Fundamentals of Data Visualization

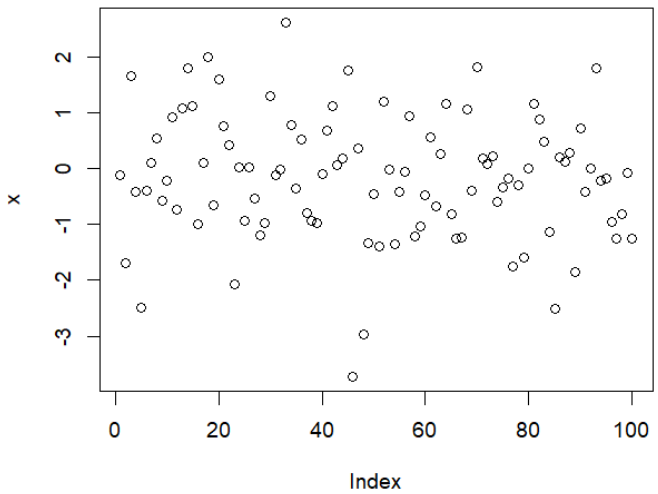
Chapter 8

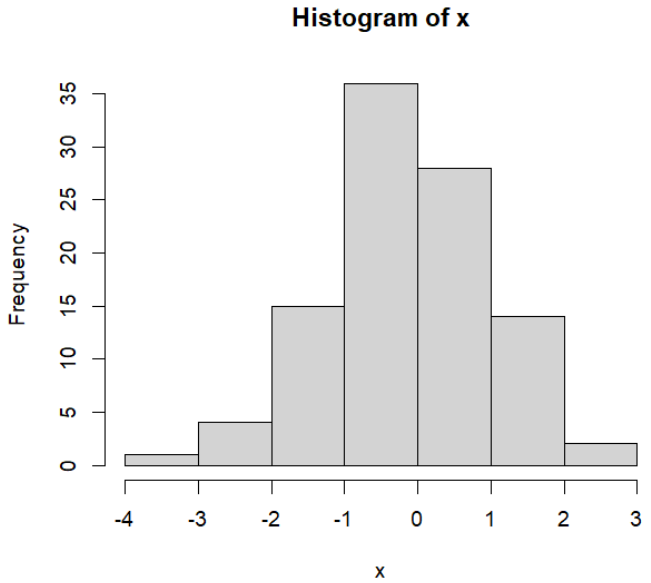
April 27, 2023

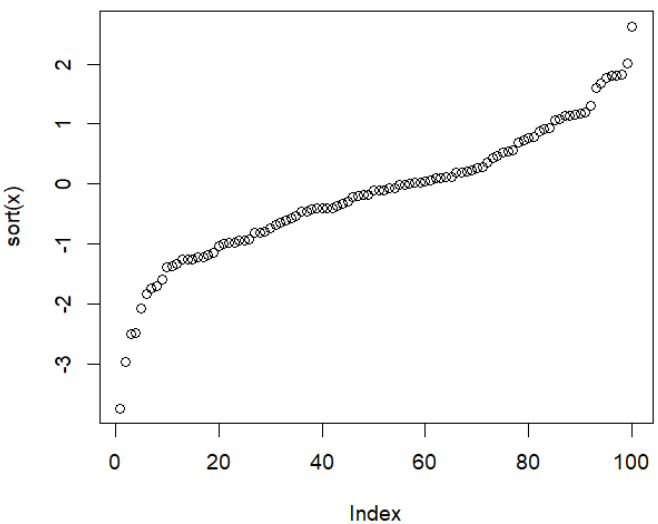
Visualizing Distributions

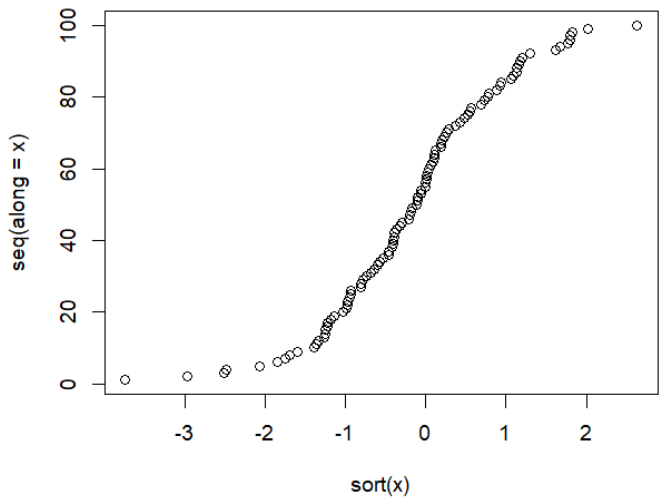
- Histograms and density plots are highly intuitive and visually appealing.
- Both arbitrarily depend on choice of parameters.
  - bin width
  - kernel function
  - bandwidth
- Empirical cumulative distribution functions (ecdfs) and quantile-quantile (q-q) plots solve these problems.
- No parameter choices involved.
- Less intuitive than histograms and density plots.
- Quite popular among statisticians.
- Deserve more widespread use.

100 normally distributed numbers

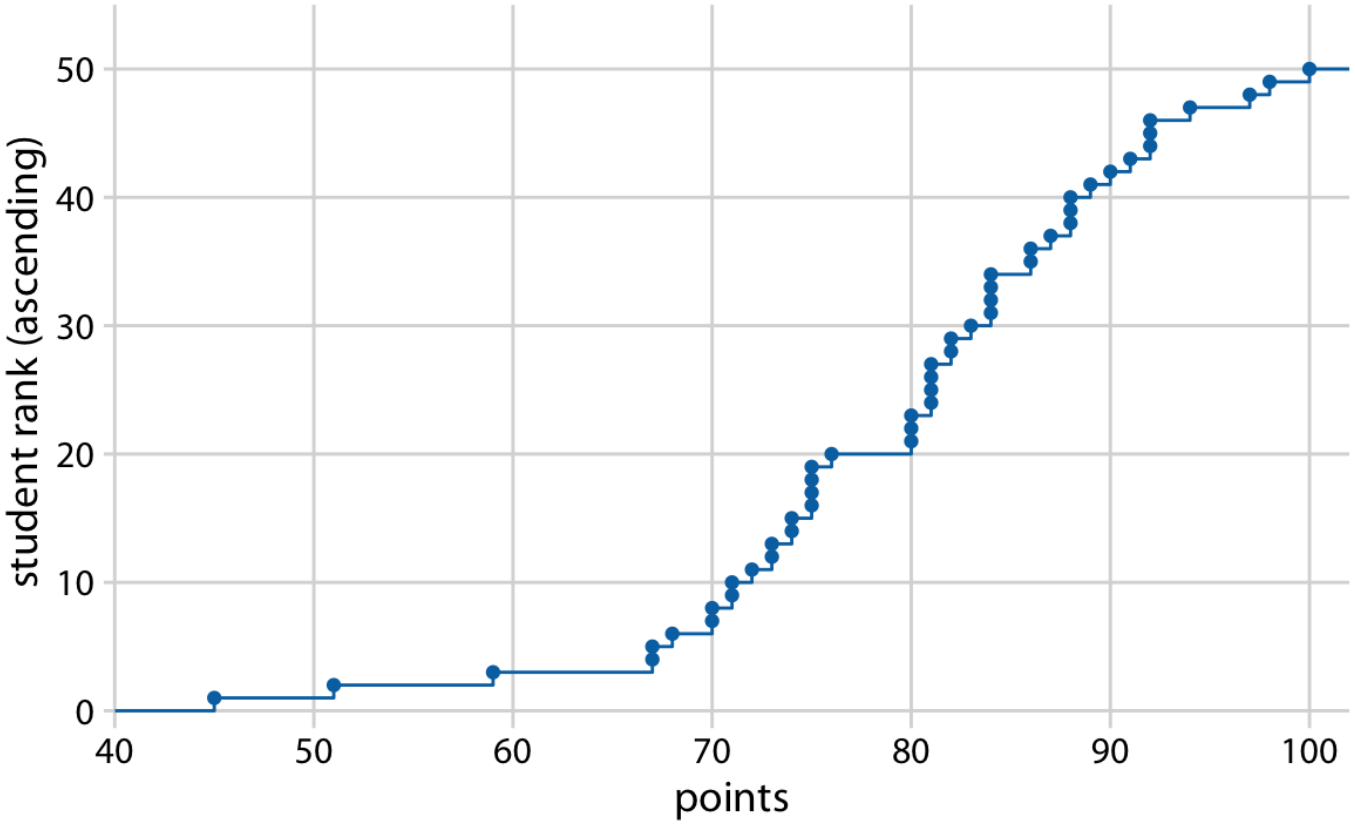






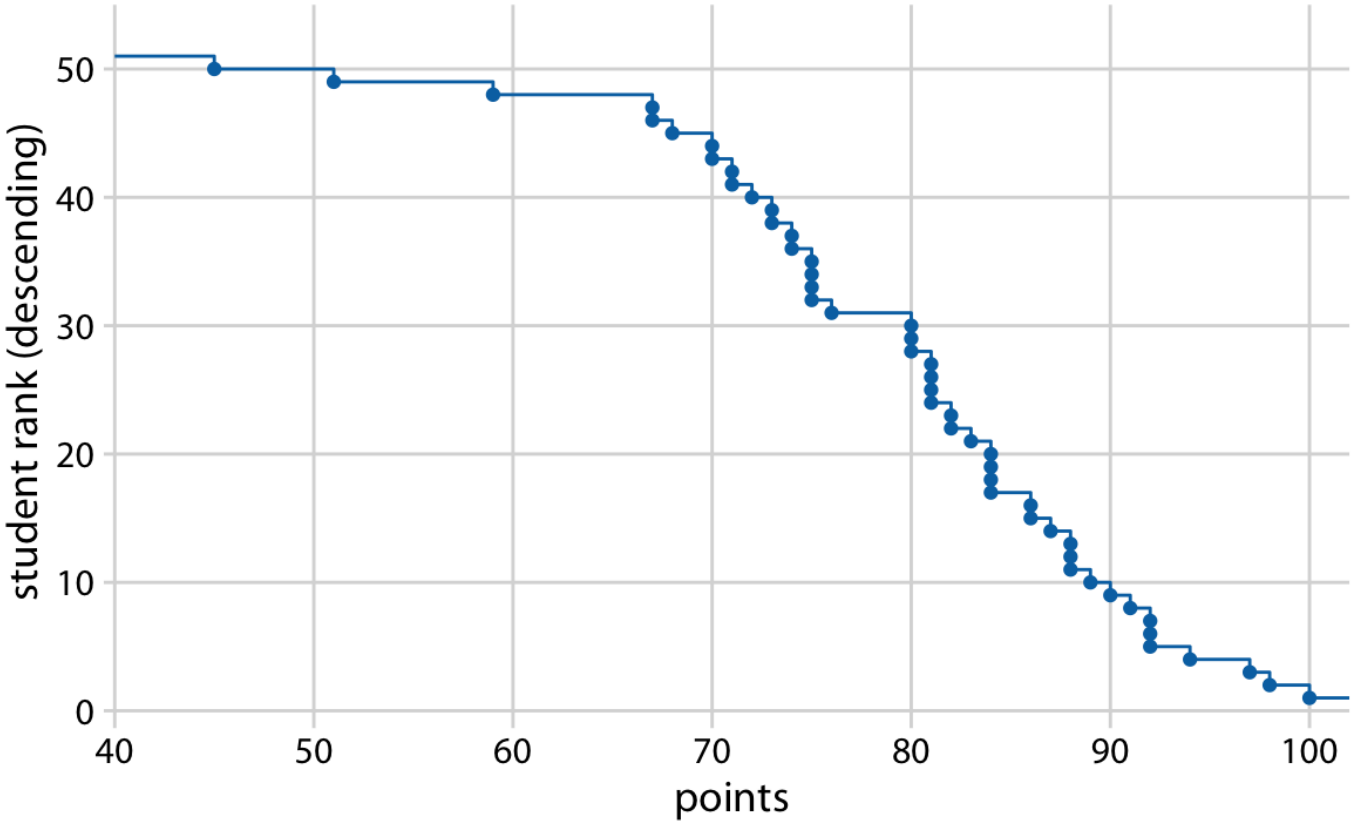


Student grades, an ecdf



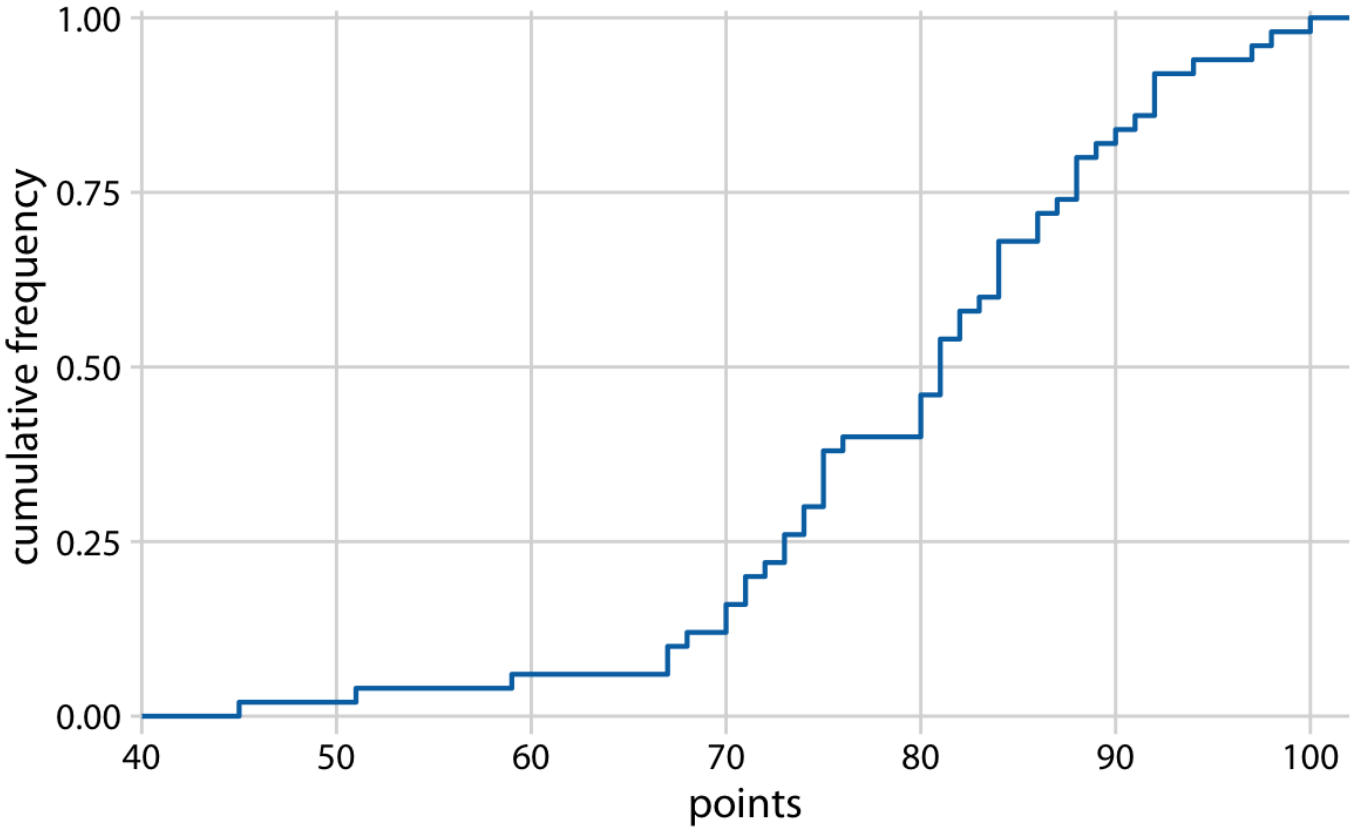
- Each dot is one student
- x axis is points on exam
- y axis is rank in class on exam
- y is proportional to the cumulative density

Student grades, an ecdf sorted the other way



- What portion of students scored above this?

Student grades, an ecdf



- Omit the points and scale y to 1
- A quarter of the students (25%) received less than 75 points.
- The median point value (corresponding to a cumulative frequency of 0.5) is 81.
- Approximately 20% of the students received 90 points or more.
- Helps find cut points of minimum unhappiness.

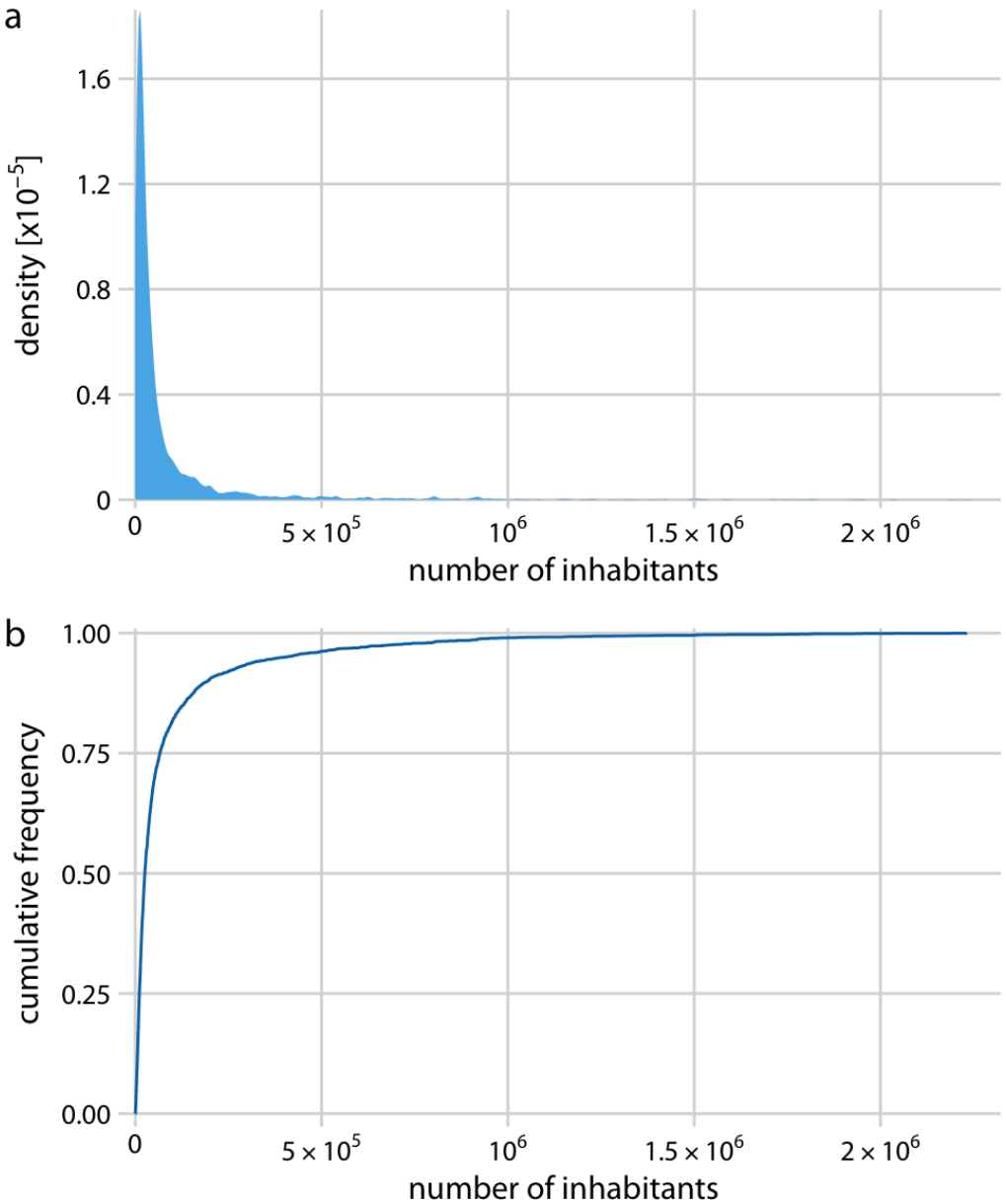
Highly skewed distributions

- The right tail decays slower than exponentially.
- Very large values are not that rare, even if the mean of the distribution is small.
- Examples:
  - the number of people living in different cities or counties,
  - the number of contacts in a social network,
  - the frequency with which individual words appear in a book,
  - the number of academic papers written by different authors,
  - the net worth of individuals,
  - the number of interaction partners of individual proteins in protein–protein interaction networks

Power law distributions

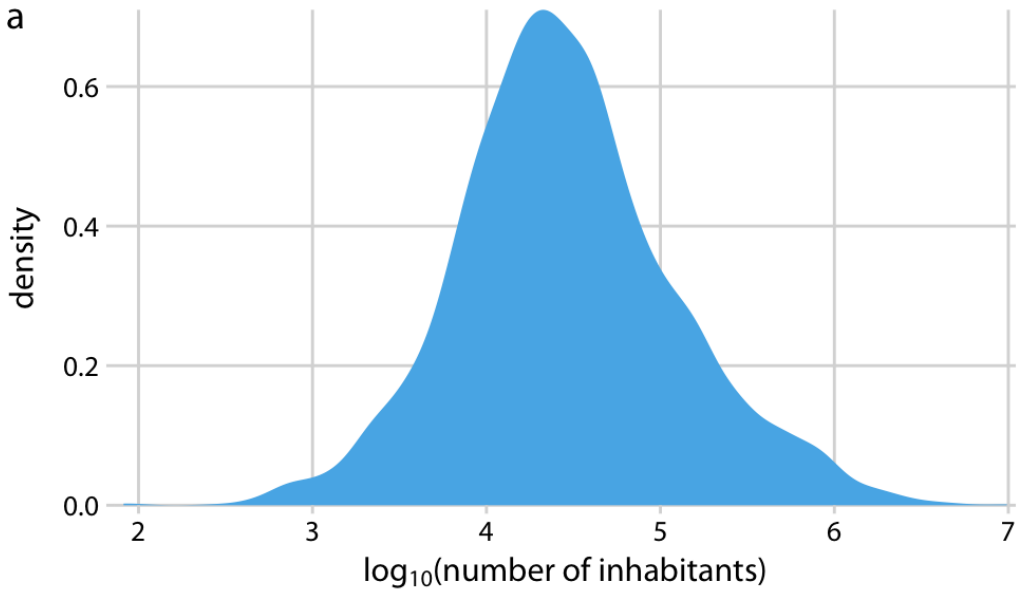
- The likelihood to observe a value that is x times larger than some reference point declines as a power of x.
- Net worth in the US, which is distributed according to a power law with exponent 2.
- At any given level of net worth (say, \$1 million), people with half that net worth are four times as frequent, and people with twice that net worth are one-fourth as frequent.
- The same relationship holds if we use \$10,000 as reference point or if we use \$100 million.
- Also called **scale-free** distributions.

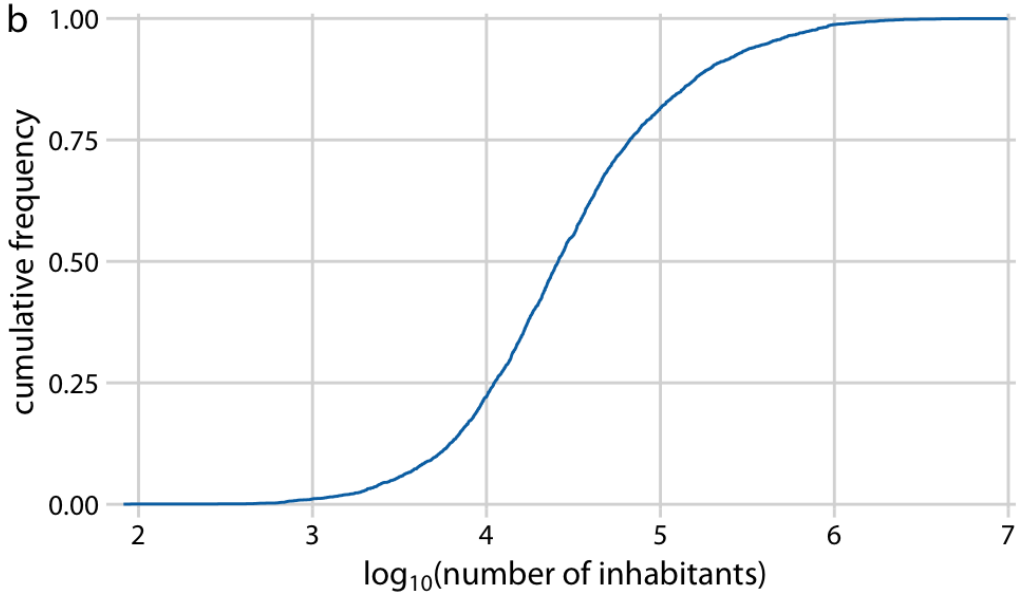
County populations, 2010 US Census



- Median is 25,857
- Los Angeles county is 9,818,605

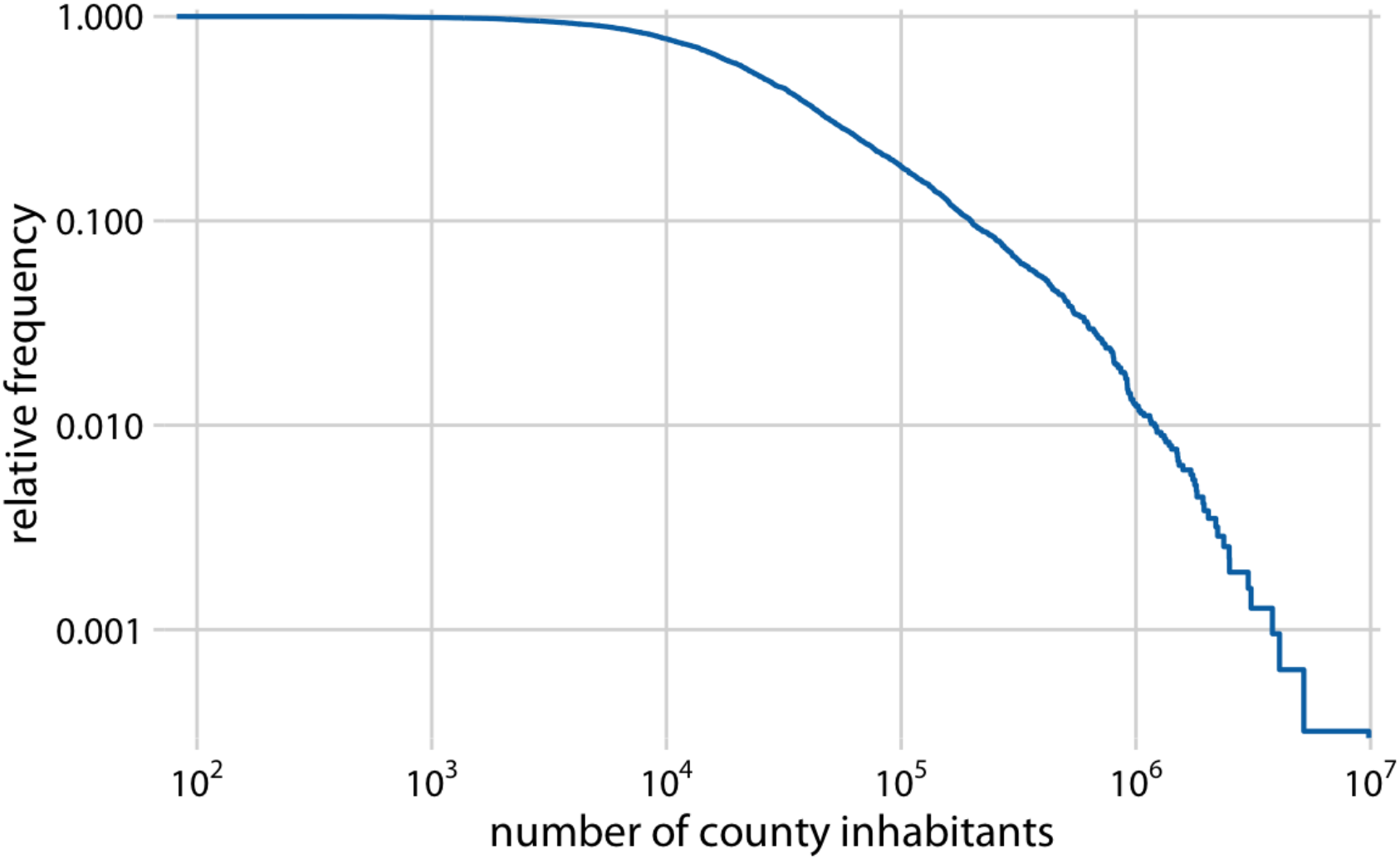
Log transformed data





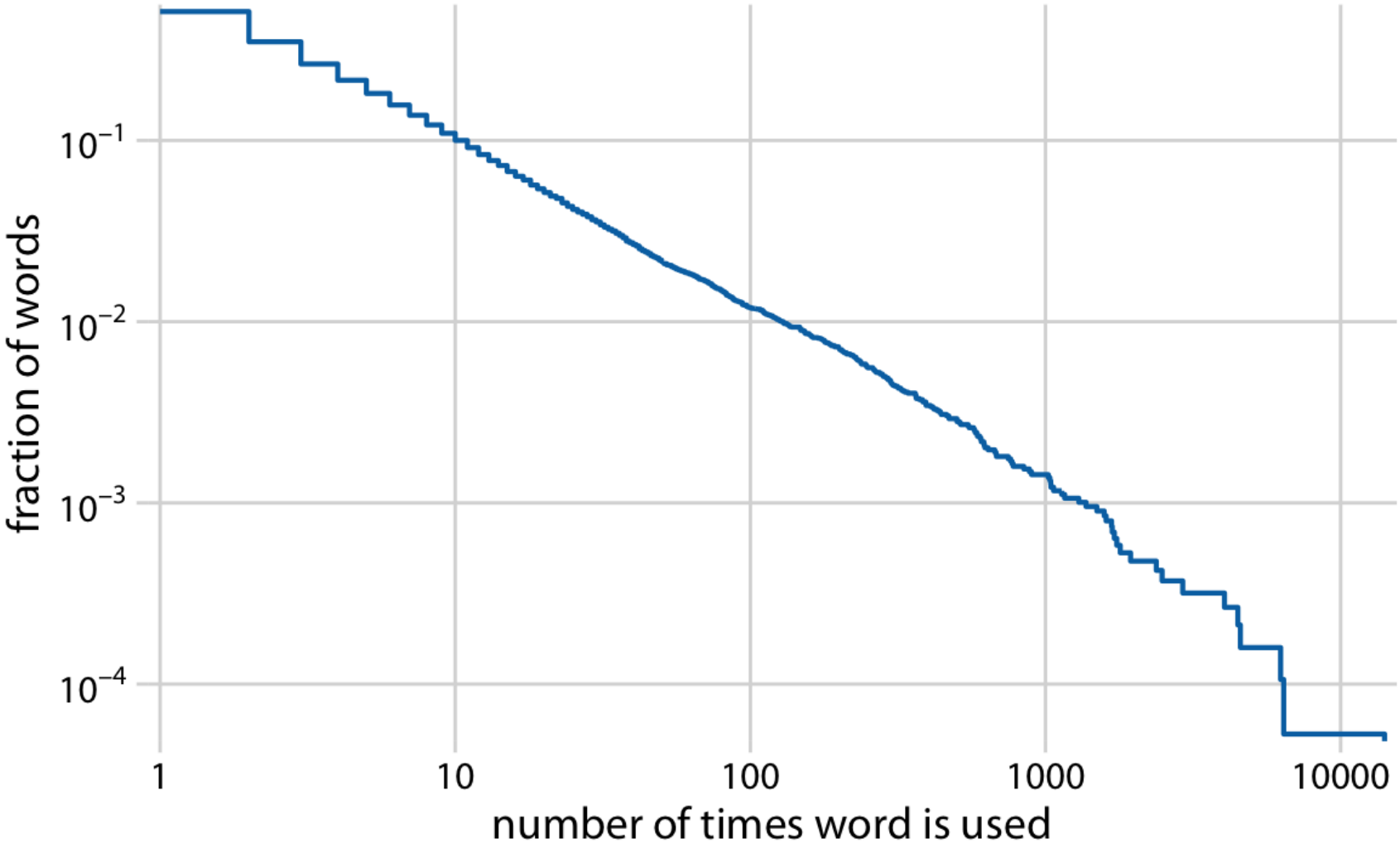
- Not a power law, but a log-normal distribution.

Plot ecdf on log-log axes



- Power law would be straight line
- Right tail is almost a straight line

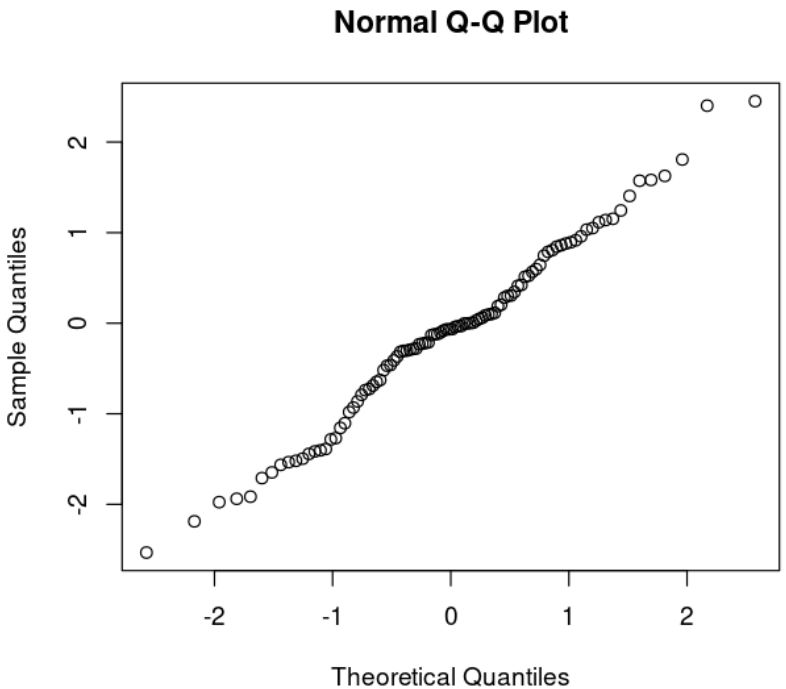
log-log ecdf of word counts in *Moby Dick*

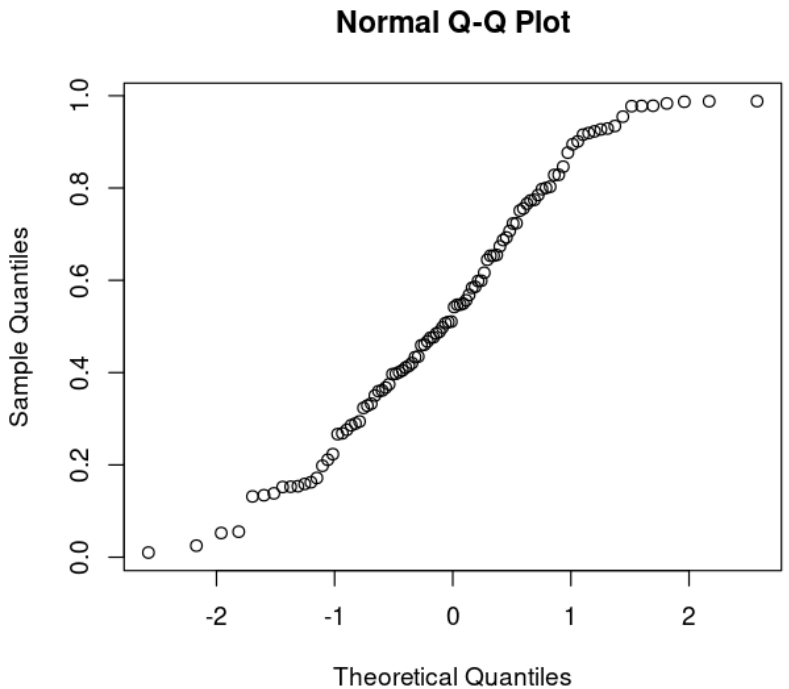


Quantile-quantile (q-q) plots

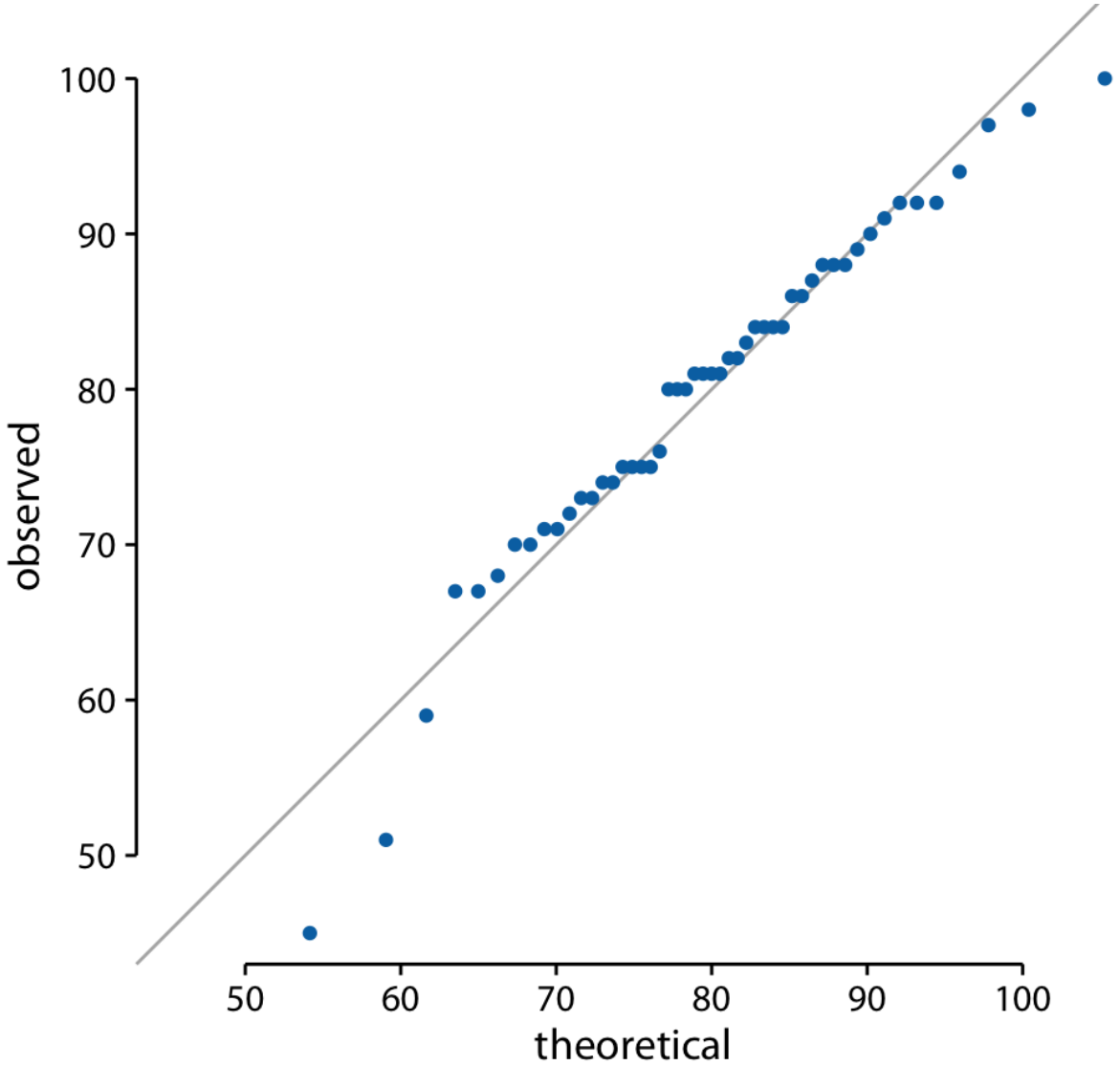
- Assume the actual data values have a mean of 10 and a standard deviation of 3.
- Then, assuming a normal distribution, we would expect:
  - a data point ranked at the 50th percentile to lie at position 10 (the mean),
  - a data point at the 84th percentile to lie at position 13 (one standard deviation above from the mean),
  - and a data point at the 2.3rd percentile to lie at position 4 (two standard deviations below the mean).
- We can carry out this calculation for all points in the dataset and then plot the observed values (i.e., values in the dataset) against the theoretical values (i.e., values expected given each data point's rank and the assumed reference distribution).

q-q plots of a normal and uniform data





q-q plot of student grade data



Do county populations follow a log normal distribution?