Economics 7720 J. Pepper
Spring 2024

Useful Theorems About Expectations

T1: Expectation and Variance of Linear Functions:

Let $Z = a + bX + cY$. Then

$$E[Z] = a + bE(x) + cE(y)$$
$$V(Z) = b^2V(x) + c^2V(y) + 2bcC(x,y).$$

What about nonlinear functions?

Jensen's Inequality: Suppose $Y = h(x)$ is convave and $E(X) = \mu$, then $E(y) \leq h(\mu)$.
For example, $E(\log(x)) \leq \log(E(x))$ and $E(x^2) \geq E^2(x)$ (see T3).

T2: Covariance between a pair of linear functions:

Let $z1 = a1 + b1X + c1Y$ and $z2 = a2 + b2X + c2Y$. Then

$$C(z1,z2) = b1b2V(x) + c1c2V(y) + (b1c2 + c1b2)*C(x,y).$$

T3: Covariance and Variance

$$C(x,y) = E(xy) - E(x)E(y)$$
$$V(x) = E(x^2) - E^2(x)$$

T4: Law of Iterated Expectations:

Let $Z = h(x,y)$. Then

$$E[Z] = E[E(Z|X)].$$

e.g., $E[Y] = E(E[Y | X])$, $E[XY] = E(E[XY | X]) = E(X E[Y | X])$

T5: Analysis of Variance:

$$V(y) = E[v(y|x)] + V[E(y|x)].$$

Proof: Write $Z = (y - \mu_y)^2 = [(y - E(y|x)) + (E(y|x) - \mu_y)]^2$. Square and take the expectation conditional on X to learn $E\{ Z | X)$. Then, using the law of iterated expectations, take the expectation over Z to learn $E(Z) = V(Y)$.

Economics 7720 J. Pepper
Spring 2024

Best Predictors

Let (y,x) have a known joint distribution P.[1] We are interested in predicting Y – that is, guess the value of Y that will occur on the next draw – so that we minimize the expected squared prediction error. Let U = Y – C(x) be the prediction error, where C(x) is our predictor that may vary by X.

We have considered the following predictors:

i. Constant Predictor: $C(x) = \theta$
ii. Linear Predictor: $C(x) = \alpha + \beta X$
iii. General Predictor: $C(x) = \theta(x)$, where $\theta(X)$ is any arbitrary function.

Best Constant Predictor: $\theta = E[Y]$.

Proof: $Min_c E[(Y-C)^2]$

FOC: $\frac{\partial E[U^2]}{\partial c} = \frac{\partial E[U^2]}{\partial U}\frac{\partial EU}{\partial c}$ = -2E[U] = 0 → E[U] = 0 → $\theta = E[Y]$.

Best Linear Predictor: $BLP[Y|X] = \alpha + \beta X$, where $\alpha = E[Y] - \beta E[X]$ and $\beta = C(X,Y) / V(X)$.

Proof: $Min_{c0,c_1} E[(Y-(c_0+c_1X))^2]$

FOC: -2E[U] = 0 → E[U] = E[Y – (c_0 + c_1X)] = 0
-2E[XU] = 0 → E[U] = E[X(Y – (c_0 + c_1X))] =0

Thus, the population best linear predictor solves the equations

1. $E(Y) = \alpha + ßE(X)$
2. $E(XY) = \alpha E(X) + ßE(X^2)$.

Solving these two equations reveals that

$\alpha = E(Y) - ßE(X)$

$$ß = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} = \frac{C(X,Y)}{V(X)}.$$

Best Predictor: $\theta(x) = E[Y|X]$

Proof: Conditional on a particular value of X, say $X = X_j$, the value of c that minimizes the expected squared prediction error is the mean of Y in the subpopulation where $X = X_j$. Since this holds for all j, E[Y | X] is the best predictor.

[1] For now, assume y and x are both scalar random variables. Soon, we will allow x to be a 1xK random vector.

Economics 7720 J. Pepper
Spring 2024

Inference About a Mean (Part 1)

How should we proceed to learn about the CEF in a bivariate population, when all we have in hand is a sample from that population? Because conditional expectations are generalizations of marginal expectations, we begin by reviewing methods of inference about a mean. How do we make inferences about the population mean, given a sample?

Let $E(Y) = \mu$ and $V(Y) = \sigma^2$.

Random sampling: In random sampling, a researcher draws N independent observations on Y:

Y_i , i. = 1, … N, where

$E(Yi) = \mu$, $V(Yi) = \sigma^2$ and $C(Y_h, Y_k) = 0$ for all i and $h \neq k$. .

In random sampling, independence ensures that all covariances are zero; while identical distribution ensures that all N expectations and variances are the same.

Sample Mean: Consider using the sample mean, $\bar{Y} = (1/N)\sum_i Y_i$, as an estimator of E[Y].

How informative is the sample mean? In practice, we will only have one sample, and thus only one estimate. The key here is to recognize that this one estimate reflects only a single draw from a distribution. Here, we focus on the sampling distribution of the sample mean. As before, we will focus on two features of this distribution --- the mean and the variance.

$$E[\bar{Y}] = \mathrm{E}[(1/N)\sum_i Y_i] = \frac{1}{N}\{E[Y_1] + E[Y_2] + \ldots + E[Y_N]\} = \frac{1}{N}(N\mu) = \mu.$$

$$V[\bar{Y}] = \mathrm{V}[(1/N)\sum_i Y_i] = \frac{1}{N^2}\{V[Y_1] + V[Y_2] + \ldots + V[Y_N]\} = \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}$$

The Sample Mean Theorem: Thus, we have shown that in a random sample, sample size n, from any population, the sample mean has an expectation equal to the population mean (that is, the sample mean is an unbiased estimator of the population mean), and a variance equal to the population variance divided by the sample size N.

Food for Thought: Where are we using independence? Where are we using identical distribution?

Economics 7720 J. Pepper
Spring 2024

Inference About a Mean (Part 2)

In random sampling, the sample mean is an unbiased estimator of the population mean. While unbiasedness is evidently an attractive property for an estimator, there are many other unbiased estimators. Why then should we prefer the sample mean to other unbiased estimators?

Gauss-Markov Theorem: The sample mean is the best linear unbiased estimator (BLUE) of the population mean. That is, in random sampling, sample size n, form any population, among all linear functions of the observations that are unbiased estimators of E(Y), the sample mean has the smallest variance.

Proof:
Consider the class of linear estimators, $m = \Sigma_i K_i Y_i$, where the K_i's are constants. If, for example, $K_i = 1/N$ for each observation, then m would be the sample mean.

We can compute the mean and variance of our estimator:

i. $E[m] = E(Y) * \Sigma_i K_i$ and
ii. $V[m] = V(Y) \Sigma_i K_i^2$

Let $K_i = 1/N + Ci$ and rewrite the mean and variance as

i. $E[m] = E(Y) * \Sigma_i (1/N + C_i) = E(Y) + E(Y) \Sigma_i C_i$ and
ii. $V[m] = V(Y) \{ \Sigma_i C_i^2 + \Sigma_i (1/N)^2 + 2/N \Sigma_i C_i \}$

Unbiasedness requires $\Sigma_i K_i = 1$ or equivalently that $\Sigma_i Ci = 0$. So, any for linear unbiased estimator m,

$$V[m] = V(Y) (\Sigma_i C_i^2 + \Sigma_i (1/N)^2) = V(Y)\{ \Sigma_i C_i^2 + 1/N \}$$

$$= V(Y)* \Sigma_i C_i^2 + V(\bar{Y}).$$

To minimize the variance of m, one should set $\Sigma_i C_i^2 = 0$, which means that all C_i's are set equal to zero, and all K_i's equal 1/N.

Example: Let N = 2. $E[m] = (k_1+k_2) E(Y)$ and $V(m) = (k_1^2 + k_2^2)V(Y)$. We want to minimize V(m) subject to the constraint the $k_1+k_2 = 1$. Note that $k_2 = 1-k_1$. So, we want to minimize $k_1^2 + (1- k_1)^2$. The first order condition is $2k_1 -2(1- k_1) = 0$ which implies that $k_1 = ½ = k_2$ achieves the minimum.

Food For Thought:
1.) What do we mean by linear estimator? Linear Conditional Expectation Function?
2.) What is the difference between estimation versus prediction?
3.) What is the difference between an estimator and a parameter?
4.) What is the difference between an estimator and an estimate?

5.) What is the difference between the variance of the sample mean, $V(\bar{Y})$, and the sample variance, $S^2 = \frac{1}{N}\sum_i (Y_i - \bar{Y})^2$.

Economics 7720 J. Pepper
Spring 2024

Inference About a Mean (Part 3)

Asymptotic Properties: For some purposes, it may be useful to consider the distribution of the sample statistic as the sample size gets larger.

Recall that in random sampling, sample size N, from any population with E[y] = μ and V(y) =σ^2,

$$E[\bar{Y}] = \mathrm{E}[(1/\mathrm{N})\sum_i \mathrm{Y}_i] = \frac{1}{\mathrm{N}}\{E[Y_1]+E[Y_2]+\ldots+E[Y_N]\} = \frac{1}{\mathrm{N}}(N\mu) = \mu, \text{ and}$$

$$V[\bar{Y}] = \mathrm{V}[(1/\mathrm{N})\sum_i \mathrm{Y}_i] = \frac{1}{\mathrm{N}^2}\{V[Y_1]+V[Y_2]+\ldots+V[Y_N]\} = \frac{1}{\mathrm{N}^2}(N\sigma^2) = \frac{\sigma^2}{N}.$$

What happens to the distribution of the sample mean as the sample size gets large? As N → 4

E[$\bar{Y}$] = μ and

V[$\bar{Y}$] → 0.

Thus, as the sample size gets large, the distribution of the sample mean becomes entirely concentrated at the point μ. This implies the

Law of Large Numbers (LLN): *In random sampling from any population with E[y] = μ and V(y) =σ^2, the sample mean, $\bar{Y}$, converges in probability to the population mean. We also say that $\bar{Y}$ →$_p$ μ or that $\bar{Y}$ is a <u>consistent</u> estimator of μ.*

An less intuitive asymptotic result is the

Central Limit Theorem (CLT): *In random sampling from any population with E[y] = μ and V(y) =σ^2, the standardized sample mean, $Z = \frac{\sqrt{N}(\bar{Y}-\mu)}{\sigma}$, converges in distribution to N (0,1). We also say that Z →$_d$ N (0,1)*

These results are quite useful (and somewhat remarkable) in that they apply regardless of the shape of the parent distribution.

Food For Thought:

1.) What is the relation between unbiased and consistent estimators? Are all unbiased estimators consistent? Are all consistent estimators unbiased?
2.) Besides being a remarkable result, why is the CLT useful?

Economics 7720 J. Pepper
Spring 2024

Estimators

In practice, the joint distribution, f(y,x), is not observed. Instead, we observe a single sample from the population and we are interested in some feature of the population, say a parameter θ. Suppose we have a single random sample, $\{y_i\ x_i\}$, i.=1,…, N, drawn from some unknown population. How shall we proceed? That is, what function of the data shall we use?

In this setting, the sample is a single observation on $\{Y,X\}= \{\ \{y_i\ x_i\}$ i.=1,…, N, $\}$ and the estimate we compute, t_n, will be a single observation on the random variable $T_n= h(y,x)$. T_N is the estimator, as distinguished from the estimate t_N.

I. How should we select our estimator, T_n:

Analogy Principle: The most natural rule for selecting an estimator is to use the analogy principle. A population parameter is some feature of the population (mean, median, variance, covariance,…). To estimate this parameter, use the corresponding feature of the sample. Equivalently, replace the population distribution P, with the empirical distribution P_n.

So, for example, suppose we are interest in the best constant predictor under squared loss, namely θ = E[Y]. Using the analogy principle, our estimator of the population expectation is the sample mean. Likewise, for the best linear predictor in a bivariate regression,

$$T_n = \begin{bmatrix} \alpha_N \\ \beta_N \end{bmatrix} = \begin{bmatrix} \bar{y} - \beta_N \bar{x} \\ \dfrac{S_{xy}}{S_{xx}} \end{bmatrix}.$$

II. Criteria for an Estimator: We would like T_N to be close to θ. Are analog estimators sensible (from a statistical point of view)? What should we do in the case where there are several estimators?

a. Small sample criteria for an estimator

General Criteria for an Estimator: Let θ ≡ best predictor(CEF) and T_n ≡ estimator.
A natural criteria is to find the estimator that minimizes the Mean Squared Error (MSE):

$$E[(T_N-\theta)^2] = E[(T_N-E(T_N)) - (\theta - E(T_N))^2\ ] = V(T_N) + (\theta - E(T_N))^2\ + 0.$$

In general, the variance and bias depend on the unknown parameter θ, and it is not feasible to find a T_n that minimizes MSE for all θ. Still, a small MSE is desirable.

If T_N is an unbiased estimator $\{E[\ T_N\ - \theta] = 0\}$, the MSE criteria is equivalent to minimizing the variance.

An estimator T_N is minimum variance unbiased if $V(T_N) \le V(T_N^*)$ for all T_N^* such that $E[\ T_N\ - \theta] = 0$ and $E[\ T_N^*\ - \theta] = 0$.

Gauss-MarkovTheorem:
In random sampling, sample size N, from any population, the sample mean is the *minimum variance linear unbiased estimator* of the population mean.

In the Classical Regression Model, the least-squares coefficient vector, b_n, is *minimum variance linear unbiased estimator* of the parameter vector β.

b. Asymptotic Criteria: The small sample properties of estimators can be difficult to evaluate as they can vary by specifics of the unknown population distribution and the sample. General results can be found if we use asymptotic theory.

Consistent Estimators: $T_N \rightarrow_{p,\,as} \theta$**.**

Best Asymptotically Normal (asymptotically efficient) : T_n is a BAN estimator (or asymptotically efficient) estimator of θ iff

i. $\sqrt{N}(T_n - \theta) \rightarrow_d N(0, \phi^2)$ and $\sqrt{N}(T_n^* - \theta) \rightarrow_d N(0, \phi^{*2})$

ii. $\phi^2 \leq \phi^{*2}$ for all T_N^*.

The BAN criteria is the asymptotic version of the minimum variance unbiased small sample criteria (why?). Except in situations where exact small sample results are available, asymptotic efficiency is the customary criteria of choice in evaluating estimators.

III. Review Basics of Asymptotic Theory: LLN, CLT, Slutsky Theorems, and Delta-Method

Economics 7720 J. Pepper
Spring 2024

The Multivariate Delta Method

We are often interested in learning the limiting distribution for functions of sample means. The least squares slope from a linear regression, for example, is a ratio of various sample means. To derive the limiting distribution of sample means, we have:

Central Limit Theorem:

- Univariate: In Random Sampling from any population with E(y) =μ and and V(y) = σ^2, $\sqrt{N}(\bar{y}-\mu) \rightarrow_d N(0,\sigma^2)$.
- Multivariate: In Random Sampling from any multivariate population with E(y) =μ and finite positive definite variance V(y) = Σ, $\sqrt{N}(\bar{y}-\mu) \rightarrow_d N(0,\Sigma)$, where y and μ are kx1, and Σ is kxk.

The Delta Method:

- Univariate: if $\sqrt{N}(\bar{y}-\mu) \rightarrow_d N(0,\sigma^2)$ and $\theta_n = h(\bar{y})$ is continuoulsly differentialbe at μ , then $\sqrt{N}(\theta_n - h(\mu)) \rightarrow_d N(0, \frac{\partial h(\mu)}{\partial \bar{y}}^2 \sigma^2)$.
- Multivariate: if $\sqrt{N}(\bar{y}-\mu) \rightarrow_d N(0,\Sigma)$ and $\theta_n = h(\bar{y})$ is a set of J continuously differentialbe functions at μ (kx1), then $\sqrt{N}(\theta_n - h(\mu)) \rightarrow_d N(0, \frac{\partial h(\mu)}{\partial \bar{y}'} \Sigma \frac{\partial h(\mu)}{\partial \bar{y}})$.

Intuition: Mean value theorem gives us the following linear approximation:
$h(\bar{y}) = h(\mu) + \frac{\partial h(\lambda_n)}{\partial \bar{y}}(\bar{y}-\mu)$ where $\lambda_n \in [\mu, \bar{y}]$. Thus, $\sqrt{N}\,(h(\bar{y}) - h(\mu)) = \frac{\partial h(\lambda_n)}{\partial \bar{y}}\sqrt{N}(\bar{y}-\mu)$. Because $\bar{y} \rightarrow_p \mu, \lambda_N \rightarrow_p \mu.$ Hence, by continuity, $h'(\lambda_N)$ converges in probability to $h'(\mu)$.

Economics 7720 J. Pepper
Spring 2024

Review of Sampling Distributions

The basic classical description of an estimate (or statistic) is its sampling distribution; that is, its probability distribution under independent repetitions of the sampling process. The sampling distribution is determined by the form of the estimate and by the sampling process.

1. Estimation of the Sampling Distribution

The sampling distribution fully characterizes the estimate, but the sampling distribution is typically not known. In this case, inference on the sampling distribution is itself an auxiliary estimation problem, requiring maintained assumptions. Two general approximations are commonly used:

A. Asymptotic Normal approximations

Under maintained assumptions, suppose that $\sqrt{N}\ (b_N - \beta) \rightarrow N(0, V)$. Then informally use $N(\beta, V/n)$ as an approximate sampling distribution for b_N. Since (β, V) are unknown, use (b_N, V_N). Thus, act as if b_N is distributed $N(b_N, V_N/N)$.

Note that the first use of b_N treats this quantity as a random variable with a sampling distribution. The second use treats it as the numerical estimate obtained from the actual sample drawn. Separate notation for these two uses should be maintained, but typically is not because it is cumbersome.

B. Bootstrap approximations

Treat the empirical distribution of the data as if it is the population distribution, and compute the sampling distribution of b_N under the empirical distribution.

2. Measurement of Precision

This is simply a matter of selecting some summary measure of the spread of the sampling distribution to use as an index of precision. If the sampling distribution is normal, the standard deviation suffices as any other measure of spread can be generated from it. In general, one can use an interquantile range. It is conventional to use the .025 - .975 quantile range.

3. Confidence Intervals (interval estimates)

The confidence interval reveals the set of plausible values of the parameter b having observed b_N
A *confidence interval* may be obtained by specifying two quantiles of the sampling distribution. These two numbers measure the location of the distribution (e.g., through the midpoint of the interval) and spread, through the interquantile range. When using an asymptotic normal approximation, it is conventional to use a "centered" confidence interval $b_{Nj} \pm c\sigma_{jN}$, where c determines the desired quantiles. In particular, setting c = 1.96 yields the .025 and .975 quantiles; setting c = 1.64 yields the 0.05 and 0.95 quantiles.

4. Classical Test of Hypothesis

The question asked is: *If the hypothesis is correct, is the numerical value obtained for the estimate, b_n, improbable?* If the numerical value is deemed improbable, then the hypothesis is "rejected." Otherwise, we say the hypothesis is "not rejected," or "accepted."

A. Rejection Region

Improbability of the estimate is operationalized by setting a "critical region" A. The hypothesis is rejected if the numerical value falls in region A. The convention is to choose A such that, under the null hypothesis, the probability α_a of the estimate falling in the region A is some specified value α_o, termed the "size" or the "significance level" of the test. This is the probability of rejecting the null hypothesis when the hypothesis is correct (i.e., the probability of a Type I error).

Generally, many alternative regions satisfy this criterion. To select among them, one considers the probability of falling in a region when the hypothesis is not correct. For this to make sense, one must specify an "alternative hypothesis" to be the negation of the null hypothesis. Let β_a denote the probability the estimate falls in A under the alternative hypothesis (i.e., $(1-\beta_a)$ probability of a Type II error). Among those regions A that satisfy the size criterion, on wishes to find a region that minimizes κ_a. Such a region is said to have maximal "power," and this test statistic is referred to as the most powerful.

Typically, the rejection region A is defined as values outsides the centered interquantile range that defines the *confidence interval*. For example, consider a test on a single parameter, $H_o:\beta_j = \beta^o$ against the alternative that $H_A:\beta_j \neq \beta^o$. When using an asymptotic normal approximation, one rejects the null at the α_o percent ***significance level*** if β^o lies outsider of the confidence interval, $b_{Nj} \pm c\sigma_{jN}$, where c_α is the ***critical value*** that defines the $(1-\alpha_o)$ centered interquantile range. Equivalently, reject the null if $|z_o^{'}| = |\sqrt{N}(\frac{b_{jn}-\beta_o}{\sigma_{jn}})| \geq c_\alpha$ .For a test at the 5% significance level, we set the critical value equal to 1.96.

B. Pretesting

Researches often use sequential testing/estimation processes. An initial estimate is obtained, a hypothesis is conducted, and then a further estimate with a different specification may be obtained depending on the result of the hypothesis test. The process is repeated until the result of the hypothesis test leads one to stop. Such sequential processes are called ***pretesting*** procedures. In general, pretesting alters the sampling distribution of statistics calculated after the initial step in the process. The subsequent estimates are calculated only if the earlier tests yield results leading the researcher to estimate the next specification. So, sampling distribution must condition on this event.

C. Economic versus Statistical Significance

A common practice is to focus attention on the "t-statistic" -- the ratio of an estimate to its standard error. If this statistic is large, researchers claim the estimator is "significant," meaning statistically different than zero. Undoubtedly, it is desirable to know how reliable a coefficient estimate is (i.e., the standard error), but focusing on statistical significance, over economic significance, can be misleading. A coefficient estimate may be statistically different from zero (or whatever other null is of interest), while the difference is economically trivial. Or the difference may be insignificant (statistically) but economically important. Test statistics measure the estimated coefficient in standard error units which are not meaningful measures of economics importance.