

# **Applied Data Analytics**

## **Statistics — Measures for bivariate data**

### **Strategies for summarising bivariate data**

Hans-Martin von Gaudecker and Aapo Stenhammar

# Histograms

Raw data

|   | x | y |
|---|---|---|
| 0 | c | R |
| 1 | a | S |
| 2 | b | S |
| 3 | a | T |
| 4 | b | R |
| 5 | b | S |

(Marginal) Frequencies

| x | count |
|---|-------|
| a | 2     |
| b | 3     |
| c | 1     |

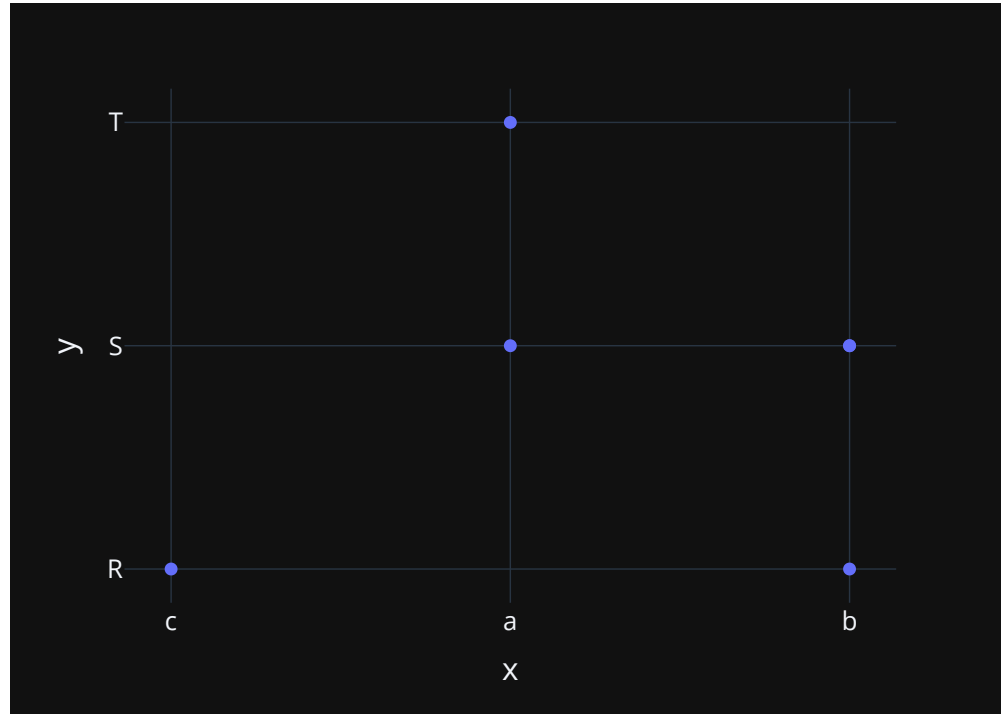
| y | count |
|---|-------|
| R | 2     |
| S | 3     |
| T | 1     |

# Joint frequencies

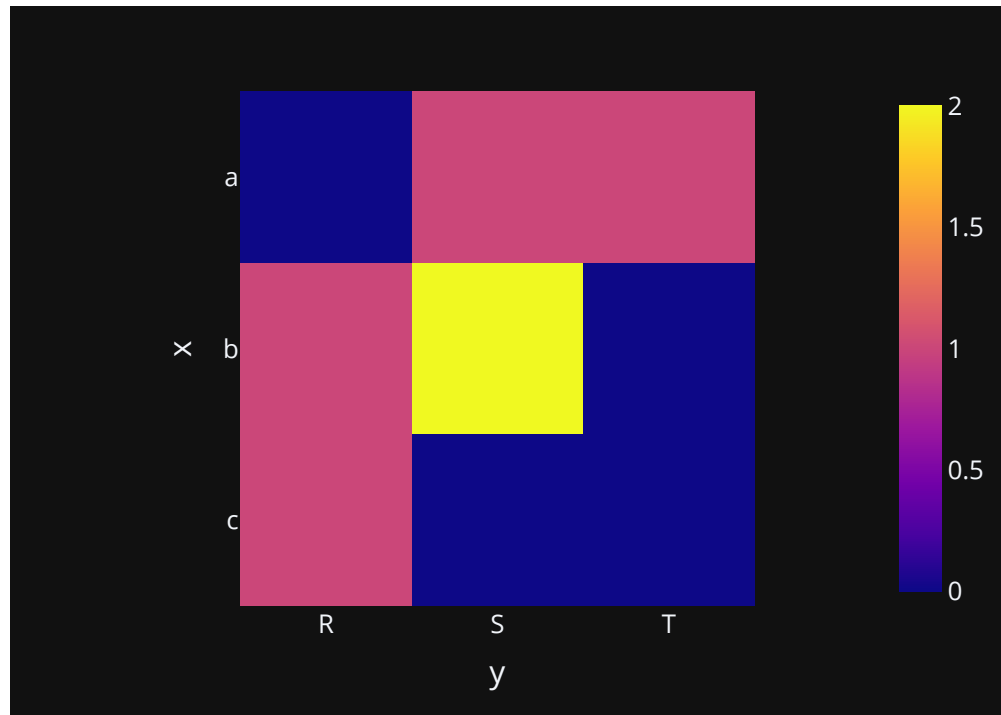
```
pd.crosstab(df["x"], df["y"])
```

| x | R | S | T |
|---|---|---|---|
| a | 0 | 1 | 1 |
| b | 1 | 2 | 0 |
| c | 1 | 0 | 0 |

# Scatter plot, discrete



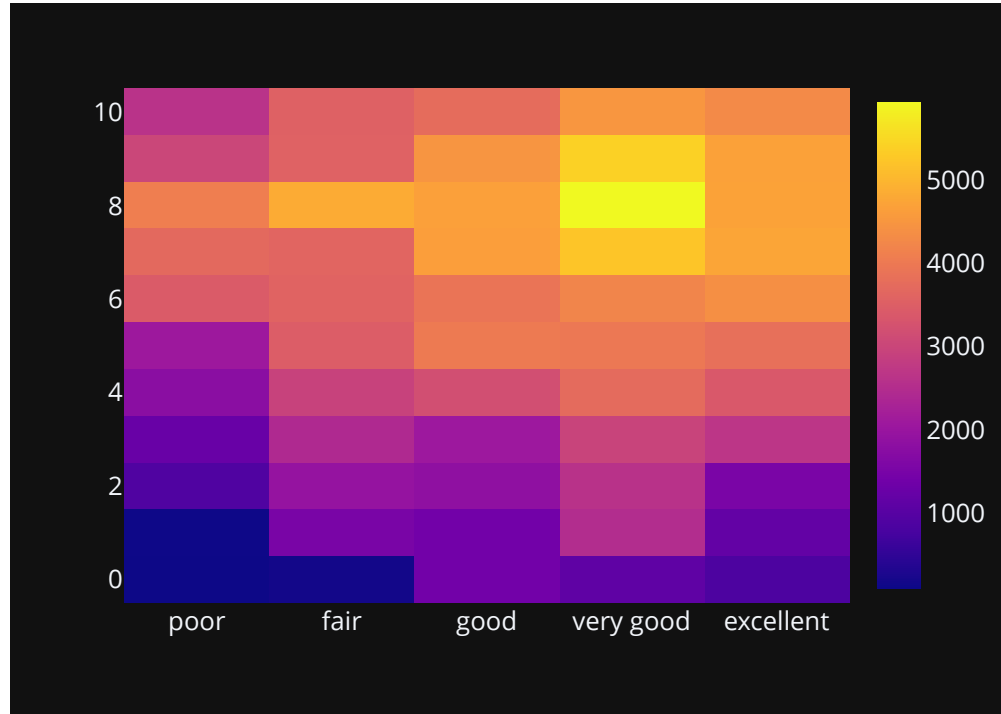
# Heatmap, small



# Survey on SRH & life satisfaction

|    | poor | fair | good | very good | excellent |
|----|------|------|------|-----------|-----------|
| 0  | 179  | 407  | 879  | 1013      | 1411      |
| 1  | 853  | 1351 | 1031 | 2028      | 1498      |
| 2  | 1319 | 1636 | 2291 | 2803      | 2017      |
| 3  | 1355 | 1610 | 2817 | 3358      | 2436      |
| 4  | 2370 | 2872 | 2761 | 3930      | 2751      |
| 5  | 2213 | 3172 | 3148 | 3513      | 3454      |
| 6  | 2955 | 3619 | 3786 | 4711      | 4331      |
| 7  | 3930 | 4482 | 4424 | 4683      | 4913      |
| 8  | 4201 | 4474 | 5074 | 5646      | 5499      |
| 9  | 3606 | 4115 | 4488 | 5075      | 4649      |
| 10 | 2979 | 3314 | 3682 | 4850      | 4297      |

# Heatmap, large



# Heatmap including contingency table

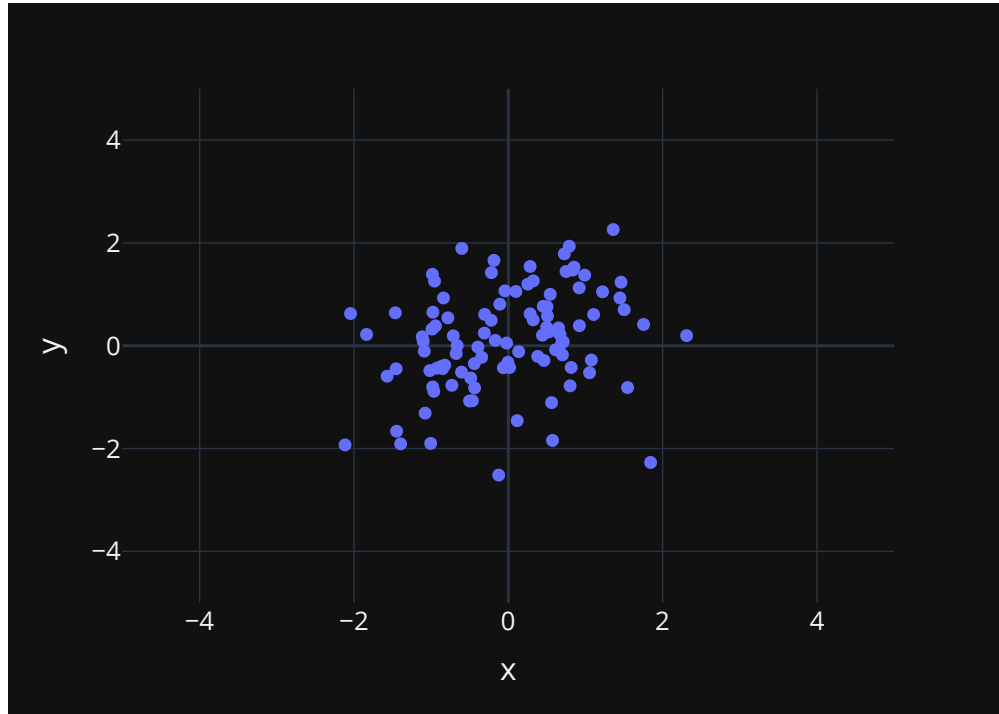


# Continuous data

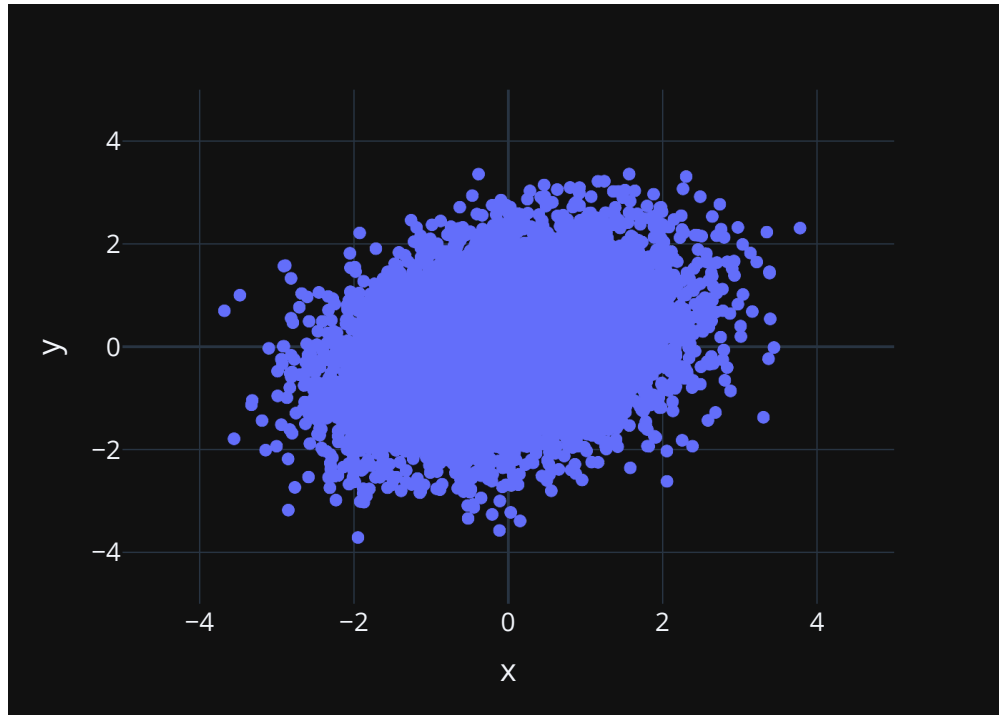
Raw data

|     | x         | y          |
|-----|-----------|------------|
| 0   | -0.231075 | 0.103152   |
| 1   | 0.999491  | -0.0753982 |
| 2   | 0.41889   | 0.115341   |
| ... | ...       | ...        |
| 97  | 1.36987   | -1.22625   |
| 98  | -0.413597 | 0.507763   |
| 99  | -1.45686  | -0.151736  |

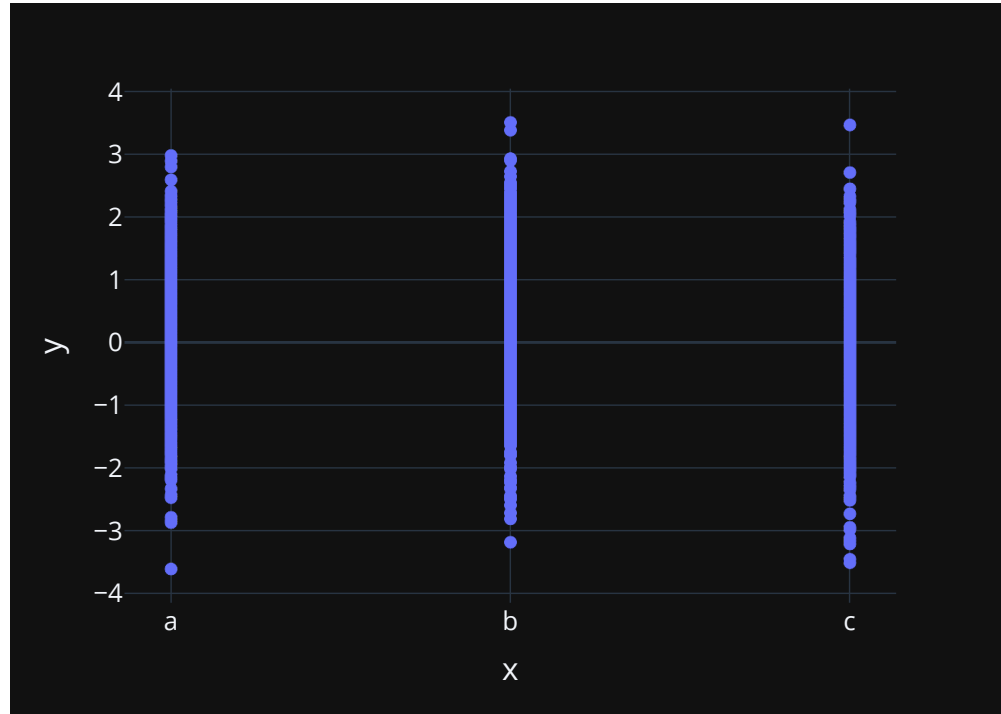
# Scatter plot, continuous, $n=100$



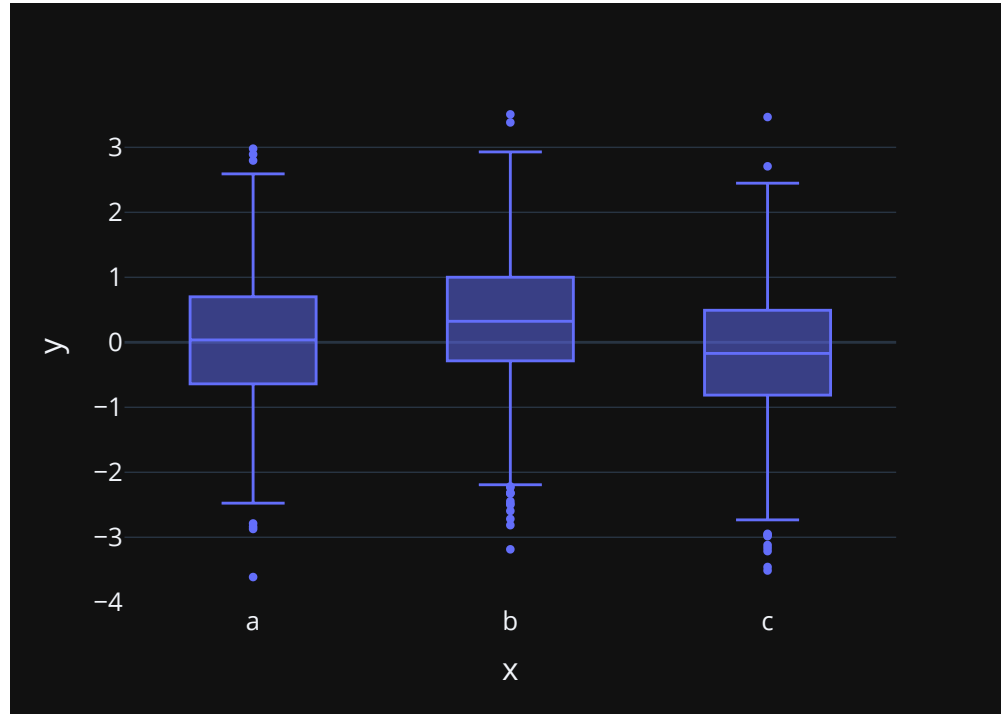
# Scatter plot, continuous, $n=10,000$



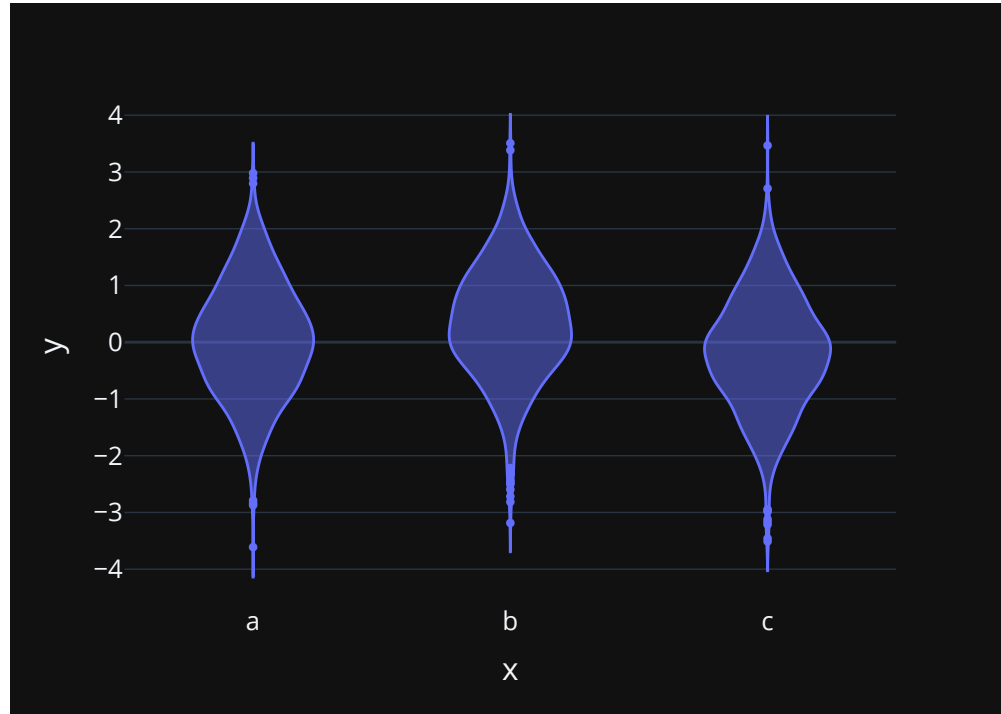
# Scatter plot, discrete + continuous



# Scatter plot, discrete + continuous



# Scatter plot, discrete + continuous



# Which strategy to use?

- **X** discrete
  - **Y** discrete: Contingency table, heatmaps
  - **Y** continuous: Box plots, violin plots
- **X** continuous
  - **Y** continuous: Scatter plots, correlations
- Always look at marginal distributions, too!