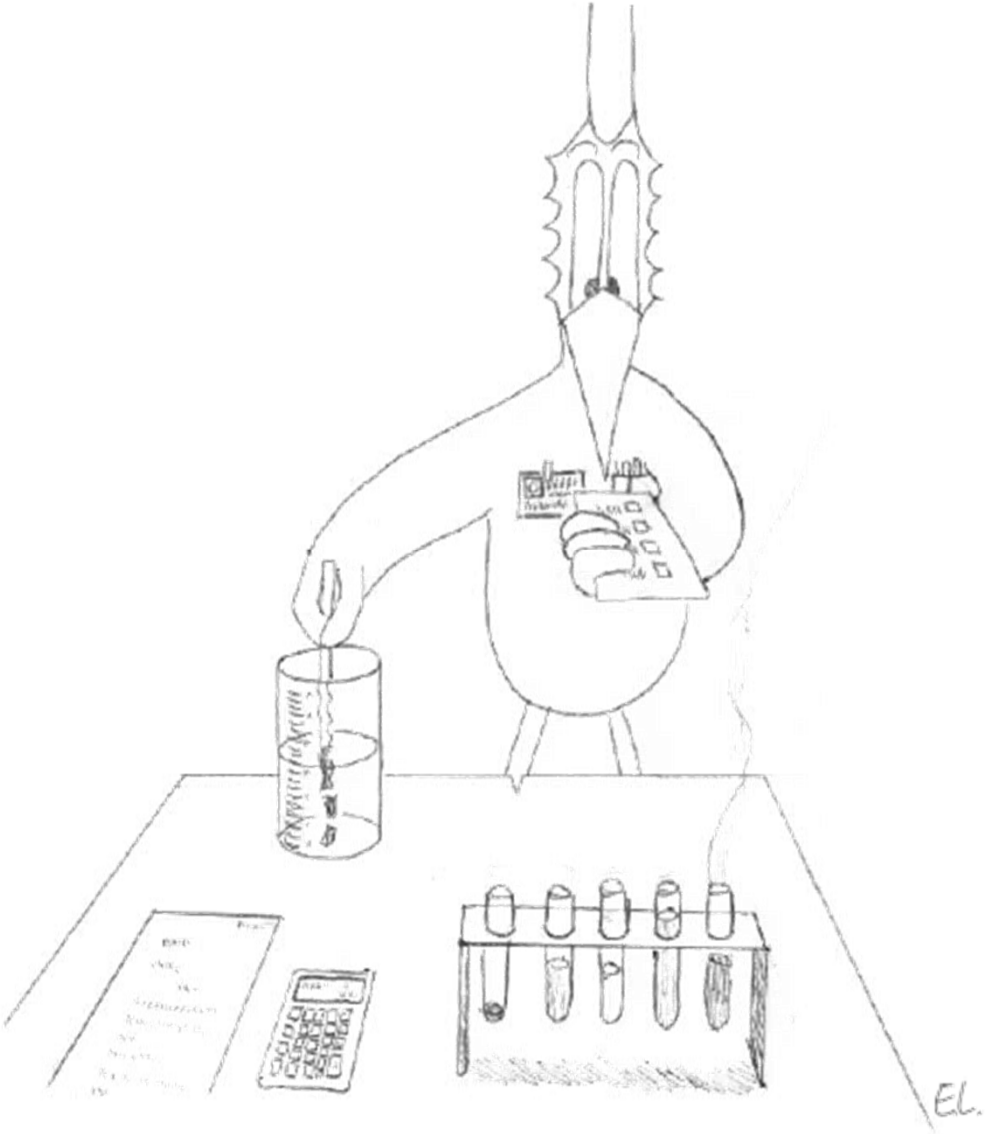


Tests

Jean-Yves Le Boudec

April 2019

Contents

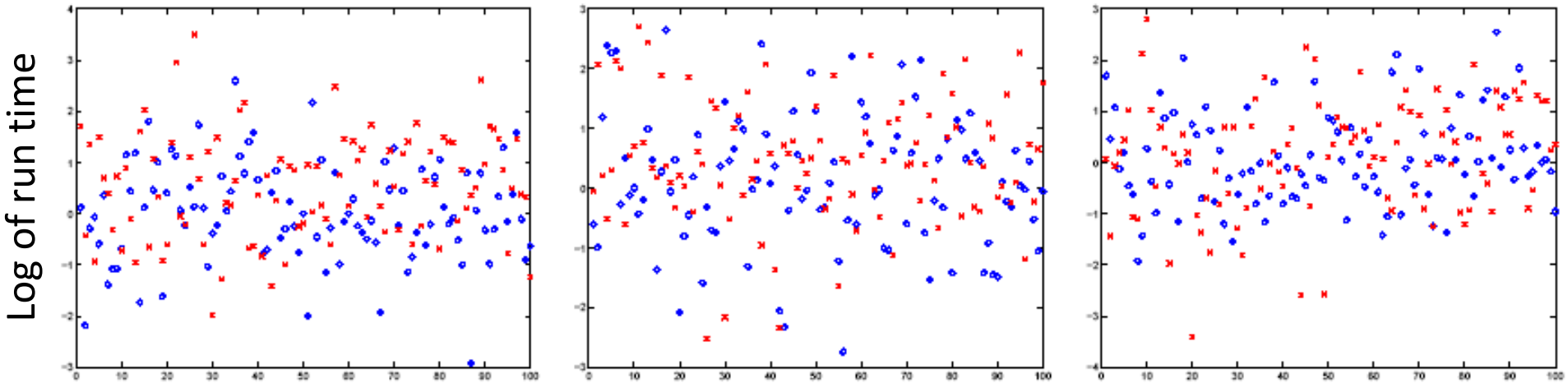
Tests

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 ?

Example: Non Paired Data

Is blue (option 0) better than red (option 1) ?



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

Parameter Set	Compiler Option 0	Compiler Option 1
1	$[-0.1669; 0.2148]$	$[0.3360; 0.7400]$
2	$[-0.0945; 0.3475]$	$[0.2575; 0.6647]$
3	$[-0.1150; 0.2472]$	$[-0.0925; 0.3477]$

For Parameter Set 1 answer is clear (by inspection of confidence interval) no test required

Is this data normal ?

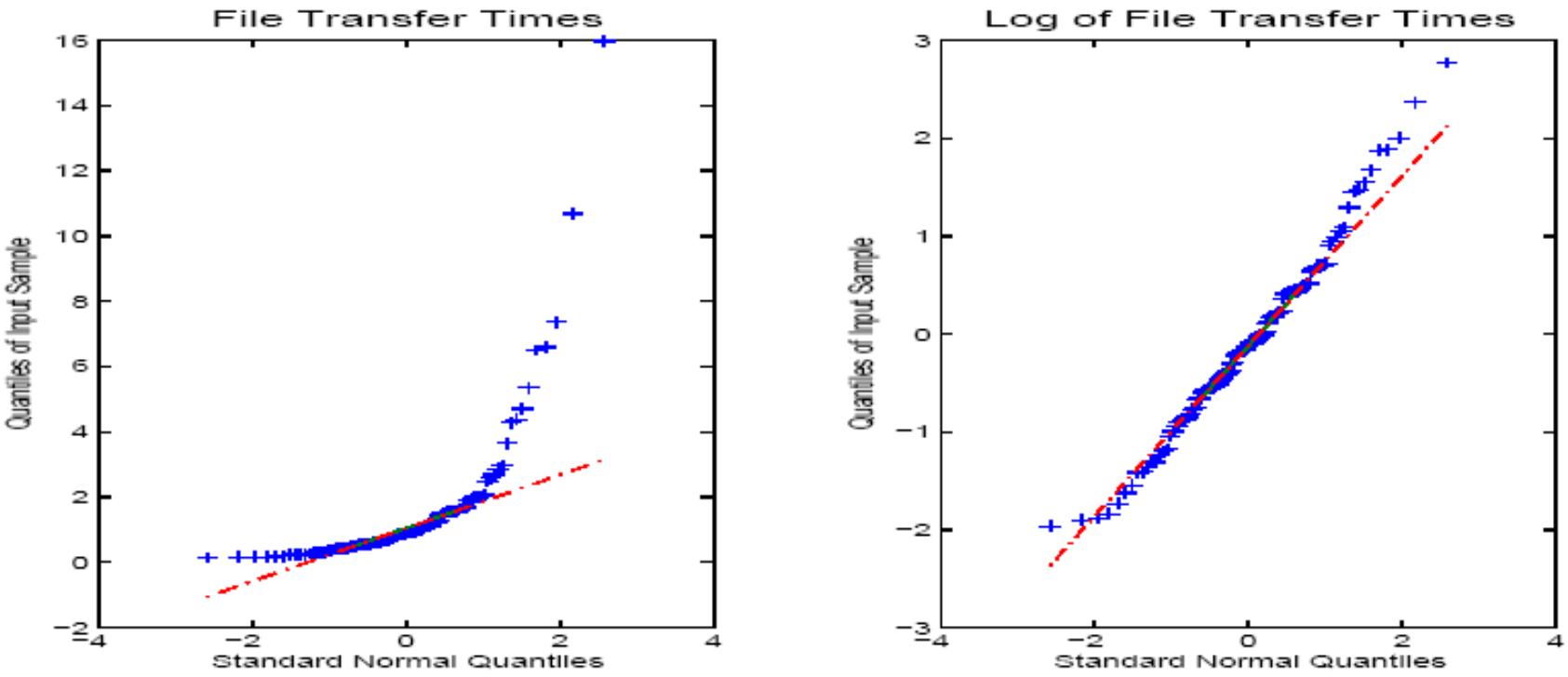


Figure 7.3: Normal qqplots of file transfer data and its logarithm.

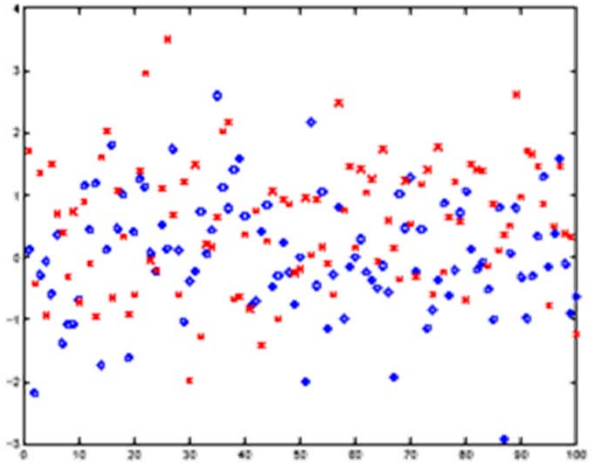
1. The Neyman-Pearson Framework

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

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

Nested model: $\Theta_0 \subset \Theta$ is a set of smaller dimension than Θ

Example: Non Paired Data; Is Red better than Blue ?



(a) Parameter set 1

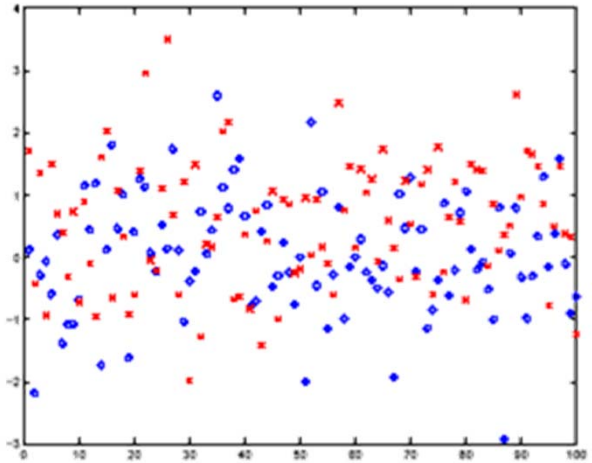
$$H_0: F_0 = F_1$$
$$H_1: F_0 \neq F_1$$

$$\Theta_0 = \{(F_0, F_0), F_0 \text{is a CDF}\}$$

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

$$\Theta = \{(F_0, F_1), F_0 \text{ and } F_1 \text{ are CDF}s\}$$

Example: Non Paired Data; Is Red better than Blue ? ANOVA Model



(a) Parameter set 1

$$H_0: \mu_0 = \mu_1$$

$$H_1: \mu_0 \neq \mu_1$$

$$\Theta_0 = \{(\mu_0, \mu_0, \sigma), \sigma > 0\}$$

Model: x_i and y_i are two independent iid samples $x_i \sim N_{\mu_0,\sigma^2}$ and $y_i \sim N_{\mu_1,\sigma^2}$

$\Theta = \{(\mu_0, \mu_1, \sigma), \sigma > 0\}$

A Test is defined by its Critical Region and has a Size and Power

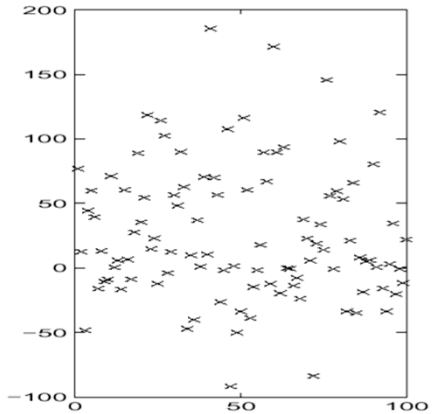
Critical Region: as set C of possible data values $(x_1, \dots, x_n)$ such that

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
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 = Reduction in execution time

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

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

$H_1: \mu > 0$

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

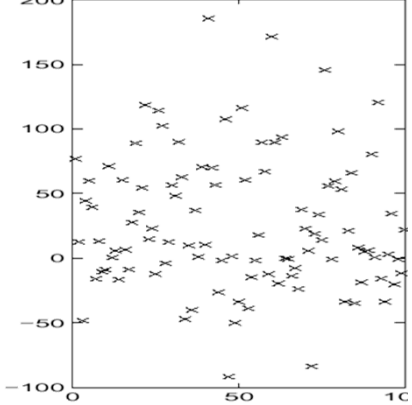
Size of this test = $\sup_{\theta \in \Theta_0} P(\text{data} \in C)$

$$= \sup_{\sigma > 0} P\left(\frac{X_1 + \cdots + X_n}{n} > c\right) =$$

$$\sup_{\sigma > 0} \left\{1 - N_{0,\frac{\sigma^2}{n}}(c)\right\} = \sup_{\sigma > 0} \left\{1 - N_{0,1}\left(\frac{c}{\sigma/\sqrt{n}}\right)\right\} = 1 \text{ !!!}$$

This definition of the rejection region does not work !

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: $C = \left\{(x_1, \dots x_n), \frac{x_1+\cdots+x_n}{ns_n/\sqrt{n}} > c\right\}$ 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

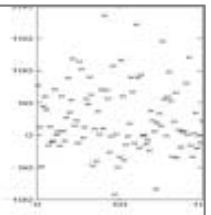
C. 1.9600

D. 2.32632

E. None of the above

F. I don't know

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: $C = \left\{(x_1, \dots x_n), \frac{x_1+\cdots+x_n}{ns_n/\sqrt{n}} > c\right\}$ 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"

11

Power function

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

Here:

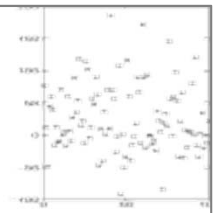
$$P_{\mu,\sigma}(data \in C) =$$

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

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

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

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



X_i = Reduction in execution time

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

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

$H_1: \mu > 0$

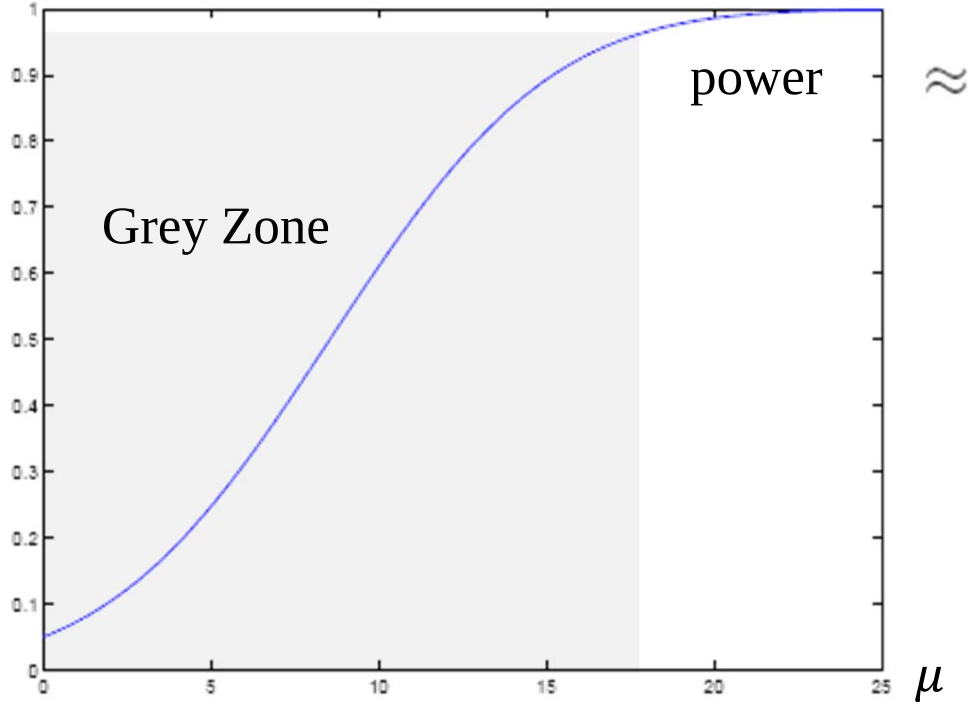
Second attempt: $C = \left\{(x_1, \ldots x_n), \frac{x_1+\cdots+x_n}{ns_n/\sqrt{n}} > c\right\}$ for some c, where s_n^2 is an estimator of variance

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"

11



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

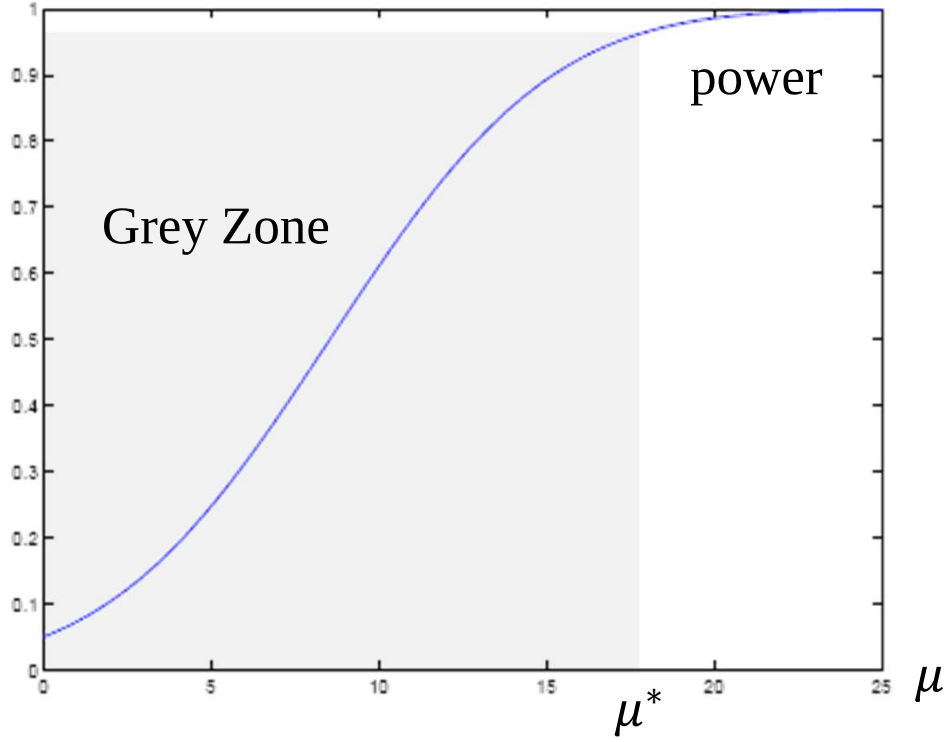
σ is approximated by s_n on the plot

$\mu \approx 0 \Rightarrow$ power ≈ 0.05 (bad but unavoidable)

Grey zone: for $\mu \leq \mu^* = 18$ power $\leq 95\%$

If true μ is in grey zone, test will likely declare $\mu = 0$

For data at hand: power = 0.9997, Proba of type 2 error = 0.0003



We can interpret μ^* as the *statistical significance* value of the test at size 0.05. The test is unable to distinguish between 0 and values much smaller than μ^*

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

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 μ^*. We want that the probability of deciding H_0 when the true value equal to μ^* or more is similar to the size α, i.e. we want to balance the two types of errors. This gives the equations

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

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

resolution μ^*	optimal threshold c^*	size α^*
10	0.97	0.17
20	1.93	0.02
40	3.87	5.38e-005

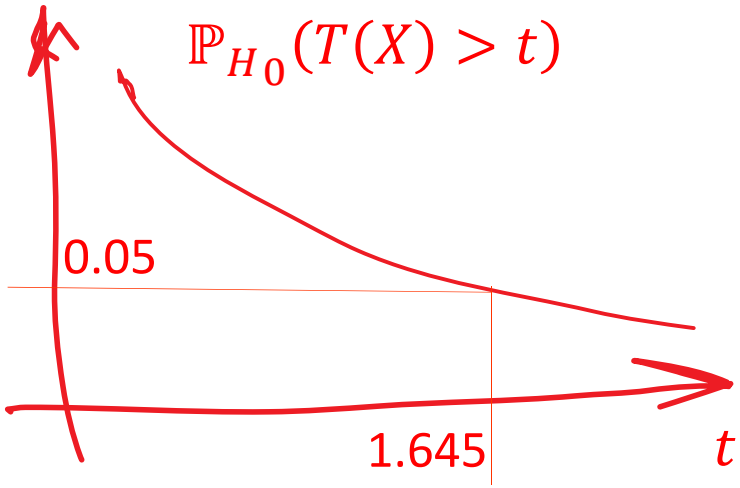
We see that if we care about validly detecting reductions in run time as small as $\mu^* = 10$ms, we should have a test size of 17% or more. In contrast, if the resolution μ^* is 20ms, then a test size of 2% is appropriate.

p-value of a Test

For the previous example, $C = \{T(x) > 1.645\}$ 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 σ and we can plot it →



Saying $t > 1.645$ is the same as saying $P_{H_0}(T(X) > t) < 0.05$

Definition: P-value of this test = $p^*(x) = P_{H_0}\big(T(X) > T(x)\big)$

We reject H_0 if p-value is small (e.g. smaller than 0.05 at size 0.05).

The p-value is defined independently of the size of the test.

Assume rejection region is $C = \{x,\ T(x) > m_0\}$ for some constant m_0. Let $F^c(\cdot\,|\theta)$ be $1 - CDF$ of $T(X)$.

p −value is

$$p^* = \sup_{\theta \in \Theta_0} (T(X) > T(x) \mid \theta) = \sup_{\theta \in \Theta_0} F^c(t(x)|\theta)$$

where x is the data and X is a hypothetical replay.

Reject H_0 if p −value is small.

Software usually returns p −value rather than decision.

EXAMPLE: CONTINUATION OF EXAMPLE 5.2. The $p-$value is

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

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

The critical region of a test has the form $\{\boldsymbol{T}(\boldsymbol{x}) > \boldsymbol{c}\}$ where $\boldsymbol{T}()$ is a pivot. $\boldsymbol{x}$ is the data. $\boldsymbol{X}$ 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 any $\theta \in \Theta_0$

C. A and B

D. None of the above

E. I don't know

In many practical cases, the test statistic $T(X)$ is such that its distribution under H_0 is independent of $\theta \in \Theta_0$. In such cases $p^*(x) = F^c(T(x)|\,\theta)$ for any $\theta \in \Theta_0$.

Observe that $p^*(X) = F^c(T(X)|\,\theta)$ and since $1 - F^c$ is CDF of $T(X)$ under H_0, $p^*(X)$ is uniformly distributed under H_0 i.e. $P_{H_0}(p^*(X) < \alpha) = \alpha$

The rejection region $(p^*(X) < \alpha)$ gives a test of size α

2. Likelihood Ratio Test

A special case of Neyman-Pearson

A Systematic Method to define tests, of general applicability

Assume the nested model: $H_0: \theta \in \Theta_0,\ H_1: \theta \in \Theta \setminus \Theta_0$

$l_x(H_0)$ = log-likelihood of the data under H_0 = log of maximum likelihood obtained assuming $\theta \in \Theta_0$

$l_x(H_1)$ = log-likelihood of the data under H_1 = log of maximum likelihood obtained assuming $\theta \in \Theta$

We always have $l_x(H_1) - l_x(H_0) \geq 0$ (why ?)

If H_1 is true the difference should be large

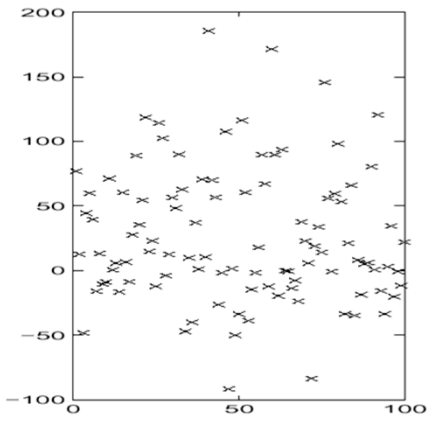
DEFINITION 4.2. *The likelihood ratio test is defined by the rejection region*

$$C = \{l_{\vec{x}}(H_1) - l_{\vec{x}}(H_0) > k\}$$

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

The test statistic $l_{\vec{x}}(H_1) - l_{\vec{x}}(H_0)$ 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.

Step 1: compute the log likelihood

$$X_1, \dots, X_n \sim iid\ N(\mu, \sigma^2)$$

$$\log f_X(x|\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (1)$$

Step 2: compute $l_x(H_1)$ = MLE when we assume $\theta \in \Theta$
i.e. maximize (1) over $\mu \geq 0, \sigma > 0$

We find at the optimum $\hat{\mu}_1 = \max(\bar{x}, 0)$ and $\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^2$
with $\bar{x} = 1/n \sum_{i=1}^{n} x_i$
Thus $l_x(H_1) = -\frac{n}{2}\log(2\pi) - n\log\hat{\sigma}_1 - \frac{n}{2}$

Step 3: compute $l_x(H_0)$ = MLE when we assume $\theta \in \Theta_0$
i.e. maximize (1) over $\mu = 0, \sigma > 0$

We find at the optimum $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Thus $l_x(H_0) = -\frac{n}{2}\log(2\pi) - n\log\hat{\sigma}_0 - \frac{n}{2}$

Step 4: log likelihood ratio : $l_x(H_1) = -n\log\hat{\sigma}_1 + n\log\hat{\sigma}_0$

The rejection region is of the form

$$-n \log \hat{\sigma}_1 + n \log \hat{\sigma}_0 > k$$

for some constant k, which is equivalent to

$$\log \frac{\hat{\sigma}_0}{\hat{\sigma}_1} > \frac{k}{n}$$

which is also equivalent to

$$\hat{\sigma}_1 < K \hat{\sigma}_0 \qquad (2)$$

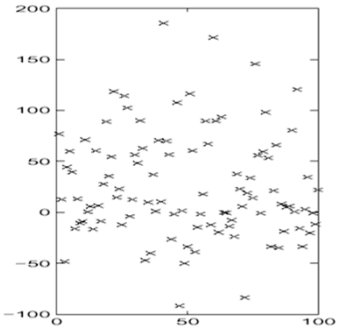
for some constant K (= $e^{k/n}$).

After some algebra, we find that (2) is equivalent to

$$\boxed{C = \left\{ x_1, \dots, x_n : \ \frac{\sqrt{n}\bar{x}}{s_n} > c \right\}}$$

for some c, determined by the size of the test. This is the same *single sided* rejection region as with the ad-hoc, test we had seen before. At size 0.05 we take $c = 1.645$.

A Two-Sided Variant of the Previous Test



X_i = Reduction in execution time.

Model: $X_i \sim N_{\mu,\sigma^2}$

$\Theta = \{(\mu, \sigma), \sigma > 0\}$

$H_0: \mu = 0$

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

$H_1: \mu \neq 0$

i.e. we test $\mu = 0$ versus $\mu \neq 0$ (previously $\mu > 0$)

Derive a likelihood ratio test:

$$l_x(H_1) - l_x(H_0) = \frac{n}{2}\log\left(1 + \frac{T(x)^2}{n-1}\right)$$

with $T(x) = \frac{\sqrt{n}\bar{x}}{\hat{\sigma}}, \bar{x} = \frac{x_1+\cdots+x_n}{n},\ \hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$.

The rejection region is, after some algebra:

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

for some η determined by the size of the test. Using the fact that $T(X)$ has a student-$(n-1)$ distribution, we find $\eta = 1.98$ (at size $\alpha = 0.05$ and for $n = 100$) (and we say we have a "student test").

What is the p-value of this test ?

The rejection region is, after some algebra:

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

for some η determined by the size of the test. Using the fact that $T(X)$ has a student-$(n-1)$ distribution, we find $\eta = 1.98$ (at size $\alpha = 0.05$ and for $n = 100$) (and we say we have a "student test").

A. $p = t_{n-1}\big(T(x)\big)$ where t_{n-1} is the CDF of student-(n-1)

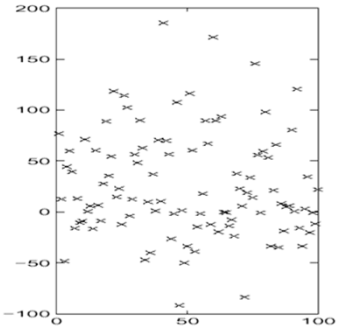
B. $p = 2\Big(1 - t_{n-1}\big(T(x)\big)\Big)$

C. $p = \left|t_{n-1}\big(T(x)\big)\right|$

D. None of the above

E. I don't know

The Two-Sided Test



X_i = Reduction in execution time

Model: $X_i \sim N_{\mu,\sigma^2}$

$$\Theta = \{(\mu, \sigma), \sigma > 0\}$$

$$H_0: \mu = 0$$

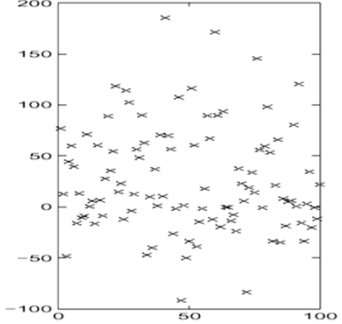
$$\Theta_0 = \{(0, \sigma), \sigma > 0\}$$

$$H_1: \mu \neq 0$$

We have obtained the two sided test $C = \{|T(X)| > 1.98\}$

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

H_1 matters ! (i.e. H_1 is implicitly present in the form of the rejection region)



Here, the two-sided test is the same as a Conf. Interval

Instead of testing $\mu = 0$ vs $\mu \neq 0$ we could estimate μ with a confidence interval and see whether the confidence interval contains 0:

Confidence interval is $\bar{x} \pm 1.98 \frac{\hat{\sigma}}{\sqrt{n}}$ at level 0.95

We reject $\mu = 0$ when $0 \notin [\bar{x} - 1.98 \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + 1.98 \frac{\hat{\sigma}}{\sqrt{n}}]$. This is equivalent to: $|\bar{x}| > 1.98 \frac{\hat{\sigma}}{\sqrt{n}}$ i.e. $\frac{\sqrt{n}|\bar{x}|}{\hat{\sigma}} > 1.98$. This is the same as the two-sided student test at size 0.05 !

Test versus Confidence Intervals

Assume a data model with parameter $\theta = (\theta_1, \theta_2, \dots \theta_k)$.

Assume you can compute a 95% confidence interval for θ_1

$$\hat{\theta}_1 \pm c$$

To test $H_0: \theta_1 = \theta_1^0$ versus $H_1: \theta_1 \neq \theta_1^0$ at size 5% you can take as rejection region

$$\{|\hat{\theta}_1 - \theta_1^0| > c\}$$

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

The "Simple Goodness of Fit" Test

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

Model:

To compute the empirical histogram, we partition the set of values of $\vec{X}$ into *bins* B_i. Let $N_i = \sum_{k=1}^{n} \mathbf{1}_{\{B_i\}}(X_k)$ (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 *multinomial* $M_{n,\vec{q}}$, i.e.

$$\mathbb{P}\{N_1 = n_1, \ldots, N_k = n_k\} = \left(\begin{array}{c} n! \\ n_1! \ldots n_k! \end{array} \right) q_1^{n_1} \ldots q_k^{n_k} \tag{4.7}$$

Hypotheses

H_0: N_i comes from the multinomial distribution $M_{n,\vec{q}}$

against

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

Compute likelihood ratio statistic

$$\mathbb{P}\{N_1 = n_1, ..., N_k = n_k\} = \left(\begin{array}{c} n! \\ n_1!...n_k! \end{array} \right) q_1^{n_1} ... q_k^{n_k}$$

Likelihood of an observation is $l_x(p) = C + \sum_i^k n_i \log p_i$

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 Lagrangian) $\hat{p}_i = \frac{n_i}{n}$

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

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

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

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.

type	observed count	predicted frequency	expected count
smooth yellow	315	9/16	312.75
smooth green	108	3/16	104.25
wrinkled yellow	102	3/16	104.25
wrinkled green	31	1/16	34.75

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

p −value = 0.92 ± 0.05 => Accept H_0

We design a likelihood ratio test with $\boldsymbol{H_0}: \boldsymbol{\theta} = \boldsymbol{\theta_0}, \boldsymbol{H_1}: \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \{\boldsymbol{\theta_0}\}$. 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

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 sub-problems

1. find $\hat{\theta} \in \Theta$ that maximizes the likelihood $l_x(\theta)$
2. find $\hat{\theta}_0 \in \Theta_0$ 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* Θ_1, Θ_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, $2lrs \sim \chi^2_{q_2}$, *where* lrs *is the likelihood ratio statistic.*

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

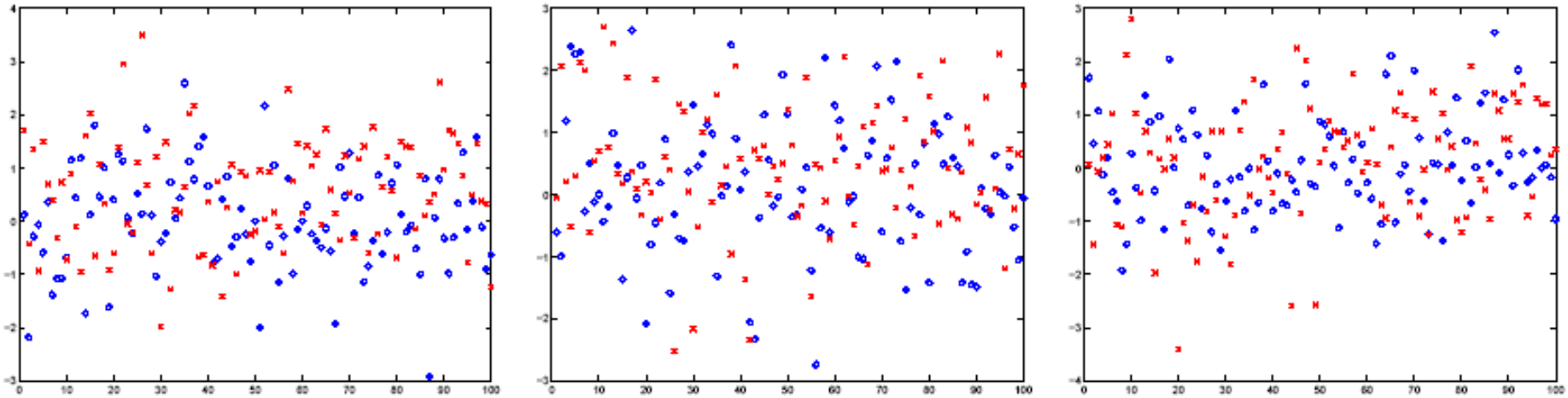
$$p^* \approx 1 - \chi^2_{q_2}(2lrs) \tag{4.25}$$

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

2 x Likelihood ratio statistic

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

Example



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

EXAMPLE: NON PAIRED DATA.(Continuation of Example 7.1 on page 144) Consider the data for one parameter set. The model is

$$X_i = \mu_1 + \epsilon_{1,i} \quad Y_j = \mu_2 + \epsilon_{2,j} \tag{7.15}$$

with $\epsilon_{i,j} \sim$ iid N_{0,σ^2}.

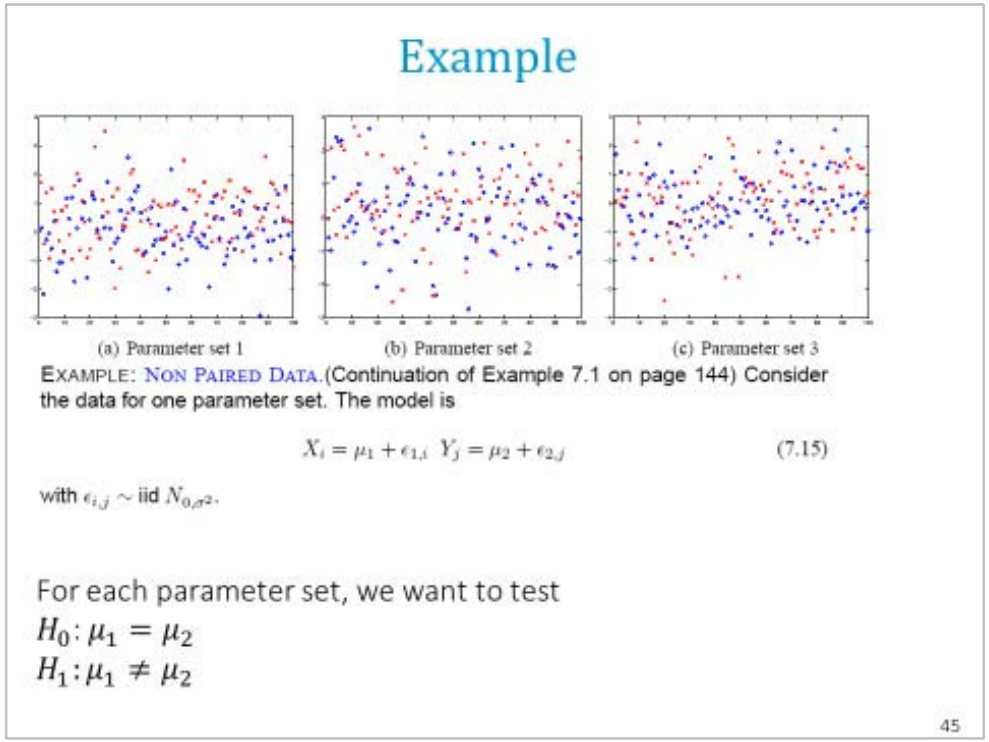
For each parameter set, we want to test

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Estimate $\mu_1 = \mu_2$ under H_0 and compute the likelihood this is a linear regression model

Similarly, we can estimate the μ_1, μ_2 under H_1 and compute the likelihood

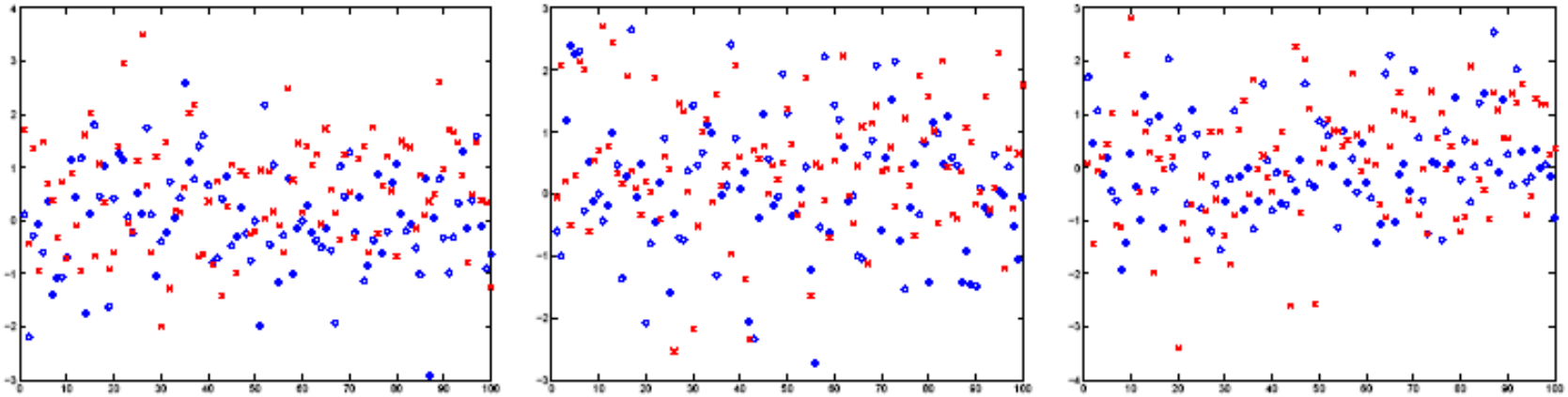


We find $lrs = \frac{n}{2}\log(\frac{SS0}{SS1})$
with $SS0 = \ell^2$norm of residuals under H_0
and $SS1 = \ell^2$norm of residuals under H_1
(See section ANOVA for details)

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

$$p = 3 - 2 = 1$$

$$p - \text{value } p^* \approx 1 - \chi_1^2(2\ lrs)$$



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

Approximate p-values

The corresponding p-values are:

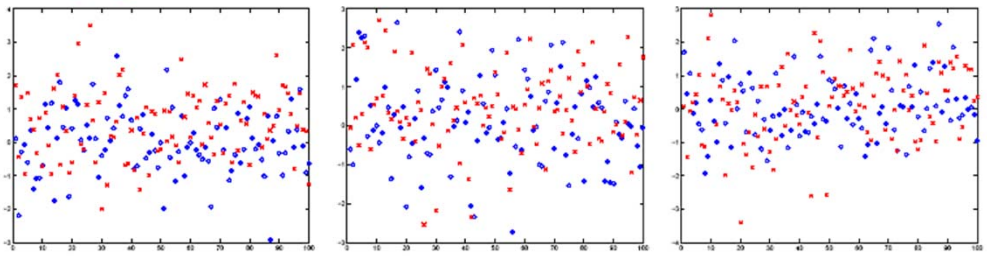
```
Parameter Set 1 pchi2 = 0.0002854
Parameter Set 2 pchi2 = 0.02731
Parameter Set 3 pchi2 = 0.6669
```

Exact p-values (ANOVA)

Parameter Set 1	SS	df	MS	F	Prob>F
Columns	13.2120	1	13.2120	13.4705	0.0003116
Errors	194.2003	198	0.9808		
total	207.4123	199			
Parameter Set 2	**SS**	**df**	**MS**	**F**	**Prob>F**
Columns	5.5975	1	5.5975	4.8813	0.0283
Errors	227.0525	198	1.1467		
total	232.6500	199			
Parameter Set 3	**SS**	**df**	**MS**	**F**	**Prob>F**
Columns	0.1892	1	0.1892	0.1835	0.6689
Errors	204.2256	198	1.0314		
total	204.4148	199			

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

```
Parameter Set 1 pchi2 = 0.0002854
Parameter Set 2 pchi2 = 0.02731
Parameter Set 3 pchi2 = 0.6669
```

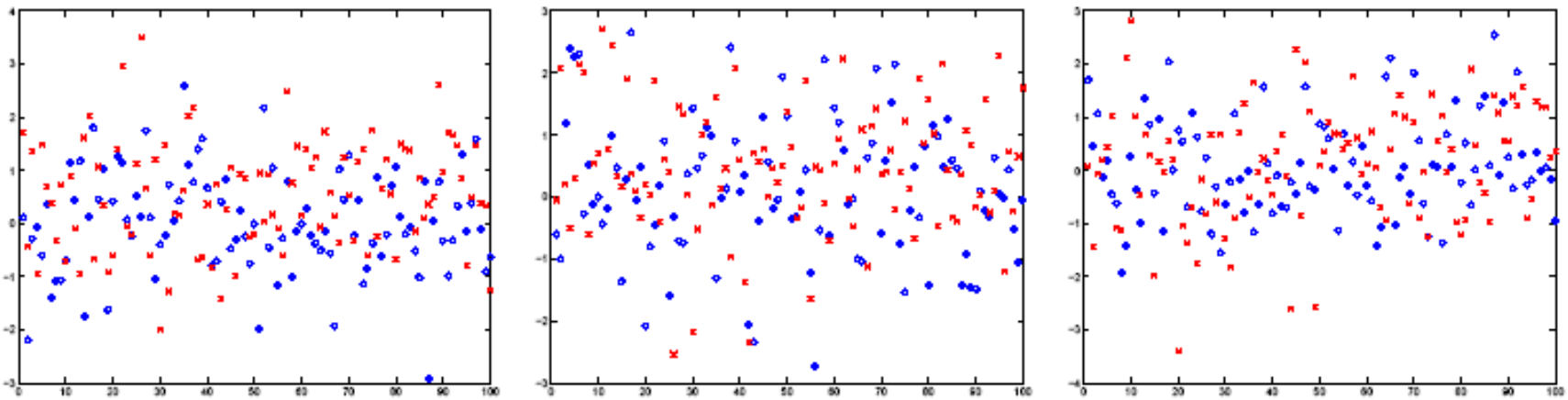
A. 1
B. 2
C. 3
D. 1 and 2
E. 1 and 3
F. 2 and 3
G. All
H. None
I. I don’t know

Compare Test to Confidence Intervals

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



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

Parameter Set	Compiler Option 0	Compiler Option 1
1	$[-0.1669; 0.2148]$	$[0.3360; 0.7400]$
2	$[-0.0945; 0.3475]$	$[0.2575; 0.6647]$
3	$[-0.1150; 0.2472]$	$[-0.0925; 0.3477]$

Parameter Set 1	SS	df	MS	F	Prob>F
Columns	13.2120	1	13.2120	13.4705	0.0003116
Errors	194.2003	198	0.9808		
total	207.4123	199			
Parameter Set 2	**SS**	**df**	**MS**	**F**	**Prob>F**
Columns	5.5975	1	5.5975	4.8813	0.0283
Errors	227.0525	198	1.1467		
total	232.6500	199			
Parameter Set 3	**SS**	**df**	**MS**	**F**	**Prob>F**
Columns	0.1892	1	0.1892	0.1835	0.6689
Errors	204.2256	198	1.0314		
total	204.4148	199			

Table 7.1: ANOVA Tests for Example 7.1 on page 142 (Non Paired Data)

The chi-square distribution

14.1.5 *Chi-Square*

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

% points of χ^2_n

n	0.99	0.975	0.95	0.9
1	6.63	5.02	3.84	2.71
2	9.21	7.38	5.99	4.61
3	11.34	9.35	7.81	6.25
4	13.28	11.14	9.49	7.78
5	15.09	12.83	11.07	9.24
6	16.81	14.45	12.59	10.64
7	18.48	16.01	14.07	12.02
8	20.09	17.53	15.51	13.36
9	21.67	19.02	16.92	14.68

Composite Goodness of Fit Test

We want to test the hypothesis that an iid sample $X_{j,j=1:n}$ has a distribution that comes from a given parametric family $F(\cdot\,|\theta)$.

Partition the set of possible values of the data into k bins $B_1, \dots, B_k$

$$N_i = \sum_{j=1:n} 1_{\{X_j \in B_i\}}$$

If data comes from $F(\cdot\,|\theta)$ then $\vec{N} \sim M_{n,\vec{q}(\theta)}$ (multinomial).

Likelihood ratio test:

$$H_0: \vec{N} \sim M_{n,\vec{q}(\theta)} \text{ for some } \theta \in \Theta$$

$$H_1: \vec{N} \sim M_{n,\vec{p}} \text{ for some } \vec{p} \geq 0, \sum_i p_i = 1$$

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 $lrs = \sum_i n_i \log \frac{n_i}{nq_i(\hat{\theta})}$

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: $p^* \approx 1 - \chi_p^2(2\ lrs)$

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.

type	observed count	predicted frequency	expected count
smooth yellow	315	9/16	312.75
smooth green	108	3/16	104.25
wrinkled yellow	102	3/16	104.25
wrinkled green	31	1/16	34.75

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

Monte Carlo: p= 0.92 ± 0.05 => Accept H_0

Asymptotic Result: p=0.8922

$p^* \approx 1 - \chi^2_p(2\ lrs)$: what is the order $\boldsymbol{p}$?

$k_0 = \dim \Theta_0$
$I =$ nb bins

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

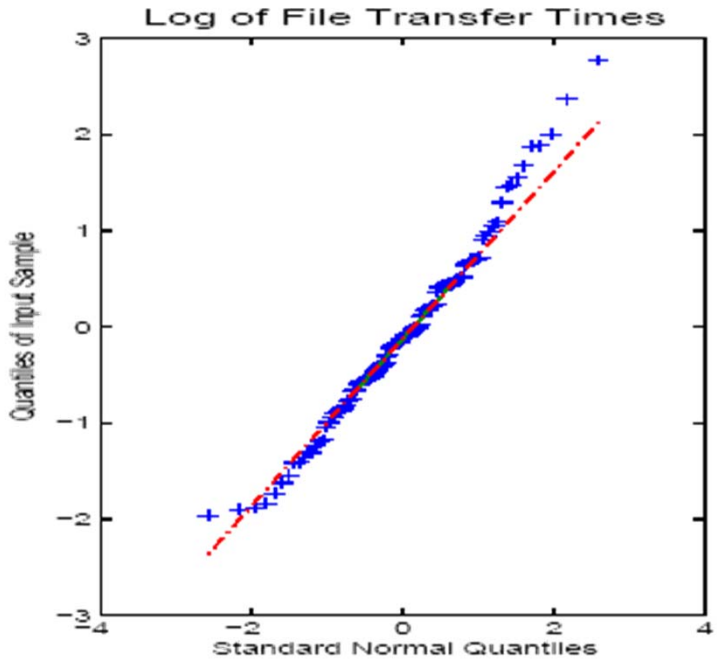
against

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

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

Is it Normal ?

Let us apply the composite goodness of fit test to decide whether the data is normal. We bin the data into a histogram with 10 bins.



Homer's method:  First he estimates μ, σ of the distribution by least-square fitting a straight line to the qqplot. Homer obtains $\hat{\mu} = -0.2652, \hat{\sigma} = 0.8709$. Then he compares the histogram to the theoretical value that would be obtained with $\hat{\mu}$ and $\hat{\sigma}$.

The likelihood ratio statistic is $lrs = 7.6352$; the p-value, obtained from χ^2_7 is $p1 = 0.0327$

Homer rejects normality.

Theoretical values	Observed values
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

Lisa's method

Lisa observes that Homer did not apply a composite goodness of fit test, since he first fitted the data to obtain μ, σ then compared the data to the fitted distribution. Instead, the composite goodness of fit test requires that the parameters q_i of every bin be fitted by MLE.

H_0: the values of the 10 bins are generated by multinomial $n, 10, q$ with $q_i = q_i(\mu, \sigma)$ is the proba that a normal (μ, σ^2) random variable falls in bin i

H_1: the values of the 10 bins are generated by multinomial $n, 10, q$ for some arbitrary distribution q

Lisa's job is to fit the model under H_0 and H_1 and compute the log-likelihoods.

Under H_0: the problem is to maximize $\sum_i n_i \log q_i(\mu, \sigma)$. Lisa uses the fact that $q_i(\mu, \sigma) = N\left(\frac{b_i - \mu}{\sigma}\right) - N\left(\frac{b_{i-1} - \mu}{\sigma}\right)$ where bin i is $[b_{i-1}, b_i]$ and N is the standard normal distribution. She solves it using fminsearch and finds $\hat{\mu} = -0.0725, \hat{\sigma} = 1.0269$. The corresponding values of nq_i are called the "theoretical values" of the histogram →

Theoretical values	Observed values
8.3309	7.0000
9.5028	9.0000
14.4317	17.0000
17.7801	21.0000
17.7709	14.0000
14.4093	17.0000
9.4783	6.0000
5.0577	4.0000
2.1892	3.0000
1.0491	2.0000

Under H_1 the estimation is straightforward and gives $nq_i =$ observed values.

The likelihood ratio statistic is now $lrs = 2.5973$. The p-value, obtained using a χ_7^2 distribution is now p1 = 0.6362.

Lisa says that the data is normal.

Bart's method

Bart cheated on Lisa and copied the beginning of her method. He saw that Lisa's distribution is for $\hat{\mu} = -0.0725, \hat{\sigma} = 1.0269$ and does a simple goodness of fit test:

H_0: the values of the 10 bins are generated by a multinomial $n, 10, q$ with $q_i = q_i(\mu, \sigma)$ is the proba that a normal random variable with $\mu = -0.0725, \sigma = 1.0269$ falls in bin i.

H_1: the values of the 10 bins are generated by multinomial $n, 10, q$ for some arbitrary distribution q.

Bart's p-value is derived from χ_9^2 gives $p2 = 0.8170$, 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.

4 Other Tests
Simple Goodness of Fit

Model: Assume iid data from some distribution

H_0: the distribution is some specified $F()$

H_1: the distribution is anything

Definition: empirical distribution $\hat{F}(x) = \sum_{i=1}^{n} 1_{x_i \leq x}$

Kolmogorov-Smirnov test:

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

Rejection region: $C = \{T > c\}$

Under H_0, the distribution of T is independent of $F()$ (i.e. T is a pivot). This distribution is known in software packages.

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 $\mathbb{P}(T > \tau) = \alpha$.

Example: Kolmogorov-Smirnov with the fitted data (least square)

We used fitted distribution therefore the p-value is overestimated.

Possible corrections: Lilliefors test (empirical correction);
here: same values

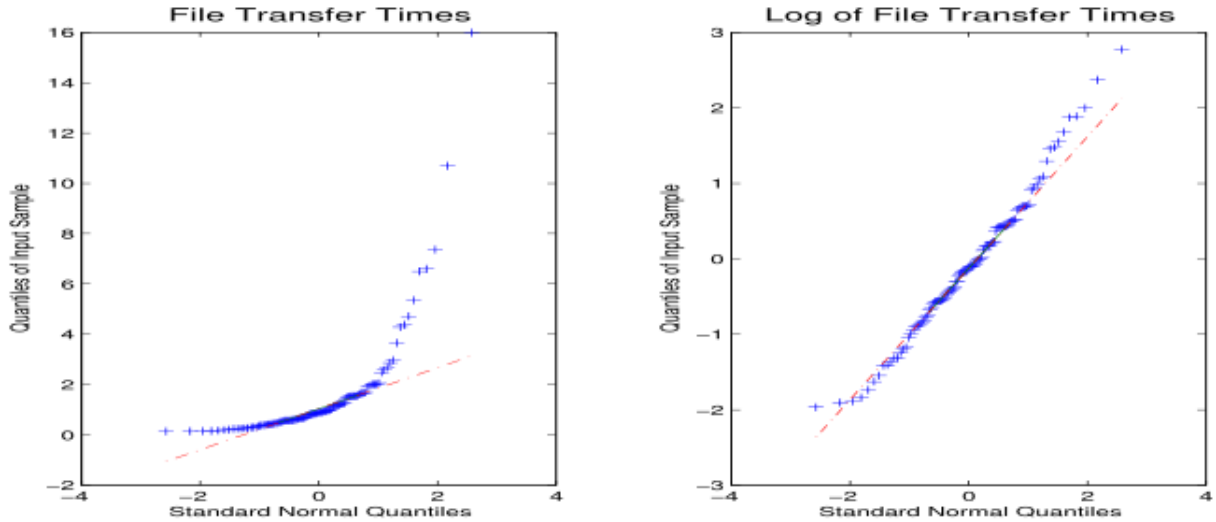


Figure 4.4: Normal qqplots of file transfer data and its logarithm.

```
Original Data
slope     =    0.8155
intercept =    1.0421

Transformed Data
slope     =    0.8709
intercept =   -0.2652
```

```
Original Data
h =     1            p =    0.0000

Transformed Data
h =     0            p =    0.2964
```

Robust Tests : Median Test

Assume an iid sample X_i

H_0: the common distribution of X_i has median $= \ 0$

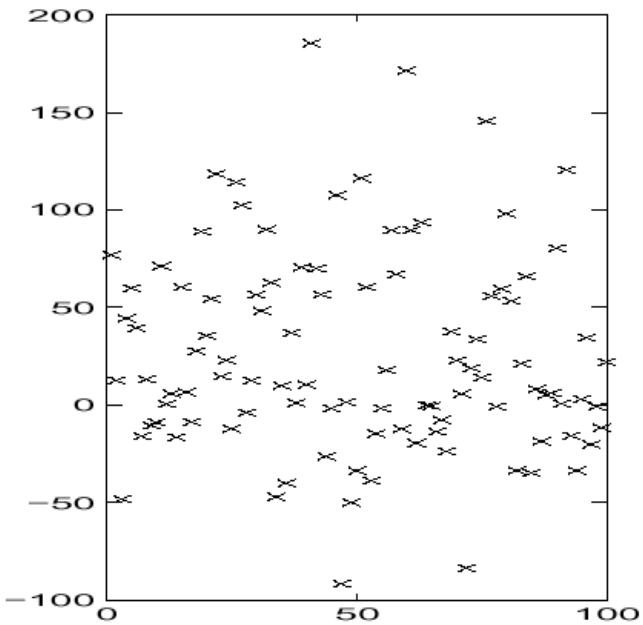
H_1: the common distribution of X_i has median $\neq \ 0$

Assume $I(x)$ is a confidence interval at level 95% for the median

We reject H_0 when $0 \notin I(x)$

Called robust because makes no distributional assumption (other than independence)

Median Test



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

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

which does not contain 0 so we reject 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

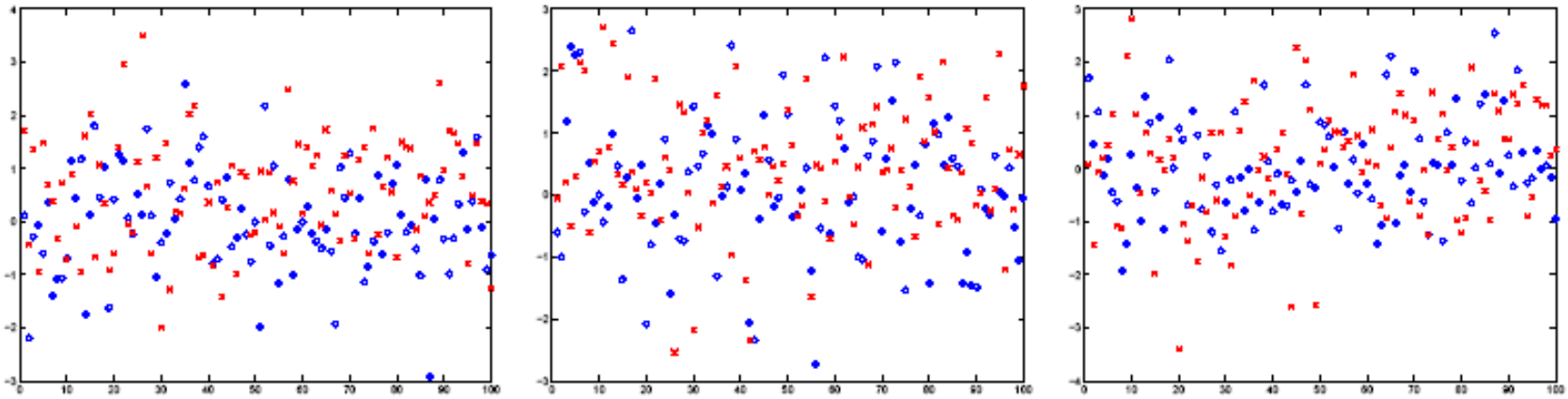
$$m_{n_1,n_2} = \frac{n_1(n_1+n_2+1)}{2} \tag{4.41}$$

$$v_{n_1,n_2} = \frac{n_1 n_2(n_1+n_2+1)}{12} \tag{4.42}$$

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

$$p = 2\left(1 - N_{0,1}\left(\frac{|R - m_{n_1,n_2}|}{\sqrt{v_{n_1,n_2}}}\right)\right) \tag{4.43}$$

Wilcoxon Rank Sum Test



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

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.

Turning Point

$H_0: X_1, \ldots, X_n$ is real and iid

$H_1: X_1, \ldots, X_n$ is real not iid

Compare X_{i-1}, X_i, X_{i+1}: 6 possible cases

X_{i-1} X_i X_{i+1}	X_{i-1} X_{i+1} X_i	X_i X_{i-1} X_{i+1}	X_i X_{i+1} X_{i-1}	X_{i+1} X_{i-1} X_i	X_{i+1} X_i X_{i-1}

Turning points in 4 out of 6 cases

Under H_0 all 6 cases are equally probable. For large n, the number of turning points $T \sim N_{\frac{2n-4}{3}, \frac{16n-29}{90}}$ approximately. Approximate p-value is

$$p = 2\left(1 - N_{0,1}\left(\frac{\left|T - \frac{2n-4}{3}\right|}{\sqrt{\frac{16n-29}{90}}}\right)\right)$$

Conclusions

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

The size of a test should ideally suit the resolution of the data

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

Don’t abuse tests; providing confidence intervals may be more sufficient and more robust