MIT OpenCourseWare
http://ocw.mit.edu

1.010 Uncertainty in Engineering
Fall 2008

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu.terms.

1.010 - Brief Notes # 9

Point and Interval Estimation of Distribution Parameters

(a) Some Common Distributions in Statistics

.

- **Chi-square distribution**

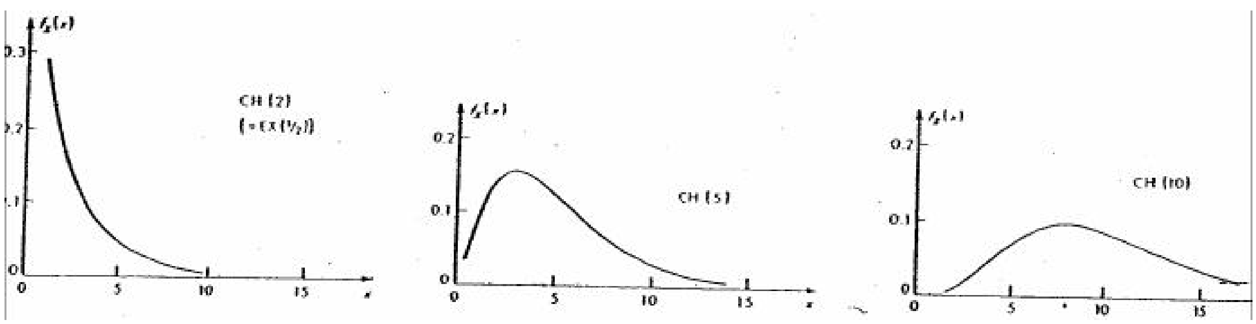
  Let $Z_1, Z_2, \ldots, Z_n$ be iid standard normal variables. The distribution of

  $\chi^2_n = \sum_{i=1}^{n} Z_i^2$

  is called the Chi-square distribution with n degrees of freedom.

  $E[\chi^2_n] = n$

  $Var[\chi^2_n] = 2n$



Probability density function of χ^2_n for $n = 2, 5, 10$.

- **t distribution**

  Let $Z, Z_1, Z_2, \ldots, Z_n$ be iid standard normal variables. The distribution of

  $$t_n = \frac{Z}{\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^2\right)^{1/2}}$$

  is called the Student's t distribution with n degrees of freedom.

$E[t_n] = 0$

$$Var[t_n] = \begin{cases} \dfrac{n}{n-2}, & n > 2 \\ \infty, & n \le 2 \end{cases}$$

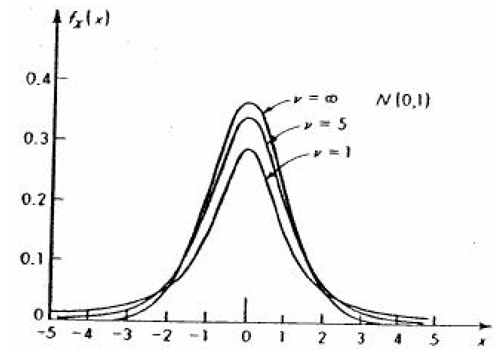


Fig. 3.4.4 The t distribution. (From A. Hald, "Statistical Theory with Engineering Applications," John Wiley and Sons, New York, 1952.)

Probability density function of t_n for $n = 1, 5, \infty$.
Note: $t_\infty = N(0,1)$.

- **F distribution**

Let $W_1, W_2, \ldots, W_m, Z_1, Z_2, \ldots, Z_n$ be iid standard normal variables. The distribution of

$$F_{m,n} = \frac{\frac{1}{m}\sum_{i=1}^{m} W_i^2}{\frac{1}{n}\sum_{i=1}^{n} Z_i^2} = \frac{\frac{1}{m}\chi_m^2}{\frac{1}{n}\chi_n^2}$$

is called the F distribution with m and n degrees of freedom.

As $n \to \infty$, $\quad mF_{m,n} \to \chi_m^2$

(b) Point Estimation of Distribution Parameters: Objective and Criteria

.

- **Definition of (point) estimator**

  Let θ be an unknown *parameter* of the distribution F_X of a random variable X, for example the mean m of the variance σ^2. Consider a random *sample* of size n from the statistical *population* of $X, \{X_1, X_2, \ldots, X_n\}$. An estimator $\widehat{\Theta}$ of θ is a function $\widehat{\Theta}(X_1, X_2, \ldots, X_n)$ that produces a numerical estimate of θ for each realization $x_1, x_2, \ldots, x_n$ of $X_1, X_2, \ldots, X_n$. Notice: $\widehat{\Theta}$ is a random variable whose distribution depends on θ.

- **Desirable properties of estimators**

1. *Unbiasedness:*
   $\widehat{\Theta}$ is said to be an unbiased estimator of θ if, for any given θ, $E_{\text{sample}}[\widehat{\Theta}|\theta] = \theta$. The bias $b_{\widehat{\Theta}}(\theta)$ of $\widehat{\Theta}$ is defined as:

   $$b_{\widehat{\Theta}}(\theta) = E_{\text{sample}}[\widehat{\Theta}|\theta] - \theta$$

2. *Mean Squared Error (MSE):*
   The mean squared error of $\widehat{\Theta}$ is the second initial moment of the estimation error $\mathrm{e} = \widehat{\Theta} - \theta$, i.e.,

   $$MSE_{\widehat{\Theta}}(\theta) = E[(\widehat{\Theta} - \theta)^2] = b^2_{\widehat{\Theta}}(\theta) + Var[\widehat{\Theta}|\theta]$$

   One would like the mean squared error of an estimator to be as small as possible.

(c) Point Estimation of Distribution Parameters: Methods

.

1. **Method of moments**

   Suppose that F_X has unknown parameters $\theta_1, \theta_2, \ldots, \theta_r$. The idea behind the method of moments is to estimate $\theta_1, \theta_2, \ldots, \theta_r$ so that r selected characteristics of the distribution match their sample values. The characteristics are often taken to be the initial moments:

   $$\mu_i = E[X^i], \quad i = 1, \ldots, r$$

   The method is described below for the case $r = 2$.

The first and second initial moments of X are, in general, functions of the unknown parameters, θ_1 and θ_2:

$$\mu_1(\theta_1, \theta_2) = E[X|\theta_1, \theta_2] = \int x f_{X|\theta_1,\theta_2}(x)dx$$

$$\mu_2(\theta_1, \theta_2) = E[X^2|\theta_1, \theta_2] = \int x^2 f_{X|\theta_1,\theta_2}(x)dx$$

The sample values of these moments are:

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

$$\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Estimators of θ_1 and θ_2 are obtained by solving the equations for $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$:

$$\mu_1(\widehat{\Theta}_1, \widehat{\Theta}_2) = \widehat{\mu}_1$$

$$\mu_2(\widehat{\Theta}_1, \widehat{\Theta}_2) = \widehat{\mu}_2$$

This method is often simple to apply, but may produce estimators that have higher MSE than other methods, e.g. maximum likelihood.

Example:

If $\theta_1 = m$ and $\theta_2 = \sigma^2$, then:

$\mu_1 = m$ and $\mu_2 = m^2 + \sigma^2$

$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ and $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

The estimators $\widehat{m}$ and $\widehat{\sigma}^2$ are obtained by solving:

$$\widehat{m} = \overline{X}$$

$$\widehat{m}^2 + \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

which gives:

$$\widehat{m} = \overline{X}$$

$$\widehat{\sigma}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_1 - \overline{X})^2$$

Notice that $\widehat{\sigma}^2$ is a biased estimator since its expected value is $\frac{n-1}{n}\sigma^2$. For this reason, one typically uses the modified estimator:

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

which is unbiased.

2. **Method of maximum likelihood:**

Consider again the case $r = 2$. The likelihood function of θ_1 and θ_2 given a sample, $L(\theta_1, \theta_2|\text{ sample})$, is defined as:

$$L(\theta_1, \theta_2|\text{ sample}) \propto P[\text{sample }|\theta_1, \theta_2]$$

Where P is either probability or probability density and is regarded for a given sample as a function of θ_1 and θ_2. In the case when X is a continuous variable:

$$P[\text{sample }|\theta_1, \theta_2] = \prod_{i=1}^{n} f_X(x_i|\theta_1, \theta_2)$$

The maximum likelihood estimators $(\widehat{\Theta}_1)_{ML}$ and $(\widehat{\Theta}_2)_{ML}$ are the values of θ_1 and θ_2 that maximize the likelihood, i.e.,

$L(\theta_1, \theta_2|\text{ sample})$ is maximum for $\theta_1 = (\widehat{\Theta}_1)_{ML}$ and $\theta_2 = (\widehat{\Theta}_2)_{ML}$

In many cases, $(\widehat{\Theta}_1)_{ML}$ and $(\widehat{\Theta}_2)_{ML}$ can be found by imposing the stationarity conditions:

$$\frac{\partial L[(\widehat{\Theta}_1, \widehat{\Theta}_2)|\text{ sample}]}{\partial \widehat{\Theta}_1} = 0 \quad \text{and} \quad \frac{\partial L[(\widehat{\Theta}_1, \widehat{\Theta}_2)|\text{ sample}]}{\partial \widehat{\Theta}_2} = 0$$

or, more frequently, the equivalent conditions in terms of the log-likelihood:

$$\frac{\partial(\ln L[(\widehat{\Theta}_1, \widehat{\Theta}_2)|\text{ sample}])}{\partial\widehat{\Theta}_1} = 0 \quad \text{and} \quad \frac{\partial(\ln L[(\widehat{\Theta}_1, \widehat{\Theta}_2)|\text{ sample}])}{\partial\widehat{\Theta}_2} = 0$$

- Properties of maximum likelihood estimators:

  As the sample size $n \to \infty$, maximum likelihood estimators:
  1. are unbiased;
  2. have the smallest possible value of MSE.

- Example:

  For $X \sim N(m, \sigma^2)$ with unknown parameters m and σ^2, the maximum likelihood estimators of the parameters are:

$$\widehat{m}_{ML} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$$

$$\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{m}_{ML})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \frac{\sigma^2}{n}\chi^2_{(n-1)}$$

  Notice that in this case the ML estimators m and σ^2 are the same as the estimators produced by the method of moments. This is not true in general.

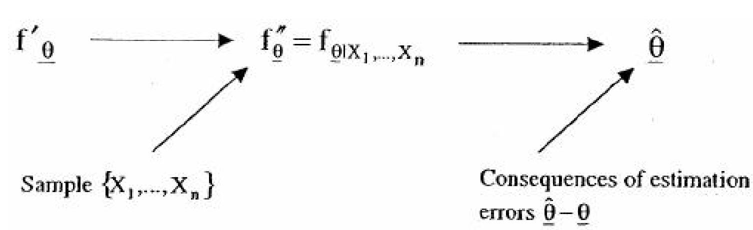
3. **Bayesian estimation**

The previous two methods of point estimation are based on the classical statistical approach which assumes that the distribution parameters $\theta_1, \theta_2, \ldots, \theta_r$ are constants but unknown. In Bayesian estimation, $\theta_1, \theta_2, \ldots, \theta_r$ are viewed as uncertain (random variables) and their uncertainty is quantified through probability distributions. There are 3 steps in Bayesian estimation:

Step 1: Quantify initial uncertainty on $\underline{\theta}$ in the form of a prior distribution, $f'_{\underline{\theta}}$

Step 2: Use sample information to update uncertainty $\to$ posterior distribution, $f''_{\underline{\theta}}$

Step 3: Choose a single value estimate of $\underline{\theta}$



The various steps are described below in the order 2, 3, 1.

<u>Step 2:</u> **How to update prior uncertainty given a sample**

Recall that for random variables,

$f_{\underline{\theta}|\underline{X}} \propto f_{\underline{\theta}}(\underline{\theta}) \cdot f_{\underline{X}|\underline{\theta}}(\underline{x})$

Here, $f'_{\underline{\theta}} = f_{\underline{\theta}}$ and $f''_{\underline{\theta}} = f_{\underline{\theta}|\underline{X}}$. Further, using $\ell(\underline{\theta}|\underline{X}) \propto f_{\underline{X}|\underline{\theta}}(\underline{x})$, one obtains:

$f''_{\underline{\theta}}(\underline{\theta}) \propto f'_{\underline{\theta}}(\underline{\theta})\ell(\underline{\theta}|\underline{X})$

<u>Step 3:</u> **How to choose $\widehat{\underline{\theta}}$**

Two main methods:

1. *Use some characteristic of* $f''_{\underline{\theta}}$, such as the mean or the mode. The choice is rather arbitrary. Note that the mode corresponds in a sense to the maximum likelihood, applied to the posterior distribution rather than the likelihood.

2. *Decision theoretic approach: (more objective and preferable)*
   θ by $\widehat{\theta}$.
   - Define a loss function $\$(\widehat{\underline{\theta}}|\underline{\theta})$ which is the loss if the estimate is $\widehat{\underline{\theta}}$ and the true value is $\underline{\theta}$.

   - Calculate the expected posterior loss or "Risk" of $\widehat{\underline{\theta}}$ as:
     $R(\widehat{\underline{\theta}}) = E''[\$(\widehat{\underline{\theta}}|\underline{\theta})] = \int_{-\infty}^{\infty} \$(\widehat{\underline{\theta}}|\underline{\theta}) f''_{\underline{\theta}}(\underline{\theta}) d\underline{\theta}$

   - Choose $\widehat{\underline{\theta}}$ such that $R(\widehat{\underline{\theta}})$ is **minimum**.
     - If $\$(\widehat{\underline{\theta}}|\underline{\theta})$ is a quadratic function of $(\widehat{\theta}_i - \theta_i)$, then $R(\widehat{\underline{\theta}})$ is minimum for $\widehat{\underline{\theta}} = E''[\underline{\theta}]$
     - If $\$(\widehat{\underline{\theta}}|\underline{\theta}) = \begin{cases} 0, & \text{if } \widehat{\underline{\theta}} = \underline{\theta} \\ c > 0, & \text{if } \widehat{\underline{\theta}} \neq \underline{\theta} \end{cases}$, then $\widehat{\underline{\theta}}$ is the mode of $f''_{\underline{\theta}}$.

<u>Step 1:</u> **How to select $f'_{\underline{\theta}}$**

1. *Judgementally.* This approach is especially useful in engineering design, where subjective judgement is often neccessary. This is how subjective judgement is formally incorporated in the decision process.

2. Based on *prior data* e.g. a "sample" of $\underline{\theta}$'s from other data sets.

3. To *reflect ignorance*, "non-informative prior".
   For example, if θ is a scalar parameter that can attain values from $-\infty$ to $+\infty$,
   then $f'_\theta(\theta)d\theta \propto d\theta$ ("flat") and $f''_\theta(\theta) \propto \ell(\theta|\text{ sample})$ i.e. the posterior reflects only the likelihood.

   If $\theta > 0$, then one typically takes $f'_{\ln\theta}(\ln\theta)d\ln\theta \propto d\ln\theta$. In this case, $f'_\theta(\theta) \propto \dfrac{1}{\theta}$.

4. *Conjugate prior.* There are distribution types such that if $f'_\theta(\theta)$ is of that type, then $f''_\theta(\theta) \propto f'_\theta(\theta)\ell(\theta)$ is also of the same type. Such distributions are called conjugate distributions.

<u>Example:</u>

Let:

$X \sim N(m, \sigma^2)$ with σ^2 known. $\theta = m$ unknown.

Suppose: $f'_m \sim N(m', \sigma'^2)$

It can be shown that $\ell(m|X_1, \ldots, X_n) \propto$ density of $N(\overline{X}, \sigma^2/n)$

From $f''_m \propto f'_m \ell(m|\text{ sample})$, one obtains

$$f''_m \sim N\left(m'' = \frac{m'(\sigma^2/n) + \overline{X}\sigma'^2}{(\sigma^2/n) + \sigma'^2}, \frac{1}{\sigma''^2} = \frac{1}{\sigma'^2} + \frac{n}{\sigma^2}\right)$$

In this case, $f'_m \sim N(m', \sigma'^2)$ is an example of a conjugate prior, since f''_m is also normal, of the type $N(m'', \sigma''^2)$.

If one writes $\sigma'^2 = \dfrac{\sigma^2}{n'}$, then n' has the meaning of equivalent prior sample size and m' has the meaning of equivalent prior sample average.

(d) Approximate Confidence Intervals for Distribution Parameters

•

1. *Classical Approach*

   Problem: θ is an unknown distribution parameter. Define two sample statistics $\widehat{\Theta}_1(X_1, \ldots, X_n)$ and $\widehat{\Theta}_2(X_1, \ldots, X_n)$ such that:

   $P[\widehat{\Theta}_1(X_1, \ldots, X_n) < \theta < \widehat{\Theta}_2(X_1, \ldots, X_n)] = P^*$

   where P^* is a given probability.

An interval $[\widehat{\Theta}_1(X_1, \ldots, X_n), \widehat{\Theta}_2(X_1, \ldots, X_n)]$ with the above property is called a confidence interval of θ at confidence level P^*.

A simple method to obtain confidence intervals is as follows. Consider a point estimation $\widehat{\Theta}$ such that, exactly or in approximation, $\widehat{\Theta} \sim N(\theta, \sigma^2(\theta))$. If the variance $\sigma^2(\theta)$ depends on θ, one replaces $\sigma^2(\theta)$ with $\sigma^2(\widehat{\Theta})$. Then:

$$\frac{\widehat{\Theta} - \theta}{\sigma(\widehat{\Theta})} \sim N(0,1)$$

$$\Rightarrow P[\widehat{\Theta} - \sigma(\widehat{\Theta})Z_{P^*/2} < \theta < \widehat{\Theta} + \sigma(\widehat{\Theta})Z_{P^*/2}] = P^*$$

where Z_α is the value exceeded with probability α by a standard normal variable.

Example:

$\theta = m =$ mean of an exponential distribution.

In this case, $\widehat{\Theta} = \overline{X} \sim \dfrac{1}{n}\text{Gamma}(m, n)$, where $\text{Gamma}(m, n)$ is the distribution of the sum of n iid exponential variables, each with mean value m. The mean and variance of $\text{Gamma}(m, n)$ are nm and nm^2, respectively. Moreover, for large n, $\text{Gamma}(m, n)$ is close to $N(nm, nm^2)$. Therefore, in approximation,

$$\overline{X} \sim N\left(m, \frac{m^2}{n}\right)$$

Using the previous method, an approximate confidence interval for m at confidence level P^* is

$$\left[\overline{X} - \frac{\overline{X}}{\sqrt{n}} \cdot Z_{P^*/2}, \quad \overline{X} + \frac{\overline{X}}{\sqrt{n}} \cdot Z_{P^*/2}\right]$$

2. *Bayesian Approach*

In Bayesian analysis, intervals $[\widehat{\theta}_1, \widehat{\theta}_2]$ that contain θ with a given probability P^* are simply obtained from the condition that:

$$F''_\theta(\widehat{\theta}_2) - F''_\theta(\widehat{\theta}_1) = P^*$$

where F''_θ is the posterior CDF of θ.