Lecture 3

Intro to Statistics. Inferences for normal families.

1 Random Sampling

Often, the data collected in an experiment consist of several observations of a variable of interest. This situation is often described by a model called random sampling.

Definition 1. Random variables $X_1, ..., X_n$ are called *a random sample* of size n from population distribution f if (a) $X_1, ..., X_n$ are mutually independent and (b) $X_i \sim f$.

In general, f means the cdf of the population distribution. In the case of continuously distributed random variables, f can be understood as its pdf. Thus, by random sample we mean an iid sample from the same population distribution.

If we have a random sample $X_1, ..., X_n$ of size n from a population distribution with pdf f, then, by definition of independence, we can write the joint pdf as

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

With some abuse of notation we often denote $f_{X_1,...,X_n}(x_1, ..., x_n)$ simply by $f(x_1, ..., x_n)$.

We refer to any function of a random sample as a *statistic*. Thus, if $g : \mathbb{R}^n \to \mathbb{R}$ is some function, then $Y = g(X_1, ..., X_n)$ is a statistic. Its distribution is called the *sampling distribution*.

1.1 Sample mean and sample variance.

The two most commonly used statistics are the *sample mean* ($\overline{X}_n = \sum_{i=1}^n X_i/n$) and the *sample variance* ($s^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$). These statistics have attractive properties as described in the lemma below:

Lemma 2. *If $X_1, ..., X_n$ is a random sample of size n from a population distribution with mean μ and variance σ^2, then $E[\overline{X}_n] = \mu$ and $E[s^2] = \sigma^2$.*

Proof. By linearity of expectation,

$$E[\overline{X}_n] = E[\sum_{i=1}^{n} X_i/n] = \sum_{i=1}^{n} E[X_i]/n = \sum_{i=1}^{n} \mu/n = \mu$$

To show the second part of the lemma, denote $Y_i = X_i - \mu$ and $\overline{Y}_n = \sum_{i=1}^n Y_i/n$. Note that $E[Y_i] = 0$. Thus, $E[Y_i^2] = V(Y_i) = V(X_i) = \sigma^2$ and $V(\overline{Y}_n) = \sigma^2/n$. Then

$$\begin{aligned}
E[s^2] &= E[\sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)] \\
&= E[\sum_{i=1}^n ((X_i - \mu) - (\overline{X}_n - \mu))^2/(n-1)] \\
&= E[\sum_{i=1}^n (Y_i - \overline{Y}_n)^2/(n-1)] \\
&= E[\sum_{i=1}^n (Y_i^2 - 2Y_i\overline{Y}_n + \overline{Y}_n^2)]/(n-1) \\
&= E[\sum_{i=1}^n Y_i^2 - 2n\overline{Y}_n^2 + n\overline{Y}_n^2]/(n-1) \\
&= E[\sum_{i=1}^n Y_i^2 - n\overline{Y}_n^2]/(n-1) \\
&= (\sum_{i=1}^n E[Y_i^2] - nE[\overline{Y}_n^2])/(n-1) \\
&= (n\sigma^2 - \sigma^2)/(n-1) \\
&= \sigma^2
\end{aligned}$$

□

1.2 Empirical Distribution Function

If we have a random sample $X_1, ..., X_n$ of size n, then empirical distribution function $\hat{F}_n$ is the cdf of the distribution that puts mass $1/n$ at each data point X_i. Thus, by definition,

$$\hat{F}_n(x) = \sum_{i=1}^n I(X_i \leq x)/n,$$

where $I(\cdot)$ stands for the indicator function, i.e. the function which equals 1 if the statement in brackets is true, and 0 otherwise. In other words, $\hat{F}_n(x)$ shows the fraction of observations with a value smaller or equal than x. An important property of an empirical distribution function is given in the lemma below.

Lemma 3. *If we have a random sample $X_1, ..., X_n$ of size n from a distribution with cdf F, then for any $x \in \mathbb{R}$, $E[\hat{F}_n(x)] = F(x)$ and $V(\hat{F}_n(x)) \to 0$ as $n \to \infty$. As a consequence, $\hat{F}_n(x) \to F(x)$ in L_2 and $\hat{F}_n(x) \to_p F(x)$ as $n \to \infty$.*

Proof. Note that $I(X_i \leq x)$ equals 1 with probability $P\{X \leq x\}$ and 0 otherwise. Thus, $E[I(X_i \leq x)] = P\{X \leq x\} = F(x)$. Hence $E[\hat{F}_n(x)] = F(x)$ by linearity of expectation. In addition, $V(I(X_i \leq x)) =$

$F(x)(1-F(x))$ by the formula for variance of a Bernoulli $(F(x))$ distribution. Therefore,

$$V(\hat{F}_n(x)) = \sum_{i=1}^{n} V(I(X_i \leq x))/n^2 = F(x)(1-F(x))/n \to 0$$

To prove the second part of the lemma, we have$E[(\hat{F}_n(x) - F(x))^2] = V(\hat{F}_n(x)) \to 0$ as $n \to \infty$ since $E[\hat{F}_n(x)] = F(x)$. Convergence in probability then follows from convergence in quadratic mean. □

Actually a much more strong result holds as well:

Theorem 4 (Glivenko-Cantelli)**.** *If $X_1, ..., X_n$ is a random sample from a distribution with cdf F, then*

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to_p 0$$

(and almost surely).

1.3 Plug-in estimators

Suppose we have a random sample $X_1, ..., X_n$ of size n from population distribution with cdf F. Suppose T is some function on the space of possible cdfs. Suppose we do not know F but we are interested in $T(F)$. Then we can use some statistic $g(X_1, ..., X_n)$ to estimate $T(F)$. In this case $g(X_1, ..., X_n)$ is called an estimator of $T(F)$. Its realization $g(x_1, ..., x_n)$ is called an estimate of $T(F)$. Here $x_1, ..., x_n$ stand for realizations of $X_1, ..., X_n$. What is a good estimator of $T(F)$? By common sense, a good estimator $g(X_1, ..., X_n)$ should be such that $g(X_1, ..., X_n) \approx T(F)$, at least with large probability. One possible estimator is $T(\hat{F}_n)$ where $\hat{F}_n$ is the empirical cdf. $T(\hat{F}_n)$ is called a plug-in estimator. From the Glivenko-Cantelli theorem we know that $\hat{F}_n$ will be close to F with large probability in large samples. Thus, if T is continuous, then $T(\hat{F}_n)$ will be close to $T(F)$.

As an example, suppose we are interested in the mean of the population distribution, i.e. $\mu = T(F) = E[X] = \int_{-\infty}^{+\infty} x dF(x)$.Then

$$\hat{\mu} = T(\hat{F}_n) = \int_{-\infty}^{+\infty} x d\hat{F}_n(x) = \sum X_i/n = \overline{X}_n.$$

Thus, the plug-in estimator of the population mean is just the sample average. Next, suppose we are interested in the variance of the population distribution, i.e. $\sigma^2 = T(F) = E[(X - E[X])^2] = \int_{-\infty}^{+\infty}(x -$

$\int_{-\infty}^{+\infty} x dF(x))^2 dF(x)$. Then

$$\begin{aligned}
\hat{\sigma}^2 &= T(\hat{F}_n) \\
&= \int_{-\infty}^{+\infty} (x - \int_{-\infty}^{+\infty} x d\hat{F}_n(x))^2 d\hat{F}_n(x) \\
&= \int_{-\infty}^{+\infty} (x - \overline{X}_n)^2 d\hat{F}_n(x) \\
&= \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 / n
\end{aligned}$$

Thus, the plug-in estimator of the population variance does not coincide with the sample variance. The reason we use $n-1$ in the denominator of the sample variance instead of n is to make it unbiased for the population variance, i.e. $E[s^2] = \sigma^2$. Note that $E[\hat{\sigma}^2] = (n-1)\sigma^2/n \neq \sigma^2$.

Finally, consider the plug-in estimator of quantiles. We already defined quantile of the distribution in lecture 1 as $q_p = \inf\{x : F(x) \geq p\}$ so that q_p is the p-th quantile of distribution F. Thus, plug-in estimator of the p-th quantile is $\hat{q}_p = \inf\{x : \hat{F}_n(x) \geq p\}$.

1.4 Order Statistics

Let $X_1, ..., X_n$ be a random sample from some distribution F. Let us order elements of $X_1, ..., X_n$ such that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Thus, $X_{(1)}$ is the minimal element among $X_1, ..., X_n$ and $X_{(n)}$ is the maximal one. In general, $X_{(i)}$ for all $i = 1, ...n$ are called order statistics. $X_{(n)} - X_{(1)}$ is called the sample range. $X_{\lceil n/2 \rceil}$ is called the sample median. Here $\lceil x \rceil$ denotes the smallest integer larger than x. Thus, $X_{\lceil n/2 \rceil} = \hat{q}_{1/2}$, i.e. the sample median equals plug-in estimator of 1/2-th quantile. $X_{\lceil n/4 \rceil}$ is called the lower quartile and $X_{\lceil 3n/4 \rceil}$ the upper quartile. Finally, $X_{\lceil 3n/4 \rceil} - X_{\lceil n/4 \rceil}$ is the interquartile range.

The theorem below shows how one can calculate the density of order statistics.

Theorem 5. *Let $X_1, ..., X_n$ be a random sample from a distribution with cdf F and pdf f. Then the pdf of $X_{(j)}$ is*

$$f_j(x) = \frac{n!}{(j-1)!(n-j)!} f(x)(F(x))^{j-1}(1-F(x))^{n-j}$$

for all $j = 1, ..., n$.

Proof. By definition of order statistics, the cdf of $X_{(j)}$ is

$$\begin{aligned}
F_j(x) &= P\{X_{(j)} \leq x\} \\
&= P\{\text{at least } j \text{ observations} \leq x\} \\
&= \sum_{i=j}^{n} P\{\text{exactly } i \text{ observations} \leq x\} \\
&= \sum_{i=j}^{n} \{n!(F(x))^i (1-F(x))^{n-i} / (i!(n-i)!)\}
\end{aligned}$$

Differentiating $F_j(x)$ yields

$$\begin{aligned}
f_j(x) &= dF_j(x)/dx \\
&= \sum_{i=j}^{n}\{n!(F(x))^{i-1}(1-F(x))^{n-i}f(x)/((i-1)!(n-i)!)\} \\
&- \sum_{i=j}^{n}\{n!(F(x))^{i}(1-F(x))^{n-i-1}f(x)/(i!(n-i-1)!)\} \\
&= \sum_{i=j}^{n}\{n!(F(x))^{i-1}(1-F(x))^{n-i}f(x)/((i-1)!(n-i)!)\} \\
&- \sum_{k=j+1}^{n}\{n!(F(x))^{k-1}(1-F(x))^{n-k}f(x)/((k-1)!(n-k)!)\} \\
&= n!(F(x))^{j-1}(1-F(x))^{n-j}f(x)/((j-1)!(n-j)!)
\end{aligned}$$

□

As an example, let $X_1, ..., X_n$ be a random sample from $U[0,1]$. Then $f(x) = 1$ if $x \in [0,1]$ and 0 otherwise. Thus, $F(x) = 0$ if $x < 0$, $F(x) = x$ if $x \in [0,1]$, and $F(x) = 1$ otherwise. By the theorem above,

$$f_j(x) = n!x^{j-1}(1-x)^{n-j}/((j-1)!(n-j)!)$$

if $x \in [0,1]$ and $f_j(x) = 0$ otherwise. So, by elementary algebra, $E[X_{(j)}] = \int_0^1 x f_j(x)dx = j/(n+1)$ and $V(X_{(j)}) = j(n-j+1)/((n+1)^2(n+2))$.

2 Parametric Families: Normal

The plug-in estimator considered above is a generic nonparametric estimator of some function $T(F)$ of distribution F in the sense that it does not use any information about the class of possible distributions. However, in practice, it is often assumed that the class of possible distributions form some parametric family. In other words, it is assumed that $F = F(\theta)$ with $\theta \in \Theta$ where Θ is some finite-dimensional set. Then θ is called a parameter and Θ is a parameter space. In this case the cdf F and the corresponding pdf f are often denoted by $F(x|\theta)$ and $f(x|\theta)$. If $X_1, ..., X_n$ is a random sample from a distribution with pdf $f(x|\theta)$, then joint pdf $f(x_1, ..., x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$. For fixed $x_1, ..., x_n$, $f(x_1, ..., x_n|\theta)$ as a function of θ is called the likelihood function.

One of the most important parametric families is a normal family when $\theta = (\mu, \sigma^2)$ and the population distribution is $N(\mu, \sigma^2)$. Before considering normal family, let us give some definitions related to normal distributions.

If $X_1, ..., X_n$ is a random sample from $N(0,1)$, then random variable $\chi^2_n = \sum_{i=1}^n X_i^2$ is called a χ^2 random variable with n degrees of freedom. Its distribution is called a χ^2 distribution with n degrees of freedom. It is known that its pdf is given by $f(x) = x^{p/2-1}e^{-x/2}/(\Gamma(p/2)2^{p/2})$ if $x > 0$ and 0 otherwise. Here $\Gamma(x)$ denotes the gamma function. Its values can be found in special tables.

Next, if X_0 is $N(0,1)$ and independent of $X_1, ..., X_n$, then $t_n = X_0/\sqrt{\chi^2_n/n}$ is called a t random variable with n degrees of freedom. Its distribution is called a t-distribution or a *Student distribution*.

Finally, if χ^2_n and χ^2_m are independent χ^2 random variables with n and m degrees of freedom correspondingly, then $F_{n,m} = (\chi^2_n/n)/(\chi^2_m/m)$ is called a *Fisher random variable* with (n,m) degrees of freedom. Its distribution is called a Fisher distribution with (n,m) degrees of freedom.

The following theorem gives some basic facts about the sample mean and the sample variance for random sample from normal distribution:

Theorem 6. *If $X_1, ..., X_n$ are iid random variables with $N(\mu, \sigma^2)$ distribution, then (1) $\overline{X}_n$ and s^2_n are independent, (2) $\overline{X}_n \sim N(\mu, \sigma^2/n)$, and (3) $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.*

Proof. Let $Z = \overline{X}_n$, $Y_1 = X_1 - \overline{X}_n$, $Y_2 = X_2 - \overline{X}_n$, ..., $Y_n = X_n - \overline{X}_n$. Then $Z, Y_1, ..., Y_n$ are jointly normal. Obviously, $E[Z] = \mu$ and $E[Y_i] = \mu - \mu = 0$ for all $i = 1, ..., n$. In addition, $V(Z) = \sigma^2/n$. Thus statement (2) holds.

For any $j = 1, ..., n$,

$$\begin{aligned} \text{cov}(Z, Y_j) &= \text{cov}(\overline{X}_n, X_j - \overline{X}_n) \\ &= \text{cov}(\overline{X}_n, X_j) - V(\overline{X}_n) \\ &= \sigma^2/n - \sigma^2/n \\ &= 0 \end{aligned}$$

Since uncorrelated jointly normal random variables are independent, we conclude that Z is independent of $Y_1, Y_2, ..., Y_n$. Moreover, $s^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1) = \sum_{i=1}^n Y_i^2/(n-1)$ and statement (1) holds since any functions of independent random variables are independent as well.

The proof of statement (3) is left for Problem set 1. □

By definition, $t = (\overline{X}_n - \mu)/(s/\sqrt{n})$ is called the t-statistic. Using the theorem above,

$$t = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{s^2/\sigma^2}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1}$$

since $N(0,1)$ and χ^2_{n-1} in the display above are independent. Thus, we proved that if $X_1, ..., X_n$ is a random sample from $N(\mu, \sigma^2)$, then t-statistic has t-distribution with $n-1$ degrees of freedom.

Finally, let $X_1, ..., X_n$ be a random sample from $N(\mu_x, \sigma^2_x)$ and $Y_1, ..., Y_m$ be a random sample from $N(\mu_y, \sigma^2_y)$. Assume that $X_1, ..., X_n$ are independent of $Y_1, ..., Y_m$. Then $F = (s^2_x/s^2_y)/(\sigma^2_x/\sigma^2_y)$ is called a F-statistic. Using the theorem above,

$$F = \frac{s^2_x/s^2_y}{\sigma^2_x/\sigma^2_y} \sim \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} = F_{n-1,m-1}$$

Thus, the F-statistic has the F-distribution with $(n-1, m-1)$ degrees of freedom.

14.381 Statistical Method in Economics
Fall 2013