Statistical Inference

- Data generated from unknown probability distribution and statement on the unknown distribution are warranted. Determine parameters (e.g. β for exponential distribution, μ and σ for normal distribution)
- Prediction of new experiments

Estimation of parameters

- Notation: $f(x|\theta)$ is the probability density of sampling x given (conditioned on) parameters θ.
- For a set of n independent and identically distributed samples the probability density is:

$$f\left(x_1,\ldots,x_n \mid \theta\right) = \prod_{i=1,\ldots,n} f\left(x_i \mid \theta\right) \equiv f\left(\mathbf{x} \mid \theta\right)$$

- However, what we want to determine now are the parameters… For example assuming the distribution is normal, we seek the mean μ and the variance σ^2

$$f\left(x \mid \mu,\sigma^2\right) = \sqrt{\frac{1}{2\pi\sigma}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

Bayesian arguments

- What we want is the function $f\left(\theta \mid \mathbf{x}\right)$ given a set of observations **x**, what is the probability that the set of parameters is θ?
- Bayesian statistics: Think of the parameters like other random variables with probability $\xi(\theta)$.

The joint probability $f\left(\mathbf{x},\theta\right) \equiv f\left(\mathbf{x} \mid \theta\right)\xi\left(\theta\right)$ is also

$$f\left(\mathbf{x},\theta\right) \equiv f\left(\theta \mid \mathbf{x}\right)g\left(\mathbf{x}\right)$$

The likelihood function

- We can formally write $\xi(\theta \mid \mathbf{x}) = \dfrac{f(\mathbf{x} \mid \theta)\,\xi(\theta)}{g(\mathbf{x})}$

which is the probability of having a particular set of parameter for the p.d.f provided a set of observation (what we wanted). Note that our prime interest here is in the parameter set *θ* and the samples of *x* is given. Since *g(x)* is independent of *θ* we can write <u>the likelihood function</u>

$$\xi(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)\,\xi(\theta)$$

Example: Likelihood function I

- Consider the exponential distribution

$$f(x \mid \beta) = \begin{Bmatrix} \beta \exp[-\beta x] & \text{for } x > 0 \\ 0 & \text{otherwise} \end{Bmatrix}$$

$$f(\mathbf{x} \mid \beta) = \begin{Bmatrix} \beta^n \exp\left[-\beta \sum_{i=1,\ldots,n} x_i\right] & \text{for } x > 0 \\ 0 & \text{otherwise} \end{Bmatrix}$$

- And assume the p.d.f. of the parameter β is a Gaussian with a mean and variance of 1.

$$\xi(\beta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\beta^2}{2}\right)$$

Example: Likelihood function II

$$\xi\left(\beta \mid \mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{1/2}} \exp\left(-\frac{\beta^2}{2}\right) \beta^n \exp\left[-\beta \sum_{i=1,\ldots,n} x_i\right]$$

Maximum Likelihood

We look for a maximum of the function $L(\theta) = \log\left(f_n\left(\mathbf{x} \mid \theta\right)\right)$

as a function o fthe parameters θ

As a concrete example we consider the normal distribution

$$L(\theta) = \log\left[f_n\left(\mathbf{x} \mid \mu, \theta\right)\right]$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1,\ldots,n}\left(x_i - \mu\right)^2$$

To find the most likely set of parameters we determine the maximum of *L(θ)*

Maximum of $L(\theta)$ for normal distribution

$$\frac{dL}{d\mu} = 0 = -\frac{1}{2\sigma^2} \sum_{i=1,\dots,n} 2\left(x_i - \mu\right) = -\frac{1}{\sigma^2}\left(\sum_{i=1,\dots,n} x_i - n\mu\right)$$

$$\Rightarrow \mu = \frac{1}{n} \sum_{i=1,\dots,n} x_i$$

$$\frac{dL}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2} \sum_{i=1,\dots,n} \left(x_i - \mu\right)^2 = 0$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1,\dots,n} \left(x_i - \mu\right)^2$$

Determine a mot likely parameter for the uniform distribution

$$f\left(x \mid \theta\right) = \begin{Bmatrix} \dfrac{1}{\theta} & \text{for } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{Bmatrix}$$

$$f\left(\mathbf{x} \mid \theta\right) = \begin{Bmatrix} \dfrac{1}{\theta^n} & \text{for } 0 \le x_i \le \theta \ \left(i = 1, \dots, n\right) \\ 0 & \text{otherwise} \end{Bmatrix}$$

It is clear that θ must be larger than all the x_i and at the same time maximizes the monotonically decreasing function $1/\theta^n$, hence

$$\theta = \max\left[x_1, \dots, x_n\right]$$

Potential problems in maximum likelihood procedure

• Value of θ is underestimated (note that θ should be larger than all *x*, not only the ones we sample so far)

•No guarantee that a solution exists for the distribution below θ must be large than any *x* but at the same time equal to the maximal *x.* This is not possible and hence, no solution

$$f\left(x \mid \theta\right) = \begin{cases} \dfrac{1}{\theta} & \text{for } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

•The solution is not necessarily unique

$$f_n\left(\mathbf{x} \mid \theta\right) = \begin{cases} 1 & \text{for } \theta \le x_i \le \theta + 1 \ \left(\text{i=1,...,n}\right) \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow f_n\left(\mathbf{x} \mid \theta\right) = \begin{cases} 1 & \text{for } \max\left(x_1,\ldots,x_n\right) - 1 \le \theta \le \min\left(x_1,\ldots,x_n\right) \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \max\left(x_1,\ldots,x_n\right) - 1 \le \theta \le \min\left(x_1,\ldots,x_n\right)$$

The χ^2 distribution with *n* degrees of freedom

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} \exp(-x/2) \qquad x > 0$$

$$E(x) = n \qquad \mathrm{var}(x) = 2n$$

There is a useful relation between the χ^2 and the normal distributions

Theorem connecting χ^2 and normal distributions

If the random variables $X_1, \ldots, X_n$ *are i.i.d. and* if each of these variables has standard normal distribution, then the sum of the squares

$$Y^2 = X_1^2 + \ldots + X_n^k$$

Has a χ^2 distribution with n degrees of freedom

The distribution functions

$$F(y) = \Pr(Y \le y) = \Pr\left(X^2 \le y\right) = \Pr\left(-y^{1/2} \le X \le y^{1/2}\right)$$

$$= \Phi\left(y^{1/2}\right) - \Phi\left(-y^{1/2}\right)$$

The p.d.f is obtained by differentiating both side $f(y) = F'(y)$

$\phi(y) = \Phi'(y)$. Note $\phi\left(y^{1/2}\right) = (2\pi)^{-1/2} \exp(-y/2)$. We have

$$f(y) = \phi\left(y^{1/2}\right)\left(1/2y^{-1/2}\right) + \phi\left(-y^{1/2}\right)\left(1/2y^{-1/2}\right)$$

$$f(y) = (2\pi)^{-1/2} y^{-1/2} \exp(-y/2)$$

which is the χ^2 distribution with one degree of freedom

Normal distribution: Parameters

Let $X_1, \ldots, X_n$ be a random sample from normal distribution having mean μ and variance σ^2. Then the sample mean (hat denotes M.L.E)

$$\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1,\ldots,n} X_i$$

and the sample variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1,\ldots,n} \left(X_i - \bar{X}_n \right)^2$$

are independent random variables.

$\hat{\mu}$ has a normal distribution with a mean μ and variance σ^2/n.

$n\hat{\sigma}^2 / \sigma^2$ has a chi-square distribution of $n-1$ degrees of freedom

Why $n-1$? (next slide)

Parameters of the normal distribution: Note 1

- Let $x_1,...,x_n$ be a vector of random number of length *n* sampled from the normal distribtuion
- Let $y_1,...,y_n$ be another vector of *n* random numbers, related to the previous vector by linear transformation *A* *(AA^t=I)*

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

- Consider now the calculation of the variance (next slide)

Variance

- The formula we should use for the variance

$$\mathrm{var}(X) = \frac{1}{n} \sum_{i=1,\dots,n} (X_i - \mu)^2$$

- However, we do not know the exact mean, and therefore we use

$$\mathrm{var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

- What are the consequences of using this approximation?

Variance is not changing upon linear transformations

- Consider the expression

$$\sum_{i=1,\ldots,n}\left(Y_i-\bar{Y}_n\right)^2=\sum_{i=1,\ldots,n}\left(\mathbf{A}X_i-\mathbf{A}\bar{X}_n\right)^t\left(\mathbf{A}X_i-\mathbf{A}\bar{X}_n\right)$$

$$\sum_{i=1,\ldots,n}\left(X_i^t-\bar{X}_n\right)^t\mathbf{A}^t\mathbf{A}\left(X_i^t-\bar{X}_n\right)=\sum_{i=1,\ldots,n}\left(X_i^t-\bar{X}_n\right)^2$$

- The analysis is based on the unitarity of **A**. Hence, linear transformation dos change the variance of he distribution. This makes it possible to exploit the difference between

$$\bar{X}_n \text{ and } \mu$$

The *n-1* (versus *n*) factor

- Since **A** is arbitrary (as long as it is unitary). We can choose one of the transformation vectors **a** to be $(1,\ldots,1)/n^{1/2}$
- The scalar product

$$X^t\mathbf{a} - \bar{X}_n\mathbf{a} = 0$$

- Is identically zero (remember how we compute the mean?)
- Hence since we computed the average from the same sample we computed the variance, the variance lost one degree of freedom.

The *n-1 factor II*

- Note that the *n-1* makes sense. Consider only a single sample point, which is of course very poor and leaves a high degree of uncertainty regarding the value sof the parameters. If we use *n* then the estimated variance becomes zero, while if we use *n-1* we obtain infinite, which is more appropriate to the problem at hand, for which we have no information to determine the variance

The *t* distribution (in preparation for confidence intervals)

- Consider two random variables *Y* and *Z*, such that *Y* has chi-2 distribution with n degrees of freedom and *Z* has a standard normal distribution the variable *X* is defined by

$$X = Z \Big/ \left(\frac{Y}{n^{1/2}} \right)$$

Then the distribution of *X* is the *t* distribution with *n* degrees of freedom.

The t distribution

- The function is tabulated and can be written in terms of Γ function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x)\, dx$$

$$t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{1/2}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \qquad \text{for } -\infty < x < \infty$$

- The t distribution is approaching the normal distribution as $n \to \infty$. It has the same mean but longer tails.

Confidence Interval

- Confidence interval provide an alternative to the use of estimator instead of the actual value of an unknown parameter. We can find an interval *(A,B) that we think has high probability of containing the desired parameter. The length of the interval gives us an idea how well we can estimate the parameter value.*

Confidence interval for the mean of the normal distribution

- Let $X_1,\ldots,X_n$ for a random sample from a normal distribution with unknown mean and unknown variance. Let $t_{n-1}(x)$ denote the p.d.f of the t distribution with $n-1$ degrees of freedom, and let c be a constant such that

$$\int_{-c}^{c} t_{n-1}(x)\,dx = \gamma$$

- For every value of n, the value of c can be found from the table of the t distribution to fit the confidence (probability) γ