Summarizing Categorical Data: Tables and Plots

Emile Latour, Nicky Wakim, Meike Niederhausen

January 12, 2026

Today's plan

1. Summarizing categorical data (tables + bar plots)
2. R basics: writing and running code in Quarto

Summarizing categorical data

- Last week we looked at:
  - Variable types
  - Summarizing numerical variables
- Today, we focus on summarizing and *describing* categorical variables.

What do we mean by categorical?

- Values are labels or categories
- Counts and proportions matter more than averages
- Examples from biomedicine:
  - Treatment group
  - Diagnosis
  - Sex
  - Response category (e.g., improved / unchanged / worsened)

- **Nominal** – labels, no order
- **Ordinal** – labels, with order

Example dataset: FAMuSS

Functional SNPs Associated with Muscle Size and Strength (FAMuSS)

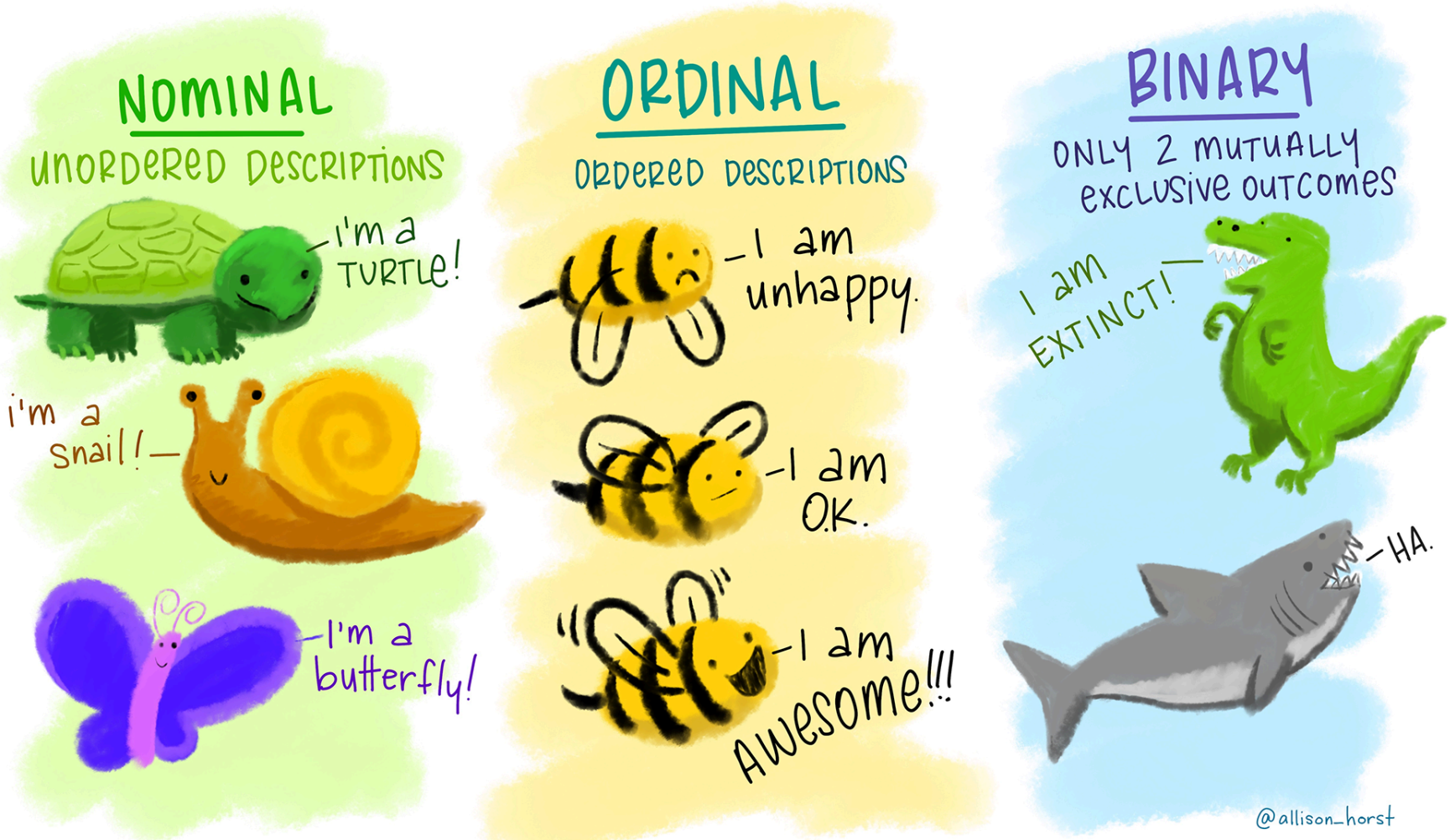
- Study goal: examine how **demographic, physiological, and genetic factors** are associated with muscle strength
- Strength measured in:
  - Dominant arm
  - Non-dominant arm
  - Before and after resistance training
- Key gene of interest:
  - **ACTN3** ("the sports gene")
- Data frame with **595 participants**

Variables we will focus on today:

- `sex` (Female, Male)
- `race` (African Am, Asian, Caucasian, Hispanic, Other)
- `actn3.r577x` (CC, CT, TT genotype)

(We will use other variables later in the course.)

One categorical variable (Section 1.5)



Artwork by @allison_horst

Frequency tables

- A frequency table shows the **count** in each category
- Often the first summary we compute for categorical data

- **Questions** it helps answer:
  - How many observations are in each category?
  - Are some categories rare?

Relative frequency tables

- A relative frequency table shows **proportions** instead of counts
- Proportions often make comparisons easier
- Especially useful when:
  - Group sizes differ
  - We want to compare across studies or samples

FAMuSS example

Counts

actn3.r577x	n
CC	173
CT	261
TT	161
Total	595

Proportions

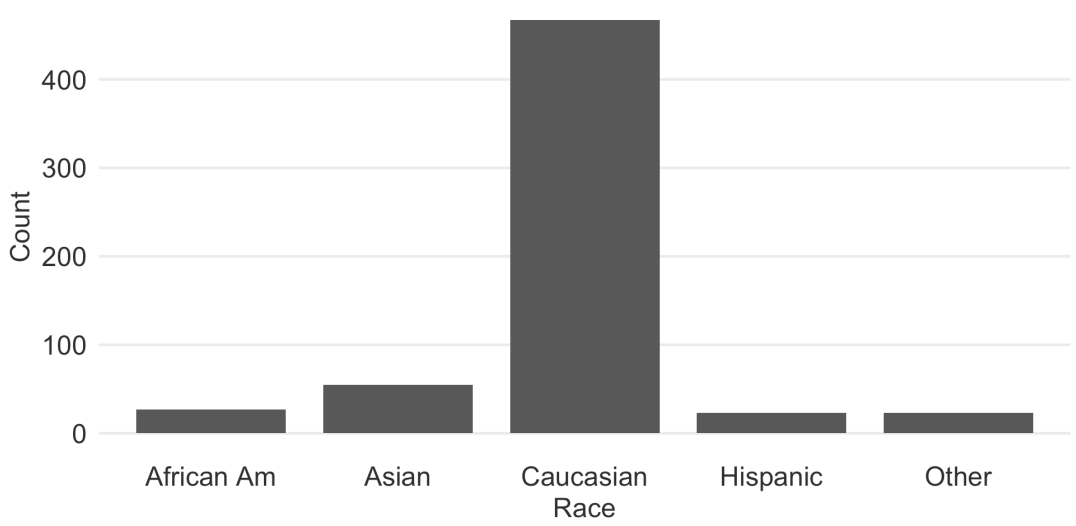
actn3.r577x	n	percent
CC	173	29.1%
CT	261	43.9%
TT	161	27.1%
Total	595	100.0%

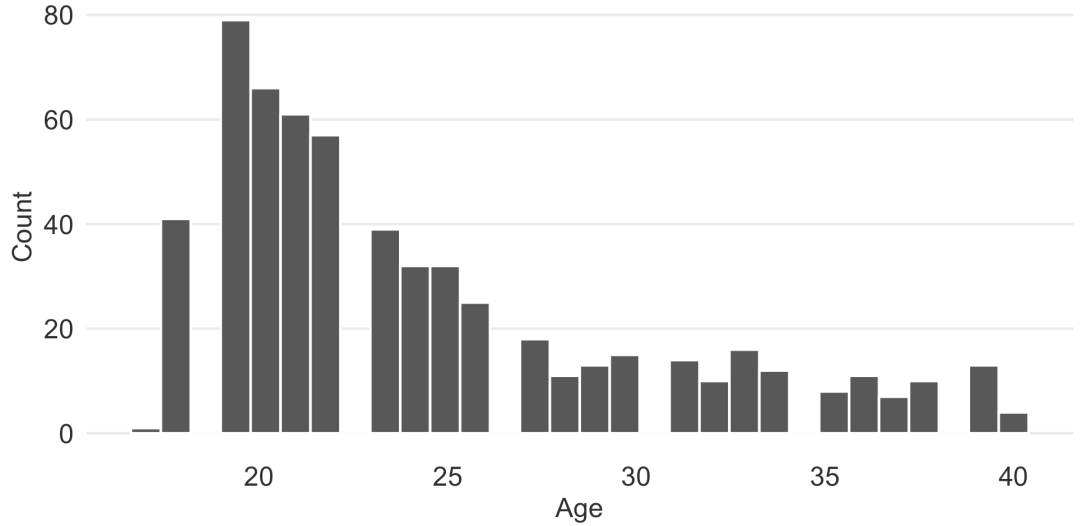
Bar plots for categorical data

- Bar plots visualize counts or proportions
- Each bar represents a category
- Bar height reflects frequency or proportion

Bar plots are used for **categorical** data

Histograms are used for **numerical** data

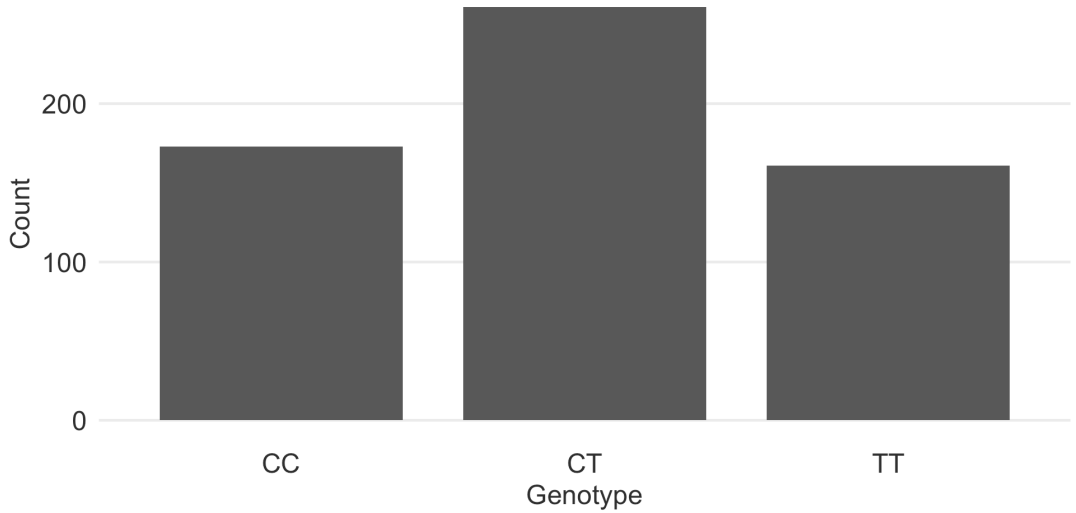




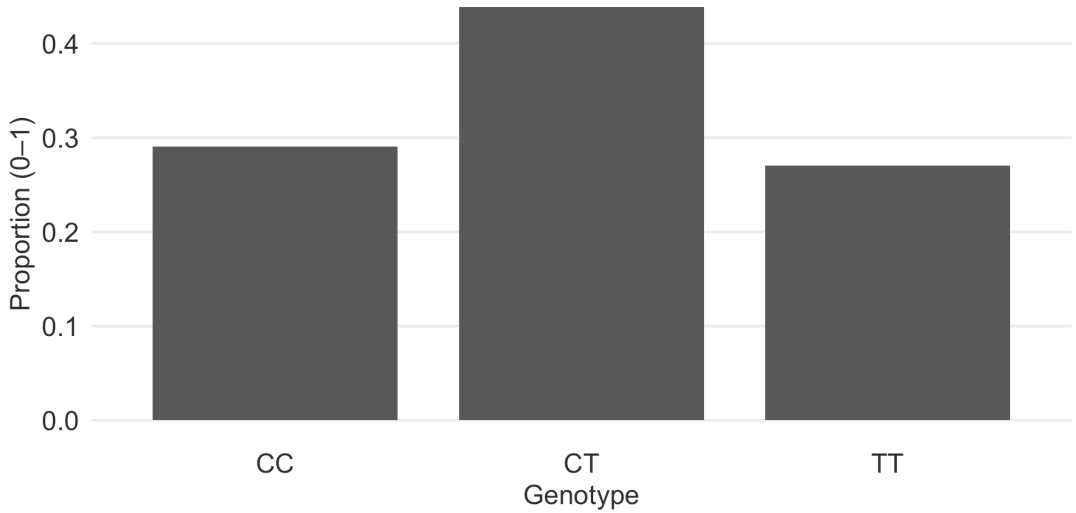
Visualizing a categorical variable: bar plots

Height = **count** or **proportion**

Counts



Proportions



Two categorical variables (Section 1.6.2)

- So far, we have summarized **one categorical variable at a time**
  - Counts
  - Proportions
  - Bar plots
- Often, we want to understand the relationship between **two categorical variables**

- **Examples**:
  - Genotype and sex
  - Treatment group and response
  - Exposure and disease status

Contingency tables

- When we have **two categorical variables**, we summarize them with a **contingency table** (also called a **two-way table**)

- Each cell shows the **count** for a combination of categories

- Rows represent one variable
- Columns represent the other variable

Contingency table example (counts)

- **Example question**:
  - Does the distribution of genotypes differ by sex?

- These are **counts**, not proportions
- Totals appear along the margins

sex	CC	CT	TT	Total
Female	106	149	98	353
Male	67	112	63	242
Total	173	261	161	595

Marginal totals vs conditional distributions

- **Marginal totals**
  - Summarize **one variable at a time**
  - Ignore the other variable
  - Found in the **row totals** or **column totals**

- **Conditional distributions**
  - Describe one variable **within levels of the other**
  - Require computing **proportions**

sex/actn3.r577x	CC	CT	TT	Total
Female	106	149	98	353
Male	67	112	63	242
Total	173	261	161	595

Row proportions vs column proportions (1/2)

- Which one you use depends on the **question**
- **Row proportions**
  - Condition on the **row variable**
  - Each row sums to 1 (or 100%)
- **Example question**: Among females, what proportion have genotype CC?

sex/actn3.r577x	CC	CT	TT	Total
Female	106	149	98	353
Male	67	112	63	242

sex/actn3.r577x	CC	CT	TT	Total
Female	0.30	0.42	0.28	1.00
Male	0.28	0.46	0.26	1.00

Row proportions vs column proportions (2/2)

- **Column proportions**
  - Condition on the **column variable**
  - Each column sums to 1 (or 100%)
- **Example question**: Among those with genotype CC, what proportion are female?

sex/actn3.r577x	CC	CT	TT
Female	106	149	98
Male	67	112	63
Total	173	261	161

sex/actn3.r577x	CC	CT	TT
Female	0.61	0.57	0.61
Male	0.39	0.43	0.39
Total	1.00	1.00	1.00

Interpreting contingency tables

- Always ask:
  - What are the **rows**?
  - What are the **columns**?
  - What is being **held fixed**?

- Interpretation depends on:
  - The research question
  - Which variable you condition on

Common interpretation pitfalls

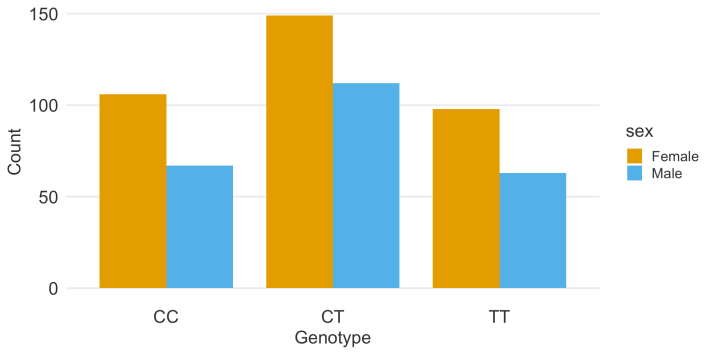
- Confusing marginal totals with conditional distributions
  - Marginal totals describe the sample overall
  - Conditional distributions describe relationships (what happens within groups)
- Comparing counts when group sizes differ
- Forgetting which variable is being conditioned on

Visualizing two categorical variables

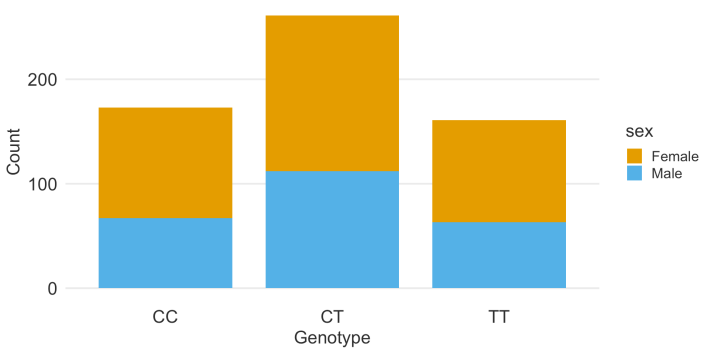
- Contingency tables show the numbers
- Plots help reveal patterns

Bar plots example: Sex by genotype

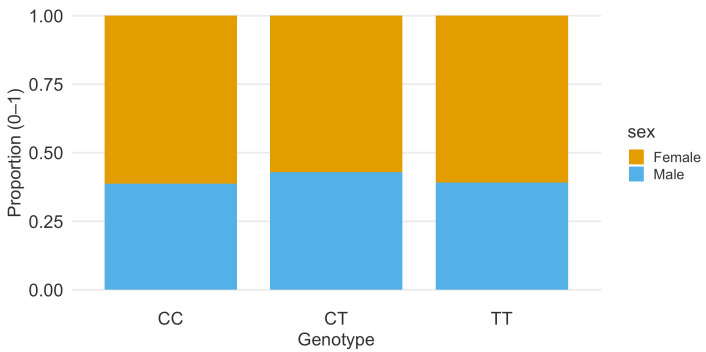
Grouped bar plot



Stacked bar plot

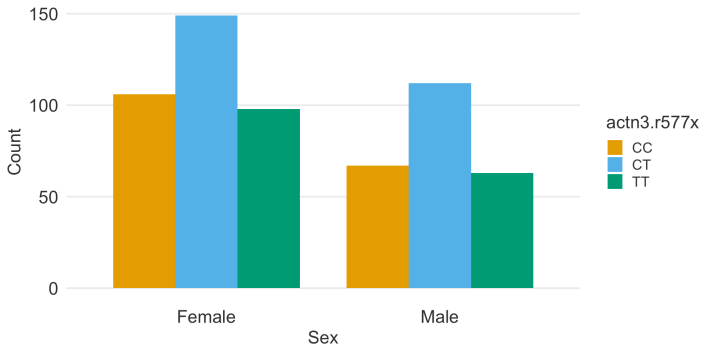


Percent stacked bar plot

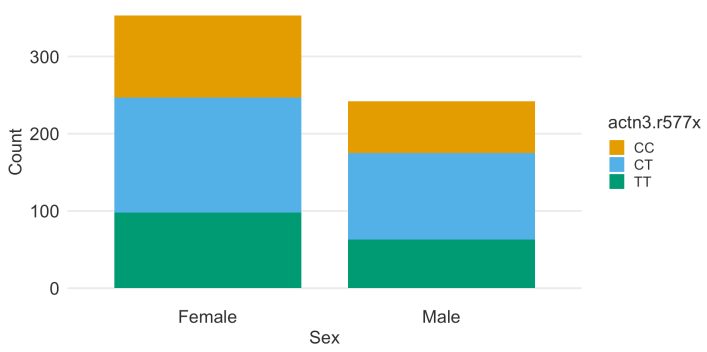


Bar plots example: Genotype by sex

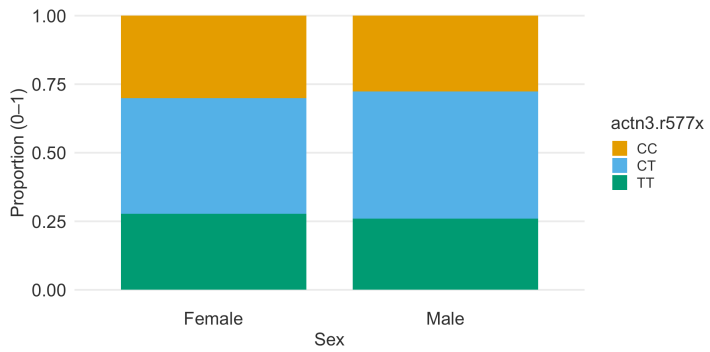
Grouped bar plot



Stacked bar plot



Percent stacked bar plot



Special case: two-by-two tables

- A **two-by-two table** is a contingency table with:
  - Two levels of one variable
  - Two levels of another variable

- Very common in biomedical research:
  - Exposure (Yes / No) × Outcome (Yes / No)
  - Treatment (Drug / Control) × Response (Improved / Not improved)
  - Test result (Positive / Negative) × Disease status (Present / Absent)

- Today:
  - Focus on **structure and interpretation**
- Later:
  - Risk, odds, probability, and inference

A quick note on numeric variables

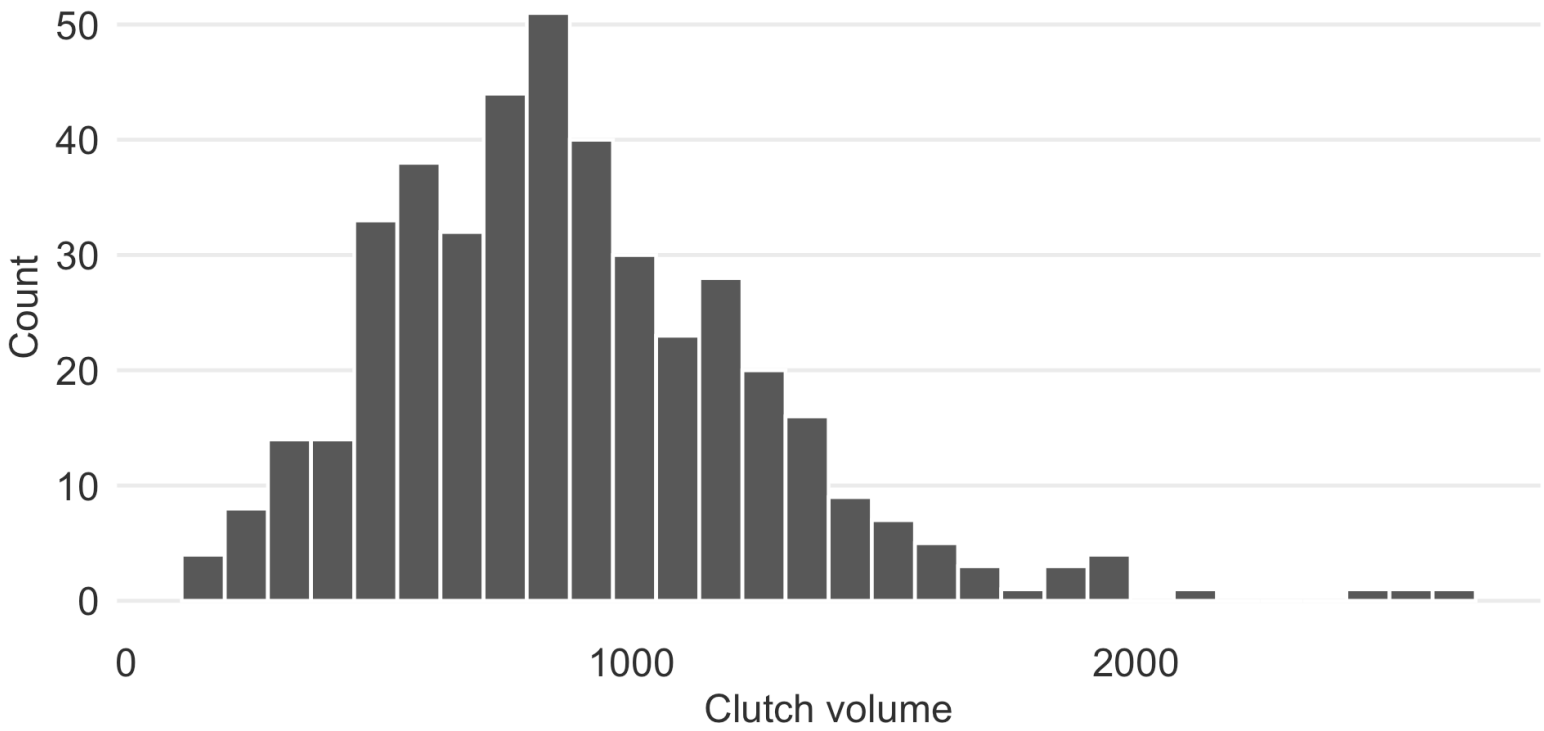
Last week, we summarized numeric variables using:

- mean and standard deviation
- median and IQR

We can also **summarize numeric variables visually**, just like we did for categorical data.

Visualizing a numeric variable: histogram

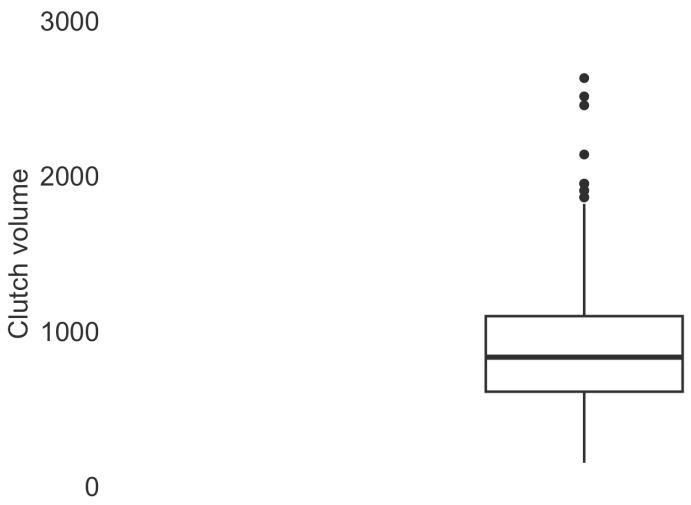
- Histograms show the **distribution** of a numeric variable
- Useful for seeing:
  - shape (symmetric vs skewed)
  - outliers
  - clusters



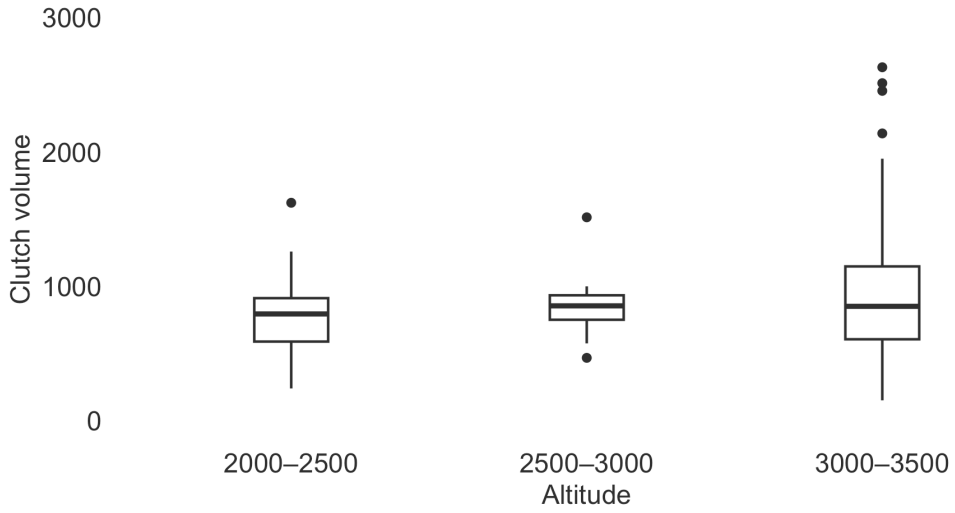
Visualizing a numeric variable: box plot

- Box plots summarize a numeric variable using:
  - median
  - IQR
  - potential outliers

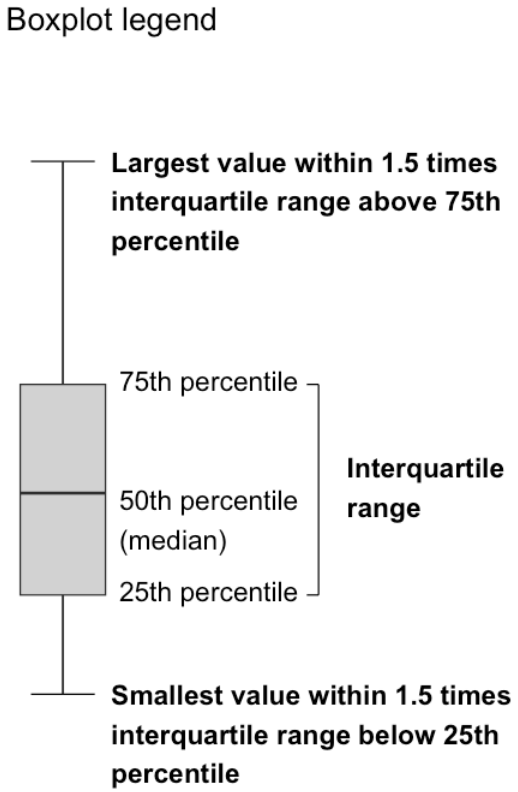
Box plot



Box plot by groups



Box plot legend



- Potential outlier; value more than 1.5 times and less than 3 times the interquartile range beyond either end of the box

Wrap-up

Today you learned how to:

- Summarize categorical variables (counts + proportions)
- Compare two categorical variables (contingency tables)
- Interpret row vs column percentages
- Use bar plots, histograms, and box plots as quick visual summaries