**2025-3-4 STATS707 Summary**

**英文原文 & 中文翻译**

1. 课程与学习安排 (Course and Learning Structure)

中文概述：

• 本课程采用混合式教学 (线上视频+线下Workshop)。

• 一周两次课（周二与周五），其中第一周有两次线下授课，之后多数内容线上讲解，线下仅保留研讨课 (Workshop)。

• 研讨课以动手练习和探讨为主，侧重通过问题和练习来加深对知识的掌握。

• 课程评估包含：两次测试 (各 25%)、四次作业 (共 40%)、研讨课参与 (若出勤则加分)。

• 作业/测试注重对概念的理解与应用，可能包含选择题与简答题，需结合统计理论和实际解释。

English Summary:

• The course adopts a hybrid format (online videos + in-person workshops).

• There are two sessions per week (Tuesday and Friday). The first week features two in-person lectures; thereafter, most lectures are online, with only workshops held on campus.

• Workshops focus on hands-on practice and discussion, using targeted exercises to reinforce learning.

• Assessment includes two tests (25% each), four assignments (total 40%), and workshop attendance credit.

• Both tests and assignments emphasize conceptual understanding and practical application, possibly including multiple-choice and short-answer questions requiring theoretical knowledge and interpretations.

2. R 简介与数据操作 (Introduction to R and Data Handling)

中文概述：

• R 是一款在统计学界极为常见的开源软件，也是在新西兰统计系发源。

• 需要同时安装 R 与 RStudio（后者为前端界面，便于写代码、查看环境等）。

• 关键操作包括：导入数据 (read.csv 或 file.choose())，查看数据 (如 head() / summary() / View())，以及基本统计分析 (如 mean(), sd(), summary() 等)。

• 可以使用 plot()、hist()、boxplot() 等函数快速可视化数据。

• 注意 R 中的缺失值用 NA 表示，与其他常规数据不同；如需统计处理，得先排除或指定参数来忽略 NA。

补充说明（老师简单带过的点）：

• R 的对象类型包括向量 (vector)、矩阵 (matrix)、数据框 (data.frame) 等，不同类型在后续分析中使用方式也略有区别。

• 通过 ?functionName 或 help(functionName) 可以查看相关函数的用法。

English Summary:

• R is a widely used open-source software in statistics, originally developed in New Zealand’s Statistics Department.

• Installation typically involves both R (the backend) and RStudio (the user-friendly interface).

• Key operations: importing data (e.g., read.csv or file.choose()), viewing data (head(), summary(), View()), and basic analyses (mean(), sd(), summary()).

• Visualization can be quickly achieved via functions such as plot(), hist(), boxplot(), etc.

• Missing values in R are represented by NA, which requires special handling or the use of parameters to omit them in calculations.

Additional Points (briefly mentioned by lecturer):

• Common R object types include vectors, matrices, and data frames; each requires slightly different handling.

• You can use ?functionName or help(functionName) to access documentation on any R function.

3. 探索性数据分析 (Exploratory Data Analysis)

中文概述：

• summary() 可查看最小值、最大值、四分位数、中位数及均值，但不包含方差或标准差。需使用 var() 或 sd()。

• 箱线图 (boxplot) 可观察分布形态、是否有离群值；直方图 (hist) 也能直观展示分布的偏态及集中程度。

• 数据分布形态常与 “平均值 vs. 中位数” 的关系有关：若均值 > 中位数，多为右偏 (long tail on the right)；均值 < 中位数，多为左偏；若两者接近，则更趋于对称。

• 在实际数据中，列如学校师生比 (student-to-staff ratio) 可能出现小数，因部分教师为兼职；出现极大或极小值时，会影响总体均值和分布形态。

补充说明（老师简单带过的点）：

• 对于右偏(positive skew)或左偏(negative skew)的数据，有时可以通过对数变换 (log) 或平方根变换 (sqrt) 使分布更趋于正态，方便后续分析(如线性回归)。

• 方差与标准差都能度量数据的离散程度；标准差更常用，单位和原数据一致。

English Summary:

• The summary() function displays minimum, maximum, quartiles (Q1 and Q3), median, and mean but does not include variance or standard deviation; use var() or sd() for those.

• Boxplots reveal distribution shape and outliers; histograms offer visual indications of skewness and central tendency.

• A distribution’s shape often relates to the mean-median relationship. If mean > median, it often indicates a right skew (long right tail). If mean < median, likely a left skew; if they’re close, the distribution tends toward symmetry.

• Real datasets (e.g., a school’s student-to-staff ratio) can be non-integer (part-time staff), and extreme values heavily influence the mean and the shape of the distribution.

Additional Points (briefly mentioned by lecturer):

• For skewed data, transformations like log or square root can improve normality, which is often useful before certain analyses (e.g., linear regression).

• Both variance and standard deviation measure data spread; standard deviation is more frequently used because it shares the same unit as the original data.

4. 随机变量与概率 (Random Variables and Probability)

中文概述：

• 随机变量 (Random Variable) 与一般变量不同，其取值具不确定性，只能通过概率分布来描述。

• 离散型随机变量 (Discrete) 如掷硬币、掷骰子，只有有限且可列举的取值；连续型随机变量 (Continuous) 则有无限多可取值，如身高、体重等。

• 样本空间 (Sample Space, Ω) 包含所有可能结果；空事件 (null event) 或概率为 0 的事件并不等于“不可能事件”，它仍然可以在现实中发生（例如飞镖落在无限个点的某个具体点上）。

• 概率计算基础：

• 经典方法 (relative frequency) 适用于简单有限结果情况，如骰子六面等可能；

• 更复杂的情形通常通过概率模型 (modeling) 或主观概率 (subjective probability) 来估计。

补充说明（老师简单带过的点）：

• 概率为 0 的事件并不一定“绝不会发生”，只是在数学度量上接近于“测不出”；同理，概率为 1 的事件也并非必然发生。

• 主观概率在实际决策（如专家评估风险）中相当常见。

English Summary:

• A random variable differs from an ordinary variable in that its value is uncertain and must be described through a probability distribution.

• Discrete random variables (e.g., coin flips, dice rolls) have a finite set of possible outcomes; continuous random variables (e.g., height, weight) have infinitely many.

• The sample space (Ω) consists of all possible outcomes. Null events (or events with probability zero) are not necessarily “impossible”; in practice, they can still occur (e.g., a dart landing on a specific point among infinitely many).

• Basic probability approaches:

• Classical or relative frequency for simple finite-outcome cases (like a fair six-sided die).

• More complex situations generally require modeling or subjective probability estimation.

Additional Points (briefly mentioned by lecturer):

• Probability zero does not imply “cannot happen”; it is simply too small to measure. Likewise, probability one does not guarantee occurrence.

• Subjective probability is widely used in expert judgment and decision-making.

5. 条件概率与后续内容 (Conditional Probability and Further Topics)

中文概述：

• 条件概率 (Conditional Probability) 描述已知一个事件发生后，另一个事件发生的概率，常写作 P(A \mid B)。

• 其计算公式：P(A \mid B) = \frac{P(A \cap B)}{P(B)}

• 实际案例：

• 已知掷骰子结果是奇数，那么它为 1 的概率就是 1/3；

• 已知天气阴天，则下雨的概率可能大于未给定阴天时的概率。

• 后续会学习 Bayes 定理 (Bayes’ Theorem)：用于将“先验概率”更新为“后验概率”，在许多机器学习和统计推断中均非常重要。

补充说明（老师简单带过的点）：

• 条件概率可以反映“已知部分信息对事件的影响程度”。如果事件互相独立，则 P(A \mid B) = P(A)，说明 B 的发生不影响 A 的发生概率。

English Summary:

• Conditional probability describes the probability of one event occurring given that another event is known to have occurred, typically written as P(A \mid B).

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

• Practical examples:

• If a die roll is known to be odd, the probability it is 1 is 1/3.

• If the weather is already cloudy, the probability of rain is typically higher than without that condition.

• Upcoming topic: Bayes’ theorem, which updates “prior probabilities” into “posterior probabilities,” widely used in machine learning and statistical inference.

Additional Points (briefly mentioned by lecturer):

• Conditional probability captures how partial information can affect the likelihood of events. If events are independent, P(A \mid B) = P(A), meaning B’s occurrence does not alter A’s probability.

总结 (Final Note)

中文概述：

• 课程重点在于巩固基础统计概念，并熟悉如何用 R 进行简单的数据探索与概率分析。

• 理解分布形态、正确解读统计指标（如均值、方差、中位数、分位数等）对于分析实际问题非常关键。

• 概率部分则奠定了后续学习估计、假设检验、回归等更深层次统计方法的基础；扎实掌握条件概率和分布概念有助于后续更高阶的学习。

English Summary:

• The main goal of the course is to solidify foundational statistical concepts and gain familiarity with R for data exploration and probability-based analyses.

• Understanding distribution shapes and correctly interpreting statistical metrics (mean, variance, median, quantiles, etc.) is crucial for real-world problem-solving.

• The probability component lays the groundwork for advanced topics like estimation, hypothesis testing, and regression; a firm grasp of conditional probability and probability distributions will significantly aid in deeper studies.