FOUNDATION LESSON

WEEK 5 · LESSON 9

# Distributions, histograms and box plots

A distribution describes how values are spread, including unusual observations.

THE MENTAL MODEL

MENTAL MODEL

# The mental model

A distribution describes how values are spread, including unusual observations.

# HISTOGRAM WITH STATED BINS

CODE WALKTHROUGH W05-L09-S03

```
01  import matplotlib
02  matplotlib.use("Agg")
03  import matplotlib.pyplot as plt
04  plt.rcParams.update({"font.size": 16})
05  values=[10,11,12,13,100]
06  fig,ax=plt.subplots(figsize=(12,6.75))
07  ax.hist(values,bins=[0,20,40,60,80,120],edgecolor="black")
08  ax.set(title="Invoice distribution · 5 observations",
09         xlabel="Invoice amount (EUR)",ylabel="Invoice count")
10  fig.tight_layout()
11  fig.savefig("figure.png", dpi=320)
12  plt.close(fig)
13  print("Saved figure.png: 3840 x 2160")
```

## WHAT TO TRACK

Histogram with stated bins

## PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

## HISTOGRAM WITH STATED BINS · RESULT

CODE WALKTHROUGH W05-L09-S04

```
01  import matplotlib
02  matplotlib.use("Agg")
03  import matplotlib.pyplot as plt
04  plt.rcParams.update({"font.size": 16})
05  values=[10,11,12,13,100]
06  fig,ax=plt.subplots(figsize=(12,6.75))
07  ax.hist(values,bins=[0,20,40,60,80,120],edgecolor="black")
08  ax.set(title="Invoice distribution · 5 observations",
09         xlabel="Invoice amount (EUR)",ylabel="Invoice count")
10  fig.tight_layout()
11  fig.savefig("figure.png", dpi=320)
12  plt.close(fig)
13  print("Saved figure.png: 3840 x 2160")
```

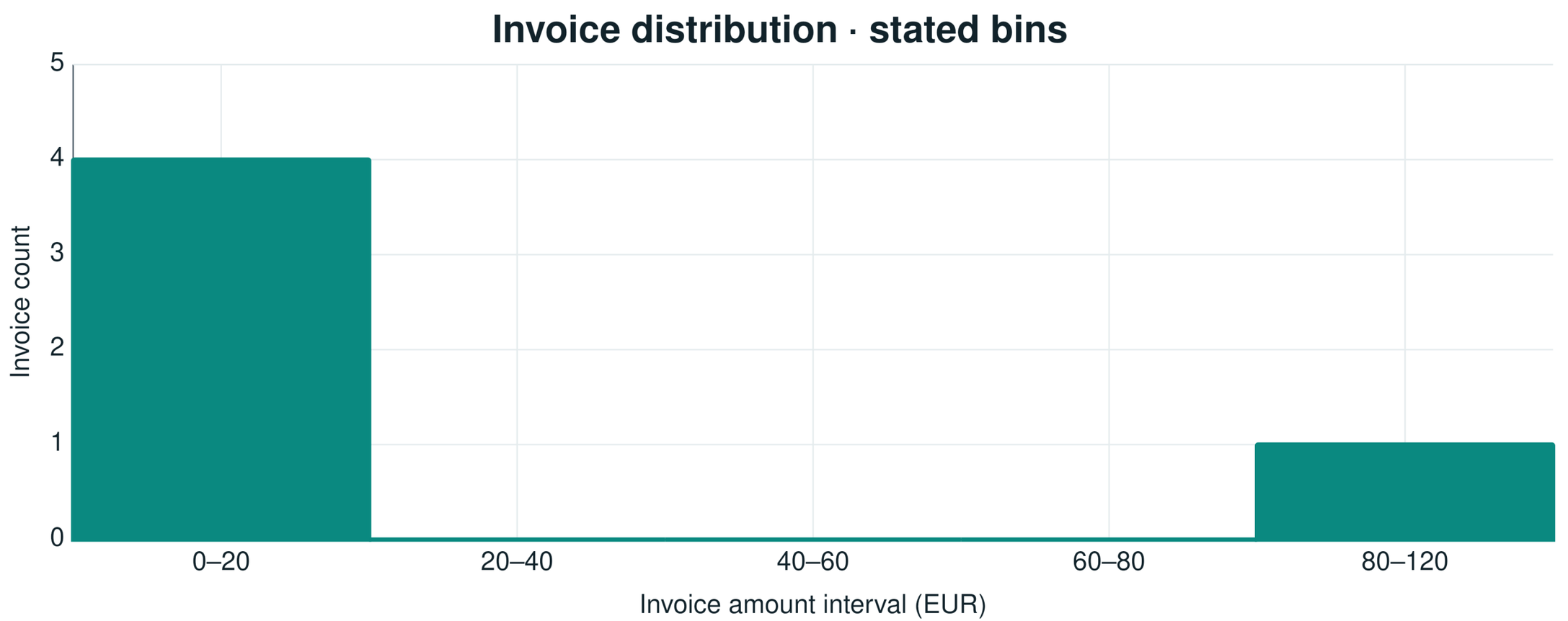
### WHAT TO TRACK

Histogram with stated bins

### VALIDATE THE RESULT

```
Saved figure.png: 3840 x
2160
```

VISIBLE CHART RESULT



# BOX PLOT AND ROBUST SUMMARY

CODE WALKTHROUGH W05-L09-S06

```
01  import matplotlib
02  matplotlib.use("Agg")
03  import matplotlib.pyplot as plt
04  plt.rcParams.update({"font.size": 16})
05  import pandas as pd
06  values=[10,11,12,13,100]
07  fig,ax=plt.subplots(figsize=(12,6.75))
08  ax.boxplot(values,orientation="horizontal")
09  ax.set(title="Invoice amounts · extreme value retained",
10         xlabel="Invoice amount (EUR)")
11  print("Median:",pd.Series(values).median())
12  print("Mean:",pd.Series(values).mean())
13  fig.tight_layout()
14  fig.savefig("figure.png", dpi=320)
15  plt.close(fig)
16  print("Saved figure.png: 3840 x 2160")
```

WHAT TO TRACK

Box plot and robust summary

PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

## BOX PLOT AND ROBUST SUMMARY · RESULT

CODE WALKTHROUGH W05-L09-S07

```
01 import matplotlib
02 matplotlib.use("Agg")
03 import matplotlib.pyplot as plt
04 plt.rcParams.update({"font.size": 16})
05 import pandas as pd
06 values=[10,11,12,13,100]
07 fig,ax=plt.subplots(figsize=(12,6.75))
08 ax.boxplot(values,orientation="horizontal")
09 ax.set(title="Invoice amounts · extreme value retained",
10        xlabel="Invoice amount (EUR)")
11 print("Median:",pd.Series(values).median())
12 print("Mean:",pd.Series(values).mean())
13 fig.tight_layout()
14 fig.savefig("figure.png", dpi=320)
15 plt.close(fig)
16 print("Saved figure.png: 3840 x 2160")
```

### WHAT TO TRACK

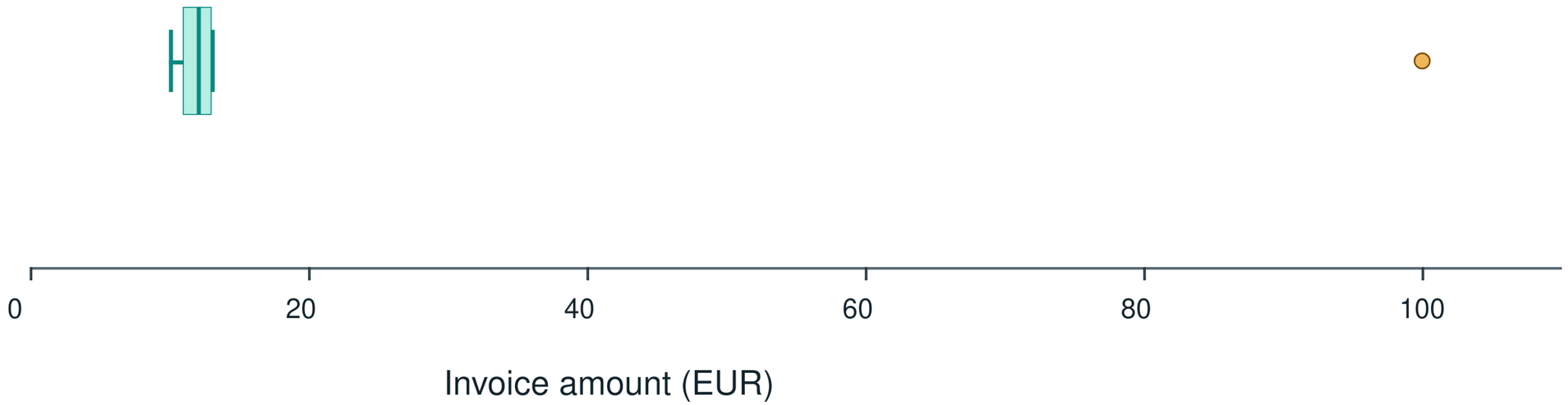
Box plot and robust summary

### VALIDATE THE RESULT

```
Median: 12.0
Mean: 29.2
Saved figure.png: 3840 x
2160
```

VISIBLE CHART RESULT

# Invoice amounts · extreme value retained




Q1=11 · Median=12 · Q3=13 · Whiskers=10,13 · Outlier=100 retained

# ONE OBSERVATION IS INSUFFICIENT FOR SPREAD

CODE WALKTHROUGH W05-L09-S09

```
01  import pandas as pd
02  s=pd.Series([12.0])
03  print("Count:",s.count())
04  print("Median:",s.median())
05  print("Sample SD:",s.std(ddof=1))
```

WHAT TO TRACK

One observation is insufficient for spread

PREDICT THE RESULT

Predict the output.
Explain each operation.
Pause before continuing.

# ONE OBSERVATION IS INSUFFICIENT FOR SPREAD · RESULT

CODE WALKTHROUGH W05-L09-S10

```
01  import pandas as pd
02  s=pd.Series([12.0])
03  print("Count:",s.count())
04  print("Median:",s.median())
05  print("Sample SD:",s.std(ddof=1))
```

WHAT TO TRACK

One observation is insufficient for spread

VALIDATE THE RESULT

```
Count: 1
Median: 12.0
Sample SD: nan
```

GUIDED PRACTICE

LEARNER PRACTICE

# Guided practice

For [2,3,4,5,50], compute mean, median, first and third quartiles. Keep every observation.

PREDICT · MODIFY · RUN

01 Restate the supplied rules.

02 Predict the result.

03 Modify and run the code.

04 Compare actual output.

05 Explain any difference.

INDEPENDENT PRACTICE

LEARNER PRACTICE

# Independent practice

Create a histogram for [1,1,2,3,8] with bin boundaries [0,2,4,6,10]. Label counts and measurement units.

PREDICT · MODIFY · RUN

01 Restate the supplied rules.

02 Predict the result.

03 Modify and run the code.

04 Compare actual output.

05 Explain any difference.

CHECK YOUR UNDERSTANDING

MENTAL MODEL

# Check your understanding

Diagnostic: Does a box-plot outlier prove a data error? No. It is a rule-based flag that needs contextual investigation.

RECAP AND NEXT STEP

# Recap and next step

Inspect distributions, state histogram bins, compare mean with median, and investigate extreme observations before changing them.

UNDERSTAND APPLY VALIDATE EXPLAIN