ENGG2430A Probability and Statistics for Engineers

Chapter 9: Classical Statistical Inference

Instructor: Shengyu Zhang

Preceding chapter: Bayesian inference

- Preceding chapter: Bayesian approach to inference.
  - Unknown parameters are modeled as random variables.
  - Work within a single, fully-specified probabilistic model.
  - Compute posterior distribution by judicious application of Bayes' rule.

This chapter: classical inference

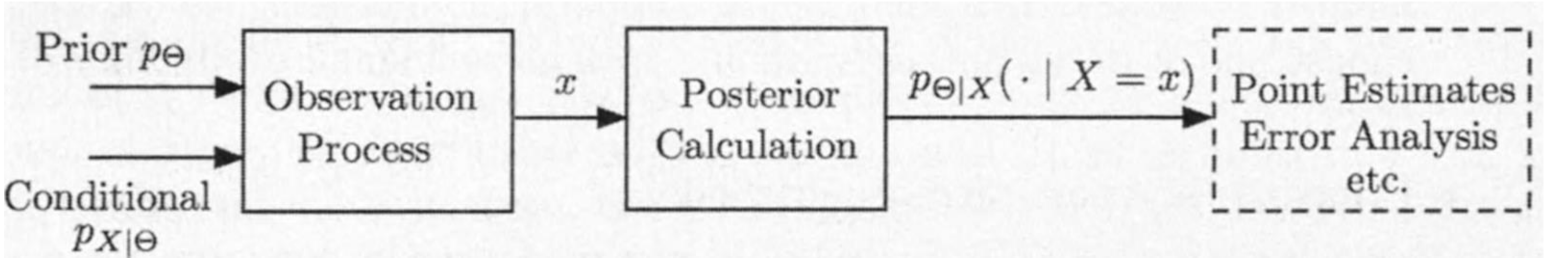
- We view the unknown parameter θ as a deterministic (not random!) but unknown quantity.
- The observation X is random and its distribution
  - $p_X(x;\theta)$ if X is discrete
  - $f_X(x;\theta)$ if X is continuous

depends on the value of θ.

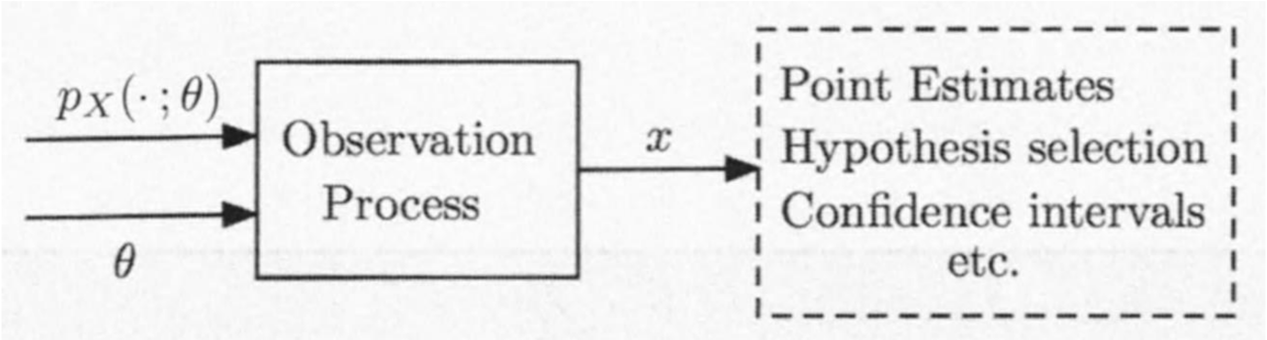
Classical inference

- Deal simultaneously with multiple candidate models, one model for each possible value of θ.
- A "good" hypothesis testing or estimation procedure will be one that possesses certain desirable properties under every candidate model.
  - i.e. for every possible value of θ.

- Bayesian:



- Classical:



Notation

- Our notation will generally indicate the dependence of probabilities and expected values on θ.
- For example, we will denote by $E_\theta[h(X)]$ the expected value of a random variable $h(X)$ as a function of θ.
- Similarly, we will use the notation $P_\theta(A)$ to denote the probability of an event A.

Content

- Classical Parameter Estimation
- Linear Regression
- Binary Hypothesis Testing
- Significance Testing

- Given observations $X = (X_1, \ldots, X_n)$, an estimator is a random variable of the form $\widehat{\Theta} = g(X)$, for some function g.
- Note that since the distribution of X depends on θ, the same is true for the distribution of $\widehat{\Theta}$.
- We use the term estimate to refer to an actual realized value of $\widehat{\Theta}$.

- Sometimes, particularly when we are interested in the role of the number of observations n, we use the notation $\widehat{\Theta}_n$ for an estimator.
- It is then also appropriate to view $\widehat{\Theta}_n$ as a sequence of estimators.
  - One for each value of n.
- The mean and variance of $\widehat{\Theta}_n$ are denoted $E_\theta\left[\widehat{\Theta}_n\right]$ and $var_\theta\left[\widehat{\Theta}_n\right]$, respectively.
  - We sometimes drop this subscript θ when the context is clear.

Terminology regarding estimators

- Estimator: $\widehat{\Theta}_n$, a function of n observations for an $(X_1, \ldots, X_n)$ whose distribution depends on θ.
- Estimation error: $\widetilde{\Theta}_n = \widehat{\Theta}_n - \theta$.
- Bias of the estimator: $b_\theta(\widehat{\Theta}_n) = E_\theta[\widehat{\Theta}_n] - \theta$, is the expected value of the estimation error.

bias

- $\widehat{\Theta}_n$ is unbiased if $b_\theta\left(\widehat{\Theta}_n\right) = 0$.
  - a desirable property.
- $\widehat{\Theta}_n$ is asymptotically unbiased if $\lim_{n\to\infty} E_\theta\left[\widehat{\Theta}_n\right] = \theta$, for every possible value of θ.
  - $\widehat{\Theta}_n$ becomes unbiased as the number n of observations increases,
  - this is desirable when n is large.

Consistent

- $\widehat{\Theta}_n$ is consistent if the sequence $\widehat{\Theta}_n$ converges to the true value θ, in probability, for every possible value of θ.
- Recall:
  - X_n converges to a in probability if
    $\forall \epsilon > 0, \mathrm{P}(|X_n - a| \geq \epsilon) \to 0,$ as $n \to \infty$.
  - X_n converges to a with probability 1 (or almost surely) if
$$\mathrm{P}\left(\lim_{n\to\infty} X_n = a\right) = 1$$

- Mean squared error: $E_\theta\left[\tilde{\Theta}_n^2\right]$.
- This is related to the bias and the variance of $\hat{\Theta}_n$: $E_\theta\left[\tilde{\Theta}_n^2\right] = b_\theta^2\left(\hat{\Theta}_n\right) + var_\theta\left[\hat{\Theta}_n\right]$.
  - Reason: $E[X^2] = (E[X])^2 + var(X)$, $X = \tilde{\Theta}_n = \hat{\Theta}_n - \theta$.
- In many statistical problems, there is a tradeoff between the two terms on the right-hand-side.
- Often a reduction in the variance is accompanied by an increase in the bias.
- Of course, a good estimator is one that manages to keep both terms small.

Maximum Likelihood Estimation (MLE)

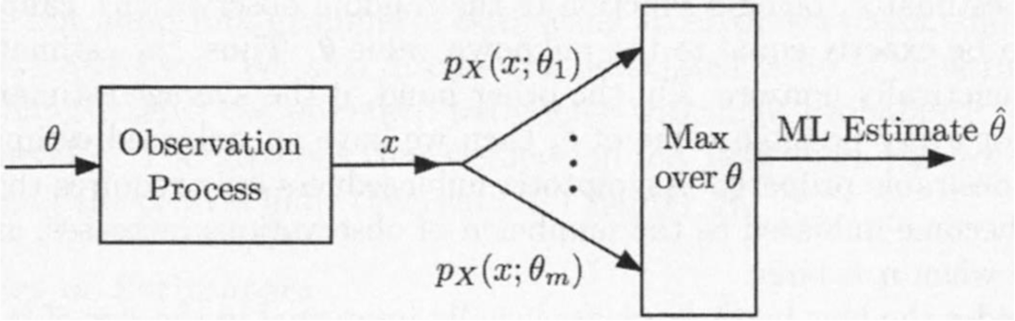
- Let the vector of observations $X = (X_1, \dots, X_n)$ be described by a joint PMF $p_X(x;\theta)$
  - Note that $p_X(x;\theta)$ is PMF for X only, not joint distribution for X and θ.
    - Recall θ is just a fixed parameter, not a random variable.
    - $p_X(x;\theta)$ depends on θ.
- Suppose we observe a particular value $x = (x_1, \dots, x_n)$ of X.

- A maximum likelihood estimate (MLE) is a value of the parameter that maximizes the numerical function $p_X(x_1, \ldots, x_n; \theta)$ over all θ.

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\, p_X(x_1, \ldots, x_n; \theta)$$

- The above is for the case of discrete X. If X is continuous, then MLE is

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\, f_X(x_1, \ldots, x_n; \theta)$$



- In many applications, the observations X_i are assumed to be independent.
- Then $p_X(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p_{X_i}(x_i; \theta)$.
- It is often analytically or computationally convenient to maximize its logarithm, called the log-likelihood function (over θ)

$$\log p_X(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta)$$

- The term "likelihood" needs to be interpreted properly.
- Having observed the value x of X, $p_X(x, \theta)$ is not the probability that the unknown parameter is equal to θ.
- It is the *probability that the observed value x can arise when the parameter is equal to θ*.

- Thus, in maximizing the likelihood, we are asking the following question:

- *"What is the value of θ under which the observations we have seen are most likely to arise?"*

Comparison with Bayesian MAP

- Recall MAP: $\max_\theta p_\Theta(\theta)\, p_{X|\Theta}(x|\theta)$.
- Thus we can interpret MLE as MAP estimation with a flat prior.
  - i.e., a prior which is the same for all θ,
  - indicating the absence of any useful prior knowledge.
- In the case of continuous θ with a bounded range, MLE is MAP with a uniform prior: $f_\Theta(\theta) = c$ for all θ and some constant c.

Estimating parameter of exponential

- Customers arrive to a facility, with the ith customer arriving at time Y_i.
- We assume that the ith interarrival time,
$$X_i = Y_i - Y_{i-1}$$
is exponentially distributed with parameter θ,
  - with the convention $Y_0 = 0$
- Assume that $X_1, \dots, X_n$ are independent.
- We wish to estimate the value of θ (interpreted as the arrival rate), on the basis of the observations $X_1, \dots, X_n$.

- The corresponding likelihood function is

$$f_X(x;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

- Thus the log-likelihood function is

$$\log f_X(x;\theta) = \sum_i \log\left(\theta e^{-\theta x_i}\right) = n \log \theta - \theta y_n$$

where $y_n = \sum_{i=1}^{n} x_i$.

- Setting the derivative (wrt θ) to be 0:

$$(n/\theta) - y_n = 0$$

- We get $\hat{\theta} = n/y_n$.
- That is, $\widehat{\Theta}_n = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^{-1}$
- It is the inverse of the sample mean of the interarrival times.
- Can be interpreted as an empirical arrival rate.

Estimating parameters of normal

- Estimating the mean μ and variance σ of a normal distribution using n independent observations $X_1, \ldots, X_n$.
- Simple calculation yields that the log likelihood function is

$$\log f_X(x; \mu, \sigma) = -\frac{n}{2}\left(\log(2\pi\sigma) + \frac{s_n^2}{\sigma} + \frac{(m_n - \mu)^2}{\sigma}\right)$$

- $\log f_X(x;\mu,\sigma) = -\frac{n}{2}\left(\log(2\pi\sigma) + \frac{s_n^2}{\sigma} + \frac{(m_n-\mu)^2}{\sigma}\right)$
- Here m_n and s_n^2 are the realized values of the random variables

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$$

  - The sample mean and sample variance, resp.
- The maximizer is $\hat{\theta} = \left(m_n, s_n^2\right)$.
- "*The MLE of normal is just sample mean and sample variance.*"

Properties of MLE

- Invariance principle: if $\widehat{\Theta}_n$ is the ML estimate of θ, then for any one-to-one function h of θ, the MLE of the parameter $\xi = h(\theta)$ is $h(\widehat{\Theta}_n)$.
- Consistency: MLE is consistent for i.i.d. observations
  - under some mild assumptions,
- Asymptotic normality property: When θ is a scalar, the distribution of $(\widehat{\Theta}_n - \theta)/\sigma(\widehat{\Theta}_n)$ approaches $N(0,1)$.
  - under some mild conditions

Estimation of the Mean

- Suppose that the observations $X_1, \ldots, X_n$ are i.i.d., with an unknown common mean μ and common variance σ^2.
- The sample mean $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased.
- Its mean squared error is
  $E[(M_n - \mu)^2] = \frac{1}{n^2} E[(\sum_{i=1}^{n} (X_i - \mu))^2]$
  $= \frac{1}{n^2}\sum_{i=1}^{n} E[(X_i - \mu)^2] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$.
  - Doesn't depend on μ.

Estimation of the Variance

- Consider the sample variance

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$$

- Let's compute its bias.
- $E\left[X_i^2\right] = \mu^2 + \sigma^2,\ E\left[M_n^2\right] = \mu^2 + \frac{\sigma^2}{n}.$
- $$\begin{aligned} E\left[\bar{S}_n^2\right] &= (1/n)E\left[\sum_{i=1}^{n} X_i^2 - 2M_n \sum_{i=1}^{n} X_i^2 + nM_n^2\right] \\ &= E\left[(1/n)\sum_{i=1}^{n} X_i^2 - 2M_n^2 + M_n^2\right] \\ &= E\left[(1/n)\sum_{i=1}^{n} X_i^2 - M_n^2\right] \\ &= \mu^2 + \sigma^2 - \left(\mu^2 + \sigma^2/n\right) \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

- Last slide: $E\left[\bar{S}_n^2\right] = \frac{n-1}{n}\sigma^2$
- The sample variance $\bar{S}_n^2$ is not an unbiased estimator of σ^2, although it is asymptotically unbiased.
- Define $\hat{S}_n^2 = \frac{n}{n-1}\bar{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$, then $\hat{S}_n^2$ is unbiased.
  - For large n, however, $\hat{S}_n^2$ and $\bar{S}_n^2$ are almost the same.

Confidence Intervals

- Consider an estimator $\widehat{\Theta}_n$ of an unknown parameter θ.
- Besides the numerical value provided by an estimate, we are often interested in constructing a so-called confidence interval.
- Roughly speaking, this is an interval that contains θ with a certain high probability, for every possible value of θ.

- Let us first fix a desired confidence level, $1-\alpha$, where α is typically a small number.
- We then replace the point estimator $\widehat{\Theta}_n$ by a lower estimator $\widehat{\Theta}_n^-$ and an upper estimator $\widehat{\Theta}_n^+$, s.t.

$$P\left(\widehat{\Theta}_n^- \leq \theta \leq \widehat{\Theta}_n^+\right) \geq 1-\alpha$$

  for every possible value of θ.
- We call $\left[\widehat{\Theta}_n^-, \widehat{\Theta}_n^+\right]$ a $(1-\alpha)$ confidence interval.

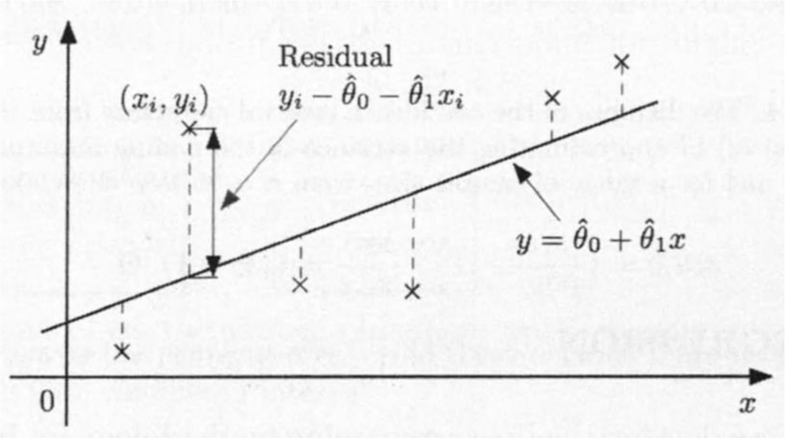
Content

- Classical Parameter Estimation
- Linear Regression
- Binary Hypothesis Testing
- Significance Testing

- We consider the case of only two variables for illustration.
- We wish to model the relation between two variables of interest, x and y
  - e.g., years of education and income.
- based on a collection of data pairs (x_i, y_i), $i = 1, \ldots, n$.
  - e.g. x_i is the years of education, and y_i the annual income

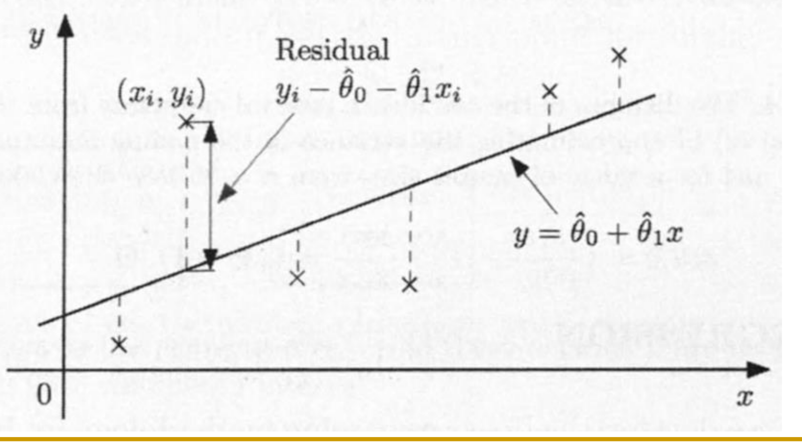
- Often a two-dimensional plot of these samples indicates a systematic, approximately linear relation between x_i and y_i.
- Then, it is natural to attempt to build a linear model of the form $y \approx \theta_0 + \theta_1 x$.
  - θ_0 and θ_1 are unknown parameters to be estimated.
- Given some estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the resulting parameters, the value y_i corresponding to x_i, as predicted by the model, is $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$.

- Generally, $\hat{y}_i$ will be different from the given value y_i, and the corresponding difference $\tilde{y}_i = \hat{y}_i - y_i$ is called the ith residual.
- A choice of estimates that results in small residuals is considered to provide a good fit to the data.



- The linear regression approach chooses the parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ that minimize the sum of the squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2$$



- Note that the postulated linear model may or may not be true.
- The true relation between the two variables may be nonlinear.
- In practice, there is often an additional phase where we examine whether the hypothesis of a linear model is supported by the data and try to validate the estimated model.

- Given n data pairs (x_i, y_i), the estimates that minimize the sum of the squared residuals are given by

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

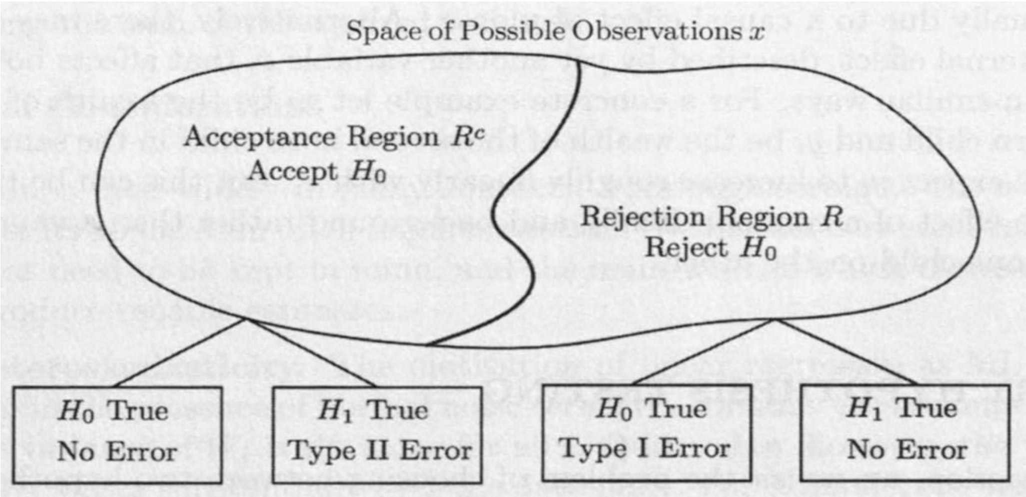
Content

- Classical Parameter Estimation
- Linear Regression
- Binary Hypothesis Testing
- Significance Testing

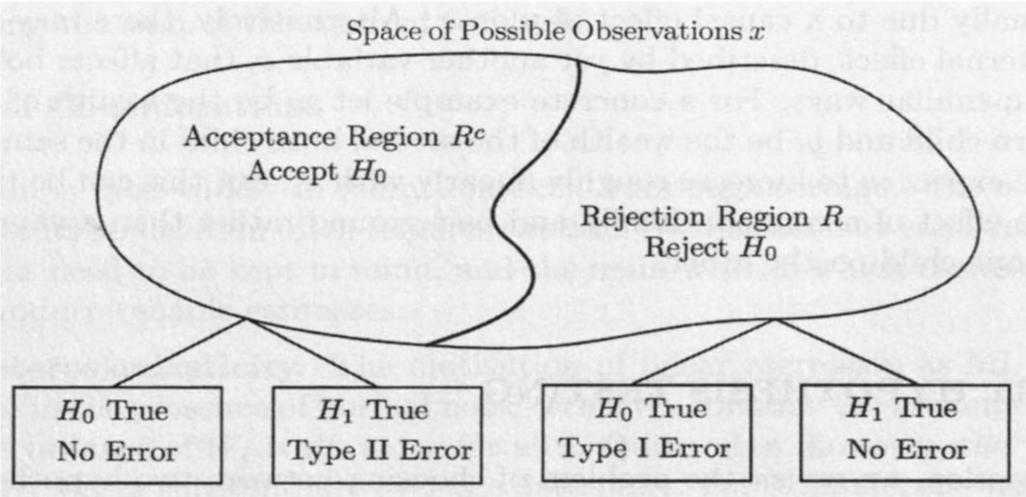
- We revisit the problem of choosing between two hypotheses.
- But unlike the Bayesian formulation, we will assume no prior probabilities.
- Two hypotheses: H_0 and H_1.
- In traditional statistical language, hypothesis H_0 is often called the null hypothesis and H_1 the alternative hypothesis.
  - H_0 plays the role of a default model, to be proved or disproved on the basis of available data.

- The available observation is a vector $X = (X_1, \dots, X_n)$ of random variables whose distribution depends on the hypothesis.
- Note that consistent with the classical inference framework, these are not conditional probabilities, because the true hypothesis is not treated as a random variable.

- Notation: $P(X \in A; H_j)$ is the probability that the observation X belongs to a set A when hypothesis H_j is true.
- $p_X(x; H_j)$ or $f_X(x; H_j)$ to denote the PMF or PDF, respectively, of the vector X, under hypothesis H_j.



- Any decision rule can be represented by a partition of the set of all possible $X = (X_1, \ldots, X_n)$ into two subsets.
  - the rejection region R,
  - the acceptance region R^c.
- The choice of a decision rule is equivalent to choosing the rejection region.



- For a particular choice of the rejection region R, there are two possible types of errors.
- Type I error, or a false rejection: Reject H_0 even though H_0 is true.
  - This happens with probability $\alpha(R) = P(X \in R; H_0)$.
- Type II error, or a false acceptance: Accept H_0 even though H_0 is false.
  - This happens with probability $\beta(R) = P(X \notin R; H_1)$.

- To motivate a particular form of rejection region, we draw an analogy with Bayesian hypothesis testing.
- Two hypotheses $\Theta = \theta_0$ and $\Theta = \theta_1$ are involved, with respective prior probabilities $p_\Theta(\theta_0)$ and $p_\Theta(\theta_1)$.
- The overall probability of error is minimized by using the MAP rule.

- Given the observed value x of X, declare $\Theta = \theta_1$ be true if

$$p_\Theta(\theta_0)p_{X|\Theta}(x|\theta_0) < p_\Theta(\theta_1)p_{X|\Theta}(x|\theta_1)$$

- This decision rule can be rewritten as follows.
- Define the likelihood ratio $L(x)$ by

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)}$$

- Declare $\Theta = \theta_1$ to be true if the realized value x of the observation vector X satisfies $L(x) \geq \xi$.

- Here ξ is the critical value defined by
$$\xi = \frac{p_\Theta(\theta_0)}{p_\Theta(\theta_1)}$$
- If X is continuous, the approach is the same, except that the likelihood ratio is defined as a ratio of PDFs: $L(x) = \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_0)}$.

- Motivated by the preceding form of the MAP rule, we are led to consider rejection regions of the form

$$R = \{x | L(x) > \xi\},$$

where the likelihood ratio $L(x)$ is denned similar to the Bayesian case:

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)}, \quad \text{or} \quad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)}.$$

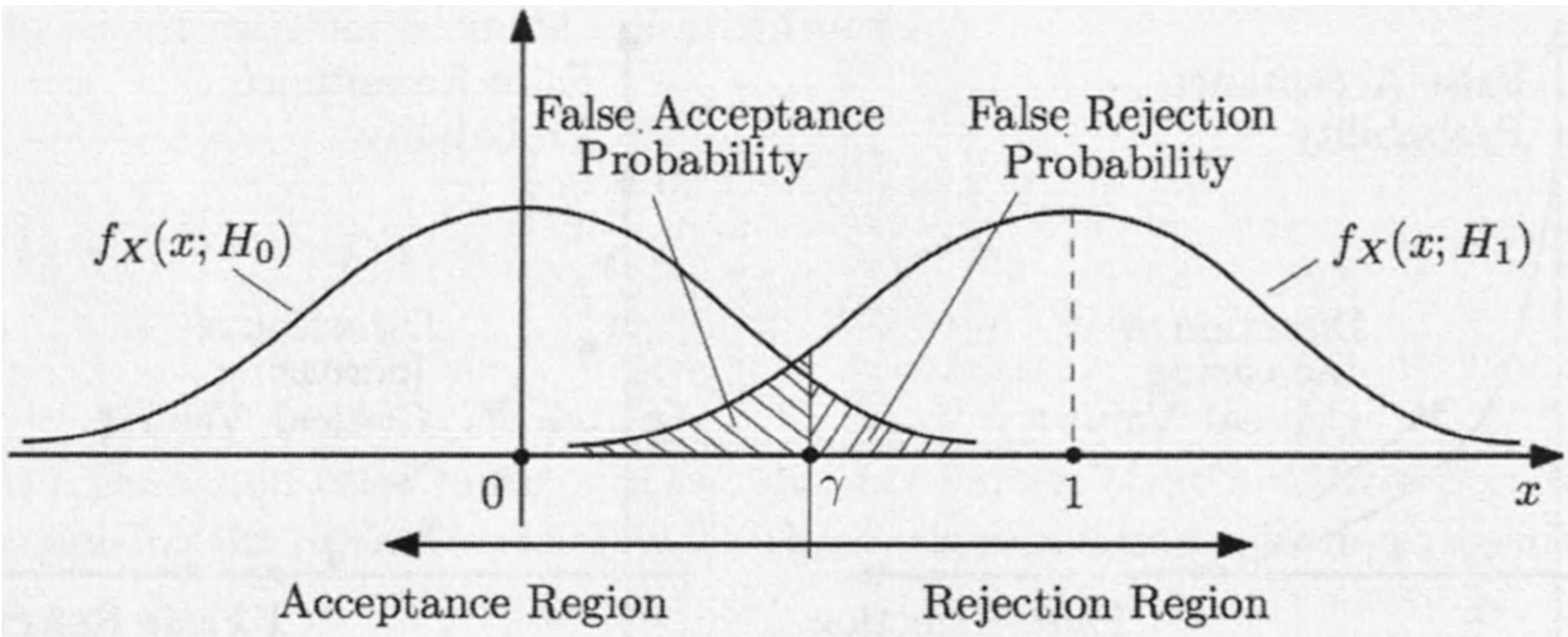
- The critical value ξ remains free to be chosen on the basis of other considerations.

Likelihood Ratio Test (LRT)

- Start with a target value α for the false rejection probability.
- Choose a value for ξ such that the false rejection probability is equal to α:
$$P(L(X) > \xi; H_0) = \alpha$$
- Once the value x of X is observed, reject H_0 if $L(x) > \xi$.
- Typical choices for α are $\alpha = 0.1$, $\alpha = 0.05$, or $\alpha = 0.01$, depending on the degree of undesirability of false rejection.

- Note that to be able to apply the LRT to a given problem, the following are required.
- We must be able to compute $L(x)$ for any given observation value x, so that we can compare it with the critical value ξ.
  - Fortunately, this is the case when the underlying PMFs or PDFs are given in closed form.

- We must either have a closed form expression for the distribution of $L(X)$
  - or of a related random variable such as $\log L(X)$
  - or we must be able to approximate it analytically, computationally, or through simulation.
- This is needed to determine the critical value ξ that corresponds to a given false rejection probability α.



False Acceptance
Probability
False Rejection
Probability
$f_X(x; H_0)$
$f_X(x; H_1)$
0
γ
1
x
Acceptance Region
Rejection Region

- When $L(X)$ is a continuous random variable, the probability $P(L(X) > \xi; H_0)$ moves continuously from 1 to 0 as ξ increases.
- Thus, we can find a value of ξ for which the requirement $P(L(X) > \xi; H_0) = \alpha$ is satisfied.
- If, however, $L(X)$ is a discrete random variable, it may be impossible to satisfy the equality $P(L(X) > \xi; H_0) = \alpha$ exactly, no matter how ξ is chosen.

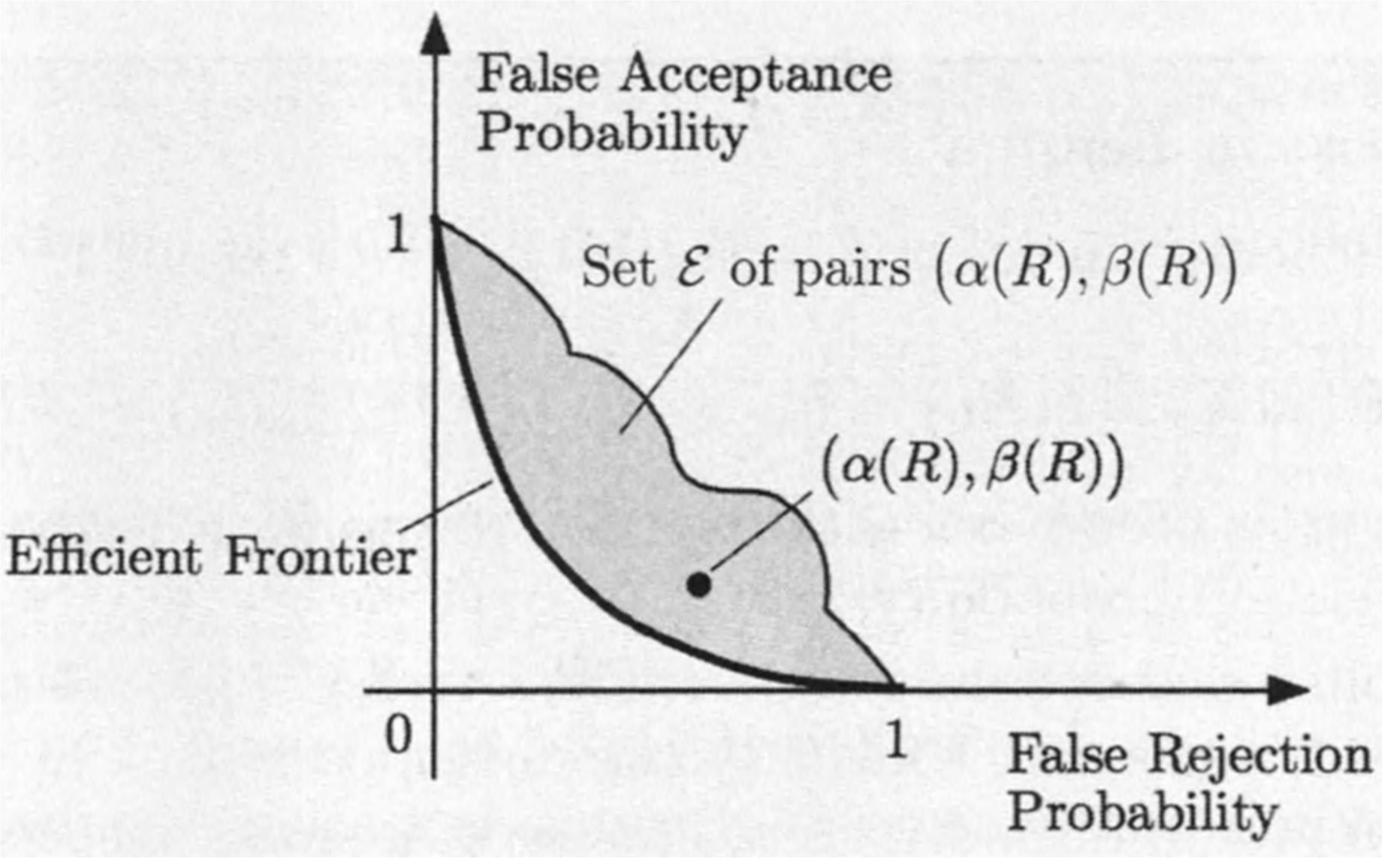
- In such cases, there are several possibilities:
- Strive for approximate equality.
- Choose the smallest value of ξ that satisfies

$$P(L(X) > \xi; H_0) \leq \alpha$$

- We have motivated so far the use of a LRT through an analogy with Bayesian inference.
- However, it also has a stronger justification.
- For a given false rejection probability, the LRT offers the smallest possible false acceptance probability.

Neyman-Pearson Lemma

- Consider a particular choice of ξ in the LRT, which results in error probabilities
$$P(L(X) > \xi; H_0) = \alpha, \ P(L(X) \le \xi; H_1) = \beta.$$
- Suppose that some other test, with rejection region R, achieves a smaller or equal false rejection probability:
$$P(X \in R; H_0) \le \alpha \qquad (1)$$
- Then, $\quad P(X \notin R; H_1) \ge \beta. \qquad (2)$
- In addition, if (1) is strict, so is (2).


False Acceptance
Probability
1
Set $\mathcal{E}$ of pairs $(\alpha(R), \beta(R))$
$(\alpha(R), \beta(R))$
Efficient Frontier
0
1
False Rejection
Probability

Content

- Classical Parameter Estimation
- Linear Regression
- Binary Hypothesis Testing
- Significance Testing

- Hypothesis testing problems encountered in realistic settings do not always involve two well-specified alternatives.
- So the methodology in the preceding section cannot be applied.
- This section introduces an approach to this more general class of problems.
- Note: a unique or universal methodology is not available. There is a significant element of judgment and art that comes into play.

Motivation

- Consider problems such as the following:
  - A coin is tossed repeatedly and independently. Is the coin fair?
  - We observe a sequence of i.i.d. normal random variables $X_1, \ldots, X_n$. Are they standard normal?
  - Two different drug treatments are delivered to two different groups of patients with the same disease. Is the first treatment more effective than the second?

- On the basis of historical data (say, based on the last year), is the daily change of the Dow Jones Industrial Average normally distributed?
- On the basis of several sample pairs (x_i, y_i) of two random variables X and Y, can we determine whether the two random variables are independent?
- …

- In all of the above cases, we are dealing with a phenomenon that involves uncertainty,
  - presumably governed by a probabilistic model.
- We have a default hypothesis, usually called the null hypothesis, denoted by H_0,
- We wish to determine on the basis of the observations $X = (X_1, \dots, X_n)$, whether the null hypothesis should be rejected or not.

- In order to avoid obscuring the key ideas, we will mostly restrict the scope of our discussion to situations with the following characteristics.
  - Parametric models: We assume that the observations $X_1, \ldots, X_n$ have a distribution governed by a joint PMF/PDF, which is completely determined by an unknown parameter θ (scalar or vector), belonging to a given set M of possible parameters.

- Simple null hypothesis: The null hypothesis asserts that the true value of θ is equal to a given element θ_0 of M.

- Alternative hypothesis: The alternative hypothesis, denoted by H_1, is just the statement that H_0 is not true, i.e., that $\theta \neq \theta_0$.

The General Approach

- We introduce the general approach through a concrete example.
- We then summarize and comment on the various steps involved.

Example: Is my coin fair?

- A coin is tossed independently $n = 1000$ times.
- Let θ be the unknown probability of heads at each toss.
- The set of all possible parameters is $M = [0,1]$.
- The null hypothesis H_0 ("the coin is fair") is of the form $\theta = 1/2$. The alternative hypothesis is that $\theta \neq 1/2$.

- The observed data is a sequence $X_1, \ldots, X_n$
  - where X_i equals 1 or 0, depending on whether the ith toss resulted in heads or tails.
- We choose to address the problem by considering the value of $S = X_1 + \cdots + X_n$, the number of heads observed, and using a decision rule of the form:

  $$\text{reject } H_0 \text{ if } \left| S - \frac{n}{2} \right| > \xi$$

  where ξ is a suitable critical value, to be determined.

- We finally choose the critical value ξ so that the probability of false rejection is equal to a given value α:

$$P(\text{reject } H_0; H_0) = \alpha$$

- Typically, α, called the significance level, is a small number:
  - In this example, we use $\alpha = 0.05$.
- Some probabilistic calculations are now needed to determine the critical value ξ.

- Some probabilistic calculations are now needed to determine the critical value ξ.
- Under the null hypothesis, the random variable S is binomial with parameters $n = 1000$ and $p = 1/2$.
- Using the normal approximation to the binomial and the normal tables, we find that an appropriate choice is $\xi = 31$.

- If, for example, the observed value of S turns out to be $s = 472$, we have

$$|s - 500| = |472 - 500| = 28 \leq 31.$$

- And the hypothesis H_0 is not rejected at the 5% significance level.
  - "not rejected" (as opposed to "accepted“): We do not have any firm grounds to assert that θ equals ½, as opposed to, say, 0.51.
  - We can only assert that the observed value of S does not provide strong evidence against hypothesis H_0.

Significance Testing Methodology

- A statistical test of a hypothesis "$H_0: \theta = \theta^*$" is to be performed, based on the observations $X = (X_1, \dots, X_n)$.
- 1. The following steps are carried out before the data are observed.
  - 1.1 Choose a statistic S, that is, a scalar random variable that will summarize the data X. This involves the choice of a function $h: R^n \to R$, resulting in the statistic $S = h(X)$.

- 1.2 Determine the shape of the rejection region by specifying the set of values of S for which H_0 will be rejected as a function of a yet undetermined critical value ξ.
- 1.3 Choose the significance level, i.e., the desired probability α of a false rejection of H_0.
- 1.4 Choose the critical value ξ so that the probability of false rejection is equal (or approximately equal) to α. (At this point, the rejection region is completely determined.)

- 2. Once the values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ are observed:
  - 2.1 Calculate the value $s = h(x_1, \ldots, x_n)$ of the statistic S.
  - 2.2 Reject the hypothesis H_0 if s belongs to the rejection region.

Comments and interpretation

- There is no universal method for choosing the "right" statistic S.
- The set of values of S under which H_0 is not rejected is usually an interval surrounding the peak of the distribution of S under H_0.
- Typical choices for the false rejection probability a range between $\alpha = 0.10$ and $\alpha = 0.01$.
- Step 1.4 is the only place where probabilistic calculations are used.

- Given the value of α, if the hypothesis H_0 ends up being rejected, one says that H_0 is rejected at the a significance level.
- Note: It does not mean that the probability of H_0 being true is less than α.
- Instead, it means that when this particular methodology is used, we will have false rejections a fraction α of the time.

- Quite often, statisticians skip steps 1.3 and 1.4 in the above described methodology.
- Instead, once they calculate the realized value s of S, they determine and report an associated p-value defined by

$$\{\min \alpha \,|\, H_0 \text{ would be rejected at the } \alpha \text{ significance level}\}$$

- Equivalently, the p-value is the value of α for which s would be exactly at the threshold between rejection and non-rejection.

- Thus, for example, the null hypothesis would be rejected at the 5% significance level if and only if the p-value is smaller than 0.05.