Censored data

- Lecture notes are being supplemented with a clarification post on Ed. See here
- Situation when data is incomplete because we can't observe the tail end of data
  - E.g., a long-term studying examining life expectancy of a long-lived species
- Incorrect options for handling data
  - Treating data as missing
    - The data we collect is still informative because we know it falls above (or below) a cutoff
  - Replacing data with the cutoff value
    - Will introduce a bias
- **Ex**: device failure rate
  - Let time until device failure be distributed Exponentially with an unknown rate parameter
  - Let us test n devices
  - Experiment stops at 7 months
    - If the device survives 7+ months, the survival data is censored
  - If $t < 7$, then we use the PDF of an exponential $\lambda e^{-\lambda t}$
  - If $t \geq 7$, then we use the complement of the CDF $e^{-7\lambda}$
  - Why do we mix the PDF and the CDF?
    - When we observe the time, we're observing a crystallization of a continuous r.v., where it's valid to use the PDF
    - If we only observe that the time is greater than some value, we're only observing a crystallization of an indicator r.v., and we obtain the probability of that indicator by using a CDF
  - Resulting likelihood
    - $L(\lambda) = \prod_{j=1}^{n} (\lambda e^{-\lambda t_j})^{\mathbb{1}(t_j < 7)} (e^{-7\lambda})^{\mathbb{1}(t_j \geq 7)}$
  - aL: Consider what would happen if we replaced the observations of $t < 7$ with indicators
    - What would our new MLE be?
    - How would this change the SE of the estimator?