18.6501 Summary

Michael Hunt 01-11-2019

Introduction

These are notes taken largely from the slides of the Spring and Autumn runs of MITx 18.6501x Fundamentals of Statistics. They were made for my benefit, are incomplete and will contain errors.

Averages of Random Variables

Let $X_1, X_2, \dots, X_n$ be i.i.d. r.v, $\mu = \mathbb{E}[X]$, and $\sigma^2 = \mathbb{V}[X]$.

Law of Large Numbers (Weak and Strong)

$$\overline{X}_n := \frac{1}{n}\sum_1^n X_i \xrightarrow[n\to\infty]{\mathbb{P},\text{a.s.}} \mu$$

Central Limit Theorem (CLT)

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0,1)$$

Equivalently,

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow[n\to\infty]{\text{(d)}} \mathcal{N}(0,\sigma^2)$$

Inequalities

Markov

For a random variable $X > 0$ with mean $\mu > 0$, and any number $t > 0$,

$$\mathbf{P}(X \geq t) \leq \frac{\mu}{t}$$

Chebyshev

For a random variable X with finite mean μ and variance σ^2, and for any number $t > 0$,

$$\mathbf{P}\left(|X - \mu| \geq t\right) \leq \frac{\sigma^2}{t^2}$$

Hoeffding

Given $n > 0$ i.i.d random variables $X_1, X_2, \dots, X_n \overset{iid}{\sim} X$ that are almost surely **bounded**, meaning $\mathbf{P}(X \notin [a,b]) = 0$,

$$\mathbf{P}\left(\left|\overline{X}_n - \frac{b}{2}\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{b^2}} \qquad \text{for all } \epsilon > 0.$$

Unlike for the central limit theorem, here the **sample size n does not need to be large**.

Three types of convergence

$T_n (n \geq 1)$ is a sequence of random variables. T is a random variable that may be deterministic.

Almost surely (a.s.)

$$T_n \xrightarrow[n\to\infty]{\text{a.s.}} T \quad \text{iff} \quad \mathbb{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1$$

Convergence in probability

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \quad \text{iff} \quad \mathbb{P}\left[\left|T_n - T\right| \geq \epsilon\right] \xrightarrow[n\to\infty]{} 0, \quad \forall \epsilon > 0$$

Convergence in distribution

$$T_n \xrightarrow[n\to\infty]{(d)} T \quad \text{iff} \quad \mathbb{E}[f(T_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(T)]$$

for all continuous and bounded functions f.
Convergence in distribution means the convergence, at each point, of the CDFs

Operations on Sequences and Convergence

For a.s. and $\mathbb{P}$ only: Assume

- $T_n \xrightarrow[n\to\infty]{(a.s./\mathbb{P})} T$
- $U_n \xrightarrow[n\to\infty]{a.s./\mathbf{P}} U$

Then,

- $T_n + U_n \xrightarrow[n\to\infty]{a.s./\mathbf{P}} T + U$,
- $T_n U_n \xrightarrow[n\to\infty]{a.s./\mathbf{P}} TU$,
- If in addition $U \neq 0$, then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{a.s./\mathbf{P}} \frac{T}{U}$.

In general, these results **do not** apply to convergence in distribution.

Slutsky's Theorem

Let $(T_n), (U_n)$ be two sequences of random variables such that

- $T_n \xrightarrow[n\to\infty]{(d)} T$
- $U_n \xrightarrow[n\to\infty]{\mathbf{P}} u$

where T is a random variable and u is a given real number (deterministic limit $\mathbb{P}(U = u) = 1$). Then,

- $T_n + U_n \xrightarrow[n\to\infty]{(d)} T + u$,
- $T_n U_n \xrightarrow[n\to\infty]{(d)} Tu$,
- If in addition $u \neq 0$, then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{(d)} \frac{T}{u}$.

Continuous Mapping Theorem

If f is a continuous function,

$$T_n \xrightarrow[n\to\infty]{\text{a.s.}/\mathbf{P}/(d)} T \quad \Rightarrow \quad f(T_n) \xrightarrow[n\to\infty]{} f(T).$$

Statistical models

A statistical model associated to a statistical experiment is a pair:

$$\left(E, \{P_\theta\}_{\theta\in\Theta}\right)$$

, where

- E is a sample space for X, i.e. a set that contains all possible outcomes of X,
- $\{P_\theta\}_{\theta\in\Theta}$ is a family of probability distributions on E,
- Θ is a parameter set, i.e. a set consisting of some possible values of θ.

Note that E cannot itself depend on an unknown parameter.

Identifiability

A parameter θ is identifiable iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.

$$\theta \neq \theta' \implies \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently,

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies \theta = \theta'$$

Estimation

Some definitions:

- *Statistic*: Any measurable function of the sample, e.g. $\overline{X}_n$
- *Estimator*: Any statistic whose expression does not depend on the parameter θ.
- An estimator $\hat{\theta}_n$ of θ is respectively weakly or strongly *consistent* if it converges in probability or almost surely to θ
$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}\text{ resp. a.s.}} \theta \quad \text{(w.r.t. } \mathbb{P}_\theta)$$
.
- An estimator $\hat{\theta}_n$ of θ is *asymptotically normal* if
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$
The quantity σ^2 is then called the *asymptotic variance* of $\hat{\theta}_n$.

Bias of an estimator

- The *bias* of an estimator $\hat{\theta}_n$ of θ is
$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$
- If $\text{bias}(\hat{\theta}) = 0$ we say that θ is *unbiased*.
- In general, $\mathbb{E}[f(x)] \neq f(\mathbb{E}[x])$

Jensen's Inequality

Given a convex function $f(x)$,

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

Given a concave function $g(x)$,

$$\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$$

Variance of estimators

$$\begin{aligned}\text{var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Quadratic risk of estimators

The *quadratic risk* of an estimator $\hat{\theta}_n$ of a true parameter θ is

$$\begin{aligned}\textit{quadratic risk } \boldsymbol{R} &= \mathbb{E}[(\hat{\theta}_n - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + (\mathbb{E}[\hat{\theta}_n] - \theta)^2 \\ &= \text{variance} + \text{bias}^2\end{aligned}$$

Small quadratic risk means small variance *and* small bias. $\boldsymbol{R}$ is the average squared distance between the estimator and the true parameter, and is typically of order $\frac{1}{n}$.

Confidence Intervals

First we take an example. This raises a problem. We then consider three (among many) possible solutions to that problem.

Example 1: Bernoulli statistical model

Class example: proportion of couples who kiss to the left.

Let us take the example of $R_1, \dots, R_n \overset{iid}{\sim} \mathsf{Ber}(p)$ for some unknown parameter p. We estimate p using the estimator $\hat{p} = \overline{R}_n = \frac{1}{n}\sum_{i=1}^n R_i$.

We want the probability of the distance between this estimator and the true parameter p be more than x with low probability α, i.e.:

$$\mathbb{P}(|\bar{R}_n - p| \geq x)$$

Normalise the variable:

$$\mathbb{P}\left(\frac{\sqrt{n}|\bar{R}_n - p|}{\sqrt{p(1-p)}} \geq \frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right) = \alpha$$

But, by the CLT, the random variable converges in distribution to a standard normal, so

$$\mathbb{P}\left(|Z| \geq \frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right) = \alpha$$

Hence

$$2 \times \mathbb{P}\left(Z \geq \frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right) = \alpha$$

and so, taking the complement:

$$2 \times \left[1 - \mathbb{P}\left(Z < \frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right)\right] = \alpha$$

using the CDF:

$$2 \times \left[1 - \Phi\left(\frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right)\right] = \alpha$$

This is equivalent to:

$$\Phi\left(\frac{\sqrt{n}x}{\sqrt{p(1-p)}}\right) = 1 - \alpha/2$$

Inverting the CDF:

$$x = \frac{\sqrt{p(1-p)}\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}$$

or, rewriting the inverse as

$$\Phi^{-1}(1-\alpha/2) := q_{\alpha/2}$$

, we get

$$x = \frac{\sqrt{p(1-p)}q_{\alpha/2}}{\sqrt{n}}$$

So in the limit of large n, the probability that the distance between the estimator x and the true parameter is α. Finally, we can write an asymptotic 'confidence interval as:

$$\lim_{n\to\infty} \mathbf{P}\left(\left[\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right] \ni p\right) = 1 - \alpha$$

But this is *not* a confidence interval because it depends on the true parameter p, which we do not know.

Some confidence interval solutions

Many approaches possible. Here are three:

Solution 1: Conservative bound

No matter what the value of p,

$$p(1-p) \leq \frac{1}{4}$$

Hence, roughly with probability at least $1-\alpha$,

$$\overline{R}_n \in \left[p - \frac{q_{\alpha/2}}{2\sqrt{n}}, p + \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$$

so we get the asymptotic confidence interval

$$\mathcal{J}_{\text{conserv}} = \left[\overline{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$$

so

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{J}_{\text{conserv}} \ni p) \geq 1 - \alpha$$

Note: this technique does not work for an exponential statistical model, for which there is no a priori way to bound the variance.

Solution 2: Solve the quadratic equation for p

We have two inequalities in p

$$\overline{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \overline{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

Hence

$$(p - \overline{R}_n)^2 \leq \frac{q_{\alpha/2}^2 p(1-p)}{n}$$

The roots p_1, p_2 of the quadratic

$$\left(1 + \frac{q_{\alpha/2}^2}{n}\right)p^2 - \left(2\overline{R}_n + \frac{q_{\alpha/2}^2}{n}\right)p + \overline{R}_n^2 = 0$$

give us a confidence interval $\mathcal{J}_{\text{solve}} = [p_1, p_2]$ such that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{J}_{\text{solve}} \ni p) = 1 - \alpha$$

Solution 3: Plug-in

By LLN, $\hat{p} = \overline{R}_n \xrightarrow[n\to\infty]{\mathbf{P},\text{a.s.}} p$

Also,

$$\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} = \sqrt{n}\frac{\overline{R}_n - p}{\sqrt{p(1-p)}}\frac{\sqrt{p(1-p)}}{\sqrt{\hat{p}(1-\hat{p})}}$$

First term converges in distribution to $\mathcal{N}(0,1)$, second term, by LLN, converges almost surely to 1. So by Slutsky we also have

$$\sqrt{n}\frac{\overline{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

which gives a new confidence interval

$$\mathcal{J}_{\text{plug-in}} = \left[\overline{R}_n - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \overline{R}_n + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$$

such that

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{J}_{\text{plug-n}} \ni p) = 1 - \alpha$$

Example 2: Exponential statistical model

Class example: Inter-arrival times of a subway train.

In this example we do not directly observe an estimator of the parameter of interest λ, but one which is a function of it. Here we see how to deal with that.

Assumptions:

- Mutually independent arrival times
- Exponential random variables with shared parameter λ

Hence we have $T_1, \dots, T_n \overset{iid}{\sim} \mathsf{Exp}(\lambda)$

We want to estimate λ, based on the observed train inter-arrival times.

- Density of T_1: $f(t) = \lambda e^{-\lambda t} \quad \forall t \geq 0$
- $\mathbb{E}[T_1] = \frac{1}{\lambda}$
- By LLNs, Estimator for $\frac{1}{\lambda}$ is $\overline{T}_n := \frac{1}{n}\sum_i^n T_i \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \frac{1}{\lambda}$
- Thus, by CMT, the estimator for λ is $\hat{\lambda} := \frac{1}{\overline{T}_n} \xrightarrow[n\to\infty]{\text{a.s./}\mathbb{P}} \lambda$

  Note that $\hat{\lambda}$ is therefore consistent. However it is biased, since $\frac{1}{\lambda}$ is convex, so by Jensen's inequality,

  $\mathbb{E}\left[\frac{1}{\overline{T}_n}\right] \geq \frac{1}{\mathbb{E}[\overline{T}_n]} = \lambda$
- So by CLT, $\sqrt{n}\left(\overline{T}_n - \frac{1}{\lambda}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right)$

But, how does the CLT transfer to $\hat{\lambda}$? How do we find an asymptotic confidence interval for λ?

Delta Method

Let $(X_n)_{n\geq 1}$ be a sequence of r.v. that satisfies

$$\sqrt{n}(X_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$$

for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$. The sequence $(X_n)_{n\geq 1}$ is said to be *asymptotically normal* around θ. Let $g : \mathbb{R} \mapsto \mathbb{R}$ be continuously differentiable at the point θ.

Then

- $g\left((X_n)_{n\geq 1}\right)$ is also asymptotically normal around $g(\theta)$.
- $\sqrt{n}\left(g(X_n)-g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0,(g'(\theta))^2\sigma^2\right)$.

Hypothesis Testing

Consider a sample $X_1,\dots,X_n$ of i.i.d random variables and a statistical model $(E,(\mathbb{P}_\theta)_{\theta\in\Theta})$.
Let Θ_0 and Θ_1 be disjoint subsets of θ (they do not necessarily have to partition Θ.)

$$\text{Consider the two hypotheses: } \begin{cases} H_0 : \theta \in \Theta_0. \\ H_1 : \theta \in \Theta_1. \end{cases}$$

H_0 is the null hypothesis, H_1 is the alternate hypothesis. If we believe that the true θ is in either Θ_0 or Θ_1, we may want to test H_0 against H_1. We want to decide whether to reject H_0 - i.e. look for evidence against it in the data.

Asymmetry in the Hypotheses

H_0 and H_1 do not play a symmetric role. The data is only used to try to disprove H_1. Lack of evidence does not mean that H_0 is true.

A test

A *test* is a statistic $\psi \in \{0,1\}$ such that
$\psi = 0 \implies H_0$ is not rejected,
$\psi = 1 \implies H_0$ *is* rejected.
Coin example: $H_0 : p = 0.5, H_1 : p \neq 0.5$
$\psi = \mathbb{1}\left\{\sqrt{n}\frac{\left|\overline{X}_n - 0.5\right|}{\sqrt{0.5(1-0.5)}} > 0\right\}$ for some $C > 0$
How to choose the threshold C?
A test is fully defined by its rejection region.

Errors

- *Rejection Region of a test* ψ (a set)

$$R_\psi = \left\{x \in E^n : \psi = 1\right\}.$$

  So $\psi(x) = \mathbb{1}\left\{x \in R_\psi\right\}$.

- *Type 1 error of a test* ψ*:* Rejecting H_0 when it is actually true.

$$\begin{aligned} \alpha_\psi : \Theta_0 &\mapsto \mathbb{R} \quad \text{(or [0,1])} \\ \theta &\mapsto \mathbb{P}_\theta[\psi = 1] \end{aligned}$$

- *Type 2 error of a test* ψ*:* Not rejecting H_0 when H_1 is actually true.

$$\begin{aligned} \beta_\psi : \Theta_1 &\mapsto \mathbb{R} \quad \text{(or [0,1])} \\ \theta &\mapsto \mathbb{P}_\theta[\psi = 0] \end{aligned}$$

- *Power of a test* ψ*:*

$$\pi_\psi = \inf_{\theta\in\Theta_1}\left(1-\beta_\psi(\theta)\right)$$

Level, test statistic and rejection region

A test ψ has *level* α if

$$\alpha_\psi(\theta) \leq \alpha, \quad \forall\theta \in \Theta_0.$$

A test ψ has *asymptotic level* α if

$$\lim_{n\to\infty} \alpha_{\psi_n}(\theta) \leq \alpha, \quad \forall\theta \in \Theta_0.$$

In general, a *test* has the form

$$\psi = \mathbb{1}\left\{T_n > c\right\},$$

for some statistic T_n and threshold $c \in \mathbb{R}$.
T_n is called the *Test Statistic*.
The *rejection region* is $R_\psi = \left\{T_n > c\right\}$.

p-value

The asymptotic *p-value* of a test ψ_α is the smallest asymptotic level α at which ψ_α rejects H_0. It is random and depends on the sample.

Methods for estimation

Total Variation Distance

Let $(E,(\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model associated with a sample of i.i.d. r.v. $X_1,\dots,X_n$. Assume that there exists $\theta^\star \in \Theta$ such that $X_1 \sim \mathbb{P}(\theta^\star) : \theta^\star$ is the true parameter. **Statistician's Goal**: Given $X_1,\dots,X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1,\dots,X_n)$ such that $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^\star}$ for the true parameter $\theta^\star$.
This means that $|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^\star}(A)|$ is **small** for $A \in E$.
The *total variation distance* between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}'_\theta$ is

$$\mathrm{TV}(\mathbb{P}_\theta,\mathbb{P}'_\theta) = \max_{A\in E} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|$$

Total variation distance between discrete measures

Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures with a *discrete* sample space E and probability mass functions $f(x)$ and $g(x)$. Then, the total variation distance between $\mathbf{P}$ and $\mathbf{Q}$ is

$$\begin{aligned} \mathrm{TV}(\mathbf{P},\mathbf{Q}) &= \max_{A\subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|, \\ &= \frac{1}{2}\sum_{x\in E} |f(x) - g(x)|. \end{aligned}$$

Total variation distance between continuous measures

Let $\mathbf{P}$ and $\mathbf{Q}$ be probability measures with a *continuous* sample space E and probability density functions $f(x)$ and $g(x)$. Then, the total variation distance between $\mathbf{P}$ and $\mathbf{Q}$ is

$$\begin{aligned} \mathrm{TV}(\mathbf{P},\mathbf{Q}) &= \max_{A\subset E} |\mathbf{P}(A) - \mathbf{Q}(A)|, \\ &= \frac{1}{2}\int_{x\in E} |f(x) - g(x)| dx. \end{aligned}$$

Properties of Total Variation Distance

Let d be a function that takes two probability measures $\mathbf{P}$ and $\mathbf{Q}$ and maps them to a real number $d(\mathbf{P},\mathbf{Q})$. Then d is a **distance** on probability measures if the following four axioms hold. (Here, $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{V}$ are all probability measures.)

- $d(\mathbf{P},\mathbf{Q}) = d(\mathbf{Q},\mathbf{P})$ (symmetric)
- $d(\mathbf{P},\mathbf{Q}) \geq 0$ (non-negative)
- $d(\mathbf{P},\mathbf{Q}) = 0 \iff \mathbf{P} = \mathbf{Q}$ (definite)
- $d(\mathbf{P},\mathbf{Q}) \leq d(\mathbf{P},\mathbf{V}) + d(\mathbf{V},\mathbf{Q})$ (triangle inequality).

These imply that the TV is a *distance* between probability measures.

Limitations of Total Variation Distance

If two probability measures $\mathbf{P}$ and $\mathbf{Q}$ have disjoint support then $\mathrm{TV}(\mathbf{P},\mathbf{Q}) = 1$. In particular, if $\mathbf{P}$ is continuous (eg $\mathcal{N}(0,1)$) and $\mathbf{Q}$ is discrete (eg $\mathrm{Ber}(p)$), then TV=1, even if $\mathbf{Q}$ might become equal to $\mathbf{P}$ as $n\to\infty$. Hence, while TV *is* a distance, it can be trivial since it does not capture proximity. Also, it is generally hard to compute.

Kullback-Leibler (KL) Divergence

Let $\mathbf{P}$ and $\mathbf{Q}$ be discrete probability distributions with pmfs p and q respectively. Let's also assume $\mathbf{P}$ and $\mathbf{Q}$ have a common sample space E. Then the **KL divergence** (also known as **relative entropy**) between $\mathbf{P}$ and $\mathbf{Q}$ is defined by

$$\mathrm{KL}(\mathbf{P},\mathbf{Q}) = \sum_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right),$$

where the sum is only over the support of $\mathbf{P}$.
If $\mathbf{P}$ and $\mathbf{Q}$ are continuous probability distributions with pdfs p and q on a common sample space E, then

$$\mathrm{KL}(\mathbf{P},\mathbf{Q}) = \int_{x\in E} p(x)\ln\left(\frac{p(x)}{q(x)}\right)dx,$$

where the integral is again only over the support of $\mathbf{P}$.

Properties of KL divergence

$\mathrm{KL}(\mathbf{P},\mathbf{Q})$ is not a distance, but a divergence.

- $0 \leq \mathrm{KL}(\mathbf{P},\mathbf{Q}) \leq 1$
- $\mathrm{KL}(\mathbf{P},\mathbf{Q}) \geq 0$ (non-negative)
- $\mathrm{KL}(\mathbf{P},\mathbf{Q}) = 0$ only if $\mathbf{P}$ and $\mathbf{Q}$ are the same distribution (*definite* - this is important).
- Easier to compute than TV distance.
- Can be written as an expectation, so we can estimate the KL by taking averages over i.i.d. samples.

but,

- $\mathrm{KL}(\mathbf{P},\mathbf{Q}) \neq \mathrm{KL}(\mathbf{Q},\mathbf{P})$ (*not* symmetric)
- $\mathrm{KL}(\mathbf{P},\mathbf{Q}) \not\leq \mathrm{KL}(\mathbf{P},\mathbf{V}) + \mathrm{KL}(\mathbf{V},\mathbf{Q})$ (triangle inequality *not* satisfied).

We can use maximum likelihood estimation to find an estimator for the KL divergence between two distributions.

Maximum Likelihood Estimation

Estimating the KL divergence

$$\begin{aligned}\mathsf{KL}(\mathbb{P}_{\theta^\star}, \mathbb{P}_\theta) &= \sum_{x\in E} p_{\theta^\star}(x) \log\left(\frac{p_{\theta^\star}(x)}{p_\theta(x)}\right), \\ &= \mathbb{E}_{\theta^\star}\left[\log\left(\frac{p_{\theta^\star}(X)}{p_\theta(X)}\right)\right] \\ &= \mathbb{E}_{\theta^\star}\left[\log p_{\theta^\star}(X)\right] - \mathbb{E}_{\theta^\star}\left[\log p_\theta(X)\right] \\ &= \text{`constant'} - \mathbb{E}_{\theta^\star}\left[\log p_\theta(X)\right]\end{aligned}$$

which we can estimate since:

$$\mathbb{E}_{\theta^\star}\left[h(X)\right] \rightsquigarrow \frac{1}{n}\sum_{i=1}^n h(X_i) \quad \text{by LLN.}$$

so that a consistent estimator for the KL divergence is

$$\widehat{\mathsf{KL}}(\mathbb{P}_{\theta^\star}, \mathbb{P}_\theta) = \text{`constant'} - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

Maximum likelihood principle

$$\begin{aligned}\min_{\theta\in\Theta} \widehat{\mathsf{KL}}(\mathbb{P}_{\theta^\star}, \mathbb{P}_\theta) &\Leftrightarrow \min_{\theta\in\Theta}\left(-\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)\right) \\ &\Leftrightarrow \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \\ &\Leftrightarrow \max_{\theta\in\Theta} \log\left[\prod_{i=1}^n p_\theta(X_i)\right] \\ &\Leftrightarrow \max_{\theta\in\Theta} \prod_{i=1}^n p_\theta(X_i)\end{aligned}$$

($\Leftrightarrow$ means 'is the same as saying'.)
All we care about the KL is that the argument that maximises -KL is θ^*. It is instrumental in getting the maximum Fisher Information i.e. the optimal (which means, minimal) asymptotic variance.

Maximum Likelihood Estimator (1)

Minimising the estimator for the KL is the same as maximising the likelihood. The quantity

$$\hat{\theta}_n := \text{maximizer of } \prod_{i=1}^n p_\theta(X_i)$$

is called the **maximum likelihood estimator**. This is the same as

$$\hat{\theta}_n := \text{minimizer of } -\frac{1}{n}\sum_{i=1}^n \ln(p_\theta(X_i)).$$

Under certain technical conditions, the maximum likelihood estimator is guaranteed to (weakly) converge to the true parameter $\theta^\star$.

Likelihood: Discrete Case

Let $\left(E, \{P_\theta\}_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \ldots, X_n$. Assume that E is discrete (i.e., finite or countable).
The *likelihood* of the model is the map L_n defined as

$$\begin{aligned}L_n : \quad E^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n, \theta) &\mapsto \mathbb{P}_\theta[X_1, \ldots, X_n] \\ &= \prod_{i=1}^n \mathbb{P}_\theta[X_i = x_i] \quad \text{(since i.i.d.).}\end{aligned}$$

Bernouilli

$$L(x_1 \ldots x_n; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

So we note here that the likelihood function in the case of one variable is just the pmf:

$$L_1(X, \theta) = L(x_1, p) = f_\theta(x_1) = p^{x_1}(1-p)^{n-x_1}.$$

Poisson

$$L(x_1 \ldots x_n; \lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1 \ldots x_n!} e^{-n\lambda}$$

Likelihood: Continuous Case

Let $\left(E, \{P_\theta\}_{\theta\in\Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \ldots, X_n$. Assume that all the $\mathbb{P}_\theta$ have density f_θ.
The *likelihood* of the model is the map L_n defined as

$$\begin{aligned}L_n : \quad E^n \times \Theta &\to \mathbb{R} \\ (x_1, \ldots, x_n, \theta) &\mapsto \prod_{i=1}^n f_\theta(x_i).\end{aligned}$$

Gaussian

$$L(x_1 \ldots x_n; \mu, \sigma^2) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)$$

Exponential

$$L(x_1 \ldots x_n; \lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right) \quad \mathbb{1}\min_i(x_i > 0)$$

..but we could leave the indicator out here since it does not depend on the parameter λ. In a well specified model, all the x_i *must* be greater than zero.

Uniform

$$L(x_1 \ldots x_n; b) = \frac{1}{b^n} \quad \mathbb{1}\max_i(x_i \le b)$$

(note how, in the last two examples, a product of indicators is expressed as a single indicator, involving min() or max().

Maximum Likelihood Estimator (2)

Let $X_1, \ldots, X_n$ be an i.i.d. sample associated with a statistical model $\left(E, \{P_\theta\}_{\theta\in\Theta}\right)$ and let L be the corresponding likelihood. Then, we define the *maximum likelihood estimator* of θ as

$$\hat{\theta}_n^{MLE} = \operatorname*{argmax}_{\theta\in\Theta} L(X_1, \ldots, X_n; \theta),$$

provided it exists. In practice, we use the **log-likelihood estimator**:

$$\hat{\theta}_n^{MLE} = \operatorname*{argmax}_{\theta\in\Theta} \log L(X_1, \ldots, X_n; \theta),$$

which gives the same result.

Maximizing/minimizing functions

Maximising a function is the same as minimising the negative of that function:

$$\min_{\theta\in\Theta} -h(\theta) \Leftrightarrow \max_{\theta\in\Theta} h(\theta).$$

Generally difficult to do for arbitrary functions.

Concave (or convex) functions

A twice differentiable function $h : \Theta \subset \mathbb{R} \mapsto \mathbb{R}$ is said to be *concave* if its second derivative satisfies

$$h''(\theta) \le 0, \quad \forall\theta \in \Theta$$

and *strictly* concave if the inequality is strict: $h''(\theta) < 0$.
Convex and strictly convex are for the reverse.

Multivariate concave functions

For a multivariate function $h : \Theta \subset \mathbb{R}^d \mapsto \mathbb{R}$ where $d \ge 2$, we define the

- *gradient* vector $\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial\theta_1}(\theta) \\ \vdots \\ \frac{\partial h}{\partial\theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^d$
- *Hessian* matrix $\boldsymbol{H}h(\theta) \in \mathbb{R}^{d\times d}$ where

$$\boldsymbol{H}h(\theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial\theta_1\partial\theta_1}(\theta), & \cdots, & \frac{\partial^2 h}{\partial\theta_1\partial\theta_d}(\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h}{\partial\theta_d\partial\theta_1}(\theta), & \cdots, & \frac{\partial^2 h}{\partial\theta_d\partial\theta_d}(\theta) \end{pmatrix}$$

When $\boldsymbol{H}$ is *negative semi-definite* :
$\boldsymbol{x}^\top \boldsymbol{H}h(\theta)\boldsymbol{x} \le 0 \Leftrightarrow h$ is concave, $\quad \forall \boldsymbol{x} \in \mathbb{R}^d, \theta \in \Theta.$

When $\boldsymbol{H}$ is *negative definite* :
$\boldsymbol{x}^\top \boldsymbol{H}h(\theta)\boldsymbol{x} < 0 \Leftrightarrow h$ is *strictly* concave, $\quad \forall \boldsymbol{x} \in \mathbb{R}^d, \theta \in \Theta, \boldsymbol{x} \ne 0.$

Optimality Conditions

If strictly concave functions have a maximum, it is the **unique** solution to:

$$h'(\theta) = 0 \qquad h : \Theta \subset \mathbb{R} \mapsto \mathbb{R}$$

or

$$\nabla h(\theta) = 0 \qquad h : \Theta \subset \mathbb{R}^d \mapsto \mathbb{R}$$

Examples of Maximum Likelihood Estimators

For	the MLE	is
Uniform model	$\widehat{\theta}_n^{MLE}$	$\max_{1\le i\le n} X_i$.
Bernoulli trials	$\widehat{p}_n^{MLE}$	$\overline{X}_n$
Poisson model	$\widehat{\lambda}_n^{MLE}$	$\overline{X}_n$
Gaussian model	$(\widehat{\mu_n}, \widehat{\sigma_n}^2)$	$(\overline{X}_n, \widehat{S_n})$ where $\widehat{S_n} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Consistency of maximum likelihood estimator

Under mild regularity conditions we have

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^\star$$

This is because, for all $\theta \in \Theta$,

$$\frac{1}{n}\log L(X_1,\dots,X_n;\theta) \xrightarrow[n\to\infty]{P} \text{'constant'} - \mathsf{KL}(\mathbb{P}_\theta^\star, \mathbb{P}_\theta).$$

The minimiser of the RHS is $\theta^\star$ if the parameter is *identifiable*. To determine the asymptotic normality of the MLE (or otherwise) we must consider covariance.

Covariance

$$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right]\\ &= \mathbb{E}[X\cdot Y] - \mathbb{E}[X]\cdot\mathbb{E}[Y]\\ &= \mathbb{E}\left[X\cdot(Y-\mathbb{E}[Y])\right]\\ &= \mathbb{E}\left[(X-\mathbb{E}[X])\cdot Y\right]\end{aligned}$$

Note it is only necessary to centre one of the variables.

Properties of Covariance

- $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$
- $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$
- If X, Y are independent, then $\mathsf{Cov}(X,Y) = 0$ - but **the converse is not true in general.**. The exception is when $(X,Y)^\mathsf{T}$ is a Gaussian vector.
- if $\mathsf{Var}(X) = \mathsf{Var}(Y)$ then $\mathsf{Cov}(X+Y, X-Y) = 0$.

Covariance Matrix

The covariance matrix of a random vector $X = (X_1,\dots,X_n)^\mathsf{T} \in \mathbb{R}^d$ is given by

$$\Sigma = \mathsf{Cov}(X) = \mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\mathsf{T}\right] \quad \in \mathbb{R}^{d\times d}$$

i.e. by the outer product of $(X-\mathbb{E}[X])$ with itself.

- **Every rank-1 matrix can be written as an outer product. Conversely, every outer product is a rank-1 matrix**.
- However, a covariance matrix is not necessarily rank-1, since although written as an outer product, it is an outer product of random variables..
- Every covariance matrix is positive definite, every positive definite matrix is a covariance matrix.
- It's eigenvalues are positive and along the diagonal.

$$\Sigma_{ij} = \mathbb{E}\left[\left(X^{(i)} - \mathbb{E}\left[X^{(i)}\right]\right)\left(X^{(j)} - \mathbb{E}\left[X^{(j)}\right]\right)^\mathsf{T}\right] = \mathsf{Cov}\left(X^{(i)}, X^{(j)}\right).$$

on the diagonal,

$$\Sigma_{ii} = \mathsf{Cov}\left(X^{(i)}, X^{(i)}\right) = \mathsf{Var}\left(X^{(i)}\right).$$

If $\boldsymbol{X} \in \mathbb{R}^d$, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices, then

$$\mathsf{Cov}(\boldsymbol{A}X + \boldsymbol{B}) = \mathsf{Cov}(\boldsymbol{A}X) = \boldsymbol{A}\mathsf{Cov}(X)\boldsymbol{A}^\mathsf{T} = \boldsymbol{A}\Sigma\boldsymbol{A}^\mathsf{T}$$

The Multivariate Gaussian Distribution

- A vector $X \in \mathbb{R}^d$ is a Gaussian vector if $\alpha^\mathsf{T} X$ is also a Gaussian vector for any $\alpha \in \mathbb{R}^d, \alpha \neq 0$.
- A Gaussian vector $X \in \mathbb{R}^d$ is completely determined by its expected value $\mathbb{E}[X] = \mu \in \mathbb{R}^d$ and covariance matrix Σ.
- We write $X \sim \mathcal{N}_d(\mu, \Sigma)$.
- pdf: $f(x) = \dfrac{1}{\sqrt{(2\pi)^{d/2}\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right)$ where $x = (x^{(1)},\dots,x^d)$.

The Multivariate CLT

Let $X_1,\dots,X_n \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[X] = \mu, \mathsf{Cov}(X) = \Sigma$.

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0,\Sigma)$$
$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\overline{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0,\mathsf{I}_d)$$

The Multivariate Delta Method

Let $\left(T_n\right)_{n\ge1}$ be a sequence of random vectors in $\mathbb{R}^d$ such that

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0,\Sigma),$$

for some $\theta \in \mathbb{R}^d$ and some variance $\Sigma \in \mathbb{R}^{d\times d}$.
Let $g : \mathbb{R}^d \to \mathbb{R}^k (k \ge 1)$ be continuously differentiable at θ. Then,

$$\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, \nabla g(\theta)^\mathsf{T}\Sigma\nabla g(\theta)\right),$$

where $\nabla g(\theta) = \dfrac{\partial g}{\partial\theta}(\theta) = \left(\dfrac{\partial g_j}{\partial\theta_i}\right)_{\substack{1\le i\le d\\ 1\le j\le k}} \in \mathbb{R}^{d\times k}$

Fisher Information

Define the log-likelihood for one observation as

$$\ell(\theta) = \log L_1(\boldsymbol{X},\theta), \quad \theta \in \Theta \subset \mathbb{R}^d$$

Note that $L_1(\boldsymbol{X},\theta)$ is a pdf or pmf. - and so integrates/sums to 1.
Assume that ℓ is twice differentiable, then under some regularity conditions, the *Fisher Information* is defined as:

$$\begin{aligned}\mathcal{I}(\theta) &= \mathsf{Cov}\left(\nabla\ell(\theta)\right)\\ &= \mathbb{E}\left[\nabla\ell(\theta)\nabla\ell(\theta)^\mathsf{T}\right] - \mathbb{E}\left[\nabla\ell(\theta)\right]\mathbb{E}\left[\nabla\ell(\theta)^\mathsf{T}\right]\\ &= -\mathbb{E}\left[\boldsymbol{H}\ell(\theta)\right]\end{aligned}$$

if $\Theta \subset \mathbb{R}$ we get

$$\mathcal{I}(\theta) = \mathsf{Var}\left[\ell'(\theta)\right] = -\mathbb{E}\left[\ell''(\theta)\right]$$

It is usually easier to compute the second derivative than to find the variance of the first derivative.

Remark: The Fisher information tells how curved (on average) the log-likehood $\ln L_n(x_1,\dots,x_n,\theta)$ for several samples $X_1 = x_1,\dots,X_n = x_n$ is. In particular, $\mathcal{I}(\theta^*)$ tells how curved (on average) the log-likelihood is near the true parameter. As a rule of thumb, if the Fisher information $\mathcal{I}(\theta^*)$ is large, then we expect the MLE to give a good estimate for θ^*.

Fisher Information explicitly from the pdf/pmf

Let $\theta \in \Theta \subset \mathbb{R}^d$ and let $(E, \mathbf{P}\{\theta\}_{\theta\in\Theta})$ be a statistical model. Let $f_\theta(\mathbf{x})$ be the pdf of the distribution $\mathbf{P}_\theta$. Then, the Fisher information $\mathcal{I}$ of the statistical model is

$$\mathcal{I}(\theta) = \mathsf{Cov}(\nabla\ell(\theta)) = -\mathbb{E}\left[\mathbf{H}\ell(\theta)\right],$$

where $\ell(\theta) = \ln f_\theta(\mathbf{X})$
The definition when the distribution has a pmf $p_\theta(\mathbf{x})$ is almost the same, with the expectation taken with respect to the pmf.

Asymptotic normality of the MLE

Theorem: Let $\theta^* \in \Theta$ be the true parameter. Assume:

- The parameter is identifiable;
- For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ;
- θ^* is not on the boundary of Θ;
- $\mathcal{I}(\theta)$ is invertible in the neighbourhood of θ^*;
- A few more technical conditions.

Then, $\hat{\theta}_n^{\mathsf{MLE}}$ satisfies

Consistency $\quad \hat{\theta}_n^{\mathsf{MLE}} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^* \quad$ w.r.t $\mathbb{P}_{\theta^*}$

Asymptotic normality $\quad \sqrt{n}\left(\hat{\theta}_n^{\mathsf{MLE}} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(0, \mathcal{I}^{-1}(\theta^*)\right)$

So, the inverse of the Fisher information is the asymptotic variance of the MLE. The bigger the Fisher information, the smaller the asymptotic variance and thus the more precisely we can estimate the true parameter. i.e. the more information we have.

Method of Moments

Hypothesis Testing

Parametric Hypothesis Testing

The χ^2 distribution

For a positive integer d, the χ^2 distribution with d degrees of freedom is the random variable

$$Z_1^2 + Z_2^2 + \cdots + Z_d^2$$

where $Z_1, \ldots, Z_d \overset{i.i.d}{\sim} \mathcal{N}(0,1)$.

- If $Z \sim \mathcal{N}_d(0, \boldsymbol{I}_d)$, then $\|Z\|_2^2 \sim \chi_d^2$
- $\chi_2^2 \sim \mathsf{Exp}(1/2)$

if $V \sim \chi_d^2$ then

- $\mathbb{E}[V] = \mathbb{E}[Z_1^2] + \cdots + \mathbb{E}[Z_d^2] = d$
- $\mathsf{var}[V] = \mathsf{var}[Z_1^2] + \cdots + \mathsf{var}[Z_d^2] = 2d$

Sample variance and Cochran's Theorem

Recall, sample variance is

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$$

Cochran's theorem states that for $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$, if S_n is the sample variance, then

- $\overline{X}_n \perp\!\!\!\perp S_n$ for all n - so not just asymptotically.
- $\dfrac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$

An unbiased estimator $\tilde{S}_n$ is

$$\tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}S_n$$

so that

$$\mathbb{E}[\tilde{S}_n] = \frac{n}{n-1}\mathbb{E}[S_n] = \frac{n}{n-1}\mathbb{E}\left[\frac{\sigma^2\chi_{n-1}^2}{n}\right] = \frac{n}{n-1}\frac{\sigma^2(n-1)}{n} = \sigma^2$$

Student's T distribution

For a positive integer d, the *Student's T distribution* with d degrees of freedom is

$$t_d \sim \frac{Z}{\sqrt{V/d}}$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi_d^2$ and $Z \perp\!\!\!\perp V$

Student's T test

Let $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. To test:

$$H_0: \quad \mu = 0, \qquad \text{vs} \qquad H_1: \quad \mu \neq 0$$

we use the test statistic

$$T_n = \frac{\sqrt{n}\bar{X}_n}{\sqrt{\tilde{S}_n}} = \sqrt{n}\frac{\overline{X}_n/\sigma}{\sqrt{\tilde{S}_n/\sigma^2}}$$

Since $\sqrt{n}\bar{X}_n/\sigma \sim \mathcal{N}(0,1)$ (under H_0), and $\tilde{S}_n/\sigma^2 \sim \frac{\chi_{n-1}^2}{n-1}$ are ndependent by Cochran's theorem, we have

$$T_n \sim t_{n-1}.$$

Thus the students T test with non-asymptotic level $\alpha \in (0,1)$ is

$$\psi(\alpha) = \mathbb{1}\left\{|T_n| > q_{\alpha/2}\right\}$$

where $q_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of t_{n-1}.

Pros and cons of the T test

Advantage: Non-asymptotic. i.e. can be run on small samples. Will also work on large samples
Disadvantage: The sample has to be drawn from a normally distributed population.

Beyond the T test

The next two tests can be applied more generally to data that is not drawn from a gaussian distribution (but some restrictions still apply).

Wald's Test

This test is based on the MLE.

- Consider an i.i.d. sample $X_1, \ldots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$
- Consider the two hypotheses

$$\begin{cases} \mathrm{H}_0: & \theta = \theta_0 \\ \mathrm{H}_1: & \theta \neq \theta_0 \end{cases}$$

- Let $\hat{\theta}^{MLE}$ be the MLE. Assume the MLE technical conditions are satisfied.
- If H_0 is true, then

$$n(\hat{\theta}^{MLE} - \theta)^T \mathcal{I}(\hat{\theta}^{MLE})(\hat{\theta}^{MLE} - \theta) \xrightarrow[n\to\infty]{(d)} \chi_d^2$$

Likelihood Ratio Test

$$T_n = 2\ln(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_n^c))$$

This is like Wald's test in one dimension, but instead of measuring whether θ is close to θ_0 on the x axis, we are measuring whether the likelihood of θ is close to that of θ_0 on the y axis.

Welch-Satterthwaite formula

$$N = \frac{(\hat{\sigma}_d^2/n + \hat{\sigma}_c^2/m)^2}{\frac{\hat{\sigma}_d^4}{n^2(n-1)} + \frac{\hat{\sigma}_c^4}{m^2(m-1)}} \geq \min(n, m)$$