Summarise Course/Methods

Siger

11 décembre 2022

Si Dieu est infini, alors je suis une partie de Dieu sinon je serai sa limite...

Table des matières

Première partie

Mathematical background

1. Probability's points of view

Statistical inference lets us do 2 things :

1. Estimating teh parameters of a statistical model
2. Testing statistical hypotheses

1.1. Frequentist's point of view

1.1.1. Introduction

This approach avoid treating parameters like random variables, and which thus avoid the use of priors and Bayes rule.

1.1.2. Sampling distribution of an estimator

In frequentist statistic a parameter estimate $\hat{\theta}$ is computed by applying an estimator δ to some data $\mathcal{D}$, so $\hat{\theta} = \delta(\mathcal{D})$. The uncertainty in the parameter estimate can be measured by computing the *sampling distribution* of the estimator. Imagine sampling many different datasets $\mathcal{D}^{(s)}$ from some true model $p(\cdot|\theta^*)$ meaning $\mathcal{D}^{(s)} = \left\{x_i^{(s)} \hookrightarrow p(\cdot|\theta^*)\right\}_{1\leq i\leq N}$ for $1 \leq s \leq S$ and θ^* is the true parameter. Now apply the estimator $\hat{\theta}(\cdot)$ to each $\mathcal{D}^{(s)}$ to get a set of estimates $\{\hat{\theta}(\mathcal{D}^{(s)})\}_{1\leq s\leq S}$.
As we late $S \to \infty$, the distribution induced on $\hat{\theta}(\cdot)$ is the **sampling distribution of the estimator**.

Bootstrap It is a simple *Monte Carlo* technique to approximate the sampling distribution. The idea is that if we knew the true parameters θ^*, we could generate S fake datasets of size N, from the true distribution. We could then compute our estimator from each sample, and use the empirical distribution of the resulting samples as our estimate of the sampling distribution.
Since θ is unknown, the idea of the **parametric bootstrap** is to generate the samples using $\hat{\theta}(\mathcal{D})$ instead. An alternative, called **non-parametric bootstrap** is to sample the x_i^s (with replacement) from the original data $\mathcal{D}$ and then compute the induced distribution as before.

1.1.3. Fequentist decision theory

In Frequentist decision theory there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus there is no automatic

way of deriving an optimal estimator, unlike the Bayesian case.
Instead, we are free to choose any estimator or decision procedure $\delta : \mathcal{X} \to \mathcal{A}$ we want. Having chosen an estimator, we define its expected loss or risk as follows

$$R(\theta^*, \delta) \triangleq \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)} \left(L(\theta^*, \delta(\tilde{\mathcal{D}})) \right) = \int L\left(\theta^*, \delta(\tilde{\mathcal{D}}) p(\tilde{\mathcal{D}})\right) d\tilde{\mathcal{D}}$$

where $\tilde{\mathcal{D}}$ is data sampled from 'nature's distribution' which is represented by parameter θ^*. Whereas the Bayesian posterior expected loss :

$$p(a, \mathcal{D}, \pi) \triangleq \mathbb{E}_{p(\theta|\mathcal{D},\pi)} \left(L(\theta, a) = \int_{\Theta} L(\theta, \boldsymbol{a}) p(\theta|\mathcal{D}, \pi) d\theta \right)$$

We see that the Bayesian approach averages over θ, which is unknown, and conditions on $\mathcal{D}$ which is known. Unlike the frequentist approach averages over $\tilde{\mathcal{D}}$, thus ignoring the observed data, and conditions on θ^* which is unknown.

Bayes risk How to chose amongst the estimators ? We need some way to convert $R(\theta^*, \delta)$ into single measure of quality, $R(\delta)$ which does not depend on knowing θ^*. One approach is to put a prior on θ^* and then to define **Bayes risk** of an estimator as follows :

$$R_B(\delta) \triangleq \mathbb{E}_{p(\theta^*)} \left(R(\theta^*, \delta) \right) \int R(\theta^*, \delta) p(\theta^*) d\theta^*$$

A **Bayes estimator** or **Bayes decision rule** is one which minimizes the expected risk : $\delta_B \triangleq \arg\min_{\delta} R_B(\delta)$

Connection Bayesian and Frequentist approaches to decision theory.

— *Theorem 1* A Bayes estimator can be optained by minimizing the posterior expected loss for each $\boldsymbol{x}$

— *Theorem 2* Every admissible decision rule is a Bayes decision rule with respect to some possibly improper prior distribution.

Minimax risk Some frequentist statistic users avoid using Bayes risk since it requires the choice of a prior, although this is only in the evaluation of the estimator, not necessarily as part of its construction. An alternative approach is as follows :

1. Define the maximum risk of an estimator as :
   $R_{max}(\delta) \triangleq \theta^* R(\boldsymbol{\theta}^*, \delta)$
2. A **minimax rule** is one which minimizes the maximum risk : $\delta_{MM} \triangleq \arg\min_{\delta} R_{\max}(\delta)$

Minimax estimators have a certain appeal, however computing them can be hard and furthermore they are very pessimistic. In most statistical situations, excluding games theoretic ones, assuming nature is an adversary is not a reasonable assumption.

Admissible estimators The basic problem with frequentis decision theory is that it relies on knowing the true distribution $p(\cdot|\theta^*)$ in order to evaluate the risk. However it might be the case that some estimators are worse than others regardless of the value of θ^*.
In particular if for $\theta \in \Theta, R(\theta, \delta_1) \leq R(\theta, \delta_2)$ then we say that δ_1 **dominates** δ_2.
An estimator is said to be **admissible** if it is not strictly dominated by any other estimator.
Admissibility is not enough

1.1.4. Desirable properties of estimators

Consistent estimators An estimator is said to be **consistent** if it eventually recovers the true parameters that generated the data as the sample size goes to infinity.

Unbiased estimator The **bias** of an estimator is defined as

$$bias\left(\hat{\theta}(\cdot)\right) = \mathbb{E}_{p(\mathcal{D}|\theta^*)}\left(\hat{\theta}(\mathcal{D}) - \theta^*\right)$$

The estimator is unbiased when the bias is equal to 0.

Minimum variance estimators A famous result called the **Cramerè-Rao lower bound** provides a lower bound on the variance of any unbiased estimator. More precisely : Let $(X_j)_{1\leq j\leq p} \hookrightarrow p(X|\theta_0)$ and $\hat{\theta}(\cdot)$ an unbiased estimator of θ^* Then, under various smoothness assumptions on $p(X|\theta_0)$ we have

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{nI(\theta^*)}$$

where $I(\theta^*)$ is the Fisher information matrix.

Bias-Variance Trade-off As $MSE = variance + bias^2$
It might be wise to use a biased estimator, so long as it reduces our variance, assuming our goal is to minimize squared error.

1.1.5. Empirical risk minimization

Frequentist decision theory suffers from the fundamental problem that one cannot actually compute the risk function, since it relies on knowing the true data distribution. By contrast, the Bayesian posterior expected loss can always be computed since it conditions on the data rather that on θ^*.
However there is one setting which avoids this problem, it is when the task is to predict observable quantities, as opposed to estimating hidden variables or parameters.
Instead of looking at loss functions of the form $L(\boldsymbol{\theta^*}, \delta(\mathcal{D}))$ let us look at loss functions of the form $L(y, \delta(\boldsymbol{x}))$.
Then the risk becomes : $R(p_*, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\hookrightarrow p_*}\left(L(y, \delta(\boldsymbol{x}))\right) = \sum_{\boldsymbol{x}}\sum_{\boldsymbol{y}} L(y, \delta(\boldsymbol{x}))p_*(\boldsymbol{x}, y)$

Where p_* represents "nature's distribution", indeed this distribution is unknown, but a simple approach is to use the empirical distribution, derived from some training data to approximate p_* $p_{emp} \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}(\boldsymbol{x})\delta_{y_i}(y) \approx p_*(\boldsymbol{x}, y)$ We define the empirical risk as follows :

$$R_{emp}(\mathcal{D}, \mathcal{D}) \triangleq R(p_{emp}, \delta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \delta(x_i))$$

1.1.6. Tools

Introduction Avoid treating parameters as randome variables. The notion of variation across repeated trials forms the basis for modelling uncertainity.

Hypothesis Testing A *frequentist* statistics, probabilities represent the frequencies at which particular events happen.

p-value It is the heart of frequentist hypothesis testing, it tells us the probability of getting a particular test statistic t as big as the one we have or bigger under the null hypothesis (that there is actually no effect).
By convention we usually conclude an effect is *statistically significant* if the *p-value* is less than a threshold α.

Confidence intervals When we fit a model to our data we look for the *maximum of likelihood* parameters, meaning the parameters that are most consistent with our data. For each parameter we will able to construct 95% interval namely 95 of the 100 intervals generated will contain the true value of the parameter.
If $H_0 : \beta = 0$ is true, the probability of getting a 95% confidence interval that does not include 0 is less than 0.05. In other words, if the 95% confidence does not include 0, $p < 0.05$.

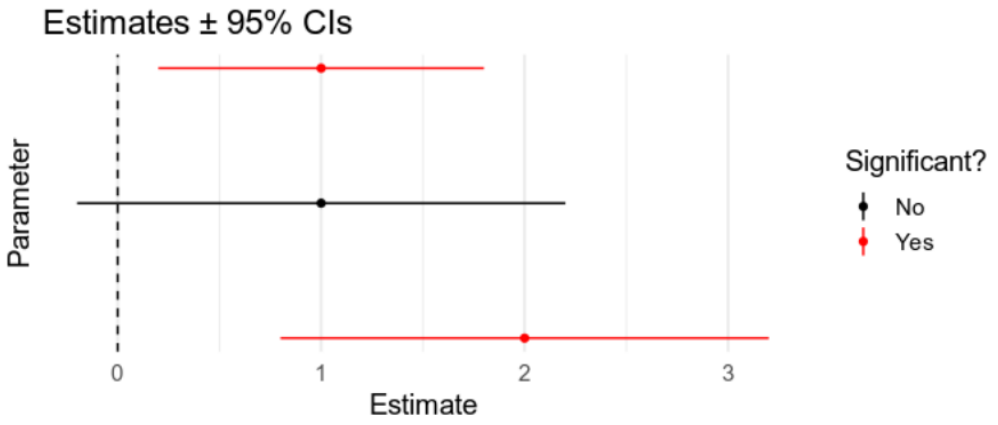

FIGURE 1.1. – Confidence interval

Multiple comparisons The more tests we run the more likely it is to we'll find at least one that is significant even though the null hypothesis is true. We can then apply a Bonferroni correction.
Let's say we are running k tests, we can either adjust :

— the threshold $\alpha_{adj} = \frac{\alpha}{k}$ OR

— the *p-value* $p_{ajd} = k \times p$

1.2. Bayesian's point of view

1.2.1. Tools

Summarizing posterior distributions

MAP Although most appropriate choice for : $\begin{cases} \text{Real valued quantity} & \rightarrow \textit{posterior median or mean} \\ \text{Discrete} & \rightarrow \textit{vector of posterior marginals} \end{cases}$
The most popular choice is *posterior mode* aka MAP, because it reduces to optimization problems for which efficient algorithms often exist.
Some point to be aware about MAP :

— No measure of uncertainty

— Plugging in the MAP estimate can result in overfitting

— The mode is an untypical point, unlike the mean or median the mode is a point of measure 0, it does not take the volume of the space into account.

— MAP estimation is not invariant to reparameterization, for example passing from centimeters to inches can break things.)

The MLE does not suffer from this since the likelihood is a function not a probability density

Credible Intervals With point estimates, we want a measure of confidence.

$$C_\alpha(\mathcal{D}) = (l, u) : \mathbb{P}(\{l \leq \theta \leq u | \mathcal{D}\})$$

In general, credible intervals are usually what people want to compute but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics.
Sometimes with central intervals there might be points be outside the CI which have higher probability density.
More formally p^* such that :

$$1 - \alpha = \int_{\theta : p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta$$

Then the HPD such that :

$$\mathcal{D} = \{\theta : p(\theta|\mathcal{D}) \geq p^*\}$$

Bayesian model selection A more efficient approach than cross-validation, meaning fitting k times each model, is to compute the posterior over models.

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(m|\mathcal{D})}$$

From this we can compute the MAP model $\hat{m} = \arg\max_m p(m|\mathcal{D})$
Then we have the marginal likelihood : $p(\mathcal{D}|\hat{m}) = \int p(\mathcal{D}|\hat{m})p(\theta|\hat{m})d\theta$

Baysian Occam's razor In integrating out the parameters rather than maximizing them we are automatically protected from overfitting : model with more parameters do not necessarily have higher marginal likelihood.
A way to understand the Bayesian Occam's razor effect is to remember that probabilities must sum to one, meaning $\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models.

Computing the marginal likelihood (evidence) For a fixed model we often write :

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m)$$

This valid since $p(\mathcal{D}|m)$ is constant. However when comparing models we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$. In general this can be quite hard,

since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute.

Let $p(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{Z_0}$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior. Let $p(\mathcal{D}|\boldsymbol{\theta}) = \frac{q(\mathcal{D}|\boldsymbol{\theta})}{Z_l}$ be the likelihood, where Z_l contains any constant factors in the likelihood. Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N}$ be our posterior where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior.

We have : $\begin{cases} p(\boldsymbol{\theta}) = \dfrac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ \dfrac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \dfrac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_l Z_0 p(\mathcal{D})} \\ p(\mathcal{D}) = \dfrac{Z_N}{Z_0 Z_l} \end{cases}$

In general $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ can be quite difficult to compute. Simpler approach

— **BIC** simple approximation : $BIC \triangleq \log(p(\mathcal{D}|\hat{\boldsymbol{\theta}})) - \frac{dof(\hat{\boldsymbol{\theta}})}{2}\log(N) \approx \log p(\mathcal{D})$

— **AIC** : $AIC(m, \mathcal{D}) \triangleq \log(p(\mathcal{D})\hat{\boldsymbol{\theta}}_{MLE}) - dof(m)$
This is derived from Frequentits framework and cannot be interpreted as an approximation to the marginal likelihood. The penalty of AIC is less than BIC, it causes AIC pick more complex models. That can be better for predictive accuracy.

— Effect of the prior.
If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is we should put a prior on the hyper-parameter α as well as the parameters $\boldsymbol{w}$. To compute the marginal likelihood we should integrate out all unknowns, we should compute : $\int\int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)p(\alpha|m)d\boldsymbol{w}d\alpha$ A computational shortcut is to optimize α rather than integrating it out. That is, we use $p(\mathcal{D}|m) \approx \int p(\mathcal{D}\boldsymbol{w})p(\boldsymbol{w}|\alpha, m)d\boldsymbol{w}$. where $\hat{\alpha} = \arg\max_{\alpha} p(\mathcal{D}|\alpha, m) = \arg\max_{\alpha} \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\hat{\alpha}, m)d\boldsymbol{w}$

Bayes Factors When prior on models is uniform, then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null hypothesis, M_0 and the alternative hypothesis, M_1.

Bayes Factor $BF(1,0)$	**Interpretation**
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Modest evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Modest evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{\frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})}}{\frac{p(M_1)}{p(M_0)}}$$

This is like a likelihood ratio, except we integrate out the parameters, which allows us to compare models of different complexity.

Jeffreys-Lindley paradox Problems can arise when we use improper priors (i.e. priors that do not integrate to 1) for model selection/ hypothesis testing, even though such priors may be acceptable for other purposes. Thus it is important to use proper priors when doing model selection.

Priors The most controversial aspect of Bayesian statistics is its reliance on priors

Uninformative priors If we do not have strong evidence on what θ should be, it is common to use an uninformative priors, to "let the data speak for itself".
One might think that the most uninformative prior would be the uniform distribution : $Beta(1,1)$, but the posterior would then be : $\mathbb{E}(\theta|\mathcal{D}) = \frac{N_1+1}{N_1+N_0+2}$, whereas the MLE is $\frac{N_1}{N_1+N_0}$.
As by decreasing the magnitude of the pseudo counts, we can lessen the impact of the prior, we can argue that the most non-informative prior is :

$$\lim_{\epsilon \to 0} Beta(\epsilon, \epsilon) = Beta(0,0)$$

Called the *Haldane prior*, it is an improper prior.
In general it is advisable to perform a some kind of sensitivity analysis, in which one checks how much one's conclusions or prediction change in response to change in the modelling assumptions which includes the choice of the prior and the likelihood as well. If the conclusion are relatively insensitive to the modelling assumption, one can have more confidence in the results.

Jeffreys priors Harold Jeffreys designed a general purpose technique for creating non-informative priors. The key observation is that if $p(\phi)$ is non-informative then any re-parametrization of the prior, such as $\theta = h(\phi)$ for some function h should also be non-informative.

— Start with a variable change : $p_\theta(\theta) = p_\phi(\phi) \left| \frac{d\phi}{d\theta} \right|$

— Consider the following constraint : $p_\phi(\phi) \propto \sqrt{\mathcal{I}(\phi)}$, where $\mathcal{I}(\phi)$ is the Fisher information.
$\mathcal{I}(\phi) \triangleq -\mathbb{E}\left(2 \times \frac{d \log\left(p(X|\phi)\right)}{d\phi}\right)$. This a measure of the curvature of the expected negative log likelihood and hence a measure of stability of the MLE.

— Now $\frac{d \log(p(x|\theta))}{d\theta} = \frac{d \log(p(X|\phi))}{d\phi} \frac{d\phi}{d\theta}$

— $\mathcal{I}(\theta) = \mathcal{I}(\phi) \left(\frac{d\phi}{d\theta}\right)^2$

— $\sqrt{\mathcal{I}(\theta)} = \sqrt{\mathcal{I}(\phi)} \left| \frac{d\phi}{d\theta} \right|$

— Finally $p_\theta(\theta) = p_\phi(\phi) \left| \frac{d\phi}{d\theta} \right| \propto \sqrt{\mathcal{I}(\phi)} \left| \frac{d\phi}{d\theta} \right| = \sqrt{\mathcal{I}(\theta)}$

Robust priors To prevent an undue influence on the result, we build priors having heavy tails, which avoids forcing things to be too close to the prior mean.

Mixture of conjugate priors Conjugate priors simplify the computation of robust priors, but are often not robust, and not flexible enough to encode our prior knowledge. However it turns out that a mixture of conjugate priors is also conjugate, and seem to be a good compromise.

Hierarchical Bayes A key requirement for computing the posterior $p(\theta|\mathcal{D})$ is the specification of a prior $p(\theta|\eta)$ where η are the hyper-parameters. A Bayesian approach is to put a prior on our priors. This is an example of a **hierarchical Bayesian Model**.

Empirical Bayes In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model, we need to compute : $p(\eta, \theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$
We can approximate the posterior on the hyper-parameters with a point-estimate, $p(\eta|\mathcal{D} \approx \delta_{\hat{\eta}}(\eta))$ where $\hat{\eta} = \arg\max_\eta p(\eta|\mathcal{D})$. Since η is typically much smaller than θ in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on η. Then the estimate becomes :

$$\hat{\eta} = \arg\max_\eta p(\mathcal{D}|\eta) = \arg\max_\eta \int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta$$

This overall approach is called **Empirical Bayes**
Empirical Bayes violates the principle that the prior should be chosen independently of the data. However, we can just view it as a computationally cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\theta \rightarrow \mathcal{D}$. In fact, we can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes :

Method	**Definition**
Maximum likelihood	$\hat{\theta} = \arg\max_\theta p(\mathcal{D}\|\theta)$
MAP estimation	$\hat{\theta} = \arg\max_\theta p(\mathcal{D}\|\theta)p(\theta\|\eta)$
ML-II (Empirical Bayes)	$\hat{\eta} = \arg\max_\eta \int p(\mathcal{D}\|\theta)p(\theta\|\eta)d\theta = \arg\max_\eta p(\mathcal{D}\|\eta)$
MAP-II	$\hat{\eta} = \arg\max_\eta \int p(\mathcal{D}\|\theta)p(\theta\|\eta)p(\eta)d\theta = \arg\max_\eta p(\mathcal{D}\|\eta)p(\eta)$
Full Bayes	$p(\theta, \eta\|\mathcal{D}) \approx p(\mathcal{D}\|\theta)p(\theta\|\eta)p(\eta)$

Bayesian decision theory We can formalize any given statistical decision problem as a game against nature (as opposed to a game against other strategic players, which is the topic of game theory). In this game, nature picks a state or parameter or label, $y \in \mathcal{Y}$, unknown to us, and then generates an observation, $\boldsymbol{x} \in \mathcal{X}$ which we get to see. We then have to make a decision, that is, we have to choose an action a from some **action space** $\mathcal{A}$.Finally we incur some **loss**, $L(y, a)$, which measures how compatible our action a is with nature's hidden state y.
Our goal is to devise a decision procedure or policy, $\delta : \mathcal{X} \rightarrow \mathcal{A}$ which specifies the optimal action for each possible input which specifies the optimal action for each possible input, meaning the action that minimizes the expected loss :

$$\delta(\boldsymbol{x}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}\left(L(y, a)\right)$$

In the Bayesian vision, the expected value of y given the data we have seen so far, whereas in the frequentist vision the expected value refers to x and y that we expect to see in the future.
In the Bayesian visionj the optimal action having observed $\boldsymbol{x}$ is defined as the action a that minimizes the **posterior expected loss** :

$$\rho(a|\boldsymbol{x}) \triangleq \mathbb{E}_{p(y|x)}\left(L(y, a)\right) = \sum_y L(y, a)p(y|x)$$

Hence the Bayes estimator also called Bayes decision rule is given by :

$$\delta(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \rho(\boldsymbol{a}|\boldsymbol{x})$$

Bayes estimators for common loss functions

— **MAP** estimate minimizes 0-1 loss : $L(y,a) = \mathbb{I}_{y \neq a} \begin{cases} 0 \text{ if } a = y \\ 1 \text{ else} \end{cases}$

— **Reject option**, in classification problems where $p(y|\boldsymbol{x})$ is very uncertain we may prefer to choose a reject action, in which we refuse to classify the example as any of the specified classes. Let choosing $a = C + 1$ correspond to picking the reject action, and choosing $a \in \{1, ..., C\}$ correspond to picking one of the classes.

$$L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, ..., C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where λ_r is the cost of the reject action, and λ_s is the cost of a substitution error.

— **Squared Error (l_2)** for a continuous parameters. $L(y,a) = (y-a)^2$

— **Absolute Error (l_1)** more robust against outliers. $L(y,a) = |y-a|$. The optimal point is the median.

— **Supervised learning** considering a prediction function $\delta : \mathcal{X} \to \mathcal{Y}$ and some cost function $l(y, \delta(x))$. Then the loss incurred by taking action δ when the unknown state of nature is θ (the parameters of the data generating the mechanism).

$$L(\boldsymbol{\theta}, \delta) \triangleq \mathbb{E}_{(\boldsymbol{x},y)\ p(\boldsymbol{x},y|\boldsymbol{\theta})} \left(l(y, \delta(\boldsymbol{x})) \right) = \sum_{\boldsymbol{x}} \sum_{y} L\left(y, \delta(\boldsymbol{x}) p(\boldsymbol{x}, y|\boldsymbol{\theta})\right)$$

— **False positive vs False negative trade-off** for binary decision problems three are 2 types of errors :

1. false positive (false alarm) if $\hat{y} = 1 \wedge y = 0$
2. false negative (missed detection) if $\hat{y} = 0 \wedge y = 1$

We can consider the loss matrix :

Headers	$\boldsymbol{y=1}$	$\boldsymbol{y=0}$
$\boldsymbol{\hat{y}=1}$	0	L_{FP}
$\boldsymbol{\hat{y}=0}$	L_{FN}	0

where L_{FN} is the cost of a false negative and L_{FP} the cost of a false positive.

— **ROC curves** From the below table

Headers		**Truth**		**Count**
Estimate	1	TP	FP	$\hat{N}_+ = TP + FP$
	0	FN	TN	$\hat{N}_- = FN + TN$
Count		$N_+ = TP + FN$	$N_- = FP + TN$	$N = N_+ + N_- = \hat{N}_+ + \hat{N}_-$

we can generate the *confusion matrix* is the below table

Headers	$\boldsymbol{y=1}$	$\boldsymbol{y=0}$
$\boldsymbol{\hat{y}=1}$	$\frac{TP}{N}$ (sensitivity/recall)	$\frac{FP}{N}$ (error type I/ false alarm)
$\boldsymbol{\hat{y}=0}$	$\frac{FN}{N}$ (error type II/ missed detection)	$\frac{TN}{N}$ (specificity)

— **Precision recall curves** When trying to detect a rare event the number of negatives is very large, hence comparing *sensitivity* and *the error of type I* is not very informative. We would then like to use a measure that only talks about positives.

1. *Probability's points of view*

   — **precision** $= \frac{TP}{\hat{N}_+}$

   — **recall** $= \frac{TP}{N_+}$

   A **precision recall curve** is a plot of *precision* vs *recall*.

— **F-scores** is the *harmonic mean of precision and recall* :
$F_1 \triangleq \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$

Hypothesis Testing A *Bayesian* statistics, probabilities represent subjective beliefs. Bayesian hypothesis testing provides rules for calculating how you should update your beliefs about different hypotheses in light of the evidence you see.

Posterior belief We use distributions to represent model parameters, we are uncertain about.
We start out with *prior distribution*, representing our belief before we have seen our data.
We then see some data, and the data will be more consistent with some parameters than others.
The rules of Bayesian inference tell us how to update our beliefs about the parameters now that we've seen the data to obtain posterior beliefs.
Bayesian distributions are easy to interpret. The mode of the distribution is the most likely value of the parameter.

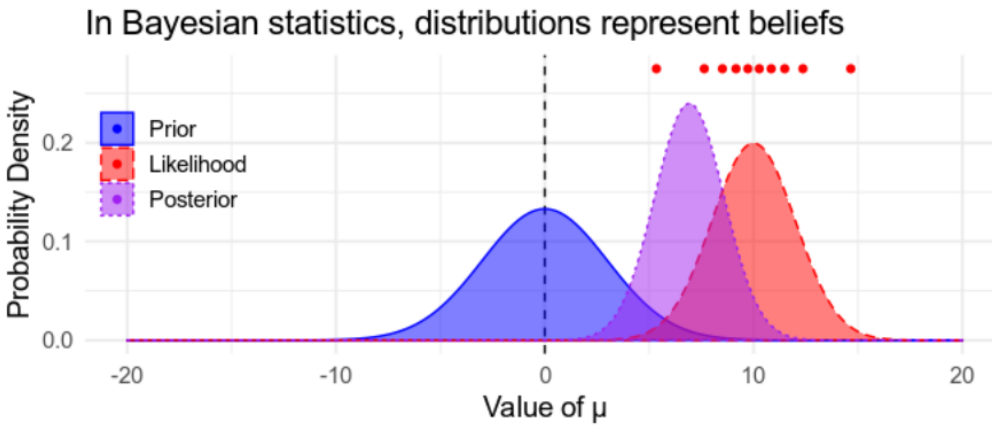

FIGURE 1.2. – Link between prior, likelihood and posterior

Bayesian credible interval If 95% of the distribution is between 2 values $\beta_{low}, \beta_{high}$ then according to the model there's a 95% probability that the parameter is somewhere

between these 2 values.
If we use uniform prior, the posterior distribution only depends on the data, and so we end up with parameters that match the frequentist maximum-likelihood estimates : the posterior mode is the same as the maximum-likelihood value, and the credible intervals match the confidence intervals.

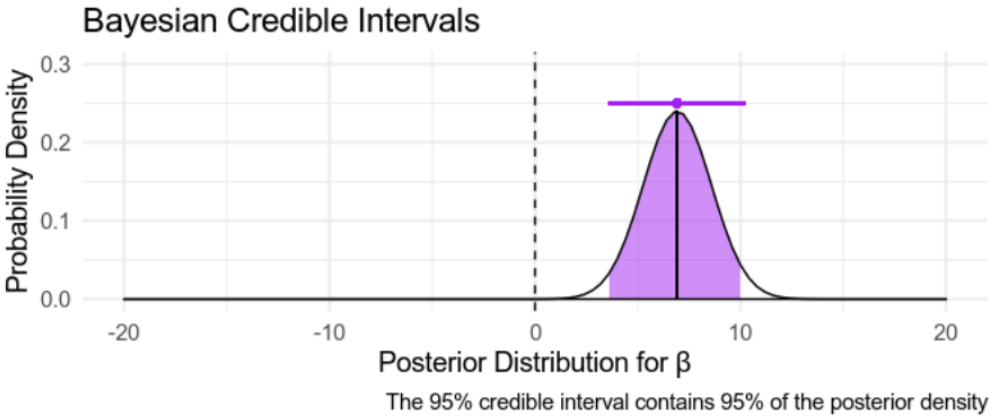

FIGURE 1.3. – Bayesian Credible Intervals

Credible interval tests 95% credible interval are commonly used as a simple way to decide whether an effect is real or not : if the credible interval does not include 0, the effect is genuine. If it includes 0, it might not be.
If we use uniform prior, we can infer on the *p-value*.

Posterior Sign tests Once we've calculated our posterior distribution for β, we can interrogate it directly. For instance if 99% of the posterior distribution is above 0, we are 99% sure that $\beta > 0$ ($\mathbb{P}(\beta > 0) = 0.99$) and 1% sure that $\beta < 0$.
However for example if our posterior distribution is centered on 0 we find that $\mathbb{P}(\beta > 0) = \mathbb{P}(\beta < 0) = 0.5$. The reason for this is that we are considering only 2 hypotheses, while ignoring that $\beta = 0$. To test the hypothesis that $\beta = 0$ we need to calculate the Bayes Factor.
Interestingly, the posterior sign test is closely

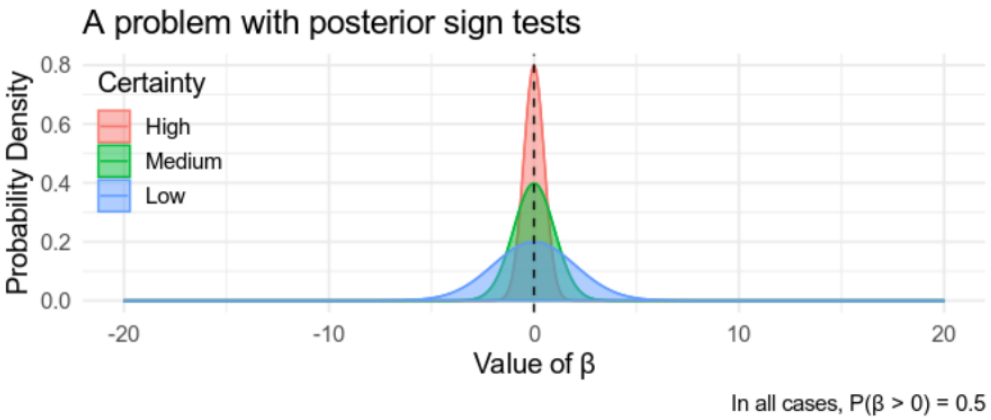

FIGURE 1.4. – Assessing certainty

Bayes Factors

Prerequisites $\mathbb{P}_{H_0}(Data)$ is a average over all possible values of β, weighted by how likely each value in according to the prior. $\begin{cases} \mathbb{P}_{H_1}(Data) = \sum_{i=1}^{n} \mathbb{P}_{\beta_i}(Data)\,\mathbb{P}(\beta_i) & \text{Discrete} \\ \mathbb{P}_{H_1}(Data) = \int \mathbb{P}_{\beta_i}(Data)\,\mathbb{P}(\beta_i) & \text{Continuous} \end{cases}$

Setting an appropriate prior is complicated, usually initial parameters are left.

Definition

$$BF_{10} = \frac{\mathbb{P}_{H_1}(Data)}{\mathbb{P}_{H_0}(Data)}$$
$$BF_{01} = \frac{1}{BF_{10}}$$

Note that in general it is easier to find evidence in favour H_0 if H_1 distribution is broad and easier to find evidence against H_0 if H_1 is narrow.

Bayes Factor	Strength of evidence
$BF = 1$	No evidence
$BF > 1$	Anecdotal evidence
$BF > 3$	Moderate
$BF > 10$	Strong
$BF > 30$	Very Strong
$BF > 100$	Extreme evidence

Interpreting **Aim** : Finding how likely it is that H_1 is true and H_0 false.
Procedure :

1. Compute BF_{10} using maximum-likelihood.
2. Deciding how likely we thought it was the one or other hypothesis was true before seeing any data : *the prior odds* $\mathbb{P}(H_1) H_0$.
   Then there are 2 kinds of priors : prior beliefs about which hypothesis or model is true, and prior beliefs about the values of the parameters in each model.
3. With the 2 above information compute the posterior distribution with Bayes' theorem
   $\mathbb{P}_{Data}(H_1) = \dfrac{\mathbb{P}_{H_1}(Data)\,\mathbb{P}(H_1)}{\mathbb{P}_{H_0}(Data)\,\mathbb{P}(H_0) + \mathbb{P}_{H_1}(Data)\,\mathbb{P}(H_1)}$
4.

For example we get the following table :

$\mathbf{BF_{10}}$	$\mathbf{BF_{01}}$	**Prior** $\mathbb{P}(H_1)$	**Posterior** $\mathbb{P}_{Data}(H_1)$
0.05	20	50%	4.8%
0.1	10	50%	9.1%
0.33	3	50%	25%
1	1	50%	50%
3	0.33	50%	75%
10	0.1	50%	90.9%
20	0.05	50%	95.2%

Multiple comparisons

Density Ratios

Posterior Estimates The more tests we run the more likely it is to we'll find at least one that is significant even though the null hypothesis is true. We can then apply a Bonferroni correction.
Let's say we are running k tests, we can either adjust :

— the threshold $\alpha_{adj} = \dfrac{\alpha}{k}$ OR

— the *p-value* $p_{ajd} = k \times p$

2. Conditional probability and Bayes' theorem

2.1. Conditional Probability

Definition of a conditional probability : Let S be a sample space associated with a random experiment. The conditional probability of an event A, given that event B has occurred, is defined by :

$$\mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

2.2. Bayes' Theorem

Law of Total Probability If the events $(B_i)_{1 \leq i \leq n}$ constitute a partition of the sample space S and $\mathbb{P}(B_i) \neq 0$ then for any event A in S :

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\, \mathbb{P}_{B_i}(A)$$

Bayes'Theorem If the events $(B_i)_{1 \leq i \leq n}$ constitute a partition of the sample space S and $\mathbb{P}(B_i) \neq 0$ then for any event A in S :

$$\mathbb{P}_A(B_k) = \frac{\mathbb{P}(B_k)\, \mathbb{P}_{B_k}(A)}{\sum_{i=1}^{n} \mathbb{P}(B_i)\, \mathbb{P}_{B_i}(A)}$$

3. Distribution Functions

3.1. Distribution Function of Discrete Variables

Definition of probability density function (pdf) : Let R_X be the space of the random variable X. The function : $f : R_X \to \mathbb{R}$ defined by :

$$f(x) = \mathbb{P}(\{X = x\})$$

is called probability density function of X.

Definition of cumulative density function (cdf) : Let R_X be the space of the random variable X. The function : $F : R_X \to \mathbb{R}$ defined by :

$$F(x) = \mathbb{P}(\{X \leq x\})$$

is called cumulative density function of X.
Then :

$$F(x) = \sum_{t \leq x} f(t)$$

3.2. Distribution Function of Continuous Variables

Definition probability density function : A random variable X is said to be a continuous random variable if there exists a continuous $f : \mathbb{R} \to [0; \infty[$ such that for every set of real numbers A :

$$\mathbb{P}(\{X \in A\}) = \int_A f(x)dx$$

Definition cumulative density function : Let $f(x)$ be the probability density function of a continuous random variable X. The cumulative distribution function $F(x)$ of X is defined as :
exists a continuous $f : \mathbb{R} \to [0; \infty[$ such that for every set of real numbers A :

$$F(x) = \mathbb{P}(\{X \leq x\}) = \int_{-\infty}^{x} f(t)dt$$

3.3. Percentile for Continuous Random Variables

Definition Let $p \in [0;1]$, a $100p^{th}$ percentile of the distribution of a random variable X is $q \in \mathbb{R}$ satisfying :

$$\mathbb{P}(\{X \leq q\}) \leq p$$

(Recall that the F is a monotonically increasing function, then it has an inverse F^{-1})

$$q = F^{-1}(p)$$

A $100p^{th}$ is a measure of location for the probability distribution in the sense that q divides the distribution of the probability mass into 2 parts, one having probability mass p and other having probability mass $1 - p$

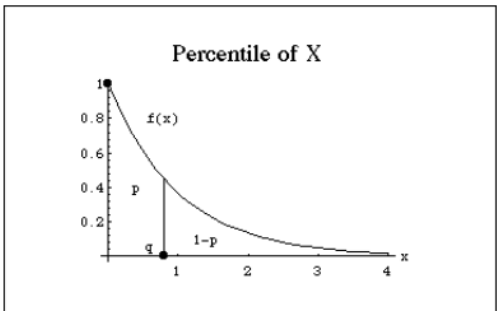

FIGURE 3.1. – Percentile

The 50^{th} percentile of any distribution is called median of the distribution.

Mode definition A mode of the distribution of a continuous random variable X is the value of x where the probability density function $f(x)$ attains a relative maximum. A mode of a random variable X is one of its most probable values.

4. Moments of Random Variables and Chebychev Inequality

4.1. Moments of Random Variables

n^{th} moments of a random variable : The n^{th} moment about the origin of a random variable X as denoted by $E(X^n)$, is defined to be :

$$\mathbb{E}(X^n) = \begin{cases} \sum\limits_{x \in R_X} x^n f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

4.2. Expected Value of Random Variables

Expected value : The expected value of a random variable X as denoted by $E(X)$, is defined to be :

$$\mathbb{E}(X) = \begin{cases} \sum\limits_{x \in R_X} x f(x) \text{ if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x) dx \text{ if } X \text{ is continuous} \end{cases}$$

4.3. Variance of Random Variables

Definition of Variance : Let X be a random variable with mean μ_X. The variance of X denoted by $\mathbb{V}(X)$ or σ_X^2 is defined by :

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mu_X]^2\right)$$

If X is a random variable with mean μ_X and variance σ_X^2 then :

$$\sigma_X^2 = \mathbb{E}\left(X^2\right) - \mu_X^2$$

And :

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

4.4. Chebychev Inequality

Theorem Chebychev inequality allows to find an estimate of the area between the values $\mu - k\sigma$ and $\mu + k\sigma$ for some given $k \neq 0$, showing that the area under $f(x)$ on the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - k^2$.
Let X be a random variable with probability density function $f(x)$. If μ and $\sigma > 0$ are the mean and standard deviation of X then :

$$\boxed{\mathbb{P}\left(\{|X - \mu| < k\sigma\}\right) \geq 1 - \frac{1}{k^2}}$$

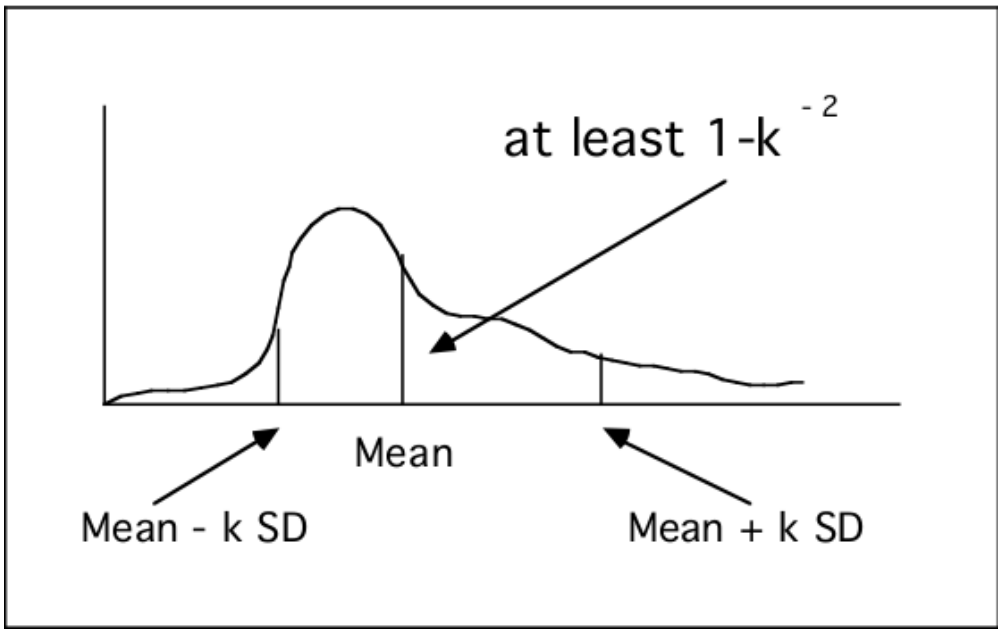

FIGURE 4.1. – Illustration of Chebychev inequality

4.5. Moment Generating Funcions

It is sometimes difficult to compute moments from the definition, a moment generating function is a real valued function from which one can generate all the moment of a given random variable.

Definition Let X be a random variable whose probability density function is $f(x)$. A real valued function $M : \mathbb{R} \to \mathbb{R}$ defined by :

$$\boxed{M(t) = \mathbb{E}\left(e^{tX}\right)}$$

if this expected value exists for all t in the interval $-h < t < h$ for some $h > 0$

Factorial moment generating function FMGF of X is denoted by $G(t)$ and defined as :

$$G(t) = E(t^X)$$

$G(t) = M(\ln(t))$

5. Important Discrete Distributions

5.1. Bernoulli Distribution

Definition It can be thought of as a model for the set of possible outcomes of any single exp X is the number of success during a Bernoulli process.

$$\forall x \in 0,1, f(x) = p^x(1-p)^{1-x} \Rightarrow X \hookrightarrow BER(\mathrm{p})$$

Theorem

$$\begin{cases} \mu_X = p \\ \sigma_X^2 = p(1-p) \\ M_X(t) = (1-p) + pe^t \end{cases}$$

5.2. Binomial Distribution

Definition

$$\forall x \in [\![0,n]\!], f(x) = \binom{n}{p} p^x(1-p)^{1-x} \Rightarrow X \hookrightarrow \mathcal{B}(n,p)$$

Theorem

$$\begin{cases} \mu_X = np \\ \sigma_X^2 = np(1-p) \\ M_X(t) = [(1-p) + pe^t]^n \end{cases}$$

5.3. Geometric Distribution

Definition

$$\forall x \in \mathbb{N}, f(x) = (1-p)^{x-1}p \Rightarrow X \hookrightarrow \mathcal{G}(n,p)$$

p denotes the probability of success in a single Bernoulli trial

Theorem

$$\begin{cases} \mu_X = \frac{1}{p} \\ \sigma_X^2 = \frac{1-p}{p^2} \\ M_X(t) = \frac{pe^t}{1-(1-p)e^t} \text{ if } t < -\ln(1-p) \end{cases}$$

Memoryless property

$$X \hookrightarrow \mathcal{G}(p) \Leftrightarrow \forall (m,n) \in \mathbb{N}^2 \mathbb{P}_{\{X>m\}} \left(\{X > m+n\}\right) = \mathbb{P}\left(\{X > m\}\right)$$

The difference between the binomial and geometric distributions is in the first the number of trials was predetermined, whereas in the last it is the random variable.

5.4. Negative Binomial Distribution

X denotes the number of trials needed to observe the r^{th} successes $\mathbb{P}\left(\{X = x\}\right) = \mathbb{P}\left(\{\text{first } x-1 \text{ trials contain : } r-1 \text{ failures and } r-1 \text{ successes}\}\right) \times \mathbb{P}\left(\left\{r^{th} \text{ success in } x^{th} \text{ trial}\right\}\right)$ where p is the probability of success in a single Bernoulli trial.

Definition

$$\forall x \in [r, \infty[\,, f(x) = \tbinom{x-1}{r-1} p^r (1-p)^{x-r} p \Rightarrow X \hookrightarrow \mathcal{NEG}(r,n)$$

p denotes the probability of success in a single Bernoulli trial

Theorem

$$\begin{cases} \mu_X = \frac{r}{p} \\ \sigma_X^2 = \frac{r(1-p)}{p^2} \\ M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right) \text{ if } t < -\ln(1-p) \end{cases}$$

5.5. Hypergeometric Distribution

Definition Suppose that there are n_1 objects in class 1 and n_2 objects in class 2. A collection of r objects is selected from these n objects at random and without replacement. We are interested in finding out the probability that exactly x of these r objects are from class 1.

$$\forall x \in [\![0, r]\!]\,, f(x) = \frac{\binom{n_1}{x}\binom{n_2}{r-x}}{\binom{n_1+n_2}{r}} \Rightarrow X \hookrightarrow \mathcal{HP}(n_1, n_2, r)$$

p denotes the probability of success in a single Bernoulli trial

Theorem

$$\begin{cases} \mu_X = r\frac{n_1}{n_1+n_2} \\ \sigma_X^2 = r\left(\frac{n_1}{n_1+n_2}\right)\left(\frac{n_2}{n_1+n_2}\right)\left(\frac{n_1+n_2-r}{n_1+n_2-1}\right) \end{cases}$$

5.6. Poisson Distribution

Use case To express the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate.

Definition

$$\forall x \in \mathbb{N}, f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \Rightarrow X \hookrightarrow \mathcal{P}(\lambda)$$

$\lambda \in \mathbb{R}_+^*$

Assumptions

— k is the number of times an event occurs in an interval and $k \in \mathbb{N}$

— Events are mutually independent.

— The average rate is constant.

— 2 events cannot occur at exactly the same instant.

Theorem

$$\begin{cases} \mu_X = \lambda \\ \sigma_X^2 = \lambda \\ M(t) = e^{\lambda(e^t - 1)} \end{cases}$$

5.7. Riemann Zeta Distribution

Definition Initially introduced by Vilfredo Pareto to study the distribution of family incomes of a given country.

$$\forall x \in \mathbb{N}, f(x) = \frac{1}{\zeta(\alpha+1)} x^{-\alpha+1} \Rightarrow X \hookrightarrow \mathcal{RZ}(\lambda)$$

$$\zeta(s) = \sum_{k=1}^{\infty} \left(\frac{1}{k}\right)^s \ \& \ \alpha > 0$$

Theorem

$$\begin{cases} \mu_X = \lambda \\ \sigma_X^2 = \lambda \\ M(t) = e^{\lambda(e^t - 1)} \end{cases}$$

5.8. Pareto Distribution

Definition Is used to model the distribution of quantities that exhibit long tails, also called heavy tails.

$$\forall x \in \mathbb{N}, f(x|k,m) = km^k x^{-(k+1)} \mathbb{1}(x \geq m) \Rightarrow X \hookrightarrow Pareto(m,k)$$

This density asserts that x must be greater than some constant m, but not too much greater, where k controls what is "too much".

Theorem

$$\begin{cases} \mu_X = \dfrac{km}{k-1} \\ \sigma_X^2 = \dfrac{m^2 k}{(k-1)^2(k-2)} \text{ if } k > 2 \\ mode = m \end{cases}$$

6. Important Continuous Distributions

6.1. Uniform Distribution

Definition

$$\forall x \in [a,b], f(x) = \tfrac{1}{b-a} \Rightarrow X \hookrightarrow \mathcal{U}(a,b)$$

$$\zeta(s) = \sum_{k=1}^{\infty} \left(\tfrac{1}{k}\right)^s \ \& \ \alpha > 0$$

Theorem

$$\begin{cases} \mu_X = \frac{b+a}{2} \\ \sigma_X^2 = \frac{(b+a)^2}{12} \\ M(t) = \begin{cases} 1 \Leftarrow t = 0 \\ \frac{e^{tb}-e^{ta}}{t(b-a)} \Leftarrow t \neq 0 \end{cases} \end{cases}$$

Theorem X is a continuous random variable with a strictly increasing cumulative distribution function $F \Rightarrow Y = F(x) \hookrightarrow \mathbb{U}(0,1)$

6.2. Gamma Distribution

Gamma distribution The *gamma* function $\Gamma(z)$ is a generalization of the notion of factorial, for $z \in \mathbb{R}_+^*$:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

Definition

$$f(x) \begin{cases} \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}} \Rightarrow x \in \mathbb{R}_+^* \\ 0 \Leftarrow \overline{x \in \mathbb{R}_+^*} \end{cases} \quad \Rightarrow X \hookrightarrow \mathcal{G}(\theta, \alpha)$$

Theorem $X \hookrightarrow \theta, \alpha \Rightarrow \begin{cases} \mathbb{E}(X) = \theta\alpha \\ \mathbb{V}(X) = \theta^2\alpha \\ M(t) = \left(\frac{1}{1-\theta t}^\alpha\right) \text{ if } t < \frac{1}{\theta} \end{cases}$

Exponential distribution

Definition

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} \Leftarrow x > 0 \\ 0 \Leftarrow x \leq 0 \end{cases}$$

Most of the information about an exponential distribution can be obtained from the gamma distribution

Chi-square distribution with r degrees of freedom

Definition

$$f(x) = \begin{cases} \frac{1}{\Gamma\frac{r}{2} 2^{\frac{r}{2}}} e^{-\frac{x}{2}} \Leftarrow x \in \mathbb{R}_+^* \\ 0 \Leftarrow x \leq 0 \end{cases} \quad \Rightarrow X \hookrightarrow \chi_2(r)$$

n-Erlang

Definition

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!} \Leftarrow x \in \mathbb{R}_+^* \\ 0 \Leftarrow x \leq 0 \end{cases} \quad \Rightarrow X \hookrightarrow n\text{-}Erlang(\lambda)$$

with $\lambda > 0$

Unified distribution

Definition

$$f(x) = \begin{cases} \dfrac{\alpha}{\theta^{\alpha^\psi} \Gamma\left(\alpha^\psi + 1\right)} x^{\alpha - 1} e^{\frac{-x^{-\left(\alpha^\psi - \alpha - 1\right)}}{\theta}} \Leftarrow x \in \mathbb{R}_+^* \\ 0 \Leftarrow x \leq 0 \end{cases} \quad \Rightarrow X \hookrightarrow n\text{-}Erlang(\lambda)$$

with $\lambda > 0$

Weibull distribution For $\alpha = 1$ we get Weibull distribution

$$\begin{cases} \mathbb{E}(X) = \theta^{\frac{1}{\alpha}} \Gamma\left(1 + \frac{1}{\alpha}\right) \\ \mathbb{V}(X) = \theta^{\frac{2}{\alpha}} \left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \left(1 + \frac{1}{\alpha}\right)^2\right) \end{cases}$$

The Weibull distribution provides probabilistic models for life-length data of components or systems.

6.3. Beta Distribution

It is a versatile distribution and as such it is used in modeling the behavior of random variables that are positive but bounded in possible values.

Beta

Theorem Let $(\alpha, \beta) \in \mathbb{R}_+^2$ then,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$
$$\text{where } \Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx \text{ is the gamma function}$$

Beta distribution

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \text{ if } 0 < x < 1 \\ 0 \text{ otherwise} \end{cases} \qquad X \hookrightarrow BETA(\alpha, \beta)$$

Beta properties

$$\begin{cases} \mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \\ \mathbb{V}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{cases}$$

Generalized Beta

Theorem

$$f(x) = \begin{cases} f(x) = \frac{1}{B(\alpha,\beta)} \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}} \text{ if } a < x < b \\ 0 \text{ otherwise} \end{cases} \qquad \Rightarrow X \hookrightarrow GBETA(\alpha, \beta, a, b)$$

Proprieties

$$\begin{cases} \mathbb{E}(X) = (b-a)\frac{\alpha}{\alpha+\beta} + a \\ \mathbb{V}(X) = (b-a)^2\frac{\alpha\beta}{\alpha+\beta+1} \end{cases}$$

6.4. Normal Distribution

Normal Distribution

Definition

$$\forall x \in \mathcal{R} f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)} \Rightarrow X \hookrightarrow \mathcal{N}(\mu, \sigma)$$

Properties

$$\begin{cases} \mathbb{E}(X) = \mu \\ \mathbb{V}(X) = \sigma^2 \\ M(t) = e^{-\mu t + \frac{1}{2}\sigma^2 t^2} \end{cases}$$

Chi Squared

Definition

$$X \hookrightarrow \mathcal{N}(\mu, \sigma^2) \Rightarrow \left(\frac{X-\mu}{\sigma}\right)^2 \hookrightarrow \chi^2(1)$$

Generalization of normal distribution

$$g(x) = \frac{\nu\varphi(\nu)}{2\sigma\Gamma(\frac{1}{\nu})} e^{-\left(\frac{\varphi(\nu)}{\sigma}|x-\mu|\right)^{\nu}} \text{ where } \varphi(\nu) = \sqrt{\frac{\Gamma\left(\frac{3}{\nu}\right)}{\Gamma(\frac{1}{\nu})}}$$

6.5. Lognormal Distribution

This distribution can be defined as the distribution of a random variable whose logarithm is normally distributed.

Definition

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2} \text{ if } x \in \mathbb{R}_+^* \\ 0 \text{ otherwise} \end{cases} \quad \Rightarrow X \hookrightarrow \wedge(\mu, \sigma)$$

Properties

$$\begin{cases} \mathbb{E}(X) = e^{\mu + \frac{1}{2}\sigma^2} \\ \mathbb{V}(X) = \left(e^{\sigma^2} - 1\right) e^{2\mu + \sigma^2} \end{cases}$$

6.6. Inverse Gaussian Distribution

The interpurchase times of toothpaste of a family, the duration of labor strikes in a geographical region, word frequency in a language, conversion time for convertible bonds, length of employee service, and crop field size follow inverse Gaussian distribution.

Definition

$$f(x) = \begin{cases} \sqrt{\frac{\lambda}{2\pi}} x^{-\frac{3}{2}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}} \text{ if } x \in \mathbb{R}_+^* \\ 0 \text{ otherwise} \end{cases} \quad \Rightarrow X \hookrightarrow IG(\mu, \lambda)$$

Properties

$$\begin{cases} \mathbb{E}(X) = \mu \\ \mathbb{V}(X) = \dfrac{\mu^3}{\lambda} \end{cases}$$

6.7. Logistic Distribution

The logistic distribution is used in modeling demographic data. It is also used as an alternative to the Weibull distribution in life-testing.

Definition

$$f(x) = \frac{\pi}{\sigma\sqrt{3}} \frac{e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}}{\left(1 + e^{-\frac{\pi}{\sqrt{3}}\left(\frac{x-\mu}{\sigma}\right)}\right)^2} \Rightarrow X \hookrightarrow LOG(\mu, \sigma)$$

Properties

$$\begin{cases} \mathbb{E}(X) = \mu \\ \mathbb{V}(X) = \sigma^2 \\ M(t) = e^{\mu t}\Gamma\left(1 + \dfrac{\sqrt{3}}{\pi}\sigma t\right)\Gamma\left(1 - \dfrac{\sqrt{3}}{\pi}\sigma t\right), |t| < \dfrac{\pi}{\sigma\sqrt{3}} \end{cases}$$

7. 2 Random Variables

7.1. Bivariate Discrete Random Variables

Joint probability density function Let $(X, Y) : (\Omega_X, \Omega_Y) \to (R_X, R_Y)$ and $f : R_X \times R_Y \to \mathbb{R}$

$$\forall (x, y) \in R_X \times R_Y, f(x, y) = \mathbb{P}(\{X = x, Y = y\}) \Leftrightarrow$$
f is the joint probability density function for X and Y

Marginal probability density function Let for all $(x, y) \in R_X \times R_Y : f(x, y)$ be the joint probability density of X and Y

$$\begin{cases} f_1(x) = \sum_{y \in R_y} f(x, y) \text{is the marginal probability density of } X \\ f_2(y) = \sum_{x \in R_x} f(x, y) \text{is the marginal probability density of } Y \end{cases}$$

Joint cumulative probability distribution function Let $F : \mathbb{R}^2 \to \mathbb{R}$

$$\forall (x, y) \in \mathbb{R}^2, F(x, y) = \mathbb{P}(\{X \leq x, Y \leq y\}) \Leftrightarrow$$
F is the joint cumulative probability density function for X and Y

7.2. Bivariate Continuous Random Variables

Marginal probability density function Let for all $(x, y) \in R_X \times R_Y : f(x, y)$ be the joint probability density of X and Y

$$\begin{cases} f_1(x) = \int_{-\infty}^{\infty} f(x, y)dy \text{is the marginal probability density of } X \\ f_2(y) = \int_{-\infty}^{\infty} f(x, y)dx \text{is the marginal probability density of } Y \end{cases}$$

Joint cumulative probability distribution function Let $F : \mathbb{R}^2 \to \mathbb{R}$

$$\forall (x, y) \in \mathbb{R}^2, F(x, y) = \mathbb{P}(\{X \leq x, Y \leq y\}) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)dudv \Leftrightarrow$$
F is the joint cumulative probability density function for X and Y

From the fundamental theorem of calculus : $f(x, y) = \dfrac{\partial^2 F(x, y)}{\partial x \partial y}$

7.3. Conditional Distributions

Keeping previous notations :

Conditional probability density function The conditional probability density function g of X given the event $\{Y = y\}$ is defined as

$$g_{\{Y=y\}}(x) = \frac{f(x,y)}{f_2(y)}$$

7.4. Independence of Random Variables

The random variables X and Y are (stochastically) independent if and only if

$$\forall (x,y) \in R_X \times R_Y, f(x,y) = f_1(x) f_2(y)$$

Conditionally independent Usually, in a set of variables, the unconditional independence is rare, because most variables can influence most variables, however the influence is mediated via other variables rather than being direct. We therefore say X and Y are conditionnally independent given Z if

$$\mathbb{P}_Z(X) = \mathbb{P}_Z(X)\,\mathbb{P}_Z(Y)$$
or
$$f_{x,y}(x,y|z) = f_x(x|z) \times f_y(y|z)$$

8. Product Moments of Bivariate Random Variabless

8.1. Covariance of Bivariate Random Variables

Product Moment The product moment of X and Y, denoted $\mathbb{E}(XY)$ is defined as :

$$\begin{cases} \sum_{x \in R_X} \sum_{y \in R_Y} xyf(x,y) \text{ if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x,y)dxdy \text{ if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Covariance

Definition Let X and Y 2 random variables with joint density function f :

$$Cov(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Properties

$$Cov(aX+b, cY+d) = acCov(X,Y)$$

8.2. Independence of Random Variables

Properties

$$\begin{cases} X\&Y : independent \Rightarrow \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \\ X\&Y : independent \Rightarrow Cov(X,Y) = 0 \end{cases}$$

8.3. Variance of Linear Combination Random Variables

Properties

$$\begin{cases} X,Y : 2RV \Rightarrow \mathbb{V}(aX+bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2abCov(X,Y) \\ X,Y : independent \Rightarrow \mathbb{V}(XY) = \mathbb{V}(X)\mathbb{V}(Y) \end{cases}$$

8.4. Correlation and Independence

Correlation coefficient To get correlation coefficient : convert each variable to standard units and compute the average of their product.

$$X, Y : 2 \text{ RV} \Rightarrow \rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

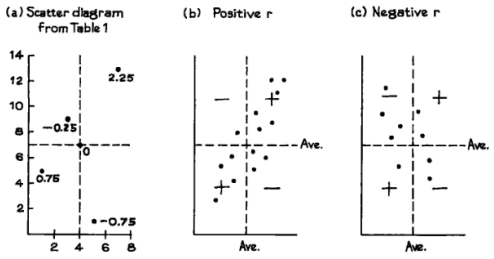

FIGURE 8.1. – + : area where r is positive.
- : area where r is negative

Associated with each of one SD in x there is an increase of only r SDs in y, on the average.

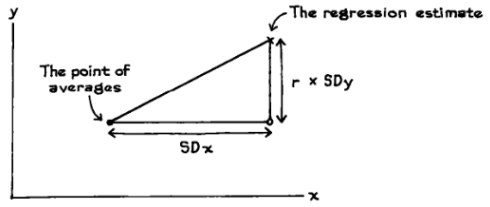

FIGURE 8.2. – Graphical representation of r

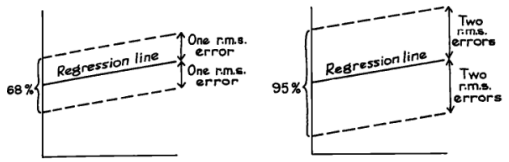

FIGURE 8.3. – Graphical interpretation of Residual Mean Squares

Property

$$X, Y : 2 \text{ RV} \Rightarrow -1 \leq \rho \leq 1$$

8.5. Moment Generating Function

One can define the moment generating function for the bivariate case to compute the various product moments

Moment generating function Let X and Y be 2 random variables with joint density function $f(x, y)$. A real valued function $M : \mathbb{R}^2 \to \mathbb{R}$

$$\forall (t, s) \in [-k, k] \times [-h, h] M(s, t) = \mathbb{E}\left(e^{sX+tY}\right) exists \Rightarrow$$
M is the joint moment generating function of X and Y

Property

$$X, Y : \text{independent} \Rightarrow M_{aX+bY}(t) = M_X(at) M_Y(bt)$$

9. Conditional Expectation of Bivariate Random Variables

9.1. Conditional Expected Values

Conditional expectation The conditional mean of X given $Y = y$ is defined as :

$$\mathbb{E}(X|y) = \begin{cases} \sum_{x \in R_X} xg(x/y) \Leftarrow X \text{ discrete} \\ \int_{-\infty}^{\infty} xg(x/y)dx \Leftarrow X \text{ continuous} \end{cases}$$

Properties

$$\begin{cases} \mathbb{E}_X\left(\mathbb{E}_{\{y|x\}}(Y|X)\right) = \mathbb{E}_{\{y\}}(Y) \\ \mathbb{E}(Y|\{X = x\}) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) \end{cases}$$

9.2. Conditional Variance

Definition

$$\begin{cases} \mathbb{V}(Y|x) = \mathbb{E}\left(Y^2|x\right) - \mathbb{E}(Y|x)^2 \\ \mathbb{E}_x\left(\mathbb{V}(Y|X) = (1 - \rho^2)\mathbb{V}(Y)\right) \end{cases}$$

9.3. Regression Curve and Scedastic Curves

Regression function of Y on X

$$\mathbb{E}(Y|\{X = x\}) = \int_{-\infty}^{\infty} yh(y/x)dx$$

The graph of this regression function of Y on X is known as the regression curve of Y on X.

Linear regression Let X and Y be two random variables with joint probability density function $f(x, y)$ and let $\mathbb{E}(Y|\{X = x\})$ be the regression function of Y on X. If this regression function is linear, then $\mathbb{E}(Y|\{X = x\})$ is called a linear regression of Y on X. Otherwise, it is called nonlinear regression of Y on X.

Scedastic function Let $h(y/x)$ is the conditional density of Y given $\{X = x\}$

$$\mathbb{V}\left(Y|\left\{X = x\right\}\right) = \int_{-\infty}^{\infty} y^2 h(y/x)dy$$

The graph of this scedastic function of Y on X is known as the scedastic curve of Y on X

10. Functions of Random Variables and Their Distribution

10.1. Transformation Method for Univariate Case

The most useful method to find the probability density function of a transformed random random variable.

S'exercer sur pdf p276

Essential theorem for transforming Let X be a continuous random variable with probability density function $f(x)$. Let $y = T(x)$ be an increasing (or decreasing) function. Then the density function of the random variable $Y = T(X)$ is given by

$$g(y) = \left|\frac{dx}{dy}\right| f\left(W(y)\right)$$

where W is the inverse function of T

10.2. Transformation Method Bivariate Case

Essential theorem to method of bivariate case Let X and Y be 2 continuous random variables with joint density $f(x, y)$. Let $U = P(X, Y)$ and $V = Q(X, Y)$ be functions of X and Y . If the functions $P(x, y)$ and $Q(x, y)$ have single valued inverses, say $X = R(U, V)$ and $Y = S(U, V)$, then the joint density $g(u, v)$ of U and V is given by

$$g(u, v) = |J| f\left(R(u, v), S(u, v)\right)$$

where $J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}$

Density functions Let the joint density of the random variables X and Y be Y $f(x, y)$. Then probability density functions of $X + Y$, XY , and $\frac{X}{Y}$ are given by

$$\begin{cases} h_{X+Y}(v) = \int_{-\infty}^{\infty} f(u, v-u)du \\ h_{XY}(v) = \int_{-\infty}^{\infty} \frac{1}{|u|} f(u, \frac{v}{u})du \\ h_{X+Y}(v) = \int_{-\infty}^{\infty} |u| f(u, vu)du \end{cases}$$

10.3. Convolution Method for Sums of Random Variables

Definition of convolution Let f and g be 2 real valued function. The convolution of f and $g$is defined as :

$$\begin{aligned}(f * g)(z) =& \int_{-\infty}^{\infty} f(z-y)g(y)dy \\ =& \int_{-\infty}^{\infty} g(z-x)f(x)dx\end{aligned}$$

Thus thanks to the previous theorem, if X and Y are 2 independent RV then $f * g$ is the density of the random variable $Z = X + Y$

10.4. Moment Method for Sums of Random Variables

This method is based on the fact that :

$$M_{X+Y}(t) = M_X(t) \times M_Y(t)$$

11. Some Special Discrete Bivariate Distributions

11.1. Bivariate Bernoulli Distribution

Definition A discrete bivariate random variable (X, Y) is said to have the bivariate Bernoulli distribution if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{1}{x!y!(1-x-y)} p_1^x p_2^y (1-p_1-p_2)^{1-x-y} \Leftarrow (x,y) \in \{0,1\}^2 \\ 0 \Leftarrow (x,y) \notin \{0,1\}^2 \end{cases}$$
$$\Rightarrow (X,Y) \hookrightarrow BER(p_1, p_2)$$

where $0 < p_1, p_2, p_1 + p_2 < 1$ and $x + y \leq 1$

Properties

$$\begin{cases} \mathbb{E}(X) = p_1 \\ \mathbb{E}(Y) = p_2 \\ \mathbb{V}(X) = p_1(1-p_1) \\ \mathbb{V}(Y) = p_2(1-p_2) \\ CovX, Y = -p_1 p_2 \\ M(s,t) = 1 - p_1 - p_2 + p_1 e^s + p_2 e^t \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{p_2(1-x)}{1-p_1} \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{p_1(1-y)}{1-p_2} \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{p_2(1-p_1-p_2)(1-x)}{(1-p_1)^2} \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{p_1(1-p_1-p_2)(1-y)}{(1-p_2)^2} \end{cases}$$

11.2. Bivariate Binomial Distribution

Definition A discrete bivariate random variable (X, Y) is said to have the bivariate binomial distribution with parameters n, p_1, p_2 if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{n!}{x!y!(n-x-y)} p_1^x p_2^y (1-p_1-p_2)^{n-x-y} \Leftarrow (x,y) \in \mathbb{N}^2 \\ 0 \Leftarrow (x,y) \notin \mathbb{N}^2 \end{cases}$$
$$\Rightarrow (X,Y) \hookrightarrow BIN(n,p_1,p_2)$$

where $0 < p_1, p_2, p_1 + p_2 < 1, x + y \leq n$ and $n \geq 0$

Properties

$$\begin{cases} \mathbb{E}(X) = np_1 \\ \mathbb{E}(Y) = np_2 \\ \mathbb{V}(X) = np_1(1-p_1) \\ \mathbb{V}(Y) = np_2(1-p_2) \\ CovX,Y = -np_1p_2 \\ M(s,t) = (1-p_1-p_2+p_1e^s+p_2e^t)^n \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{p_2(n-x)}{1-p_1} \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{p_1(n-y)}{1-p_2} \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{p_2(1-p_1-p_2)(n-x)}{(1-p_1)^2} \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{p_1(1-p_1-p_2)(n-y)}{(1-p_2)^2} \end{cases}$$

11.3. Bivariate Geometric Distribution

Definition A discrete bivariate random variable (X,Y) is said to have the bivariate Geometric distribution if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{(x+y)!}{x!y!} p_1^x p_2^y (1-p_1-p_2) \Leftarrow (x,y) \in \mathbb{N}^2 \\ 0 \Leftarrow (x,y) \notin \mathbb{N}^2 \end{cases} \quad \Rightarrow (X,Y) \hookrightarrow GEO(p_1,p_2)$$

where $0 < p_1, p_2, p_1 + p_2 < 1$

Properties

$$\begin{cases} \mathbb{E}(X) = \dfrac{p_1}{1-p_1-p_2} \\ \mathbb{E}(Y) = \dfrac{p_2}{1-p_1-p_2} \\ \mathbb{V}(X) = \dfrac{p_1(1-p_1)}{1-p_1-p_2} \\ \mathbb{V}(Y) = \dfrac{p_2(1-p_2)}{1-p_1-p_2} \\ CovX,Y = \dfrac{p_1p_2}{1-p_1-p_2} \\ M(s,t) = \dfrac{1-p_1-p_2}{1-p_1e^s-p_2e^t} \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{p_2(1+x)}{1-p_1} \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{p_1(1+y)}{1-p_2} \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{p_2(1+p_1-p_2)(1-x)}{(1-p_1)^2} \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{p_1(1+p_1-p_2)(1-y)}{(1-p_2)^2} \end{cases}$$

11.4. Bivariate Negative Binomial Distribution

Definition A discrete bivariate random variable (X,Y) is said to have the bivariate negative Binomial distribution with parameters k, p_1, p_2 if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{(x+y+k-1)!}{x!y!(k-1)!} p_1^x p_2^y (1-p_1-p_2)^k \Leftarrow (x,y) \in \mathbb{N}^2 \\ 0 \Leftarrow (x,y) \notin \mathbb{N}^2 \end{cases}$$
$$\Rightarrow (X,Y) \hookrightarrow NBIN(p_1,p_2)$$

where $0 < p_1, p_2, p_1 + p_2 < 1$

Properties

$$\begin{cases} \mathbb{E}(X) = \dfrac{kp_1}{1-p_1-p_2} \\ \mathbb{E}(Y) = \dfrac{kp_2}{1-p_1-p_2} \\ \mathbb{V}(X) = \dfrac{kp_1(1-p_1)}{1-p_1-p_2} \\ \mathbb{V}(Y) = \dfrac{kp_2(1-p_2)}{1-p_1-p_2} \\ CovX,Y = \dfrac{kp_1p_2}{1-p_1-p_2} \\ M(s,t) = \dfrac{(1-p_1-p_2)^k}{(1-p_1e^s-p_2e^t)^k} \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{p_2(k+x)}{1-p_1} \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{p_1(k+y)}{1-p_2} \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{p_2(k+p_1-p_2)(1-x)}{(1-p_1)^2} \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{p_1(k+p_1-p_2)(1-y)}{(1-p_2)^2} \end{cases}$$

11.5. Bivariate Hypergeometric Distribution

Definition A discrete bivariate random variable (X,Y) is said to have the bivariate hypergeometric distribution with parameters r, n_1, n_2, n_3 if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{\binom{n_1}{x}\binom{n_2}{y}\binom{n_3}{r-x-y}}{\binom{n_1+n_2+n_3}{r}} \Leftarrow (x,y) \in [\![0,r]\!]^2 \\ 0 \Leftarrow (x,y) \notin [\![0,r]\!]^2 \end{cases} \quad \Rightarrow (X,Y) \hookrightarrow HYP(p_1,p_2)$$

where $0 < p_1, p_2, p_1 + p_2 < 1$

Properties

$$\begin{cases} \mathbb{E}(X) = \dfrac{rn_1}{n_1+n_2+n_3} \\ \mathbb{E}(Y) = \dfrac{rn_2}{n_1+n_2+n_3} \\ \mathbb{V}(X) = \dfrac{rn_1(n_2+n_3)}{(n_1+n_2+n_3)^2}\left(\dfrac{n_1+n_2+n_3-r}{n_1+n_2+n_3-1}\right) \\ \mathbb{V}(Y) = \dfrac{rn_2(n_1+n_3)}{(n_1+n_2+n_3)^2}\left(\dfrac{n_1+n_2+n_3-r}{n_1+n_2+n_3-1}\right) \\ Cov(X,Y) = -\dfrac{rn_1n_2}{(n_1+n_2+n_3)^2}\left(\dfrac{n_1+n_2+n_3-r}{n_1+n_2+n_3-1}\right) \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{n_2(r-x)}{n_2+n_3} \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{n_1(r-y)}{n_1+n_3} \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{n_2 n_3}{n_2+n_3-1}\left(\dfrac{n_1+n_2+n_3-x}{n_2+n_3}\right)\left(\dfrac{x-n_1}{n_2+n_3}\right) \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{n_1 n_3}{n_1+n_3-1}\left(\dfrac{n_1+n_2+n_3-y}{n_1+n_3}\right)\left(\dfrac{y-n_1}{n_1+n_3}\right) \end{cases}$$

11.6. Bivariate Poisson Distribution

Unlike the previous bivariate distributions, the conditional distributions of bivariate Poisson distribution are not Poisson.

Definition A discrete bivariate random variable (X, Y) is said to have the bivariate Poisson distribution with parameters $\lambda 1, \lambda_2, \lambda_3$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{e^{-\lambda_1-\lambda_2+\lambda_3(\lambda_1-\lambda_3)^x(\lambda_2-\lambda_3)^y}}{x!y!}\psi(x,y) \Leftarrow (x,y) \in \mathbb{N}^2 \\ 0 \Leftarrow (x,y) \notin \mathbb{N}^2 \end{cases}$$
$$\Rightarrow (X,Y) \hookrightarrow POI(p_1,p_2)$$

where $\psi(x,y) := \sum_{r=0}^{min(x,y)} \dfrac{x^{(r)}y^{(r)}\lambda_3^r}{(\lambda_1-\lambda_3)^r(\lambda_2-\lambda_3)^r r!}$ and $0 < p_1, p_2, p_1+p_2 < 1$ and $x+y \leq 1$

Properties

$$\begin{cases} \mathbb{E}(X) = \lambda_1 \\ \mathbb{E}(Y) = \lambda_2 \\ \mathbb{V}(X) = \lambda_1 \\ \mathbb{V}(Y) = \lambda_2 \\ CovX, Y = \lambda_3 \\ M(s,t) = e^{-\lambda_1-\lambda_2-\lambda_3+\lambda_1 e^s+\lambda_2 e^t+\lambda_3 e^{s+t}} \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \lambda_2 - \lambda_3 + \dfrac{\lambda_3}{\lambda_1}x \\ \mathbb{E}_{\{Y=y\}}(X) = \lambda_1 - \lambda_3 + \dfrac{\lambda_3}{\lambda_2}y \\ \mathbb{V}_{\{X=x\}}(Y) = \lambda_2 - \lambda_3 + \dfrac{\lambda_3(\lambda_1 - \lambda_3)}{\lambda_1^2}x \\ \mathbb{V}_{\{Y=y\}}(X) = \lambda_1 - \lambda_3 + \dfrac{\lambda_3(\lambda_2 - \lambda_3)}{\lambda_2^2}y \end{cases}$$

12. Some Special Continuous Bivariate Distributions

12.1. Bivariate Uniform Distribution

Definition A continuous bivariate random variable (X, Y) is said to have the bivariate uniform distribution with parameters on the rectangle $[a, b]\times[c, d]$ if its joint probability density is of the form

$$f(x,y) = \begin{cases} \dfrac{1+\alpha\left(\frac{2x-2a}{b-a}-1\right)\left(\frac{2y-2c}{d-c}-1\right)}{(b-a)(d-c))} \Leftarrow (x,y) \in [a,b]\times[c,d] \\ 0 \Leftarrow (x,y) \notin [a,b]\times[c,d] \end{cases}$$
$$\Rightarrow (X,Y) \hookrightarrow POI(p_1, p_2)$$

where $\alpha \in [0, 1]$

Properties

$$\begin{cases} \mathbb{E}(X) = \dfrac{b+a}{2} \\ \mathbb{E}(Y) = \dfrac{d+c}{2} \\ \mathbb{V}(X) = \dfrac{(b-a)^2}{12} \\ \mathbb{V}(Y) = \dfrac{(d-c)^2}{12} \\ Cov(X,Y) = \dfrac{1}{36}\alpha(b-a)(d-c) \end{cases}$$

Conditional Properties

$$\begin{cases} \mathbb{E}_{\{X=x\}}(Y) = \dfrac{d+c}{2} + \dfrac{\alpha}{6(b-a)}(c^2+4cd+d^2)\left(\dfrac{2x-2a}{b-a}-1\right) \\ \mathbb{E}_{\{Y=y\}}(X) = \dfrac{b+a}{2} + \dfrac{\alpha}{6(b-a)}(a^2+4ab+b^2)\left(\dfrac{2y-2c}{d-c}-1\right) \\ \mathbb{V}_{\{X=x\}}(Y) = \dfrac{1}{36}\left(\dfrac{d-c}{b-a}\right)^2\left[\alpha^2(a+b)(4x-a-b)+3(b-a)^2-4\alpha^2x^2\right] \\ \mathbb{V}_{\{Y=y\}}(X) = \dfrac{1}{36}\left(\dfrac{b-a}{d-c}\right)^2\left[\alpha^2(c+d)(4y-c-d)+3(d-c)^2-4\alpha^2y^2\right] \end{cases}$$

12.2. Bivariate Cauchy Distribution

—

12.3. Bivariate Gamma Distribution

—

12.4. Bivariate Beta Distribution

—

12.5. Bivariate Normal Distribution

—

12.6. Bivariate Logistic Distribution

—

13. Sequences of Random Variables and Order Statistics

13.1. Distribution of Sample Mean and Variance

For a random sample,

$$\begin{cases} \overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \text{ sample mean} \\ S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \text{ sample variance} \end{cases}$$

are called statistics.

Property

$(X_i)_{1\le i\le n}$: mutually independent RV, with $(f_i(x_i))_{1\le i\le n}$ and $(\mathbb{E}(u_i(X_i)))_{1\le i\le n}$

$$\Rightarrow \mathbb{E}\left(\prod_{i=1}^{n} u_i(X_i)\right) = \prod_{i=1}^{n} \mathbb{E}(u_i(X_i))$$

$(X_i)_{1\le i\le n}$: mutually independent RV, with $(\mu_i)_{1\le i\le n}$ and $\left(\sigma_i^2\right)_{1\le i\le n}$

$$\Rightarrow \begin{cases} \mu_Y = \displaystyle\sum_{i=1}^{n} a_i\mu_i \Leftarrow Y = \sum_{i=1}^{n} a_i X_i \\ \sigma_Y^2 = \displaystyle\sum_{i=1}^{n} a_i^2\sigma_i^2 \Leftarrow Y = \sum_{i=1}^{n} a_i X_i \end{cases}$$

$(X_i)_{1\le i\le n}$: mutually independent RV, with $(M_{X_i}(t)_i)_{1\le i\le n}$

$$\left(\Rightarrow Y = \sum_{i=1}^{n} a_i X_i \Rightarrow M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)\right)$$

$(X_i)_{1\le i\le n}$: mutually independent RV, with $\left(\chi^2(r_i)\right)_{1\le i\le n}$

$$\Rightarrow \left(Y = \sum_{i=1}^{n} X_i \Rightarrow Y = \chi^2\left(\sum_{i=1}^{n} r_i\right)\right)$$

$(Z_i)_{1\le i\le n}$: mutually independent RV and follow a standard normal distribution

$$\Rightarrow \sum_{i=1}^{n} Z_i \hookrightarrow \chi^2(n)$$

$(X_i)_{1\leq i\leq n}$: mutually independent RV and follow a normal distribution

$$\Rightarrow \begin{cases} \dfrac{(n-1)S_n^2}{\sigma^2} \hookrightarrow \chi^2(n-1) \\ \overline{X_n} \& S_n^2 : \textit{independent} \end{cases}$$

13.2. Law of Large Numbers

Markov inequality

$$X \neq \underline{0} \Rightarrow \mathbb{P}\left(\{X \geq t\}\right) \leq \frac{\mathbb{E}\left(X\right)}{t}$$

Theorem weak law of large numbers : Let $(X_i)_{1\leq i\leq n}$: independent & identically distributed RV

$$\lim_{n\to\infty} \mathbb{P}\left(\left\{\left|\overline{S}_n - \mu\right| \geq \epsilon\right\}\right) = 0 \text{ with } \overline{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Convergence in probability Suppose $(X_i)_{1\leq i\leq n}$ is a sequence of random variables de- fined on a sample space S. The sequence "converges in probability" to the random variable X if, for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\{|X_n - X| < \epsilon\}\right) = 1$$

Convergence almost surely Suppose the RV X and $(X_i)_{1\leq i\leq n}$ is a sequence of random variables de- fined on a sample space S. The sequence $\overline{X_n}(\omega)$ "converges almost surely" to $X(\omega)$ if

$$\mathbb{P}\left(\left\{w \in S | \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

Properties

— For a Bernoulli distribution, $\overline{S}_n$ converges in probability to p

— For a Normal distribution, $\overline{S}_n$ converges almost surely to μ

13.3. Central Limit Theorem

The central limit therorem (Lindeberg-Levy Theorem) states that for any population distribution, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size.

Central Limit Theorem

$$\begin{cases} (X_i)_{1\leq i\leq n} \hookrightarrow ?(\mu, \sigma^2) \\ n \to \infty \end{cases} \quad \Rightarrow \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Convergence in distribution Consider X with its cumulative density function F and $(X_i)_{1\leq i\leq n}$ with their cdf $(F_i)_{1\leq i\leq n}$:

$$\lim_{n\to\infty} F_n(x) = F(x) \Rightarrow X_n \text{ "converges in distribution" to } X$$

Lévy Continuity Theorem

$$\begin{cases} (X_i)_{1\leq i\leq n} \text{ RV} \\ (F_i)_{1\leq i\leq n} \text{ distribution functions} \\ (M_{X_i})_{1\leq i\leq n} \text{ moment generating function} \end{cases}$$
$$\forall t \in [-h,h] \lim_{n\to\infty} M_{X_n}(t) = M_X(t) \Rightarrow \lim_{n\to\infty} F_n(x) = F(x)$$

13.4. Order Statistics

Order Statistics Let $(X_i)_{1\leq i\leq n}$ be observations from a random sample from a distribution f and $\left(X_{(i)}\right)_{1\leq (i)\leq n}$ the sorted observations from the smallest to the higher which constituted the **order statistics** of the sample $(X_i)_{1\leq i\leq n}$.
The sample range, R, is the distance between the smallest and the largest observation :

$$\boxed{R = X_{(n)} - X_{(1)}}$$

Let $(X_i)_{1\leq i\leq n}$ be a random sample of size n from a distribution with density function $f(x)$. Then the probability density

$$\boxed{g(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1} f(x)[1-F(X)]^{n-r}}$$

where $F(x)$ denotes the cdf of $f(x)$.

13.5. Sample Percentiles

Sample Median Let $(X_i)_{1\leq i\leq n}$ be a random sample. The sample median M is defined as :

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}\right] & \text{if } n \text{ is even} \end{cases}$$

The median is a measure of location like sample mean.

14. Sampling Distributions associated with the Normal population

Given a random sample $(X_i)_{1\leq i\leq n}$ from a population X with probability distribution $f(x;\theta)$ where θ is a parameter, a statistics is a function T of $(X_i)_{1\leq i\leq n}$ that is :

$$T = T\left((X_i)_{1\leq i\leq n}\right)$$

14.1. Monte Carlo approximation

In general, computing the distribution of a function of a random variable using the change of variables formula can be difficult. One simple but powerful alternative is to generate S samples from distribution, one popular method for higher dimensional distributions, is called Markov chain Monte Carlo or MCMC. Given the samples we can approximate the distribution of $f(X)$ by using the empirical distribution of $\{f(x_s)\}_{s=1}^{S}$ We can varying many quantities such as :

— $\mathbb{E}(X) \to \overline{x} = \dfrac{1}{S}\displaystyle\sum_{s=1}^{S} x_S$

— $\mathbb{V}(X) \to \dfrac{1}{S}\displaystyle\sum_{s=1}^{S}(x_s - \overline{x})^2$

— $\mathbb{P}(\{X \leq c\}) \to \dfrac{1}{S}Card(\{x_s \leq c\})$

— $median(X) \to median(\{x_i\}_{1\leq i\leq S})$

14.2. Chi-Square distribution

Definition A continuous variable X is said have a Chi-Square distribution with r degrees of freedom if its probability density function is of the form

$$f(x;r) = \begin{cases} \dfrac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}} x^{\frac{r}{2}-1}e^{-\frac{x}{2}} & \text{if } 0 \leq x \leq \infty \\ 0 & \text{otherwise} \end{cases}$$

Recall that a gamma distribution reduces to Chi-Square distribution if $\alpha = \frac{r}{2}$ and $\theta = 2$.

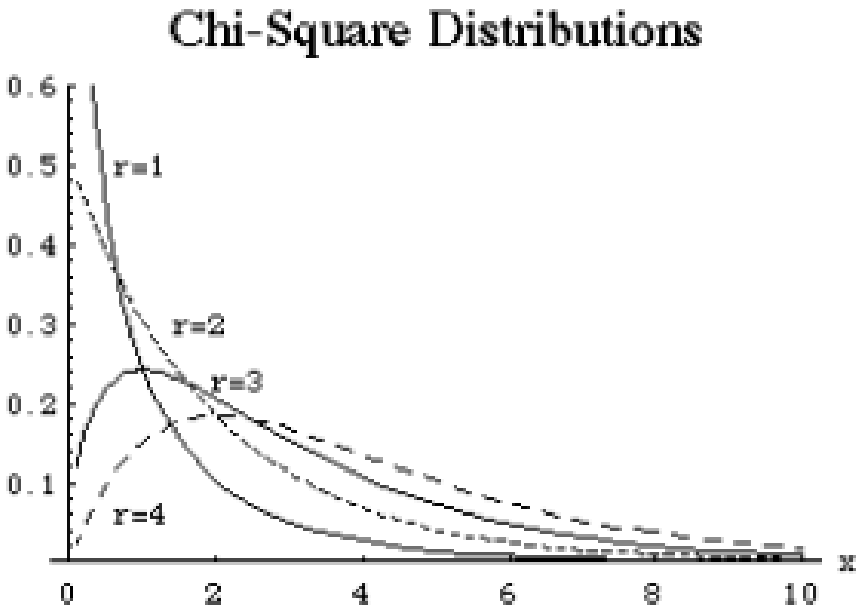

FIGURE 14.1. – If $r \to \infty$ then chi-square distribution tends to normal distribution

Properties

Population If $X \hookrightarrow N(\mu, \sigma^2)$, then $\left(\frac{X-\mu}{\sigma}\right)^2 \hookrightarrow \chi^2(1)$

Sample If $X \hookrightarrow N(\mu, \sigma^2)$ and $(X_i)_{1 \leq i \leq n}$ is a random sample from X, then :

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \hookrightarrow \chi^2(n)$$

Sample variance If $X \hookrightarrow N(\mu, \sigma^2)$ and $(X_i)_{1 \leq i \leq n}$ is a random sample from X, then :

$$\frac{(n-1)S^2}{\sigma^2} \hookrightarrow \chi^2(n-1)$$

Gamma IF $X \hookrightarrow \gamma(\theta, \alpha)$, then :

$$\frac{2}{\theta} \hookrightarrow \chi^2(2\alpha)$$

14.3. Student's t-distribution

Definition A continuous random variable X is said to have a t-distribution with ν degrees of freedom if its probability density function is of the form :

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, -\infty \leq x \leq \infty$$

where $\nu > 0$. If X has a t-distribution with ν degrees of freedom, then we denote it by writing $X \hookrightarrow t(\nu)$
The distribution is a generalization of the Cauchy distribution and the normal distribution :

$$\begin{cases} \nu = 1 \Rightarrow \forall x \in \mathbb{R} & f(x;\nu) = \dfrac{1}{\pi(1+x^2)} \\ \nu \to \infty \Rightarrow \forall x \in \mathbb{R} & \lim\limits_{\nu\to\infty} f(x;\nu) = \dfrac{1}{2\pi} e^{-\frac{1}{2}x^2} \end{cases}$$

Properties

Expected value and Variance If the random variable X has a t-distribution with ν degrees of freedom, then :

$$\mathbb{E}(X) = \begin{cases} 0 & \text{if } \nu \geq 2 \\ DNE & \text{if } \nu = 1 \end{cases}$$

$$\mathbb{V}(X) = \begin{cases} \dfrac{\nu}{\nu-2} & \text{if } \nu \geq 3 \\ DNE & \text{if } \nu \in [\![1,2]\!] \end{cases}$$

Normal and Chi-Squared distribution

$$\begin{cases} Z \hookrightarrow N(0,1) \\ U \hookrightarrow \chi^2(\nu) \\ Z \text{ and } U \text{ independents} \end{cases} \Rightarrow W = \frac{Z}{\sqrt{\frac{U}{\nu}}} \hookrightarrow t(\nu)$$

Normal variable sample

$$\begin{cases} X \hookrightarrow N(\mu,\sigma^2) \\ (X_i)_{1\leq i\leq n} \text{ a sample of the population } X \end{cases} \Rightarrow \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \hookrightarrow t(n-1)$$

14.4. Snedecor's F-distribution

Definition A continuous random variable is said to have a F-distribution with ν_1 et ν_2 degrees of freedom if its probability density function is of the form :

$$f(x,\nu_1,\nu_2) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}x\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

The F-distribution was named in honor of Sir Ronald Fisher by George Snedecor. F-distribution arises as the distribution of a ratio of variances.

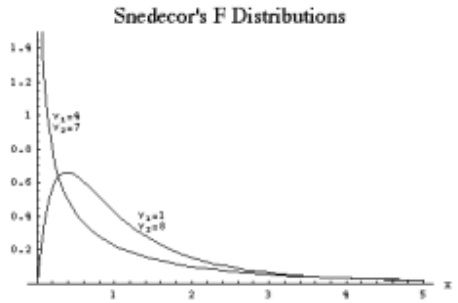

FIGURE 14.2. – Shape of the graph of F-distribution for various degrees of freedom.

Properties

Expected value and Variance $X \hookrightarrow F(\nu_{1,nu_2}) \Rightarrow$

$$E(X) = \begin{cases} \frac{nu_2}{nu_2-2} & \text{if } \nu_2 \geq 3 \\ DNE & \text{if } \nu_2 \in [\![1,2]\!] \end{cases}$$

And

$$V(X) = \begin{cases} \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} & \text{if } \nu_2 \geq 5 \\ DNE & \text{if } \nu_2 \in [\![1,4]\!] \end{cases}$$

Inverse $X \hookrightarrow F(\nu_1,\nu_2) \Rightarrow \frac{1}{X} \hookrightarrow F(\nu_2,\nu_1)$

Chi-squared and F-distributions $\begin{cases} U \hookrightarrow \chi_2(\nu_1) \\ V \hookrightarrow \chi_2(\nu_2) \\ U\&V independent \end{cases} \Rightarrow \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \hookrightarrow F(\nu_1,\nu_2)$

Quotient of Inverses Let $X \hookrightarrow N(\mu_1, \sigma_1^2)$ and $(X_i)_{1 \leq i \leq n}$ be a random sample of size n from the population X. Let $Y \hookrightarrow N(\mu_2, \sigma_2^2)$ and $(Y_i)_{1 \leq i \leq m}$ be a random sample of size n from the population Y.

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \hookrightarrow F(n-1, m-1)$$

where S_1^2, and S_2^2 denote the sample variances of the first and second sample, respectively.

15. Generative models for discrete data

15.1. Bayesian concept learning

Psychological research has shown that people can learn from positive examples alone. The **concept learning** can be thought as a binary classification. Let's define $f(x) = 1$ if x is an example of the concept $\mathcal{C}$, $f(x) = 0$ otherwise.
Then the goal is to learn the indicator function f, which just defines the concept $\mathcal{C}$. The standard binary classification techniques require positive and negative examples. By contrast we will devise a way to learn from positive examples alone. We can then propose a concept $\mathcal{C}$, for example "prime number", and then we look into a randomly selected example $\mathcal{D} = \{x_i\}_{1 \leq i \leq N}$ Some examples are more likely to belong to the concept $\mathcal{C}$.
$\mathbb{P}_{\{\mathcal{D}\}}(\{\tilde{x} \in \mathcal{C}\})$ the probability that $\tilde{x} \in \mathcal{C}$ knowing that $\tilde{x}$ comes from $\mathcal{D}$. This probability is called **posterior predictive distribution**. Sometimes given $\mathcal{D}$ we have to find the concept $\mathcal{C}$ among what we call the **hypothesis space** $\mathcal{H}$.

15.1.1. Likelihood

Let's consider the set $\mathcal{D} = \{16, 8, 2, 64\}$ we would like to find the underneath concept. Consider the following *version space*, meaning a subset of $\mathcal{H}$ that is consistent with the data in $\mathcal{D}$: {'$h_{x=2^k}$:powers of two', $h_{x=2k}$'even numbers'}. We would like to avoid *suspicious coincidences*. We compute then the probability of independently sampling N items (with replacement) from h is given by :

$$\mathbb{P}_{\{h\}}(\{\mathcal{D}\}) = \left[\frac{1}{size(h)}\right]^N.$$

This crucial equation embodies what we size principle more commonly known as **Occam's razor**, it means that the model favors the simplest hypothesis consistent with the data.

15.1.2. Prior

As prior we could choose a concept like "h' : powers of two except 32", meaning the prior might be different than mine, this subjective aspect of Bayesian reasoning is the object of lot of controversy.

15.1.3. posterior

Posterior is by default the likelihood times the prior, normalized.

$$\mathbb{P}_{\{\mathcal{D}\}}(\{h\}) = \frac{\mathbb{P}_{\{h\}}(\{\mathcal{D}\})\,\mathbb{P}(\{h\})}{\sum_{h' \in \mathcal{H}} \mathbb{P}(\{\mathcal{D} \cap h'\})} = \frac{\frac{\mathbb{P}(\{h\}) \times \mathbb{1}_{\{\mathcal{D} \in h\}}}{size(h)^N}}{\sum_{h' \in \mathcal{H}} \frac{\mathbb{P}(\{h'\})\mathbb{1}_{\{\mathcal{D} \in h'\}}}{size(h)^N}}$$

As the MAP estimate can be written as :

$$\begin{aligned}
\hat{h}_{MAP} &= \arg\max \mathbb{P}_{\{h\}}(\{\mathcal{D}\}) \times \mathbb{P}(\{h\}) \\
&= \arg\max(\log(\mathbb{P}_{\{h\}}(\{\mathcal{D}\})) + \log(\mathbb{P}(\{h\}))) \\
&= \arg\max_{h}(N \times \log\left(\frac{1}{size(h)}\right) + \log(\mathbb{P}(\{h\}))
\end{aligned}$$

We deduce that when we reach a sufficiently high level of observations the MAP converges towards the *maximum likelihood estimate (MLE)*.
Meaning, if we have enough data, the knowledge about data overwhelms the prior.

15.1.4. Posterior predictive distribution

It is the way to test our posterior, meaning our belief state about the world.

$$\mathbb{P}_{\{\mathcal{D}\}}(\{\tilde{x} \in \mathcal{C}\}) = \sum_{h} \mathbb{P}_{\{\tilde{x},h\}}(\{y = 1\})\,\mathbb{P}_{\{\mathcal{D}\}}(\{h\})$$

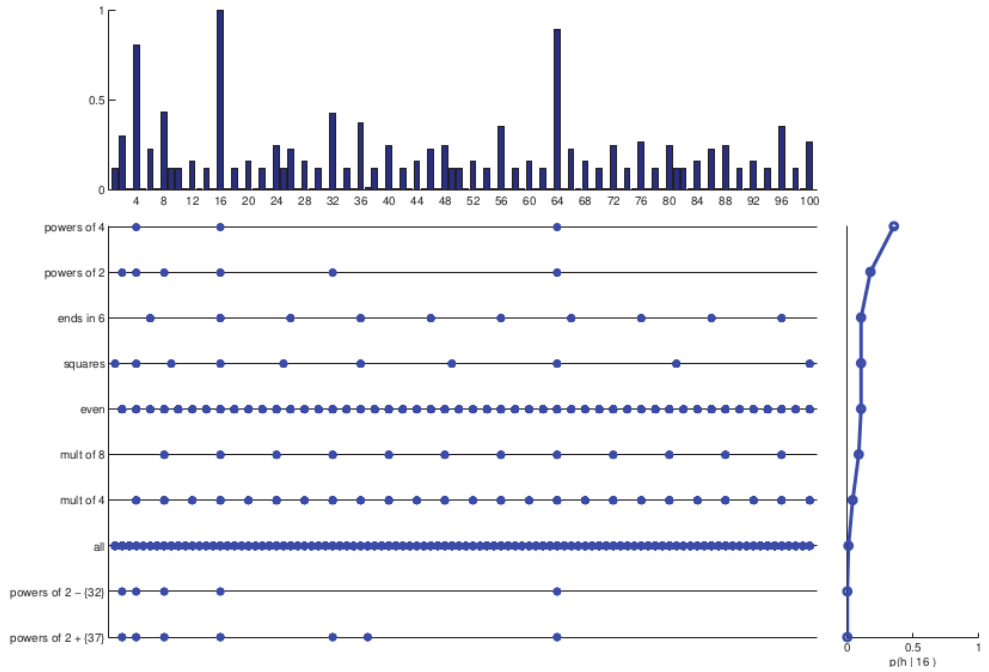

FIGURE 15.1. – First assume $\mathcal{D} = \{16\}$
Each row represents an hypothesis, the dots inside are the numbers consistent with this hypothesis. The graph on the right is the $\mathbb{P}_{\{\mathcal{D}\}}(\{h\})$ the weight given to the hypothesis h. Finally by taking a weighted sum of dots, we get $\mathbb{P}_{\{\mathcal{D}\}}(\{\tilde{x} \in \mathcal{C}\})$.

15.2. Naive Bayes classifiers

15.2.1. Naive Bayes classifiers

The aim is to classify vectors of discrete-valued features $\boldsymbol{x} \in \{x_i\}_{1 \leq i \leq K}^{D}$. Using generative approach requires us to specify the class conditional distribution $\mathbb{P}_{\{y=c\}}(\{\boldsymbol{x}\})$. We assume that the features are *conditionally independent* given the class label.

$$\mathbb{P}_{\{y=c,\theta\}}(\{\boldsymbol{x}\}) = \prod_{j=1}^{D} \mathbb{P}_{\{y=c,\boldsymbol{\theta}_{jc}\}}(\{x_j\})$$

This model is qualified of naive because actually we do not expect the features to be independent, even conditional on the class label.
Considering c parameters, and d features, the model is only $O(c \times d)$ and hence it is relatively immune to overfitting. Use cases :

— Real-valued features with Gaussian distribution
— Binary features
— Categorical features with multinoulli distribution.

Model fitting By computing the *MLE* or the *MAP*

MLE for NBC The probability for a single data case is given by

$$\begin{aligned}\mathbb{P}_{\{\theta\}}\left(\{x_i, y_i\}\right) &= \mathbb{P}_{\{\pi\}}\left(\{y_i\}\right)\prod_j^D \mathbb{P}_{\{\theta_j\}}\left(\{x_{ij}\}\right) \\ &= \prod_c^C \pi_c^{\mathbb{1}_{\{y_i=c\}}}\prod_j^D\prod_c^C \mathbb{P}_{\{\theta_{jc}\}}\left(\{x_{ij}\}\right)^{\mathbb{1}_{\{y_i=c\}}}\end{aligned}$$

Then the log-likelihood is given by

$$\log\left(\mathbb{P}_{\{\theta\}}\left(\{D\}\right)\right) = \sum_{c=1}^C N_c \log(\pi_c) + \sum_{j=1}^D\sum_{c=1}^C\sum_{i:y_i=c}\log\left(\mathbb{P}_{\{\theta_{jc}\}}\left(\{x_{ij}\}\right)\right)$$

Remember that the prior is just a ratio, in our case :

$$\hat{\pi_c} = \frac{N_c}{N}$$

With $N_c = \sum_i \mathbb{1}_{\{y_i=c\}}$ The MLE for the likelihood depends on the type of distribution we choose to use for each feature. For example, if we consider $x_j|y=c\ Bernoulli(\theta_{jc})$ Then $\theta_{jc} = \frac{N_{jc}}{N_c}$

Bayesian naive Bayes The trouble with maximum likelihood is that it can overfit. A simple solution to overfitting is to be Bayesian. For simplicity let's use a factored prior :

$$\mathbb{P}\left(\{\theta\}\right) = \mathbb{P}\left(\{\pi\}\right)\prod_{j=1}^D\prod_{c=1}^C \mathbb{P}\left(\{\theta_{jc}\}\right)$$

We will use a $Dir(\alpha)$ prior for $\boldsymbol{\pi}$ and a $Beta(\beta_0, \beta_1)$ prior for each θ_{jc} Then we get :

$$\begin{cases}\mathbb{P}_{\{D\}}\left(\{\theta\}\right) = \mathbb{P}_{\{D\}}\left(\{\pi\}\right)\prod_{j=1}^D\prod_{c=1}^C \mathbb{P}_{\{D\}}\left(\{\theta_{jc}\}\right) \\ \mathbb{P}_{\{D\}}\left(\{\pi\}\right) = Dir(\sum_{c=1}^C N_c + \alpha_c) \\ \mathbb{P}_{\{D\}}\left(\{\theta_{jc}\}\right) = Beta(N_c - N_{jc} + \beta_0, N_{jc} + \beta 1)\end{cases}$$

In other words, to compute teh posterior we just update the prior counts with the empirical counts from the likelihod.

Using the model for prediction Aim : compute $\mathbb{P}_{\{\boldsymbol{x},\mathcal{D}\}}(\{y=c\}) \propto \mathbb{P}_{\{\mathcal{D}\}}(\{y=c\}) \prod_{j=1}^{D} \mathbb{P}_{\{y=c,\mathcal{D}\}}(\{x_j\})$

The correct Bayesian procedure is to integrate out the unknown parameters

$$\begin{aligned}\mathbb{P}_{\{\boldsymbol{x},\mathcal{D}\}}(\{y=c\}) &\propto \int Cat(y=c|\boldsymbol{\pi})\mathbb{P}_{\{\mathcal{D}\}}(\{\boldsymbol{\pi}\}) \\ &= \prod_{j=1}^{D} \int Ber(x_j|y=c,\theta_{jc})\mathbb{P}_{\{\mathcal{D}\}}(\{\theta_{jc}\})\end{aligned}$$

Fortunately, this is easy to do, at least if the posterior is Dirichlet.

The log-sum-exp trick Unfortunately a naive implementation can fail due to numerical underflow, the problem is that $\mathbb{P}_{\{y=c\}}(\{\boldsymbol{x}\})$ is often very small, especially if $\boldsymbol{x}$ is a high-dimensional vector. This is because we require that $\sum_{\boldsymbol{x}} \mathbb{P}_{\{y\}}(\{\boldsymbol{x}\}) = 1$.

The obvious solution is to take logs when applying the Bayes rule : $\begin{cases} b_c \triangleq \log(\mathbb{P}_{\{y=c\}}(\{\boldsymbol{x}\})) + \log(\mathbb{P}(\{y=c\})) \\ \log(\mathbb{P}_{\{\boldsymbol{x}\}}(\{y=c\})) = b_c - \log\left(\sum_{c'=1}^{C} e^{b_{c'}}\right) \end{cases}$

This requires evaluating the last member of the first above equation, but we can factor out the largest term and just represent the remaining numbers relative to that In general

$$\begin{aligned}\log\left(\sum_c e^{b_c}\right) &= \log\left(\left(\sum_c e^{b_c - B}\right) e^B\right) \\ &= \log\left(\sum_c e^{b_c - B}\right) + B\end{aligned}$$

where $B = \max_c b_c$ This technique is widely used.

15.2.2. Feature selection using mutual information

In using

$$I(X,Y) = \sum_{x_j}\sum_{y} \mathbb{P}(\{x_j,y\}) \times \log\left(\frac{\mathbb{P}(\{x_j,y\})}{\mathbb{P}(\{x_j\})\,\mathbb{P}(\{y\})}\right)$$

The mutual information can be thought of as the reduction in entropy on the label distribution once we observe the value of feature j.

16. Some Techniques for finding point Estimators of Parameters

In point estimation, we try to find the parameter θ of the population distribution $f(x;\theta)$ from the sample information. Thus, in the parametric point estimation one assumes the functional form of the pdf $f(x;\theta)$ to be known and only estimate the unknown parameter θ of the population using information available from the sample.

16.1. Moment Method

Parameter space Let X be a population with the density function $f(x;\theta)$, where θ is an unknown parameter. The set of all admissible values of θ is called a parameter space and it is denoted by Ω that is :

$$\Omega = \{\theta \in \mathbb{R}^n | f(x;\theta) \text{ is a pdf}\}$$

Estimators Let $X \hookrightarrow f(x;\theta)$ and $(X_i)_{1\leq i\leq n}$ be a random sample from population X. Any statistic that can be used to guess the parameter θ is called an estimator of θ. The numerical value of this statistic is called an estimate of θ.
The estimator of the parameter θ is denoted by $\hat{\theta}$

Moment method Let $(X_i)_{1\leq i\leq n}$ be a random sample from a population X with probability density function $f\left(x;(\theta_j)_{1\leq j\leq m}\right)$ where $(\theta_j)_{1\leq j\leq m}$ are m unknown parameters.
Let : $\mathbb{E}\left(X^k\right) = \displaystyle\int_{-\infty}^{\infty} x^k f\left(x;(\theta_j)_{1\leq j\leq m}\right) dx$ is the k^{th} population moment about 0.
Further, let $M_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$ be the k^{th} sample moment about 0.
Then we find the estimator for the parameters $(\theta_j)_{1\leq j\leq m}$ by equating the first m population moments (if they exist) to the first m sample moments that is :

$$\begin{cases} \mathbb{E}(X) & = M_1 \\ \mathbb{E}\left(X^2\right) & = M_2 \\ \mathbb{E}\left(X^3\right) & = M_3 \\ & \cdot \\ & \cdot \\ & \cdot \\ \mathbb{E}\left(X^m\right) & = M_m \end{cases}$$

The motivation of moment method comes from the fact that the sample moments are in some sense estimates for the population moments.

16.2. Maximum Likelihood Method

Definition Let $(X_i)_{1\leq i\leq n}$ be a random sample from a population X with probability density function $f(x;\theta)$, where θ is an unknown parameter. The likelihood, $L(\theta)$, is the distribution of the sample. That is :

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

The θ that maximizes the likelihood function $L(\theta)$ is called the maximum likelihood estimator of θ and it is denoted by $\hat{\theta}$

Method

1. Obtain a random sample $(x_i)_{1\leq i\leq n}$ from the distribution X with probability density function $f(x;\theta)$
2. Define the likelihood function for the sample $(x_i)_{1\leq i\leq n}$ by $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$
3. find the expression for θ that maximizes $L(\theta)$. This can be done directly or by maximizing $ln\left(L(\theta)\right))$
4. replace θ by $\hat{\theta}$ to obtain an expression for the maximum likelihood estimator for θ
5. find the observed value of this estimator for a given sample.

Theorem Let $\hat{\theta}$ be a maximum likelihood estimator of a parameter θ and let $g(\theta)$ be a function of θ. Then the maximum likelihood estimator of $g(\theta)$ is given by $g\left(\hat{\theta}\right)$

Fisher information Let X be an observation from a population with probability density function $f(x;\theta)$. Suppose $f(x;\theta)$ is continuous, twice differentiable and it's support does not depend on θ. Then the Fisher information $I(\theta)$ in a single observation X about θ is given by :

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d\ ln\left(f(x;\theta)\right)}{d\theta}\right]^2 f(x;\theta)dx$$

It can be given alternatively as :

$$I(\theta) = -\int_{-\infty}^{\infty} \left[\frac{d^2\ ln\left(f(x;\theta)\right)}{d\theta^2}\right] f(x;\theta)dx$$

$$\begin{cases} (X_i)_{1\leq i\leq n} \text{ random sample from } X \\ X \hookrightarrow f(x;\theta) \end{cases} \Rightarrow I_n(\theta) = nI(\theta)$$

Theorem Under certain regularity conditions on the $f(x;\theta)$ the maximum likelihood estimator $\hat{estimator}$ of θ based on a random sample of size n from a population X with probability density $f(x;\theta)$ is asymptotically normally distributed with mean θ and variance $\frac{1}{nI(\theta)}$. That is :

$$\hat{\theta}_{ML} \hookrightarrow N\left(\theta, \frac{1}{nI(\theta)}\right) \text{ as } n \to \infty$$

16.3. Bayesian Method

Prior distribution Let $(X_i)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. The probability density function of the random variable θ is called the prior distribution of θ and usually denoted by $h(\theta)$

Posterior distribution Let $(X_i)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. The conditional density, $k\left(\theta;(x_i)_{1\leq i\leq n}\right)$, of θ given the sample $(x_i)_{1\leq i\leq n}$ is called the posterior distribution of θ.

Squared and Absolute Error Loss Let $(X_i)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. Let $\hat{\theta}$ of θ : $\begin{cases} \mathcal{L}_2(\theta) = \left(\hat{\theta}-\theta\right)^2 & \text{Squared error loss} \\ \mathcal{L}_1(\theta) = \left|\hat{\theta}-\theta\right| & \text{Absolute error loss} \end{cases}$

Risk Let $(X_i)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. Let $\hat{\theta}$ be an estimator of θ and let $\mathcal{L}\left(\hat{\theta}\right)$ be a given loss function. The expected value of this loss function with respect to the population distribution $f(x/\theta)$, that is $R_{\mathcal{L}}(\theta) = \int \mathcal{L}\left(\hat{\theta}\right) f(x/\theta)dx$

In Bayesian estimation of parameter one chooses an estimate $\hat{theta}$ for θ such that : $k\left(\hat{\theta}/(x_i)_{1\leq i\leq n}\right)$ is maximum subject to a loss function.

Mathematically this equivalent to minimizing the integral : $\int_\Omega \mathcal{L}\left(\hat{\theta},\theta\right) k\left(\theta/(x_i)_{1\leq i\leq n}\right) d\theta$ where Ω denotes the support of the prior density $h(\theta)$ of θ.

Estimator for squared error Let $(X_i)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. If the loss function is squared error, then the Bayes' estimator $\hat{\theta}$ of parameter θ is given by :

$$\hat{\theta} = \mathbb{E}\left(\theta/(x_i)_{1\leq i\leq n}\right)$$

where the expectation is taken with respect to density $k\left(\theta/\left(x_i\right)_{1\leq i\leq n}\right)$

Estimator for absolute error Let $\left(X_i\right)_{1\leq i\leq n}$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. If the loss function is absolute error, then the Bayes' estimator $\hat{\theta}$ of parameter θ is given by :

$$\hat{\theta} = \text{ median of } k\left(\theta/\left(x_i\right)_{1\leq i\leq n}\right)$$

where $k\left(\theta/\left(x_i\right)_{1\leq i\leq n}\right)$ is the posterior distribution.

16.4. Information Theory

Aim : representing data in a compact fashion.
Note that compactly representing data requires allocating short codewords to highly probable bit strings, and reserving longer codewords to less probable bit strings.

Entropy The **entropy** is a measure of its uncertainty, $\mathbb{H}(X)$

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K}\mathbb{P}\left(\{X=k\}\right)\log_2(\mathbb{P}\left(\{X=k\}\right))$$

KL Divergence

Equality Kullback-Leibler divergence or **relative entropy** allows to measure the dissimilarity of 2 probability distribution p and q.

$$\begin{aligned}\mathbb{KL}(p||q) &\triangleq \sum_{k=1}^{k} p_k \log\left(\frac{p_k}{q_k}\right)\\ &= \sum_k p_k \log\left(p_k\right) - \sum_k p_k \log\left(q_k\right)\\ &= -\mathbb{H}(p) + \underbrace{\mathbb{H}(p,q)}_{\text{cross entropy}}\end{aligned}$$

Inequality Let (p,q) 2 distinct probability distributions $\begin{cases}\mathbb{KL}\geq 0\\ p=q\Rightarrow\mathbb{KL}=0\end{cases}$

Mutual Information

Definition Aim : knowing how much knowing one varaible tells us about one other. Motivation : The correlation coefficient is only defined for real-valued variables and has some limitations. Purpose : Determining how similar the joint distribution $p(X, Y)$ is to the factored distribution $p(X), p(Y)$

$$\begin{aligned}\mathcal{I}(X:Y) &\triangleq \mathcal{KL}(p(X,Y)||p(X),p(Y)) \\ &= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)\end{aligned}$$

Pointwise Mutual Information This measures the discripencey between these events occuring together compared to what would be expected by chance. MI is indeed the expected value of the PMI

$PMI(x,y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \log\left(\frac{p(x|y)}{p(x)}\right) = \log\left(\frac{p(y|x)}{p(y)}\right)$a

This is the amount we learn from updating the prior $p(x)$ into the posterior $p(x|y)$, or equivalently updating the prior $p(y)$ into the posterior $p(y|x)$.

Maximal Information Coefficient For continuous random variables, it is common to **discretize** by dividing the ranges of each variable into bins. As the bin boundaries can have a significant impact, we can try many different bin sizes and location, then compute the maximum MI achieved. This statistic, appropriately normalized, is known as the Maximal Information Coefficient (MIC).

$$MIC \triangleq \max_{x,y:xy<B} \frac{\max_{G\in\mathcal{G}(x,y)} \mathcal{I}(X(G):Y(G))}{\log(\min(x,y))}$$

17. Criteria for evaluating the Goodness of Estimators

17.1. The Unbiased Estimator

The Unbiased Estimator $\hat{\theta} =$ unbiased estimator of $\theta \Leftrightarrow \mathbb{E}\left(\hat{\theta}\right) = \theta$
In statistics between 2 unbiased estimators one prefers the estimator which has the minimum variance.
There are 2 disadvantages :

1. A unbiased estimator for a parameter may not exist.
2. Unbiasedness is not invariant under functional transformation.

17.2. The relatively Efficient Estimator

Efficiency of estimators Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be 2 unbiased estimators of θ.

$$\mathbb{V}\left(\hat{\theta}_1\right) < \mathbb{V}\left(\hat{\theta}_2\right) \Rightarrow \hat{\theta}_1 \text{ is said to be more efficient than } \hat{\theta}_2$$

The ratio is given by :

$$\eta\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{\mathbb{V}\left(\hat{\theta}_2\right)}{\mathbb{V}\left(\hat{\theta}_1\right)}$$

is called relative efficiently of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$.

17.3. The Minimun Variance Unbiased Estimator

Definition Let $\hat{\theta}$ to be an unbiased estimator of θ For any unbiased estimator $\hat{T}$ of $\theta : \mathbb{V}\left(\hat{\theta}\right) \leq \mathbb{V}\left(\hat{T}\right) \Leftrightarrow \hat{\theta}$ is said to be a uniform minimum variance unbiased estimator of θ.

Definition (2) Let $\hat{\theta}$ to be an unbiased estimator of θ. $\hat{\theta}$ minimizes the variance $\mathbb{E}\left(\left[\hat{\theta} - \theta\right]^2\right) \Rightarrow \hat{\theta}$ is said to be a minimum variance unbiased estimator.

Minoration of an unbiased estimator For any $h\left((x_i)_{1\leq i\leq n}\right)$ with $\mathbb{E}\left(h\left((x_i)_{1\leq i\leq n}\right)\right) < \infty$

$$\begin{cases} X \text{ a population with pdf } f(x;\theta) \\ (X_i)_{1\leq i\leq n} \text{ a random sample of size } n \text{ from X} \\ \hat{\theta} \text{ any unbiased estimator of } \theta \\ L(\theta) \text{ the likelihood function is differentiable} \\ \frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h\left((x_i)_{1\leq i\leq n}\right) L(\theta)dx_1\cdots dx_n = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h\left((x_i)_{1\leq i\leq n}\right)\frac{d}{d\theta}L(\theta)dx_1\cdots dx_n \end{cases}$$

$$\Rightarrow \mathbb{V}\left(\hat{\theta}\right) \geq \frac{1}{\mathbb{E}\left(\left[\frac{\partial\ ln(L(\theta))}{\partial\theta}\right]^2\right)}$$

If $L(\theta)$ is twice differentiable with respect to θ, the last inequality can be stated equivalently as $\mathbb{V}\left(\hat{\theta}\right) \geq \frac{-1}{\mathbb{E}\left(\frac{\partial^2\ ln(L(\theta))}{\partial\theta^2}\right)}$

Theorem $\begin{cases} X \text{ a population with pdf } f(x;\theta) \\ (X_i)_{1\leq i\leq n} \text{ a random sample of size } n \text{ from X} \\ \hat{\theta} \text{ an unbiased estimator} \\ \mathbb{V}\left(\hat{\theta}\right) = \frac{1}{\mathbb{E}\left(\left[\frac{\partial\ ln(L(\theta))}{\partial\theta}\right]^2\right)} \end{cases}$

$\Rightarrow \hat{\theta}$ is a minimum variance unbiased estimator of θ

Cramér-Rao lower bound Let $\hat{\theta}$ to be a unbiased estimator. $\mathbb{V}\left(\hat{\theta}\right) = \frac{1}{\mathbb{E}\left(\left[\frac{\partial\ ln(L(\theta))}{\partial\theta}\right]^2\right)} \Rightarrow \hat{\theta}$ is an efficient estimator

17.4. Sufficient Estimator

Definition $\begin{cases} X \hookrightarrow f(x;\theta) be a population \\ (X_i)_{1\leq i\leq n} \text{ random sample of } X \\ \hat{\theta} \text{ an estimator of } \theta \\ \text{ the conditional distribution of the sample given } \hat{\theta} \text{ does not depend on the parameter } \theta \end{cases}$

$\Rightarrow \hat{\theta}$ is a *sufficient estimator*

Theorem $\begin{cases} X \hookrightarrow f(x;\theta) beapopulation \\ (X_i)_{1\leq i\leq n} \hookrightarrow f\left((x_i)_{1\leq i\leq n};\theta\right) \\ f\left((x_i)_{1\leq i\leq n};\theta\right) = \phi\left(\hat{\theta},\theta\right) h\left((x_i)_{1\leq i\leq n}\right) \end{cases}$

$\Rightarrow \hat{\theta}$ is sufficient for θ

Theorem $\begin{cases} X \hookrightarrow f(x;\theta) = e^{\{K(x)A(\theta)+S(x)+B(\theta)\}} \\ (X_i)_{1\leq i\leq n} \text{ a random sample from } X \end{cases}$

$\Rightarrow$ the statistic $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter θ

Necessary estimator $\hat{\theta}$ can be written as a function of every sufficient estimators $\Rightarrow$ $\hat{\theta}$ is a *necessary estimator*

17.5. Consistent Estimator

Consistent sequence of estimators $\begin{cases} X \hookrightarrow f(x;\theta) \text{ a population} \\ (X_i)_{1\leq i\leq n} \text{ a sample from } X \\ \{\theta_n\} \text{ a sequence of estimators} \\ \forall \epsilon > 0, \lim\limits_{n\to\infty} \mathbb{P}\left(\left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\}\right) = 0 \end{cases}$

$\Rightarrow \{\theta_n\}$ is said to be *consistent* for θ

Theorem $\begin{cases} X \hookrightarrow f(x;\theta) \text{ a population} \\ (X_i)_{1\leq i\leq n} \text{ a sample from } X \\ \{\theta_n\} \text{ a sequence of estimators} \\ \forall n \in \mathbb{N}, \mathbb{V}\left(\hat{\theta}_n\right) < \infty \\ \lim\limits_{n\to\infty} \mathbb{E}\left(\left[\hat{\theta}_n - \theta\right]^2\right) = 0 \end{cases} \Rightarrow \forall \epsilon > 0 \lim\limits_{n\to\infty} \mathbb{P}\left(\left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\}\right) = 0$

Limit of a sequence variance $\begin{cases} X \hookrightarrow f(x;\theta) \\ (X_i)_{1\leq i\leq n} \text{ be a random sample from } X \\ \left\{\hat{\theta}_n\right\} \text{ a sequence of estimators of } \theta \text{ based on the sample} \\ \forall n \in \mathbb{N}, \mathbb{V}\left(\hat{\theta}_n\right) \leq \infty \end{cases}$

$\Rightarrow \lim\limits_{n\to\infty} \mathbb{E}\left(\left[\hat{\theta}_n - \theta\right]^2\right) = 0$

Deuxième partie

Tests

18. Some Techniques for finding Interval Estimators of Parameters

18.1. Interval Estimators and Confidence Intervals for Parameters

Interval estimator & Interval estimate Let $(X_i)_{1\leq i\leq n}$ be a random sample from $X \hookrightarrow f(x;\theta)$.
The *interval estimator* of θ is a pair of statistics $L = L((X_i)_{1\leq i\leq n})$ and $U = U((X_i)_{1\leq i\leq n})$ with $L \leq U$ such that if $(x_i)_{1\leq i\leq n}$ is a set of sample data, then θ belongs to the interval $\left[L((X_i)_{1\leq i\leq n}), U((X_i)_{1\leq i\leq n})\right]$
The interval $[l,u]$ will be denoted as an interval estimate of θ of θ whereas the random interval [L,U] will denote the interval estimator of θ

Confidence interval $\begin{cases} X \hookrightarrow f(x;\theta) \\ (X_i)_{1\leq i\leq n} \text{ a random sample from } X \\ \mathbb{P}(\{L \leq \theta \leq U\}) = 1-\alpha \end{cases}$
$\Rightarrow$ The interval estimator of θ is called a $100(1-\alpha)\%$ *confidence interval* for θ

18.2. Pivotal Quantity Method

Definition Let $(X_i)_{1\leq i\leq n}$ be a random sample from $X \hookrightarrow f(x;\theta)$.
A *pivotal quantity* Q is a function of $(X_i)_{1\leq i\leq n}$ and θ *whose pdf is independent of the parameter* θ

Location-scale family $g : \mathbb{R} \to \mathbb{R}$ a pdf
$\Rightarrow \mathcal{F} = \left\{f(x;\mu;\sigma) = \frac{1}{\sigma}g\left(\frac{x-\mu}{\sigma}\right) | (\mu,\sigma) \in \mathbb{R}\times\mathbb{R}_+^*\right\}$ is called the *location-scale family* with standard probability function density $f(x;\theta)$
The parameter μ is called the *location parameter* and the parameter σ is called the *scale parameter*.
$\sigma = 1 \Rightarrow \mathcal{F}$ is called *location* family
$\mu = 0 \Rightarrow \mathcal{F}$ is called *scale* family

Remark Let be $\hat{\mu}$ and $\hat{\sigma}$ respectively the maximum likelihood of μ and σ :
Density $\in$ *location* family $\Rightarrow \hat{\mu} - \mu$ is the pivot for μ

Density $\in$ *scale* family $\Rightarrow \frac{\hat{\sigma}}{\sigma}$ is the pivot for σ

Density $\in$ *location-scale* family $\Rightarrow \begin{cases} \frac{\hat{\mu}-\mu}{\hat{\sigma}} \text{ is the pivot for } \mu \\ \frac{\hat{\sigma}}{\sigma} \text{ is the pivot for } \sigma \end{cases}$

18.3. Confidence Interval for Population Mean

Practice exercises

18.4. Confidence Interval for Population Variance

Practice exercises

18.5. Confidence Interval for Parameter of some Distribution not belonging to the Location-Scale Family

—

18.6. Approximate Confidence Interval for Parameter with MLE

—

18.7. The Statistical or General Method

—

18.8. Criteria for Evaluating Confidence Intervals

—

19. Gaussian models

19.1. Introduction to the gaussian models

19.1.1. Notion

paragraphBasics

MultiVariate Normal Recall that the probability density functon for a *MultiVariate Normal* in D dimensions is

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Mahalanobis corresponds to $-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, which is the *distance between* $\boldsymbol{x}$ *and the mean vector* $\boldsymbol{\mu}$.
Let look at the *eigendecomposition* of $\boldsymbol{\Sigma}$, then
$\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U^T}$, with $\boldsymbol{U}$ is an orthonormal matrix satisfying $\boldsymbol{UU^T} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ a diagonal matrix of eigenvalues.

$$\begin{aligned}\Sigma^{-1} &= U^{-T}\boldsymbol{\Lambda}^{-1}U^{-1} \\ &= \boldsymbol{U\Lambda}^{-1}\boldsymbol{U}^T \\ &= \sum_{i=1}^{D}\frac{1}{\lambda_j}\boldsymbol{u}\boldsymbol{u}_j^T\end{aligned}$$

where $\boldsymbol{u}_j$ is the jth eigenvector.
We can then write :

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{u})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{u}) &= (\boldsymbol{x}-\boldsymbol{u})^T\sum_{i=1}^{D}\frac{1}{\lambda_j}\boldsymbol{u}\boldsymbol{u}_j^T(\boldsymbol{x}-\boldsymbol{u}) \\ &= \sum_{i=1}^{D}\frac{1}{\lambda_j}(\boldsymbol{x}-\boldsymbol{u})^T\boldsymbol{u}\boldsymbol{u}_j^T(\boldsymbol{x}-\boldsymbol{u}) \\ &= \sum_{i=1}^{D}\frac{y_i^2}{\lambda_j}\end{aligned}$$

Then for $D = 2$ we have the equation of an ellipse.

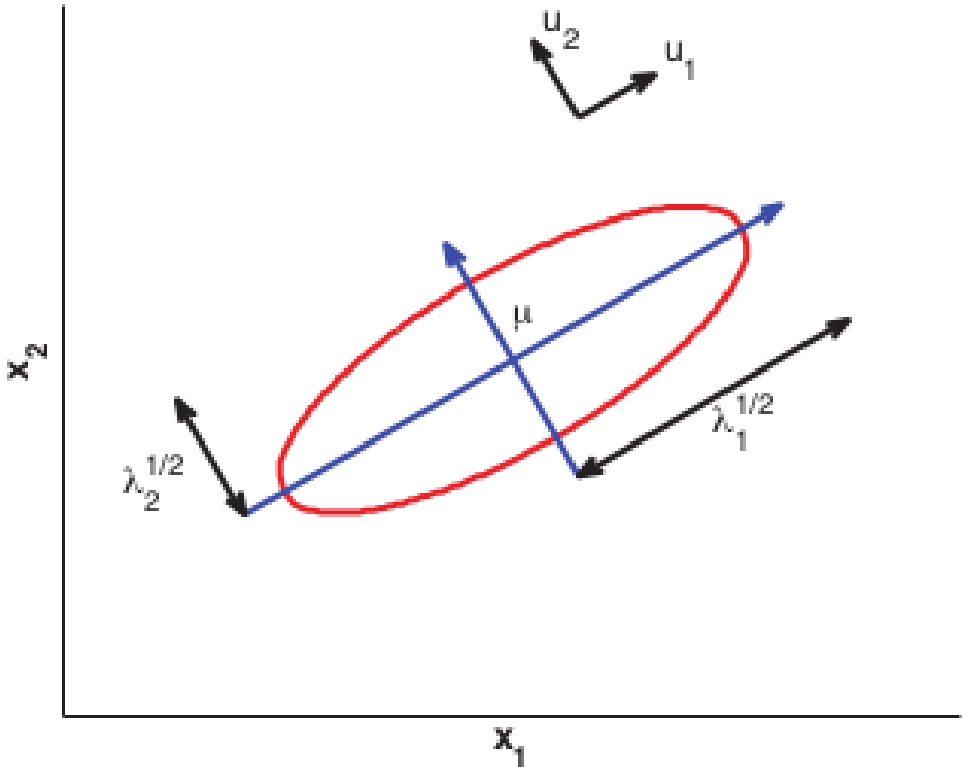

FIGURE 19.1. – Mahalanobis distance interpretation in 2D

MLE for an MVN

MLE for a Gaussian If we have n samples $\boldsymbol{x}_i \hookrightarrow \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{cases} \hat{\boldsymbol{\mu}}_{mle} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \triangleq \overline{\boldsymbol{x}} \\ \hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = \frac{1}{N}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T \end{cases}$$

That is, the *MLE* is just the empirical mean and empirical covariance.

Proof Reall that for $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{A}, \boldsymbol{B}) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathcal{M}(\mathbb{R})_{n,n} \times \mathcal{M}(\mathbb{R})_{n,n}$

$$\begin{cases} \frac{\partial(\boldsymbol{b}^T\boldsymbol{a})}{\partial \boldsymbol{a}} = \boldsymbol{b} \\ \frac{\partial(\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a})}{\partial \boldsymbol{a}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{a} \\ \frac{\partial tr(\boldsymbol{A}\boldsymbol{B})}{\partial \boldsymbol{a}} = \boldsymbol{B}^T \\ \frac{\partial \log(|\boldsymbol{A}|)}{\partial \boldsymbol{A}} = \boldsymbol{A}^{-T} \triangleq (\boldsymbol{A}^{-1})^T \\ tr(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = tr(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B}) = tr(\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}) \\ \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = tr(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}) = tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T) \end{cases}$$

Then, with $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$, log-likelihood is

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} &= \frac{\partial}{\partial \mu}\left(\frac{N\log(|\boldsymbol{\Lambda}|)}{2} - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Lambda(\boldsymbol{x}_i - \boldsymbol{\mu})\right) \\
&= -\frac{\partial\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}{2\partial\mu} \\
&= -\sum_{i=1}^{n}\frac{\partial(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}{2\partial\mu} \\
&= -\sum_{i=1}^{n}\frac{\partial\boldsymbol{x}_i^T\Sigma^{-1} - \boldsymbol{\mu}^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}{2\partial\mu} \\
&= -\sum_{i=1}^{n}\frac{\partial\boldsymbol{x}_i^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) - \boldsymbol{\mu}^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})}{2\partial\mu} \\
&= -\sum_{i=1}^{n}\frac{\partial\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i - \boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x}_i + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}}{2\partial\mu} \\
&= -\frac{1}{2}\sum_{i=1}^{n}0 - \frac{\partial\boldsymbol{x}_i}{\partial\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-T}\boldsymbol{x}_i - \frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i - \frac{\partial\boldsymbol{x}_i}{\partial\boldsymbol{\mu}}\boldsymbol{\Sigma}^{-T}\boldsymbol{\mu} + \frac{\partial\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}}{\partial\boldsymbol{\mu}} \\
&= -\frac{1}{2}\sum_{i=1}^{n} - (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-T})\boldsymbol{x}_i + (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-T})\boldsymbol{\mu} \\
&= \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-T})(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
&= \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-T})\boldsymbol{y}_i \\
&= \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i \\
&= 0_{p,p}
\end{aligned}$$

Considering that $S_{\boldsymbol{\mu}} \triangleq \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$

$$\begin{aligned}
\frac{\partial l(\Lambda)}{\partial \Lambda} &= \frac{\partial}{\partial \Lambda}\left(\frac{n}{2}\log(|\Lambda|) - \frac{1}{2}\sum_{i=1}^{n} tr\left([\boldsymbol{x}_i - \boldsymbol{\mu}][\boldsymbol{x}_i - \boldsymbol{\mu}]^T \boldsymbol{\Lambda}\right)\right) \\
&= \frac{\partial}{\partial \Lambda}\left(\frac{n}{2}\log(|\Lambda|) - \frac{1}{2} tr(\boldsymbol{S}_{\boldsymbol{\mu}} \boldsymbol{\Lambda})\right) \\
&= \frac{1}{2}\left(n\Lambda^{-T} - \boldsymbol{S}_{\boldsymbol{\mu}}^T\right) \\
&= 0_{p,p}
\end{aligned}$$

Maximum entropy derivation of the Gaussian The multivariate Gaussian is the distribution with the maximum entropy subject to having a specified mean and covariance. This is one reason the Gaussian is so widely used : the first two moments are usually all that we can reliably estimate from the data, so we want a distribution that captures these properties, but otherwise makes as few additional assumptions as possible.

Theorem Let $q(x)$ be any density satisfying $\int q(x) x_i x_j = \sum_{ij}$. Let $p = \mathcal{N}(0, \boldsymbol{\Sigma})$, then $h(q) \leq h(p)$

Proof

$$\begin{aligned}
0 &\leq \mathbb{KL}(q||p) \\
&= \int q(x) \log\left(\frac{q(x)}{p(x)}\right) dx \qquad\qquad = \int q(x)\log(q(x))dx - \int q(x)\log(p(x))dx \\
&= -h(q) - \int q(x)\log(p(x))dx \\
&= -h(q) - \int p(x)\log(p(x))dx \\
&= -h(q) + h(q)
\end{aligned}$$

The 2nd line from the bottom follows since q and p yield the same moments for the quadratic form encoded by $\log(p(x))$

19.2. Gaussian discriminant analysis

One important application of MultiVariate Normals is to define the class conditional densities in a generative classifier : $p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
The resulting technique is called discriminant analysis even though it is a generative not a discriminant classifier.

Quadratic discriminant analysis (QDA)

$$\begin{aligned} p(y=c|\boldsymbol{x},\boldsymbol{\theta}) &= \frac{p(y=c|\boldsymbol{\theta})p(\boldsymbol{x}|y=c,\boldsymbol{\theta})}{\sum_{c'} p(y=c'|\boldsymbol{\theta})p(\boldsymbol{x}|y=c',\boldsymbol{\theta})} \\ &= \frac{\pi_c|2\pi\boldsymbol{\Sigma_c}|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma_c}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right)}{\sum_{c'}\pi_{c'}|2\pi\boldsymbol{\Sigma_{c'}}|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma_{c'}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{c'})\right)} \end{aligned}$$

Linear discriminant analysis (LDA) Let us assume that the covariance matrices are shared across classes, $\boldsymbol{\Sigma}_{c=}\boldsymbol{\Sigma}$ then

$$\begin{aligned} &\pi_c exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma_{c'}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{c'})\right) \\ &\qquad = \pi_c exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \frac{1}{2}\left(\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)^T + \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right) \\ &\qquad = exp\left(\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log(\pi_c)\right) exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \end{aligned}$$

Then

$$\begin{aligned} p(y=c|\boldsymbol{x},\boldsymbol{\theta}) &= \frac{\pi_c|2\pi\boldsymbol{\Sigma_c}|^{-\frac{1}{2}} exp\left(\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log(\pi_c)\right) exp\left(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)}{\sum_{c'}\pi_{c'} exp\left(\boldsymbol{\mu}_{c'}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_{c'}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{c'} + \log(\pi_{c'})\right) exp\left(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)} \\ &= \frac{e^{\boldsymbol{\beta}_{c'}^T\boldsymbol{x}+\gamma_{c'}}}{\sum_{c'} e^{\boldsymbol{\beta}_c^T\boldsymbol{x}+\gamma_c}} \\ &= \mathcal{S}(\eta)_c \end{aligned}$$

With $\begin{cases} \gamma_c = \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mu_c + \log(\pi_c) \\ \beta_c = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c \\ \eta^T = \left[\beta_1^T\boldsymbol{x} + \lambda_1, \ldots, \beta_c^T\boldsymbol{x} + \lambda_c\right] \\ \mathcal{S} \text{ is the } \textbf{softmax} \text{ function} \end{cases}$ The softmax function is so-called since it acts a bit like the max function. Let us divide each η_c by a constant T called the **temperature**,

$$\lim_{T\to 0} \mathcal{S}\left(\frac{\eta}{T}\right)_c = \begin{cases} 1 \text{ if } c = \arg\max_{c'} \eta_{c'} \\ 0 \text{ otherwise} \end{cases}$$

In other words, at low temperatures, the distribution spends essentially all of its time in the most probable state, whereas at high temperatures, it visits all states uniformly.

Note that this terminology comes from the area of statistical physics, where it is common to use the **Boltzmann distribution**, which has the same form as the softmax function.
This terminology comes from the statistical physics it is common to use the *Boltzmann distribution*, which has the same form as the *softmax* function.
If we apply the *log* function we end up with a linear function of $\boldsymbol{x}$, thus the decision boundary between any two classes, say c and c', will be a straight line.

$$\begin{aligned} & p(y=c|\boldsymbol{x},\boldsymbol{\theta}) = p(y=c'|\boldsymbol{x},\boldsymbol{\theta}) \\ \Leftrightarrow\ & \beta_c^T \boldsymbol{x} + \gamma_c = \beta_{c'}^T \boldsymbol{x} + \gamma_{c'} \\ \Leftrightarrow\ & \boldsymbol{x}^T(\beta_{c'} + \beta_c) = \gamma_{c'} - \gamma_c \end{aligned}$$

Two-class LDA

$$\begin{aligned} \gamma_1 - \gamma_0 &= \frac{1}{2}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 + \log\left(\frac{\pi_1}{\pi_0}\right) \\ &= \frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)^T \Sigma^{-1} \left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0\right) + \log\left(\frac{\pi_1}{\pi_0}\right) \end{aligned}$$

Then in defining

$$\begin{cases} \boldsymbol{w} = \beta_1 - \beta_0 = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}1 - \boldsymbol{\mu}0\right) \\ \boldsymbol{x}_0 = \frac{1}{2}\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0\right) - \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right) \dfrac{\log\left(\frac{\pi_1}{\pi_0}\right)}{\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)^T \Sigma^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\right)} \end{cases}$$

Hence we can observe the similarity with logistic regression :

$$\boxed{p(y=1|\boldsymbol{x},\boldsymbol{\theta}) = sigm\left(\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0)\right)}$$

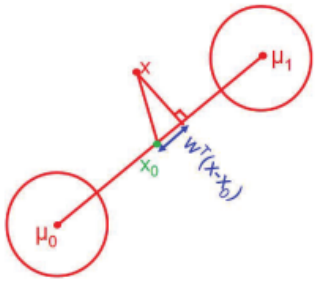

FIGURE 19.2. – Geometry of LDA in the 2 class case where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \boldsymbol{I}$

MLE for discriminant analysis The *log-likelihood* function is as follows :

$$\log\left(p(\mathcal{D}|\theta)\right) = \sum_{i=1}^{N}\sum_{c=1}^{C} \mathbb{1}(y_i = c)\log(\pi_c) + \sum_{c=1}^{C}\left[\sum_{i:y_i=c} \log\left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right)\right]$$

Strategies for preventing the overfitting Even if *speed* and *simplicity* are the attractive aspects of MLE, it can badly overfit in high dimensions.
In particular the MLE for a full covariance matrix is singular if $N_c < D$, and even when $N_c > D$ the MLE can be ill-conditioned, meaning close to singular.
Some solution to this problem :

— use a diagonal covariance matrix for each class, this assumes the features are conditionally independent, this is equivalent to using a naive Bayes classifier.

— use the full covariance matrix but force it to be the same for all classes, $\Sigma_c = \Sigma$, this is equivalent to *parameter sharing* and is equivalent to LDA.

— Use a diagonal matrix and force it to be shared.

— use a diagonal covariance, but impose a prior and then integrate it out. If we use a conjugate prior this is analogous to the Bayesian naive Bayes.

— fit a full or a diagonal covariance matrix by MAP estimation.

— Project the data into a low dimensional subspace and fit the Gaussian there.

Regularized LDA* Assume we tie the covariance matrices, so $\boldsymbol{\Sigma_c} = \boldsymbol{\Sigma}$ as in LDA and furthermore we perform MAP estimation of $\boldsymbol{\Sigma}$ using an inverse Wishart prior of the form $IW(diag(\hat{\boldsymbol{\Sigma}}_{mle}), \nu_0)$

$$\hat{\boldsymbol{\Sigma}} = \lambda diag(\hat{\boldsymbol{\Sigma}}_{mle}) = (1-\lambda)\hat{\boldsymbol{\Sigma}}_{mle}$$

where λ controls the amount of regularization which is related to the strength of the prior ν_0

Nearest shruken centroids classifier A disadvantage of *diagonal LDA* is it depends on all of the features.
An idea can be to perform MAP estimation for the diagonal LDA with a sparsity-promoting (Laplace).
More precisely define

— μ_{cj} class-specific feature mean

— m_j class-independent feature mean

— Δ_{cj} class-specific offset

Then $\mu_{cj} = m_j + \Delta_{cj}$ In putting a prior on the Δ_{cj} terms to encourage them to be strictly zero and compute a MAP estimate. If, for feature j we find that $\Delta_{cj} = 0$ for all c, then feature j will play no role in teh classification decision. Thus features that are not discriminative are automatically ignored.

Inference in jointly Gaussian distribution Suppose $\boldsymbol{x} = (x_1, x_2)$, then $\boldsymbol{\mu} = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix}\Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22}\end{pmatrix}$

Thus the marginals are given by : $\begin{cases} p(x_1) = \mathcal{N}(x_1|\mu_1, \Sigma_{11}) \\ p(x_2) = \mathcal{N}(x_2|\mu_2, \Sigma_{22}) \end{cases}$

and the posterior conditional is given by :

$$\begin{aligned} p(x_1|x_2) &= \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2}) \\ \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(x_2 - \mu_2) \\ &= \Sigma_{1|2}\left(\Lambda_{11}\mu_1 - \Lambda_{12}(x_2 - \mu_2)\right) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Lambda_{11}^{-1} \end{aligned}$$

The conditional mean is just a linear function of x_2 and the conditional covariance is just a constant matrix that is independent of x_2

19.3. Wishart distribution

Definition It is a generalization of the Gamma distribution to positive definite matrices. The Wishart distribution is for some, ranked next to the multivariate normal distribution in order of importance and usefulness in multivariate statistics. The probability density function is defined as follows :

$$Wi(\Lambda|S, \nu) = \frac{1}{Z_{Wi}}|\Lambda|^{\frac{\nu-D-1}{2}} exp\left(-\frac{1}{2}tr(\Lambda S^{-1})\right)$$

Here ν and S correspond respectively to degrees of freedom and scale matrix. The normalization constant fro this distribution is :

$$Z_{Wi} = 2^{\frac{\nu D}{2}}\Gamma_D\left(\frac{\nu}{2}\right)|S|^{\frac{\nu}{2}}$$

where $\Gamma_D : x \mapsto \pi^{\frac{D(D-1)}{4}}\prod_{i=1}^{D}\Gamma\left(x + \frac{1-i}{2}\right)$

Inverse Wishart distribution

$$IW(\boldsymbol{\Sigma}|\boldsymbol{S}, \nu) = \frac{1}{Z_{IW}}|\boldsymbol{\Sigma}|^{-\frac{\nu+D+1}{2}}\exp\left(\frac{1}{2}tr(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}^{-1})\right)$$

$$Z_{IW} = |\boldsymbol{S}|^{-\frac{\nu}{2}}2^{-\frac{D\nu}{2}}\boldsymbol{\Gamma}_D\left(\frac{\nu}{2}\right)$$

One can show that the distribution has these properties $\begin{cases} mean = \dfrac{S^{-1}}{\nu - D - 1} \\ mode = \dfrac{S^{-1}}{\nu + D + 1} \end{cases}$

19.4. Linear Gaussian systems

Condition Let $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{D_x} \times \mathbb{R}^{D_y}$ respectively a hidden variable and a noisy observation of the hidden variable.

Let assume we have the following prior and likelihood : $\begin{cases} p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_x\right) \\ p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_y\right) \end{cases}$

where $\boldsymbol{A} \in \mathcal{M}(\mathbb{R})_{D_x, D_y}$.

This schematically means $\boldsymbol{x}$ generates $\boldsymbol{y}$, in this section we show how to invert the arrow that is how to infer $\boldsymbol{x}$ from $\boldsymbol{y}$.

Bayes rule for linear Gaussian systems Given a linear Gaussian system as the previous one is given by the following :

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \mathcal{N}\left(\boldsymbol{x}|\mu_{x|y}, \boldsymbol{\Sigma}_{x|y}\right) \\ \boldsymbol{\Sigma}_{x|y}^{-1} &= \boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^T \boldsymbol{\Sigma}_y^{-1} \boldsymbol{A} \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}\left(\boldsymbol{A}^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x\right) \end{aligned}$$

19.5. Inferring the parameters of an MVN

Posterior distribution of μ For simplicity purpose we will use the conjugate prior, which in this case is Gaussian. If $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}_0, \boldsymbol{V}_0)$ then we can derive a Gaussian posterior for μ based on the results in

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{D}, \boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}_N, \boldsymbol{V}_N) \\ \boldsymbol{V}_N^{-1} &= \boldsymbol{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{m}_N &= \boldsymbol{V}_N\left(\boldsymbol{\Sigma}^{-1} N\hat{\boldsymbol{x}} + \boldsymbol{V}_0^{-1}\boldsymbol{m}_0\right) \end{aligned}$$

Posterior distribution of $\boldsymbol{\Sigma}$ Aim : compute $p(\boldsymbol{\Sigma}|\mathcal{D}, \boldsymbol{\mu})$

The likelihood has the form

$$p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} exp\left(-\frac{1}{2} tr(\boldsymbol{S}_{\boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1})\right)$$

The corresponding conjugate prior is known as the inverse Wishart distribution. Then

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left(-\frac{1}{2} tr(\boldsymbol{S}_{\boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1})\right) |\boldsymbol{\Sigma}|^{-\frac{\nu_0 + D + 1}{2}} \exp\left(-\frac{1}{2} tr(\boldsymbol{S}_0 \boldsymbol{\Sigma}^{-1})\right) \\ &= |\boldsymbol{\Sigma}|^{\frac{n + \nu_0 + D + 1}{2}} \exp\left(-\frac{1}{2} tr\left[(\boldsymbol{S}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_0)\boldsymbol{\Sigma}^{-1}\right]\right) \\ &= IW(\boldsymbol{\Sigma}|\boldsymbol{S}_N, \nu_N) \end{aligned}$$

Then $\begin{cases} \nu_n = \nu_0 + n \\ \boldsymbol{S}_n^{-1} = \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\mu}} \end{cases}$

MAP estimation To progress in case where $\hat{\boldsymbol{\Sigma}}_{mle}$ is non-invertible or ill-conditioned we can use the posterior mode (or mean). One can show that the MAP estimate is given by :

$$\hat{\boldsymbol{\Sigma}}_{map} = \frac{\boldsymbol{S}_n}{\nu_N + D + 1} = \frac{\boldsymbol{S}_0 + \boldsymbol{S}_\mu}{n_0 + n}$$

Consider now the use of a proper informative prior, which is necessary whenever D/N is large (> 0.1). We can rewrite the MAP estimate as a convex combination of the prior mode and the MLE.

$$\hat{\boldsymbol{\Sigma}}_{MAP} = \frac{\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\mu}}}{n_0 + n} = \frac{n_0}{n_0 + n}\frac{\boldsymbol{S}_0}{n_0} + \frac{n}{n_0 + n}\frac{\boldsymbol{S}}{n} = \lambda \boldsymbol{\Sigma}_0 + (1 - \lambda)\hat{\boldsymbol{\Sigma}}_{mle}$$

20. Test of Statistical Hypotheses

20.1. Introduction

Statistical hypothesis is a conjecture about the distribution $f(x;\theta)$ of a population X.
This conjecture is usually about the parameter θ.

Simple and Composite hypotheses H completely specifies the density $f(x;\theta)$ of the population $\Rightarrow$ H is *simple hypothesis*
H does not completely specify the density $f(x;\theta)$ of the population $\Rightarrow$ H is *composite hypothesis*

Null and alternative hypothesis Null hypothesis H_0 is to be tested, and correspond to the idea that an observed difference is due to chance.
Alternative hypothesis $H_a = \overline{H_0}$, corresponds to the idea that the observed difference is real.

Hypothesis test It is used to measure the difference between the data and what is expected on the null hypothesis.
Let $X \hookrightarrow f(x;\theta)$ and $(X_i)_{1\leq i\leq n}$ a sample from X and C a Borel set in $\mathbb{R}^n$ It is an ordered sequence : $\left((X_i)_{1\leq i\leq n}; H_0, H_\alpha, C\right)$
The set C is called the *critical region* in the hypothesis test. The critical region is obtained using a *test statistics* $W\left((X_i)_{1\leq i\leq n}\right)$. If the outcome of $(X_i)_{1\leq i\leq n}$ turns out to be an element of C then we decide to accept H_α otherwise we accept H_0

20.2. Tests

Z-test It says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis.

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

T-test

$$t = \frac{\text{average of draws} - \text{c}}{\text{SE}}$$

c corresponding to the constant which is fixed in H_0
As the number degrees of freedom increases the curves get closer and closer to the normal.

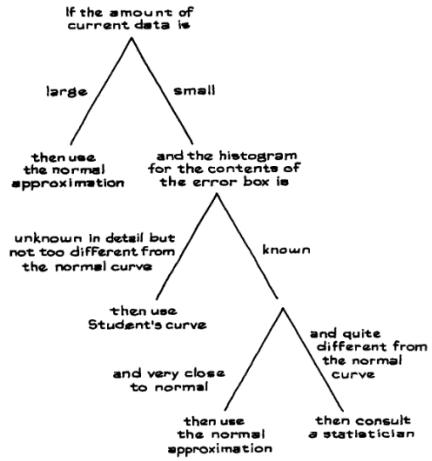

FIGURE 20.1. – Procedure when testing

Test Method

20.3. A method of Finding Tests

The likelihood ratio test statistics For testing the simple null hypothesis $H_0 : \theta \in \Sigma_0$ against the composite alternative hypothesis $H_\alpha \notin \Sigma_0$ based on a set of random sample data $(x_i)_{1\leq i\leq n}$ is defined as $W\left((x_i)_{1\leq i\leq n}\right) = \frac{\max_{\theta\in\Omega_0} L\left(\theta;(x_i)_{1\leq i\leq n}\right)}{\max_{\theta\in\Omega} L\left(\theta;(x_i)_{1\leq i\leq n}\right)}$
Let $k \in [0,1]$ a likelihood ratio test is any test that has a critical region C that is rejection region) of the form : $C = \left\{(x_i)_{1\leq i\leq n} | W\left((x_i)_{1\leq i\leq n}\right) \leq k\right\}$

20.4. Methods of Evaluating Tests

	H_0 is true	H_0 is false
Accept H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

Significance level It is denoted by $\alpha = \mathbb{P}(\{\text{Type I Error}\})$ corresponds to the probability of getting a test statistic as extreme as, or more extreme than the observed

one. This probability is computed on the basis that the H_0 is true.

P-value Is the probability of getting a big test statistic, assuming the null hypothesis to be right.

Probability of type II error It is denoted by $\beta = \mathbb{P}_{\{H_0 \text{ is false}\}}(\{\text{Accept } H_0\}) = \mathbb{P}_{\{H_\alpha \text{ is true}\}}(\{\text{Accept } H_0\})$

The power function It is the function $\pi : \Omega \to [0,1]$ defined by : $\pi(\theta) = \begin{cases} \mathbb{P}(\{\text{Type I Error}\}) \\ 1 - \mathbb{P}(\{\text{Type II Error}\}) \end{cases}$

A test of level δ Given $\delta \in [0,1]$ $\max_{\theta \in \Omega_0} \pi(\theta) \leq \delta$

Test of size δ Given $\delta \in [0,1]$ $\max_{\theta \in \Omega_0} \pi(\theta) = \delta$

Uniformly most powerful Let T be a test procedure for testing the null hypothesis. For any test W of level δ,
$\forall \theta \in \Omega_0, \pi_T(\theta) \geq \pi_W(\theta)$

Theorem Neyman-Perarson $\begin{cases} X \hookrightarrow f(x;\theta) \\ (X_i)_{1 \leq i \leq n} \text{ sample from } X \\ L\left(\theta; (x_i)_{1 \leq i \leq n} = \prod_{i=1}^{n} f(x_i;\theta)\right) \text{ likelihood function of the sample} \end{cases}$
$\Rightarrow C = \left\{ (x_i)_{1 \leq i \leq n} \mid \frac{L\left(\theta_0, (x_i)_{1 \leq i \leq n}\right)}{L\left(\theta_\alpha, (x_i)_{1 \leq i \leq n}\right)} \right\} \leq k$ for $k \in \mathbb{R}_+^*$ is best of its size for testing :
$H_0 : \theta = \theta_0$ against $H_a : \theta = \theta_a$

20.5. Some Examples of Likelihood Ratio Tests

—

21. Simple Linear Regression and Correlation Analysis

21.1. Least Squared Method

Unbiased estimators

1. $\mathbb{E}(Y_x) = a + \beta x$ so that $\mu_i = \mathbb{E}(Y_i) = \alpha + \beta x_i$
2. $(Y_i)_{1 \leq i \leq n}$ are independent ;
3. $(Y_i)_{1 \leq i \leq n}$ has the same variance σ^2

$\mathbb{E}(Y/X) = \alpha + \beta x$ are unbiased.

21.2. Normal Regression Analysis

Likelihood estimators In the normal regression analysis, the likelihood estimators $\hat{\beta}$ and $\hat{\alpha}$ are unbiased estimators of β and α respectively.

Distributions of the estimators In the normal regression analysis, the distributions of the estimators $\hat{\beta}$ and $\hat{\alpha}$ are given by :

$$\begin{cases} \hat{\beta} \hookrightarrow \mathbb{N}\left(\beta, \dfrac{\sigma^2}{S_{xx}}\right) \\ \hat{\alpha} \hookrightarrow \mathbb{N}\left(\alpha, \dfrac{\sigma^2}{n} + \dfrac{\overline{x}^2 \sigma^2}{S_{xx}}\right) \end{cases}$$

21.3. The Correlation Analysis

Correlation coefficient $\left((X_i, Y_i)_{1 \leq i \leq n}\right)$ from $(X, Y) \Rightarrow$ Sample correlation coefficient is defined as

$$R = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}^2\right)} \sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}^2\right)}}$$

And $-1 \leq R \leq 1$

22. Analysis of Variance

22.1. One-way Analysis of Variance with Equal Sample Sizes

Theorem ANOVA $\begin{cases} \text{Model is given by } \forall (i,j) \in [\![1,m]\!] \times [\![1,n]\!] \, Y_{ij} = \mu_i + \epsilon_{ij} \\ (\epsilon_{ij})_{(i,j)\in[\![1,m]\!]\times[\![1,n]\!]} \text{ are independent and normally distributed} \end{cases}$

$$\Rightarrow \begin{cases} \hat{\mu_i} = \overline{Y}_{i\bullet} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} Y_{ij} \\ \hat{\sigma^2} = \frac{1}{nm} SS_W = \frac{1}{mn}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \end{cases}$$

Lemma

$$\begin{cases} \overline{Y}_{\bullet\bullet} = \dfrac{1}{nm}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij} \\ SS_T = \dfrac{1}{nm}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \\ SS_W = \dfrac{1}{nm}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right) \\ SS_B = \dfrac{1}{nm}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n} \left(Y_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \end{cases} \Rightarrow SS_T = SS_W + SS_B$$

ANOVA model Consider ANOVA model : $\begin{cases} \forall (i,j) \in [\![1,m]\!] \times [\![1,n]\!] \, Y_{ij} = \mu_i + \epsilon_{ij} \\ Y_{ij} \hookrightarrow \mathbb{N}(\mu_i, \sigma^2) \end{cases} \Rightarrow$

$$\begin{cases} \text{(a) the random variable } \frac{SS_W}{\sigma^2} \hookrightarrow \chi^2\left(m(n-1)\right) \\ \text{(a) the statics } SS_W \text{ and } SS_B \text{ are independent} \\ H_0 : \forall (i,j) \in [\![1,m]\!]^2 \, \mu_i = \mu_j \text{ it is True} \end{cases} \Rightarrow \begin{cases} \text{the random variable } \frac{SS_B}{\sigma^2} \hookrightarrow \chi^2(m-1) \\ \text{the statics } \frac{SS_B m(n-1)}{\sigma^2(m-1)} \hookrightarrow F(m-1, m(n-1)) \\ \text{the random variable } \frac{SS_T}{\sigma^2} \hookrightarrow \chi^2(nm-1) \end{cases}$$

22.2. One-way Analysis of Variance with Unequal Sample Sizes

—

22.3. Pair Wise Comparisons

—

22.4. Test for Homogeneity of Variances

—

23. Goodness of Fits Tests

23.1. Chi-Squared test

—

23.2. Kolmogorov-Smirnov test

—