Republic of Iraq Ministry of Higher Education & Research

University of Anbar

College of Education for Pure Sciences

Department of Mathematics

محاضرات الاحصاء ١

مدرس المادة : الاستاذ المساعد الدكتور

فراس شاكر محمود

A factorization theorem: The characterization of a sufficient statistic in terms of the conditional distribution of the data given the statistic can be difficult to work with.The following factorization theorem provides a convenient means of identifying sufficient statistics.

Theorem: A necessary and sufficient condition for the statistic $T(X1,...,Xn)$ to be sufficient for a parameter θ is that the joint probability function (density function or frequency function) factors in the form $f(x_1,...,xn|\theta) = g[T(x_1,...,xn),\theta]h(x_1,...,xn)$.

Proof We shall only consider the case where the X_i's are discrete random variables. (The ideas for the proof of the continuous case is similar but is technically more challenging.)

Sufficient condition

First suppose the frequency function (or pmf) factors as

$f(x_1,...,xn|\theta)= g[T(x_1,...,xn),\theta]h(x_1,...,xn)$.

Writing $X = (X1,...,Xn)$ and $x = (x1,...,xn)$, we have $P(T = t) = \sum P(X = x)\ x:T(x)=t$

$= g(t,\theta)\ \sum h(x).\ x:T(x)=t.$ Consequently,

$$P(X = x|T = t) = \frac{P(X = x,T = t)}{P(T = t)}$$

$$= \frac{h(x)}{\sum x: T(x) = t\ h(x)}$$

This implies that $P(X = x|T = t)$ does not depend on θ and we conclude that $T(X)$ is sufficient for θ.

Necessary condition

Now suppose that $T(X)$ is sufficient for θ. Then P(X = x|T = t) does not depend on θ.

Let g(t,θ) = P(T = t|θ), h(x) = P(X = x|T = t).Then we have

$P(X = x|\theta) = P(T = t|\theta)P(X = x|T = t)= g(t,\theta)h(x)$.

This completes the proof of the factorization theorem.

Example:

Consider a sequence of Bernoulli random variables $X1,...,Xn$ where

$P(X_i = x_i) = \theta^{xi}(1-\theta)^{1-xi}$, $x_i = 0,1$. Then writing $x = (x1,...,xn)$,

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= (\frac{\theta}{1-\theta})^{\sum_{i=1}^n x_i}(1-\theta)^n \\ &= g(t,\theta)h(x), \end{aligned}$$

where

$$\begin{aligned} t &= \sum_{i=1}^{n} x_i, \\ g(t,\theta) &= (\frac{\theta}{1-\theta})^t(1-\theta)^n \\ h(x) &= 1. \end{aligned} \quad ,$$

We conclude from the factorization theorem that $T = \sum_{i=1}^n X_i$ is sufficient for θ.

Example:

Consider a random sample $X = (X1,...,Xn)$ from $N(\mu,\sigma^2)$ where μ and σ^2 are both unknown. Then

$$\begin{aligned} & f(x|\mu,\sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x_i-\mu)^2] \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2] \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp[-\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 \\ &\quad -2\mu\sum_{i=1}^{n} x_i + n\mu^2)]. \end{aligned}$$

The rhs is only a function of $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$.

It follows from the factorization theorem

that $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$ are the sufficient statistics for μ and σ^2.

$$T(X) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2),$$

is an example of a 2-dimensional sufficient statistic.

Some remarks If $T(X1,...,Xn)$ is sufficient for the parameter θ, then the M. L. E. must be a function **only** of T. Likewise in the Bayesian approach, the posterior distribution of θ given the data X = (X1,...,Xn) is equal to the posterior distribution of θ given only T(X).

F. distribution

We have motivated the *t*-distribution in part by its application to problems in which there is comparative sampling (i.e., a comparison between two sample means). For example, some of our examples in future chapters will take a more formal approach, chemical engineer collects data on two catalysts, biologist collects data on two growth media, or chemist gathers data on two methods of coating material to inhibit corrosion. While it is of interest to let sample information shed light on two population means, it is often the case that a comparison of variability is equally important, if not more so. The *F*-distribution finds enormous application in comparing sample variances. Applications of the *F*-distribution are found in problems involving two or more samples. The statistic *F* is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom. Hence, we can write $F = \frac{U/v_1}{V/v_2}$

where U and V are independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. We shall now state the sampling distribution of F.

Theorem 1: Let U and V be two independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

This is known as the ***F*-distribution** with v_1 and v_2 degrees of freedom (d.f.).

We will make considerable use of the random variable F in future chapters. However, the density function will not be used and is given only for completeness. The curve of the F-distribution depends not only on the two parameters v_1 and v_2 but also on the order in which we state them. Once these two values are given, we can identify the curve. Typical F-distributions are shown in Figure 1.

Let f_α be the f-value above which we find an area equal to α. This is illustrated by the shaded region in Figure 2. Table A.6 gives values of f_α only for $\alpha = 0.05$ and $\alpha = 0.01$ for various combinations of the degrees of freedom v_1 and v_2. Hence, the f-value with 6 and 10 degrees of freedom, leaving an area of 0.05 to the right, is $f_{0.05} = 3.22$. By means of the following theorem, Table A.6 can also be used to find values of $f_{0.95}$ and $f_{0.99}$. The proof is left for the reader.

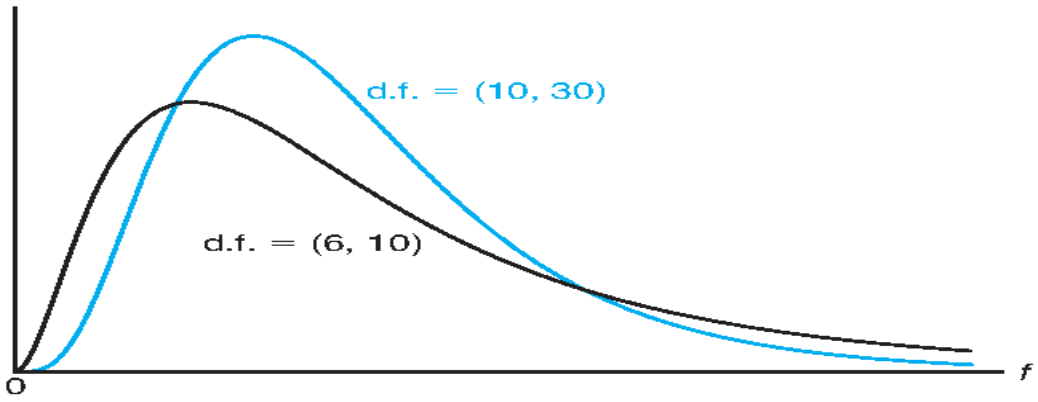

Figure1: Typical F-distributions.

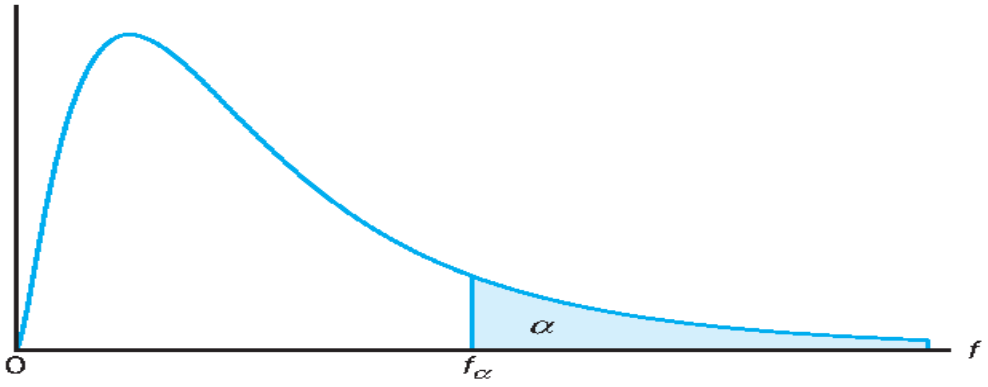

Figure2: Illustration of the f_α for the F-distribution.

Theorem2:

Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

Thus, the *f*-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right, is

$$f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246.$$

The F-Distribution with Two Sample Variances

Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variances σ_1^2 and σ_2^2, respectively. From Theorem 8.4, we know that $\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2}$ and $\chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2}$

are random variables having chi-squared distributions with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Furthermore, since the samples are selected at random, we are dealing with independent random variables. Then, using Theorem 1 with $\chi_1^2 = U$ and $\chi_2^2 = V$, we obtain the following result.

Theorem 3 :

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an *F*-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

What Is the F-Distribution Used For?

We answered this question, in part, at the beginning of this section. The *F* distribution is used in two-sample situations to draw inferences about the population variances. This involves the application of Theorem 3 . However, the

F-distribution can also be applied to many other types of problems involving sam- ple variances. In fact, the *F*-distribution is called the *variance ratio distribution*. As an illustration, consider Case Study 8.2, in which two paints, *A* and *B*, were compared with regard to mean drying time. The normal distribution applies nicely (assuming that σ_A and σ_B are known). However, suppose that there are three types of paints to compare, say *A*, *B*, and *C*. We wish to determine if the population means are equivalent. Suppose that important summary information from the experiment is as follows:

Paint	Sample Mean	Sample Variance	Sample Size
A	$\bar{X}_A = 4.5$	$s^2_A = 0.20$	10
B	$\bar{X}_B = 5.5$	$s^2_B = 0.14$	10
C	$\bar{X}_C = 6.5$	$s^2_C = 0.11$	10

The problem centers around whether or not the sample averages ($\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$) are far enough apart. The implication of "far enough apart" is very important. It would seem reasonable that if the variability between sample averages is larger than what one would expect by chance, the data do not support the

conclusion that $\mu_A = \mu_B = \mu_C$. Whether these sample averages could have occurred by chance depends on the *variability within samples*, as quantified by s^2_A, s^2_B, and s^2_C. The notion of the important components of variability is best seen through some simple graphics. Consider the plot of raw data from samples *A*, *B*, and *C*, shown in Figure 8.13. These data could easily have generated the above summary information. It appears evident that the data came from distributions with different population means, although there is some overlap between the samples. An analysis that involves all of the data would attempt to determine if the variability between the sample averages *and* the variability within the samples could have occurred jointly *if in fact the populations have a common mean*. Notice that the key to this analysis centers around the two following sources of variability.

Variability within samples (between observations in distinct samples)

Variability between samples (between sample averages)

Clearly, if the variability in (1) is considerably larger than that in (2), there will be considerable overlap in the sample data, a signal that the data could all have come

from a common distribution. An example is found in the data set shown in Figure 4. On the other hand, it is very unlikely that data from distributions with a common mean could have variability between sample averages that is considerably larger than the variability within samples. The sources of variability in (1) and (2) above generate important ratios of *sample variances*, and ratios are used in conjunction with the *F*-distribution. The general procedure involved is called **analysis of variance**. It is interesting that in the paint example described here, we are dealing with inferences on three population means, but two sources of variability are used. We will not supply details here, but in Chapters 13 through 15 we make extensive use of

analysis of variance, and, of course, the F-distribution plays an important role.