

VU LEARNING HUB

vulearninghub.xyz

HIGHLIGHTED HANDOUT

STA6310

Probability and Probability
Distributions-II

STA6310 by VU Learning HUB

Prepared for VU students • Free study resources

<https://vulearninghub.xyz>

STA6310

Probability and Probability Distributions-II

Lecture Handouts

Prepared for the students of Virtual University of Pakistan

Table of Contents

Placeholder for table of contents

0

1. Random Variables: A Review

LEARNING OBJECTIVES

- Distinguish discrete and continuous random variables
- Use pmf, pdf and cdf correctly
- Compute probabilities from distributions

A **random variable** X is a real-valued function on the sample space. **Discrete** X takes countable values with a **probability mass function** $p(x) = P(X = x)$, where $p(x) \geq 0$ and the sum over all x equals 1. **Continuous** X has a **probability density function** $f(x) \geq 0$ with total integral 1; probabilities are areas: $P(a < X < b)$ is the integral of f from a to b , and $P(X = x) = 0$ at any single point.

$$F(x) = P(X \leq x): \text{discrete: } F(x) = \sum p(t), t \leq x \mid \text{continuous: } F(x) = \int_{-\infty}^x f(t) dt$$

The **cumulative distribution function** $F(x) = P(X \leq x)$ works for both types: it is non-decreasing, right-continuous, tends to 0 at minus infinity and 1 at plus infinity. From it: $P(a < X \leq b) = F(b) - F(a)$. The **quantile** (e.g. median) inverts F . Mixed distributions (part discrete, part continuous) occur in insurance and reliability - handle each part with its own tool.

SECTION SUMMARY

pmf for countable values (sums), pdf for continua (integrals), cdf $F(x) = P(X \leq x)$ for everything. $P(X = x) = 0$ for continuous variables - always.

2. Mathematical Expectation

LEARNING OBJECTIVES

- Compute $E(X)$ for discrete and continuous variables
- Apply linearity of expectation
- Find variance and standard deviation

$$E(X) = \sum x \cdot p(x) \text{ (discrete)} \mid E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx \text{ (continuous)}$$

Expectation is the probability-weighted average - the long-run mean. **Linearity**: $E(aX + bY + c) = aE(X) + bE(Y) + c$, requiring NO independence assumption - the most-used theorem in probability. For functions: $E[g(X)] = \text{sum/integral of } g(x) \text{ times } p(x)/f(x)$ (the law of the unconscious statistician).

$$\text{Var}(X) = E[(X - \mu)^2] = E(X^2) - [E(X)]^2, \sigma = \sqrt{\text{Var}(X)}$$

Variance measures spread around the mean $\mu = E(X)$. The computational formula $E(X^2) - \mu^2$ is usually easier. Properties: $\text{Var}(aX + b) = a^2 \text{Var}(X)$ -

constants shift but do not spread; for INDEPENDENT variables, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

Standard deviation sigma restores original units. **Chebyshev's inequality**: $P(|X - \mu| \geq k \sigma) \leq 1/k^2$ - a universal tail bound needing only mean and variance.

■ VERY IMPORTANT

Linearity of E needs no independence; additivity of Var needs independence (or zero covariance).
Confusing these is the classic exam trap.

SECTION SUMMARY

$E(X)$ is the weighted mean; linearity always holds. $\text{Var} = E(X^2) - \mu^2$; $\text{SD} = \sqrt{\text{Var}}$.
Chebyshev bounds tail probabilities from just two moments.

3. Moment Generating Functions

LEARNING OBJECTIVES

- Define the mgf and derive moments from it
- Use mgfs to identify distributions
- State the uniqueness theorem

$$M(t) = E(e^{tX}) = \sum e^{tx} \cdot p(x) \text{ or } \int e^{tx} \cdot f(x) dx, M^{(k)}(0) = E(X^k)$$

The **moment generating function** $M(t) = E(e^{tX})$ encodes ALL moments: differentiate k times and set $t = 0$ to get $E(X^k)$. Example: for exponential(rate λ), $M(t) = \lambda/(\lambda - t)$, giving $E(X) = 1/\lambda$, $\text{Var} = 1/\lambda^2$. The **uniqueness theorem**: if two variables share an mgf (near 0), they share the distribution - so mgfs **identify** distributions and prove sums: the mgf of independent $X + Y$ is the **product** $M_X(t)$ times $M_Y(t)$.

Cautions: some distributions (e.g. lognormal, Cauchy) have **no mgf** (the integral diverges) - then use the **characteristic function** $\phi(t) = E(e^{itX})$, which always exists. Related: the **probability generating function** $G(s) = E(s^X)$ for non-negative integer variables, handy for branching processes. Cumulants (from $\log M$) give skewness and kurtosis cleanly.

SECTION SUMMARY

$M(t) = E(e^{tX})$; derivatives at zero give moments; products of mgfs give sums of independents; uniqueness identifies distributions.

4. Bernoulli, Binomial and Poisson

LEARNING OBJECTIVES

- Model Bernoulli trials
- Apply the binomial distribution
- Use the Poisson as a limit and a model

$$\text{Binomial}(n, p): P(X = k) = C(n, k) \cdot p^k \cdot (1-p)^{n-k}, E(X) = np, \text{Var}(X) = np(1-p)$$

A **Bernoulli(p)** trial is one success/failure experiment ($E = p$, $\text{Var} = p(1-p)$). The **Binomial(n, p)** counts successes in n INDEPENDENT identical trials: quality-control defectives, survey yes-counts. Requirements: fixed n , constant p , independence - violating independence (e.g. sampling without replacement from small populations) breaks it.

$$\text{Poisson}(\lambda): P(X = k) = e^{-\lambda} \cdot \lambda^k / k!, E(X) = \text{Var}(X) = \lambda$$

The **Poisson(lambda)** counts events in time/space at constant rate: arrivals, defects per meter, goals per match. **Poisson limit theorem**: Binomial(n, p) tends to Poisson(np) as n grows and p shrinks - use it when $n \geq 100$, $p \leq 0.01$. Mean equals variance (λ) - a quick diagnostic: real data with variance far above the mean is **overdispersed** (try negative binomial). Poisson processes have **memoryless** exponential inter-arrival times - the bridge to Section 6.

SECTION SUMMARY

Bernoulli = one trial; Binomial(n, p) = successes in n trials (np , $np(1-p)$); Poisson(λ) = event counts (mean = variance = λ), the rare-event limit of the binomial.

5. Geometric, Negative Binomial and Hypergeometric**LEARNING OBJECTIVES**

- Model waiting times with geometric/NB
- Apply hypergeometric sampling
- Distinguish with/without replacement

$$\text{Geometric}(p): P(X = k) = (1-p)^{k-1} \cdot p, E(X) = 1/p \mid \text{NB}(r, p): E(X) = r/p$$

The **Geometric(p)** counts trials UNTIL the first success ($E = 1/p$) - it is **memoryless**: $P(X > m + n \mid X > m) = P(X > n)$ - the die has no memory. The **Negative Binomial(r, p)** counts trials until the r -th success - sum of r independent geometrics, so $E = r/p$. Use for 'how long until k events?' questions.

$$\text{Hypergeometric}(N, K, n): P(X = k) = C(K, k) \cdot C(N-K, n-k) / C(N, n)$$

The **Hypergeometric** counts successes sampling n from a finite population of N containing K successes, WITHOUT replacement: committee selections, acceptance sampling, card hands.

Mean = nK/N (same as binomial); variance is binomial variance times the **finite population correction** $(N - n)/(N - 1)$. Rule of thumb: use hypergeometric when $n/N > 5\%$; otherwise binomial approximates it.

SECTION SUMMARY

Geometric/NB = waiting-time distributions (memoryless trials); Hypergeometric = binomial without replacement, with the finite-population correction shrinking variance.

6. Uniform and Exponential Distributions

LEARNING OBJECTIVES

- Use the continuous uniform distribution
- Apply the exponential and its memoryless property
- Relate exponential to Poisson processes

Uniform(a, b): $f(x) = 1/(b-a)$, $E = (a+b)/2$, $\text{Var} = (b-a)^2/12$ | Exp(λ): $f(x) = \lambda e^{(-\lambda x)}$, $E = 1/\lambda$

The **Uniform(a, b)** spreads probability evenly - random number generators, quantization error, 'arrive at a random time' problems. The **Exponential(λ)** models waiting times: bus arrivals, component lifetimes, call-center gaps. Its **memoryless property** $P(X > s + t | X > s) = P(X > t)$ is UNIQUE among continuous distributions - 'how long you have waited tells nothing.'

Exponential $\leftrightarrow$ Poisson duality: if arrivals follow Poisson(λ) per unit time, inter-arrival times are Exponential(λ). **Hazard rate** λ is constant - 'as good as new' - which is why the exponential suits electronics but NOT human lifetimes or wearing parts (use Weibull/gamma there). Minimum of independent exponentials is exponential with summed rates - the competing-risks result.

■ VERY IMPORTANT

Memoryless = only geometric (discrete) and exponential (continuous). If an exam asks 'which continuous distribution is memoryless?', the answer is always exponential.

SECTION SUMMARY

Uniform(a, b): flat density. Exponential(λ): waiting times, uniquely memoryless, dual to the Poisson; constant hazard - fine for components, wrong for aging.

7. The Normal Distribution

LEARNING OBJECTIVES

- Work with $N(\mu, \sigma^2)$
- Standardize and use normal tables
- Apply the normal approximation to the binomial

$$f(x) = (1 / (\sigma\sqrt{2\pi})) \cdot e^{-(x-\mu)^2 / (2\sigma^2)}, Z = (X - \mu) / \sigma \sim N(0, 1)$$

The **Normal** $N(\mu, \sigma^2)$ is the bell curve of measurement error, heights, and sample means. **Standardization** $Z = (X - \mu)/\sigma$ converts any normal probability to the standard $N(0,1)$ table: $P(X \leq x) = \Phi((x - \mu)/\sigma)$. The **empirical rule**: 68% within 1 sigma, 95% within 2, 99.7% within 3. Symmetry gives $\Phi(-z) = 1 - \Phi(z)$.

Normal approximation to Binomial(n, p): valid when $np \geq 10$ and $n(1-p) \geq 10$ - use the **continuity correction**: $P(X \leq k) \approx \Phi((k + 0.5 - np)/\sqrt{np(1-p)})$. Linear combinations of independent normals are normal; sums of normals stay normal - the closure property behind regression theory. **Lognormal**: if $\log X$ is normal, X models incomes, stock prices - right-skewed, positive-only.

SECTION SUMMARY

Standardize via $Z = (X - \mu)/\sigma$; 68-95-99.7 rule; continuity-corrected normal approximates binomial for large n ; linear combinations of normals stay normal.

8. Gamma, Beta, Chi-Square, t and F**LEARNING OBJECTIVES**

- Define the gamma family and its members
- Use chi-square, t and F in inference
- Relate beta to order statistics

$$\text{Gamma}(\alpha, \lambda): f(x) = \lambda^\alpha x^{\alpha-1} e^{-\lambda x} / \Gamma(\alpha), E = \alpha/\lambda \mid \chi^2(\nu) = \text{Gamma}(\nu/2, 1/2)$$

The **Gamma(alpha, lambda)** generalizes the exponential ($\alpha = 1$): sums of α iid exponentials - total waiting time for α events, insurance claims, rainfall. $E = \alpha/\lambda$, $\text{Var} = \alpha/\lambda^2$. **Chi-square with nu df** is $\text{Gamma}(\nu/2, 1/2)$: the distribution of sums of squared standard normals - hence of the **sample variance**: $(n-1)S^2/\sigma^2$ follows $\chi^2(n-1)$.

Student's t (nu): $Z/\sqrt{V/\nu}$ - used when σ is estimated: confidence intervals and tests for means with small samples; approaches $N(0,1)$ as ν grows (use z beyond ~ 30 df). **F(d1, d2)**: ratio of scaled chi-squares - ANOVA and variance-ratio tests. **Beta(a, b)** on $(0,1)$: models proportions and probabilities; the k -th **order statistic** of n uniforms is $\text{Beta}(k, n+1-k)$ - linking

beta to minima/maxima.

SECTION SUMMARY

Gamma generalizes exponential; chi-square(ν) governs sample variance; t handles estimated sigma; F compares variances; Beta lives on (0,1) and describes uniform order statistics.

9. Transformations of Random Variables

LEARNING OBJECTIVES

- Find distributions of transformed variables
- Apply the Jacobian method
- Handle min/max and sums

$$Y = g(X), g \text{ monotone: } f_Y(y) = f_X(g^{-1}(y)) \cdot |d/dy g^{-1}(y)|$$

For **monotone** transformations, the change-of-variables formula multiplies by the Jacobian $|dx/dy|$: e.g. $X \sim \text{uniform}(0,1)$, $Y = -\ln(X)/\lambda$ gives exponential - the **inverse-transform** basis of random-number generation. For non-monotone g (e.g. $Y = X^2$), split into monotone pieces and sum. The **cdf method** (find F_Y , differentiate) is the reliable fallback.

Multivariate: $f_{\{U,V\}}(u,v) = f_{\{X,Y\}}(x,y)$ times $|J|$ (Jacobian determinant) - used to derive t and F distributions. **Min/max**: for iid variables, $F_{\max}(x) = [F(x)]^n$ and $F_{\min}(x) = 1 - [1 - F(x)]^n$ - e.g. system reliability: a series system fails at the minimum component life. **Sums**: convolution for independents; mgf products shortcut it (Section 3).

SECTION SUMMARY

Transform via Jacobian (or cdf method); multivariate via Jacobian determinant; max/min from powers of F; sums of independents via convolution or mgf products.

10. Joint, Marginal and Conditional Distributions

LEARNING OBJECTIVES

- Work with joint pmf/pdf
- Derive marginals and conditionals
- Test independence

$$\text{Marginal: } f_X(x) = \int f(x,y) dy \mid \text{Conditional: } f(y|x) = f(x,y) / f_X(x)$$

The **joint** pmf/pdf $f(x, y)$ describes pairs: probabilities are sums/double integrals over regions. **Marginals** come from summing/integrating out the other variable. **Conditional** distributions $f(y|x)$ = joint/marginal renormalize given $X = x$ - the basis of regression ($E[Y|x]$) and Bayesian updating. **Independence** iff the joint factors: $f(x,y) = f_X(x)$ times $f_Y(y)$ for all x, y - equivalently conditionals equal marginals.

Example: multinomial($n; p_1..p_k$) for categorized counts (die faces, survey responses) - marginals are binomial, conditionals reduce dimension. **Bivariate transformations** and **hierarchical models** ($X|\theta$ then θ) build real-world models: e.g. accident counts Poisson(λ) with λ gamma-distributed gives the negative binomial - overdispersion explained.

SECTION SUMMARY

Joint describes pairs; marginalize by summing out; condition by renormalizing; independence = factorization. Multinomial and hierarchical models are the workhorses.

11. Covariance, Correlation and Bivariate Normal

LEARNING OBJECTIVES

- Compute covariance and correlation
- Interpret correlation correctly
- Use the bivariate normal

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E(XY) - \mu_x\mu_y, \rho = \text{Cov}(X,Y) / (\sigma_x\sigma_y), -1 \leq \rho \leq 1$$

Covariance measures joint linear variation; **correlation** ρ standardizes it to $[-1, 1]$ - unit-free and comparable. Properties: independent variables have zero covariance (converse FALSE - zero correlation does not imply independence, except in the normal family); $|\rho| = 1$ means exact linear relation. **Correlation is not causation** - confounding and coincidence abound; always the exam's conceptual checkpoint. ☰

The **bivariate normal** is parameterized by $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$: marginals normal, conditionals normal with **regression line** $E[Y|x] = \mu_y + \rho(\sigma_y/\sigma_x)(x - \mu_x)$ - the origin of linear regression. Zero correlation DOES imply independence here - the family's special property. Contours are ellipses; ρ rotates them.

SECTION SUMMARY

$\text{Cov} = E(XY) - \mu_x \mu_y$; ρ in $[-1,1]$; independence implies zero covariance, not conversely (except normal). Bivariate normal gives linear regression as conditional expectation.

12. Sampling Distributions and Order Statistics

LEARNING OBJECTIVES

- Derive the distribution of the sample mean
- Work with the t, chi-square and F sampling results
- Use order statistics

$$\bar{X} \sim \text{approx } N(\mu, \sigma^2/n); (n-1)S^2/\sigma^2 \sim \chi^2(n-1); (\bar{X}-\mu)/(S/\sqrt{n}) \sim t(n-1)$$

From a random sample: the **sample mean** $\bar{X}$ has $E = \mu$, $\text{Var} = \sigma^2/n$ - averaging shrinks variation (why polls average out noise). If the population is normal: $\bar{X}$ is exactly normal; **$(n-1)S^2/\sigma^2$ follows chi-square($n-1$)**; and $\bar{X}$ and S^2 are INDEPENDENT - the miracle behind the t-statistic **$(\bar{X} - \mu)/(S/\sqrt{n}) \sim t(n-1)$** , the engine of small-sample inference.

Order statistics $X_{(1)} \leq \dots \leq X_{(n)}$: the sample min/max/median. For iid continuous samples, the pdf of $X_{(k)}$ uses the binomial form: $n! / ((k-1)!(n-k)!) [F(x)]^{k-1} [1-F(x)]^{n-k} f(x)$. Applications: tolerance limits, auction theory (winning bid = max), reliability (k-out-of-n systems), and robust estimation (median resists outliers that destroy the mean).

SECTION SUMMARY

$\bar{X}$: mean μ , variance σ^2/n . Normal samples: chi-square for S^2 , t for the mean with estimated σ , $\bar{X}$ independent of S^2 . Order statistics model extremes and robust summaries.

13. Laws of Large Numbers and the Central Limit Theorem

LEARNING OBJECTIVES

- State the weak and strong laws
- Apply the Central Limit Theorem
- Use CLT for approximations

$$\text{CLT: } (\bar{X} - \mu) / (\sigma/\sqrt{n}) \rightarrow N(0, 1) \text{ as } n \rightarrow \infty$$

The **Weak Law of Large Numbers**: $\bar{X}$ converges in probability to μ (Chebyshev proves it given finite variance) - frequencies stabilize to probabilities, justifying all of statistics. The **Strong Law** upgrades this to almost-sure convergence. Monte Carlo simulation is the WLLN in action: average random draws to estimate integrals and expectations.

The **Central Limit Theorem**: for iid variables with finite variance, the standardized sample mean tends to $N(0,1)$ - REGARDLESS of the population's shape. This is why normal-based inference

works on non-normal data at $n \geq 30$ (roughly), why measurement errors are normal (sums of small effects), and why the binomial/proportion approximations hold. Proof sketch via characteristic functions; the mgf version needs mgfs to exist. **De Moivre-Laplace** is the binomial special case.

■ VERY IMPORTANT

LLN says the mean converges to μ (stability); CLT says the standardized mean is approximately normal (shape). One gives consistency, the other gives inference - know which is which.

SECTION SUMMARY

WLLN: $\bar{X} \rightarrow \mu$ (stability of frequencies). CLT: standardized $\bar{X} \rightarrow N(0,1)$ for ANY shape (the basis of large-sample inference). Together they justify statistics.

14. Exam Pointers

High-yield topics

- Definitions: pmf/pdf/cdf, expectation, variance, mgf, covariance, correlation.
- Formulas to memorize cold: binomial/Poisson/geometric/hypergeometric pmfs; exponential/normal/gamma/chi-square densities; mgf derivatives; transformation Jacobian; CLT standardization.
- Derivations: $E(X^2) - \mu^2$; mgf of exponential; memoryless proof; chi-square from squared normals; order-statistic pdf.
- Distinguishing: linearity of E vs additivity of Var ; correlation vs causation; with/without replacement; weak vs strong law; LLN vs CLT.
- Applications: continuity-corrected normal approximation ($np \geq 10$); Poisson limit; min/max reliability; t-intervals.
- Diagram-free but computation-heavy: practice standardization, mgf differentiation, and joint-distribution region integrals until fluent.

SECTION SUMMARY

STA6310 rewards fluent computation: memorize the pmf/pdf catalog, the mgf toolkit, and the sampling-distribution trio ($\bar{X}$, chi-square, t) - then practice until mechanical.