Lecture 1: Graphing 1 Variable

Intro Stats with Nathan Favero

American University (Washington, DC)

August 2, 2024

Some important concepts

- Observation: one unit (e.g., event, object, or person) in a dataset; often represented as a row of data
- Variable: characteristics of observations that can take on different values or amounts; often represented as a column of data

Some important concepts

- Qualitative variable (also called categorical or nominal): a variable consisting of categories that are not obviously ordered in any way
  - Examples: race (white, black, Latino, other), industry (manufacturing, agriculture, technology, etc.), pregnancy (pregnant, not pregnant)
- Quantitative variable: a variable where values have a numerical ordering (can be arranged from least to greatest)
  - Examples: age, income, education level

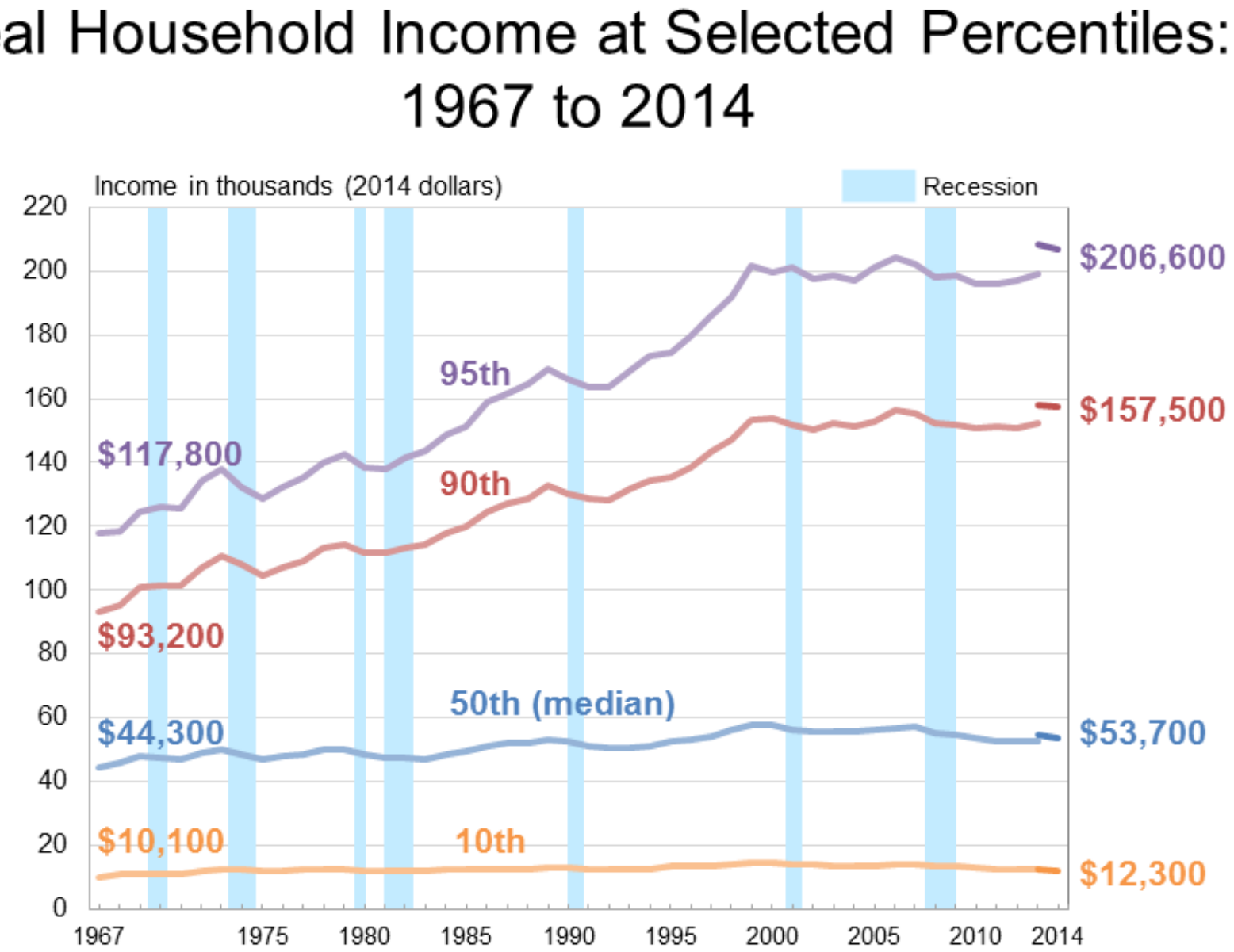
Some important concepts

- Three types of quantitative variables:
  - Ordinal variable: a variable with values that can be clearly arranged from least to greatest but where the distances between values cannot be easily quantified
    - Examples: education (high school diploma or less, some college, 4-year college degree, advanced degree), Likert scales (strongly disagree, somewhat disagree, somewhat agree, strongly agree)
  - Discrete variable: a variable that can only take on certain values within a range (e.g., whole numbers) and that is not ordinal (with a discrete variable, one can quantify the distance between different values)
    - Examples: number of children, number of visits to office
  - Continuous variable: a variable that can be measured to many decimal places (at least theoretically)
    - Examples: weight, income, response time

Percentiles

- Basic definition: *p*% of the data falls (at or) below the *p*th percentile
  - Example: if the 25th percentile of age is 30, then 25% of the observations are younger than (or equal to) 30 years old; conversely, 75% of the observations must be older than (or equal to) 30
- To find the *p*th percentile, line all values up from least to greatest and then pick the value that marks the *p*% point in this lineup (e.g, for 50th percentile, pick the middle point)
- There are various methods that can be used to deal with ties, rounding, etc., but we won't worry about learning those for this class

Percentiles: example

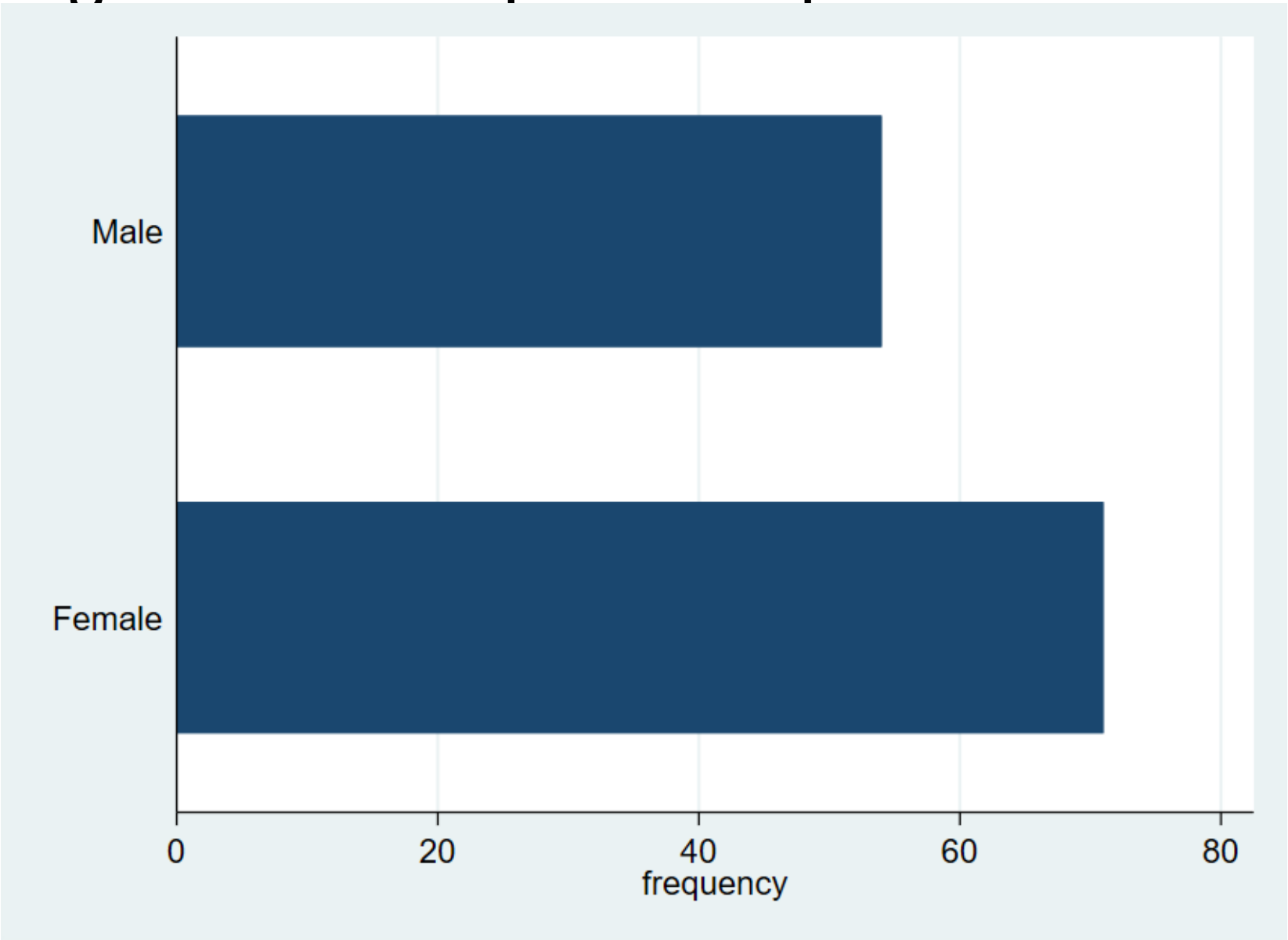


(Image: Public domain, U.S. Census Bureau)

Histograms

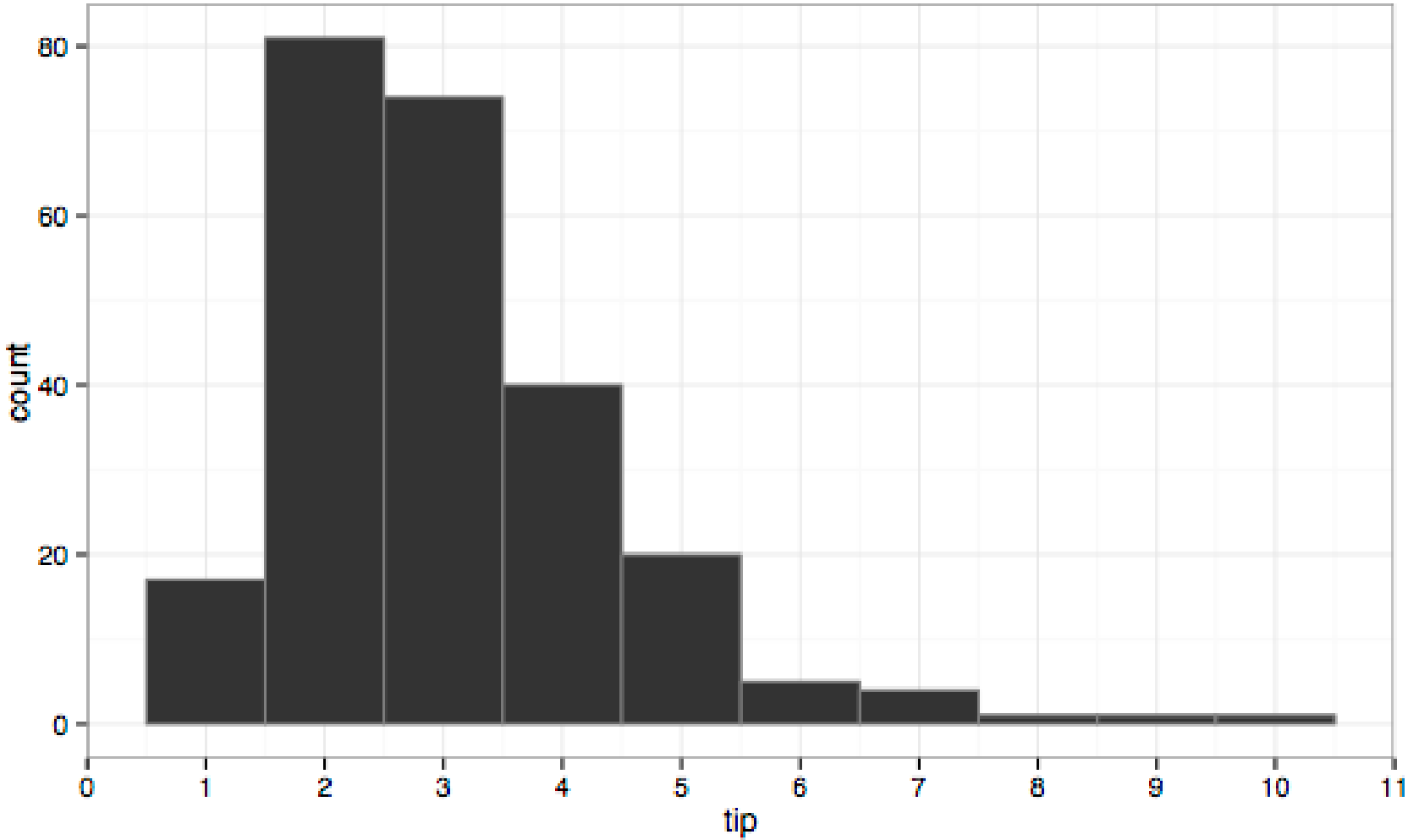
- A type of bar chart
- Each bar represents either (1) a single value (for qualitative, ordinal, or discrete variables) or (2) a range of values (known as a “bin”) (for continuous variables)
- The height of each bar represents how many data points (1) are equal to the corresponding value or (2) fall within the corresponding range of values (bin)

Histograms: example with qualitative variable



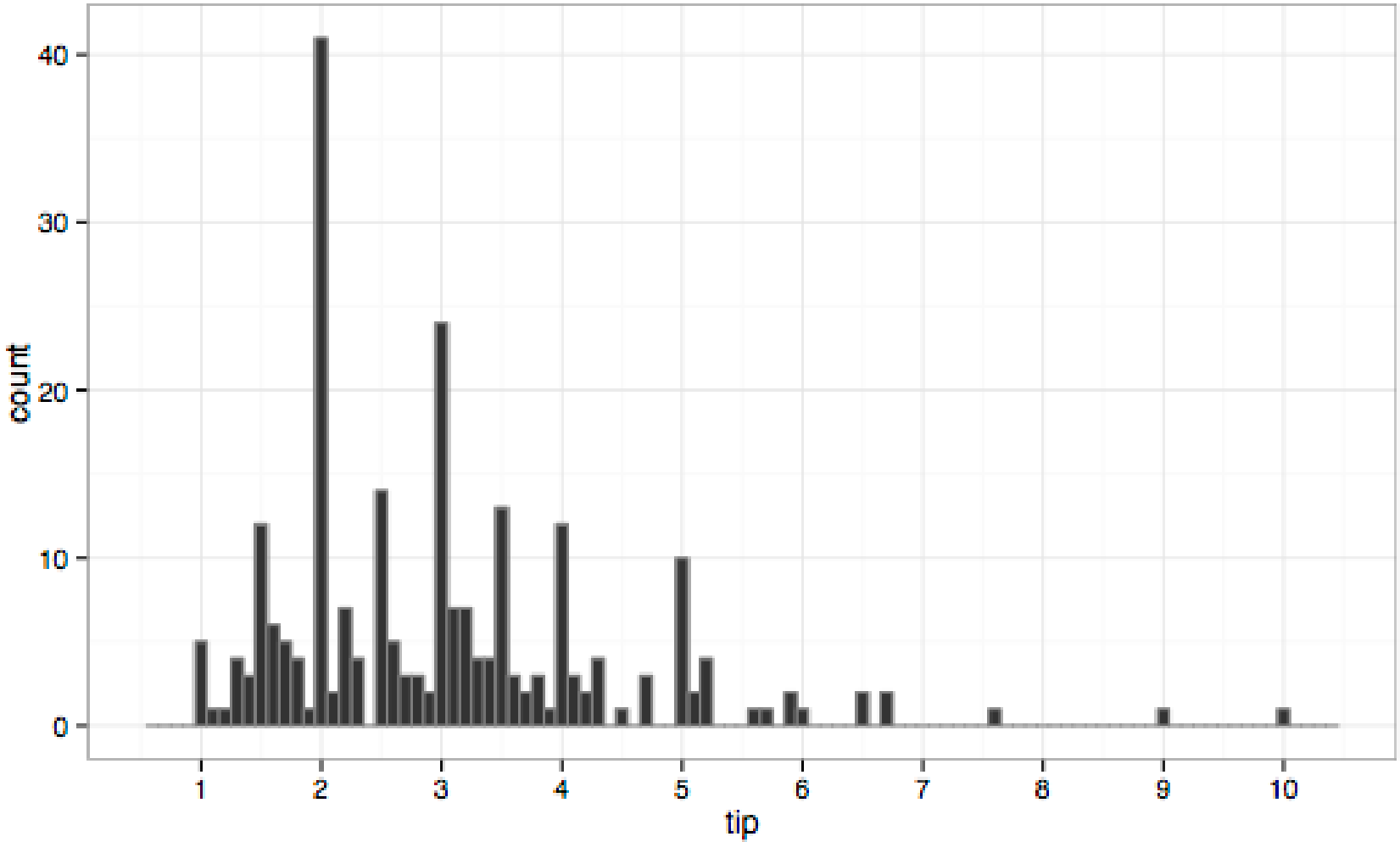
(Image: Public domain, `https://onlinestatbook.com`)

Histograms: example with continuous variable ($1 bin width)



(Image: CC BY-SA 4.0, “Tips-histogram1.png” by Visnut)

Histograms: example with continuous variable (10-cent bin width)

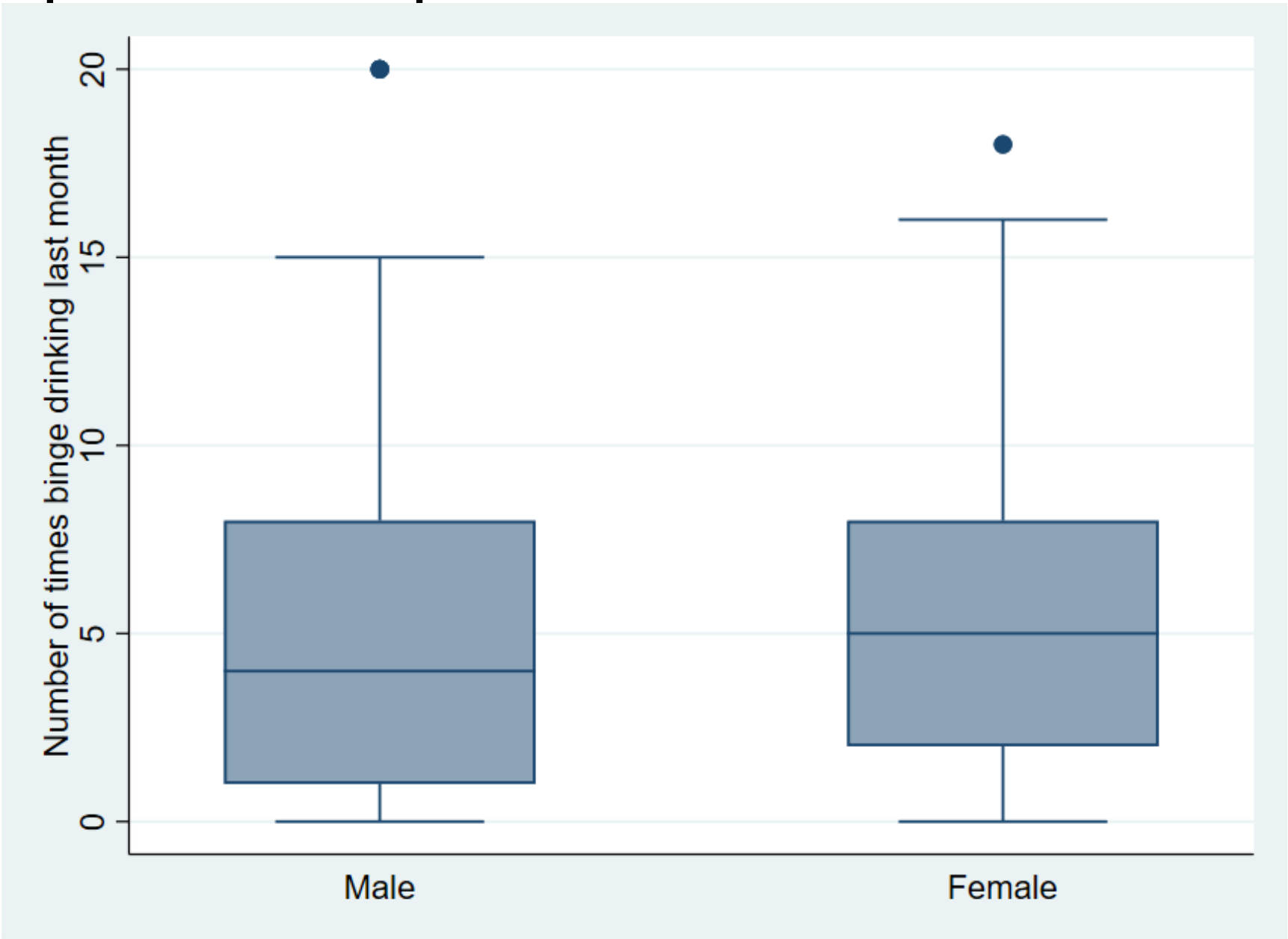


(Image: CC BY-SA 4.0, “Tips-histogram2.png” by Visnut)

Boxplots

- Another way of depicting continuous variables
- Especially useful for making comparisons (e.g., do college men binge drink more than women?)
- Middle line is the 50th percentile (median)
- Edges of box are 25th and 75th percentiles
- “Whiskers” usually indicate range (max and min), although outliers (extreme values) might be ignored
- Outliers may be shown as dots

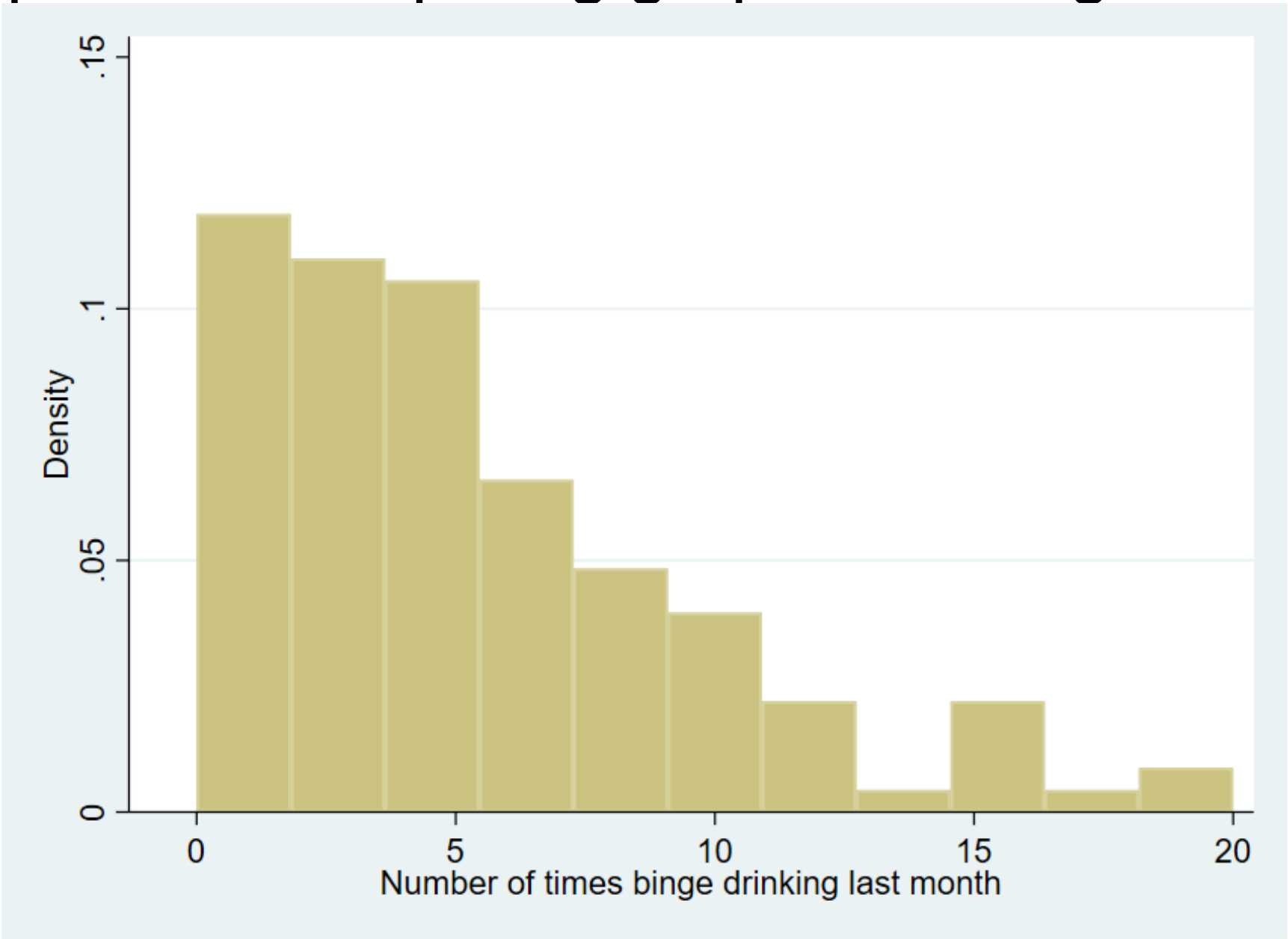
Boxplots: example



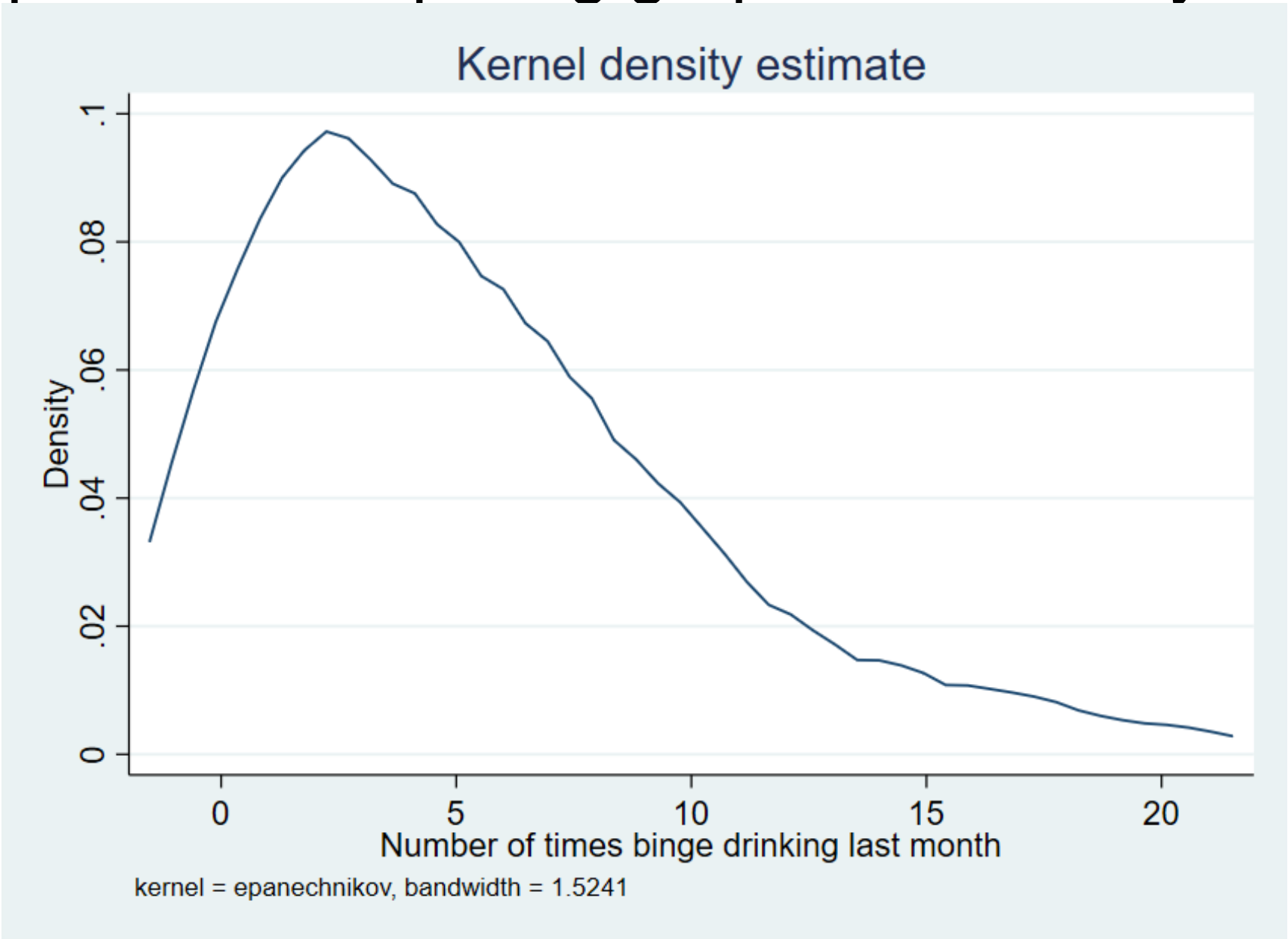
Graphing pitfalls

- Simpler is usually better (3D, etc., can be deceptive)
- Histograms: Be aware that the graph might look different depending on how bins are drawn (bin widths and starting points)
- Axes: Always pay attention to how the X and Y axes have been drawn; in general, any variation/change can be made to look very small or very big depending on how the axes are chosen; for example, see `https://peltiertech.com/tax-the-rich-or-deceptive-axis-scales/`

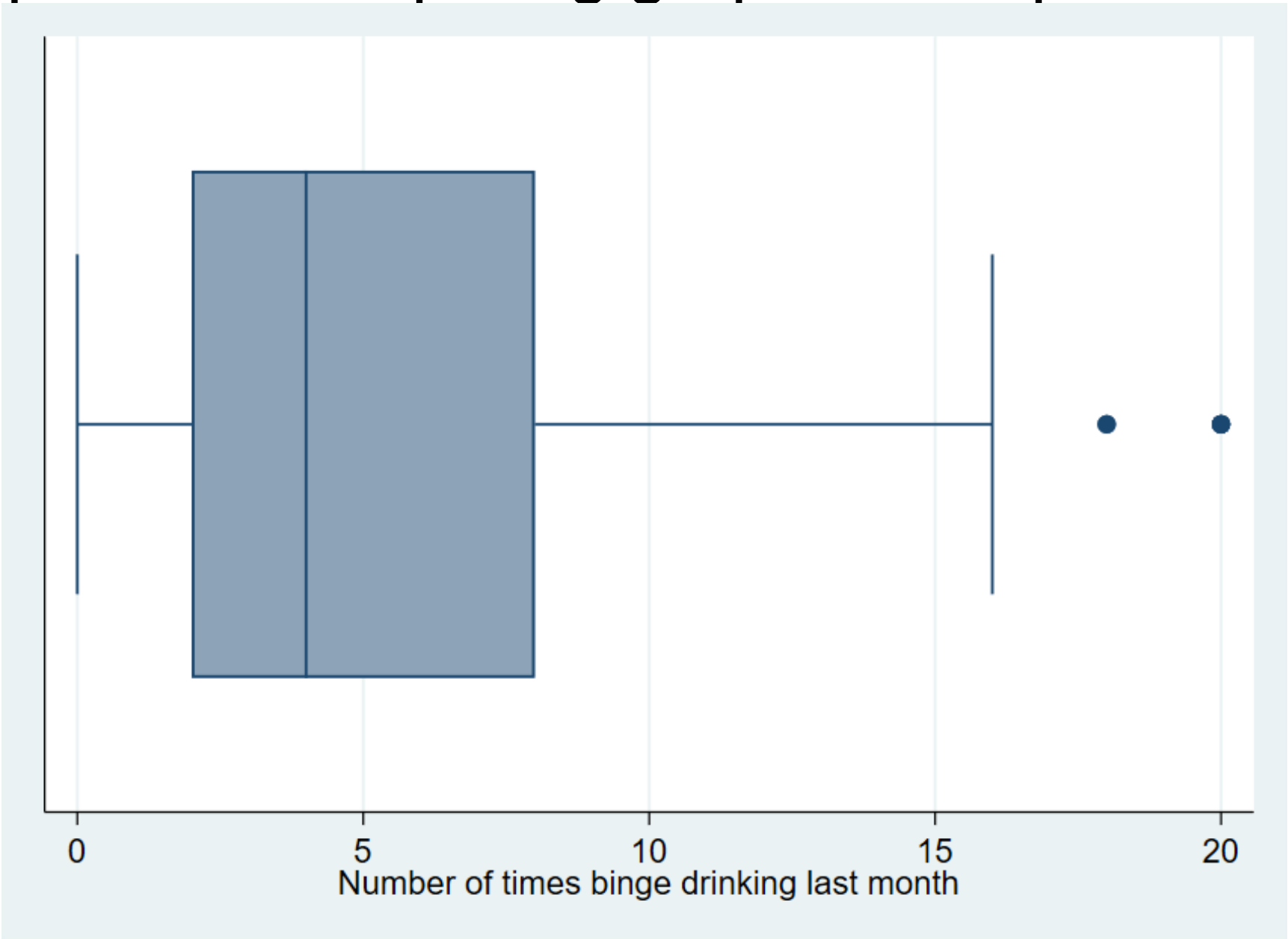
Appendix: Comparing graphs - Histogram



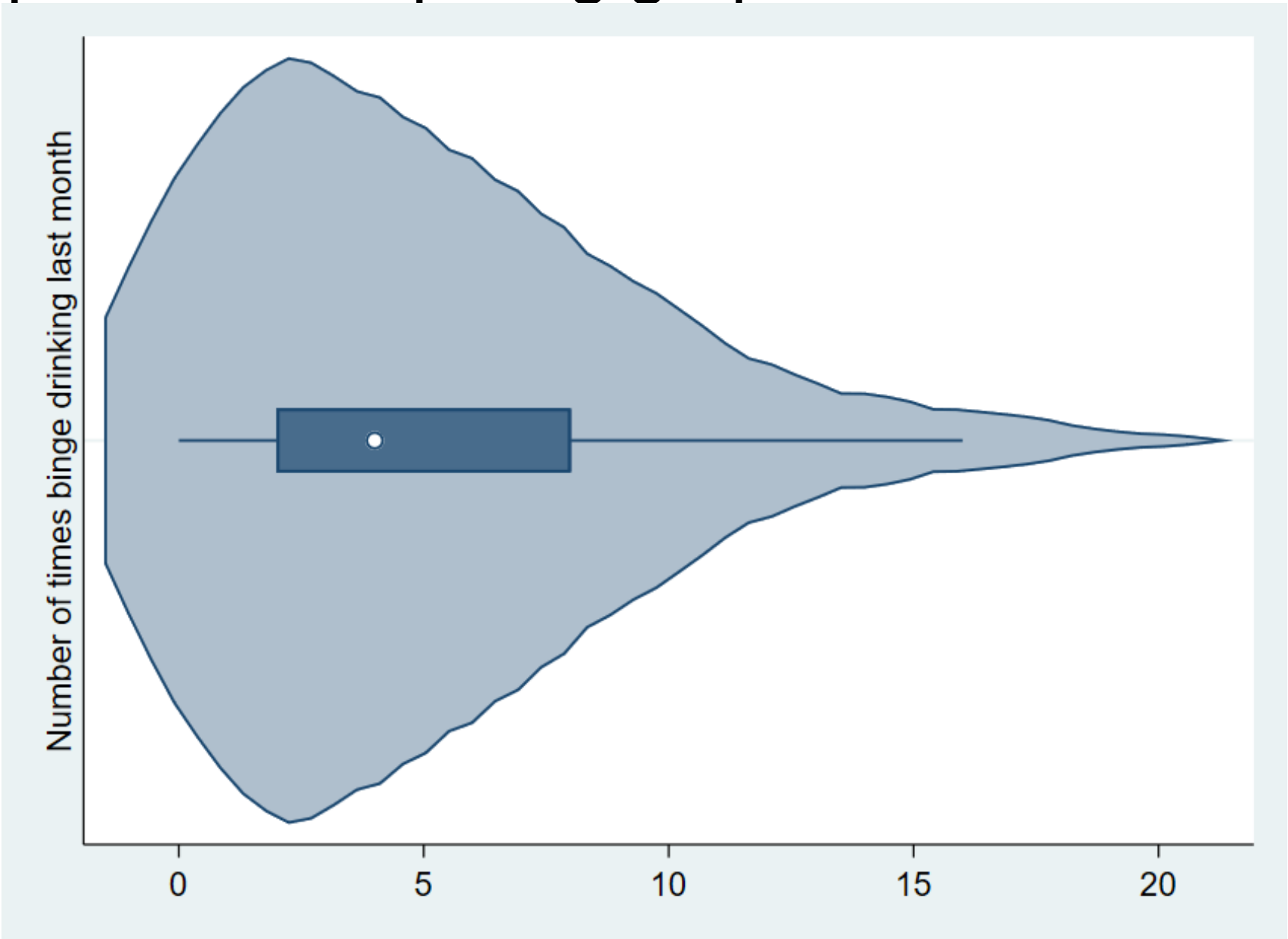
Appendix: Comparing graphs - K-density



Appendix: Comparing graphs - Boxplot



Appendix: Comparing graphs - Violin



Appendix: Comparing graphs - Strip plot