Tests

Jean-Yves Le Boudec
1. The Neyman Pearson framework
2. Likelihood Ratio Tests
3. Asymptotic Results
4. Other Tests
Tests

Tests are used to give a binary answer to hypotheses of a statistical nature

- Ex: is A better than B?
- Ex: does this data come from a normal distribution?
- Ex: does factor $n$ influence the result?
Example: Non Paired Data

Is red better than blue?

(a) Parameter set 1
(b) Parameter set 2
(c) Parameter set 3

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Compiler Option 0</th>
<th>Compiler Option 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$[-0.1669; 0.2148]$</td>
<td>$[0.3360; 0.7400]$</td>
</tr>
<tr>
<td>2</td>
<td>$[-0.0945; 0.3475]$</td>
<td>$[0.2575; 0.6647]$</td>
</tr>
<tr>
<td>3</td>
<td>$[-0.1150; 0.2472]$</td>
<td>$[-0.0925; 0.3477]$</td>
</tr>
</tbody>
</table>

For data set (a) answer is clear (by inspection of confidence interval) no test required
Is this data normal?

Figure 7.3: Normal qqplots of file transfer data and its logarithm.
1 The Neyman-Pearson Framework

- Given: data set $x_i$
  a model with parameter $\theta \in \Theta$ (which, we believe, explains the data)

- Two hypotheses on $\theta$
  $H_0: \theta \in \Theta_0$ (null hypothesis)
  $H_1: \theta \in \Theta \setminus \Theta_0$ (alternative hypothesis)

- Nested model: $\Theta_0$ is a set of smaller dimension than $\Theta$
Example: Non Paired Data; Is Red better than Blue?

- $H_0: F_0 = F_1$
- $H_1: F_0 \neq F_1$
- $\Theta_0 = \{(F_0, F_0), F_0 \text{is a CDF}\}$

Model: $x_i$ and $y_i$ are two independent iid samples $x_i \sim F_0$ and $y_i \sim F_1$

$\Theta = \{(F_0, F_1), F_0 \text{ and } F_1 \text{are CDFs}\}$
Example: Non Paired Data; Is Red better than Blue? ANOVA Model

- $H_0: \mu_0 = \mu_1$
- $H_1: \mu_0 \neq \mu_1$
- $\Theta_0 = \{(\mu_0, \mu_0, \sigma), \sigma > 0\}$

Model: $x_i$ and $y_i$ are two independent iid samples $x_i \sim N_{\mu_0, \sigma^2}$ and $y_i \sim N_{\mu_1, \sigma^2}$
$\Theta = \{(\mu_0, \mu_1, \sigma), \sigma > 0\}$
Critical Region, Size and Power

- **Critical Region**: as set $C$ of possible data values $(x_1, ..., x_n)$ such that

  if data $\in C$ then reject $H_0$

- **Type 1 error**: reject $H_0$ when $H_0$ is true
  - Size of a test = maximum proba of type 1 error
  - $\text{Size} = \sup_{\theta \in \Theta_0} P_\theta (\text{data} \in C)$ should be small

- **Type 2 error**: accept $H_0$ when $H_1$ is true
  - **Power function**: $\theta \in \Theta \setminus \Theta_0 \mapsto P_\theta (\text{data} \in C)$ should be large

- Neyman Pearson framework:
  - Design a test that maximizes power subject to size $\leq \alpha$ ($= 0.05$, e.g.)
Example : Paired Data -- Is A better than B?

\[ X_i = \text{Reduction in execution time} \]

Model: \( X_i \sim N_{\mu, \sigma^2}, \quad \Theta = \{(\mu, \sigma), \mu \geq 0, \sigma > 0\} \)

\( H_0: \mu = 0, \quad \Theta_0 = \{(0, \sigma), \sigma > 0\} \)

\( H_1: \mu > 0 \)

First attempt: \( C = \left\{(x_1, \ldots, x_n), \frac{x_1 + \cdots + x_n}{n} > c\right\} \) for some \( c \) to be computed from the required test size

Size of this test = \( \sup_{\theta \in \Theta_0} P(\text{data} \in C) \)

\[
= \sup_{\sigma > 0} P\left(\frac{X_1 + \cdots + X_n}{n} > c\right) = \sup_{\sigma > 0} \left\{1 - N_{0, \frac{\sigma^2}{n}}(c)\right\} = 1 \text{ !!!}
\]

This definition of the rejection region does not work!
Example: Paired Data -- Is A better than B?

**Model:** $X_i \sim N(\mu, \sigma^2)$, $\Theta = \{ (\mu, \sigma), \mu \geq 0, \sigma > 0 \}$

$H_0: \mu = 0$, $\Theta_0 = \{ (0, \sigma), \sigma > 0 \}$

$H_1: \mu > 0$

**Second attempt:** $C = \left\{ (x_1, \ldots, x_n), \frac{x_1 + \cdots + x_n}{ns_n/\sqrt{n}} > c \right\}$ for some $c$,

where $s_n^2$ is an estimator of variance

**Size of this test** = $\sup_{\theta \in \Theta_0} P(data \in C)$

$= \sup_{\sigma > 0} P \left( \frac{X_1 + \cdots + X_n}{ns_n/\sqrt{n}} > c \right) \approx \sup_{\sigma > 0} \left( 1 - N_{0,1}(c) \right) = 1 - N_{0,1}(c)$

**The distribution of** $\frac{x_1 + \cdots + x_n}{ns_n/\sqrt{n}}$ **under** $H_0$ **is independent of** $\theta \in \Theta_0$; this is called a “pivot”
What value of c should we choose for a test of size 5%?

A. 1.2816
B. 1.6449
C. 1.9600
D. 2.32632
E. None of the above
F. I don’t know
Solution

Answer B

We need to choose $c$ such that

$$1 - N_{0,1}(c) = 0.05$$
i.e. $N_{0,1}(c) = 0.95$
this gives $c = 1.6449$

(with $c = 1.9600$ we would have $N_0(c) = 0.975$ i.e. a test at level 2.5%)
Power function

$$\text{Power} = \sup_{\theta \in \Theta \setminus \Theta_0} P(\text{data} \in C)$$

Here:

$$P_{\mu, \sigma} \left( \text{data} \in C \right) =$$

$$P_{\mu, \sigma} \left( \frac{X_1 + \cdots + X_n}{ns_n/\sqrt{n}} > c \right) =$$

$$P_{\mu, \sigma} \left( \frac{X_1 + \cdots + X_n - n\mu}{ns_n/\sqrt{n}} > c - \frac{n\mu}{ns_n/\sqrt{n}} \right)$$

$$\approx 1 - N_{0,1} \left( c - \frac{\mu}{s_n/\sqrt{n}} \right) \approx 1 - N_{0,1} \left( c - \frac{\sqrt{n} \mu}{\sigma} \right)$$

Example: Paired Data -- Is A better than B?

- $X_i$: Reduction in execution time
- Model: $X_i \sim N_{\mu, \sigma^2}$, $\Theta = \{ (\mu, \sigma), \mu \geq 0, \sigma > 0 \}$
  - $H_0: \mu = 0$, $\Theta_0 = \{ (0, \sigma), \sigma > 0 \}$
  - $H_1: \mu > 0$
- Second attempt: $C = \{(x_1, \ldots, x_n), \frac{x_1 + \cdots + x_n}{ns_n/\sqrt{n}} > c\}$ for some $c$, where $s_n^2$ is an estimator or variance
- Size of this test: $= \sup_{\theta \in \Theta_0} P(\text{data} \in C)$

$$= \sup_{\sigma > 0} P \left( \frac{X_1 + \cdots + X_n}{ns_n/\sqrt{n}} > c \right) \approx \sup_{\sigma > 0} \left( 1 - N_{0,1}(c) \right) = 1 - N_{0,1}(c)$$

- The distribution of $\frac{X_1 + \cdots + X_n}{ns_n/\sqrt{n}}$ under $H_0$ is independent of $\theta \in \Theta_0$; this is called a “pivot”
σ is approximated by $s_n$ on the plot

- $μ \approx 0 \Rightarrow \text{power} \approx 0.05$ (bad but unavoidable)
  - Grey zone: for $μ \leq μ^* = 18$ power $\leq 95$
  - If true $μ$ is in grey zone, test will often declare $μ = 0$

- For data at hand: power $= 0.9997$, Proba of type 2 error $= 0.0003$
We can interpret $\mu^*$ as the *statistical significance* value of the test at size 0.05. The test is unable to distinguish values much smaller than $\mu^*$ from 0.

Ideally, the statistical significance should be matched with the physical resolution of the data.

\[ \approx 1 - N_{0.1} \left( c - \sqrt{\frac{n}{\sigma}} \right) \]
Example 4.3: Optimal Test Size, Continuation of Example 4.2. Say that we consider that a reduction in run time is negligible if it is below $\mu^*$. We want that the probability of deciding $H_0$ when the true value equal to $\mu^*$ or more is similar to the size $\alpha$, i.e. we want to balance the two types of errors. This gives the equations

$$1 - N_{0,1}(c^*) = \alpha$$

$$1 - N_{0,1}\left(c^* - \sqrt{n\frac{\mu^*}{s_n}}\right) = 1 - \alpha$$

thus

$$N_{0,1}(c^*) + N_{0,1}\left(c^* - \sqrt{n\frac{\mu^*}{s_n}}\right) = 1$$

By symmetry of the gaussian PDF around its mean, we have

$$\text{if } N_{0,1}(x) + N_{0,1}(y) = 1 \text{ then } x + y = 0$$

from where we derive

$$c^* = \sqrt{n\frac{\mu^*}{2s_n}}$$

The table below gives a few numerical examples, together with the corresponding test size $\alpha^* = 1 - N_{0,1}(c^*)$.

<table>
<thead>
<tr>
<th>resolution $\mu^*$</th>
<th>optimal threshold $c^*$</th>
<th>size $\alpha^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.97</td>
<td>0.17</td>
</tr>
<tr>
<td>20</td>
<td>1.93</td>
<td>0.02</td>
</tr>
<tr>
<td>40</td>
<td>3.87</td>
<td>5.38e-005</td>
</tr>
</tbody>
</table>

We see that if we care about validly detecting reductions in run time as small as $\mu^* = 10$ms, we should have a test size of 17% or more. In contrast, if the resolution $\mu^*$ is 20ms, then a test size of 2% is appropriate.
**p-value of a test**

- For the previous example, \( C = \left\{ \frac{x_1 + \cdots + x_n}{n^{\frac{s_n}{\sqrt{n}}}} > 1.645 \right\} = \{ T(x) > 1.645 \} \)
  
  with \( T(x) = \frac{x_1 + \cdots + x_n}{n^{\frac{s_n}{\sqrt{n}}}} \)

- The test consists in computing \( t = T(X) \) and see if \( t > 1.645 \)

- Consider \( P_{0,\sigma^2} (T(X) > t) \) where \( X \) is a hypothetical replay. It is independent of \( \sigma \) and we can plot it:

  \[ P(T(X) > t) < 0.05 \]

  saying \( t > 1.645 \) is the same as saying \( P(T(X) > t) < 0.05 \)

- P-value of test = \( p^*(x) = P(T(X) > t(x)) \)

- We reject \( H_0 \) if p-value is small
**$p$-value of a test**

For many tests, the rejection region has the form \( \{T(x) > m_0\} \), where \( x \) is the observation, \( T() \) some mapping, and \( m_0 \) is a parameter that depends on the size \( \alpha \) of the test. In Example 5.2 we can take \( T(x) = \sqrt{n \frac{\bar{y}_n}{s_n}} \).

**Definition 5.1.1.** The *$p$-value of an observation* \( x \) is

\[
p^*(x) = \sup_{\theta \in \Theta_0} \mathbb{P}(T(X) > T(x)|\theta)
\]

In this formula, \( X \) is a random variable that represents a hypothetical replication of the experiment, whereas \( x \) is the data that we have observed.

The mapping \( m \mapsto \sup_{\theta \in \Theta_0} \mathbb{P}(T(X) > m|\theta) \) is monotonic nonincreasing, and usually decreasing. Assuming the latter, we have the equivalence

\[
p^*(x) < \alpha \iff T(x) > m_0
\]

in other words, instead of comparing the test statistic \( T(x) \) against the threshold \( m_0 \), we can compare the \( p \)-value against the test size \( \alpha \):

**The test rejects** \( H_0 \) **when the** \( p \)-**value is smaller than the test size** \( \alpha \).
The interest of the $p$-value is that it gives more information than just a binary answer. It is in fact the minimum test size required to reject $H_0$. Very often, software packages return $p-$values rather than hard decisions ($H_0$ or $H_1$).

---

**Example: Continuation of Example 5.2.** The $p-$value is

$$p^* = 1 - N_{0,1} \left( \frac{\sqrt{n \mu_n}}{s_n} \right)$$

We find $p^* = 2.2476e - 007$ which is small, therefore we reject $H_0$. 

The critical region of a test has the form \( \{ T(x) > c \} \) where \( T() \) is a pivot. \( x \) is the data. \( X \) is a random vector having the same distribution as the data. The p-value is given by...

A. \( p^* = \sup_{\theta \in \Theta_0} P_{\theta}(T(X) > T(x)) \)
B. \( p^* = P_{\theta}(T(X) > T(x)) \) for one specific \( \theta \in \Theta_0 \)
C. A and B
D. None of the above
E. I don’t know
Solution

A is correct (definition of the p-value)
B is also correct because $T()$ is a pivot, i.e. its distribution is the same for all $\theta \in \Theta_0$

Answer C
2. Likelihood Ratio Test

A special case of Neyman-Pearson

A Systematic Method to define tests, of general applicability

**ASSUMPTIONS AND NOTATION** We assume the nested model setting, with \( H_0 \doteq \{ \theta \in \Theta_0 \} \) whereas \( H_1 \doteq \{ \theta \in \Theta \setminus \Theta_0 \} \). For a given statistic (random variable) \( \bar{X} \) and value \( \bar{x} \) of \( \bar{X} \), define:

\[
\begin{align*}
& l_{\bar{x}}(\theta) \doteq \ln f_{\bar{X}}(\bar{x}|\theta) \text{ where } f_{\bar{X}}(\cdot|\theta) \text{ is the probability density of the model, when the parameter is } \theta. \\
& l_{\bar{x}}(H_0) = \sup_{\theta \in \Theta_0} l_{\bar{x}}(\theta) \\
& l_{\bar{x}}(H_1) = \sup_{\theta \in \Theta} l_{\bar{x}}(\theta)
\end{align*}
\]

If \( H_0 \) is true, then, approximately, the likelihood is maximum for \( \theta \in \Theta_0 \) and thus \( l_{\bar{x}}(H_0) = l_{\bar{x}}(H_1) \). In the opposite case, the maximum likelihood is probably reached at some \( \theta \notin \Theta_0 \) and thus \( l_{\bar{x}}(H_1) > l_{\bar{x}}(H_0) \). This gives an idea for a generic family of tests:

**Definition 4.2.** The likelihood ratio test is defined by the rejection region

\[
C = \{l_{\bar{x}}(H_1) - l_{\bar{x}}(H_0) > k\}
\]

where \( k \) is chosen based on the required size of the test.

The test statistic \( l_{\bar{x}}(H_1) - l_{\bar{x}}(H_0) \) is called likelihood ratio for the two hypotheses \( H_0 \) and \( H_1 \).
Example: Paired Data
Is A better than B?

- $X_i =$ Reduction in execution time
- Model: $X_i \sim N_{\mu, \sigma^2}$
- $\Theta = \{(\mu, \sigma), \mu \geq 0, \sigma > 0\}$
- $H_0: \mu = 0$
- $\Theta_0 = \{(0, \sigma), \sigma > 0\}$
- $H_1: \mu > 0$

Let us compute the likelihood ratio test.
Example 4.5: Continuation of Example 4.2, Compiler Options. We want to test \( H_0: \mu = 0 \) against \( H_1: \mu > 0 \). The log-likelihood of an observation is

\[
l_x(\mu, \sigma) = -\frac{n}{2} \ln (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2
\]

and the likelihood ratio statistic is

\[
l_x(H_1) - l_x(H_0) = \sup_{\mu \geq 0, \sigma > 0} l_x(\mu, \sigma) - \sup_{\sigma > 0} l_x(0, \sigma) = -n \ln \frac{\hat{\sigma}_1}{\hat{\sigma}_0}
\]

with

\[
\hat{\sigma}_0^2 = \frac{1}{n} \sum_i x_i^2
\]

\[
\hat{\sigma}_1^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu}_n^+)^2
\]

\[
\hat{\mu}_n^+ = \max(\bar{x}, 0)
\]

The likelihood ratio test has a rejection region of the form \( l_x(H_1) - l_x(H_0) > k \), which is equivalent to

\[
\hat{\sigma}_1 < k \hat{\sigma}_0
\]

In other words, we reject \( H_0 \) if the estimated variance under \( H_1 \) is small. Such a test is called “Analysis of Variance”. We can simplify the definition of the rejection region by noting first that \( \hat{\sigma}_1 \leq \hat{\sigma}_0 \), and thus we must have \( k \leq 1 \). Second, if \( \bar{x} \geq 0 \) then Eq.(4.3) is equivalent to \( \sqrt{n} \frac{\bar{x}}{\hat{s}_n} > c \) for some \( c \). Third, if \( \bar{x} \leq 0 \) then Eq.(4.3) is never true. In summary, we have shown that this test is the same as the ad-hoc test developed in Example 4.2.
A Classical Test: Student Test

This test applies to a single sample of data, assumed to be normal with unknown mean and variance. It can also be applied to two paired samples, after computing the differences.

■ The model:

The model is: $X_1, \ldots, X_n \sim \text{iid } N_{\mu,\sigma^2}$ where $\mu$ and $\sigma$ are not known.

■ The hypotheses:

$H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$

where $\mu_0$ is a fixed value.
\[ l_x(H_1) - l_x(H_0) = \frac{n}{2} \ln \left( 1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_i (x_i - \bar{x})^2} \right) \]

Let \( T(x) = \sqrt{n \frac{\bar{x} - \mu_0}{\hat{\sigma}}} \) be the student statistic (Theorem 3.2.2), with \( \hat{\sigma} = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \). We can write the likelihood ratio statistic as

\[ l_x(H_1) - l_x(H_0) = \frac{n}{2} \ln \left( 1 + \frac{T(x)^2}{n-1} \right) \quad (5.4) \]

which is an increasing function of \(|T'(x)|\). The rejection region thus has the form

\[ C = \{|T(x)| > \eta\} \]

We compute \( \eta \) from the condition that the size of the test is \( \alpha \). Under \( H_0 \), \( T(X) \) has a student distribution \( t_{n-1} \) (Theorem 3.2.2). Thus

\[ \eta = t_{n-1}^{-1} \left( 1 - \frac{\alpha}{2} \right) \quad (5.5) \]

For example, for \( \alpha = 0.05 \) and \( n = 100 \), \( \eta = 1.98 \).

The \( p \)-value is

\[ p^* = 2(1 - t_{n-1}(T(x))) \quad (5.6) \]
Example : Paired Data
Is A better than B ?

- $X_i =$ Reduction in execution time
- Model: $X_i \sim N_{\mu, \sigma^2}$
  $\Theta = \{(\mu, \sigma), \sigma > 0\}$
- $H_0 : \mu = 0$
  $\Theta_0 = \{(0, \sigma), \sigma > 0\}$
- $H_1 : \mu \neq 0$

The likelihood ratio test is the Student test
$C = \{T(X) > 1.98\}$

Compare to one sided test: $C = \{T(X) > 1.645\}$
$H_1$ matters!

**Example 4.6: Paired Data.** This is a variant of Example 4.2. Consider again the reduction in run time due to a new compiler option, as given in Figure 2.7 on Page 32. We want to test whether the reduction is significant. We assume the data is iid normal and use the student test:

$H_0 : \mu = 0 \text{ against } H_1 : \mu \neq 0$

The test statistic is $T(\bar{x}) = 5.05$, larger than 1.98, so we reject $H_0$. Alternatively, we can compute the $p$-value and obtain $p^* = 1.80e - 006$, which is small, so we reject $H_0$. 
Here it is the same as a Conf. Interval

In some cases, tests can be avoided by the use of confidence intervals. This applies to matching pairs as in Example 4.2: a confidence interval for the mean can readily be obtained by Theorem 2.2. At level 0.05, the confidence interval is $[15.9, 36.2]$, so we can conclude that $\mu > 0$ (and more, we have a lower bound on $\mu$).

More generally, consider a generic model parameterized with some $\theta \in \Theta \subset \mathbb{R}$. For testing

$$\theta = \theta_0 \text{ against } H_1: \theta \neq \theta_0$$

we can take as rejection region

$$|\hat{\theta} - \theta_0| > c$$

If $\hat{\theta} \pm c$ is a confidence interval at level $1 - \alpha$, then the size of this test is precisely $\alpha$. For such cases, we do not need to use tests, since we can simply use confidence intervals as discussed in Chapter 2. However, it is not always as simple, or even possible, to reduce a test to the computation of confidence intervals, as for example with unpaired data in Example 4.1 (though it is possible to use confidence sets rather than confidence intervals).
Test versus Confidence Intervals

- If you can have a confidence interval, use it instead of a test

More generally, consider a generic model parameterized with some $\theta \in \Theta \subset \mathbb{R}$. For testing

$$\theta = \theta_0 \text{ against } H_1: \theta \neq \theta_0$$

we can take as rejection region

$$\left| \hat{\theta} - \theta_0 \right| > c$$

If $\hat{\theta} \pm c$ is a confidence interval at level $1 - \alpha$, then the size of this test is precisely $\alpha$. For such cases, we do not need a general theory of tests, since we can simply use confidence intervals as discussed in Chapter 3. However, there are many tests that cannot be put in this form.
The “Simple Goodness of Fit” Test

- Goal: test whether \( X_1, ..., X_n \), assumed to be iid, comes from distribution \( F() \)

- Model:

To compute the empirical histogram, we partition the set of values of \( X \) into \textit{bins} \( B_i \). Let \( N_i = \sum_{k=1}^{n} 1_{\{B_i\}}(X_k) \) (number of observation that fall in bin \( B_i \)) and \( q_i = \mathbb{P}\{X_1 \in B_i\} \). If the data comes from the distribution \( F() \) the distribution of \( N \) is \textit{multinomial} \( M_{n,q} \), i.e.

\[
\mathbb{P} \{N_1 = n_1, ..., N_k = n_k\} = \binom{n}{n_1!...n_k!} q_1^{n_1} ... q_k^{n_k}
\]  

(4.7)

- Hypotheses

\( H_0: \ N_i \) comes from the multinomial distribution \( M_{n,q} \)

against

\( H_1: \ N_i \) comes from a multinomial distribution \( M_{n,p} \) for some arbitrary \( p \).
Compute likelihood ratio statistic

\[ p \{ N_1 = n_1, \ldots, N_k = n_k \} = \binom{n!}{n_1! \ldots n_k!} q_1^{n_1} \ldots q_k^{n_k} \]

- Likelihood of an observation is
  \[ l_x(\vec{p}) = C + \sum_{i=1}^{k} n_i \ln(p_i) \]

- Under \( H_0 \): \( l_{H_0} = \sup l_x(p) = l_x(q) = C + \sum_i n_i \log q_i \)

- Under \( H_1 \): maximize \( \sum_i n_i \log p_i \) over \( p \) s.t. \( p_i \geq 0 \) and \( \sum_i p_i = 1 \)
  
  We find (with Lagrange’s technique) \( \hat{p}_i = \frac{n_i}{n} \)
  
  \[ l_{H_1} = C + \sum_i n_i \log \frac{n_i}{n} \]

- Likelihood ratio is \( \sum_i n_i \log \frac{n_i}{nq_i} \)
Compute $p$-value

We now compute the $p$-value. It is equal to

$$
P \left( \sum_{i=1}^{k} N_i \ln \frac{N_i}{n q_i} > \sum_{i=1}^{k} n_i \ln \frac{n_i}{n q_i} \right)
$$

where $\vec{N}$ has the multinomial distribution $M_{n,\vec{q}}$.

- How can we compute the p-value?
  - No exact closed form -> Monte Carlo
  - An approximate exists for large $n$ (see later)
**Mendel’s Peas**

**Example 6.2** In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions based on genetic theory.

<table>
<thead>
<tr>
<th>type</th>
<th>observed count</th>
<th>predicted frequency</th>
<th>expected count</th>
</tr>
</thead>
<tbody>
<tr>
<td>smooth yellow</td>
<td>315</td>
<td>9/16</td>
<td>312.75</td>
</tr>
<tr>
<td>smooth green</td>
<td>108</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>wrinkled yellow</td>
<td>102</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>wrinkled green</td>
<td>31</td>
<td>1/16</td>
<td>34.75</td>
</tr>
</tbody>
</table>

Is there any evidence in this data to reject the hypothesis that theory is correct?

- \( p \text{-value} = 0.92 \pm 0.05 \Rightarrow \text{Accept } H_0 \)
We design a likelihood ratio test with $H_0: \theta = \theta_0$, $H_1: \theta \in \Theta \setminus \{\theta_0\}$. The data is $x$. The p-value is equal to …

A. $P_{\theta_0}(lrs(X) > lrs(x))$ where $X$ is a replay experiment

B. $P_{\theta_0}(lrs(X) < lrs(x))$ where $X$ is a replay experiment

C. $\sup_{\theta \in \Theta} P_{\theta_0}(lrs(X) > lrs(x))$ where $X$ is a replay experiment

D. None of the above

E. I don’t know
Solution

The critical region is \( \{lrs(x) > c\} \) for some \( c \). Therefore the definition of a p-value applies and we obtain Answer A
We design a likelihood ratio test with $H_0: \theta = \theta_0$, $H_1: \theta \in \Theta \setminus \{\theta_0\}$. We evaluate the $p$-value of a test using Monte-Carlo simulation with $R=400$ runs. We find $p=0$. At size 0.05...

A. We reject $H_0$
B. We accept $H_0$
C. None of the above
D. I don’t know
Solution

We evaluate $p^* = P_{\theta_0}(lrs(X) > c)$ by simulation, where $c = lrs(x)$. We simulate 400 replicates of $X$ and obtain 0, which means that the 400 simulations produced $lrs(X^r) \leq c, r = 1: 400$

The problem is that of estimating a failure probability $p^*$ out of 100 trials. We have the 95% confidence interval: $p^* \in \left[0; \frac{3.689}{400}\right]$.

We can say that $p^* < 0.05$ with great certainty and therefore reject $H_0$. 
3. Asymptotic Result

- Applicable to a likelihood ratio test when central limit theorem holds, i.e. when distributions have finite variances and $n$ is large.
- If applicable, radically simple:
  - Compute likelihood ratio statistic $lrs$
    - This is equivalent to 2 optimization subproblems:
      1. find $\hat{\theta} \in \Theta$ that maximizes the likelihood $l_x(\theta)$
      2. find $\hat{\theta} \in \Theta_0$ that maximizes the likelihood $l_x(\theta)$

\[ lrs = l_x(\hat{\theta}) - l_x(\hat{\theta}_0) \]

- Inspect and find the order $p$ (nb of dimensions that H1 adds to H0)
- The p-value is $p^* \approx 1 - \chi_p^2(2 \ lrs)$
Theorem 4.3. [32] Consider a likelihood ratio test (Section 4.2) with $\Theta = \Theta_1 \times \Theta_2$, where $\Theta_1, \Theta_2$ are open subsets of $\mathbb{R}^{q_1}, \mathbb{R}^{q_2}$ and denote $\theta = (\theta_1, \theta_2)$. Consider the likelihood ratio test of $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 \neq 0$. Assume that the conditions in Definition B.1 hold. Then, approximately, for large sample sizes, under $H_0$, $2\text{lr}s \sim \chi^2_{q_2}$, where $\text{lr}s$ is the likelihood ratio statistic.

It follows that the p-value of the likelihood ratio test can be approximated for large sample sizes by

$$p^* \approx 1 - \chi^2_{q_2} \left(2\text{lr}s\right)$$

(4.25)

where $q_2$ is the number of degrees of freedom that $H_1$ adds to $H_0$. 

- For a likelihood ratio test, the likelihood ratio statistic is an approximate pivot
Example

Example: Non Paired Data. (Continuation of Example 7.1 on page 144) Consider the data for one parameter set. The model is

\[ X_i = \mu_1 + \epsilon_{1,i} \quad Y_j = \mu_2 + \epsilon_{2,j} \quad (7.15) \]

with \( \epsilon_{i,j} \sim \text{iid } N_0,\sigma^2 \).

For each parameter set, we want to test

\[ H_0: \mu_1 = \mu_2 \]
\[ H_1: \mu_1 \neq \mu_2 \]
Estimate $\mu_1 = \mu_2$ under $H_0$ and compute the likelihood; this is a linear regression model.

Similarly, we can estimate the $\mu_1, \mu_2$ under $H_1$ and compute the likelihood.

We find $lrs = \frac{n}{2} \log \left( \frac{SS0}{SS1} \right)$

with $SS0 = \ell^2$ norm of residuals under $H_0$ and $SS1 = \ell^2$ norm of residuals under $H_1$ (See section ANOVA for details).

The order $p$ is the number of degrees of freedom added to $H_0$

$$p = 3 - 2 = 1$$

$p$-value $p^* \approx 1 - \chi^2_1(2 \ lrs)$
Approximate p-values

The corresponding p-values are:

Parameter Set 1 $p_{\text{chi2}} = 0.0002854$
Parameter Set 1 $p_{\text{chi2}} = 0.02731$
Parameter Set 1 $p_{\text{chi2}} = 0.6669$

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns</td>
<td>13.2120</td>
<td>1</td>
<td>13.2120</td>
<td>13.4705</td>
<td>0.0003116</td>
</tr>
<tr>
<td>Errors</td>
<td>194.2003</td>
<td>198</td>
<td>0.9808</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>207.4123</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Set 2</td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>Prob&gt;F</td>
</tr>
<tr>
<td>Columns</td>
<td>5.5975</td>
<td>1</td>
<td>5.5975</td>
<td>4.8813</td>
<td>0.0283</td>
</tr>
<tr>
<td>Errors</td>
<td>227.0525</td>
<td>198</td>
<td>1.1467</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>232.6500</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Set 3</td>
<td>SS</td>
<td>df</td>
<td>MS</td>
<td>F</td>
<td>Prob&gt;F</td>
</tr>
<tr>
<td>Columns</td>
<td>0.1892</td>
<td>1</td>
<td>0.1892</td>
<td>0.1835</td>
<td>0.6689</td>
</tr>
<tr>
<td>Errors</td>
<td>204.2256</td>
<td>198</td>
<td>1.0314</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>204.4148</td>
<td>199</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: ANOVA Tests for Example 7.1 on page 142 (Non Paired Data)
For which parameter sets do we conclude that there is a significant difference?

A. 1  
B. 2  
C. 3  
D. 1 and 2  
E. 1 and 3  
F. 2 and 3  
G. All  
H. None  
I. I don’t know
Solution

The test rejects $H_0$ when $p$ is small.

$H_0$ means: there is no difference.

At size 0.05, we reject $H_0$ for parameter sets 1 and 2.

We say that there is a significant difference for parameter sets 1 and 2.
For non paired data, we cannot simply compute the difference.

However CI is sufficient for parameter set 1.

Tests disambiguate parameter sets 2 and 3.

<table>
<thead>
<tr>
<th>Parameter Set</th>
<th>Compiler Option 0</th>
<th>Compiler Option 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$[-0.1669; 0.2148]$</td>
<td>$[0.3360; 0.7400]$</td>
</tr>
<tr>
<td>2</td>
<td>$[-0.0945; 0.3475]$</td>
<td>$[0.2575; 0.6647]$</td>
</tr>
<tr>
<td>3</td>
<td>$[-0.1150; 0.2472]$</td>
<td>$[-0.0925; 0.3477]$</td>
</tr>
</tbody>
</table>

Table 7.1: ANOVA Tests for Example 7.1 on page 142 (Non Paired Data)
The chi-square distribution

14.1.5 Chi-Square

$\chi^2_n$ is the distribution of the sum of the squares of $n$ independent random variables with distribution $N_{0,1}$. Expectation: $n$; Variance: $2n$

<table>
<thead>
<tr>
<th>$n$</th>
<th>0.99</th>
<th>0.975</th>
<th>0.95</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.63</td>
<td>5.02</td>
<td>3.84</td>
<td>2.71</td>
</tr>
<tr>
<td>2</td>
<td>9.21</td>
<td>7.38</td>
<td>5.99</td>
<td>4.61</td>
</tr>
<tr>
<td>3</td>
<td>11.34</td>
<td>9.35</td>
<td>7.81</td>
<td>6.25</td>
</tr>
<tr>
<td>4</td>
<td>13.28</td>
<td>11.14</td>
<td>9.49</td>
<td>7.78</td>
</tr>
<tr>
<td>5</td>
<td>15.09</td>
<td>12.83</td>
<td>11.07</td>
<td>9.24</td>
</tr>
<tr>
<td>6</td>
<td>16.81</td>
<td>14.45</td>
<td>12.59</td>
<td>10.64</td>
</tr>
<tr>
<td>7</td>
<td>18.48</td>
<td>16.01</td>
<td>14.07</td>
<td>12.02</td>
</tr>
<tr>
<td>8</td>
<td>20.09</td>
<td>17.53</td>
<td>15.51</td>
<td>13.36</td>
</tr>
<tr>
<td>9</td>
<td>21.67</td>
<td>19.02</td>
<td>16.92</td>
<td>14.68</td>
</tr>
<tr>
<td>10</td>
<td>23.21</td>
<td>20.45</td>
<td>18.31</td>
<td>15.96</td>
</tr>
</tbody>
</table>
We want to test the hypothesis that an iid sample has a distribution that comes from a given parametric family.

**Composite Goodness of Fit** Similar to Section 4.2.3, assume we are given \( n \) data points \( x_1, ..., x_n \), generated from an iid sequence, and we want to verify whether their common distribution comes from a given family of distributions \( F(\mid \theta) \) where the parameter \( \theta \) is in some set \( \Theta_0 \). We say that the test is composite because the null hypothesis has several possible values of \( \theta \). We compare the empirical histograms: we partition the set of values of \( \bar{X} \) into bins \( B_i, i = 1...I \). Let

\[
N_i = \sum_{k=1}^{n} 1\{B_i\}(X_k) \text{ (number of observation that fall in bin } B_i \text{)} \quad \text{and} \quad q_i = \mathbb{P}_\theta\{X_1 \in B_i\}.
\]

If the data comes from a distribution \( F(\mid \theta) \) the distribution of \( N_i \) is multinomial \( M_{n,q(\theta)} \). The likelihood ratio statistic test is

\[
H_0: \ N_i \text{ comes from a multinomial distribution } M_{n,q(\theta)}, \text{ with } \theta \in \Theta_0
\]

against

\[
H_1: \ N_i \text{ comes from a multinomial distribution } M_{n,\bar{p}} \text{ for some arbitrary } \bar{p}.
\]
Likelihood of an observation is

Under $H_0$: $l_{H_0} = \sup l_x(q(\theta)) = l_x(q(\hat{\theta})) = C + \sum_i n_i \log q_i(\hat{\theta})$

Under $H_1$: $l_{H_1} = C + \sum_i n_i \log \frac{n_i}{n}$

Likelihood ratio is $l_{RS} = \sum_i n_i \log \frac{n_i}{nq_i(\hat{\theta})}$

The $p$-value is

$$\sup_{\theta \in \Theta_0} \mathbb{P} \left( \sum_{i=1}^{k} N_i \ln \frac{N_i}{nq_i} > \sum_{i=1}^{k} n_i \ln \frac{n_i}{nq_i(\hat{\theta})} \right)$$

Exact computation of $p$-value is hard, even with Monte Carlo, because we need to consider all possible cases in $H_0$

We use the asymptotic result instead: $p^* \approx 1 - \chi^2_p(2 \cdot l_{RS})$
$p^* \approx 1 - \chi^2_p(2 \text{ lrs})$ : what is the order $p$?

$k_0 = \dim \Theta_0$

$I = \text{nb bins}$

$H_0$: $N_i$ comes from a multinomial distribution $M_{n,q(\theta)}$, with $\theta \in \Theta_0$ against

$H_1$: $N_i$ comes from a multinomial distribution $M_{n,p}$ for some arbitrary $p$.

A. 1

B. I

C. $I - k_0 - 1$

D. $I - k_0$

E. $I - k_0 + 1$

F. $2(I - k_0) - 1$

G. $2(I - k_0)$

H. $2(I - k_0) + 1$

I. I don’t know
Solution

The dimension of $\Theta_0$ is $k_0$
The dimension of $\Theta$ is $I - 1$ since we need to pick $I$ numbers $p_i$
with the constraint $p_I = 1 - p_1 - \cdots - p_{I-1}$
$p = (I - 1) - k_0$

Answer C
Is it normal?

**Example:** Impact of estimation of \((\mu, \sigma)\). We want to test whether the data set on the right of Figure 7.3 has a normal distribution. We use a histogram with 10 bins. We need first to estimate \(\hat{\theta} = (\hat{\mu}, \hat{\sigma})\).

1. Assume we do this by fitting a line to the qqplot. We obtain \(\hat{\mu} = -0.2652, \hat{\sigma} = 0.8709\). The values of \(nq_i(\hat{\theta})\) and \(n_i\) are:

| \(7.9297\) | \(7.0000\) |
| \(11.4034\) | \(9.0000\) |
| \(18.0564\) | \(17.0000\) |
| \(21.4172\) | \(21.0000\) |
| \(19.0305\) | \(14.0000\) |
| \(12.6672\) | \(17.0000\) |
| \(6.3156\) | \(6.0000\) |
| \(2.3583\) | \(4.0000\) |
| \(0.6594\) | \(3.0000\) |
| \(0.1624\) | \(2.0000\) |

The likelihood ratio statistic as in Equation (7.24) is \(l_{rs} = 7.6352\). The \(p\)-value is obtained using a \(\chi^2\) distribution \((m = 10 - 2 - 1)\): \(p_1 = 0.0327\), thus we would reject normality at size \(0.05\).
2. It is not correct to simply fit \((\mu, \sigma)\) on the qqplot. The theory says that we should find \((\mu, \sigma)\) that maximizes the log likelihood of the model. This is equivalent to minimizing the likelihood ratio statistic \(l_{H_1}(x) - l_{\mu,\sigma}(x)\) (note that the value of \(l_{H_1}(x)\) is easy to compute). We do this with a numerical optimization procedure and find now \(\hat{\mu} = -0.0725, \hat{\sigma} = 1.0269\). The corresponding values of \(nq_i(\hat{\theta})\) and \(n_i\) are now:

\[
\begin{align*}
8.3309 & \quad 7.0000 \\
9.5028 & \quad 9.0000 \\
14.4317 & \quad 17.0000 \\
17.7801 & \quad 21.0000 \\
17.7709 & \quad 14.0000 \\
14.4093 & \quad 17.0000 \\
9.4783 & \quad 6.0000 \\
5.0577 & \quad 4.0000 \\
2.1892 & \quad 3.0000 \\
1.0491 & \quad 2.0000
\end{align*}
\]

Note how the true value of \(\hat{\mu}, \hat{\sigma}\) provides a better fit to the tail of the histogram. The likelihood ratio statistic is now \(lrs = 2.5973\), which also shows a much better fit. The \(p\)-value, obtained using a \(\chi^2_i\) distribution is now \(p1 = 0.6362\), thus we accept that the data is normal.
3. Assume we would ignore that \((\mu, \sigma)\) is estimated from the data, but would do as if the test were a simple goodness of fit test, with \(H_0: \text{"The distribution is } N_{-0.0725, 1.0269}\) instead of \(H_0: \text{"The distribution is normal"}\). We would compute the \(p\) value using a \(\chi^2_9\) distribution \((m = 10 - 1)\) and would obtain: \(p_2 = 0.8170\), a value larger than the true \(p\)-value. This is quite general: if we estimate some parameter and pretend it is a priori known, then we overestimate the \(p\)-value.
Mendel’s Peas

Example 6.2 In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions based on genetic theory.

<table>
<thead>
<tr>
<th>type</th>
<th>observed count</th>
<th>predicted frequency</th>
<th>expected count</th>
</tr>
</thead>
<tbody>
<tr>
<td>smooth yellow</td>
<td>315</td>
<td>9/16</td>
<td>312.75</td>
</tr>
<tr>
<td>smooth green</td>
<td>108</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>wrinkled yellow</td>
<td>102</td>
<td>3/16</td>
<td>104.25</td>
</tr>
<tr>
<td>wrinkled green</td>
<td>31</td>
<td>1/16</td>
<td>34.75</td>
</tr>
</tbody>
</table>

Is there any evidence in this data to reject the hypothesis that theory is correct?

- $P = 0.92 \pm 0.05 \Rightarrow \text{Accept } H_0$

Example: Mendel’s peas, continuation of Example 7.4 on page 152. The likelihood ratio statistic is $lrs = 0.3092$ and we found by Monte Carlo a $p$-value $p^* = 0.9191 \pm 0.0458$. By the asymptotic result, we can approximate the $p$-value by $\chi^2_3(2lrs) = 0.8922$.

The Pearson Chi-squared statistic is $pcs = 0.6043$, very close to $2lrs = 0.618$. The corresponding $p$ value is 0.8954.
Test of Independence

■ Model

A sequence \((x_k, y_k)\), which we interpret as a sample of the sequence \((X_k, Y_k), k = 1, \ldots, n\). The sequence is iid \(((X_k, Y_k)\) is independent of \((X_k, Y_k')\) and has the same distribution). We are interested in knowing whether \(X_k\) is independent of \(Y_k\).

■ Hypotheses

To this end, we compute an empirical histogram of \((X, Y)\), as follows. We partition the set of values of \(X\) [resp. \(Y\)] into \(I\) [resp. \(J\)] bins \(B_i\) [resp. \(C_j\)]. Let \(N_{i,j} = \sum_{k=1}^{n} 1_{B_i}(X_k)1_{C_j}(Y_k)\) (number of observation that fall in bin \((B_i, C_j)\)) and \(p_{i,j} = \mathbb{P}\{X_1 \in B_i \text{ and } Y_1 \in C_j\}\). The distribution of \(N\) is multinomial. The test of independence is

\[
H_0: \text{“}p_{i,j} = q_ir_j\text{ for some }q \text{ and } r\text{ such that }\sum_i q_i = \sum_j r_j = 1\text{”}
\]

against

\[
H_1: \text{“}p_{i,j}\text{ is arbitrary”}
\]
Apply the generic method

The maximum likelihood estimator under $H_0$ is \( \hat{p}_{i,j}^0 = \frac{n_i}{n} \frac{n_j}{n} \) where \( n_{i,j} = \sum_{k=1}^{n} 1_{\{B_i\}}(x_k)1_{\{C_j\}}(y_k) \)

\[
\begin{aligned}
&\quad \begin{cases}
 n_i = \sum_j n_{i,j} \\
 n_j = \sum_i n_{i,j}
\end{cases}
\end{aligned}
\] (4.32)

The maximum likelihood estimator under $H_1$ is \( \hat{p}_{i,j}^1 = \frac{n_{i,j}}{n} \). The likelihood ratio statistic is thus

\[
lrs = \sum_{i,j} n_{i,j} \ln \left( \frac{n n_{i,j}}{n_i n_j} \right)
\] (4.33)

To compute the $p$-value, we use, for large $n$, a $\chi^2_{q_2}$ distribution. The numbers of degrees of freedom under $H_1$ is $IJ - 1$, under $H_0$ it is $(I - 1) + (J - 1)$, thus $q_2 = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$. The $p$-value is thus

\[
p^* = \left(1 - \chi^2_{(I-1)(J-1)}\right)(2lrs)
\] (4.34)
**Example 4.8:** *Brassica Oleracea Gemmifera.* A survey was conducted at the campus cafeteria, where customers were asked whether they like Brussels sprouts. The answers are:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likes</td>
<td>454 (44.69%)</td>
<td>251 (48.08%)</td>
<td>705 (45.84%)</td>
</tr>
<tr>
<td>Dislikes</td>
<td>295 (29.04%)</td>
<td>123 (23.56%)</td>
<td>418 (27.18%)</td>
</tr>
<tr>
<td>No Answer / Neutral</td>
<td>267 (26.28%)</td>
<td>148 (28.35%)</td>
<td>415 (26.98%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1016 (100%)</td>
<td>522 (100%)</td>
<td>1538 (100%)</td>
</tr>
</tbody>
</table>

We would like to test whether affinity to Brussels sprouts is independent of customer’s gender. Here we have $I = 3$ and $J = 2$, so we use a $\chi^2$ distribution with $q_2 = 2$ degrees of freedom. The likelihood ratio statistic and the $p$-value are

$$lrs = 2.6489, \quad p = 0.0707$$  \hspace{1cm} (4.36)

so we accept $H_0$, i.e. affinity to Brussels sprouts is independent of gender. Note that the Pearson Chi-squared statistic is

$$pcs = 5.2178$$  \hspace{1cm} (4.37)

which is very close to $2lrs$. 

---

58
4 Other Tests

Simple Goodness of Fit

- Model: iid data
- Hypotheses: $H_0$ common distrib has cdf $F()$
  $H_1$ common distrib is anything

Define the empirical distribution $F$ by

$$\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$$

- Kolmogorov-Smirnov: under $H_0$, the distribution of

$$T = \sup_{x} |\hat{F}(x) - F(x)|$$

is independent of $F()$
That the distribution of this random variable is independent of \( F \) is not entirely obvious, but can be derived easily in the case where \( F \) is continuous and strictly increasing, as follows. The idea is to change the scale on the \( x \)-axis by \( u = F(x) \). Formally, define

\[
U_i = F(X_i)
\]

so that \( U_i \sim U(0, 1) \). Also

\[
\hat{F}(x) = \frac{1}{n} \sum_i 1\{X_i \leq x\} = \frac{1}{n} \sum_i 1\{U_i \leq F(x)\} = \hat{G}(F(x))
\]

where \( \hat{G} \) is the empirical distribution of the sample \( U_i, i = 1, ..., n \). By the change of variable \( u = F(x) \), it comes

\[
T = \sup_{u \in [0,1]} |\hat{G}(u) - u|
\]

which shows that the distribution of \( T \) is independent of \( F \). Its distribution is tabulated in statistical software packages. For a large \( n \), its tail can be approximated by \( \tau \approx \sqrt{-\ln \alpha / 2} \) where \( \mathbb{P}(T > \tau) = \alpha \). 
Anderson-Darling

An alternative to K-S, less sensitive to “outliers”

\[ A = n \int_{\mathbb{R}} \frac{\left( \hat{F}(x) - F(x) \right)^2}{F(x)(1 - F(x))} dF(x) \]

The test is similar to K-S but is less sensitive to outliers.
Original Data
slope = 0.8155
intercept = 1.0421

Transformed Data
slope = 0.8709
intercept = -0.2652

Figure 4.4: Normal qqplots of file transfer data and its logarithm.

Original Data
h = 1
p = 0.0000

Transformed Data
h = 0
p = 0.2964
Thus the test rejects the normality assumption for the original data and accepts it for the transformed data.

This way of doing is approximate in that we used estimated parameters for $H_0$. This introduces some bias, similar to using the normal statistic instead of student when we have a normal sample. The bias should be small when the data sample is large, which is the case here.

A fix to this problem is to use a variant of KS, for example the Lilliefors, or to use different normality tests such as Jarque Bera (see Example 6.1 on page 127) or Shapiro-Wilk. The Lilliefors test is a heuristic that corrects the $p$-value of the KS to account for the uncertainty due to estimation. In this specific example, with the Lilliefors test we obtain the same results as previously.
Jarque Bera test of normality (Chapter 4)

Based on Kurtosis and Skewness

Should be 0 for normal distribution

**Jarque-Bera.** The *Jarque-Bera* statistic is used to test whether an iid sample comes from a normal distribution. It is equal to \( \frac{n}{6} \left( \hat{\gamma}_1^2 + \frac{\hat{\gamma}_2^2}{4} \right) \), the distribution of which is asymptotically \( \chi_2^2 \) for large sample size \( n \). In the formula, \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \) are the sample indices of skewness and kurtosis, obtained by replacing expectations by sample averages in Equation (6.3).

\[
\begin{align*}
\kappa_1 &= \mathbb{E}(X) \\
\kappa_2 &= \mathbb{E}(X - \mathbb{E}(X))^2 = \text{var}(X) \\
\kappa_3 &= \mathbb{E}(X - \mathbb{E}(X))^3 \\
\kappa_4 &= \mathbb{E}(X - \mathbb{E}(X))^4 - 3\text{var}(X)^2
\end{align*}
\]
Figure 4.4: Normal qqplots of file transfer data and its logarithm.

**Example 4.18: Application to Example 4.17.** We would like to test whether the data in Example 4.17 and its transform are normal.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Critical Value $h$</th>
<th>Probability $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Data</td>
<td>$h = 1$</td>
<td>$p = 0.0010$</td>
</tr>
<tr>
<td>Transformed Data</td>
<td>$h = 0$</td>
<td>$p = 0.1913$</td>
</tr>
</tbody>
</table>

The conclusions are the same as in Example 4.17, but for the original data the normality assumption is clearly rejected, whereas it was borderline in Example 4.17.
Robust Tests
Median Test

- Model: iid sample
- Hypotheses

\[ H_0: \text{“the median of } F \text{ is 0”} \text{ against } H_1: \text{“unspecified”} \]

A simple test is based on confidence interval, as mentioned in Section 4.1.4. Let \( I(\bar{x}) \) be a confidence interval for the median (Theorem 2.1). We reject \( H_0 \) if

\[
0 \not\in I(\bar{x})
\]  

(4.39)

This test is robust in the sense that it makes no assumption other than independence.
**Example 4.19: Paired Data.** This is a variant of Example 4.2. Consider again the reduction in run time due to a new compiler option, as given in Figure 2.7 on Page 32. We want to test whether the reduction is significant. We assume the data is iid, but not necessarily normal. The median test gives a confidence interval

\[ I(\bar{x}) = [2.9127; 33.7597] \]

which does not contain 0 so we reject \( H_0 \).
Alternatively, let us use the Wilcoxon Signed Rank test. We obtain the $p$-value

$$p = 2.3103e - 005$$

and thus this test also rejects $H_0$. 
Wilcoxon Rank Sum Test

- Model: $X_i$ and $Y_j$ independent samples, each is iid
- Hypotheses:
  - $H_0$: both have same distribution
  - $H_1$: the distributions differ by a location shift

Let $X_i^1$, $i = 1...n_1$ and $X_i^2$, $i = 1...n_2$ be the two iid sequences that the data is assumed to be a sample of. The **Wilcoxon Rank Sum Statistic** $R$ is the sum of the ranks of the first sample in the concatenated sample.

As for the Wilcoxon signed rank test, its distribution under the null hypothesis depends only on the sample sizes and can be tabulated or, for a large sample size, approximated by a normal distribution. The mean and variance under $H_0$ are

\[
m_{n_1,n_2} = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{4.41}
\]
\[
v_{n_1,n_2} = \frac{n_1n_2(n_1 + n_2 + 1)}{12} \tag{4.42}
\]

We reject $H_0$ when the rank sum statistic deviates largely from its expectation under $H_0$. For large $n_1$ and $n_2$, the $p$-value is

\[
p = 2 \left(1 - N_{0,1} \left( \frac{|R - m_{n_1,n_2}|}{\sqrt{v_{n_1,n_2}}} \right) \right) \tag{4.43}
\]
EXAMPLE 4.20: **Non Paired Data.** The Wilcoxon rank sum test applied to Example 4.1 gives the following $p$-values:

Parameter Set 1 $p = 0.0002854$
Parameter Set 2 $p = 0.02731$
Parameter Set 3 $p = 0.6669$

The results are the same as with ANOVA. $H_0$ (same distribution) is accepted for the 3rd data set only, at size= 0.05.

The **Kruskal-Wallis** test is a generalization of Wilcoxon Rank Sum to more than 2 non paired data series. It tests ($H_0$): the samples come from the same distribution against ($H_1$): the distributions may differ by a location shift.
This is a test of iid-ness. It tests

\[ H_0: \ X_1, \ldots, X_n \ \text{is iid} \]
against
\[ H_1: \ X_1, \ldots, X_n \ \text{is not iid} \]

We say that the vector \( X_1, \ldots, X_n \) is monotonic at index \( i \) (\( i \in \{2, \ldots, n - 1\} \)) if

\[ X_{i-1} \leq X_i \leq X_{i+1} \quad \text{or} \quad X_{i-1} \geq X_i \geq X_{i+1} \]

and we say that there is a \textbf{turning point} at \( i \) if the vector \( X_1, \ldots, X_n \) is not monotonic at \( i \). Under \( H_0 \), the probability of a turning point at \( i \) is \( 2/3 \).

The indices of \( X_{(1)}, X_{(2)}, X_{(3)} \) form a permutation, uniform among 6 values.

More precisely, let \( T \) be the number of turning points in \( X_1, \ldots, X_n \). It can be shown [18, 105] that, for large \( n \), \( T \) is approximately \( \mathcal{N}_{\frac{2n-4}{3}, \frac{16n-29}{90}} \). Thus the \( p \)-value is, approximatively for large \( n \):

\[
p = 2 \left( 1 - \mathcal{N}_{0,1} \left( \frac{|T - \frac{2n-4}{3}|}{\sqrt{\frac{16n-29}{90}}} \right) \right) \tag{4.44}
\]
Conclusions

- Tests are useful to quantify whether a (small) difference is significant or not.

- Tests are only tests, they all contain assumptions (in the model) that must be discussed.

- Don’t abuse tests; providing confidence intervals is often sufficient and more robust.