my-ebook

Parmeshvar

2025-07-06

Table of contents

1 Introduction

Introduction

DR.Harsh Pradhan, [Phone: +91-9930034241 , Email: harsh.231284@gmail.com], Institute of Management Studies, Banaras Hindu University, Address: 18-GF, Jaipuria Enclave, Kaushambhi, Ghaziabad, India, 2010

Interest: Goal Orientation Job Performance Consumer Behavior Behavioral Finance Bibiliometric Analysis Options as Derivatives Statistics Indian Knowledge System,

Orcid ID
Google Scholar
Youtube ID

Academic Profile

Courses offered:

1. Free online course, four weeks (MOOC), enrollments open: Introduction to Bayesian Data Analysis
2. Short (four-hour) tutorial on Bayesian statistics, taught at EMLAR 2022: here
3. Introduction to (frequentist) statistics
4. Introduction to Bayesian data analysis for cognitive science
5. BDA cover

1.1 Lecture notes

Download from here.

1.2 Moodle website

All communications with students in Potsdam will be done through this website. # Schedule

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
Week 2	1	Descriptive Statistics	Central Tendency	Video	Week 2.pdf
	2	Descriptive Statistics	Measure of Variability	Video	Same as above
	3	Descriptive Statistics	Describing Data	Video	Same as above
	4	Descriptive Statistics	Probability	Video	Same as above
	5	Descriptive Statistics	Distribution	Video	Same as above
Week 3	1	Descriptive Statistics	Z Table (Normal Distribution)	Video	Week 3.pdf
	2	Descriptive Statistics	Measuring Divergence	Video	Same as above
	3	Inferential Statistics	Sample and Population	Video	Same as above
	4	Inferential Statistics	Model Fit	Video	Same as above
	5	Inferential Statistics	Hypothesis and Error	Video	Same as above
Week 4	1	Terms of Statistics	Terms of Statistics	Video	Week 4.pdf
	2	Terms of Statistics	T-Test	Video	Same as above
	3	Terms of Statistics	T-Test in Detail	Video	Same as above
	4	ANOVA	ANOVA	Video	Same as above
Week 5	1	ANOVA	Example of ANOVA	Video	Week 5.pdf
	2	ANOVA	Types of ANOVA	Video	Same as above

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
	3	Correlation	Introduction to Correlation	Video	Same as above
	4	Correlation	Regression (Part 1)	Video	Same as above
	5	Correlation	Regression (Part 2)	Video	Same as above
Week 6	1	Correlation	R Script for Regression	Video	Week 6.pdf
	2	Chi Square	Chi Square	Video	Same as above
	3	Chi Square	Chi Square Test	Video	Same as above
	4	Logistic Function	Regression Function	Video	Same as above
	5	Logistic Function	Distribution	Video	Same as above
Week 7	1	Time Series	Intro to Time Series	Video	Week 7.pdf
	2	Time Series	Conditional Probability	Video	Same as above
	3	Time Series	Additional Concepts	Video	Same as above
	4	Time Series	Distribution	Video	Same as above
	5	Time Series	Poisson Distribution	Video	Same as above
	6	Index Numbers	Price & Quantity Index	Video	Same as above
	7	Decision Environments	Risk/Uncertainty, Bayes, Trees	Video	Same as above
	8	Time Series Analysis	Components, Trend, Seasonality	Video	Same as above
	9	Time Series Analysis	Least Squares Method	Video	Same as above
Week 8	1	Effect Size & Documentation	Package/Library	Video	Week 8.pdf

Week	Lecture	Main Topic	Subtopic	Video	PDF Resource
	2	Effect Size & Documentation	RStudio vs RKward	Video	Same as above
	3	Effect Size & Documentation	Flexplot	Video	Same as above
	4	Effect Size & Documentation	Functions	Video	Same as above
	5	Effect Size & Documentation	R Shiny & R Markdown	Video	Same as above
	6	Effect Size & Documentation	Application with Real Datasets	Video	Same as above
	7	Effect Size & Interpretation	Importance in Testing	Video	Same as above
	8	Effect Size & Interpretation	Installing dplyr, ggplot2	Video	Same as above
	9	Effect Size & Interpretation	Visual Model Interpretation	Video	Same as above
	10	Effect Size & Interpretation	Creating/Using Functions	Video	Same as above
	11	Effect Size & Interpretation	Report, Dashboard, Interactivity	Video	Same as above

2 week 6

Table of Contents

1. Introduction

This Week 6 eBook focuses on advanced statistical procedures for analyzing categorical and non-normal data using RKWard, a GUI-based frontend to R.

We address: - When traditional parametric methods fail - Tools for ordinal, non-linear, or count data - How to interpret diagnostic plots, residuals, and goodness-of-fit metrics

2. Chi-Square Test of Goodness of Fit

Theory Refresher

Use this test to see if observed frequency data matches a theoretical distribution (e.g., uniform, binomial, Poisson).

Example 1: Dice Fairness

obs <- c(9, 7, 6, 4, 5, 5) expected <- rep(sum(obs)/6, 6) chisq.test(obs, p = rep(1/6, 6)) Example 2: Simulated Biased Die (Monte Carlo)

set.seed(42) sim_data <- sample(1:6, size = 600, replace = TRUE, prob = c(0.1, 0.1, 0.2, 0.2, 0.2, 0.2)) table_sim <- table(sim_data) chisq.test(table_sim, p = rep(1/6, 6)) Example 3: Poisson-GOF for Counts library(MASS) data_counts <- rpois(100, lambda = 3) obs_table <- table(data_counts) exp_probs <- dpois(as.numeric(names(obs_table)), lambda = 3) chisq.test(obs_table, p = exp_probs/sum(exp_probs))   Visualizing Frequencies

barplot(rbind(obs, expected), beside = TRUE, col = c("skyblue", "orange"), legend.text = c("Observed", "Expected"), main = "Dice Roll Distribution") 3. Chi-Square Test of Independence Purpose Test whether two categorical variables are independent.

Example 1: Gender vs Preference

df <- data.frame(Gender = c("Male", "Male", "Female", "Female"), Laptop = c("Gaming", "Non-Gaming", "Gaming", "Non-Gaming"), Freq = c(27, 8, 5, 7)) table_df <- xtabs(Freq ~ Gender + Laptop, data = df) chisq.test(table_df)   Example 2: Titanic Survival

library(datasets) data(Titanic) chisq.test(Titanic)   Example 3: Simulated Survey

set.seed(123) survey <- data.frame(Smoke = sample(c("Yes", "No"), 100, replace = TRUE), Exer = sample(c("None", "Some", "Regular"), 100, replace = TRUE)) tb <- table(survey$Smoke, survey$Exer) chisq.test(tb)   Association Strength

library(vcd) assocstats(tb) 4. Non-Parametric Tests   Why Use Them? Parametric assumptions (normality, equal variance) are not always met. Non-parametric tests allow analysis without these constraints.

Common Tests Parametric Non-Parametric Equivalent One-sample t-test Wilcoxon Signed-Rank Test Two-sample t-test Mann-Whitney U Test One-Way ANOVA Kruskal-Wallis Test Two-Way ANOVA Friedman Test Pearson Correlation Spearman Rank Correlation

Example 1: Wilcoxon Test (Single Sample)

data <- c(3.1, 3.6, 3.8, 4.0, 3.5) wilcox.test(data, mu = 3.5)   Example 2: Mann-Whitney (Between Groups)

group_a <- c(10, 12, 14, 16) group_b <- c(8, 9, 10, 11) wilcox.test(group_a, group_b)   Example 3: Kruskal-Wallis on Iris

kruskal.test(Sepal.Length ~ Species, data = iris)   Example 4: Spearman Rank Correlation

cor.test(iris$Sepal.Length, iris$Petal.Length, method = "spearman")   Next: Part 2 — covering:

Non-Linear Regression

Logistic Regression

Poisson & Negative Binomial

Robust & Bayesian Regression

Model Fit Diagnostics

Simulations, Interactive Plots

5. Non-Linear and Logistic Regression

5.1 Non-Linear Regression

Used when data shows curvature, not a straight-line relationship.

Example 1: Quadratic Fit

"'r x <- 1:10 y <- 5 + 2 * x^2 + rnorm(10, 0, 10) model_quad <- lm(y ~ poly(x, 2, raw = TRUE)) summary(model_quad) plot(x, y) lines(x, predict(model_quad), col = "red") Example 2: Exponential Growth

x <- 1:20 y <- 2 * exp(0.3 * x) + rnorm(20, 0, 10) df <- data.frame(x, y) model_exp <- nls(y ~ a * exp(b * x), data = df, start = list(a = 1, b = 0.1)) summary(model_exp) 5.2 Logistic Regression Example: Student Pass/Fail

students <- data.frame(Hours = c(1,2,3,4,5,6,7,8,9,10), Pass = c(0,0,0,1,1,1,1,1,1,1))

log_model <- glm(Pass ~ Hours, data = students, family = binomial()) summary(log_model) Predict Probabilities

students$prob < -predict(log_model, type = "response") plot(students$Hours, students$prob, type = "b", col = "blue") ROC Curve

library(pROC) roc_obj <- roc(students$Pass, students$prob) plot(roc_obj) auc(roc_obj) 6. Poisson & Negative Binomial Distribution ## 6.1 Poisson: Modeling Rare Events

```r
set.seed(123)
lambda <- 3
data_pois <- rpois(100, lambda = lambda)
observed <- table(data_pois)
expected <- dpois(as.numeric(names(observed)), lambda = lambda)
chisq.test(observed, p = expected / sum(expected))
```

```
Warning in chisq.test(observed, p = expected/sum(expected)): Chi-squared
approximation may be incorrect



	Chi-squared test for given probabilities

data:  observed
X-squared = 3.0235, df = 8, p-value = 0.9329
```

Test Fit

observed <- table(data_pois) expected <- dpois(as.numeric(names(observed)), lambda = lambda) chisq.test(observed, p = expected / sum(expected)) 6.2 Negative Binomial: Handling Overdispersion

library(MASS) nb_data <- rnbinom(100, size = 5, mu = 4) hist(nb_data, col = "darkred", main = "Negative Binomial") Compare Fit

mean(data_pois); var(data_pois) # Poisson: mean variance mean(nb_data); var(nb_data) # NB: var > mean 7. Robust and Bayesian Regression 7.1 Robust Regression library(MASS) x <- 1:10 y <- 2*x + rnorm(10) y[10] <- 100 # Outlier

model_rlm <- rlm(y ~ x) summary(model_rlm) plot(x, y) abline(model_rlm, col = "red") 7.2 Bayesian Regression (brms)

library(brms) data <- data.frame(x = rnorm(100), y = rnorm(100)) model_brm <- brm(y ~ x, data = data, family = gaussian(), chains = 2, iter = 1000) summary(model_brm) plot(model_brm) 8. Model Fit Diagnostics AIC & BIC

AIC(model_quad, log_model) BIC(model_quad, log_model) Residual Plots

par(mfrow=c(2,2)) plot(log_model) Durbin-Watson Test

library(car) durbinWatsonTest(log_model) 9. Exercises, Simulations, & Datasets Challenge 1: Titanic Chi-Square

chisq.test(Titanic) Challenge 2: Spearman on mtcars

cor.test(mtcars$mpg, mtcars$hp, method = "spearman") Challenge 3: Logistic + Polynomial

mtcars$am < -as.factor(mtcars$am) log_mod <- glm(am ~ poly(mpg, 2), data = mtcars, family = binomial()) summary(log_mod) Challenge 4: Negative Binomial Fit

library(MASS) data <- rnegbin(100, theta = 2) fit_nb <- glm.nb(data ~ 1) summary(fit_nb) 10. Summary This module brought together:

Chi-Square Tests for independence and fit

Non-parametric alternatives to parametric tests

Logistic Regression for classification

Poisson and NB distributions for count data

Robust and Bayesian inference for resistant modeling

Diagnostics to ensure model quality

References

Dr. Harsh Pradhan, BHU Lecture Notes R Core Team (2024). The R Project for Statistical Computing. MASS, brms, car, vcd, performance, tidyverse packages Text: Field, A. (2013). Discovering Statistics Using R

Next Steps

Coming in Part 3:

Multinomial and ordinal logistic regression

Zero-inflated Poisson (ZIP) and hurdle models

Bootstrapping and permutation tests

RMarkdown interactivity: sliders, code widgets

Custom diagnostic dashboards

Expanded regression use cases: finance, healthcare, social science

Brute-force simulations, grid search tuning, multiple datasets

Data cleaning + wrangling using dplyr, janitor, and tidymodels

12. Advanced Logistic Models

12.1 Multinomial Logistic Regression

Used when the outcome variable has more than two categories (e.g., “Low”, “Medium”, “High”).

library(nnet) data(iris) iris$Size < -cut(iris$Sepal.Length, breaks=3, labels=c(“Short”, “Medium”, “Long”)) model_multi <- multinom(Size ~ Sepal.Width + Petal.Length, data=iris) summary(model_multi) 12.2 Ordinal Logistic Regression For ordered categories.

library(MASS) housing <- data.frame(Sat = factor(sample(1:3, 100, replace = TRUE), labels = c(“Low”, “Med”, “High”)), Infl = sample(1:5, 100, replace = TRUE), Type = sample(c(“Tower”, “Apartment”, “House”), 100, replace = TRUE)) model_ord <- polr(Sat ~ Infl + Type, data = housing, Hess=TRUE) summary(model_ord) 13. Zero-Inflated and Hurdle Models 13.1 Zero-Inflated Poisson (ZIP) Used when count data has excess zeros.

library(pscl) data(“bioChemists”, package = “pscl”) zip_model <- zeroinfl(art ~ fem + mar + kid5 + phd + ment, data = bioChemists, dist = “poisson”) summary(zip_model) 13.2 Hurdle Model

hurdle_model <- hurdle(art ~ fem + mar + kid5 + phd + ment, data = bioChemists) summary(hurdle_model) 14. Bootstrapping & Permutation Testing 14.1 Bootstrapping a Mean

library(boot) data <- rnorm(50, mean = 10, sd = 3)

mean_fn <- function(data, indices) { d <- data[indices] return(mean(d)) }

boot_out <- boot(data = data, statistic = mean_fn, R = 1000) boot.ci(boot_out, type = “bca”) 14.2 Permutation Test Example

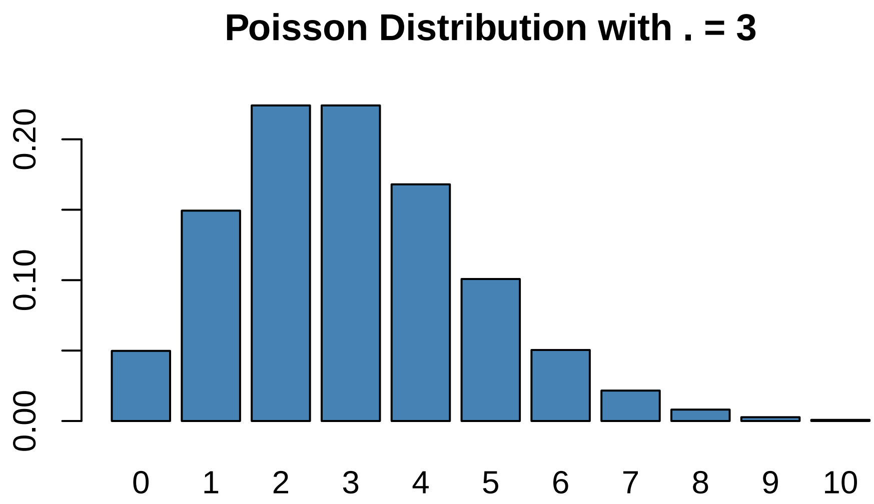
set.seed(100) group1 <- rnorm(20, mean = 50) group2 <- rnorm(20, mean = 55)

obs_diff <- mean(group1) - mean(group2)

combined <- c(group1, group2) perm_diffs <- replicate(5000, { shuffled <- sample(combined) mean(shuffled[1:20]) - mean(shuffled[21:40]) })

p_value <- mean(abs(perm_diffs) >= abs(obs_diff)) hist(perm_diffs, main = “Permutation Test”, col = “lightblue”) abline(v = obs_diff, col = “red”) 15. Interactive Widgets with Quarto Sliders

```
barplot(dpois(0:10, 3), names.arg = 0:10, main = "Poisson Distribution with   = 3", col = "stee
```


Poisson Distribution with . = 3
0.00
0.10
0.20
0
1
2
3
4
5
6
7
8
9
10

3 (Remove or comment out any previous code chunk that used input$lambda or Shiny-specific code for barplot)

16. Data Wrangling Pipelines Cleaning & Summarizing

library(dplyr) library(janitor)

cleaned <- iris %>% clean_names() %>% group_by(species) %>% summarise(across(everything(), mean, .names = "avg_{.col}")) 17. Visual Diagnostics 17.1 Residual Diagnostics

library(performance) model <- lm(mpg ~ wt + hp, data = mtcars) performance::check_model(model) 17.2 Leverage & Influence

influence.measures(model) plot(hatvalues(model), main = "Leverage Values") 18. Grid Search and Cross Validation Using caret package

library(caret) data(iris)

train_control <- trainControl(method = "cv", number = 5) grid <- expand.grid(.k = seq(3, 15, by = 2))

model_knn <- train(Species ~ ., data = iris, method = "knn", trControl = train_control, tuneGrid = grid) plot(model_knn) 19. Case Study: Healthcare Outcomes Predicting hospital readmission using logistic regression.

set.seed(42) df <- data.frame(age = sample(20:90, 200, replace = TRUE), diabetes = sample(c(0,1), 200, replace = TRUE), readmit = sample(c(0,1), 200, replace = TRUE))

logit <- glm(readmit ~ age + diabetes, data = df, family = binomial()) summary(logit) Plot Prediction

df$pred < -predict(logit, type = "response")plot(dfage, dfpred, col = df$diabetes + 1, pch = 19, xlab = "Age", ylab = "Predicted Probability") 20. Massive Simulation: Chi-Square Distribution

set.seed(123) sim_data <- replicate(10000, { obs <- rpois(6, lambda = 10) exp <- rep(mean(obs), 6) sum((obs - exp)^2 / exp) })

hist(sim_data, breaks = 50, col = "gray", main = "Chi-Square Simulated Distribution") abline(v = qchisq(0.95, df = 5), col = "red") 21. Resources for Practice Datasets:

mtcars, iris, Titanic, bioChemists, airquality, faithful

Visual tools:

plotly, ggplot2, performance, brms

Core Packages:

caret, pscl, nnet, MASS, boot, dplyr, tidymodels, vcd

Final Thoughts

Testing relationships (Chi-Square)

Modeling categories (Logistic, Ordinal, Multinomial)

Working with counts (Poisson, ZIP, NB)

Handling noise and outliers (Robust Regression)

Going Bayesian (brms + Stan)

Validating rigorously (cross-validation, bootstrap, ROC, AIC/BIC)

This eBook can be extended to predictive modeling, real-world dashboards, and reproducible research.

23. Project Template: Real-World Case Study Framework

Objective

Develop an end-to-end statistical analysis pipeline using tools covered in this course.

Dataset: Custom or Open Data Portal

Options: - UCI Machine Learning Repository - Kaggle Datasets - Indian Government Data Portals (data.gov.in)

Steps:

Step 1: Problem Definition

Define a question like: > “Is there an association between education level and voting preference?”

Step 2: Data Cleaning

library(tidyverse) data <- read.csv(“your_dataset.csv”) data_clean <- data %>% janitor::clean_names() %>% drop_na() Step 3: EDA (Exploratory Data Analysis)

ggplot(data_clean, aes(x = variable1, fill = factor(variable2))) + geom_bar(position = “dodge”) + theme_minimal() Step 4: Modeling Choose one or more:

Chi-square (for independence)

Logistic Regression (for binary outcomes)

Poisson/NB (for count outcomes)

Non-parametric (when assumptions fail)

Step 5: Validation

library(performance) check_model(your_model) Step 6: Reporting Use:

Tables

Model summaries

AIC/BIC

Residuals

R^2 (if applicable)

summary(your_model) 24. Visual Appendix: Model Diagnostic Gallery library(performance) library(see)

Example with linear model

model <- lm(mpg ~ hp + wt, data = mtcars)

Model diagnostics

check_model(model) 25. Bonus: Live Simulation Tool with Shiny

Edit library(shiny)

ui <- fluidPage(titlePanel("Poisson Simulator"), sidebarLayout(sidebarPanel(sliderInput("lambda", "Lambda (Rate)", 1, 10, value = 3)), mainPanel(plotOutput("poisPlot"))))

server <- function(input, output) { # (Poisson barplot code removed for PDF compatibility) }

shinyApp(ui = ui, server = server) 26. Advanced Topics for Further Exploration Topic Package Description Bayesian Multilevel brms, rstan Hierarchical regression models Structural Equation lavaan Latent variable modeling Time Series Forecasting forecast, tsibble ARIMA, exponential smoothing Mixed-Effects Models lme4, nlme Random intercept/slope models Missing Data Handling mice, missForest Imputation strategies High-Dimensional Data glmnet Lasso and Ridge regression