Summary

0) The big picture: how statistical work is done

Most of the course follows a consistent workflow:

Data preparation (types, duplicates, missing values, outliers, transformations, encoding, scaling, splitting)

Descriptive statistics / EDA

Choose the right test/model + check assumptions

Inference and interpretation, then present results clearly

A key idea repeated in the regression material: regression can predict, but does not automatically prove causality unless you have design that supports it (random assignment, good controls, etc.).

1) Hypothesis testing: H0, H1, p-value, “reject”

Even if it wasn’t on the lines we pulled directly, the ANOVA and nonparametric lectures assume you know this framing:

H0 (Null Hypothesis): “no effect / no difference”

H1 (Alternative): “there is an effect / difference”

You compute a test statistic (t, F, χ^2, etc.) and a p-value

Decision rule: if $p < \alpha$ (often 0.05), you reject H0

How to think about it in real life:

If H0 is “the new feature does not change conversion”, rejecting H0 means “we have evidence that conversion changed”.

Not rejecting H0 does not mean “no effect exists”. It means “not enough evidence in this data”.

2) ANOVA: Analysis of Variance (Topic 1)

Why ANOVA exists

When you want to compare more than 2 groups, doing many pairwise t-tests inflates false positives. ANOVA is the standard global test for “are any group means different?”.

One-way ANOVA

One categorical factor (k groups), one numeric outcome.

Tests whether there is any overall difference across groups.

Interpretation:

Significant ANOVA says: at least one mean differs.

Then you usually do post-hoc comparisons (which pairs differ).

Two-way ANOVA (+ interaction)

Two categorical factors A and B. You test:

Main effect of A

Main effect of B

Interaction A×B: does the effect of A depend on B?
The slides explicitly define interaction sums of squares and how it fits into the ANOVA table.

3) Repeated Measures ANOVA (Topic 2)

Used when the same subjects are measured multiple times (timepoints, conditions). This violates independence, and repeated-measures designs handle that structure. The slides include real paper-style examples (e.g., longitudinal outcomes).

Practical example:

Same patients measured at baseline, 3 months, 6 months.
You need a model that respects “within-subject” dependence.

4) Linear Regression (Topic 3) and Multiple Regression (Topic 4)
Simple linear regression

Goal: model a numeric outcome Y using predictor X, and quantify the relationship.

A key conceptual warning appears in the regression slides:

“Even if X and Y are related, you cannot conclude X causes Y” unless design supports it.

Multiple linear regression

Why multiple?

To estimate the effect of each predictor while controlling for others.

The multiple regression deck frames the goal as estimating the effect of explanatory variables on Y and their relative importance, and gives the standard model form and estimation objective (minimize squared residuals).

Typical outputs you interpret:

coefficients (β), standard errors, p-values

overall model fit and error (RMSE, R^2 depending on context)

5) GLM + Model comparison + GEE (Topic 6)
Why GLM

Linear regression has limitations when Y is not “nice” (binary, counts, etc.). The GLM lecture organizes the topic as: limitations of linear regression, GLM intro, exponential-family distributions, link function, parameter estimation.

Model comparison: likelihood vs RMSE

The deck compares Log-Likelihood / LRT vs RMSE and notes that likelihood-based metrics support formal model comparison tools (LRT/AIC/BIC) while RMSE is prediction-error style.

When observations are dependent: GEE

If you have repeated measures per subject, the deck says the measurements of the same subject are dependent and introduces GEE (Generalized Estimating Equations) and correlation structures.

6) Poisson Regression (Topic 6b)

For count outcomes (0,1,2,…) such as number of calls, crashes, events per time/space.

The Poisson distribution assumptions are given:

events occur randomly

constant average rate

independence

Poisson regression is presented as a special case of GLM for count prediction, with concrete examples like predicting service calls or weekly road accidents.

7) Logistic Regression + Odds Ratio (Topic 6c)

For binary outcomes (0/1).
A big emphasis is Odds and Odds Ratio (OR); the deck is structured heavily around OR and then simple logistic regression.

Core interpretation you should know:

OR > 1 increases odds of outcome

OR < 1 decreases odds

OR = exp(β) for a one-unit change in predictor (standard logistic interpretation)

8) Variable selection + Model criteria + Regularization (Topic 7)

This topic is about “how to choose a model that generalizes”.

AIC vs BIC

The deck defines:

AIC = -2·log(L) + 2k

BIC = -2·log(L) + k·log(n)

And explains the practical difference:

AIC leans toward predictive performance

BIC penalizes complexity more strongly and leans toward finding the “true” model under assumptions

Regularization: Ridge, LASSO, Elastic Net

The deck covers:

Ridge (L2) shrinks coefficients, helps with multicollinearity

LASSO (L1) can shrink some coefficients to exactly zero (feature selection)

Elastic Net combines L1 and L2, and is recommended when you want both stability and feature selection, especially with correlated predictors

9) Dimension reduction: PCA, FA, (and PLS) (Topic 8)
Factor Analysis (FA)

FA assumes correlations among observed variables are explained by a smaller set of latent factors (hidden variables).

The deck uses a real-life analogy: restaurant reviews where many observed variables (waiting time, cleanliness, service, taste) can be explained by latent factors like “food quality” and “service”.

It also notes that FA is a latent-variable statistical model and contrasts it with PCA.

PCA vs FA

The slides summarize:

PCA focuses on compressing information (variance explained)

FA focuses on uncovering meaningful latent constructs

PLS (mentioned later in the deck)

PLS is presented as a supervised dimension reduction method (unlike PCA which ignores Y).

10) Survival Analysis (Topic 9)

Used when the outcome is time until event, often with censoring.

The survival deck includes examples like comparing survival curves and using Cox Proportional Hazards Model (CPHM) while adjusting for covariates like age, and interpreting hazard-related regression outputs (e.g., effect sizes, significance).

Practical mental model:

It's not "did the event happen", it's "when did it happen", and many subjects might not have experienced it yet by end of study (censoring).

11) Nonparametric tests (Topic 10)

When assumptions for parametric tests are weak (non-normality, ordinal scales, heavy outliers), use rank-based methods.

The deck lists core tools:

Mann–Whitney U / Wilcoxon rank-sum (alternative to two-sample t-test) and notes it doesn't depend on the data distribution

Wilcoxon signed-rank (paired)

Kruskal–Wallis (alternative to one-way ANOVA)

Friedman test (repeated measures alternative)

Kendall's Tau for monotonic association in ordinal settings

12) Tree-based learning: Decision Trees + Random Forest (Topic 11)

The Random Trees deck frames ensemble learning:

bootstrap sampling and feature subsampling to reduce correlation between trees and reduce variance

Random Forest prediction: majority vote for classification, average for regression

common hyperparameters like number of trees, max features, max depth, min samples split/leaf

feature importance is highlighted as a key output

13) Data preparation (Topic 5): what you're expected to do in projects

This topic is very "hands on":

check data sanity (units, ranges, duplicates)

fix variable types

handle missing values

detect outliers

encode categorical variables (One-hot, Target Encoding, etc.)

standardize/normalize

merge rare categories

feature engineering

class balancing (if needed)

train/val/test split

It also gives guidance for which encoding fits which model family (linear/logistic vs trees vs neural nets).

A super practical "which method do I use?" cheat sheet

Numeric Y, 2 groups: t-test (or Mann–Whitney)

Numeric Y, 3+ groups: One-way ANOVA (or Kruskal–Wallis)

Numeric Y with predictors: Linear / Multiple Regression

Binary Y: Logistic Regression + OR interpretation

Count Y: Poisson Regression (GLM)

Dependent observations (repeated measures): Repeated Measures ANOVA / GEE

Many correlated predictors: PCA/FA/PLS

Time-to-event: Survival analysis (Cox / curves)

Nonlinear, mixed-type, strong baseline model: Random Forest