ECO 602 Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

Key concepts

1. Monotonic functions
2. Sigmoidal functions
3. Central tendency and spread
4. Statistics and parameters
5. Symmetry and skew

McGarigal Chapter 3

Very dense chapter, we’ll cover the material in 3 lectures:

1. sections 1 – 4
2. sections 4 – 6
3. Sections 7 - 9

Today's Agenda

1. Lightning talks
2. Data types
3. Sampling units and data
4. Data exploration overview
5. Data visualization: One variable

ECo Lightning Talks

If you attended, did you notice any concepts from this course?

ECo Lightning Talks

If you attended, did you notice any concepts from this course?

******Model Thinking******

Data Types

Integers, rational numbers, real numbers

Quantitative, qualitative

Continuous, discrete, directional

Categorical, nominal, ordinal, binary

Counts, ratios, time to death

Data types

1. What have you encountered in your real data sets and/or models?
2. What interesting data types have you worked with in the past?

Mycorrhizae

What are they?

Why should we care about them?

Sampling units and data

System: mycorrhizal fungi and microbial communities in soils in mid to high altitude conifer forests in the western US.

Question: Are higher elevations and proximity to roads associated with lower mycorrhizal species richness on ponderosa pines in the Deschutes National Forest?

Mycorrhizae Example

Question: Are higher elevations and proximity to roads associated with lower mycorrhizal species richness on ponderosa pines in the Deschutes National Forest?

1. What are sampling units?
2. What is the statistical population?
3. What is the ecological population

Mycorrhizae Example

Research question modified to:

Are higher elevations and proximity to roads associated with lower mycorrhizal species richness on ponderosa pines ?

1. What are possible sampling units?
2. What is the statistical population?
3. What is the ecological population

Warning

The path to understanding the following concepts is not linear: You need a baseline understanding of multiple concepts to understand the whole.

We'll introduce ideas, and circle back as needed to see how they fit together.

The learning process is iterative, much like model building.

Data exploration, screening, and adjustments

Data exploration: reality check before you begin formal analyses:

1. Are your data entered and coded correctly?
2. Do you have missing or extreme observations?
3. Do the data look the way you expected?
4. Are there unexpected patterns?

Data Exploration

Our brains are really really really good at detecting some kinds of features or patterns

We are better (for now) than computers at identifying objects in images.

We often interpret patterns on inanimate objects as faces.

Data Exploration

Our pattern-detecting skills are powerful, but subjective.

We need evidence from models and statistics to support, or refute, our visual intuition.

Data Exploration

Classical, frequentist statistical inference assumes an a priori model, i.e. you should have defined your hypotheses and models before collecting data.

Is this reasonable?

Data Snooping

Using exploratory data analyses to guide model selection, hypothesis building, suggest novel relationships, etc. is a fact of life but…

…models, hypotheses, relationships, patterns, etc. that we detect via post-collection data exploration require further data collection efforts.

Data Snooping

You can’t validate a model with the data that was used to create it.

- We’ll need to remember this concept later when we look at resampling methods.

Data snooping can increase the false positive, i.e. Type I error rate.

Statistics

Statistics are quantities that we calculate directly from measurements of our samples.

Uncertainty is minimized*, but…

Statistics: sources of uncertainty

1. Accuracy/precision of measurements
2. Miscounting, data transcription errors
3. Some statistics are computationally intensive, and we must settle for approximations
4. Other issues?
5. D. Rumsfeld: known and unknown unknowns

Some common statistics

Measures of central tendency: what is a ‘typical’ value for the data?

- Mean, median, model
- Expected value
- Deterministic component of our data model
- What is our best guess for the value of a random sample from the population?

Some common statistics

Measures of spread: how variable is the sample data?

- Min, max, range
- Variance, standard deviation

Symmetry: are the data distributed similarly above and below the mean?

Normality: Could the sample be from a Normally-distributed population?

Parameters and statistics

We use the **statistics** we calculate from samples to **infer parameters** of the overall statistical population.

Exploratory data analysis consists of examining **statistics** calculated from our existing data.

Probability Density Functions: PDF

X-axis is the value of the predictor

Y-axis is the probability density

- (approximately) a measure of likeliness of observing the corresponding x

The total area under the PDF is 1.0, PDF is always > 0

Probability Density Functions: PDF

Called Probability Mass Functions for discrete distributions.

PDF/PMF are NOT monotonic.

* Don’t worry too much about this now, we’ll revisit when we cover probability distributions.

Cumulative Probability Functions: CDF

X-axis: predictor value

Y-axis: sum of probability density (or mass) for all values less than the corresponding x.

Monotonic increasing, approaches infinity as x grows

Integral of the PDF

* Don’t worry too much about this now, we’ll revisit when we cover probability distributions.

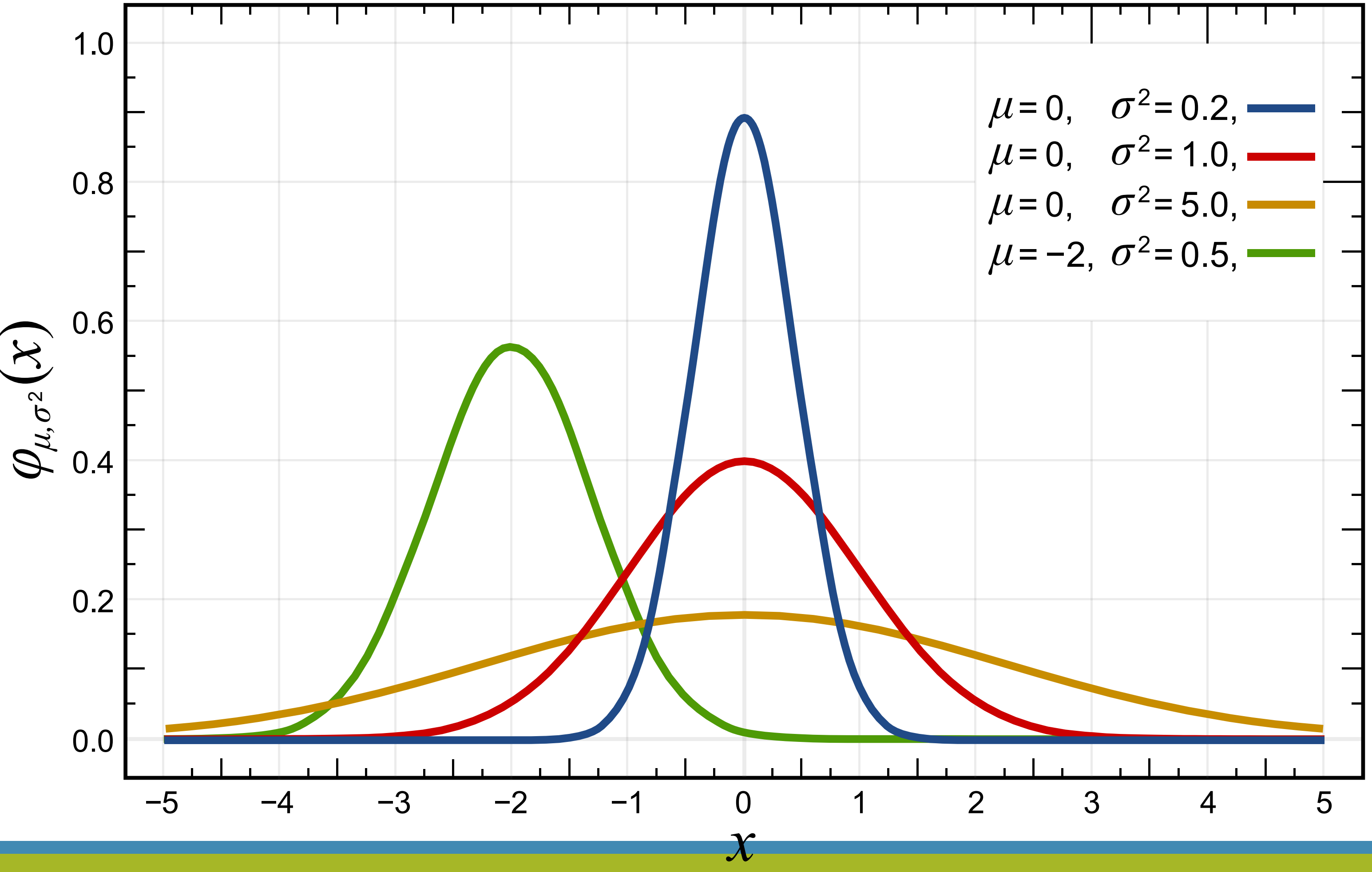
The Normal distribution

- Familiar bell-shaped, unimodal curve.
- Very common in analyses
  - Mathematically tractable, arises in theory, Central Limit Theorem, etc.

The Normal distribution

  - Can often approximate other distributions
- Two parameters
  - Mean: Greek mu - μ
  - Standard deviation: Greek sigma - σ

Normal Distribution: probability density function



Normal distribution: important properties

Symmetrical

Domain is = –infinity to infinity

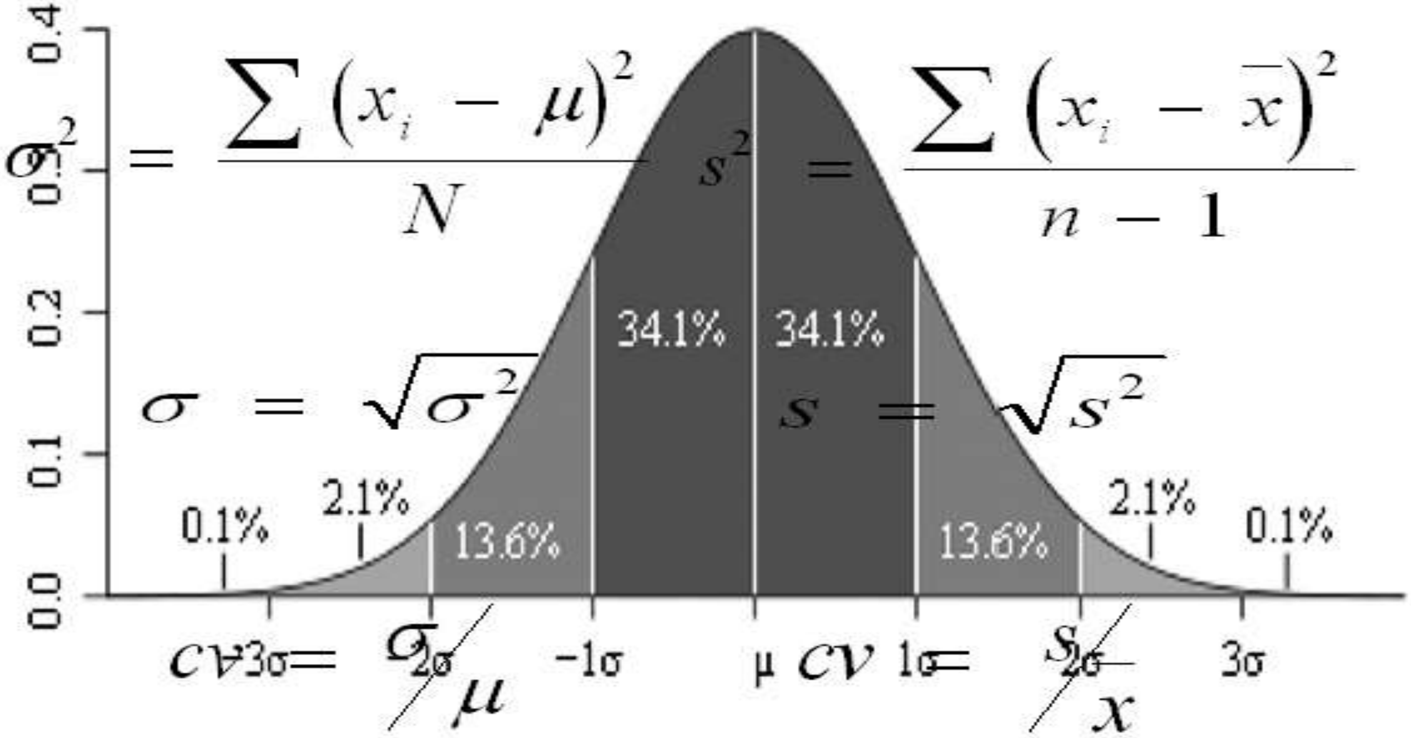
Continuous

Standard deviation has nice, intuitive interpretation

Normal distribution: standard deviation interpretation

~ 68% of density is within +/- 1 SD

~ 95% of density is within +/- 2SD



$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

$$cv = \sigma / \mu \qquad cv = s / \bar{x}$$

Normal distribution: standard deviation interpretation

In a sample from Normally-distributed population we would expect:

1. Ca. 68% of samples to be within +/- 1 SD
2. Ca. 95% of samples to be within +/- 2SD

These facts make inference with the Normal distribution more intuitive than with many other distributions.

Some single-variable plots

Empirical Distribution Plot

Histogram

Boxplot

Quantile-quantile plots

Empirical (Cumulative) Distribution Plots

A sample analogue of probability density and cumulative density functions.

These can be conceptually difficult. We’ll return to them later.

They’ll make more sense after we talk about probability distributions.

Histograms

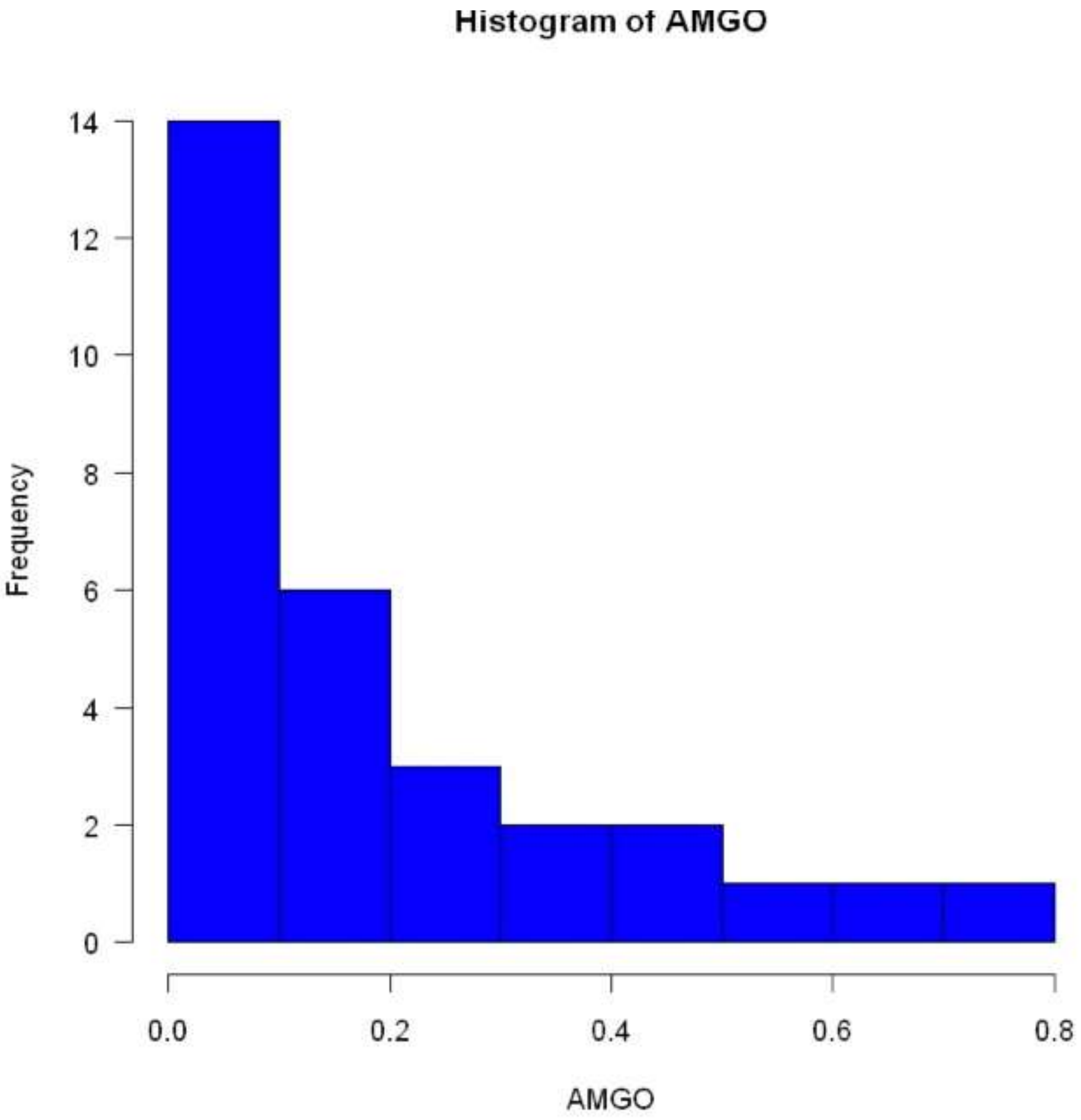
Frequency of observations in bins

Similar to Probability Density (or Mass) Functions.

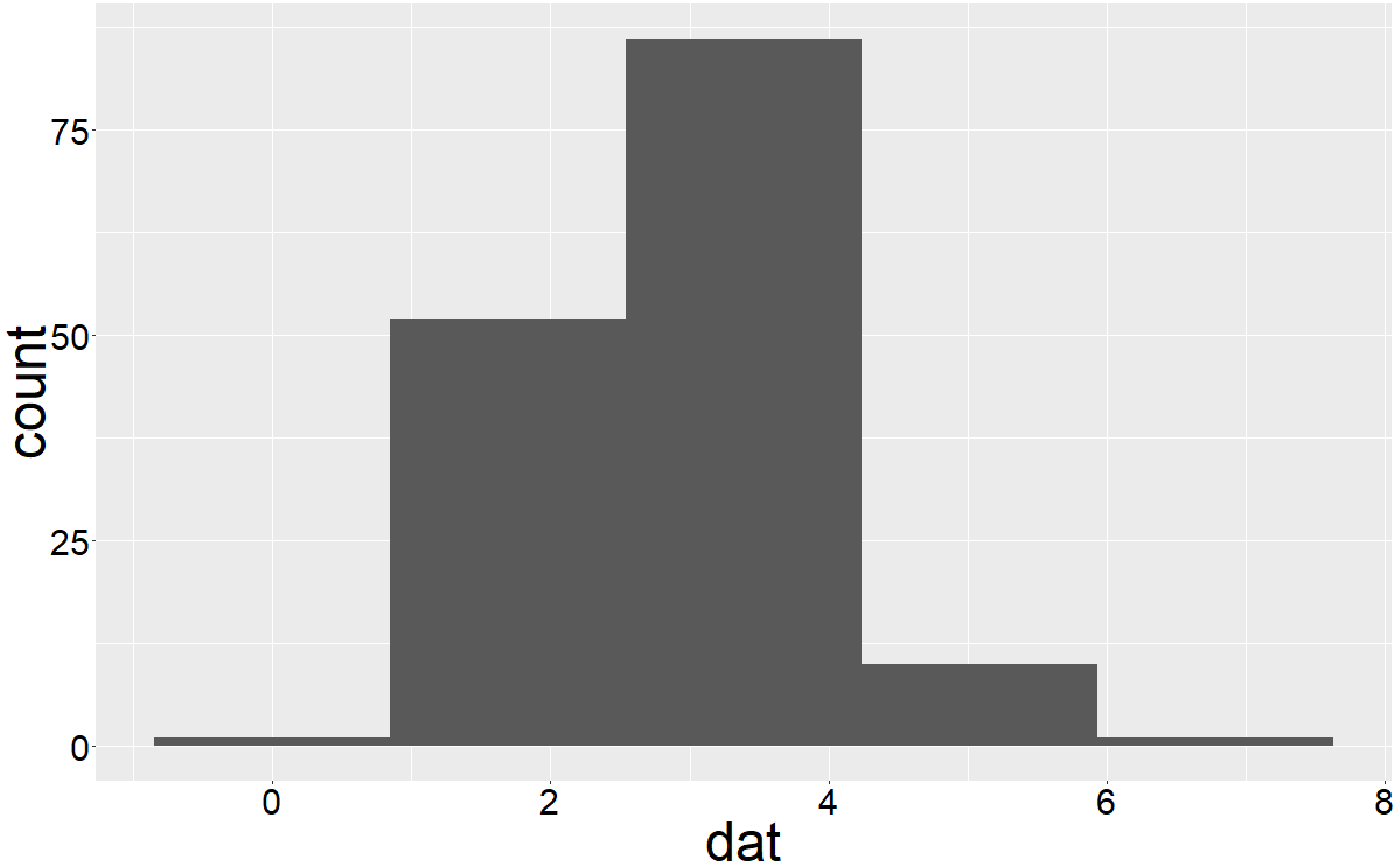
Useful to assess symmetry, central tendency.

Be aware of bin sizes

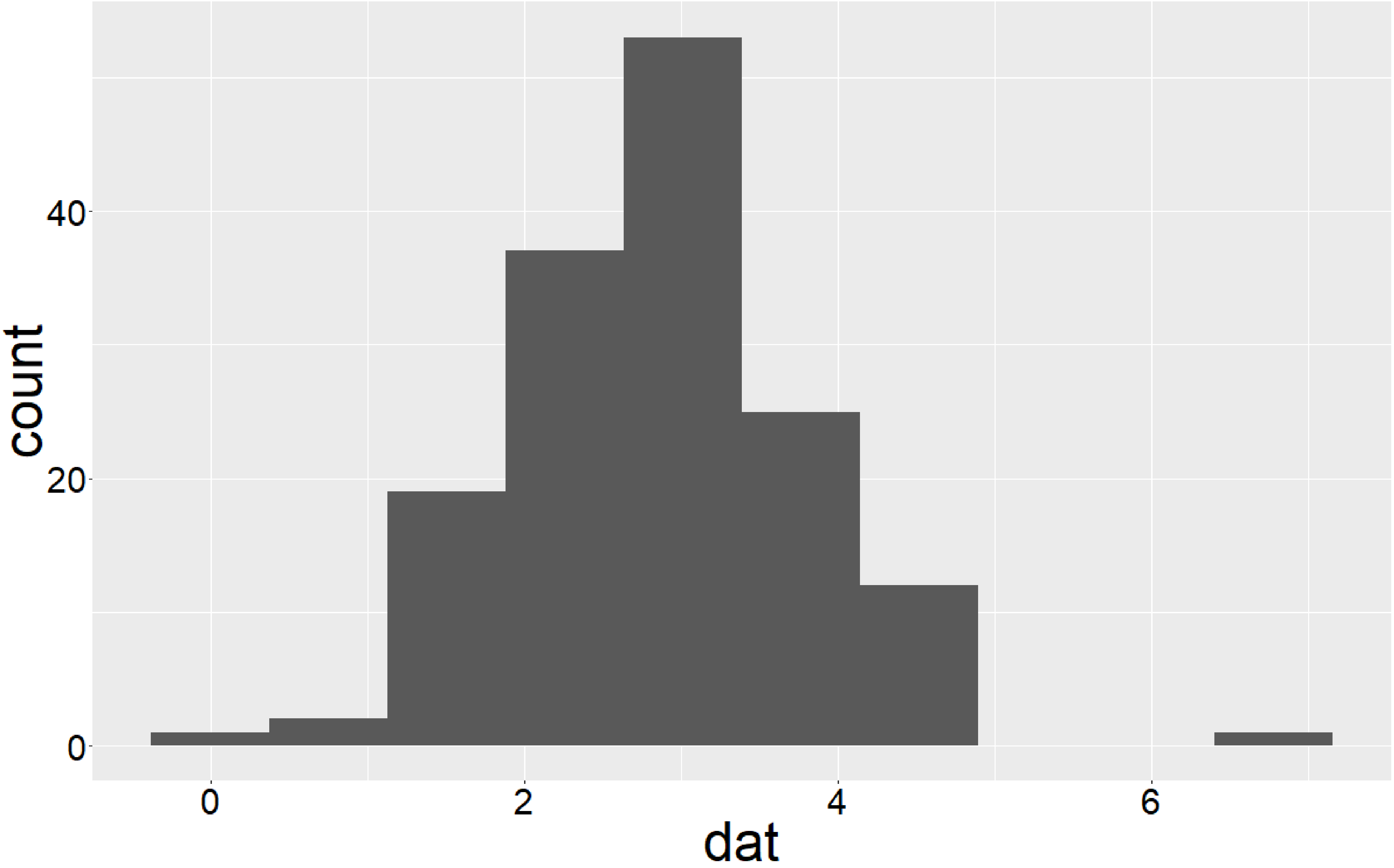
- Graphical display of tabulated frequencies (or probabilities), shown as bars
- Shows what proportion of cases fall into each of several adjacent non-overlapping categories
- A way of binning the data



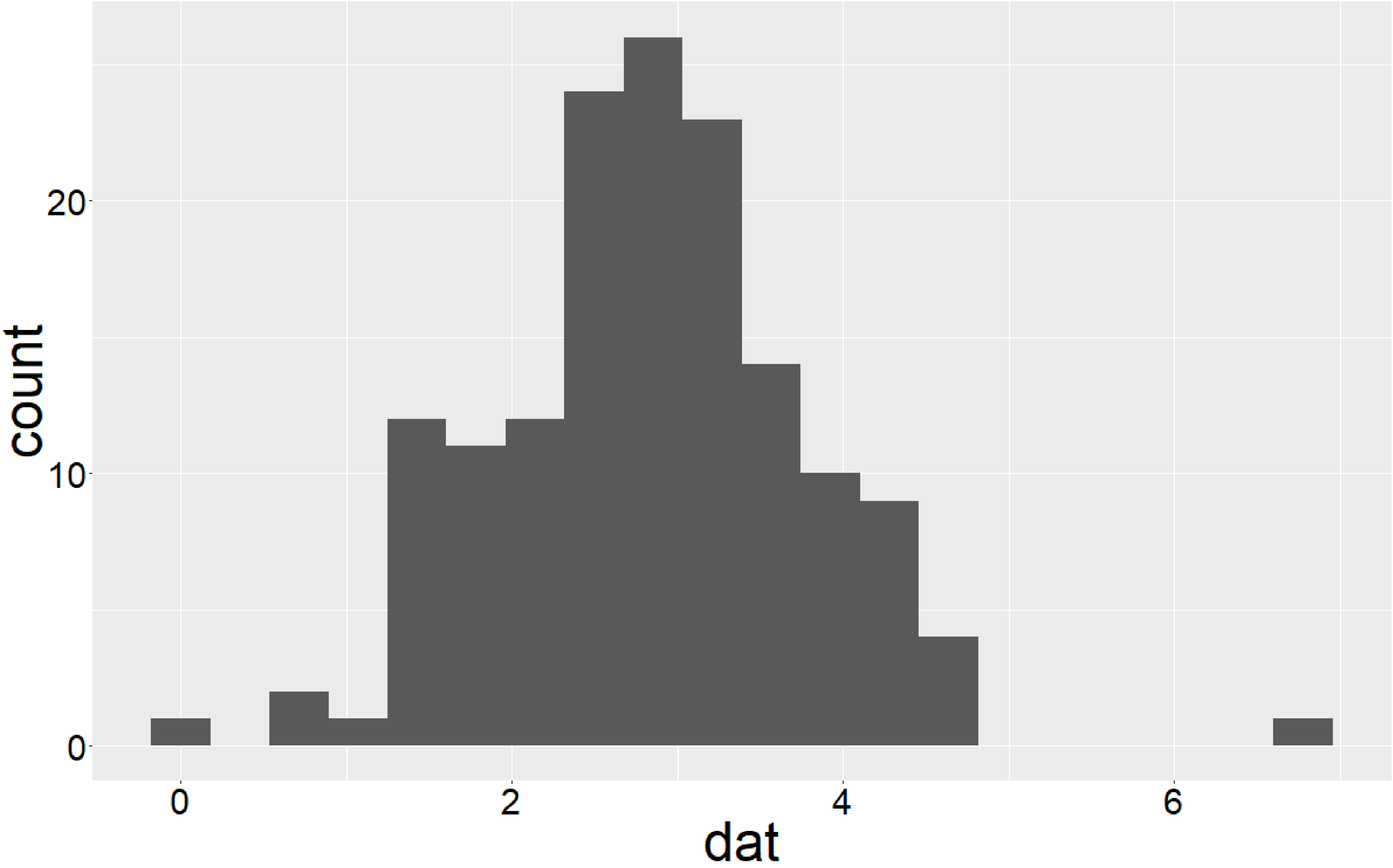
Normal 5 bins



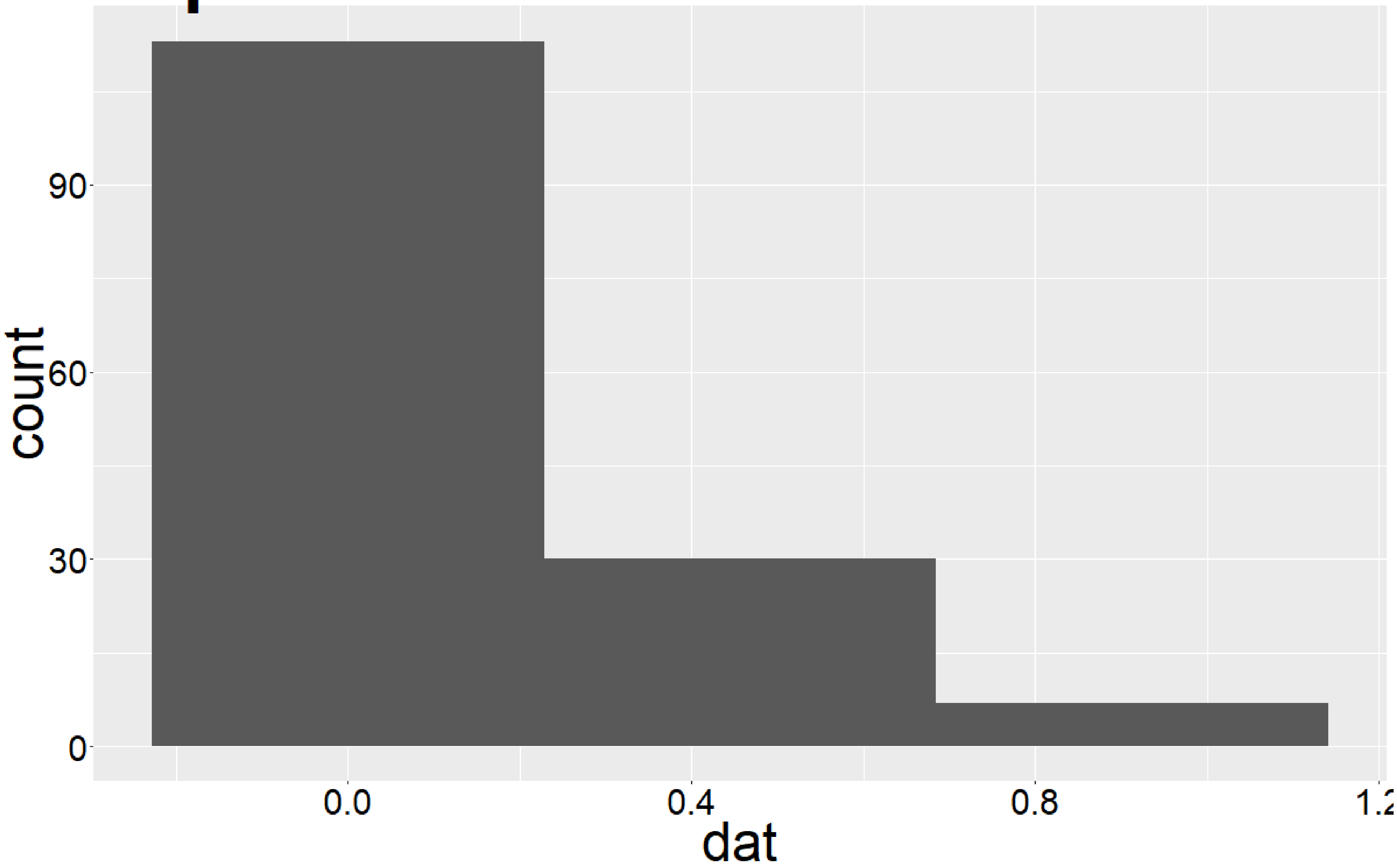
Normal 10 bins



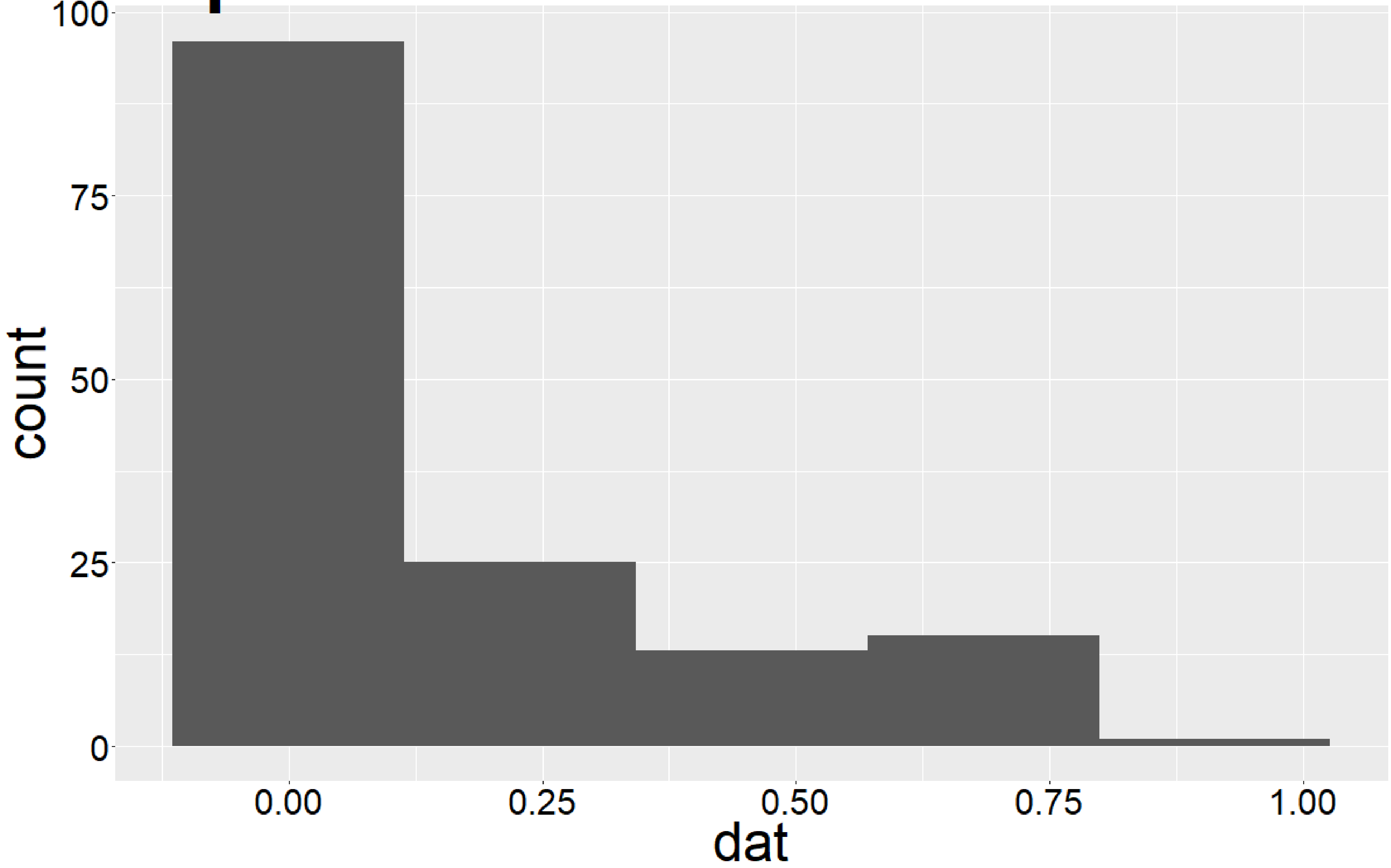
Normal 20 bins

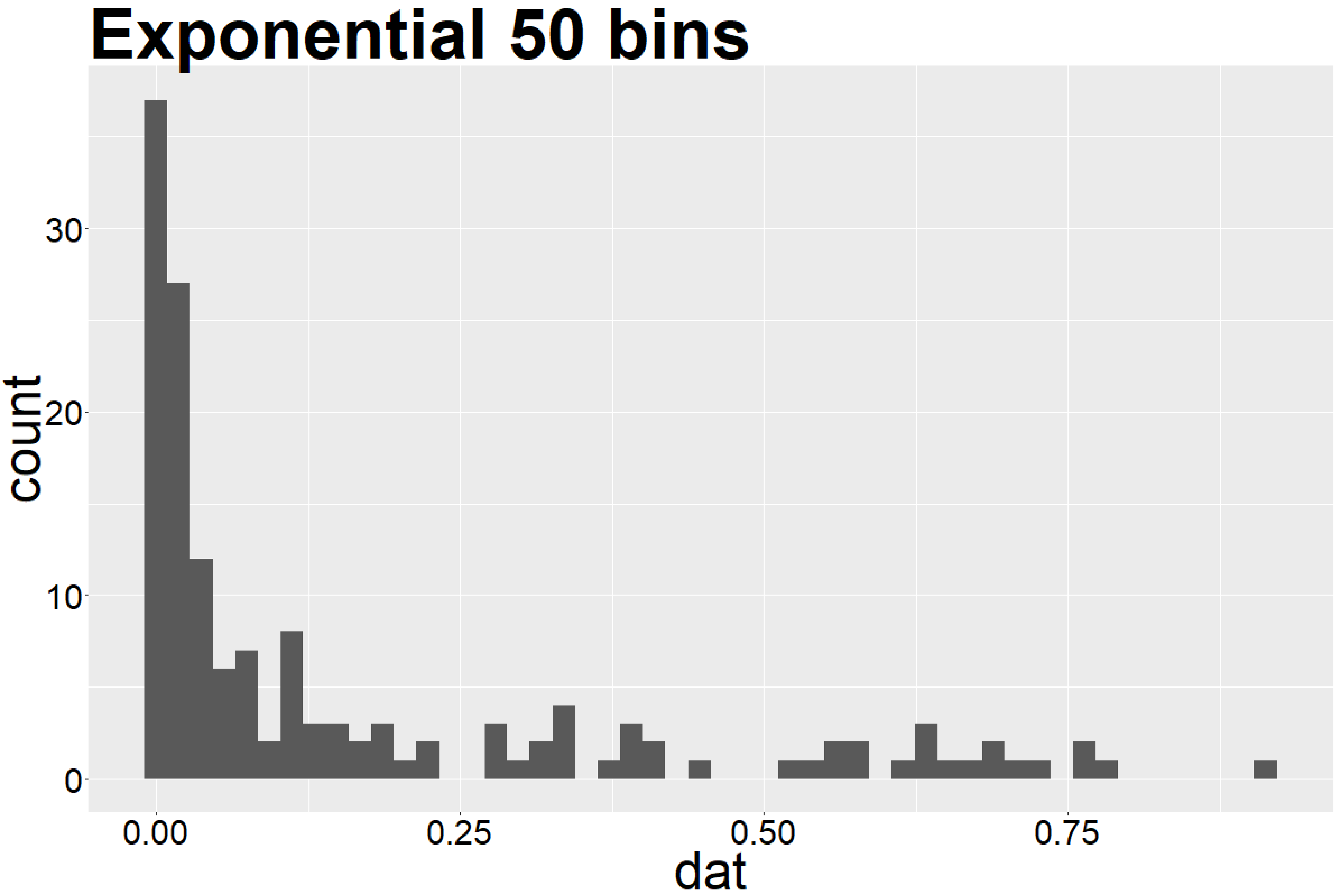


Exponential 3 bins



Exponential 5 bins

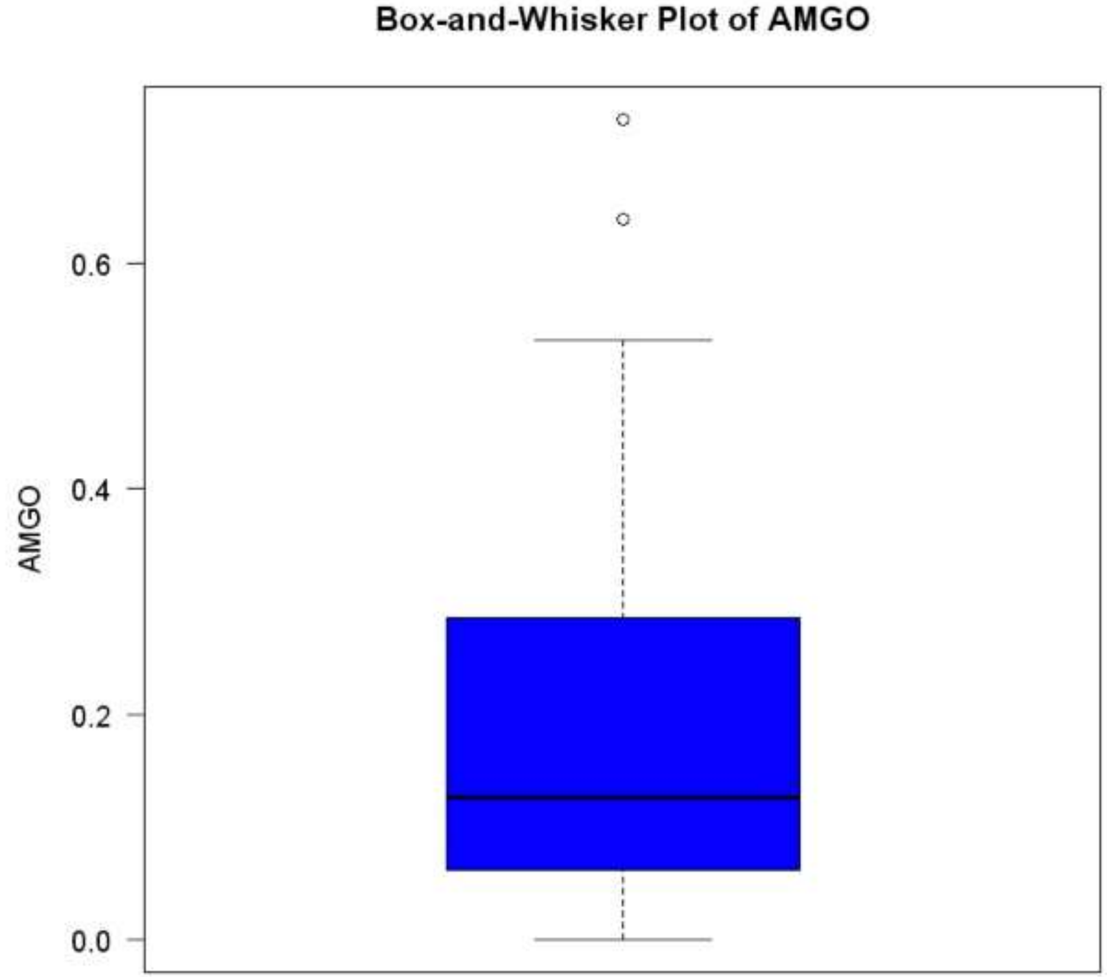




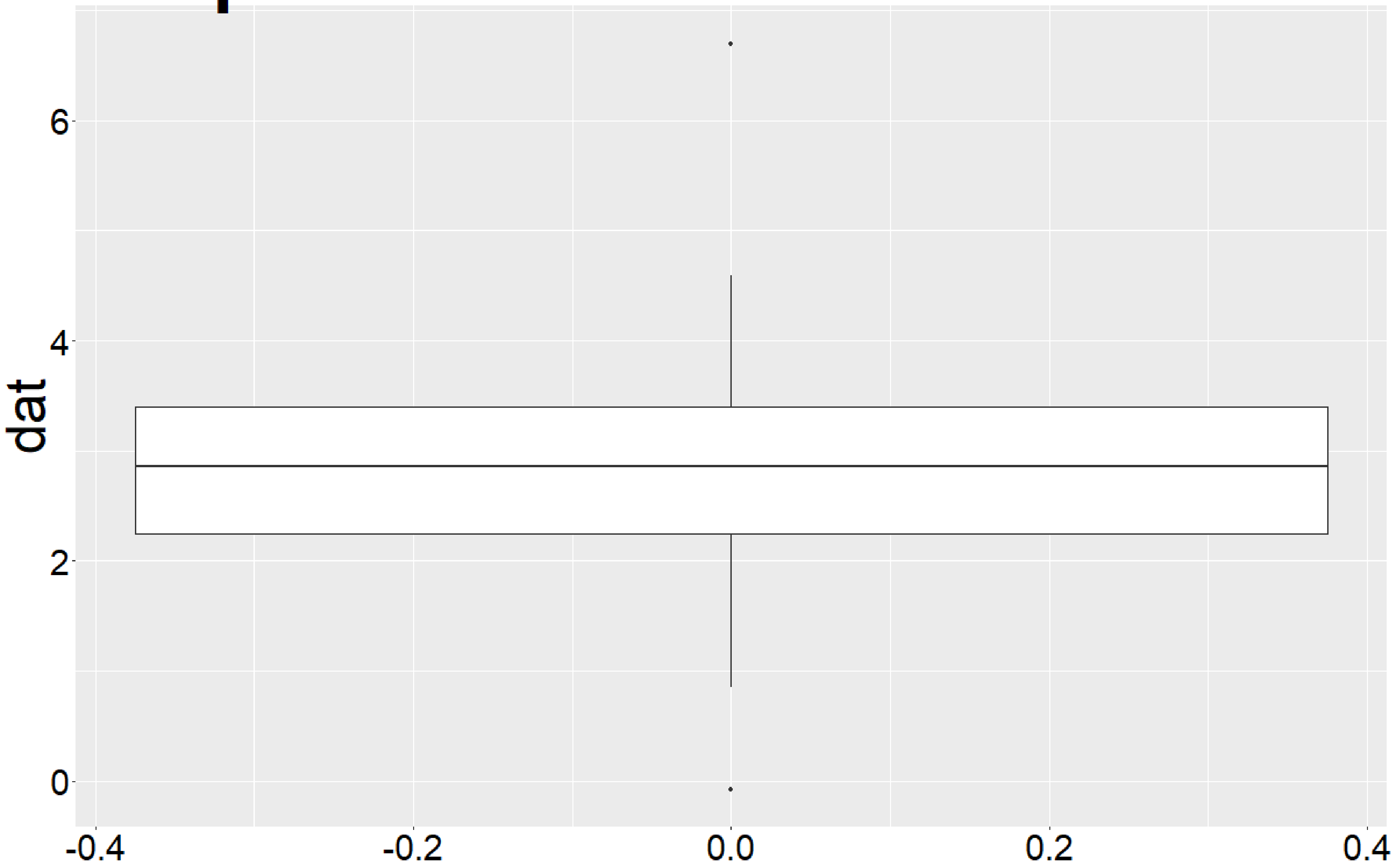
Exponential 50 bins

Box and Whisker Plot

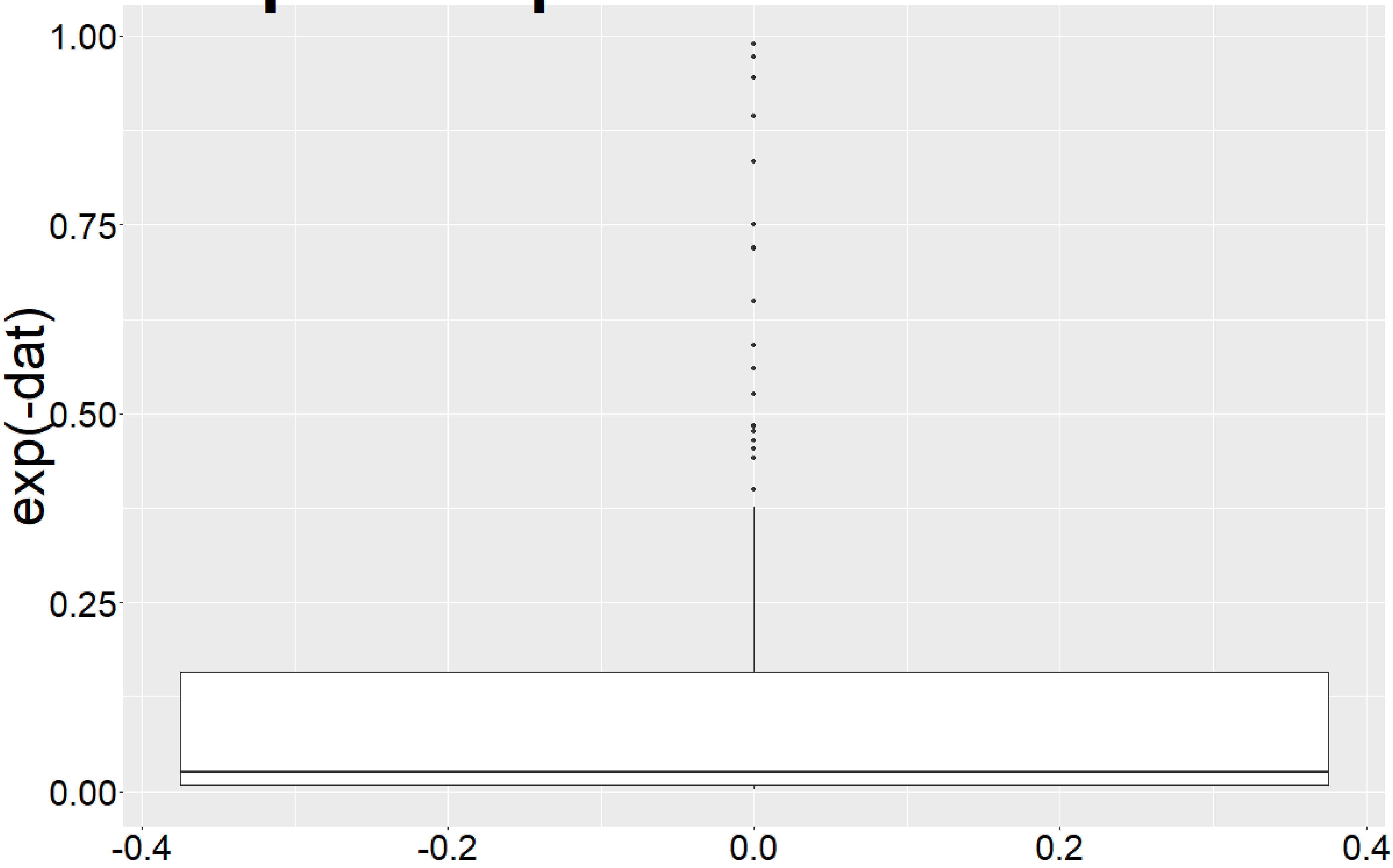
- Graphical display of the spread of a variable
  - *Solid line* depicts the median (50th quantile)
  - *Box* depicts the inter-quartile (25-75th quantiles) range
  - *Whiskers* depict the range of the data up to 1.5xIQR
  - *Isolated points* depict "extreme" values



Boxplot: Normal distribution



Boxplot: Exponential distribution



Normal quantile-quantile plots

Concept: a comparison of real sample data to predicted values from a Normal PDF

X-axis expected quantile value

Y-axis observed quantile value

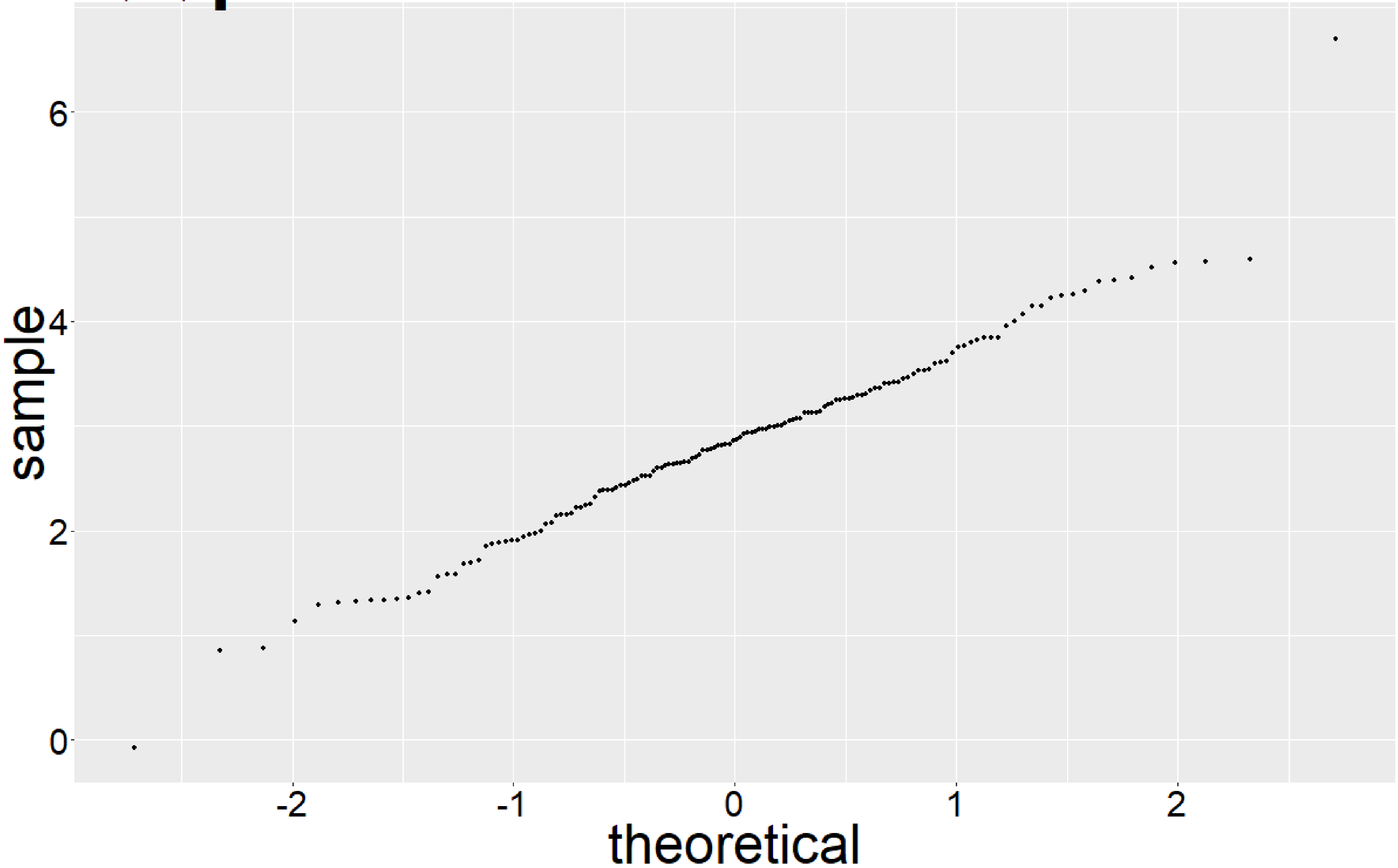
Calculation: can be tricky, endpoints need special treatment, we won’t cover the details.

Normal quantile-quantile plots

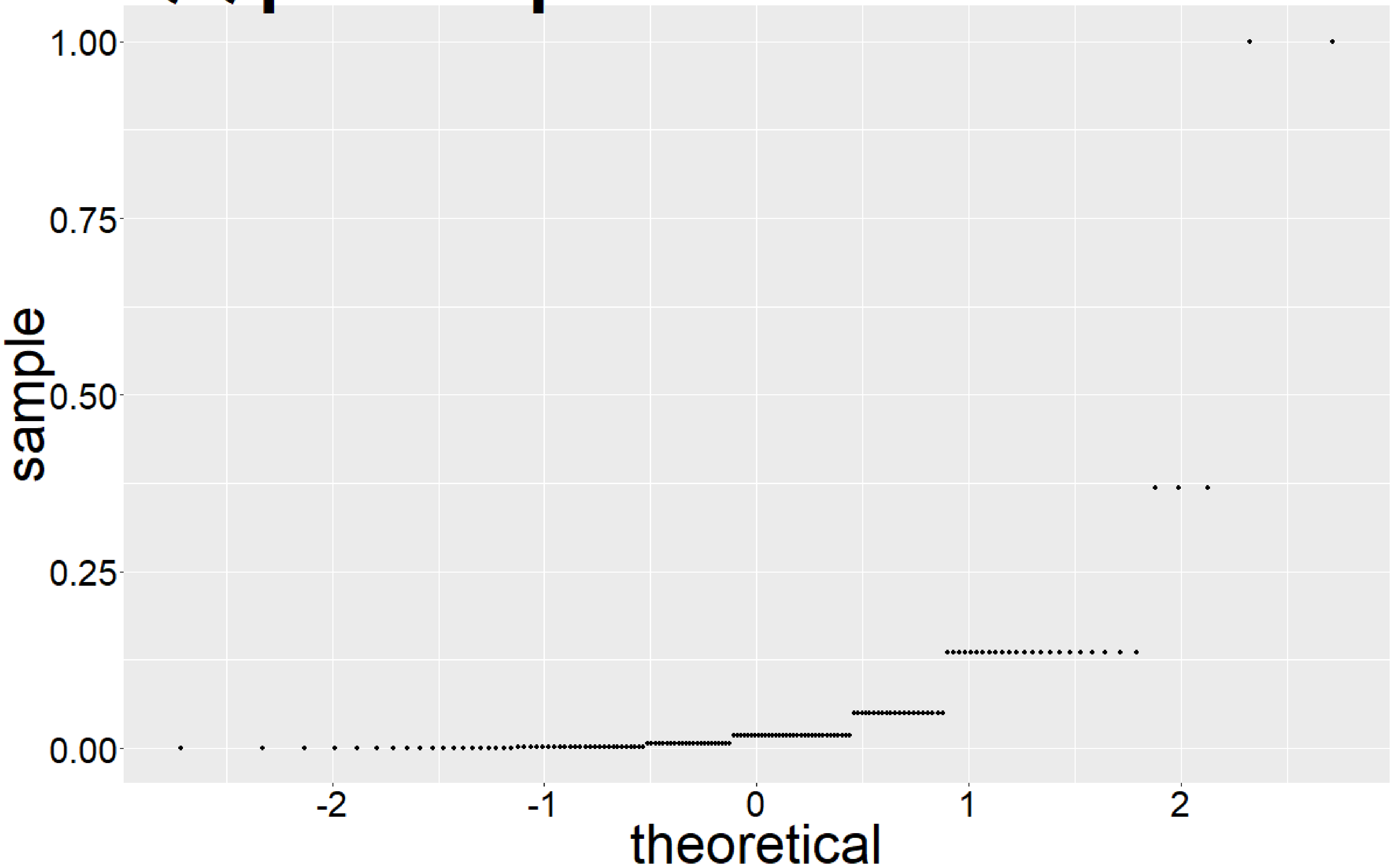
Interpretation: samples from a Normally-distributed population should be close to a diagonal line

Calculation: can be tricky, endpoints need special treatment, we won’t cover the details. Usually software will do it for you.

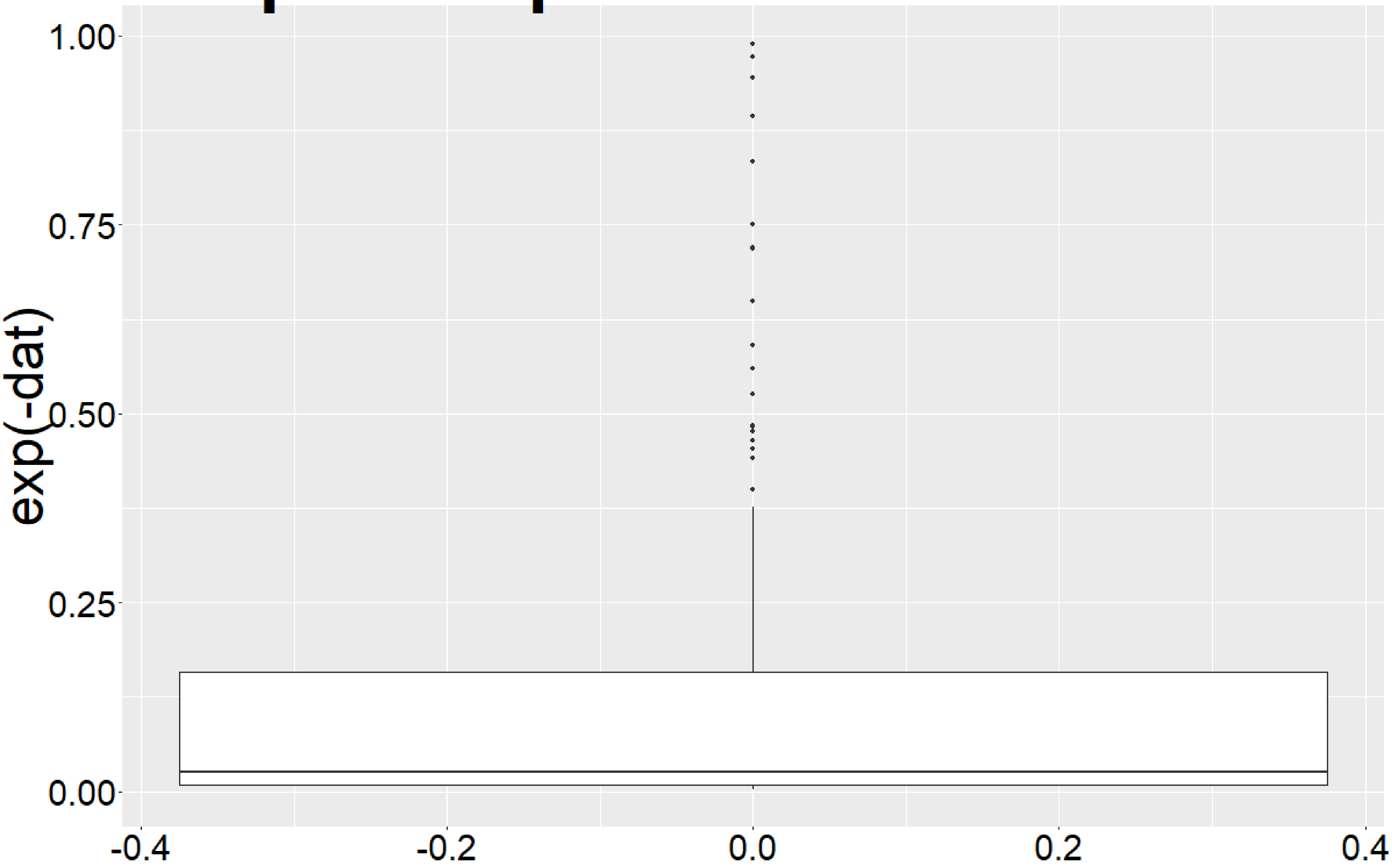
QQ plot: Normal distribution



QQ plot: Exponential distribution



Boxplot: Exponential distribution



Single variable plot overview

Histogram: skewness and symmetry, center, spread, good indicator of normality

Boxplot: 5 number summary, skewness, center, spread, extreme values, not a great depiction of normality

Empirical probability plots: assessing normality

QQ plots: assessing normality, useful for model diagnostics (more on this later)

For next lecture: association

More than 1 variable: >1 measured attributes per sample unit

Do sample units with high values of variable 1 also tend to have high values for variable 2?

In a sampling unit, does the value of variable 1 influence the expected value of variable 2?

For next lecture: key take-homes

Histograms: importance of bin sizes

Boxplots: extreme values

Symmetrical and skewed distributions

2 parameters of Normal distribution: mean, SD

Monotonic and sigmoidal functions