Statistics and Numerical Method, Tsinghua University

Classical Statistical Inference

Xuening Bai (白雪宁)

Institute for Advanced Study (IASTU) &

Department of Astronomy (DoA)

清華大學
Tsinghua University

Nov. 12, 2024

Parametric vs non-parametric models

A statistic model is a set of distributions, among them there are:

A parametric model is one that can be parameterized by a finite number of parameters $\boldsymbol{\theta}$.

- Available models are restricted within parameter space $\boldsymbol{\Theta}$.
- There are also cases where only one or some parameters within $\boldsymbol{\Theta}$ needs to be estimated.

A nonparametric model one that can not be parameterized by a finite number of parameters.

- Directly estimate the CDF/PDF without parameterization.
- Estimate some statistical functional of the CDF without parameterization.

Frequentist vs Bayesian inference

Frequentists consider model parameters θ as an unknown constant. Given θ, observed data are a realization (sampling) of some random variables.

This lecture.

Bayesians consider θ as some random variables, whereas data are considered as known constants.

Next lecture

Fundamental concepts

Let $X_1, \ldots, X_n \sim F$ (the CDF) be an IID sample. A point estimator for some parameter θ is usually of the form:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n)$$

The distribution of $\hat{\theta}_n$ is called sampling distribution.

The bias is defined as: $\mathrm{bias}(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$

An estimator is said to be unbiased if its bias is 0 for any n.

An estimator is said to be consistent if $\hat{\theta}_n$ converges to θ in probability.

The standard deviation of $\hat{\theta}_n$ is called the standard error.

$$\mathrm{se}_{\hat{\theta}_n} = \sqrt{\mathrm{Var}(\hat{\theta}_n)}$$

Fundamental concepts

The efficacy of a point estimate can be assessed by the mean squared error:

$$\text{MSE} = E(\hat{\theta}_n - \theta)^2$$

It is straightforward to show

$$\text{MSE} = \text{bias}^2(\hat{\theta}_n) + \text{Var}(\hat{\theta}_n)$$

A (1-α) confidence interval for a parameter θ is an interval C_n=(a,b), where a, b are functions of $X_1,\ldots,X_n$ such that

$$P(\theta \in C_n) \geq 1 - \alpha$$

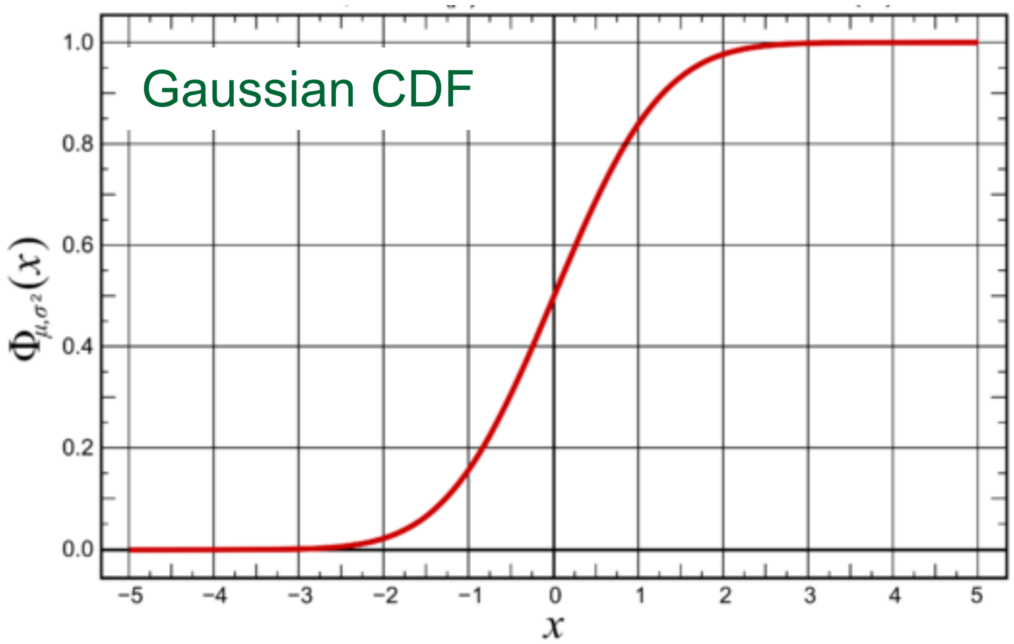
It is important to note that C_n is random but θ is fixed (frequentists' view).

If the sample distribution is asymptotically normal, then confidence interval can be approximately determined from $\text{se}_{\hat{\theta}_n}$ based on the normal distribution.

Confidence interval for a Gaussian distribution

It often occurs that that the sample distribution of $\hat{\theta}_n$ is a Gaussian:

$$\hat{\theta}_n \sim \mathcal{N}(\theta, \hat{\mathsf{se}}^2_{\hat{\theta}_n})$$



Let Φ be the CDF of a standard normal distribution. We define :

$$z_{\alpha/2} \equiv \Phi^{-1}(1-\alpha/2)$$

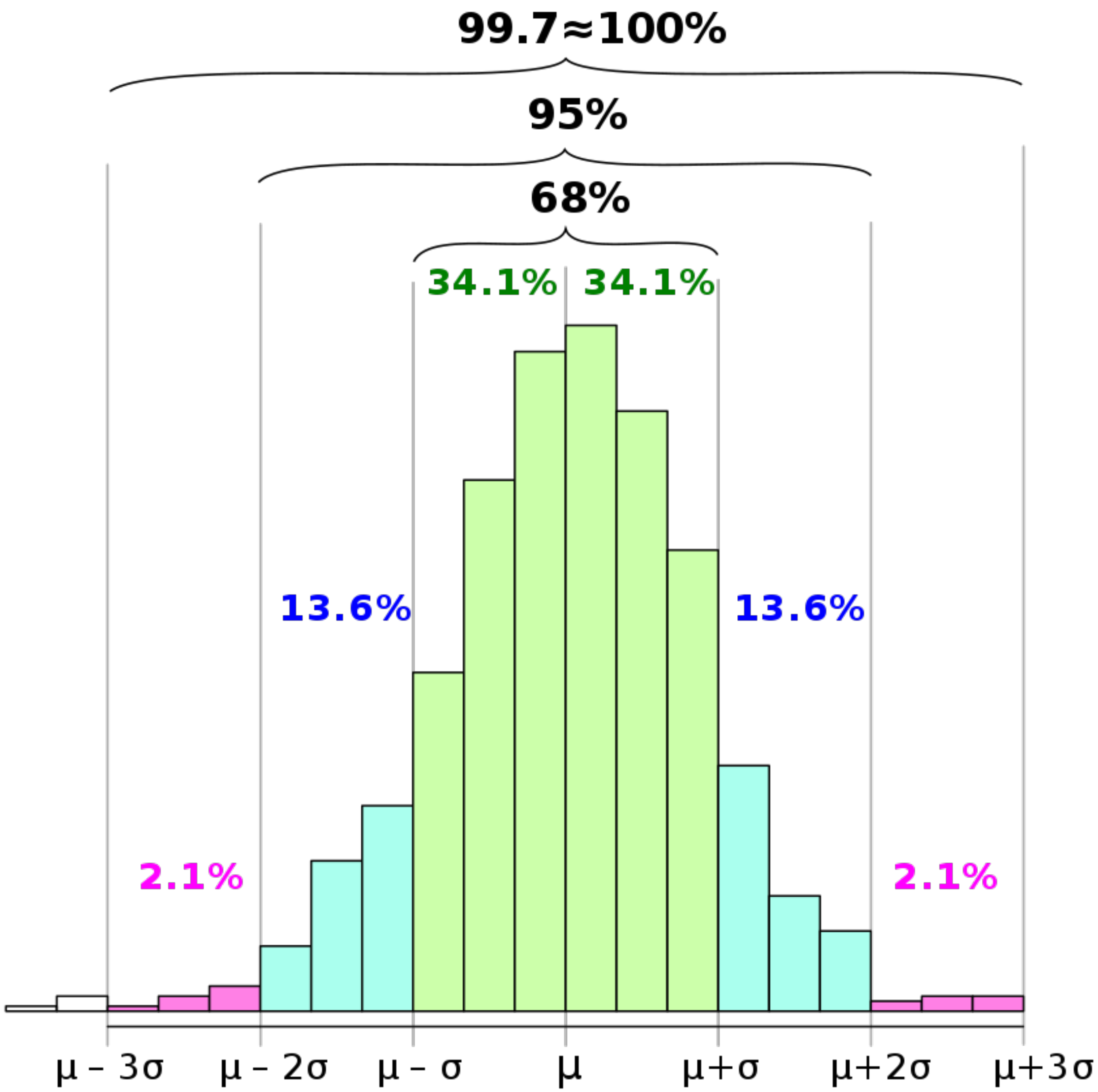
Then $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ for Z~$\mathcal{N}$(0,1).

The confidence interval for $\hat{\theta}_n$ is then given by

$$C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{\mathsf{se}}_{\hat{\theta}_n}, \hat{\theta}_n + z_{\alpha/2}\hat{\mathsf{se}}_{\hat{\theta}_n})$$

For 95% confidence level, α=0.05, we have $z_{\alpha/2} = 1.96 \approx 2$.

Confidence interval for a Gaussian distribution



1σ: 68%

2σ: 95%

3σ: 99.7%

5σ: 99.999943%

Non-parametric estimate of the CDF

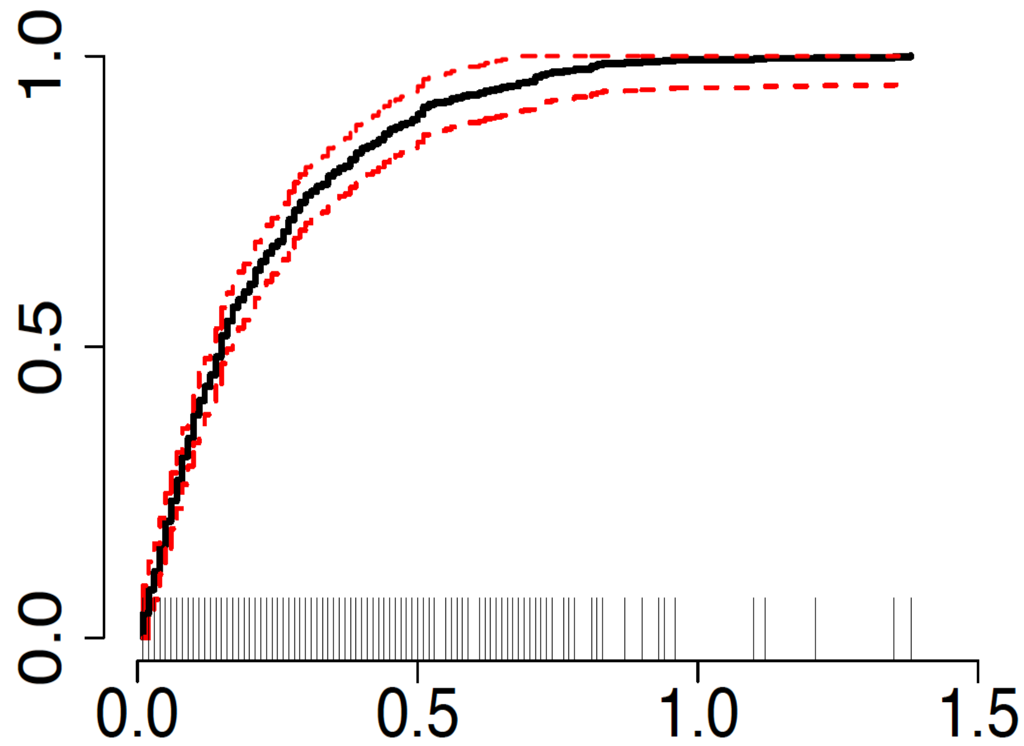
Let $X_1, X_2, \ldots, X_n \sim F$ be an IID sample, where $F(x)$ is the CDF, and they take the values $x_1, x_2, \ldots x_n$.

We can estimate F via the so-called empirical distribution function (EDF):

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

where

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$



Non-parametric estimate of the CDF

One can show that

$$E(\hat{F}_n(x)) = F(x) \ , \ \mathrm{Var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$$

and that the EDF *almost surely* converges to the CDF as $n \to \infty$.

There is the Dvorezky-Kiefer-Wolfowitz (DKW) inequality, which allows one to find the nonparametric confidence interval:

$$L(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\}$$

$$U(x) = \max\{\hat{F}_n(x) + \epsilon_n, 1\}$$

where $\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}$

so that

$$P(L(x) \leq F(x) \leq U(x)) \geq 1 - \alpha \quad \text{(for all } x\text{)}$$

Plug-in estimator

With the EDF, one can estimate any statistical function of *F* as follows.

For $\theta = T(F)$, the plug-in estimator is defined as $\hat{\theta} = T(\hat{F})$.

Examples:

The mean: $\mu = T(F) = \int x dF(x)$ => $\hat{\mu} = \int x d\hat{F}_n(x) = \bar{x}_n$

The variance:

$$\sigma^2 = T(F) = \mathrm{Var}_X = \int x^2 dF(x) - \left(\int x dF(x) \right)^2$$

$$\hat{\sigma}^2 = \int x^2 d\hat{F}(x) - \left(\int x d\hat{F}(x) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}_n^2 = \boxed{\frac{1}{n}} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$$

Bootstrap

A resampling method for estimating standard errors and confidence intervals.

We have a data set $\{x_i\}$ coming from some distribution with its CDF being $F(x)$.

We do not know $F(x)$, but we can use the EDF, $\hat{F}_n(x)$ that is drawn from n IID samples, as an approximation. Note, the PDF of this function is essentially

$$\hat{f}_n(x) = \frac{1}{N} \sum_{i=1}^{n} \delta(x - x_i)$$

Suppose we are interested in some statistic $T(F)$, and with a finite sample it is given by $T_n = g(X_1, \ldots, X_n)$, e.g., based on a plug-in estimator.

Further question: what is the variance and confidence interval of our estimate?

Bootstrap

Key insight from bootstrap:

Drawing an observation from $\hat{F}_n(x)$ is equivalent to drawing one point at random from the original data set.

	draw random sample		compute statistic	
Real world:	F	$\Longrightarrow$ $X_1, \ldots, X_n$	$\Longrightarrow$	$T_n = g(X_1, \ldots, X_n)$
Bootstrap:	$\widehat{F}_n$	$\Longrightarrow$ $X_1^*, \ldots, X_n^*$	$\Longrightarrow$	$T_n^* = g(X_1^*, \ldots, X_n^*)$

In bootstrap, this can be done for large number of times, which allows one to compute the variance and confidence intervals.

Jackknife

Another resampling method in similar spirit to bootstrap.

Instead of drawing a data set of the same size as original, now we drop one or more observations at a time to compute the statistic of interest.

Let $T_n=g(X_1,\ldots,X_n)$ be a statistic, and T_{-i} be the statistic with ith observation removed. Further define:

$$\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{-i}$$

Then the jackknife estimate of Var(T_n) is

$$V_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}(T_{-i} - \bar{T}_n)^2$$

It is easier to implement than bootstrap, while the latter provides more reliable confidence intervals.

Prelude: method of moments

Express population moments as a function of parameters of interest.

$$\alpha_j = E(X^j) = g(\theta_1, \ldots, \theta_k)\ ,\ j = 1, \ldots, k$$

Approximate them with sample moments to and solve for the parameters.

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Example: Normal distribution with $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$, then

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

Easy to calculate, but results are often not optimal.

Can be used as starting values for other methods that requires iteration.

Maximum likelihood estimation (MLE)

The most common method for parameter estimation in a parametric model.

Let $X_1, \ldots, X_n$ be IID with PDF $f(x;\boldsymbol{\theta})$, with observed values being $x_1,\ldots,x_n$.

The likelihood function is defined as:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

Often times, we use the log-likelihood function:

$$l_n(\boldsymbol{\theta}) \equiv \log \mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(x_i;\boldsymbol{\theta})$$

The maximum likelihood estimator, MLE, is the value of θ, denoted by $\hat{\theta}_n$, that maximizes the (log) likelihood function.

Properties of the MLE

Under certain regularity conditions:

- The MLE is consistent:

  Converge to the true parameter in probability.

- The MLE is equivariant:

  If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.

- The MLE is asymptotically normal:

  This is to say, $\frac{\hat{\theta}_n - \theta}{\hat{\text{se}}^2_{\hat{\theta}_n}} \sim \mathcal{N}(0,1)$ at large n.

- The MLE is asymptotically optimal/efficient:

  Among all well-behaved estimators, the MLE achieves the smallest possible variance (known as *Cramer-Rao* bound), at least for large samples.

Fisher information

First introduce the score function:

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$$

It can be seen that $E[s(X;\theta)] = 0$

The *expected* Fisher information (for IID) is defined as

$$I_n(\theta) = \mathrm{Var}_\theta\left(\sum_{i=1}^{n} s(X_i;\theta)\right) = \sum_{i=1}^{n} \mathrm{Var}_\theta[s(X_i;\theta)] = nI(\theta)$$

It can be shown that

$$I(\theta) = \mathrm{Var}_\theta[s(X;\theta)] = E_\theta[s^2(X;\theta)]$$

$$I(\theta) = -E_\theta(-s'(X;\theta)) = -\int \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right) f(x;\theta)dx$$

MLE confidence intervals

The Fisher information can be used to estimate the standard deviation and prove asymptotic normality (based on the central limit theorem).

Let $\hat{\theta}_n$ be the MLE, then asymptotically, its standard error is

$$\mathsf{se}^2_{\hat{\theta}_n} \approx \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}$$

In practice, this standard error can be estimated as

$$\hat{\mathsf{se}}^2_{\hat{\theta}_n} \approx \frac{1}{I_n(\hat{\theta}_n)} = \frac{1}{nI(\hat{\theta}_n)}$$

The asymptotic distribution of $\hat{\theta}_n$ is given by $\hat{\theta}_n \sim \mathcal{N}(\theta, \mathsf{se}^2_{\hat{\theta}_n}) \sim \mathcal{N}(\theta, \hat{\mathsf{se}}^2_{\hat{\theta}_n})$

This allows one to estimate the approximate confidence interval as

$$C_n \approx (\hat{\theta}_n - z_{\alpha/2}\hat{\mathsf{se}}_{\hat{\theta}_n}, \hat{\theta}_n + z_{\alpha/2}\hat{\mathsf{se}}_{\hat{\theta}_n})$$

Example: Bernoulli distribution

Recall that the probability function:

$$f(x;p) = p^x(1-p)^{1-x}\ ,\ (x=0,1)$$

$$\log f(x;p) = x\log p + (1-x)\log(1-p)$$

Now compute the score function:

$$s(X;p) = \frac{X}{p} - \frac{1-X}{1-p}\ ,\quad -s'(X;p) = \frac{X}{p^2} + \frac{1-X}{(1-p)^2}$$

Fisher information:

$$I(p) = E_p(-s'(X;p)) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(p-1)}$$

Therefore, the standard error is approximately

$$\hat{\mathsf{se}}^2_{\hat{p}_n} = \frac{1}{nI(\hat{p}_n)} = \frac{\hat{p}_n(\hat{p}_n-1)}{n}$$

Multiparameter estimation

Let $\theta = (\theta_1, \ldots, \theta_k)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ be the MLE.

The log likelihood function is $l_n = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta})$, and define

$$H_{jj} \equiv \frac{\partial^2 l_n}{\partial \theta_j^2}, \quad H_{jk} \equiv \frac{\partial^2 l_n}{\partial \theta_j \partial \theta_k}$$

The Fisher information matrix is the defined as

$$I_n(\theta) = - \begin{bmatrix} E_\theta(H_{11}) & E_\theta(H_{12}) & \ldots & E_\theta(H_{1k}) \\ E_\theta(H_{21}) & E_\theta(H_{22}) & \ldots & E_\theta(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ E_\theta(H_{k1}) & E_\theta(H_{k2}) & \ldots & E_\theta(H_{kk}) \end{bmatrix}$$

Let $J_n(\theta) = I_n^{-1}(\theta)$ be its inverse.

Asymptotic normality

Under appropriate regularity conditions, we have

$$\hat{\theta} \sim \mathcal{N}(\theta, J_n(\theta))$$

Also, for individual components, we have

$$\hat{\theta}_j \sim \mathcal{N}(\theta_j, J_{n,jj}(\theta))$$

The errors in different parameters are not necessarily independent. In general, they are correlated, and the covariance is given by

$$\mathrm{Cov}(\hat{\theta}_j, \hat{\theta}_k) = J_{n,jk}(\theta)$$

Example: normal distribution

For a normal distribution with two parameters μ, σ, we have

$$l_n(\mu,\sigma) = -n\log\sigma - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial l_n}{\partial \mu} = \sum_{i=1}^{n}\frac{1}{\sigma^2}(X_i-\mu)\,, \qquad \frac{\partial l_n}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^3}$$

One can find that at MLE values, the Fisher information matrix reads:

$$\left.\frac{\partial^2 l_n}{\partial \mu^2}\right|_{\hat\mu,\hat\sigma} = -\frac{n}{\hat\sigma^2}\,, \qquad \left.\frac{\partial^2 l_n}{\partial \sigma^2}\right|_{\hat\mu,\hat\sigma} = \frac{n}{\hat\sigma^2} - 3\frac{n}{\hat\sigma^2} = -\frac{2n}{\hat\sigma^2}\,, \qquad \left.\frac{\partial^2 l_n}{\partial \mu\partial\sigma}\right|_{\hat\mu,\hat\sigma} = 0$$

$$I_n(\hat\mu,\hat\sigma) = \frac{n}{\hat\sigma^2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix} \quad\Rightarrow\quad J_n(\hat\mu,\hat\sigma) = \frac{\hat\sigma^2}{n}\begin{pmatrix}1 & 0\\ 0 & 1/2\end{pmatrix}$$

Therefore, errors for μ, σ are uncorrelated, and are given by

$$\hat{\text{se}}(\hat\mu) = \frac{\hat\sigma}{\sqrt{n}}\,, \quad \hat{\text{se}}(\hat\sigma) = \frac{\hat\sigma}{\sqrt{2n}}$$

Outline

- Parameter estimation and confidence interval
  - Fundamental concepts
  - Non-parametric estimations
  - Bootstrap and jackknife
  - Maximum likelihood method
- Hypothesis testing
  - Basic concepts
  - Likelihood ratio test and applications to Gaussian samples
  - Comparing distributions

Hypothesis testing: basic concepts

Hypothesis testing is a procedure for comparing observed data with a hypothesis whose plausibility is to be assessed.

Suppose we partition the parameter space into two disjoint sets $\mathbf{\Theta}_0$ and $\mathbf{\Theta}_1$, and we wish to test:

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1$$

We call H_0 the null hypothesis, and H_1 the alternative hypothesis.

Three ingredients:

- Significance level α: (roughly) the probability of rejecting H_0 when it is correct.
- Test statistic T: calculated from data to measure its compatibility with H_0.
- A rejection rule: specify the range of T to reject H_0.

Example: how to pose the problem?

Suppose you made an amazing discovery, e.g., detect a dark matter particle!

Before announcing this exciting result, you don't want to make a fool of yourself.

How would you pose the hypothesis test problem?

Null hypothesis:	~~This is consistent with dark matter.~~	This is inconsistent with dark matter.
Alternative hypothesis:	~~There is definitely NO dark matter.~~	There is definitely dark matter!

Read: data suggest the Null hypothesis is rejected at 5% significance level.

Hypothesis testing: basic concepts

Let X be a random variable and $\mathcal{X}$ be its range.

We test a hypothesis by finding an appropriate subset of outcomes $W \subset \mathcal{X}$ called the rejection region so that:

$$X \in W \quad \Longrightarrow \quad \text{reject } H_0$$

$$X \notin W \quad \Longrightarrow \quad \text{retain (do not reject) } H_0$$

Usually, the rejection region W is of the form

$$W = \left\{ x : \; T(x) > c \right\}$$

Here, $T(X)$ is called test statistic, and c is called a critical value.

The problem in hypothesis test is to find an appropriate test statistic T and an appropriate critical value c (equivalently, a rejection region).

Example: source detection

Suppose your detector is subject to Poisson noise, with background count rate being λ (known). Over time interval t, total count satisfies the Poisson distribution:

$$f_0(n,t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Let N be the number of photons gathered in an experiment. We want to ask whether there is a signal (which would give an excess).

Null hypothesis: there is no signal, so that data is consistent with $f = f_0$.

Test statistic: $T(N) = N$ (this is a trivial case: there is no parameter)

The rejection region can be defined as: $W : \{n : T(n) > c\}$

We need to specify a significance level α, say 5%. Thus, we require

$$P(X \in W) = P(N > c) = \alpha = 0.05$$

In other words, we should choose c such that $\sum_{k=0}^{c} \frac{(\lambda t)^k}{k!} e^{-\lambda t} \approx 0.95$.

Type I/II errors and significance level

The conclusion drawn from hypothesis test can be false, leading to an error.

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

false positive

false negative

We are generally more concerned with type I error.

Define the power function: $\rho_W(\theta) \equiv P(X \in W\ ;\ \theta)$

Given θ, what is the probability of being rejected?

If H_0 is true, that is, $\theta \in \Theta_0$, how likely is it rejected (type-I error)?

Define significance level: $\alpha = \sup_{\theta \in \Theta_0} \rho_W(\theta)$. (i.e., maximum type I error rate)

Example: normal distribution

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ is known.

We want to test H_0: $\mu \leq 0$ versus H_1: $\mu > 0$, that is, $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$

Consider using $T = \bar{X}$ as a test statistic, and

$$\text{reject } H_0 \text{ if } T > c$$ for some c TBD.

We known that T should satisfy $T \sim \mathcal{N}(\mu, \sigma^2/n)$. That is,

$$Z \equiv \frac{\sqrt{n}(T - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

Therefore,

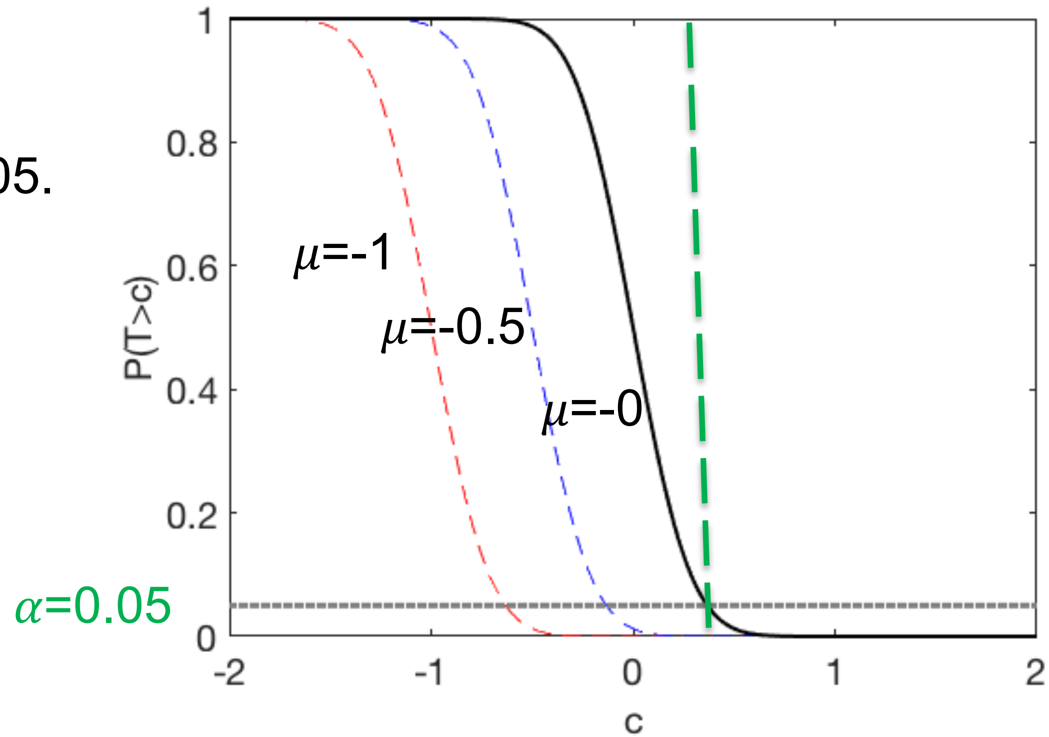
$$P(T > c; \mu) = P\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}; \mu\right) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right)$$

Example: normal distribution

We want significance level α=0.05.

This should apply to all μ≤0.

Clearly, the case that maximizes type-I error corresponds to μ=0.



Therefore, we should determine c according to:

$$P(T > c; \mu = 0) = \alpha \text{ , or } \Phi\left(\frac{\sqrt{n}c}{\sigma}\right) = 1 - \alpha$$

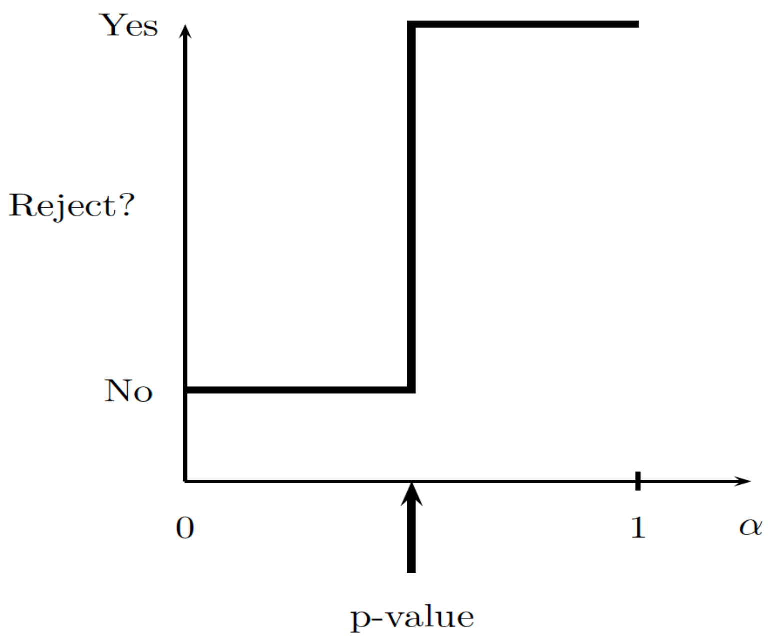
This yields: $c = \dfrac{\sigma}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$

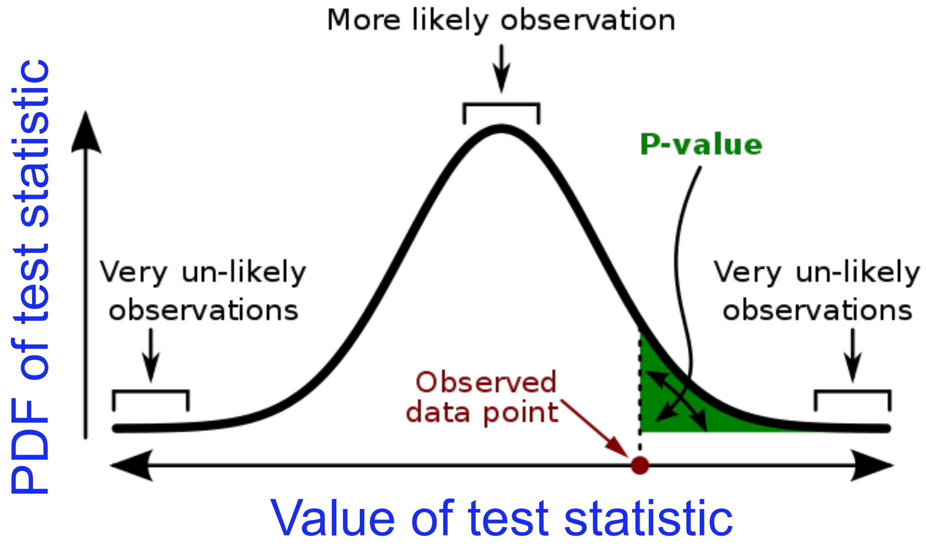
p-value

Given data, whether we reject the hypothesis depends on the significance level α.

We can further ask that given the data, at what level that the null hypothesis can be rejected.

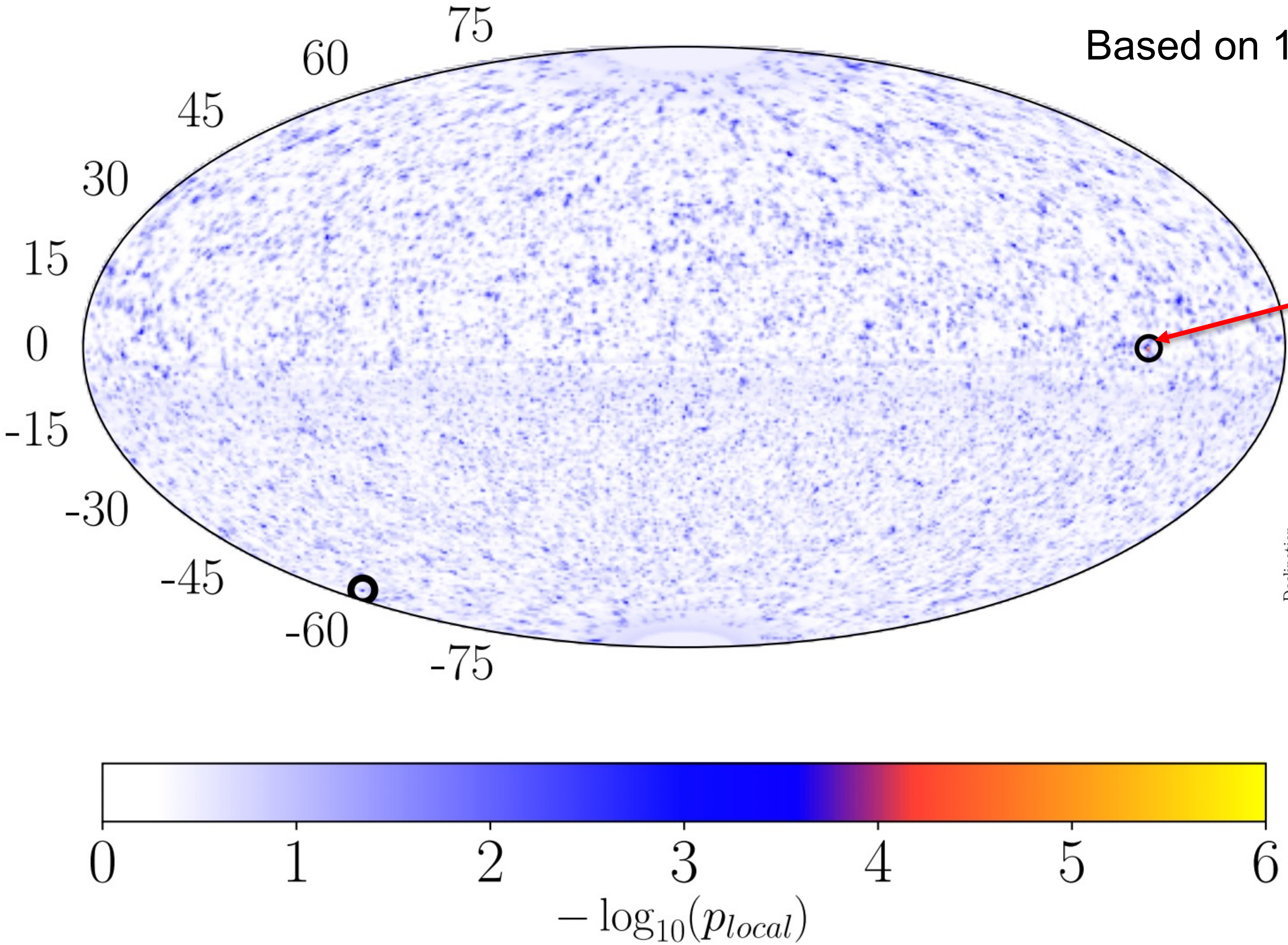
This level is called the p-value: the probability of observing the test statistic at current or more extreme values, assuming the null hypothesis is true.





Example: astro neutrino sources

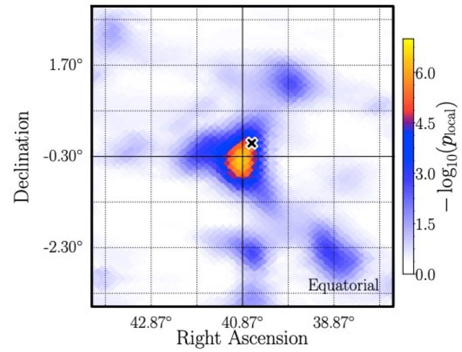
Based on 10-yr IceCube data



Best source (NGC 1068):

log(p)=-4.74

=>4.1 sigma



Aartsen et al. 2020, PRL

Outline

- Parameter estimation and confidence interval
  - Fundamental concepts
  - Non-parametric estimations
  - Bootstrap and jackknife
  - Maximum likelihood method
- Hypothesis testing
  - Basic concepts
  - Likelihood ratio test and applications to Gaussian samples
  - Comparing distributions

Likelihood ratio test

Consider testing: $H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta / \Theta_0$

The likelihood ratio statistic is defined as

parameter that achieves maximum likelihood

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)} \right)$$

Clearly, one should have $\lambda \geq 0$.

The log likelihood ratio should be close to 0 if H_0 is true, and larger otherwise.

The rejection region should be determined as

$$W_0 = \{ \boldsymbol{x} : \lambda(\boldsymbol{x}) > \lambda_0 \}$$

data values

critical value set by significance level

Example

Let $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ where σ is known.

Now test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

With σ readily known, the likelihood function reads:

$$\mathcal{L}_n(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

The likelihood ratio for this problem is thus given by

$$\lambda = 2\log\frac{\mathcal{L}_n(\bar{x})}{\mathcal{L}_n(\mu_0)} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu_0)^2 - \sum_{i=1}^{n}(x_i-\bar{x})^2\right) = \frac{n}{\sigma^2}(\bar{x}-\mu_0)^2$$

Effectively, one can simply choose $\bar{X}$ as test statistic, which satisfies a Gaussian distribution, with rejection region given by

$$W_0 = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) > \lambda_0\} = \{\boldsymbol{x} : |\bar{x}-\mu_0| > c\}$$

with c determined by $P(|\bar{X}-\mu_0| > c) = \alpha = 0.05$

Example

Let $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ where σ is unknown.

Now test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

With σ unknown, the likelihood function reads:

$$\mathcal{L}_n(\mu,\sigma) \propto \frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = \frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]\right)$$

Recall that the maximum likelihood estimate of parameters are

$$\hat{\mu} = \bar{x}\ ,\ \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

or when $\mu = \mu_0$ is known:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$$

For full parameter space:

$$\mathcal{L}_n(\hat{\mu},\hat{\sigma}) \propto \frac{1}{\hat{\sigma}^n}e^{-n/2} \propto \left[\frac{n}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{\frac{n}{2}}$$

For null hypothesis, similarly:

$$\mathcal{L}_n(\mu_0,\hat{\sigma_0}) \propto \left[\frac{n}{\sum_{i=1}^{n}(x_i-\mu_0)^2}\right]^{\frac{n}{2}}$$

Example

Let $X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ where σ is unknown.

Now test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

Therefore, the likelihood ratio is

$$\frac{\mathcal{L}_n(\hat{\mu},\hat{\sigma})}{\mathcal{L}_n(\mu_0,\hat{\sigma_0})} = \left[\frac{\sum_{i=1}^n (x_i-\mu_0)^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]^{\frac{n}{2}} = \left(1+\frac{T^2}{n-1}\right)^{\frac{n}{2}}$$

where $T = \dfrac{\sqrt{n(n-1)}(\bar{x}-\mu_0)}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}$ satisfying the student's t distribution.

(with n-1 degrees of freedom)

Effectively, we can choose T as the test statistic, with rejection region given by

$$W_0 = \{\boldsymbol{x} : |T| > c\}$$

with c determined by $P(|T| > c) = \alpha = 0.05$.

General cases with Gaussian distribution

One-sample test:

Situation	Test statistic	Distribution	Dof	
Test on μ with $\sigma=\sigma_0$ known	$\bar{X}$	$\mathcal{N}(\mu, \sigma_0^2/n)$		a.k.a. Z-test
Test on μ with σ unknown	$T = \frac{\sqrt{n(n-1)}(\bar{X}-\mu_0)}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$	t	n-1	
Test on σ with $\mu=\mu_0$ known	$\chi^2 = \sum_{i=1}^{n} \frac{(X_i-\mu_0)^2}{\sigma_0^2}$	χ^2	n	
Test on σ with μ unknown	$\chi^2 = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{\sigma_0^2}$	χ^2	n-1	

Applications:

- Quality check: is the weight of the chocolate bars the same as claimed?

  One-sample test with t statistic (one-sample t-test).

- How effective is a certain method to reduce weight?

  One-sample test with t statistic (one-sample t-test).

- The city bus is supposed to run every 10 minutes. How can we tell if the bus is sufficiently on time (i.e., variance less than say 2 minutes)?

  One-sample test with χ^2 statistic.

- Are men on average taller than women?

  Two-sample test with t statistic (two-sample t-test).

- Goodness of fit: is there a linear relation between X and Y?

  Read: fit y=ax+b, does a=0? => Does linear fit significantly improve the residuals?

  Two-sample test with F statistic (two-sample F-test).

Two-sample tests

Let $X_1,\ldots,X_{n1} \sim \mathcal{N}(\mu_1,\sigma_1^2)$, $Y_1,\ldots,Y_{n2} \sim \mathcal{N}(\mu_2,\sigma_2^2)$ where $\mu_{1,2}$ are unknown.

Now test H_0: $\sigma_1^2=\sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$.

One can go through similar (but tedious) procedures to derive the likelihood ratio and arrive at the test statistic to be

$$F = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2/(n_1 - 1)}{\sum_{j=1}^{n_2}(y_j - \bar{y})^2/(n_2 - 1)}$$

which satisfies the F distribution with (n_1-1,n_2-1) degrees of freedom.

The rejection region can be set by requiring: $P(F < c_1) = P(F > c_2) = \alpha/2$

The values of c_1, c_2 can be obtained by consulting standard tables.

This test is known as the F-test, which is important in analysis of variance (ANOVA).

General cases with Gaussian distribution

Two-sample test:

Situation	Test statistic	Distribution	Dof
Compare σ_1^2/σ_2^2 with $\mathbf{\Delta}_0$, with $\mu_{1,2}$ unknown	$F = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2}{\sum_{j=1}^{n_2}(Y_j - \bar{Y})^2} \frac{1}{\Delta_0} \frac{n_2 - 1}{n_1 - 1}$	F	(n_1-1,n_2-1)
Compare μ_1, μ_2 knowing σ_1=σ_2	$T = \frac{\sqrt{n_1 n_2 (n_1 + n_2 - 2)/(n_1 + n_2)}(\bar{X} - \bar{Y})}{\sqrt{\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{j=1}^{n_2}(Y_j - \bar{Y})^2}}$	t	n_1+n_2-2
Compare μ_1, μ_2 with $\sigma_1 \neq \sigma_2$	$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$	t (approximately)	round(m*)

$$S_{1,2}^2 = \frac{1}{n_{1,2}} \sum_{1}^{n_{1,2}} (X_i - \bar{X})^2 \text{ or } (Y_j - \bar{Y})^2$$

where $m^* = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left[\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2\right]^{-1}$

Comparing two distributions

Given two sets of data, we want to ask, are they drawn from the same distribution function?

- Are the arrival directions of ultra-high-energy cosmic-rays consistent from coming uniformly in the sky?
- Does the galaxy luminosity function evolve with redshift?
- Does my source property change between two measurements?
- How does the architecture of the solar system compare with other known exoplanetary systems?

Null hypothesis: two distributions are the same.

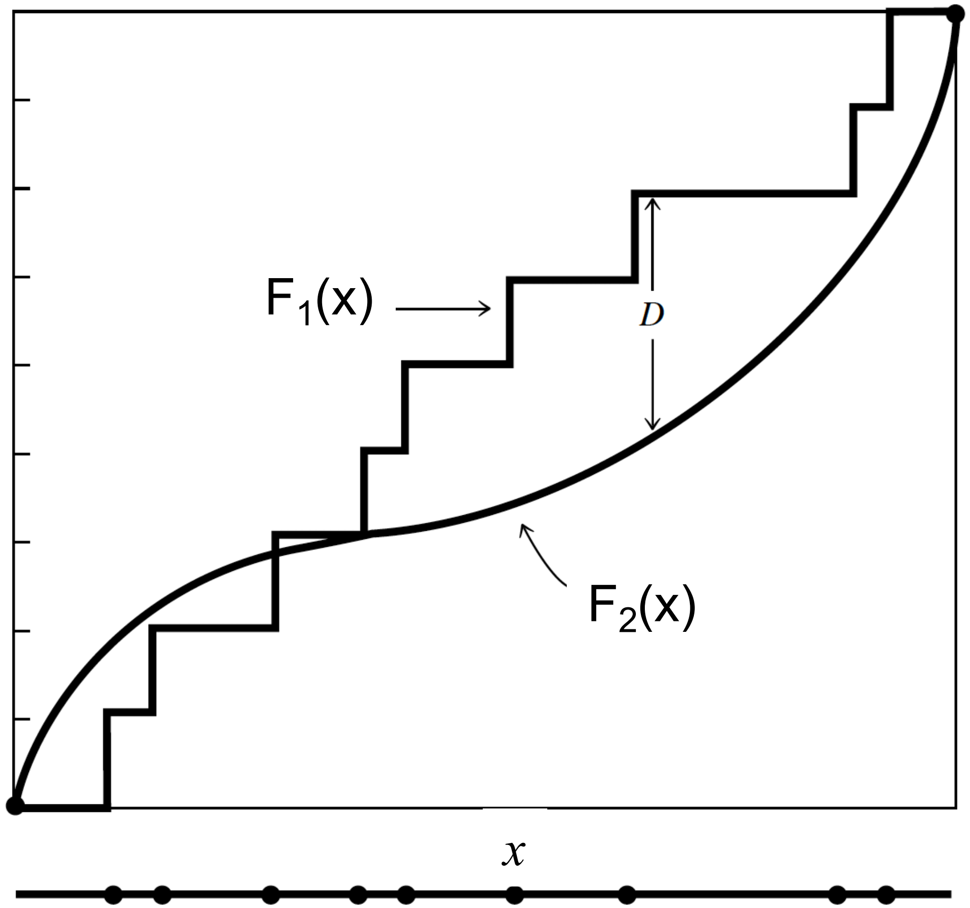
Note: data can be either continuous or discrete/binned, and we either compare data to a known distribution, or to another data set.

Kolmogorov-Smirnov (K-S) test

The most popular non-parametric test for comparing two distributions.

The test is based on the K-S statistic, which measures the maximum distance between two CDFs (EDFs):

$$D = \max_{-\infty < x < \infty} |F_1(x) - F_2(x)|$$



One-sample K-S test: F_2 is a reference distribution;

Two-sample K-S test: F_2 is the CDF/EDF from another sample.

Kolmogorov-Smirnov (K-S) test

The probability to obtain a sample distribution with D larger than measured is

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}\, e^{-2j^2\lambda^2}$$

[Note: $Q_{KS}(0)=1$, $Q_{KS}(\infty)=0$]

where

$$\lambda = \left(0.12 + \sqrt{N_e} + \frac{0.11}{\sqrt{N_e}}\right)D \quad \text{with} \quad N_e = \frac{N_1 N_2}{N_1 + N_2}$$

For one-sample distribution, take $N_2= \infty$, so that $N_e=N_1$.

It is not the best choice for Gaussian distribution (not sensitive to the tails).

There are ways to generalize it to multi-D, though not very straightforward.

Other methods

Kuiper test:

$$D^* = \max[F_1(x) - F_2(x)] + \max[F_2(x) - F_1(x)]$$

It is invariant under cyclic transformation.

Mann-Whitney U test and Wilcoxon signed-rank test:

Non-parametric analog of t-test of the mean for Gaussian distributions, involving using ranks (sorting the data).

Anderson-Darling test and Shapiro-Wilk test:

More sensitive to differences in the tail of the distribution, and are used primarily for testing whether data is drawn from a normal distribution.

Pearson's χ^2 test (and G-test)

Statistical test applied to sets of categorical data (taking finite # of values) to evaluate how likely the observed differences between the sets arise by chance.

The test statistic is defined as

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Superseded by (likelihood ratio): $G = 2\sum_{i=1}^{n} O_i \ln\left(\frac{O_i}{E_i}\right)$

# of observations of type i   expected # of type i

Expected to approximately satisfy χ^2 distribution. The # of degree of freedom is:

Total # of data points – total # of constraints (e.g., parameters)

Three types of comparison: goodness of fit, homogeneity, and independence.

Require large sample size (~1000) to be exact.

Example: fairness of dice

A 6-sided dice is thrown 60 times. Given the result, is it fair?

i	O_i	E_i	O_i-E_i	$(O_i-E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
1	5	10	−5	25	2.5
2	8	10	−2	4	0.4
3	9	10	−1	1	0.1
4	8	10	−2	4	0.4
5	10	10	0	0	0
6	20	10	10	100	10
				Sum	13.4

Null hypothesis: it is fair.

Value of test statistic: 13.4

Dof: 6-1=5.

We find a p-value of 0.02.

Therefore, fairness is rejected at α=0.05 level, but retained at α=0.01 level.

Summary

- Parameter estimation and confidence interval
  - Concepts: sample distribution, standard error, confidence set.
  - Non-parametric estimations: EDF and plug-in estimator.
  - Bootstrap and jackknife: important resampling methods.
  - Maximum likelihood method: asymptotic normality, Fisher information, asymptotically optimal.
- Hypothesis testing
  - Basic concepts: null/alternative hypothesis, type 1/2 error, significance level, p-value.
  - Likelihood ratio test and applications to Gaussian samples: contains several cases of χ^2 test, *t*-test, *F*-test, etc.
  - Comparing distributions: KS test, Pearson's χ^2/G -test, etc.