Sampling and Standard Error

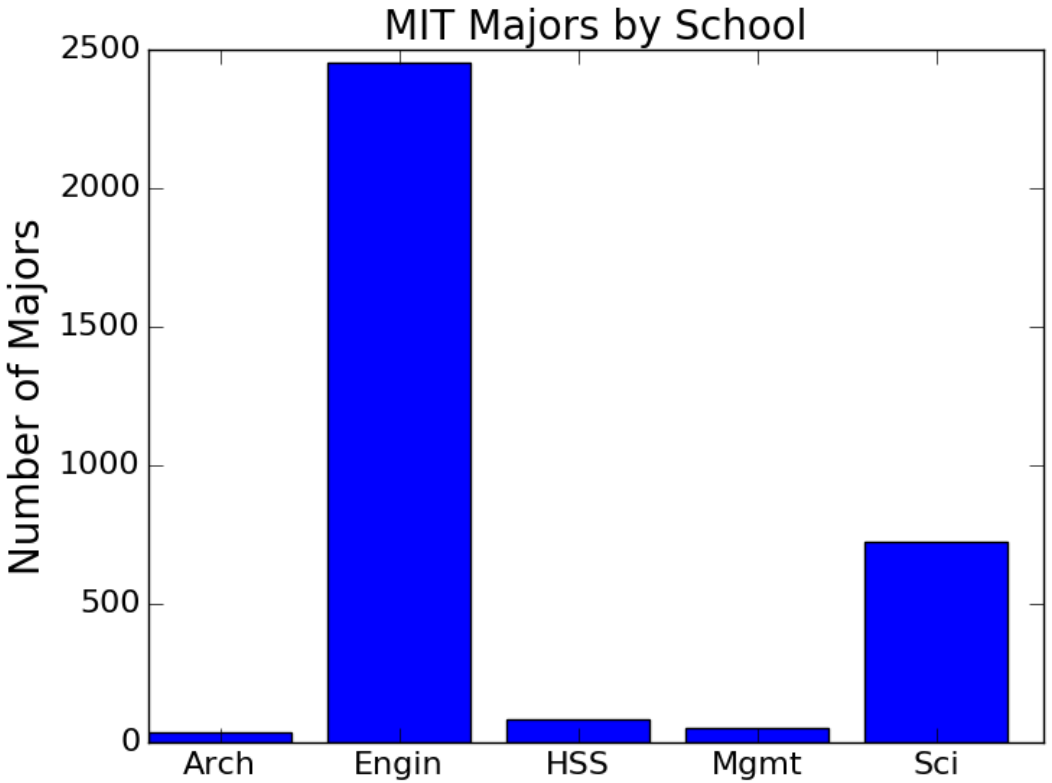
Recall Inferential Statistics

- Inferential statistics: making inferences about a populations by examining one or more random samples drawn from that population
- With Monte Carlo simulation we can generate lots of random samples, and use them to compute confidence intervals
- But suppose we can't create samples by simulation?
  - "According to the most recent poll Clinton leads Trump by 3.7 percentage points in swing states. The registered voter sample is 835 with with a margin of error of plus or minus 4 percentage points."

Probability Sampling

- Each member of the population has a nonzero probability of being included in a sample
- Simple random sampling: each member has an equal chance of being chosen
- Not always appropriate

Stratified Sampling



- Stratified sampling
  - Partition population into subgroups
  - Take a simple random sample from each subgroup

Stratified Sampling

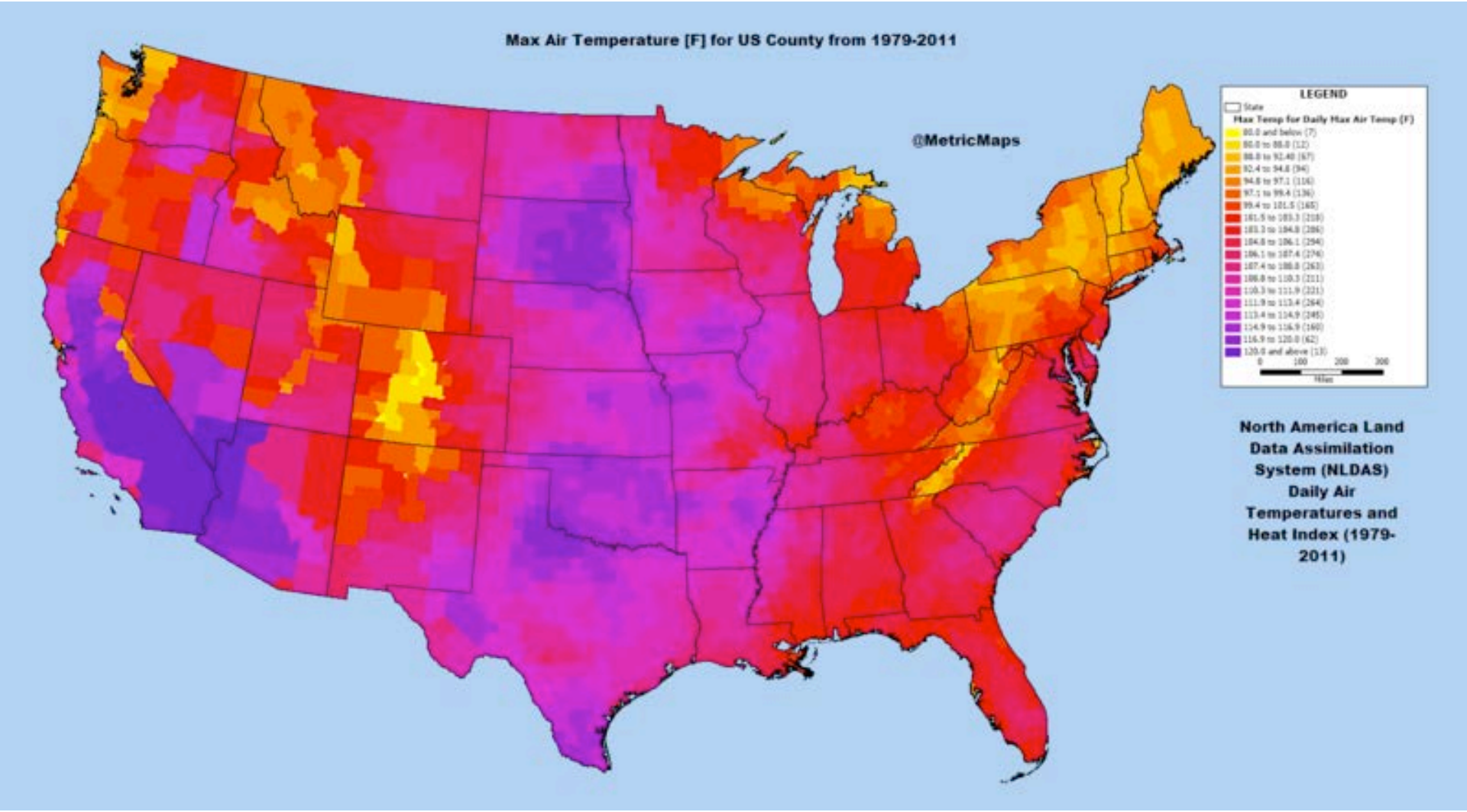
layered

- When there are small subgroups that should be represented
- When it is important that subgroups be represented proportionally to their size in the population
- Can be used to reduced the needed size of sample
  - Variability of subgroups less than of entire population
- Requires care to do properly
- Well stick to simple random samples

Predicting Outcome of an Election

- Approaches
  - Ask every voter → ground truth
  - Draw multiple random samples and compute mean and confidence interval
  - Draw one sample and estimate mean weight and confidence interval using that
- Can't actually ask every voter, so no obvious way to evaluate sampling techniques
- Let's look at an example where we have ground truth

Temperatures in the U.S.



Data

- From U.S. National Centers for Environmental Information (NCEI)
- Daily high and low temperatures for
  - 21 different US cities
    - ALBUQUERQUE, BALTIMORE, BOSTON, CHARLOTTE, CHICAGO, DALLAS, DETROIT, LAS VEGAS, LOS ANGELES, MIAMI, NEW ORLEANS, NEW YORK, PHILADELPHIA, PHOENIX, PORTLAND, SAN DIEGO, SAN FRANCISCO, SAN JUAN, SEATTLE, ST LOUIS, TAMPA
  - 1961 – 2015
  - 421,848 data points (examples)

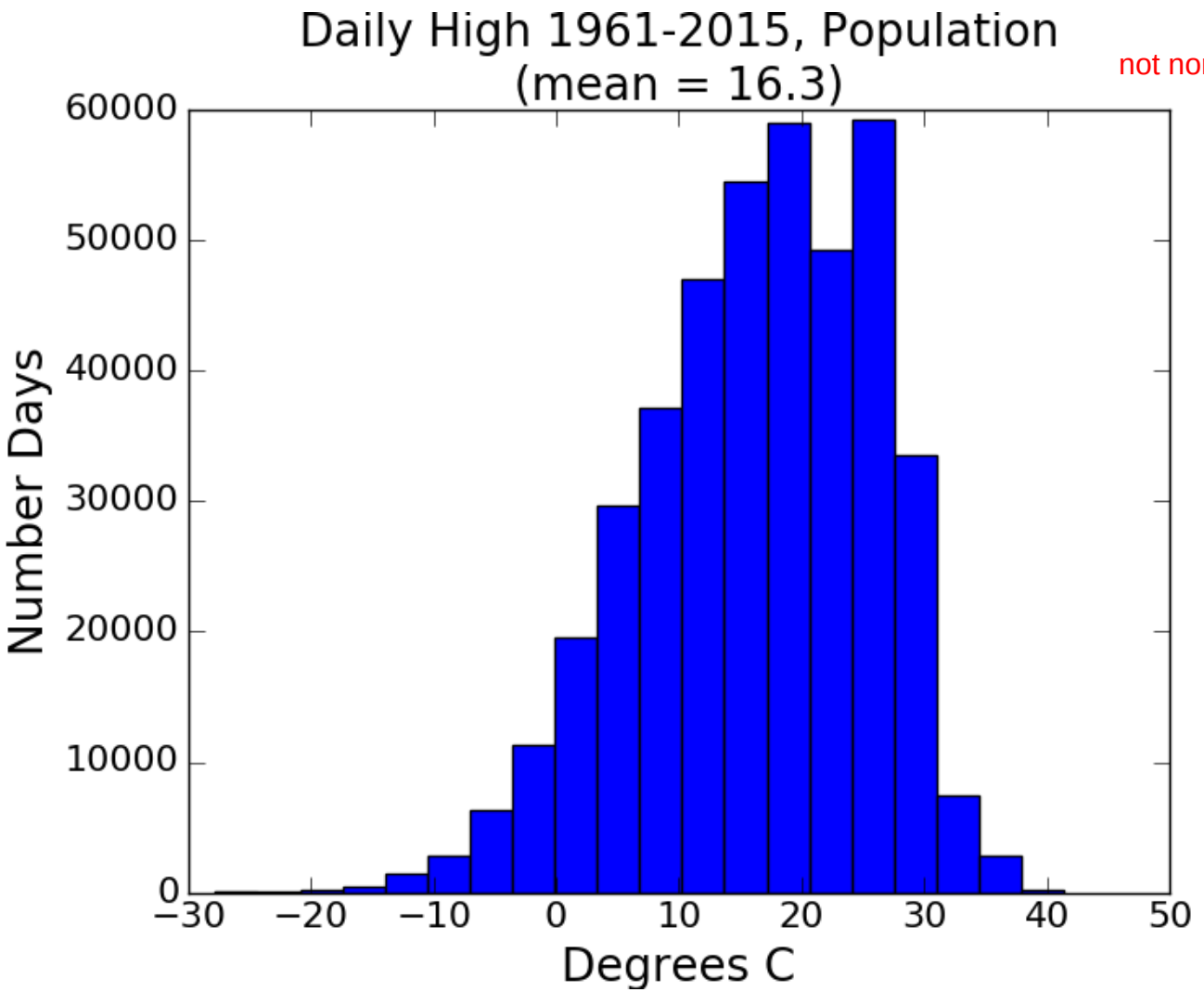
covered nearly all the US except for Alaska and Hawaii

- Let's use some code to look at the data

To Notice in Code

- Function `makeHist` there because I expect too make a lot of histograms, and I wanted to do it with one line of code
- `numpy.std` is function in the numpy module that returns the standard deviation
- `random.sample(population, sampleSize)` returns a list containing `sampleSize` randomly chosen distinct elements of `population`
  - Sampling without replacement

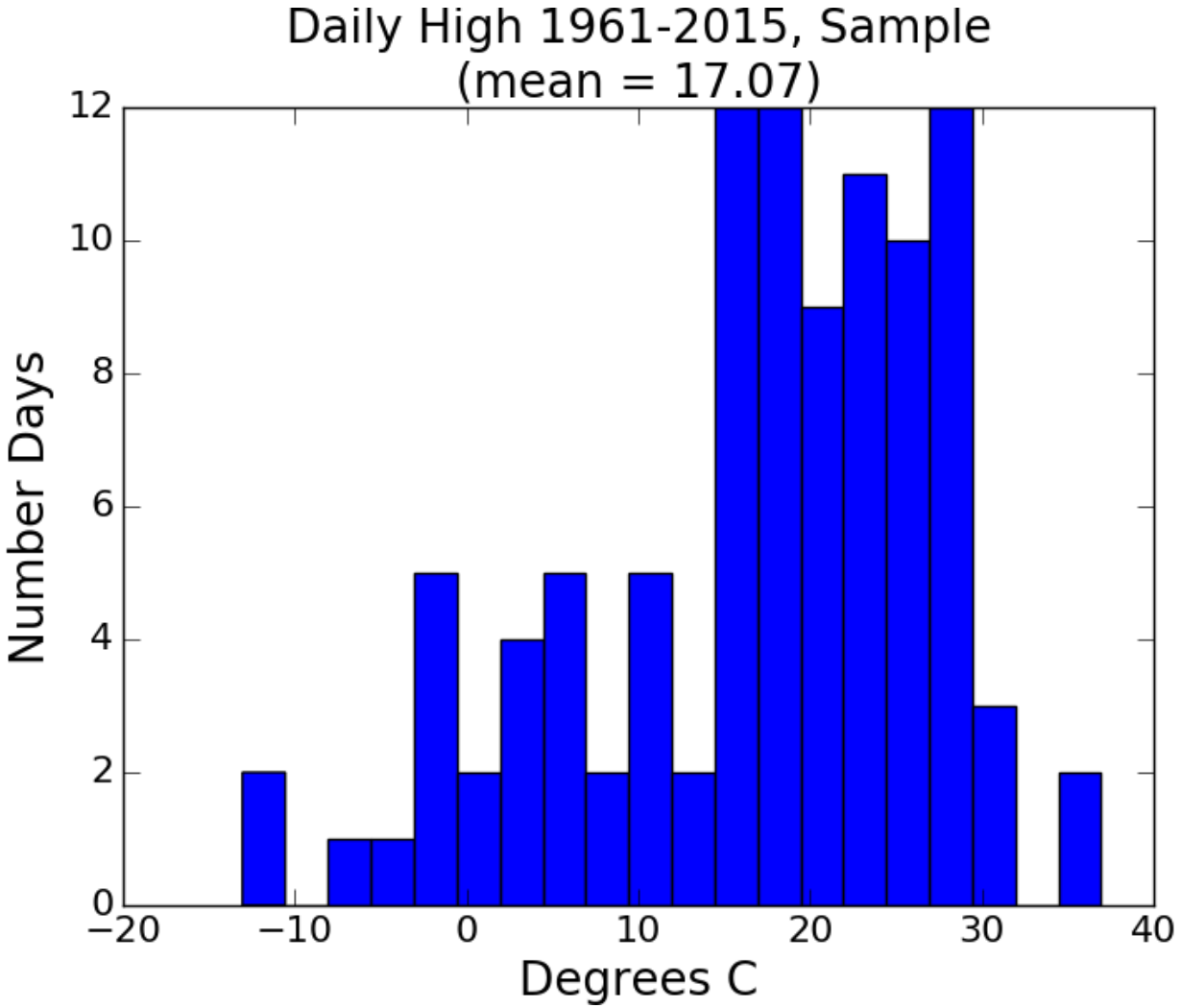
Histogram of Entire Population



not normal distribution

σ = ~9.4

Histogram of Sample of Size 100



σ = ~10.4

Means and Standard Deviations

- Population mean = 16.3
- Sample mean = 17.1
- Standard deviation of population = 9.44
- Standard deviation of sample = 10.4
- A happy accident, or something we should expect?