UGC-NET & SET Economics: Statistics and Econometrics

Prepared by Dr.Dubba Ranjith

1. Probability Theory: Concepts, Distributions, Moments, Central Limit Theorem

Basic concepts

- **Random experiment, sample space (S), event.** Events can be mutually exclusive, exhaustive, independent, equally likely.
- **Classical definition:** $P(A)$ = favourable outcomes / total equally likely outcomes.
- **Relative frequency (empirical) definition:** limit of relative frequency as trials $\to \infty$.
- **Axiomatic definition (Kolmogorov, 1933):** $P(A) \geq 0$; $P(S) = 1$; $P(A \cup B) = P(A) + P(B)$ for mutually exclusive events.
- **Addition theorem:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Multiplication theorem:** $P(A \cap B) = P(A) \cdot P(B|A)$; independent events: $P(A \cap B) = P(A) \cdot P(B)$.
- **Conditional probability:** $P(A|B) = P(A \cap B)/P(B)$.
- **Bayes' theorem:** $P(A_i|B) = P(B|A_i)P(A_i) / \Sigma P(B|A_j)P(A_j)$ — revises prior probabilities into posterior probabilities.

Random variables and expectation

- Discrete (probability mass function) vs continuous (probability density function); cumulative distribution function $F(x) = P(X \leq x)$.
- $E(X) = \Sigma x \cdot p(x)$; $Var(X) = E(X^2) - [E(X)]^2$.
- $E(aX + b) = aE(X) + b$; $Var(aX + b) = a^2Var(X)$.
- Covariance: $Cov(X, Y) = E(XY) - E(X)E(Y)$; zero for independent variables (converse not necessarily true).

Important distributions

Distribution	Type	Mean	Variance	Remarks
Bernoulli	Discrete	p	pq	Single trial
Binomial B(n, p)	Discrete	np	npq	n independent trials; mean > variance
Poisson (λ)	Discrete	λ	λ	Rare events; limit of binomial ($n \to \infty$, $p \to 0$); mean = variance
Uniform (a, b)	Continuous	$(a+b)/2$	$(b-a)^2/12$	
Normal $N(\mu, \sigma^2)$	Continuous	μ	σ^2	Bell-shaped, symmetric; mean = median = mode
Exponential (λ)	Continuous	$1/\lambda$	$1/\lambda^2$	Waiting times; memoryless
Chi-square $\chi^2(k)$	Continuous	k	$2k$	Sum of squares of k standard normals
Student's t (k)	Continuous	0	$k/(k-2)$	Heavier tails; tends to normal as k grows
F (k_1, k_2)	Continuous	$k_2/(k_2-2)$	—	Ratio of two independent chi-squares divided by their df

- **Normal distribution:** about 68.27% of values within $\mu \pm 1\sigma$, 95.45% within $\pm 2\sigma$, 99.73% within $\pm 3\sigma$. Standard normal $Z = (X - \mu)/\sigma$.

Moments

- **Raw moments** $\mu'_r = E(X^r)$; **central moments** $\mu_r = E(X - \mu)^r$. $\mu_1 = 0$, μ_2 = variance.
- **Skewness:** $\beta_1 = \mu_3^2/\mu_2^3$ ($\gamma_1 = \sqrt{\beta_1}$). Karl Pearson's coefficient = (Mean − Mode)/σ. Positive skew: mean > median > mode.
- **Kurtosis:** $\beta_2 = \mu_4/\mu_2^2$. Normal (mesokurtic) $\beta_2 = 3$; leptokurtic > 3; platykurtic < 3.
- Moment generating function $M(t) = E(e^{tx})$.

Laws of large numbers and Central Limit Theorem

- **Chebyshev's inequality:** $P(|X - \mu| \geq k\sigma) \leq 1/k^2$.
- **Law of Large Numbers:** sample mean converges to population mean as $n \to \infty$.
- **Central Limit Theorem (Lindeberg–Lévy):** the distribution of the mean of n independent, identically distributed variables with finite variance approaches normal $N(\mu, \sigma^2/n)$ as n becomes large, whatever the parent distribution. Justifies use of normal tests for large samples ($n \geq 30$).

Exam pointers: Bayes' theorem; Poisson mean = variance; normal area rules; $\beta_2 = 3$ for normal; CLT statement.

2. Descriptive Statistics: Central Tendency, Dispersion, Correlation, Index Numbers

Measures of central tendency

Measure	Formula / idea	Merits and uses
Arithmetic mean	$\Sigma x/n$	Uses all values; affected by extremes
Median	Middle value	Not affected by extremes; open-ended classes
Mode	Most frequent value; grouped: $L + (f_1-f_0)/(2f_1-f_0-f_2) \times h$	Qualitative data, fashion
Geometric mean	$(x_1 \cdot x_2 \ldots x_n)^{\wedge}(1/n)$	Growth rates, ratios, index numbers
Harmonic mean	$n / \Sigma(1/x)$	Average speeds, rates

- **AM ≥ GM ≥ HM** (equal only if all values are equal); for two values, $GM^2 = AM \times HM$.
- Empirical relation (moderately skewed): Mode = 3 Median − 2 Mean.
- Partition values: quartiles, deciles, percentiles.

Measures of dispersion

- **Absolute:** range, quartile deviation $(Q_3 - Q_1)/2$, mean deviation, standard deviation $\sigma = \sqrt{[\Sigma(x - \bar{x})^2/n]}$.
- **Relative:** coefficient of range, coefficient of QD, **coefficient of variation CV = (σ/x̄) × 100** (compare variability).
- Variance is independent of change of origin but not of scale.
- **Lorenz curve and Gini coefficient** measure inequality (Gini = area between line of equality and Lorenz curve ÷ total area under line of equality; 0 to 1).

Correlation

- **Karl Pearson's coefficient:** $r = Cov(X, Y)/(\sigma x \cdot \sigma y)$; $-1 \le r \le +1$; independent of change of origin and scale.
- **Spearman's rank correlation:** $R = 1 - 6\Sigma d^2/[n(n^2 - 1)]$.
- **Coefficient of determination r^2:** share of variation in Y explained by X.
- Regression coefficients: $r = \pm\sqrt{(byx \cdot bxy)}$; both regression lines pass through $(\bar{x}, \bar{y})$.
- Partial and multiple correlation; spurious correlation; probable error of $r = 0.6745(1 - r^2)/\sqrt{n}$.

Index numbers

Index	Formula (weights)	Bias
Laspeyres	$\Sigma p_1 q_0 / \Sigma p_0 q_0 \times 100$ (base-year quantities)	Upward bias
Paasche	$\Sigma p_1 q_1 / \Sigma p_0 q_1 \times 100$ (current-year quantities)	Downward bias
Fisher's ideal	√(Laspeyres × Paasche)	Satisfies time-reversal and factor-reversal tests
Marshall–Edgeworth	$\Sigma p_1(q_0+q_1) / \Sigma p_0(q_0+q_1) \times 100$	Satisfies time-reversal
Dorbish–Bowley	(L + P)/2	

- **Tests of adequacy (Fisher):** unit test, time-reversal test, factor-reversal test, circular test.
- Chain-base vs fixed-base indices; splicing; deflating (real wages = money wages/price index × 100).
- Indian indices: WPI (Office of Economic Adviser, DPIIT), CPI-Combined, CPI-IW, CPI-AL/RL (Labour Bureau), Index of Industrial Production (NSO).

Exam pointers: AM ≥ GM ≥ HM; CV formula; Spearman formula; Fisher's ideal index tests; Laspeyres vs Paasche bias.

3. Sampling Methods and Sampling Distribution

Census vs sample

- Census: every unit studied — accurate but costly. Sample: subset; cheaper, faster, allows deeper study.
- **Sampling errors** (from studying a part) decrease with sample size; **non-sampling errors** (measurement, non-response, processing) can occur in both.

Methods of sampling

Probability (random) sampling	Non-probability sampling
Simple random (lottery, random numbers; with or without replacement)	Purposive / judgement
Stratified (homogeneous strata; proportional or Neyman optimum allocation)	Convenience
Systematic (every kth unit)	Quota
Cluster (groups selected randomly)	Snowball
Multi-stage (e.g., NSS: villages → households)	

- Stratified sampling gives more precise estimates when strata are internally homogeneous.

Sampling distribution

- Distribution of a statistic (mean, proportion) over all possible samples.
- **Standard error of mean:** SE = $\sigma/\sqrt{n}$ (infinite population); with finite population correction $\sqrt{[(N - n)/(N - 1)]}$.
- **SE of proportion:** $\sqrt{(pq/n)}$.
- Sampling distributions of t, χ^2 and F used for small samples.
- **Properties of estimators:** unbiasedness ($E(\hat{\theta}) = \theta$), efficiency (minimum variance), consistency ($\hat{\theta} \to \theta$ as $n \to \infty$), sufficiency.
- Sample variance with divisor $(n - 1)$ is unbiased.

Exam pointers: types of random sampling; standard error formula; properties of estimators; sampling vs non-sampling errors.

4. Statistical Inference and Hypothesis Testing

Estimation

- **Point estimation:** methods — moments, maximum likelihood (Fisher), least squares.
- **Interval estimation:** 95% confidence interval for mean = $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$ (large sample); $\bar{x} \pm t(\alpha/2, n-1) \cdot s/\sqrt{n}$ (small sample).

Hypothesis testing procedure

1. State null hypothesis H_0 and alternative H_1 (one-tailed or two-tailed).
2. Choose level of significance α (commonly 5% or 1%).
3. Compute test statistic.
4. Compare with critical value (or p-value with α).
5. Reject or do not reject H_0.

	H_0 true	**H_0 false**
Reject H_0	**Type I error** (α)	Correct decision (Power = $1 - \beta$)
Accept H_0	Correct decision	**Type II error** (β)

- Critical Z values: two-tailed 1.96 (5%), 2.58 (1%); one-tailed 1.645 (5%), 2.33 (1%).
- **Neyman–Pearson lemma:** most powerful test for given α.

Test	Use
Z-test	Large samples; mean or proportion with known σ
t-test (W.S. Gosset, "Student", 1908)	Small samples ($n < 30$); difference of means; paired t; significance of regression coefficients
χ^2-test (Karl Pearson)	Goodness of fit; independence of attributes (contingency tables); population variance
F-test (R.A. Fisher)	Equality of two variances; ANOVA; overall significance of regression
ANOVA	Comparing means of more than two groups (one-way, two-way)
Non-parametric	Sign test, Mann–Whitney U, Wilcoxon signed-rank, Kruskal–Wallis, run test

Common tests

- Degrees of freedom: χ^2 in a contingency table = $(r - 1)(c - 1)$.

Exam pointers: Type I vs Type II error; power of test; critical values; which test for which situation; χ^2 degrees of freedom.

5. Linear Regression Models and their Properties — BLUE

Classical Linear Regression Model (CLRM)

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- **OLS estimators:** $\hat{\beta}_2 = \Sigma x_i y_i / \Sigma x_i^2$ (deviations from means); $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$.
- OLS minimises the residual sum of squares $\Sigma \hat{u}_i^2$.
- Matrix form: $\hat{\beta} = (X'X)^{-1}X'Y$.

Assumptions of CLRM

1. Linear in parameters.
2. X values fixed in repeated sampling (non-stochastic) or independent of u.
3. Zero mean of disturbance: $E(u_i) = 0$.
4. **Homoscedasticity:** $\text{Var}(u_i) = \sigma^2$ constant.
5. **No autocorrelation:** $\text{Cov}(u_i, u_j) = 0$.
6. n > number of parameters; variation in X.
7. **No perfect multicollinearity** among regressors.
8. Correct specification.
9. (For inference) $u_i \sim$ Normal.

Gauss–Markov theorem

Under assumptions 1–8, OLS estimators are **BLUE — Best Linear Unbiased Estimators**: among all linear unbiased estimators they have the minimum variance. Normality is not needed for BLUE; it is needed for t and F tests. With normality, OLS = maximum likelihood estimators.

Goodness of fit and tests

- **$R^2 = ESS/TSS = 1 - RSS/TSS$** (TSS = ESS + RSS); lies between 0 and 1.
- **Adjusted $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k)$**; can be negative; penalises extra regressors.
- t-test for individual coefficients: $t = \beta/SE(\beta)$.
- F-test for overall significance: $F = [R^2/(k - 1)] / [(1 - R^2)/(n - k)]$.
- Information criteria: AIC, SIC (Schwarz).

Violations of assumptions

Problem	Consequence for OLS	Detection	Remedy
Multicollinearity	Still BLUE, but large variances; high R^2 with few significant t's	VIF > 10, correlation matrix, Farrar–Glauber	Drop variable, more data, ridge regression, principal components
Heteroscedasticity	Unbiased but not efficient; SEs biased	Park, Glejser, Goldfeld–Quandt, Breusch–Pagan, White test	WLS / GLS, White robust SEs, log transformation
Autocorrelation	Unbiased but not efficient; t and F invalid	**Durbin–Watson d** (≈ 2 means no autocorrelation; range 0–4), Breusch–Godfrey LM test, Durbin h (lagged dependent variable)	GLS, Cochrane–Orcutt, Newey–West SEs, first differencing
Specification error	Omitting relevant variable → biased and inconsistent; including irrelevant → unbiased but inefficient	Ramsey RESET	Correct model
Errors in variables (in X)	Biased and inconsistent	Hausman test	Instrumental variables

- **Dummy variables:** for qualitative regressors; with m categories use m − 1 dummies (dummy variable trap otherwise). **Chow test** for structural break.
- Functional forms: log-linear (coefficient = elasticity), semi-log (growth rate), reciprocal, polynomial.
- Distributed lag (Koyck) and autoregressive models.

Exam pointers: CLRM assumptions; Gauss–Markov (BLUE); R^2 vs adjusted R^2; DW range and interpretation; consequences of each violation.

6. Identification Problem

In a simultaneous equation system, an equation is **identified** if its structural parameters can be uniquely derived from the reduced-form coefficients.

- **Under-identified:** structural parameters cannot be estimated.
- **Exactly identified:** unique values — estimated by **Indirect Least Squares (ILS)**.
- **Over-identified:** more than one value — estimated by **2SLS**.

Rules (for an equation in a model with M endogenous variables and K predetermined variables)

- m = endogenous variables in the equation; k = predetermined variables in the equation.
- **Order condition (necessary):** $K - k \geq m - 1$.
  - $K - k = m - 1 \rightarrow$ exactly identified.
  - $K - k > m - 1 \rightarrow$ over-identified.
  - $K - k < m - 1 \rightarrow$ under-identified.
- **Rank condition (necessary and sufficient):** at least one non-zero determinant of order $(M - 1)(M - 1)$ can be formed from the coefficients of variables excluded from the equation but included in other equations.

Example: demand–supply model

- Demand: $Q = \alpha_0 + \alpha_1 P + u_1$; Supply: $Q = \beta_0 + \beta_1 P + u_2$ — both under-identified (only equilibrium points observed).
- Adding income (Y) to demand identifies the supply equation; adding weather (W) to supply identifies demand.
- Historical: Working (1927), "What do statistical demand curves show?".

Exam pointers: order condition formula; rank condition is necessary and sufficient; ILS for exact, 2SLS for over-identified.

7. Simultaneous Equation Models — Recursive and Non-Recursive

Nature

- Two-way causation: endogenous variables appear as regressors in other equations (e.g., Keynesian model: $C = a + bY + u$; $Y = C + I$).
- **Simultaneity bias:** OLS estimates are **biased and inconsistent** because the endogenous regressor is correlated with the error term.
- **Structural form** vs **reduced form** (endogenous variables expressed only in terms of predetermined variables).
- **Hausman specification test** checks for simultaneity / endogeneity.

Recursive (triangular) models

- Causation runs one way: Y_1 depends on exogenous variables only; Y_2 depends on Y_1 and exogenous variables; Y_3 on Y_1, Y_2 … The coefficient matrix of endogenous variables is triangular and error covariances are zero (diagonal).
- **OLS can be applied equation by equation** and gives consistent estimates.
- Associated with Herman Wold's causal chain models.

Non-recursive (interdependent) models

- Feedback among endogenous variables. OLS is inconsistent.

Estimation methods

Method	Type	Use
Indirect Least Squares (ILS)	Single equation	Exactly identified equations
Two-Stage Least Squares (2SLS) — Theil, Basmann	Single equation	Over-identified (also exactly identified); replace endogenous regressors by fitted values from reduced form
Instrumental Variables (IV)	Single equation	Instrument correlated with regressor, uncorrelated with error
LIML	Single equation	Limited information maximum likelihood
Three-Stage Least Squares (3SLS) — Zellner & Theil	System	Uses cross-equation error correlation
FIML	System	Full information maximum likelihood

Exam pointers: OLS inconsistent in non-recursive systems but valid in recursive ones; 2SLS steps; system vs single-equation methods.

8. Discrete Choice Models

Used when the dependent variable is qualitative (e.g., buy/not buy, work/not work).

Binary choice models

Model	Form	Features
Linear Probability Model (LPM)	$Y_i = \beta_1 + \beta_2 X_i + u_i$, Y = 0/1, estimated by OLS	Predicted probabilities may lie outside 0–1; heteroscedastic errors; non-normal errors; low R^2
Logit	$P_i = 1/(1 + e^{-(\beta_1+\beta_2 X_i)})$; log-odds $\ln[P/(1-P)] = \beta_1 + \beta_2 X$	Logistic CDF; coefficient = change in log-odds; estimated by maximum likelihood
Probit (normit)	$P_i = \Phi(\beta_1 + \beta_2 X_i)$, standard normal CDF	Similar results to logit; logit coefficients ≈ 1.81 × probit coefficients

- Goodness of fit: pseudo-R^2 (McFadden), count R^2, likelihood ratio test.
- Marginal effects are not constant; depend on X values.

Other models

- **Multinomial logit/probit:** more than two unordered choices (transport mode).
- **Ordered logit/probit:** ranked choices (ratings).
- **Tobit model (James Tobin, 1958):** censored dependent variable (e.g., expenditure on durables, zero for many).
- **Heckman two-step selection model:** sample selection bias.
- **Poisson and negative binomial regression:** count data.

Exam pointers: drawbacks of LPM; logit = logistic, probit = normal CDF; Tobit for censored data; estimation by MLE.

9. Time Series Analysis

Components of a time series

- **Secular trend (T), seasonal variation (S), cyclical variation (C), irregular (random) variation (I).**
- Models: additive ($Y = T + S + C + I$) and multiplicative ($Y = T \times S \times C \times I$).
- Trend measured by freehand, semi-averages, moving averages, least squares; seasonal indices by ratio-to-trend, ratio-to-moving-average, link relatives.

Stationarity

- A series is (weakly) **stationary** if its mean, variance and autocovariance do not change over time.
- Non-stationary series → **spurious regression** (Granger & Newbold, 1974): high R^2 but low DW (rule of thumb: $R^2 > d$).
- **Random walk:** $Y_t = Y_{t-1} + u_t$ (unit root); with drift: $Y_t = \delta + Y_{t-1} + u_t$.
- **Order of integration:** I(d) — needs d differencings to become stationary.

Tests

- Correlogram (ACF, PACF); Ljung–Box Q statistic.
- **Unit root tests:** Dickey–Fuller (DF), **Augmented Dickey–Fuller (ADF)**, Phillips–Perron, KPSS (null = stationarity).
- **Cointegration (Engle & Granger, 1987):** a linear combination of I(1) series is I(0) — long-run equilibrium relationship. Engle–Granger two-step test, **Johansen** test (multiple vectors).
- **Error Correction Model (ECM):** short-run dynamics with adjustment towards long-run equilibrium (Granger representation theorem).
- **Granger causality test:** whether past values of X help predict Y.

Forecasting models

Model	Description
AR(p)	Y_t depends on its own p lagged values
MA(q)	Y_t depends on current and q lagged error terms
ARMA(p, q)	Combination, for stationary series
ARIMA(p, d, q)	Differenced d times; **Box–Jenkins methodology:** identification → estimation → diagnostic checking → forecasting
VAR	System of equations, each variable regressed on lags of all (Sims, 1980); impulse response functions
ARCH / GARCH	Time-varying volatility (Engle, 1982; Bollerslev, 1986) — financial series

- Exponential smoothing (Holt–Winters).
- Nobel Prize 2003: Robert Engle (ARCH) and Clive Granger (cointegration).

Exam pointers: components of time series; unit root and ADF test; cointegration and ECM; Box–Jenkins steps; ARCH/GARCH authors.