Penguins: distributions, groups and hidden associations

A Course on Data Visualization · Homework

Livio Finos, University of Padova

Contents

Getting started 2

1. Understand the observations and analysis samples **3**

2. Boxplots: unusual relative to which group? **3**

3. KDE: bandwidth and population mixtures **3**

4. The same scatterplot, with and without species **4**

5. Compare islands within more comparable groups **4**

6. Explain what conditioning changes **4**

7. Optional: bivariate KDE **5**

Looking ahead: from simple to multiple linear regression **5**

Submission and reflection **5**

Data and references **5**

Main question: How does accounting for species and sex change our description of penguin body measurements and the relationships between them?

You will combine boxplots, one-dimensional kernel density estimates (KDEs), stratified scatterplots, Pearson correlations and comparisons within groups. All analyses are descriptive: focus on the direction, strength and shape of the observed patterns. A final optional section introduces bivariate density contours. Work through Sections 1–6 before attempting Section 7.

Student task — work with your group.

The observational unit is a penguin in the Palmer Archipelago dataset. The measurements were collected in 2007–2009. The dataset contains three species and both numerical and categorical variables.

Before each analysis, write a prediction. Then add your code and explain whether the result supports it. Keep the supplied setup; replace the TODO comments with your work. The uncompleted template can be knitted, but it will not contain your answers.

Getting started

Open this file in RStudio and choose **Knit** to produce an HTML report. Install the packages below once, from the console. Place this Rmd in the folder that contains the `data` folder; the supplied dataset must be at `data/penguins.csv`. When running chunks interactively, use that same folder as the working directory. The CSV is a local copy of the Palmer Penguins data distributed by the `palmerpenguins` project; its provenance and citation are given below. You do not need to install that data package. `MASS` is used only for the optional bivariate KDE section.

```
install.packages(c("ggplot2", "MASS", "knitr", "rmarkdown"))
```

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = TRUE,
  fig.width = 8, fig.height = 5, dpi = 120)
library(ggplot2)
theme_set(theme_minimal(base_size = 12))
set.seed(314159)
data_file <- file.path("data", "penguins.csv")
if (!file.exists(data_file)) {
  stop("Cannot find data/penguins.csv. Put the data folder next to this Rmd file.")
}
p_raw <- read.csv(data_file, stringsAsFactors = FALSE,
                  na.strings = c("", "NA"), strip.white = TRUE)
required <- c("species", "island", "sex", "body_mass_g",
              "bill_length_mm", "bill_depth_mm")
missing_columns <- setdiff(required, names(p_raw))
if (length(missing_columns)) {
  stop("Missing CSV columns: ", paste(missing_columns, collapse = ", "))
}
# CSV files do not preserve R factor levels: recreate them explicitly.
p_raw$species <- trimws(p_raw$species)
p_raw$sex <- tolower(trimws(p_raw$sex))
stopifnot(all(na.omit(p_raw$species) %in% c("Adelie", "Chinstrap", "Gentoo")),
          all(na.omit(p_raw$sex) %in% c("female", "male")))
p_raw$species <- factor(p_raw$species, levels = c("Adelie", "Chinstrap", "Gentoo"))
p_raw$sex <- factor(p_raw$sex, levels = c("female", "male"))
p_raw$island <- factor(trimws(p_raw$island))
numeric_columns <- c("body_mass_g", "bill_length_mm", "bill_depth_mm")
stopifnot(all(vapply(p_raw[numeric_columns], is.numeric, logical(1))))
p_raw$row_id <- seq_len(nrow(p_raw))
species_colours <- c(Adelie = "#D55E00", Chinstrap = "#0072B2", Gentoo = "#009E73")

# Analysis populations: each excludes only the variables needed for that task.
p_mass <- p_raw[complete.cases(p_raw[c("body_mass_g", "species", "sex")]), ]
p_bill <- p_raw[complete.cases(p_raw[c("bill_length_mm", "bill_depth_mm", "species")]), ]

# We use sample quartiles (R quantile type 7), not boxplot.stats() hinges.
flag_iqr <- function(x) {
  q <- quantile(x, c(0.25, 0.75), type = 7, names = FALSE)
  spread <- q[2] - q[1]
  x < q[1] - 1.5 * spread | x > q[2] + 1.5 * spread
}
```

Keep the same analysis rows when comparing graphical summaries. Record any exclusions and their reasons.

Use units in axis labels, common scales for comparable panels, and captions stating the population or subgroup shown.

1. Understand the observations and analysis samples

Inspect `p_raw`. Make a compact codebook for `species`, `island`, `sex`, `body_mass_g`, `bill_length_mm` and `bill_depth_mm`.

1. Count observations and missing values. Report the sizes of `p_mass` and `p_bill` and explain why they differ.
2. Tabulate species by sex, including missing sex values, and species by island.
3. Can you compare every species on every island? Explain why this matters when comparing islands while holding species fixed.

Use `p_mass` in Sections 2–3 and 5, and `p_bill` in Sections 4 and 7. These samples have different purposes; do not interpret differences between their summaries as an effect of missingness alone.

```
# TODO: add your analysis here.
```

Your interpretation: write here.

2. Boxplots: unusual relative to which group?

Plot body mass by species, then by sex within species. Overlay observations with horizontal jitter only. Use a common vertical scale.

Apply the supplied `flag_iqr()` rule to body mass (a) in the pooled sample, (b) within species and (c) within species–sex groups. Compare the number and identities of the flagged observations. Do not remove them.

Does every flagged point indicate a measurement error? Does a point have to be extreme overall to be extreme within its group? Explain what can be learned even if a particular rule flags no observations.

```
# TODO: add your analysis here.
```

Your interpretation: write here.

3. KDE: bandwidth and population mixtures

Using `p_mass`, compute `h <- bw.nrd0(p_mass$body_mass_g)`. Compare pooled KDEs using bandwidths `h/2`, `h` and `2*h`. Keep the observations fixed and use the same plotting range.

Next plot species-specific KDEs with the same numerical bandwidth `h`. State the units of density. Does a taller curve necessarily identify a species with more observations? Explain why the number of visible peaks need not equal the number of species.

Write the relationship between the pooled density and the conditional densities using the sample species proportions. Explain what changes if you give the species equal weights.

```
# TODO: add your analysis here.
```

Your interpretation: write here.

4. The same scatterplot, with and without species

Plot bill depth against bill length using `p_bill`. First show all points in one colour. Then colour and shape the points by species. Finally make one panel per species, using the same axis limits in all panels. Do not add fitted lines or smoothers.

Describe the direction of the overall point cloud and the direction within each species. Compare the locations of the species clusters with their internal patterns. Can both descriptions be correct?

Calculate Pearson's correlation between bill length and bill depth for the pooled sample and separately for each species, using `cor()`. Report the number of complete pairs alongside each correlation. Use the same `p_bill` rows for the pooled and within-species analyses.

Compare the signs and absolute values of the correlations and relate them to the scatterplots. Does the pooled correlation describe the relationship within each species? Explain what a correlation close to zero would mean for a linear association, and why it would not exclude a non-linear relationship. Use descriptive statements about these observations; no tests, p-values or confidence intervals are requested.

```
# TODO: add your plots and summaries here.
```

Your interpretation: write here.

5. Compare islands within more comparable groups

Question: Do differences in body mass between islands necessarily describe an island effect?

Use `p_mass`, retaining rows with known island, to create `p_island`. Report how many rows are retained. Make a boxplot of mass by island, overlaying points coloured by species. Tabulate the species composition of each island as proportions within island.

Next restrict the comparison to Adelie penguins, which are observed on all three islands. Plot body mass by island again. Finally compare the islands separately within female and male Adelie penguins, showing subgroup sizes and using a common vertical scale.

Describe which comparisons changed when you held species and then sex fixed. Why can you not repeat the three-island comparison for every species? Can these plots establish a causal effect of island?

```
# TODO: add your plots and summaries here.
```

Your interpretation: write here.

6. Explain what conditioning changes

Choose three figures from your analysis and write a short explanation for another student. Your explanation should connect:

1. How the chosen reference group affects a boxplot flag.

2. How bandwidth and group composition affect the shape of a KDE.
3. How the pooled correlation differs from the correlations within species, and how overall island comparisons differ from comparisons within species and sex. Refer to the observed correlation values and their signs.

Finish with one conclusion supported by the plots and one causal conclusion that they do not justify. No new calculation or statistical model is required.

Your interpretation: write here.

7. Optional: bivariate KDE

Overlay bivariate density contours on the bill scatterplot. Compare one pooled KDE with species-specific KDEs. Use the same bandwidth settings and coordinate limits in both plots. Investigate whether the pooled contours conceal the within-species orientation.

Explain what a contour of constant estimated density represents. Does a contour height directly give the proportion of observations inside it? Can the number of visible peaks alone identify the number of species?

```
# TODO: add your analysis here.
```

Your interpretation: write here.

Looking ahead: from simple to multiple linear regression

Remark — connecting with what you already know. You have studied simple linear regression. Here we describe the direction and strength of linear associations using Pearson correlations and scatterplots. Later, multiple linear regression will allow you to consider several variables together and handle continuous covariates without dividing them into many small groups. The comparisons within species and sex provide a graphical introduction to that idea; no multiple regression is needed in this activity.

Submission and reflection

Submit the completed Rmd and its rendered HTML. Include the requested analyses, readable figures and a short final answer to the main question. Explain one conclusion that changes when you condition on another variable, and one conclusion that the data do not justify. Label the optional extension separately.

Data and references

- Horst, A. M., Hill, A. P., & Gorman, K. B. (2022). *Palmer Archipelago Penguins Data in the palmerpenguins R Package — An Alternative to Anderson's Irises.* The R Journal.
- Dataset documentation and provenance. Data are available under CC0. Original field data were collected by Kristen Gorman and the Palmer Station Long Term Ecological Research programme. This activity uses a CSV copy of the curated `penguins` dataset, retaining its original variable names. Providing a CSV changes the access method, not the source of the data; no installation of `palmerpenguins` is needed.
- Univariate KDE in ggplot2 and bivariate density contours.