1, \ldots, X_n$ of unknown mean $m$

**Example**

The $X_i$’s are the weights of people in this room, $m$ is the mean weight of the French population.
Are we heavier than the recommended weight (70kg) given by the health insurance?
Test on the mean

Observe $X_1, \ldots, X_n$ of unknown mean $m$

Example

The $X_i$’s are the weights of people in this room, $m$ is the mean weight of the French population.
Are we heavier than the recommended weight (70kg) given by the health insurance?

**Step 1 : Find $H_0$ and $H_1$**

If you want to show that we are heavier (warning against risk):

$$H_0 : "m = 70" \text{ versus } H_1 : "m > 70"$$
Test on the mean

Step 2 : The estimate

\[ \bar{X} = \frac{X_1 + \ldots + X_n}{n} \]
Test on the mean

Step 2: The estimate

\[ \bar{X} = \frac{X_1 + \ldots + X_n}{n} \]

Step 3: The qualitative rejection

One rejects \( H_0 \) "\( m = 70 \)" if \( \bar{X} \) is too big i.e. \( \hat{X} > c \). Remains to find \( c \)
Test on the mean

Step 2 : The estimate

\[ \bar{X} = \frac{X_1 + \ldots + X_n}{n} \]

Step 3 : The qualitative rejection
One rejects \( H_0 \) " \( m = 70 \)" if \( \bar{X} \) is too big i.e. \( \hat{X} > c \). Remains to find \( c \)
Test on the mean

Step 4 : The distribution
Classically, one assumes:
$X_1, \ldots, X_n$ are independent Gaussian variables with mean $m$ and variance $\nu$ (both unknown usually) ($\mathcal{N}(m, \nu)$).
Hence $\bar{X}$ is $\mathcal{N}(m, \nu/n)$.

Problem : under $H_0$ $m = 70$ is known, but $\nu$ isn't.
One can estimate it by $\hat{\nu} = (X_1 - \bar{X} + \ldots + X_n - \bar{X}) / (n - 1)$.
This modifies the distribution:
$\sqrt{n} (\bar{X} - m) / \hat{\nu}$ is Student with $n - 1$ degrees of freedom.
Hence under $H_0$, we know the distribution of $T = \sqrt{n} (\bar{X} - 70) / \hat{\nu}$. 

$\frac{25}{56}$
Test on the mean

Step 4 : The distribution
Classically, one assumes :
$X_1, \ldots, X_n$ are independent Gaussian variables with mean $m$ and variance $\nu$ (both unknown usually) ($\mathcal{N}(m, \nu)$).
Hence $\bar{X}$ is $\mathcal{N}(m, \nu/n)$.
Problem : under $H_0$ $m = 70$ is known, but $\nu$ isn’t.
**Test on the mean**

**Step 4 : The distribution**
Classically, one assumes:
$X_1, \ldots, X_n$ are independent Gaussian variables with mean $m$ and variance $\nu$ (both unknown usually) ($\mathcal{N}(m, \nu)$).
Hence $\bar{X}$ is $\mathcal{N}(m, \nu/n)$.

Problem: under $H_0$ $m = 70$ is known, but $\nu$ isn’t.
One can estimate it by

$$\hat{\nu} = \frac{(X_1 - \bar{X}) + \ldots + (X_n - \bar{X})}{n - 1}. $$
**Test on the mean**

**Step 4 : The distribution**

Classically, one assumes:

$X_1, \ldots, X_n$ are independent Gaussian variables with mean $m$ and variance $\nu$ (both unknown usually) ($\mathcal{N}(m, \nu)$).

Hence $\bar{X}$ is $\mathcal{N}(m, \nu/n)$.

Problem: under $H_0$ $m = 70$ is known, but $\nu$ isn’t.

One can estimate it by

$$\hat{\nu} = \frac{(X_1 - \bar{X}) + \ldots + (X_n - \bar{X})}{n - 1}.$$ 

This modifies the distribution:

$$\sqrt{n} \frac{\bar{X} - m}{\sqrt{\hat{\nu}}}$$

is Student with $n - 1$ degrees of freedom.
Test on the mean

Step 4 : The distribution
Classically, one assumes:

$X_1, \ldots, X_n$ are independent Gaussian variables with mean $m$ and variance $\nu$ (both unknown usually) ($\mathcal{N}(m, \nu)$).

Hence $\bar{X}$ is $\mathcal{N}(m, \nu/n)$.

Problem : under $H_0$ $m = 70$ is known, but $\nu$ isn’t.

One can estimate it by

$$\hat{\nu} = \frac{(X_1 - \bar{X}) + \ldots + (X_n - \bar{X})}{n - 1}.$$

This modifies the distribution:

$$\sqrt{n}\frac{\bar{X} - m}{\sqrt{\hat{\nu}}}$$

is Student with $n - 1$ degrees of freedom.

Hence under $H_0$, we know the distribution of

$$T = \sqrt{n}\frac{\bar{X} - 70}{\sqrt{\hat{\nu}}}.$$
Quick illustration

Probability to be there = 5%
**Test on the mean**

**Step 5 : The quantitative rejection** Let $q_{1-\alpha}$ be the $1 - \alpha$ quantile of the Student distribution. One rejects $H_0$ at level $\alpha$ if $T > q$, that is

$$\bar{X} > c = 70 + q_{1-\alpha} \frac{\sqrt{\hat{v}}}{\sqrt{n}}.$$
Test on the mean

**Step 5 : The quantitative rejection** Let $q_{1-\alpha}$ be the $1 - \alpha$ quantile of the Student distribution.
One rejects $H_0$ at level $\alpha$ if $T > q$, that is

$$\bar{X} > c = 70 + q_{1-\alpha} \frac{\sqrt{\hat{V}}}{\sqrt{n}}.$$ 

**NB :**
- if $n$ large ($n > 120$), the Student distribution is almost $\mathcal{N}(0, 1)$. 
-
Test on the mean

**Step 5 : The quantitative rejection** Let \( q_{1-\alpha} \) be the \( 1 - \alpha \) quantile of the Student distribution.

One rejects \( H_0 \) at level \( \alpha \) if \( T > q \), that is

\[
\bar{X} > c = 70 + q_{1-\alpha} \frac{\sqrt{\hat{v}}}{\sqrt{n}}.
\]

**NB :**
- if \( n \) large (\( n > 120 \)), the Student distribution is almost \( \mathcal{N}(0, 1) \).
- if the \( X_i \)'s are not Gaussian and if \( n \) large enough, it is also true that one can do as if \( T \) is \( \mathcal{N}(0, 1) \).
Test on the mean

**Step 5 : The quantitative rejection** Let $q_{1-\alpha}$ be the $1 - \alpha$ quantile of the Student distribution. One rejects $H_0$ at level $\alpha$ if $T > q$, that is

$$\bar{X} > c = 70 + q_{1-\alpha} \frac{\sqrt{\hat{v}}}{\sqrt{n}}.$$

NB :
- if $n$ large ($n > 120$), the Student distribution is almost $\mathcal{N}(0, 1)$.
- if the $X_i$’s are not Gaussian and if $n$ large enough, it is also true that one can do as if $T$ is $\mathcal{N}(0, 1)$. Main problem, it depends on the distribution of $X_i$ : for instance if $X_i$ Bernoulli (Ex : the egg hatches or not with probability $p$), the approximation holds as long as $np \geq 5$ and $n(1 - p) \geq 5.$
Test on the mean

Step 6: p-value
It is the $\tilde{\alpha}$ such that $T = q_{1-\tilde{\alpha}}$.

Step 7: Conclusion We accept $H_0$ at level $\alpha$ if p-value $\leq \alpha$
- p-value $= 10^{-10}$: we are definitely heavier
- p-value $= 0.0023$: it is likely that we are heavier
- p-value $= 0.12$: no evidence that we are heavier
Verification that the test is OK
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(\text{p-value} \leq \alpha) \leq \alpha$
meaning that the first kind error is less than $\alpha$. 
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(\text{p-value} \leq \alpha) \leq \alpha$ meaning that the first kind error is less than $\alpha$.

If we know how to simulate under $H_0$ on a computer (at least some particular representative cases),

- Fix level $\alpha = 5\%$ and count the frequency over $N_{simu} = 5000$ that the p-value is less than $\alpha$ (ie the test rejects at level $\alpha$).
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(\text{p-value} \leq \alpha) \leq \alpha$ meaning that the first kind error is less than $\alpha$. If we know how to simulate under $H_0$ on a computer (at least some particular representative cases),

- Fix level $\alpha = 5\%$ and count the frequency over $N_{simu} = 5000$ that the p-value is less than $\alpha$ (ie the test rejects at level $\alpha$).
- If $freq \approx 0.05$ (at the third decimal), then you’re fine!
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(p\text{-value} \leq \alpha) \leq \alpha$ meaning that the first kind error is less than $\alpha$.

If we know how to simulate under $H_0$ on a computer (at least some particular representative cases),

- Fix level $\alpha = 5\%$ and count the frequency over $N_{simu} = 5000$ that the p-value is less than $\alpha$ (i.e. the test rejects at level $\alpha$).
  - If $freq \simeq 0.05$ (at the third decimal), then you’re fine!
  - If $freq << 0.05$, you’re fine too but it is likely that you are too conservative (i.e. you will not reject that much even when you have the right to do so)
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(\text{p-value} \leq \alpha) \leq \alpha$ meaning that the first kind error is less than $\alpha$.

If we know how to simulate under $H_0$ on a computer (at least some particular representative cases),

- Fix level $\alpha = 5\%$ and count the frequency over $N_{simu} = 5000$ that the p-value is less than $\alpha$ (ie the test rejects at level $\alpha$).
  - If $freq \simeq 0.05$ (at the third decimal), then you’re fine!
  - If $freq << 0.05$, you’re fine too but it is likely that you are too conservative (i.e. you will not reject that much even when you have the right to do so)
  - If $freq >> 0.05$, the level is not controlled, you cannot trust your test when it rejects.
How can we verify that nothing bad happens?

The overall procedure is OK if under $H_0$, $P(p\text{-value} \leq \alpha) \leq \alpha$
meaning that the first kind error is less than $\alpha$.
If we know how to simulate under $H_0$ on a computer (at least some particular representative cases),

- Fix level $\alpha = 5\%$ and count the frequency over $N_{simu} = 5000$
  that the p-value is less than $\alpha$ (ie the test rejects at level $\alpha$).
  - If $freq \simeq 0.05$ (at the third decimal), then you’re fine!
  - If $freq << 0.05$, you’re fine too but it is likely that you are too conservative (i.e. you will not reject that much even when you have the right to do so)
  - If $freq >> 0.05$, the level is not controlled, you cannot trust your test when it rejects.

- If you want to be good whatever the level, draw the curve

$\alpha \rightarrow \frac{\#(p\text{-value} \leq \alpha)}{N_{simu}}$,

that is the cumulative distribution function of the p-values
Example of interpretation of the cdf of the p-values

Graphs showing the cumulative distribution function (cdf) of p-values under different hypotheses $(H_0)$ and $(H_1)$.
Other tests on the mean
Variations

- $H_0 : "m \leq 70"$ against $H_1 : "m > 70"

The same test! It works ;-}
Variations

- $H_0 : "m \leq 70"$ against $H_1 : "m > 70"

The same test! It works ;-)  
Usually it is sufficient to do as if $H_0 = \text{the border}$ and not the whole interval, because the first kind error is maximal close to the border.
Variations

- $H_0 : "m \leq 70\"$ against $H_1 : "m > 70\"
  
The same test! It works ;-)
  
  Usually it is sufficient to do as if $H_0= \text{the border}$ and not the whole interval, because the first kind error is maximal close to the border.

- $H_0 : "m = 70\"$ versus $H_1 : "m \neq 70\"
  
  **Main difference : Step 3**
  
  One qualitatively rejects when $|\bar{X} - 70|$ is large.
Variations

- $H_0 : "m \leq 70"$ against $H_1 : "m > 70"
  The same test! It works ;-) 
  Usually it is sufficient to do as if $H_0 =$ the border and not the whole interval, because the first kind error is maximal close to the border.

- $H_0 : "m = 70"$ versus $H_1 : "m \neq 70"
  Main difference: Step 3
  One qualitatively rejects when $|\bar{X} - 70|$ is large.
  leads to (Step 5) rejection when $|T| > q_{1-\alpha/2}$ (only if distribution symmetric)
Variations

- \( H_0 : "m \leq 70" \) against \( H_1 : "m > 70" \)
  The same test! It works ;-)
  Usually it is sufficient to do as if \( H_0 = \) the border and not the whole interval, because the first kind error is maximal close to the border.

- \( H_0 : "m = 70" \) versus \( H_1 : "m \neq 70" \)
  **Main difference**: **Step 3**
  One qualitatively rejects when \( |\bar{X} - 70| \) is large.
  leads to (**Step 5**) rejection when \( |T| > q_{1-\alpha/2} \) (only if distribution symmetric)

- \( H_0 : "m \neq 70" \) versus \( H_1 : "m = 70" \)
  This test is **NOT** possible.
  Usually not possible to test big space against points or space of much smaller dimension (see hereafter).
Test of equality (Two sample problem)

Two samples: $X_1, ..., X_n$ with mean $m_A$ and $Y_1, ..., Y_k$ with mean $m_B$.

**Step 1:** $H_0$: "$m_A = m_B$" against $H_1$: "$m_A \neq m_B$"
Test of equality (Two sample problem)

Two samples: \(X_1, \ldots, X_n\) with mean \(m_A\) and \(Y_1, \ldots, Y_k\) with mean \(m_B\).

Step 1: \(H_0: \ "m_A = m_B" \) against \(H_1: \ "m_A \neq m_B" \)

Step 2: The parameter of interest is \(m_A - m_B\)

It is estimated by \(\bar{X} - \bar{Y}\).
Test of equality (Two sample problem)

Two samples: $X_1, ..., X_n$ with mean $m_A$ and $Y_1, ..., Y_k$ with mean $m_B$.

Step 1: $H_0: "m_A = m_B"$ against $H_1: "m_A \neq m_B"

Step 2: The parameter of interest is $m_A - m_B$
It is estimated by $\bar{X} - \bar{Y}$.

Step 3: Qualitatively, one rejects when $|\bar{X} - \bar{Y}|$ too large.
Test of equality (Two sample problem)

Two samples: $X_1, \ldots, X_n$ with mean $m_A$ and $Y_1, \ldots, Y_k$ with mean $m_B$.

Step 1: $H_0: "m_A = m_B"$ against $H_1: "m_A \neq m_B"$

Step 2: The parameter of interest is $m_A - m_B$

It is estimated by $\bar{X} - \bar{Y}$.

Step 3: Qualitatively, one rejects when $|\bar{X} - \bar{Y}|$ too large.

Step 4: Assuming $n, m$ large enough for all the Gaussian approximations to hold,

$$\bar{X} - \bar{Y} \simeq \mathcal{N} \left( m_A - m_B, \frac{\nu_A}{n} + \frac{\nu_B}{m} \right)$$

where $\nu_A$ and $\nu_B$ are the respective variance.
Test of equality (Two sample problem)

Two samples: \( X_1, \ldots, X_n \) with mean \( m_A \) and \( Y_1, \ldots, Y_k \) with mean \( m_B \).

**Step 1**: \( H_0 : \ "m_A = m_B" \) against \( H_1 : \ "m_A \neq m_B" \)

**Step 2**: The parameter of interest is \( m_A - m_B \)
It is estimated by \( \bar{X} - \bar{Y} \).

**Step 3**: Qualitatively, one rejects when \( |\bar{X} - \bar{Y}| \) too large.

**Step 4**: Assuming \( n, m \) large enough for all the Gaussian approximations to hold, under \( H_0 \),

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{v}_A}{n} + \frac{\hat{v}_B}{m}}} \sim \mathcal{N}(0, 1).
\]

**Step 5** Rejection if \( |T| > q_{1-\alpha/2} \)
Test of equality (Two sample problem)

Two samples: $X_1, ..., X_n$ with mean $m_A$ and $Y_1, ..., Y_k$ with mean $m_B$.

**Step 1**: $H_0$: "$m_A = m_B$" against $H_1$: "$m_A \neq m_B$"

**Step 2**: The parameter of interest is $m_A - m_B$.
It is estimated by $\bar{X} - \bar{Y}$.

**Step 3**: Qualitatively, one rejects when $|\bar{X} - \bar{Y}|$ too large.

**Step 4**: Assuming $n, m$ large enough for all the Gaussian approximations to hold, under $H_0$,

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{v}_A}{n} + \frac{\hat{v}_B}{m}}} \approx \mathcal{N}(0, 1).$$

**Step 5**: Rejection if $|T| > q_{1-\alpha/2}$

Nb : Not possible to exchange $H_0$ and $H_1$!
Several tests at once:
What is the problem?
Performing several tests

- It is quite usual to ask several YES/NO questions at the same time.
Performing several tests

- It is quite usual to ask several YES/NO questions at the same time.

Examples
  - does this particular drug as any impact on this particular organ? → What if several organs? several drugs?
Performing several tests

- It is quite usual to ask several YES/NO questions at the same time.

Examples

- does this particular drug as any impact on this particular organ? → What if several organs? several drugs?
- is this gene expressed differently between healthy and sick people? (15000 genes, 100 people)
Performing several tests

It is quite usual to ask several YES/NO questions at the same time.

Examples

- does this particular drug have any impact on this particular organ? → What if several organs? several drugs?
- is this gene expressed differently between healthy and sick people? (15000 genes, 100 people)
- When is there any difference between both signals? (sliding windows, one test per window)
Performing several tests

- It is quite usual to ask several YES/NO questions at the same time.

Examples

- does this particular drug as any impact on this particular organ? → What if several organs? several drugs?
- is this gene expressed differently between healthy and sick people? (15000 genes, 100 people)
- When is there any difference between both signals? (sliding windows, one test per window)
- I don’t like this test (because it accepts), then let us use another one (or another data set!) (or several of those) until I find the conclusion that I like;-)
Performing several tests

- It is quite usual to ask several YES/NO questions at the same time.
Examples
  - does this particular drug have any impact on this particular organ? → What if several organs? several drugs?
  - is this gene expressed differently between healthy and sick people? (15000 genes, 100 people)
  - When is there any difference between both signals? (sliding windows, one test per window)
  - I don’t like this test (because it accepts), then let us use another one (or another data set!) (or several of those) until I find the conclusion that I like ;-) 

- **BE CAREFUL!!!!** one cannot perform all the tests at the same level, because errors add up.
Intuitively

**JELLY BEANS CAUSE ACNE!**

**SCIENTISTS! INVESTIGATE!**

**BUT WE'RE PLAYING MINECRAFT!**

...**FINE.**

**WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).**

xkcd
Tests that you can build by yourself

"Goodness-of-fit" tests

Independence testing

Intuitively

JELLY BEANS CAUSE ACNE!

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

THAT SETTLES THAT.

I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.

xkcd
Intuitively

xkcd
Intuitively
More mathematically

If the $K$ tests are independent, if all the null hypotheses hold and if all the first kind errors are equal to $\alpha$, then
More mathematically

If the $K$ tests are independent, if all the null hypotheses hold and if all the first kind errors are equal to $\alpha$, then

$$P(\text{no first kind error}) = (1 - \alpha)^K$$
More mathematically

If the $K$ tests are independent, if all the null hypotheses hold and if all the first kind errors are equal to $\alpha$, then

$$\mathbb{P}(\text{no first kind error}) = (1 - \alpha)^K$$
Several tests at once: How to control the mistakes?
What can we do?

a rejected test = a detection or a discovery or a positive
because only a ”NO” i.e. a rejection has a scientific meaning.
What can we do?

A rejected test = a detection or a discovery or a positive because only a "NO" i.e. a rejection has a scientific meaning.

False positive or False Discovery

A false positive happens when the corresponding test wrongly rejects.
What can we do?

a rejected test = a detection or a discovery or a positive
because only a ”NO” i.e. a rejection has a scientific meaning.

**False positive or False Discovery**

A false positive happens when the corresponding test wrongly rejects.

Even 1 false positive is ”bad” if fundamental decision based on it.
What can we do?

A rejected test = a detection or a discovery or a positive because only a "NO" i.e. a rejection has a scientific meaning.

False positive or False Discovery

A false positive happens when the corresponding test wrongly rejects.

Even 1 false positive is "bad" if fundamental decision based on it. For instance, if we change the data set until "discovery" and if there is nothing to see, after 15 tests at level 5%, we have 50% of chance to "discover" something.
What can we do?

A rejected test is a detection or a discovery or a positive because only a "NO" i.e. a rejection has a scientific meaning.

**False positive or False Discovery**

A false positive happens when the corresponding test wrongly rejects.

Even 1 false positive is "bad" if fundamental decision based on it.

**Bonferroni’s choice**

The $K$ tests are performed at level $\alpha/K$. The probability of having a false positive is then ALWAYS controlled by $\alpha$. 
What can we do?

a rejected test = a detection or a discovery or a positive
because only a "NO" i.e. a rejection has a scientific meaning.

**False positive or False Discovery**

A false positive happens when the corresponding test wrongly rejects.

Even 1 false positive is "bad" if fundamental decision based on it.

**Bonferroni’s choice**

The $K$ tests are performed at level $\alpha/K$.
The probability of having a false positive is then ALWAYS controlled by $\alpha$.

Problem: it is sometimes too conservative, the procedure generally discovers nothing.
False Discovery Rate

If no fundamental decision, but just a sorting of the interesting features,
**False Discovery Rate**

If no fundamental decision, but just a sorting of the interesting features,
→ only a control of the proportion of false positives among the discoveries of the procedure
False Discovery Rate

If no fundamental decision, but just a sorting of the interesting features,
→ only a control of the proportion of false positives among the discoveries of the procedure

**False Discovery Rate (FDR)**

\[ E \left( \frac{V}{R} \right), \]

where \( R \) nbr of discoveries (rejected tests), \( V \) nbr of false positives (wrongly rejected tests).

NB : if \( R = 0 \), the convention is \( V/R = 0 \)
False Discovery Rate

If no fundamental decision, but just a sorting of the interesting features,
→ only a control of the proportion of false positives among the discoveries of the procedure

False Discovery Rate (FDR)

\[ \mathbb{E}\left( \frac{V}{R} \right), \]

where \( R \) nbr of discoveries (rejected tests), \( V \) nbr of false positives (wrongly rejected tests).

NB : if \( R = 0 \), the convention is \( V/R = 0 \)

If all the null hypotheses hold, controlling \( FDR \leq \alpha \) amounts to

\[ P( \text{a false positive exists}) \leq \alpha. \]
Benjamini and Hochberg procedure

- The $K$ p-values of the $K$ tests:

$$p^{(1)} \leq \ldots \leq p^{(K)}$$
Benjamini and Hochberg procedure

- The $K$ p-values of the $K$ tests:
  \[ p^{(1)} \leq ... \leq p^{(K)} \]

- For a given $\alpha$ (5%), let
  \[ k = \max \left\{ \ell / p^{(\ell)} \leq \frac{\alpha}{K} \ell \right\} . \]
Benjamini and Hochberg procedure

- The $K$ p-values of the $K$ tests:

$$p^{(1)} \leq \ldots \leq p^{(K)}$$

- For a given $\alpha$ (5%), let

$$k = \max \left\{ \ell / p^{(\ell)} \leq \frac{\alpha}{K} \ell \right\}.$$ 

- The $k$ tests corresponding to the $k$ smallest p-values are rejected (discoveries).
Benjamini and Hochberg procedure

- The $K$ p-values of the $K$ tests:

$$p^{(1)} \leq \ldots \leq p^{(K)}$$

- For a given $\alpha$ (5%), let

$$k = \max \left\{ \ell / p^{(\ell)} \leq \frac{\alpha}{K} \ell \right\}.$$ 

- The $k$ tests corresponding to the $k$ smallest p-values are rejected (discoveries).

- Then (most of the time) $FDR \leq \alpha$. 
Scheme of BH procedure

\[ p_{value} = \frac{\alpha}{K} \times \text{rank} \]
Goodness-of-fit tests
Goodness-of-fit

- How can we verify that a model is good?
  Example: are the data Gaussian? Are the spike trains Poisson processes?
Goodness-of-fit

- How can we verify that a model is good?
- (...) Under the model, one can compute many distributions.
How can we verify that a model is good?

(...): Under the model, one can compute many distributions.

Hence "the model is true" = $H_0$
Goodness-of-fit

- How can we verify that a model is good?
- (...) Under the model, one can compute many distributions.
- Hence "the model is true" = $H_0$
- We take our favorite statistics and rejects if it is larger than the $1 - \alpha$ quantile of the known distribution (then p-values etc).
Goodness-of-fit

- How can we verify that a model is good?
- (…) Under the model, one can compute many distributions.
- Hence ”the model is true” = $H_0$
- We take our favorite statistics and rejects if it is larger than the $1 - \alpha$ quantile of the known distribution (then p-values etc).

Example: for Gaussianity the most powerful test is the one of Shapiro and Wilk (\texttt{shapiro.test} in R - see also Lilliefors test or other tests of the \texttt{nortest} package)
How can we verify that a model is good?

(... ) Under the model, one can compute many distributions.

Hence ”the model is true” = $H_0$

We take our favorite statistics and rejects if it is larger than the $1 - \alpha$ quantile of the known distribution (then p-values etc).

If we want to confirm the model, we want the test to accept.
Goodness-of-fit

- How can we verify that a model is good?
- (...) Under the model, one can compute many distributions.
- Hence ”the model is true” = $H_0$
- We take our favorite statistics and rejects if it is larger than the $1 - \alpha$ quantile of the known distribution (then p-values etc).
- If we want to confirm the model, we want the test to accept.
- → we can never be sure! → a plausible model but not a confirmed one.
If several models to test

- There are many models and many tests per model
If several models to test

- There are many models and many tests per model
  Example: Gaussian, Exponential, Nested models, very large models (inhomogeneous Poisson processes) etc
- To perform a correct goodness-of-fit analysis, one needs to test everything that we want and be careful about multiplicity of the tests!!!
If several models to test

- There are many models and many tests per model
  Example: Gaussian, Exponential, Nested models, very large models (inhomogeneous Poisson processes) etc
- To perform a correct goodness-of-fit analysis, one needs to test everything that we want and be careful about multiplicity of the tests!!!
- We will therefore detect unrealistic models (the discoveries). The models whose tests are accepted are just ”plausible” without being true.
If several models to test

- There are many models and many tests per model
  Example: Gaussian, Exponential, Nested models, very large models (inhomogeneous Poisson processes) etc
- To perform a correct goodness-of-fit analysis, one needs to test everything that we want and be careful about multiplicity of the tests!!!
- We will therefore detect unrealistic models (the discoveries). The models whose tests are accepted are just "plausible" without being true.
  Example: it is possible that one some data one accepts both Gaussian and Exponential model. It just means that both models do not have any particular contradiction with this particular data set.
  Providing more data should help to distinguish (except if models are nested or with non empty intersection or if the tests are particularly non powerful)
Tests of independence
### Independence

**X and Y are independent**

if for all $A$ and $B$,

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B).$$
Independence

X and Y are independent
if for all $A$ and $B$,

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B).$$

Example:
- 1/2 chance of catching a cold
- 1/2 chance to forgot my umbrella when it's raining
Independence

**X and Y are independent**

if for all $A$ and $B$,

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B).$$

Example:
- 1/2 chance of catching a cold
- 1/2 chance to forgot my umbrella when it’s raining

If I have 1/4 chance of doing both (...), it’s independent.
If each time I forgot the umbrella, I catch a cold, I have 1/2 chance of doing both : it is not independent.
Testing independence between discrete variables

Observations: \((X_i, Y_i)\) for \(i = 1, \ldots, n\) with \(X_i\) taking values in \(\{1, \ldots, r\}\) and \(Y_i\) taking values in \(\{1, \ldots, s\}\)
Testing independence between discrete variables

**Observations**: \((X_i, Y_i)\) for \(i = 1, \ldots, n\) with \(X_i\) taking values in \(\{1, \ldots, r\}\) and \(Y_i\) taking values in \(\{1, \ldots, s\}\)

**Example**: (color of the eyes, color of the hair) with value in
\(\{\text{brown, blue, green}\} \times \{\text{black, brown, blond}\}\)
Testing independence between discrete variables

**Observations**: $(X_i, Y_i)$ for $i = 1, ..., n$ with $X_i$ taking values in \{1, ..., $r$\} and $Y_i$ taking values in \{1, ..., $s$\}

**Example**: (color of the eyes, color of the hair) with value in \{brown, blue, green\} × \{black, brown, blond\}

**Estimating probabilities**:
An estimate of $P(X = j$ and $Y = k)$ is

$$\frac{\text{number of } (X_i, Y_i) = (j, k)}{n} = \frac{N_{j,k}}{n}$$
Testing independence between discrete variables

Observations: $(X_i, Y_i)$ for $i = 1, \ldots, n$ with $X_i$ taking values in \{1, \ldots, r\} and $Y_i$ taking values in \{1, \ldots, s\}

Example: (color of the eyes, color of the hair) with value in \{brown, blue, green\} $\times$ \{black, brown, blond\}

Estimating probabilities:
An estimate of $P(X = j \text{ and } Y = k)$ is

$$\frac{\text{number of } (X_i, Y_i) = (j, k)}{n} = \frac{N_{j,k}}{n}$$

An estimate of $P(X = j)$ is

$$\frac{\text{number of } X_i = j}{n} = \frac{N_{j,\cdot}}{n}$$
Testing independence between discrete variables

Observations: \((X_i, Y_i)\) for \(i = 1, ..., n\) with \(X_i\) taking values in \(\{1, ..., r\}\) and \(Y_i\) taking values in \(\{1, ..., s\}\)

Example: (color of the eyes, color of the hair) with value in \(\{\text{brown, blue, green}\} \times \{\text{black, brown, blond}\}\)

Estimating probabilities:

An estimate of \(P(X = j \text{ and } Y = k)\) is

\[
\frac{\text{number of } (X_i, Y_i) = (j, k)}{n} = \frac{N_{j,k}}{n}
\]

An estimate of \(P(X = j)\) is

\[
\frac{\text{number of } X_i = j}{n} = \frac{N_{j,\bullet}}{n}
\]

An estimate of \(P(Y = k)\) is

\[
\frac{\text{number of } Y_i = k}{n} = \frac{N_{\bullet,k}}{n}
\]
Chi-square test of independence

**Qualitatively**: One should reject

\[ H_0 : \text{independence between } X \text{ and } Y \]

versus \( H_1 : \text{they are not independent} \)
Chi-square test of independence

**Qualitatively**: One should reject

\[ H_0 : \text{independence between } X \text{ and } Y \]

versus \( H_1 : \text{they are not independent} \)

if \( \frac{N_{j,k}}{n} \) too different from \( \frac{N_{j,\bullet}N_{\bullet,k}}{n^2} \)
Chi-square test of independence

**Qualitatively**: One should reject

\[ H_0 : \text{independence between } X \text{ and } Y \]

versus \( H_1 : \text{they are not independent} \)

if \( \frac{N_{j,k}}{n} \) too different from \( \frac{N_{j,\bullet}N_{\bullet,k}}{n^2} \)

or if \( N_{j,k} \) (the observed number) too different from \( \frac{N_{j,\bullet}N_{\bullet,k}}{n} \) (the expected number under independence),
Chi-square test of independence

**Qualitatively**: One should reject

\[ \mathcal{H}_0 : \text{independence between } X \text{ and } Y \]

versus \( \mathcal{H}_1 : \text{they are not independent} \)

\[
\text{if } \frac{N_{j,k}}{n} \text{ too different from } \frac{N_{j,\star}N_{\star,k}}{n^2} \\
\text{or if } N_{j,k} \text{ (the observed number) too different from } \frac{N_{j,\star}N_{\star,k}}{n} \text{ (the expected number under independence), at least for a } j \text{ and a } k.
\]
Chi-square test of independence

Qualitatively: One should reject

\[ H_0 : \text{independence between } X \text{ and } Y \]

versus \[ H_1 : \text{they are not independent} \]

if \( \frac{N_{j,k}}{n} \) too different from \( \frac{N_{j,\cdot}N_{\cdot,k}}{n^2} \)

or if \( N_{j,k} \) (the observed number) too different from \( \frac{N_{j,\cdot}N_{\cdot,k}}{n} \) (the expected number under independence), at least for a \( j \) and a \( k \).

Quantitatively \((\text{chisq.test in R})\)

\[
T = \sum_{j=1}^{r} \sum_{k=1}^{s} \left( \frac{N_{j,k} - \frac{N_{j,\cdot}N_{\cdot,k}}{n}}{\frac{N_{j,\cdot}N_{\cdot,k}}{n}} \right)^2
\]

One rejects (approximately) at level \( \alpha \) if \( T \geq 1 - \alpha \) quantile of a chi-square distribution with \((r - 1)(s - 1)\) degrees of freedom.
Chi-square test of independence

Qualitatively: One should reject

\[ H_0 : \text{independence between } X \text{ and } Y \]
versus \( H_1 : \text{they are not independent} \)

if \( \frac{N_{j,k}}{n} \) too different from \( \frac{N_{j,\cdot}N_{\cdot,k}}{n^2} \)

or if \( N_{j,k} \) (the observed number) too different from \( \frac{N_{j,\cdot}N_{\cdot,k}}{n} \) (the expected number under independence), at least for a \( j \) and a \( k \).

Quantitatively (chisq.test in R)

\[
T = \sum_{j=1}^{r} \sum_{k=1}^{s} \left( \frac{N_{j,k} - \frac{N_{j,\cdot}N_{\cdot,k}}{n}}{\frac{N_{j,\cdot}N_{\cdot,k}}{n}} \right)^2
\]

One rejects (approximately) at level \( \alpha \) if \( T \geq 1 - \alpha \) quantile of a chi-square distribution with \((r - 1)(s - 1)\) degrees of freedom.

Valid approximation as soon as
ALL the expected numbers under independence \( \geq 5 \)
For other variables

if $X$ and $Y$ have not a finite number of possibilities

- Make some classes and perform chi-square test. The classes should at least satisfy the rule
  
  all the expected numbers per class under independence $\geq 5$
For other variables

if $X$ and $Y$ have not a finite number of possibilities

- Make some classes and perform **chi-square** test. The classes should at least satisfy the rule
  
  all the expected numbers per class under independence $\geq 5$

- Or **Kendall’s tau** : statistic based on the number of pairs $(i, i’)$ such that $X_i < X_{i’}$ and $Y_i < Y_{i’}$ (or $Y_i > Y_{i’}$) - need $X$ and $Y$ to be real.

Kendall of the Kendall package in R
For other variables

if $X$ and $Y$ have not a finite number of possibilities

- Make some classes and perform chi-square test. The classes should at least satisfy the rule
  
  all the expected numbers per class under independence $\geq 5$

- Or Kendall’s tau: statistic based on the number of pairs $(i, i')$
  such that $X_i < X_{i'}$ and $Y_i < Y_{i'}$ (or $Y_i > Y_{i'}$) - need $X$ and $Y$ to be real.
  Kendall of the Kendall package in R

- Or Correlation tests
Autocorrelation tests

One observed a time series: $X_1, ..., X_t, ... X_n$
Autocorrelation tests

One observed a time series: $X_1, ..., X_t, ... X_n$

Example: amount of raining per month, interest rates etc
Autocorrelation tests

One observed a time series: \(X_1, \ldots, X_t, \ldots X_n\)
Example: amount of raining per month, interest rates etc
Are the variables correlated or not at a given lag \(h\)?
Autocorrelation tests

One observed a time series: $X_1, ..., X_t, ... X_n$
Example: amount of raining per month, interest rates etc
Are the variables correlated or not at a given lag $h$?
NB: if they are independent, they are not correlated.
   The reverse is wrong.

Estimation autocovariance:
$c_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X})$

Autocorrelations:
$r_h = \frac{c_h}{c_0}$
Autocorrelation tests

One observed a time series: $X_1, \ldots, X_t, \ldots X_n$

Example: amount of raining per month, interest rates etc.

Are the variables correlated or not at a given lag $h$?

NB: if they are independent, they are not correlated.

The reverse is wrong.

$H_0$: No correlation at lag $h$ versus $H_1$: correlation at lag $h$.

Estimation

autocovariance: $c_h = \frac{1}{n} \sum_{t=1}^{n-h}(X_t - \bar{X})(X_{t+h} - \bar{X})$
Autocorrelation tests

One observed a time series: \( X_1, \ldots, X_t, \ldots, X_n \)
Example: amount of raining per month, interest rates etc
Are the variables correlated or not at a given lag \( h \)?
NB: if they are independent, they are not correlated.
    The reverse is wrong.

\( H_0 \): No correlation at lag \( h \) versus \( H_1 \): correlation at lag \( h \).

\textbf{Estimation}

autocovariance: \( c_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X}) \)

autocorrelations: \( r_h = \frac{c_h}{c_0} \)
Autocorrelation tests

Assuming that the $X_t$ have the same distribution if they are in addition independent, then for $h > 0$, $r_h$ is approximately $\mathcal{N}(0, 1/n)$. Hence the test for lag $h$ rejects at level $\alpha$ if $\sqrt{n}|r_h| > 1 - \alpha/2$ quantile of $\mathcal{N}(0, 1)$. Problem: Be careful!! If you do it for several $h$, it is a multiple testing problem!
Illustration

Data

No correlation

Existing correlation
The lagplot by default in R

acf(data)

No correlation, acf by default

Clear correlation, acf by default
The lagplot corrected for multiplicity

\[ \text{acf}(\text{data}, ci=(1-0.05/26)) \]

No correlation, acf with modified level

Existing correlation, acf with modified level