Unit 2 - Parametric Estimation and Confidence Intervals

In this unit, we introduce a mathematical formalization of statistical modeling to make a principled sense of the Trinity of Statistical inference.

1. Estimation: $\hat{p} = \overline{R_n}$ is an estimator for the real parameter p^*.

2. Confidence intervals: $[0.56, 0.73]$ is a 95% confidence interval for p.

3. Hypothesis testing: "We found *statistical evidence* that more couples turn their head to the right."

Chapter 4 - Inequalities

Inequalities are used to bound quantities that might otherwise be hard to compute.
Markov's inequality: let X be a non-negative r.v. and suppose $\mathbb{E}(X)$ exists. For any $t > 0$,

$$\mathbb{P}(X > t) = \frac{\mathbb{E}[X]}{t}$$

Chebyshev's inequality: let $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var(X)$. Then,

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

Hoeffding's inequality: let $X_1, \ldots, X_n \sim Ber(p)$ then for any $\epsilon > 0$,

$$\mathbb{P}(|\overline{X_n} - p| > \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

Where $\overline{X_n}$ is the sample average and $X \in [a, b]$.
Jensen's inequality: If g is convex, then

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

The reverse is true if g is concave.

Chapter 5 - Convergence of r.v.'s

There are 2 main ideas in this chapter. First, the **LLN** says that the *sample average* converges in probability to $\mathbb{E}$. Second, the **CTL** says that $\sqrt{n}(\overline{X_n} - \mu)$ converges in distribution to a Normal distribution for large enough n.
Let $X_1, \ldots, X_n$ be a sequence of r.v.'s, X is another r.v. Let $F_n(t)$ denote the CDF of X_n an $F(t)$ denote the CDF of X.

Convergence in probability: $X_n \xrightarrow{P} X$ if, for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0$$

as $n \to \infty$. **Warning !** One might think that $X_n \xrightarrow{P} b$ implies $\mathbb{E}[X] \to b$. This is **not** true.
Convergence in distribution: $X_n \xrightarrow{d} X$ if,

$$\lim_{n \to \infty} F_n(t) = F(t)$$

for all t for which F is continuous.

Properties of convergence

The following relationships hold:

1. $X_n \xrightarrow{P} X$ implies that $X_n \xrightarrow{d} X$.

2. If $X_n \xrightarrow{d} X$ and if $\mathbb{P}(X = c) = 1$ for some real number c, then $X_n \xrightarrow{P} X$.

3. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then, $X_n + Y_n \xrightarrow{P} X + Y$.

4. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ then, $X_n + Y_n \xrightarrow{c} X + c$.

5. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then, $X_n Y_n \xrightarrow{P} XY$.

6. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ then, $X_n Y_n \xrightarrow{d} cX$.

7. If $X_n \xrightarrow{P} X$ then, $g(X_n) \xrightarrow{P} g(X)$.

8. If $X_n \xrightarrow{d} X$ then, $g(X_n) \xrightarrow{d} g(X)$.

Parts (4) and (6) are known as the **Slutzky's theorem**. Note that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does not imply that $X_n = Y_n \xrightarrow{d} X + Y$.

Weak LLN & CTL

Interpretation of the WLLN: The distribution of $\overline{X_n}$ becomes more concentrated around μ as n get larger. From the Chebyshev's inequality: $\mathbb{P}(|\overline{X_n} - \mu| > \epsilon) \leq \frac{Var(\overline{X_n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$.
Central Limit Theorem: let $X_1, \ldots, X_n$ be i.i.d with mean μ and variance σ^2. Then

$$Z_n \equiv \frac{\overline{X_n} - \mu}{\sqrt{Var(\overline{X_n})}} = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{(d)} Z$$

where Z is a standard Normal.
Multivariate version of the CTL. Let $X_1, \ldots, X_n$ be i.i.d random *vectors*, $X_i = [X_1, X_2, \ldots, X_k]^T$ with mean $\mu = [\mu_1, \mu_2, \ldots, \mu_k]^T$ and covariance matrix Σ. Then,

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{(d)} N(0, \Sigma)$$

The Delta Method

The Delta Method allows us to find the limiting distribution of $g(Y_n)$ where g is a smooth function. Suppose that $\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{(d)} N(0, 1)$ and g is a differentiable function such that $g'(\mu) \neq 0$. Then,

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{(d)} N(0, g'(\mu)|^2 \sigma^2)$$

Multivariate version of the Delta method. Let $Y_n = (Y_{n1}, \ldots, Y_{nk})$ be a sequence random *vectors* such that $\sqrt{n}(Y_n - \mu) \xrightarrow{(d)} N(0, \Sigma)$. In this case $g : \mathbb{R}^k \to \mathbb{R}$. Then,

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{(d)} N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu))$$

Chapter 6 Statistical Inference

"...is the process of using data to infer the distribution that generated that data."

Point estimation:

Provides a *best guess* of some quantity of interest. The point estimate of θ, a fixed number, is $\hat{\theta}$, which depends on the data, therefore a r.v. A point estimator $\hat{\theta}_n$ of the parameter θ is **consistent** if $\hat{\theta}_n \xrightarrow{P} \theta$. The **bias** of an estimator is defined as $bias(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta$. An estimator is unbiased if $\mathbb{E}_\theta[\hat{\theta}_n] = 0$. Following the bias, the **mean squared error** (MSE) is $MSE = bias^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n)$.

Confidence Intervals:

Is an interval $I = (a, b)$ such that $\mathbb{P}_\theta(\theta \in I) \geq 1 - \alpha$, for all $\theta \in \Theta$. I is random, but θ is not random. Suppose that $\hat{\theta}_n \approx N(\theta, \sigma^2)$. The confidence interval will be of the form

$$I = \left[\hat{\theta}_n - q_{\alpha/2} \frac{\sqrt{\sigma^2}}{\sqrt{n}}, \hat{\theta}_n + q_{\alpha/2} \frac{\sqrt{\sigma^2}}{\sqrt{n}} \right]$$

However, most of the time, this confidence interval will depend on the true parameter θ. There are 3 methods to fix this problem.
Conservative Bound: Suppose we're calculating a confidence interval for the parameter p of a Bernoulli r.v. where $\sigma = \sqrt{p(1 - p)}$. In this case $p(1 - p)$ has the *upper bound* $1/4$. We can replace this value in I. Then $I = [\hat{p} - q_{\alpha/2} \frac{1}{2\sqrt{n}}, \hat{p} + q_{\alpha/2} \frac{1}{2\sqrt{n}}]$.
Plug-in: Since $\hat{\theta}$ is a consistent estimator of θ, we can simply replace θ with $\hat{\theta}$.
Solve the equation: A more elaborate method is to solve the system of two Inequalities w.r.t θ.

$$\hat{\theta} - q_{\alpha/2} \frac{\sqrt{\sigma^2}}{\sqrt{n}} \leq \theta \leq \hat{\theta} + q_{\alpha/2} \frac{\sqrt{\sigma^2}}{\sqrt{n}}$$

$$(\theta - \hat{\theta})^2 \leq q_{\alpha/2}^2 \frac{\sigma^2}{n}$$

Finding the roots of the quadratic equation will lead us to $I_{solve} = (\theta_1, \theta_2)$.

Unit 3 Mehods of Estimation

Distance TV and KL divergence

The **total variation** distance between two probabilities is defined as

$$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 1/2 \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|$$

We replace the $\sum$ for an integral in the case of continuous r.v.'s. The goal is to create an estimator that minimizes this distance, however it's unclear how to proceed.
Kullback-Leibler Also known as the relative entropy, mesure the divergence between two probabilities.

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \sum_{x \in E} p_\theta(x) log\left(\frac{p_\theta(x)}{p_{\theta'}(x)} \right)$$

Estimating the KL divergence yield the maximum likelihood principal presented bellow.

Chapter 9 Parametric Inference

Maximum Likelihood Estimator

The likelihood function is defined as

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

and the log-likelihood is defined as $l_n(\theta) = log(\mathcal{L}_n(\theta))$.
The **maximum likelihood estimator** is denoted by $\widehat{\theta}_n$ that maximizes $\mathcal{L}_n(\theta)$. Usually we follow theses steps to calculate the MLE of a parameter:

1. Calculate $\mathcal{L}_n(\theta)$

2. Calculate the log-likelihood, $l_n(\theta)$

3. Derivate w.r.t the parameter, θ, and set the derivative to 0

4. Solve for θ.

Here's a table of likelihood function for popular distributions.

Distribution	Likelihood Function
Bernoulli	$p^{\sum x_i}(1-p)^{n-\sum x_i}$
Poisson	$(\lambda^{\sum x_i}e^{-n\lambda})/\prod x_i$
Gaussian	$(1/2\pi\sigma^2)^{n/2}e^{-(\sum(x_i-\mu)^2)/(2\sigma^2))}$
Exponential	$\lambda^n e^{-\lambda \sum x_i}$
Uni[0, b]	$(1/b^n)\mathbb{1}(max(x_i)) \leq b)$

Asymptotic Normality of the MLE:

- The parameter is identifiable

- $\forall \theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ

- $\theta_\star$ is not on the boundary of Θ, otherwise $l(\theta)$ may not be differentiable

- $I(\theta)$ is invertible in the neighborhod of $\theta_\star$

- A few other technical conditions

Then, $\widehat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{P} \theta^*$ w.r.t. $\mathbb{P}_\theta^*$ and
$\sqrt{n}(\widehat{\theta}_n^{MLE} - \theta^*) \xrightarrow[n\to\infty]{d} N(0, I(\theta^*)^{-1})$, where $I(\theta^*)$ is the
Fisher Information, defined as $I(\theta) = -\mathbb{E}_\theta[\mathbb{H}l(\theta)] = -\mathbb{E}_\theta[l"(\theta)]$.

Method of Moments

Let $X_1, \ldots, X_n$ be an i.i.d sample associated with $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$.
$\longrightarrow$ Population moments: $m_k(\theta) = \mathbb{E}_\theta[X_1^k]$
$\longrightarrow$ Empirical moments: $\widehat{m}_k(\theta) = \overline{X}_n^k$
Example: We have a normal distribution model, $N(\mu, \sigma^2)$. Given data, the method of moments will estimate the parameters as:

$$\widehat{m}_1 = \widehat{\mu} = \overline{X}_n$$
$$\widehat{m}_2 = \widehat{\mu^2} + \widehat{\sigma^2} = \overline{X_n^2}$$

All that is left is to solve the equations system.

M-Estimators

Goal: estimate the parameter μ^* associated with some unknown distribution $\mathbb{P}$, e.g.: it's mean, median, variance, etc. Find some function $\rho : E \times M \to \mathbb{R}$, where M is the set of all possible values of μ^* such that:

$$\mathcal{Q}(\mu) := \mathbb{E}[\rho(X_1, \mu)]$$

achives its minimum at $\mu = \mu^*$.
Examples:
If $E = M = \mathbb{R}$ and $\rho(x, \mu) = (x - \mu)^2; \forall(x, \mu) \in \mathbb{R}; \mu^* = \mathbb{E}[X]$
If $E = M = \mathbb{R}$ and $\rho(x, \mu) = |x - \mu|; \forall(x, \mu) \in \mathbb{R}; \mu^*$ is the median of $\mathbb{P}$
The empirical median is always more robust than the mean, meaning that it's less affected by missing or wrong data.

All the matrices/vectors

Covariance Matrix(Σ): formed by the $var(X_i)$ in the diagonals and $Cov(X_i, X_j)$ for the other terms.

$$\Sigma = \begin{bmatrix} var(X_1) & \ldots & cov(X_1, X_d) \\ \vdots & \vdots & \vdots \\ cov(X_d, X_1) & \ldots & var(X_d) \end{bmatrix}_{d\times d}$$

Gradient(∇): first derivative matrix of a function matrix $g(\theta)$

$$\nabla g(\theta) = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \theta_1} & & \frac{\partial g_k(\theta)}{\partial \theta_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial g_1(\theta)}{\partial \theta_d} & & \frac{\partial g_k(\theta)}{\partial \theta_d} \end{bmatrix}_{d\times k}$$

Hessian($\mathbb{H}$): secong derivative matrix of a function matrix $g(\theta)$

$$\mathbb{H}g(\theta) = \begin{bmatrix} \frac{\partial^2 g(\theta)}{\partial \theta_1^2} & & \frac{\partial^2 g(\theta)}{\partial \theta_d \theta_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 g(\theta)}{\partial \theta_1 \theta_d} & & \frac{\partial^2 g(\theta)}{\partial \theta_d^2} \end{bmatrix}_{d\times d}$$

Note: the function $g(\theta)$ is said to be concave if $x^T\mathbb{H}g(\theta)x \leq 0, \forall x \in \mathbb{R}^d$.
Fisher Information Matrix ($I(\theta)$): is defined as $-\mathbb{E}_\theta[\mathbb{H}l(\theta)]$

Multivariate Distributions
Normal

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Huber's Loss

The Huber's Loss function allow us to render absolute functions differentiable, for example $|x|$, which is not differentiable at 0.

$$h_\delta(a) = \begin{cases} a^2/2, & |a| \leq \delta \\ \delta(|a| - \delta/2), & \text{otherwise} \end{cases}$$

$$h'_\delta(a) = \begin{cases} 2a, & |a| \leq \delta \\ \pm\delta, & \text{otherwise} \end{cases}$$

Unit 4 Hypothesis Testing

Chapter 10

Let X be a r.v. and let χ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \chi$ called rejection region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject it. The null hypothesis is called $H_0 : \theta \in \Theta_0$ and the alternative hypothesis is called $H_1 : \theta \in \Theta_1$ and we retain H_0 unless there is strong evidence to reject it. Θ_0 *and* Θ_1 are disjoint subsets of the sample space. We can make 2 errors when testing a hypothesis:
Type 1 error: Rejecting H_0 when H_0 is true.
Type 2 error: Retaining H_0 when H_1 is true.
The **power function** of a test is defined as
$\beta(\theta) = \mathbb{P}_\theta(X \in R)$, and the **size** of a test is defined to be
$\alpha = \max_{\theta\in\Theta_0}\beta(\theta)$. The test is then said to have level α.

Wald's Test

Consider testting $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ Assuming that $\widehat{\theta}$ is asymptotically Normal we have :

$$W = \frac{(\widehat{\theta} - \theta_0)}{\widehat{se}}, \text{ where } \widehat{se} = \sqrt{\sigma_0^2/n}$$

The size α Wald test will reject H_0 when $|W| > z_{\alpha/2}$.
Note: in the case of a two samples test (e.g.: comparing two means), $\widehat{se} = \sqrt{(\sigma_1^2/n) + (\sigma_2^2/m)}$, where m and n are the sample sizes.

p–values

Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, the p–value is the smallest level α at which we can reject H_0. Therefore, the smaller the p–value, the stronger the evidence against H_0. **Example:** We toss a coin 30 times and get 13 heads. We want to test $H_0 : \theta = 1/2$ vs. $H_1 : \theta \neq 1/2$. By CTL:

$$\sqrt{n}\frac{\overline{X}_n - \theta_{H_0}}{\sigma_{H_0}} \approx 0.77$$

Then the p–value is $\mathbb{P}_\theta(|Z| > 0.77) = 2\mathbb{P}_\theta(Z < -0.77) \approx 0.44$

The χ^2 Distribution

The χ^2 distribution is defined as $V = \sum_{i=1}^{k} Z_i^2$, where the Z_i are standard normal r.v.'s. We say that V has k degrees of freedom, which coincides with the $\mathbb{E}[V]$, while the $var(V) = k^2$. The PDF of V is:

$$f(v) = \frac{v^{k/2-1}e^{-v/2}}{2^{k/2}\Gamma(k/2)}, \text{ for } v > 0$$

Cochran's Theorem

For $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then $\overline{X_n} \perp\!\!\!\perp S_n$, $\forall n$, where S_n is the sample variance. Furthermore, $\frac{nS_n}{\sigma^2} \sim \chi_{n-1}^2$. We often prefer the unbiased variance estimator $\tilde{S}_n = \frac{nS_n}{n-1}$.

Student-T Distribution

The Student-T distribution with d degrees of freedom, denoted t_d, is the law of r.v.'s that follow

$$\frac{Z}{\sqrt{V/d}},$$

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

From the student-T distribution, we derive the **t-Test**. Specially usefull when the sample sizes are small and one can not make use of the large numbers theorems (CTL, Slutsky's, etc.).

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, with unknown parameters. We want to test $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. Our test statistic is :

$$T_n = \frac{\overline{X_n} - \mu}{\sqrt{\tilde{S}_n/n}} = \sqrt{n}\frac{\overline{X_n}/\sigma}{\sqrt{\tilde{S}_n/\sigma^2}} \sim \frac{N(0,1)}{\chi_n^2 - 1/(n-1)}$$

Since the numerator of the equation above $\sim N(0,1)$, under H_0, and the denominator $\sim \chi_n^2 - 1/(n-1)$ they are independent by Cochran's theorem. So finally we have a Student-T distribution $T_n \sim t_{n-1}$, and the test with non-asymptotic level α is $\psi = \mathbb{1}(|T_n| > q_{\alpha/2})$ where $q_{\alpha/2}$ is the $\alpha/2$ quantile of t_{n-1}.

In the case of a two samples T-test T_n has the following form:

$$T_n = \frac{\overline{X_n} - \overline{Y_m} - (\mu_x - \mu_y)}{\sqrt{(\sigma_x^2/n) + (\sigma_y^2/m)}} \sim t_N$$

where N is given by the Welch-Satterthwaite formula :

$$N = \frac{(\sigma_x^2/n + \sigma_y^2/m)^2}{\frac{\sigma_x^4}{n^2(n-1)} + \frac{\sigma_y^4}{m^2(m-1)}} \geq min(n,m)$$

Some advantages and drawbacks of the T-test are:

- Non asymptotic
- Can be run on small samples sizes
- Can also be run on large sample sizes
- Assumes the data is Gaussian

MLE based test

Let $\widehat{\theta}^{MLE}$ be the MLE of θ and assume that it's technical conditions are satisfied. From the Wald's test we have under H_0 :

$$\sqrt{n}\, I(\theta)^{1/2}(\widehat{\theta}^{MLE} - \theta_0) \xrightarrow[n\to\infty]{(d)} N_d(0, I_d)$$

The goal of the test is to measure the Eucledian distance between two vectors, we end up with :

$$T_n = n\,(\widehat{\theta}^{MLE} - \theta_0)^T\, I(\widehat{\theta}^{MLE})\,(\widehat{\theta}^{MLE} - \theta_0) \to \chi_d^2$$

Note: The Wald's test is still valid if the test is one sided, however it's less powerfull.

Log-Likelihood Test

The likelihood test is usefull for testing vector-valued parameters. Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. The test statistic is :

$$T_n = 2(l_n(\widehat{\theta_n}) - l_n(\widehat{\theta_0})) \xrightarrow[n\to\infty]{(d)} \chi_{r-q}^2$$

Here θ is of the form $(\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$ and the parameter space $\Theta_0 = \{\theta : (\theta_{q+1}, \ldots, \theta_r) = (\theta_{q+1}^{(0)}, \ldots, \theta_r^{(0)})\}$. Both $\widehat{\theta}$'s are MLE's, but $\widehat{\theta_0}$ is constrained to the parameter space Θ_0. In english, we're testing whether a subset of the parameter vector θ equals the vector Θ_0. For example, testing whether $\theta_4 = \theta_5 = 0$ in $\theta = (\theta_1, \ldots, \theta_5)$.

Goodness of Fit Tests

GOF's tests are used to check if the data come from an assumed parametric model. For example if the data is Gaussian, uniform or comes from a multinomial distribution.

Chi-Squared Test - Discrete distributions

The multinomial likelihood is $L_n(X_1, \ldots, X_n; \vec{p}) = p_1^{N_1} p_2^{N_2} \ldots p_k^{N_k}$, where N_k is the number of times that $x = X_i$. The MLE is $\widehat{p}_j = N_j/n$ and $\sqrt{n}(\widehat{p} - \vec{p_0})$ is asymptotically normal under H_0. Therefore the following theorem holds true :

$$T_n = n\sum_{j=1}^{K} \frac{(\widehat{p}_j - p_j^0)^2}{p_j^0} \xrightarrow[n\to\infty]{(d)} \chi_{K-1}^2$$

Note: this test does not work for continuous distribution. For example, to test a normal distribution we would need to "bin" our data and test the probability of falling under on of the bins. To fully analyse continuous distributions we need the CDF and the Empirical CDF.

Empirical CDF

Let $X_1, \ldots, X_n$ be real r.v., the CDF is defined as $F(t) = \mathbb{P}[X_i \leq t] \,\forall t \in \mathbb{R}$. Now the empirical CDF (a.k.a sample CDF) is defined as :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq t) \,\forall t \in \mathbb{R}$$

By the strong law of large numbers $F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$, but this is only a pointwise convergence. To have a *uniform convergence* of the function we need the Glivenko-Canteli theorem (also known as the fundamental theorem of statistics), defined as :

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$$

By the CTL, $\sqrt{n}(F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2)$ and since the empirical CDF is a indicator functio, it's asymptotic variance is $F_n(t)(1 - F_n(t))$. Once again, the goal is to measure the distance between two functions.

Kolmogorov-Smirnov Test

Considering $H_0 : F = F_0$ v.s $H_1 : F \neq F_0$, the KS test is defined as :

$$T_n = \sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F(t)_0| \xrightarrow[n\to\infty]{(d)} \sup_{0\leq t\leq 1} |\mathbb{B}(t)|$$

where $\mathbb{B}(t)$ is a Brownian bridge distribution (Donker's Theorem), which is a pivotal distribution and it's quantiles can be obtained in tables. The Ks test with asymptotic level α is $\delta_\alpha^{KS} = \mathbb{1}(T_n > q_\alpha)$, where q_α is the (1-α) of $\sup |\mathbb{B}(t)|$.

In real life, we compute T_n as follows :

1. F_0 is non-decreasing function, F_n is piecewise constant, with jumps at $t_i = X_i$, $i = 1, \ldots, n$.

2. Let X be the *reordered* sample.

3. The expression for T_n reduces to the following pratical formula:

$$\sqrt{n} \max_{i=1,\ldots,n} \left\{ max\left(\left|\frac{i-1}{n} - F_0(X_i)\right|, \left|\frac{i}{n} - F_0(X_i)\right| \right) \right\}$$

4. Under H_0, T_n is a *pivotal statistic* and does not depend on the distribution of the X_i's but it depends on n. It is easy to reproduce in simulations and find the values in KS tables.

Other GOFs test exist such as the Cramer-Von Mises (L_2 distance) and the Anderson-Darling tests.

Kolmogorov-Lillofors Test

Now, if we want to check if our data is Gaussin and we use a plug-in estimators for μ and σ^2 in the KS test, Donker's theorem *is no longer valid*. Instead, we compute que quantiles for the test statistic:

$$T_n = \sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - \psi_{\widehat{\mu},\widehat{\sigma^2}}(t)|$$

This test statistic does not depend on the unknown parameters and it's quantiles can be obtained in K-L tables for different values of n.

Unit 5

Bayesian Inference

In the Bayesian approach we use our belief or prior knowlage of what the parameter vector θ might be. Here's how it is carried out :

1. We choose a *prior distribution* that expresses our beliefs about θ before seeing the data.

2. We define the a statistical model, $f(x|\theta)$ that reflects our data given our parameter.

3. After observing the data we update our belief and calculate a *posterior distribution* $f(\theta|X_1, \ldots, X_n)$.

Bayes' Method

From Bayes' Theorem we can compute the prior distribution as follows :

$$\pi(\theta|X^n) = \frac{\mathcal{L}(X^n|\theta)\pi(\theta)}{f(X^n)} \propto \mathcal{L}(X^n|\theta)\pi(\theta)$$

In the formula above, $f(X^n)$ is a *normalizing constant* and it can be removed from the equation since it does not depend on θ. The posterior is the said to be *proportional* to the likelihood times (or weigthed by) the prior. Since the posterior distribution has to integrate to 1 on it's sample space, we can always recover the constant later if needed.

Bayes' Interval Estimate

One can estimate a Bayesian interval by finding a and b such that $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$. Let $C = (a, b)$, then :

$$\mathbb{P}(\theta \in C|x^n) \int_{b}^{a} f(\theta|x^n)d\theta = 1 - \alpha$$

so C is a $1 - \alpha$ *posterior interval*.

Improper Priors & "Noninformative" Priors

If we have no information about the parameter, how to pick a prior ? A good candidate is the *constant prior*, for example $\pi(\theta) \propto 1$. If the parameter space Θ is bounded ($[0, 1]$), then this is juste the *uniform prior* on Θ. However, if Θ is not bounded, then this is not a proper prior since it doesn't integrate to 1. We define an *improper prior* as a mesurable, nonnegative function $\pi(.)$ defined on Θ that is not integrable. However we can still define a posterior with an improper prior.

Conjugate Priors

When the posterior distribution belongs to the same distribution family as the prior distribution, the prior is called a *conjugate prior* to the likelihood model. Here're some common conjugate priors: and models :

Prior	Model	Posterior Parameters
Beta	Bernoulli	$(\alpha + \sum x_i, \beta + n - \sum x_i)$
Beta	Binomial	$(\alpha + \sum x_i, \beta + N_i - \sum x_i)$
Beta	Geometric	$(\alpha + n, \beta + \sum x_i)$
Gamma	Poisson	$(\alpha + \sum x_i, \beta + n)$
Gamma	Exponential	$(\alpha + n, \beta + \sum x_i)$
Normal	Normal	See Wikipedia table.

Jeffreys' Prior

Jeffrey created a rule to choose priors and that is as follows :

$$\pi_J(\theta) \propto \sqrt{det(I(\theta))}$$

Since a *higher* Fisher Information is associated with an *easier point to estimate* (less randomness in the data) than one with lower Fisher Information, we're trying to balance that by using it as a prior in the Bayes approach. The effect is that we're putting more weight to the points that are easier to estimate. The Jeffreys' prior satisfies a *representation invariance principal.* If $\eta = \psi(\theta)$ in an invertible function then the PDF $\tilde{\pi}(.)$ of η satisfies $\pi(\eta) \propto \sqrt{\tilde{I}(\eta)}$. We can write $\tilde{\pi}(\eta)$ in term of $\phi'(.)$ and $\phi^{-1}(.)$:

$$\tilde{\pi}(\eta) = \frac{\pi(\phi^{-1}(\eta))}{|\phi'(\phi^{-1}(\eta))|}$$

Some applications of this property would be :

- Compute the prior for a $Ber(q^{10})$ instead of $Ber(p)$
- Compute the prior for $Exp(1/\lambda)$ instead of $Exp(\lambda)$

Bayesian Estimation

The Bayes approach can also be used to estimate the true underlined parameter, however in a frequentist manner. One would start bu calculating the posterior, nd then the posterior *mean, meadian or mode* can be used as estimator. Another popular choice is to use the Maximum A Posteriori (MAP) estimator.

$$\widehat{\theta}^{MAP} = \underset{\theta \in \Theta}{argmax}\, \pi(\theta|X_1, \ldots, X_n)$$

Unit 6 - Regression

Linear Regression

Regression is a method for studying the relationship between a *response* variable Y and a *covariate* variable X (often called a feature is the ML world). One important fact about regression is that it exhibits *correlations*, **not** *casuality*. Here we summarize this relationship via the *regression function*:

$$\mu(x) = \mathbb{E}[Y|X = x] = \int y f(y|x)dy$$

The goal is to estimate the regression function from the data of the form $(X_1, Y_1), \ldots, (X_i, Y_i) \sim F_{Y,X}$. When we assume $\mu(x)$ is linear this is called *linear regression* and it's simplest form is $\mu(x) = \beta_0 + \beta_1 x$. To estimate the regression function we need some modeling assumptions:

- (X_i, Y_i) are i.i.d from some unknown joint distribution;
- $var(X) > 0$ and $var(Y|X = x) > 0$;
- This distribution is best described by $h(y|x) = \frac{h(x,y)}{h(x)}$. It contains all the information of Y given X.

From above it's clear that if $var(Y|X = x) > 0$, the points do not follow a perfect line. Therefore we introduce a *noise* (ϵ) term to the theorical linear regression function, which gives $Y = a^* + b^*x + \epsilon$.

Least Squares Estimator(LSE)

The LSE are the values of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimizes the *residual sums of squares* or $RSS = \sum_{i=1}^{n} \widehat{\epsilon}_i^2$. Where the residuals, $\widehat{\epsilon}_i$ are the vertical distance between the regression function and the Y_i's. We can also use the following form:

$$\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

The LSE's are given by:

$$\widehat{\beta}_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}$$

$$\widehat{\beta}_0 = \overline{Y_n} - \widehat{\beta}_1 \overline{X_n}$$

When we add the assumption that $\epsilon \sim N(0, \sigma^2)$ given X, $Y_i|X_i \sim N(\mu_i, \sigma^2)$. Under assumption of *Normality*, the LSE is the same as the MLE where $\widehat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}\widehat{\epsilon}_i^2$ is an unbiased estimator of the σ^2.
Note: one can use Goodness of Fit to test whether the residuals are Gaussian.

Multivariate Regression and Matrix form

The covariates may come in a vector form and we'll en up with a $\mathbb{X}_{(n \times p)}$ *design matrix* of covariates samples. The regression function still works in a multi dimentional space and is of the form:

$$\mathbf{Y}_i = \mathbb{X}_i^T \beta + \epsilon_{\mathbf{i}}, \text{ and satisfies}$$
$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{argmax}||\mathbf{Y} - \mathbb{X}^T \beta||_2^2$$

Assuming that $rank(\mathbb{X}) = p$, which means that $\mathbb{X}$ is invertable, we have a closed form solution for the LSE as $\widehat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1}\mathbb{X}^T \mathbf{Y}$.
The *geometric* interpretation of the LSE: $\mathbb{X}\beta$ is the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbb{X}$.

LSE Inference and Testing

However, to make further inference we need to make additional assumptions:

- The design matrix $\mathbb{X}$ is **deterministic** and invertible
- The model is *homoscedastic*: ε_i are i.i.d
- The noise vector is Gaussian $N(0, \sigma^2)$, which leads to $Y_i|X_i \sim N(\mathbb{X}(\beta)^*, \sigma^2 I_p)$

*Note : if $\mathbb{X}$ is not deterministic, all the above can be understood **conditionally** on $\mathbb{X}$.*
Theses assumptions give rise to the following properties:

- LSE = MLE
- $\widehat{\beta} \sim N(\beta^*, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$
- Quadratic risk of $\widehat{\beta}$ is $\sigma^2 tr((\mathbb{X}^T\mathbb{X})^{-1})$
- The prediction error is $\sigma^2(n - p)$

By Cochran's theorem, $\widehat{\beta} \perp\!\!\!\perp \widehat{\sigma}^2$, therefore we can use a modified T-test to test the significance of the j-th explanatory variable in the linear regression. For example, let $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$.

$$T_n^{(j)} = \left|\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \gamma_j}}\right| \sim t_{n-p}$$

where γ_j is the j-th *diagonal* element is the $(\mathbb{X}^T\mathbb{X})^{-1}$ matrix. The test with non asymptotic level α is $\psi_j = \{T_n^{(j)} > q_{\alpha/2}(t_{n-p})\}$. In the case where we're testing multiple explanatory variable et use the **Bonferroni's Test**. Let $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$, $(\forall j \in S)$, where $S \subseteq \{1, \ldots, p\}$. To have non asymptotic level α one has to use α/k, where $k = |S|$ (number of tests).