Lecture 3

Intro to Statistics. Inferences for normal families.

1 Basic Concepts: Population, Sample, Parameter, Statistics

In statistics we usually have a data set (or a *sample*), which can be described as a single draw of data from all potential realizations of that data. We may describe it as a realization x of random vector $\mathcal{X}$. The distribution F_X of data vector $\mathcal{X}$ is often referred as *population*.

In Econometrics one encounters 3 types of data: cross-section, time series and panel. Cross-section is usually described as a set of iid (independent and identically distributed) random vectors $X_1, ..., X_n$ (that is, $\mathcal{X} = (X_1, ..., X_n), x = (x_1, ..., x_n)$). If we assume that $X_i \sim F$, then $F_X(x) = \prod_{i=1}^{n} F(x_i)$. Time-series data $X_t, t = 1, ..., T$ usually allow dependency between consecutive observations and describe $(X_1, ..., X_T)$ as one realization of a path that results from a dynamic process. Panel data usually consider $\mathcal{X} = \{X_{it}, i = 1, .., n, t = 1, ..., T\}$ that we have a draw from n independent identically distributed dynamic processes.

The object of interest here is usually some functional of the unknown distribution F_X. Any such function is known as a *parameter*. In the case of cross-sectional data, it is usually some function of distribution of one observation F; for time series, the parameter may be related to the dependence between observations as well as marginal distributions of observation. Notice that parameter is a population concept.

The goal of Statistics generally is to render some judgement about a parameter (or population F_X) based on a single draw from this population. This is called *inference*. We will see three types of inference: estimation, confidence set construction and testing. In performing each task we will sometimes make mistakes, and the quality of the procedure will be related to minimizing the size and/or probability of mistakes.

We refer to any function of a random sample as a *statistic*. Thus, $Y = g(\mathcal{X}) = g(X_1, ..., X_n)$ is a statistic. By construction, it is random variable. When calculated for our specific data set $y = g(x) = g(x_1, ..., x_n)$ it produces a single realization of this random variable. The distribution of a statistic is called the *sampling distribution*.

Example 1. Assume you want to figure out whether the penny you have is a fair coin by flipping it $n = 10$ times and recording 0 for each tail and 1 for each head. In this case your data $x = (x_1, ..., x_{10})$ is a sequence of of 0s and 1s of the length 10, where x_i is the result of i-th experiment. This is a single realization of random vector $X = (X_1, ..., X_{10})$, where $X_i \sim$ *i.i.d. Bernoulli*(p). The population here is described as $P(x_1, ..., x_{10}) = \prod_{i=1}^{10} p^{x_i}(1-p)^{1-x_i}$, and is known up to parameter p. So, the only goal is to make some judgement about p. Remember, statistics refers to any function of the data set. We may consider many different functions: say,

$$Y_1 = \{\text{number of heads}\} = \sum_{i=1}^{10} X_i,$$

$Y_2 = \{$the order number of the first experiment resulting in heads, with 0 if no heads$\} = X_1 + 2(1 - X_1)X_2 + 3(1 - X_1)(1 - X_2)X_3 + ...,$

$Y_3 = \{$the difference between the number of heads in the first 6 flips and the number of heads in the remaining 4$\} = \sum_{i=1}^{6} X_i - \sum_{i=7}^{10} X_i$.

If one has data=(0,1,1,1,0,0,1,1,0,1), then $y_1 = 6, y_2 = 2, y_3 = 0$. All statistics Y_1, Y_2, Y_3 are random variables with a distribution depending on p.

Example 2. Assume you want to estimate the average income of a man aged between 25 and 65, who resides in Massachusetts. You were able to get a random sample of income for n such men. Your data is represented by random vector $X = (X_1, ..., X_n)$, where $X_i \sim i.i.d.F(\cdot)$, where $F(\cdot)$ is the unknown distribution of incomes. The population is described by $F_X(x_1, ..., x_n) = \prod_{i=1}^{n} F(x_i)$. The parameter of interest is $\mu = \int u dF(u)$- the mean of the unknown income distribution. You have one realization of this data set $x = (x_1, ..., x_n)$, statistic is any function of the data set. Some examples:

$Y_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ (average);

$Y_2 = (n/2)-$th highest value among $(X_1, ..., X_n)$; (median)

Y_3= average of 80% of middle values (drop 10% smallest and 10% largest values)- a trimmed mean.

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 1. *If $X_1, ..., X_n \sim i.i.dF$ is a random sample of size n from a population distribution with mean $\mu = EX_i$ and variance $\sigma^2 = Var(X_i)$, 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

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

$$= E[\sum_{i=1}^{n} Y_i^2 - 2n\overline{Y}_n^2 + n\overline{Y}_n^2]/(n-1) = E[\sum_{i=1}^{n} Y_i^2 - n\overline{Y}_n^2]/(n-1)$$
$$= (\sum_{i=1}^{n} E[Y_i^2] - nE[\overline{Y}_n^2])/(n-1) = (n\sigma^2 - \sigma^2)/(n-1) = \sigma^2$$

□

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 places 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 than or equal to x. An important property of an empirical distribution function is given in the lemma below.

Lemma 2. *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, $\hat{F}_n(x) \to_p F(x)$ 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 Chebyshev's inequality. □

Actually, a much more strong result holds as well:

Theorem 3 (Glivenko-Cantelli)**.** *If $X_1, ..., X_n$ is a random sample from a distribution with cdf F, then*

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

2 Ways to find the distribution of a statistic

In order to make inferences we often will need to know the distribution of different statistics. There are several ways of getting them.

2.1 Exact distribution

In rare cases one can actually calculate the exact distribution of a statistic.

Example 1(cont.) Consider statistic $Y_1 = \sum_{i=1}^n X_i$, it take integer values between 0 and 10 with

$$P\{Y_1 = k\} = \frac{10!}{k!(10-k)!} p^k (1-p)^{10-k}.$$

The distribution depends on an unknown p; once we know (or postulate p) we can calculate exact distribution of Y_1.

Example 2(cont) Let us try to figure out the distribution of $Y_1 = \frac{1}{n}\sum_{i=1}^n X_i$. If $X_i \sim F(\cdot)$ with pdf f then $X_1 + X_2 \sim f_2(u)$ where $f_2(u) = \int f(x)f(u-x)dx$. So, Y_1 has pdf

$$f_Y(y) = \int ... \int f(10y - y_1 - ... - y_{n-1}) f(y_1)...f(y_{n-1}) dy_1...dy_{n-1},$$

which is a complicated expression and is not very helpful in most situations. It depends in a significant way on the unknown distribution $F(\cdot)$. If we make some strong assumptions about the distribution of our data we may end up with an exact distribution of some statistics. For example, assume $X_i \sim i.i.d.N(\mu, \sigma^2)$, then $Y_1 \sim N(\mu, \sigma^2/n)$. However, this is a strange assumption to make about income distribution (why?).

2.2 Monte-Carlo Method.

Let us consider Example 2, but now with an alternative assumption of the log-normal distribution. Assume that X_i has the following pdf:

$$f(x; \mu, \sigma^2, \gamma) = (2\pi\sigma^2)^{-1/2}(x-\gamma)^{-1} e^{-\frac{(\log(x-\gamma)-\mu)^2}{2\sigma^2}} \mathbb{I}\{x > \gamma\}.$$

Apparently, getting a (closed-form) exact distribution of $Y_1 = \frac{1}{n}\sum_{i=1}^n X_i$ is not that easy. But we can do this numerically by the use of Monte-Carlo method. A typical algorithm for given (μ, σ^2, γ):

- For $b = 1, ..., B$, simulate $\mathcal{X}_b^* = (X_{1b}^*, ..., X_{nb}^*)$, where the X_{ib}^* are independently drawn from $f(x; \mu, \sigma^2, \gamma)$;
- Calculate $Y_b^* = g(\mathcal{X}_b^*)$, in our case $\frac{1}{n}\sum_{i=1}^n X_{ib}^*$;
- If you are interested in:
  - cdf of Y, then $F_Y(s) \approx \frac{1}{B}\sum_{b=1}^B \mathbb{I}\{Y_b^* \leq s\}$;
  - probability Y gets into the set A: $P\{Y \in A\} \approx \frac{1}{B}\sum_{b=1}^B \mathbb{I}\{Y_b^* \in A\}$;
  - α-quantile of Y: $q_Y(\alpha) \approx Y^*_{(\lfloor \alpha B \rfloor)}$, here $(\cdot)$- stands for the order statistics;
  - mean of Y: $EY \approx \frac{1}{B}\sum_{b=1}^B Y_b^*$;
  - variance of Y: $V(Y) \approx \frac{1}{B-1}\sum_{b=1}^B \left(Y_b^* - \frac{1}{B}\sum_{s=1}^B Y_s^*\right)^2$.

Here all the Y_b^* are i.i.d. from the correct distribution. We control the accuracy here: larger B leads to better accuracy; we are bounded only by power of our computers.

The important assumption here is we assume that we know (or postulate) the distribution of the data.

2.3 Asymptotic approximation

Exact finite-sample distributions of many statistics have complicated forms and cannot be calculated directly if we do not know (and do not want to assume) the distribution of the data. But at the same time that often may be well-approximated if the sample size is large. This is known as asymptotic approximation and relies on CLT, delta-methods, Slutsky theorem and alike.

Example 2(cont) If we are unwilling to assume the distribution of X_i at all, but are willing to assume that it has finite variance, then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - EX_i) \Rightarrow N(0, V(X_i)).$$

Thus the distribution of $Y_1 = \sum_{i=1}^{n} X_i$ is approximately gaussian with mean EX_i and variance $V(X_i)/n$. We do not control the quality of this approximation and cannot improve it, but as the sample size grows the approximation should become more accurate.

Asymptotic approximation is probably the most often used way of figuring out the distribution of a statistic in econometrics.

2.4 Bootstrap

The distribution of statistic $Y = g(\mathcal{X})$ is a function of $g(\cdot)$ and the population distribution F_X. The typical issue, as we have seen above, is that we do not know F_X. One idea may be to approximate F_X by some close distribution. If statistic Y depends on the distribution of the data in a continuous enough fashion, we may get a good enough approximation. For example, in the i.i.d case $F_X(x) = \prod_{i=1}^{n} F(x_i)$, all we need is to approximate F- the cdf of one observation.

As one example, Glivenko-Cantelli' theorem above suggests that empirical distribution $\hat{F}_n(x)$ may be a good approximation to $F(x)$. So, the idea is to run the same algorithm as one does with Monte-Carlo simulations, but simulate X_{ib}^* from distribution $\hat{F}_n(x)$ rather than from an unknown F. This specific procedure is called the non-parametric bootstrap:

- For $b = 1, ..., B$, simulate $\mathcal{X}_b^* = (X_{1b}^*, ..., X_{nb}^*)$, where the X_{ib}^* are drawn independently and uniformly with replacement from the set of initial observations $\{x_i, i = 1, ..., n\}$ (each x_i has the same probability to be drawn);
- Calculate $Y_b^* = g(\mathcal{X}_b^*)$, in our case $\frac{1}{n}\sum_{i=1}^{n} X_{ib}^*$;
- If you are interested in:
  - the cdf of Y, then $F_Y(s) \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \leq s\}$;

- probability Y gets into the set A: $P\{Y \in A\} \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{Y_b^* \in A\}$.
- α-quantile of Y: $q_Y(\alpha) \approx Y^*_{(\lfloor \alpha B \rfloor)}$, here $(\cdot)$ - stands for the order statistics.
- mean of Y: $EY \approx \frac{1}{B}\sum_{b=1}^{B} Y_b^*$.
- variance of Y: $V(Y) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(Y_b^* - \frac{1}{B}\sum_{s=1}^{B} Y_s^*\right)^2$.

Larger B will help eliminate some part of approximation error (simulation error), but some part (due to approximating F) cannot be controlled.

We will see many other ways to approximate $F(x)$, which would lead to a number of different bootstrap procedures.

There are often two ways to justify using the bootstrap. Typically we wish to claim that the distance between true distribution of statistic F_Y and the bootstrapped one F_{Y^*} is converging to zero in some sense as the sample size increases to infinity. One way is to show that F_Y is continuous in F in some sense (we have to be accurate as there are different metrics that can be introduced on the space of these distributions) and show that whatever approximation $\hat{F}$ we use, it converges to F in the proper metrics. The other way is to show that the distribution of Y converges somewhere (allow for asymptotic approximation) and that Y^* converges to the same distribution. We will discuss this many times in what follows.

3 Plug-in estimators

Suppose we have a random sample $X_1, ..., X_n$ of size n from a 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 the $x_1, ..., x_n$ 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's 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. $\sigma^2 = T(F) = E[(X - E[X])^2] = \int_{-\infty}^{+\infty}(x -$

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

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

Thus, the plug-in estimator of the population variance does not coincide with the sample variance. The reason we use $n-1$ instead of n in the denominator of the sample variance 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 the 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, the plug-in estimator of the p-th quantile is $\hat{q}_p=\inf\{x: \hat{F}_n(x)\geq p\}$.

4 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 sometimes 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 $\chi_n^2=\sum_{i=1}^n X_i^2$ is called a χ^2 random variable with n degrees of freedom. Its distribution is known as 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 the $\Gamma(x)$ 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 $t_n=X_0/\sqrt{\chi_n^2/n}$ 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_n^2 and χ_m^2 are independent χ^2 random variables with n and m degrees of freedom correspondingly, then $F_{n,m}=(\chi_n^2/n)/(\chi_m^2/m)$ is called a *Fisher random variable* with (n,m) degrees of freedom. This 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 4. *If $X_1, ..., X_n$ are iid random variables with $N(\mu, \sigma^2)$ distribution, then (1) $\overline{X}_n$ and s_n^2 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 $j = 1, ..., n$,

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

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 $N(\mu_x, \sigma_x^2)$ and $Y_1, ..., Y_m$ be a random sample from $N(\mu_y, \sigma_y^2)$. Assume that $X_1, ..., X_n$ are independent of $Y_1, ..., Y_m$. Then $F = (s_x^2/s_y^2)/(\sigma_x^2/\sigma_y^2)$ is called a F-statistic. Using the theorem above,

$$F = \frac{s_x^2/s_y^2}{\sigma_x^2/\sigma_y^2} \sim \frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)} = F_{n-1,m-1}$$

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