**2025-3-7 STATS707 Summary**

**英文原文 & 中文翻译**

1. Probability Basics & Notation（概率基础与符号）

English Version:

• Sample Space (S): The set of all possible outcomes of an experiment (e.g., for a 6-sided die: S = {1, 2, 3, 4, 5, 6}).

• Events (A, B, etc.): Subsets of the sample space. Example: A = {1, 3, 5} (odd outcomes).

• Intersection (A ∩ B): Outcomes belonging to both A and B.

• Union (A ∪ B): Outcomes belonging to either A or B (or both).

• Complement (Aᶜ): The set of outcomes not in A (i.e., S \ A).

• Basic Probability Rules:

1. \( P(A^\complement) = 1 - P(A) \)

2. P(A \cup B) = P(A) + P(B) - P(A \cap B)

Chinese Version（中文）:

1. 样本空间 (S)：一个试验所有可能结果的集合（例如，掷一次六面骰子的样本空间 S = {1, 2, 3, 4, 5, 6}）。

2. 事件 (A, B 等)：样本空间的子集。例如，A = {1, 3, 5}（表示掷骰子得到奇数）。

3. 交集 (A ∩ B)：属于 A 且属于 B 的所有结果。

4. 并集 (A ∪ B)：属于 A 或属于 B（或两者都属于）的所有结果。

5. 补集 (Aᶜ)：不在 A 中而在样本空间 S 中的所有结果，即 S \ A。

6. 概率基本规则：

7. \( P(A^\complement) = 1 - P(A) \)

8. P(A \cup B) = P(A) + P(B) - P(A \cap B)

2. Conditional Probability & Independence（条件概率与独立性）

English Version:

• Conditional Probability: P(A \mid B) = \frac{P(A \cap B)}{P(B)}.

• This tells us the probability of A happening given that B has already happened.

• Independence: Two events A and B are independent if P(A \cap B) = P(A)\,P(B). Equivalently, P(A \mid B) = P(A) or P(B \mid A) = P(B).

• Example: Rolling a die and getting an odd number (A) vs. rolling a die and getting a number > 4 (B). We say A and B are independent if the occurrence of A doesn’t affect the probability of B.

Chinese Version（中文）:

• 条件概率：P(A \mid B) = \frac{P(A \cap B)}{P(B)}。

• 它表示在事件 B 已发生的条件下，事件 A 发生的概率。

• 独立性：如果 P(A \cap B) = P(A)\,P(B)，我们就称事件 A 和事件 B 独立。等价地可写为 P(A \mid B) = P(A) 或 P(B \mid A) = P(B)。

• 示例：掷骰子时，A 表示结果为奇数，B 表示结果大于 4。如果 A 的发生并不影响 B 的发生概率（也就是 A 发生与否都不改变 B 的概率），则称它们独立。

3. Bayes’ Theorem & Total Probability（贝叶斯定理与全概率公式）

English Version:

• Bayes’ Theorem:

P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}.

• It allows us to “flip” conditional probabilities (e.g., if we know P(B \mid A), we can find P(A \mid B)).

• Extensively used in medical testing (sensitivity, specificity) and machine learning (Naive Bayes classifiers).

• Law of Total Probability: If A and \( A^\complement \) are complementary events, then

\[

P(B) = P(B \mid A)\,P(A) \;+\; P(B \mid A^\complement)\,P(A^\complement).

\]

• This generalizes to partitions of the sample space as well.

Chinese Version（中文）:

• 贝叶斯定理：

P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}.

• 该公式让我们能够在已知 P(B \mid A) 的情况下，推导出 P(A \mid B)。

• 广泛用于医学检测（灵敏度、特异度）以及机器学习（如朴素贝叶斯分类器）。

• 全概率公式：若 A 与 \( A^\complement \) 互为补集，则

\[

P(B) = P(B \mid A)\,P(A) \;+\; P(B \mid A^\complement)\,P(A^\complement).

\]

• 这一公式也可以推广到对样本空间的任意划分。

4. Discrete vs. Continuous Distributions（离散分布与连续分布）

English Version:

• Discrete: The random variable takes countable (often finite) distinct values (e.g., {1, 2, 3, 4, 5, 6} for a die). Probability is given by a sum of point masses P(X = x).

• Continuous: The random variable takes an uncountably infinite set of values (e.g., real numbers in an interval). Probability is represented by the area under a density curve (PDF). For a single point, P(X = x) = 0; only intervals have nonzero probability.

• Reading Plots:

• Discrete: Bar charts or probability mass functions (PMF).

• Continuous: Probability density functions (PDF) and cumulative distribution functions (CDF).

Chinese Version（中文）:

• 离散分布：随机变量只能取可数（通常有限）个离散值（例如掷骰子时 X ∈ {1, 2, 3, 4, 5, 6}）。其概率由一系列点概率 P(X = x) 决定。

• 连续分布：随机变量取值为某个区间内的无穷多实数。其概率通过密度曲线（PDF）下面的面积来表示。对于单点，概率 P(X = x) = 0；只有区间才有非零概率。

• 读图方法：

• 离散分布：常用柱状图或概率质量函数 (PMF)。

• 连续分布：常用概率密度函数 (PDF) 和累积分布函数 (CDF) 来进行可视化。

5. Mean & Variance（期望与方差）

English Version:

• Mean (Expectation):

E[X] = \mu.

• For a discrete variable: E[X] = \sum x\_i P(X = x\_i).

• For a continuous variable: E[X] = \int\_{-\infty}^{\infty} x f(x)\, dx.

• Variance:

\mathrm{Var}(X) = E[(X - \mu)^2].

• Standard Deviation (SD) is the square root of the variance, which is easier to interpret.

• Interpretation: Variance measures how spread out the data are around the mean. A higher variance indicates data points are more dispersed.

Chinese Version（中文）:

• 期望（平均值）：

E[X] = \mu.

• 离散变量：E[X] = \sum x\_i P(X = x\_i).

• 连续变量：E[X] = \int\_{-\infty}^{\infty} x\, f(x)\, dx.

• 方差：

\mathrm{Var}(X) = E[(X - \mu)^2].

• 标准差（SD）是方差的平方根，更便于理解其数值含义。

• 含义：方差用来衡量数据值围绕平均值的离散程度。方差越大，数据点越分散。

6. Common Distributions & Their Uses（常见分布及用途）

English Version:

1. Bernoulli Distribution: Models a single trial with only two outcomes (success/failure). Parameter p.

2. Binomial Distribution: Models the number of successes in n independent Bernoulli trials. Probability mass function:

P(X = x) = \binom{n}{x} p^x (1-p)^{n - x}.

3. Poisson Distribution: Models the number of rare events in a given interval. Parameter \lambda. Mean = Variance = \lambda.

4. Exponential Distribution: Models the waiting time between occurrences of events (often ties in with Poisson processes). Parameter \lambda.

5. Normal Distribution: Symmetric “bell curve” used widely when data are roughly symmetric or from Central Limit Theorem contexts. Parameterized by mean \mu and standard deviation \sigma.

Chinese Version（中文）:

1. 伯努利分布（Bernoulli Distribution）：描述一次只有两种结果（成功/失败）的试验，参数为 p。

2. 二项分布（Binomial Distribution）：描述 n 次相互独立的伯努利试验中成功的次数，其概率质量函数为

P(X = x) = \binom{n}{x} p^x (1-p)^{n - x}.

3. 泊松分布（Poisson Distribution）：描述在给定区间内发生的稀有事件次数，参数为 \lambda。其均值和方差都等于 \lambda。

4. 指数分布（Exponential Distribution）：描述两个独立事件之间的等待时间，经常与泊松过程相联系。参数为 \lambda。

5. 正态分布（Normal Distribution）：对称的“钟形”曲线，广泛应用于数据近似对称或符合中心极限定理的场合，以 \mu（均值）和 \sigma（标准差）为参数。

7. Using Probability Distributions in R（在 R 中使用概率分布）

English Version:

• Key Functions (for a given distribution xxx):

1. dxxx(x, ...): Gives the density or PMF at x.

2. pxxx(q, ...): Gives the cumulative distribution up to q.

3. qxxx(p, ...): Gives the quantile corresponding to cumulative probability p.

4. rxxx(n, ...): Generates n random samples from the distribution.

• Examples:

• dnorm(x, mean=10, sd=1): Probability density at x for Normal(10, 1).

• pnorm(q, mean=10, sd=1): CDF evaluated at q.

• qnorm(p, mean=10, sd=1): p-th quantile.

• rnorm(n, mean=10, sd=1): Generate n samples.

• Similar Functions for Others: dbinom / pbinom / qbinom / rbinom (Binomial), dpois / ppois / qpois / rpois (Poisson), etc.

Chinese Version（中文）:

• 主要函数（对某一分布记作 xxx）：

1. dxxx(x, ...)：给出该分布在 x 处的密度或 PMF 值。

2. pxxx(q, ...)：给出该分布在 q 处的累积分布函数值。

3. qxxx(p, ...)：给出分布在累积概率为 p 时对应的分位数。

4. rxxx(n, ...)：从该分布中生成 n 个随机样本。

• 示例：

• dnorm(x, mean=10, sd=1)：在 Normal(10, 1) 分布下计算 x 的概率密度。

• pnorm(q, mean=10, sd=1)：计算在点 q 处的 CDF 值。

• qnorm(p, mean=10, sd=1)：求分位数，即累计概率为 p 时的 x 值。

• rnorm(n, mean=10, sd=1)：从 Normal(10,1) 中生成 n 个样本值。

• 其他分布的类似函数：二项分布用 dbinom / pbinom / qbinom / rbinom，泊松分布用 dpois / ppois / qpois / rpois，等等。

8. Additional Notes & Expanded Points（补充说明与扩展内容）

English Version:

1. Connection to Hypothesis Testing:

• The lecture briefly mentioned t-distribution, F-distribution, and chi-square distribution. These are commonly used in statistical inference to test hypotheses about means, variances, and relationships between variables.

• Example: The t-distribution appears in small-sample mean comparisons (e.g., Student’s t-test).

2. Simulation & Randomness:

• set.seed() ensures reproducibility in R. If not set, each run may yield different results due to different random draws.

• The concept of “random sampling” is crucial for understanding how sample means vary around the population mean (Central Limit Theorem).

3. Choosing a Distribution:

• In practice, the choice of distribution depends on context. If data are “count data with rare events,” Poisson might be suitable. If continuous and time-between-events, exponential is used. If large sample size and data roughly symmetrical, normal distribution is often used.

4. Interpretation of p-values & Confidence Intervals (briefly touched by the lecturer):

• 95% confidence often is tied to ±2 standard deviations in a normal distribution.

• In inference, these intervals and p-values are heavily dependent on which underlying distribution assumption is used.

Chinese Version（中文）:

1. 与假设检验的关联：

• 课上提到的 t 分布、F 分布和卡方分布在统计推断中应用广泛，用来检验关于均值、方差以及变量之间关系等的假设。

• 示例：t 分布经常用于小样本的均值检验（如 Student’s t-test）。

2. 模拟与随机性：

• set.seed() 能使 R 中的随机过程可重复。如果不设置，每次运行都会因随机数不同而产生不同结果。

• 理解“随机抽样”对于理解样本均值在总体均值附近的波动（中心极限定理）非常关键。

3. 如何选择合适的分布：

• 在实际问题中，分布的选择取决于情境。如果是稀有事件的计数数据，可以选泊松分布；若要描述事件之间的时间间隔，可用指数分布；若样本量大且数据近似对称时，常选正态分布。

4. 关于 p 值与置信区间的解释（老师简单提及）：

• 在正态分布中，“95% 置信度”通常对应在均值±2 个标准差的区域。

• 在统计推断中，这些区间和 p 值往往取决于假设所用的分布。

9. Conclusion & Study Tips（结论与学习建议）

English Version:

• Mastering probability requires a solid grasp of basic set operations (union, intersection, complement) and the associated formulas.

• Understand how conditional probability works and why independence is defined the way it is.

• Practice using the Bayes’ theorem and total probability to handle real-world problems (medical test accuracy, classification tasks).

• Learn the difference between discrete and continuous distributions, and how to interpret PMFs vs. PDFs.

• Familiarize yourself with the major distributions (Bernoulli, Binomial, Poisson, Normal, Exponential, etc.) and practice implementing them in R.

• Explore advanced distributions (t, F, chi-square) when diving into hypothesis testing and confidence intervals.

Chinese Version（中文）:

• 要掌握概率，需要对基本的集合运算（并、交、补）及其相关公式非常熟悉。

• 理解条件概率的内涵，以及为什么“独立”要用 P(A \cap B) = P(A)P(B) 这样的形式来定义。

• 多练习使用贝叶斯定理和全概率公式来处理实际问题（如医学检测准确度、分类任务等）。

• 分清离散型与连续型分布的区别，理解 PMF 和 PDF 分别如何解释。

• 熟悉主要分布（伯努利、二项、泊松、正态、指数等），并在 R 环境下多加练习这些分布的函数使用。

• 如果要深入到假设检验和置信区间领域，则需进一步学习 t 分布、F 分布、卡方分布等高阶分布。