Overview of some Basic Statistical Concepts

Presentation 2

Random variables

- A variable whose value is determined by the outcome of a chance experiment is called a **random variable** (r.v.).
- Random variables are usually denoted by the capital letters X, Y, Z, and so on, and the values taken by them are denoted by small letters x, y, z, and so on.

Random variables

- A random variable may be either **discrete** or **continuous.** A discrete r.v. takes on only a finite (or countably infinite) number of values.
- For example, in throwing two dice, each numbered 1 to 6, if we define the random variable *X* as the sum of the numbers showing on the dice, then *X* will take one of these values: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12. Hence it is a discrete random variable.

Random variables

- A continuous r.v., on the other hand, is one that can take on any value in some interval of values.

Random variables

- Let X be a discrete r.v. taking distinct values $x1$, $x2, \ldots, xn, \ldots$. Then the function

  $f(x) = P(X = xi)$ for $i = 1, 2, \ldots, n$,

  is called the **discrete probability density function** (PDF) of X, where $P(X = xi)$ means the probability that the discrete r.v. X takes the value of xi.

Expected values

- Usually we use definite values like mathematical expectation and variance to describe random variables.
- These values are determined by the Mathematical expectation operator *E(X)*.

Expected values

- In the case of discrete r.v. that takes values of $X_1, X_2 \ldots X_n$ with probability p_1, p_2, … p_n we define:
- The Variance is a measure for scattering of the r.v. values around the mathematical expectation

$$\mu_X = E(X) = \sum_{i=1}^{n} p_i X_i$$

$$Var(X) = \sigma_X^2 = \sum_{i=1}^{N} p_i [X_i - E(X)]^2 =$$

$$= E[X - E(X)]^2$$

Expected values

- If ***a*** and ***b*** are constants, X is a r.v., we can define the following important properties of the mathematical expectation operator:

$$E(aX + b) = aE(X) + b$$

$$E[(aX)^2] = a^2 E(X^2)$$

$$Var(aX + b) = a^2 Var(X)$$

Joint probability distributions

- Let denote the probability two r.v. to take definite values simultaneously with $\boldsymbol{p}_{ij}$.
- Important characteristics of the joint distributions are covariance and correlation.

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] =$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}(X_i - E(X))(Y_i - E(Y))$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Joint probability distributions

- There are some important properties of the mathematical expectation operator in respect of the joint distribution of r.v.

$$E(X+Y)=E(X)+E(Y)$$

$$Var(X+Y)=$$
$$=Var(X)+Var(Y)+2Cov(X,Y)$$

Independence and correlation

- In some cases the values that r.v. X takes does not depend on the values, that takes r.v. Y.
- We call such r.v. **independent r.v.**

Independence and correlation

$$If\ \ X\ \ and\ Y\ are\ independent, \quad E(XY)=E(X)E(Y)$$

$$If\ \ X\ \ and\ Y\ are\ independent, \quad Cov(X,Y)=0$$

$$If\ \ X\ \ and\ Y\ are\ independent, \quad Var(X+Y)=Var(X)+Var(Y)$$

- The independence predetermine the lack of correlation, but the opposite is not true.
- The covariance is a measure for linear dependence between two r.v.
- Examples 1,2

Estimation

- In the case of incomplete information (we have just a sample, not the whole population), we can calculate only approximate values of the r.v. characteristics, called estimators.
- The important thing is these estimators to be as close to the real values as possible.

Estimation

- As the estimators vary in accordance with the samples, we may consider them as a r.v.
- For the estimators we can define: probability distribution, mathematical expectation, variance, covariance etc.

Estimation (Example 3)

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

$$\overline{Cov}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

Estimators properties

- **Unbiasedness.** An estimator is said to be an unbiased estimator if the expected value of the estimator is equal to the *true value* of the parameter.

$$Bias = E(\hat{\beta}) - \beta = 0$$

Estimators properties

- **Minimum Variance.** An estimator is said to be a minimum-variance estimator if it's variance is smaller than or at most equal to the variance of any other estimator of that parameter.
- **Best Unbiased, or Efficient, Estimator.** If we have *unbiased* estimator and it's variance is smaller than or at most equal to the variance of any other unbiased estimator, then we have **minimum-variance unbiased**, or **best unbiased**, or **efficient,** estimator.

Estimators properties

- **Minimum Mean-Square-Error (MSE) Estimator.** The MSE of an estimator is defined as:

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

- There is a tradeoff involved—to obtain minimum variance you may have to accept some bias.

Estimators properties

- **Consistency**: Asymptotic property – the value of the estimator approaches the real value with the increase of the sample size

$$\Pr\left|\beta - \hat{\beta}\right| < \delta \to 1,\ \textit{for all}\ \delta > 0, \quad \textit{when}\ N \to \infty$$

Probability distributions

- The Normal distribution depends on two parameters: mathematical expectation and variance.

$$p(X = X_i) = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\left[-\frac{1}{2\sigma_X^2}(X_i - \mu_X)^2\right]}$$

$$\Pr(\mu_X - 1.96\sigma_X < X_i < \mu_X + 1.96\sigma_X) \approx .95$$

$$\Pr(\mu_X - 2.57\sigma_X < X_i < \mu_X + 2.57\sigma_X) \approx .99$$

Probability distributions

- Chi-square distribution: sum of the squares of *N* standard normal r.v.
- *N* determines the degrees of freedom of the distribution
- Is used for hypothesis testing about the variances of r.v. or estimators.

Probability distributions

- *t* distribution is used when, the variance of a r.v. **is unknown**.
- Let *X* is standard normally distributed r.v. and *Z* is chi-square r.v. with *N* degrees of freedom. Then we may define t distributed r.v. with N d.f. (Example 4).

$$\frac{X}{\sqrt{Z/N}} \approx t_N$$

Probability distributions

- Sometimes it is necessary to test hypotheses about two r.v. F distribution is one of the appropriate distributions in that respect. It has two parameters: the first is connected to the number of estimated parameters and the second is related to the degrees of freedom of the data.
- If X and Z are Chi-square r.v. with N_1 and N_2 degrees of freedom, the r.v. $(X/N_1)/(Z/N_2)$ is with F distribution with N_1 and N_2 degrees of freedom.

Hypothesis testing and confidence intervals

- The hypotheses that we test in Econometrics usually concern the parameters of the regression models.
- The confidence intervals give some probabilistic guarantee about the obtained results.
- Every time the real value of the parameter belongs or not to the c.i., but after repeated sampling we have on average belonging with the predefined confidence level.

Example 5 – mean

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$Var(\bar{X}) = \frac{\sigma_X^2}{N}$$

$$\Pr(\bar{X} - \delta \leq \mu_X \leq \bar{X} + \delta) = 1 - \alpha,$$

$$(\alpha \text{ is called level of significance})$$

$$\Pr(\bar{X} - 1.96 \frac{\sigma_x}{\sqrt{N}} \leq \mu_X \leq \bar{X} + 1.96 \frac{\sigma_x}{\sqrt{N}}) = .95$$