Exploratory Data Analysis

Stat 250

Click here for PDF version

Warm up questions

Take a little time to discuss the warm-up questions with your neighbors.

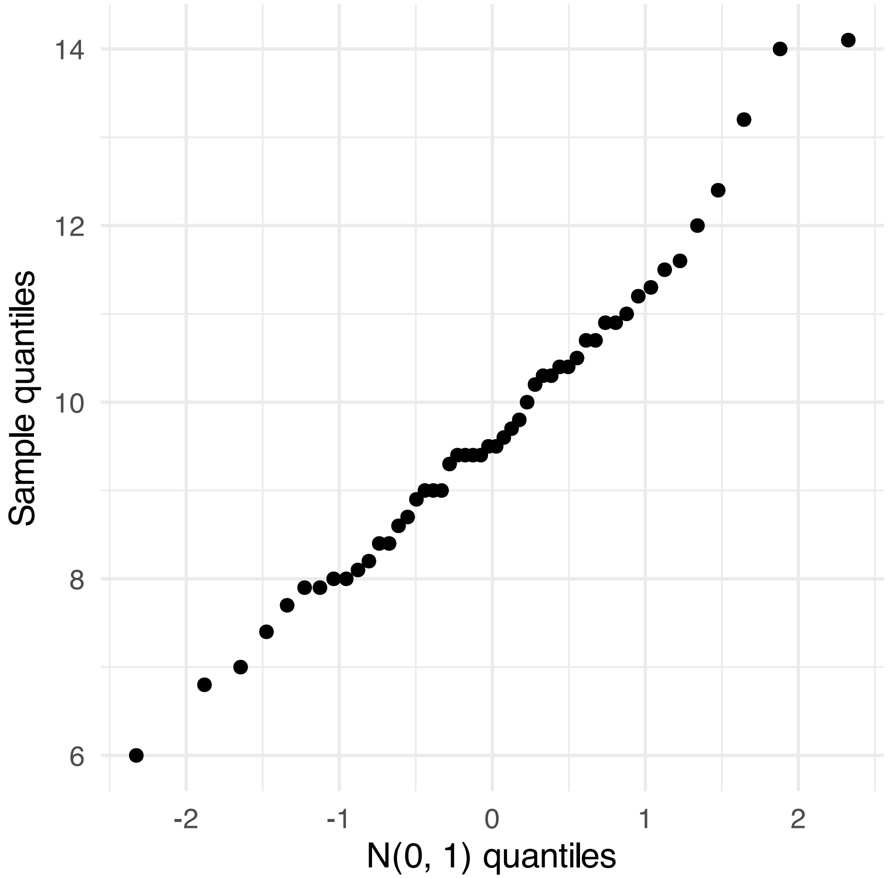
Example

- An ecologist draws a random sample of 50 organisms and measures the length of a certain feature
- The ecologist is interest to see whether these measures could follow a normal distribution

Q-Q plot

- Compare two sets of quantiles to see if the distributions could be the same
    - sample vs. sample
    - sample vs. theoretical (← our focus)
- If the two distributions are the same, then the quantiles should roughly agree/align

Normal Q-Q plot

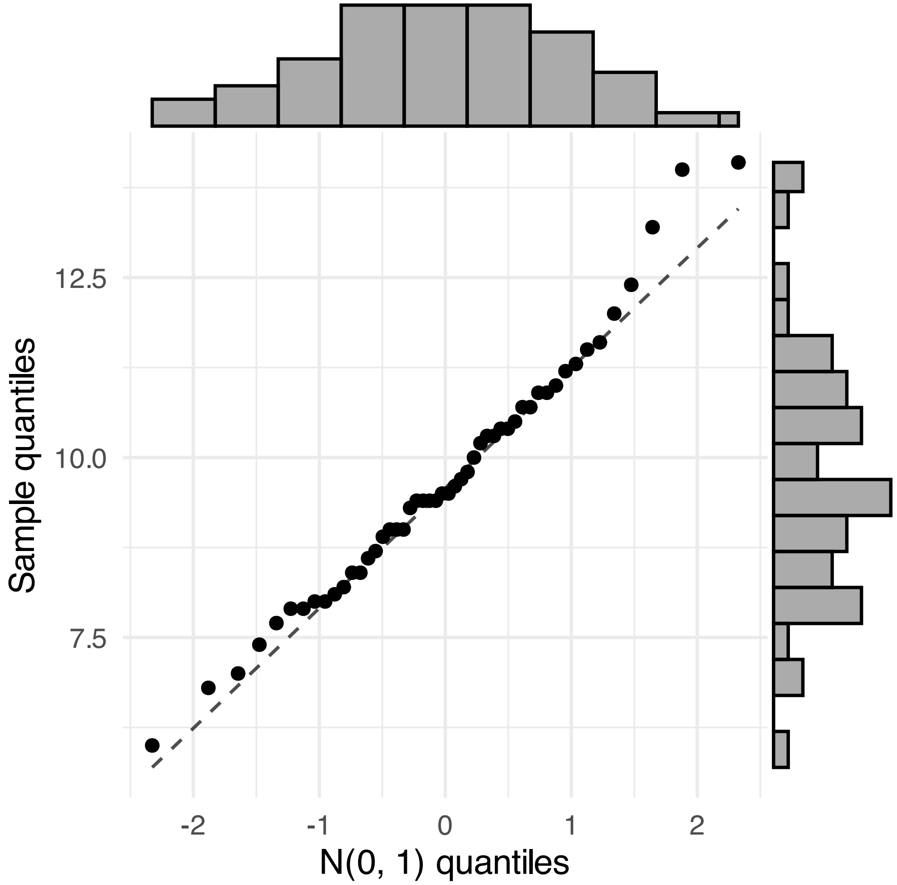


1. Sort observed values
2. Decide on what quantiles are in the data set (R does this for us)

   $$0.01, 0.03, 0.05, \ldots, 0.97, 0.99$$

3. Calculate quantiles from $N(0,1)$
4. Plot ordered pairs: (x_i, q_{p_i})

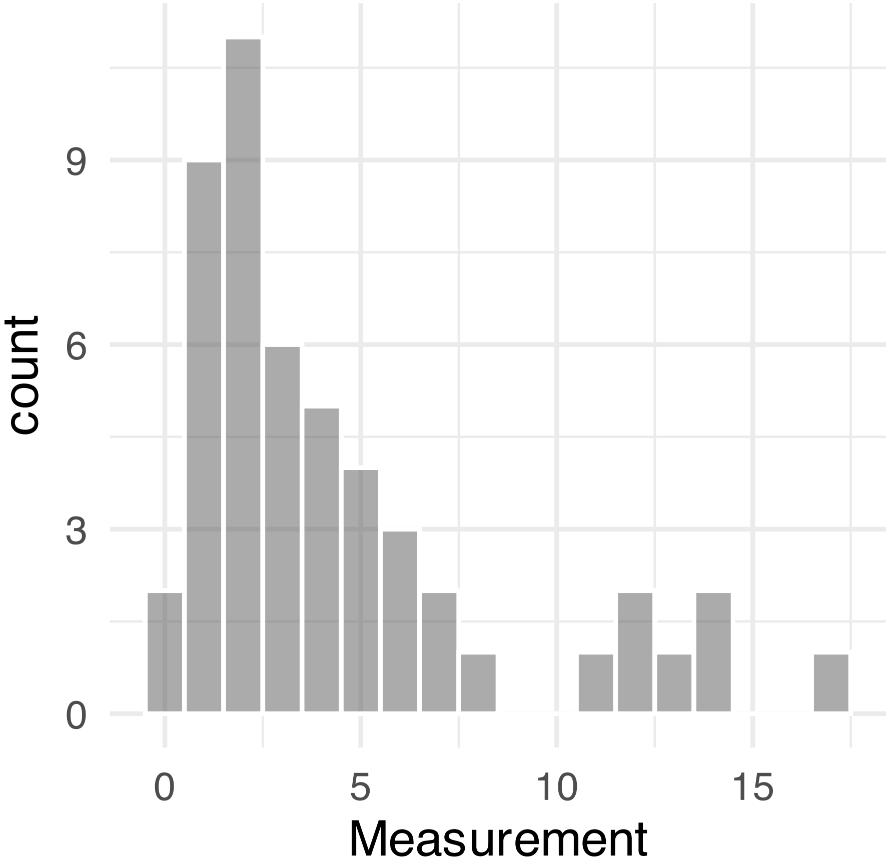
Normal Q-Q plot

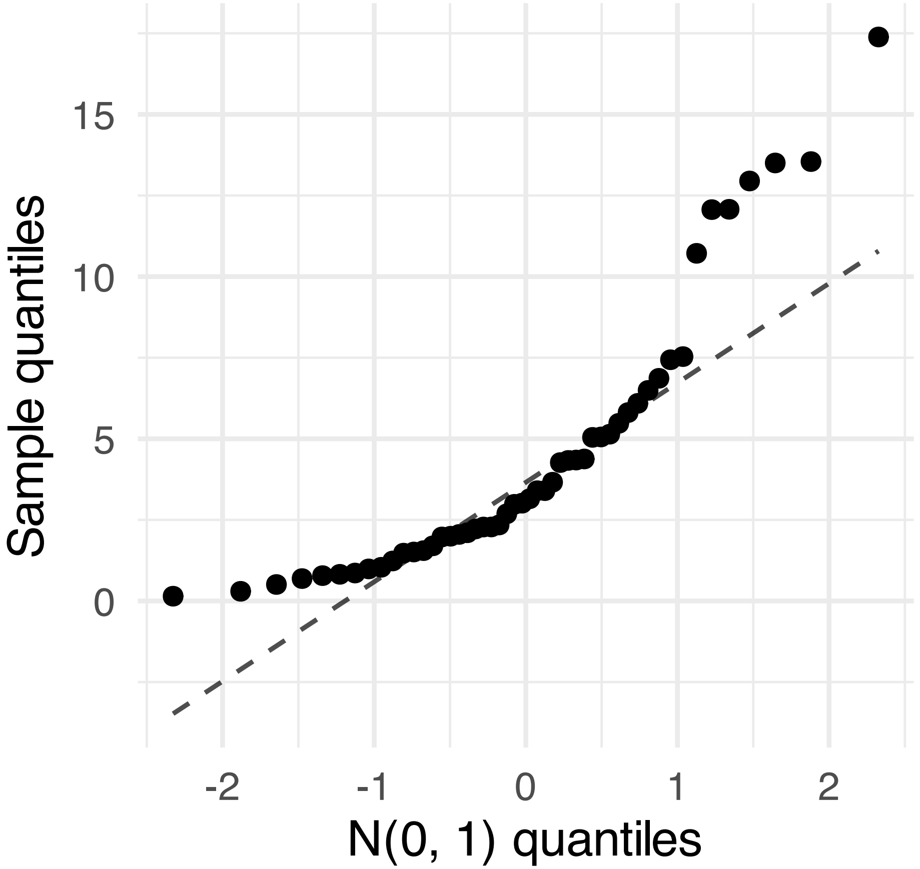


- Comparing shapes of the distributions
- Perfect agreement = line
- Deviations from the shape appear as vertical departure from the line
- Minor deviations are not troubling (sampling variability)

Your turn 1

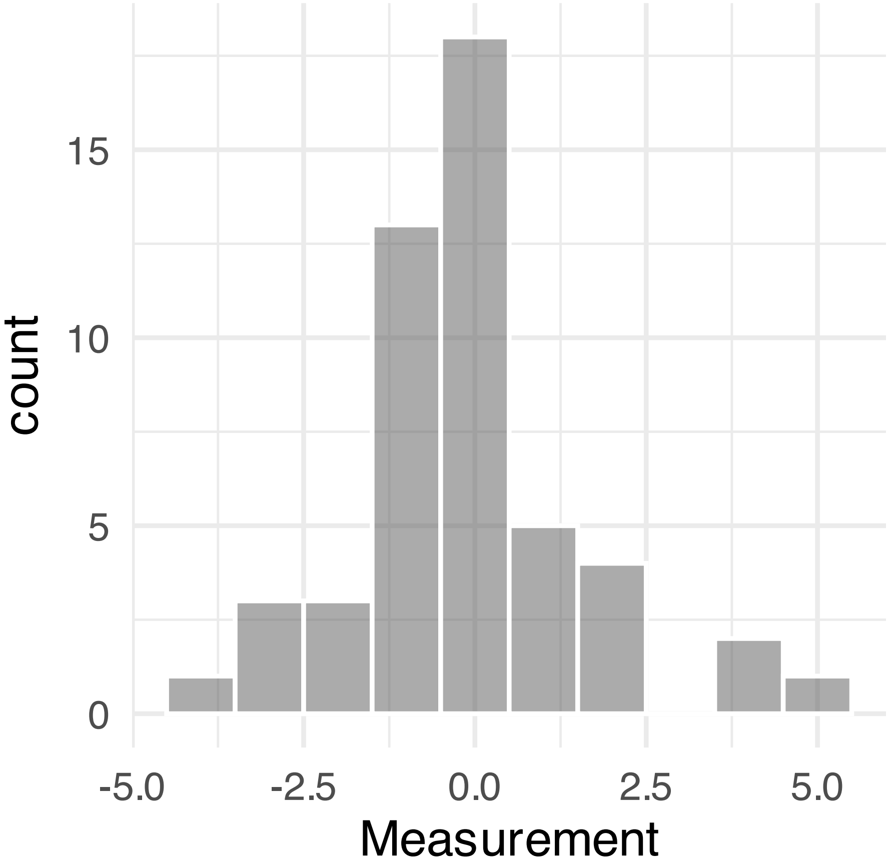
Do you think the sample could arise from a normal distribution?

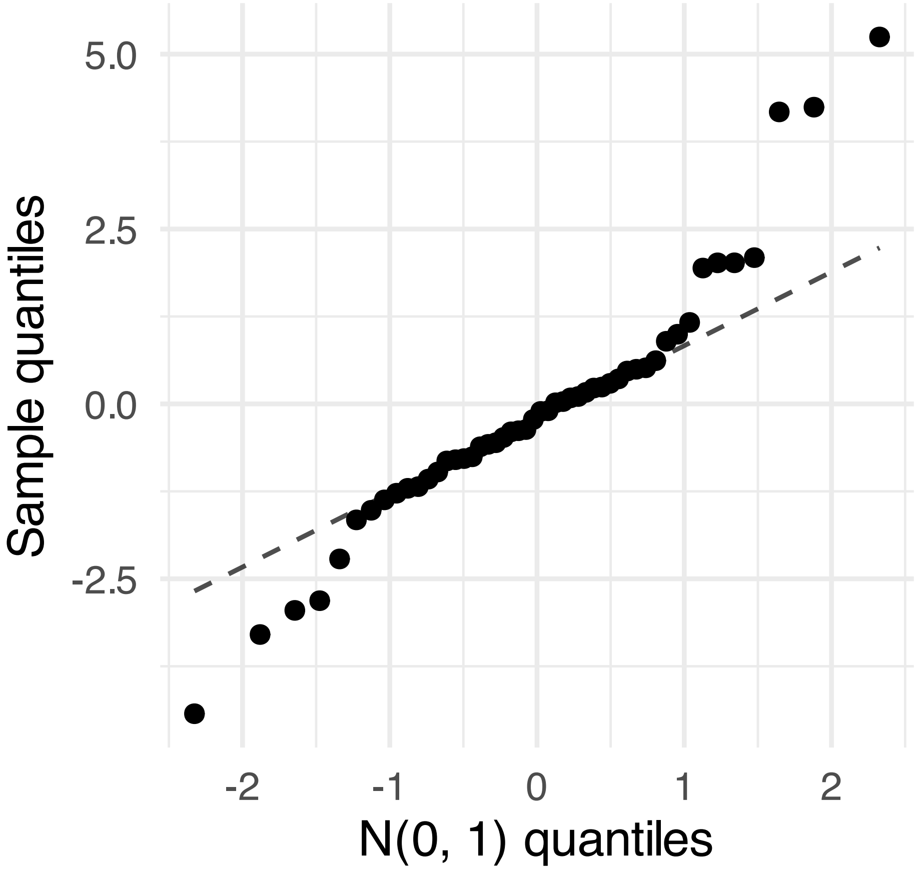




Your turn 2

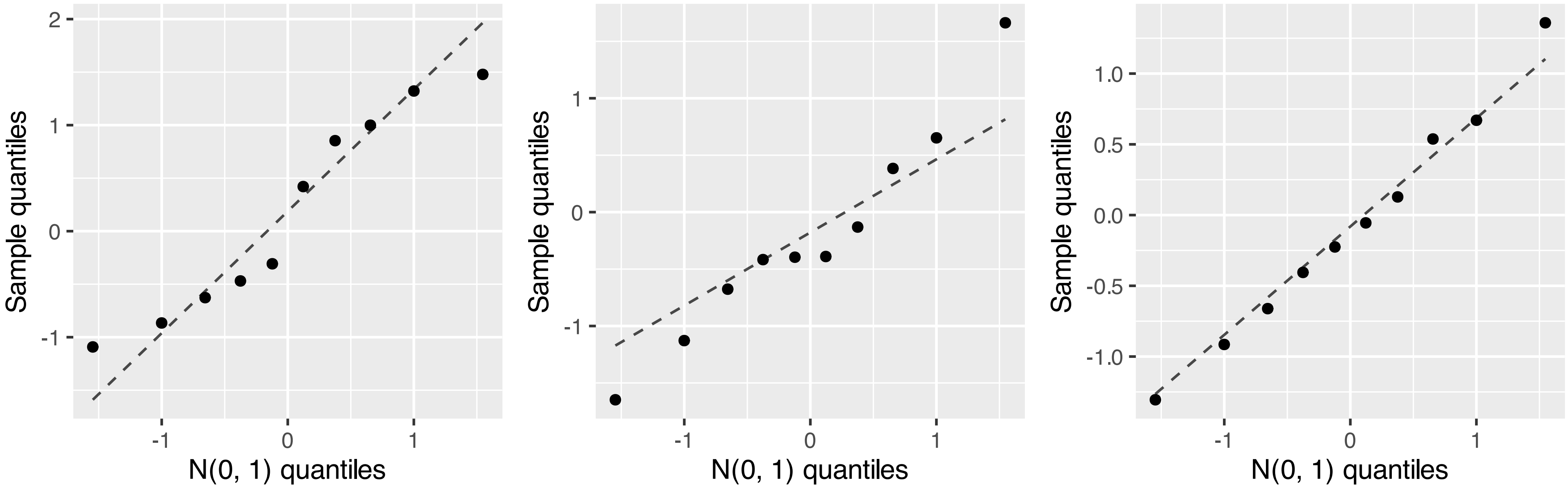
Do you think the sample could arise from a normal distribution?





Small sample sizes are tricky

Here we have three samples of $n = 10$



Intro to R

Function application syntax

```
function_name(arg1, arg2, arg3)
```

Creating objects

```
my_object <- function_name(arg1, arg2, arg3)
```

Things to remember

- R is case sensitive
- R only does what you ask
- Always close parentheses
- Separate arguments with commas

Your turn

- Work through the EDA in R tutorial with your neighbors
- R has a learning curve, stick with it and ask questions when you’re confused!

Tips for EDA

1. Experiment with histogram binwidth or number of bins
2. Bar charts start at zero
3. Use histograms, not pie charts
4. Label your axes, including your units
5. Give context in a title or caption
6. Be able to describe every graph that you use