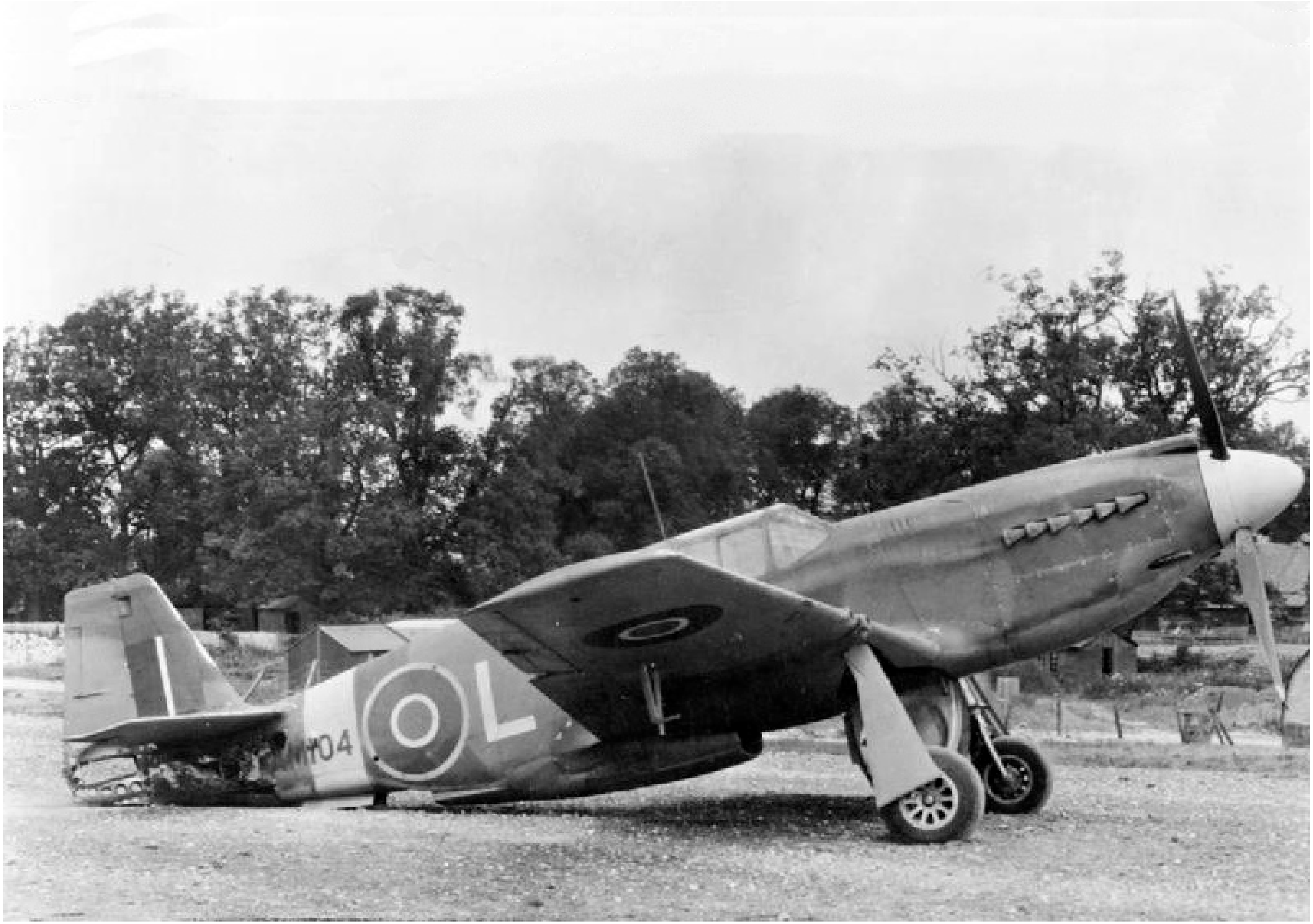
Statistical Abuses and Course Wrap Up

A Mustang Fighter Plane



Sampling

- All statistical techniques are based upon the assumption that by sampling a subset of a population we can infer things about the population as a whole
- As we have seen, *if random sampling is used*, one can make meaningful mathematical statements about the expected relation of the sample to the entire population
- Easy to get random samples in simulations
- Not so easy in the field, where some examples are more convenient to acquire than others

Non-representative Sampling

- "Convenience sampling" not usually random, e.g.,
  - Survivor bias, e.g., course evaluations at end of course or grading final exam in 6.00.2x on a curve
  - Non-response bias, e.g., opinion polls conducted by mail or online
- When samples not random and independent, we can still do things like computer means and standard deviations, but **we shouldn't draw conclusions from them** using things like the empirical rule and central limit theorem.
- Moral: Understand how data was collected, and whether assumptions used in the analysis are satisfied. If not, be wary.

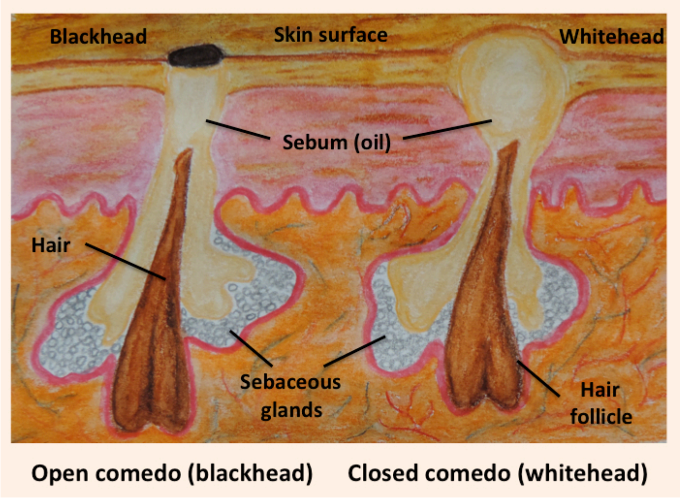
A Comforting Statistic?

- 99.8% of the firearms in the U.S. will not be used to commit a violent crime in any given year
- How many privately owned firearms in U.S.?
- 300,000,000
- 300,000,000*0.002 = 600,000

- Moral: Context matters. A number means little without context.

Relative to What?

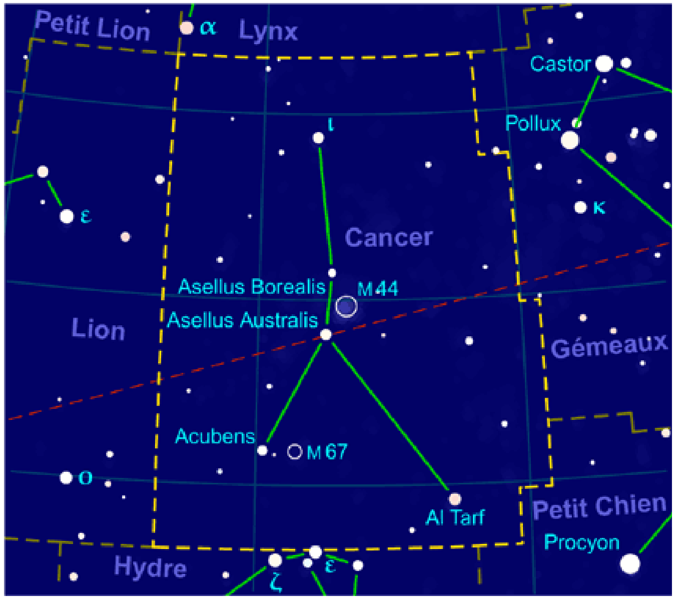
- Consider drugs X and Y for treating acne
  - X cures acne twice as well as Y
  - X kills twice as many acne patients as Y
- Do you want to take X or Y?
  - Suppose Y kills 0.00001% of cases, and cures 50% of them
- Moral: Beware of percentages when you don’t know the baseline



By Hilda Bastian

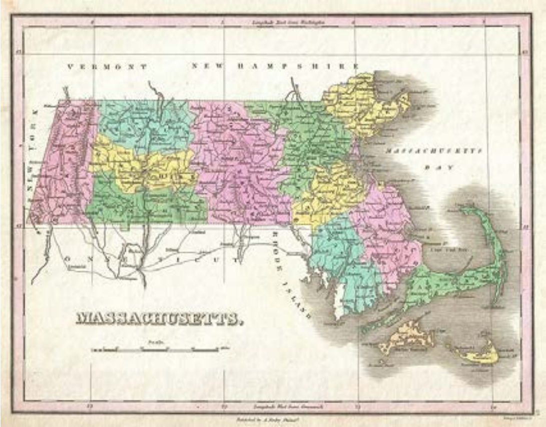
Cancer Clusters

- A **cancer cluster** is defined by the CDC as “a greater-than-expected number of cancer cases that occurs within a group of people in a geographic area over a period of time”
- About 1000 ”cancer clusters” per year are reported to health authorities in the U.S.



A Hypothetical Example

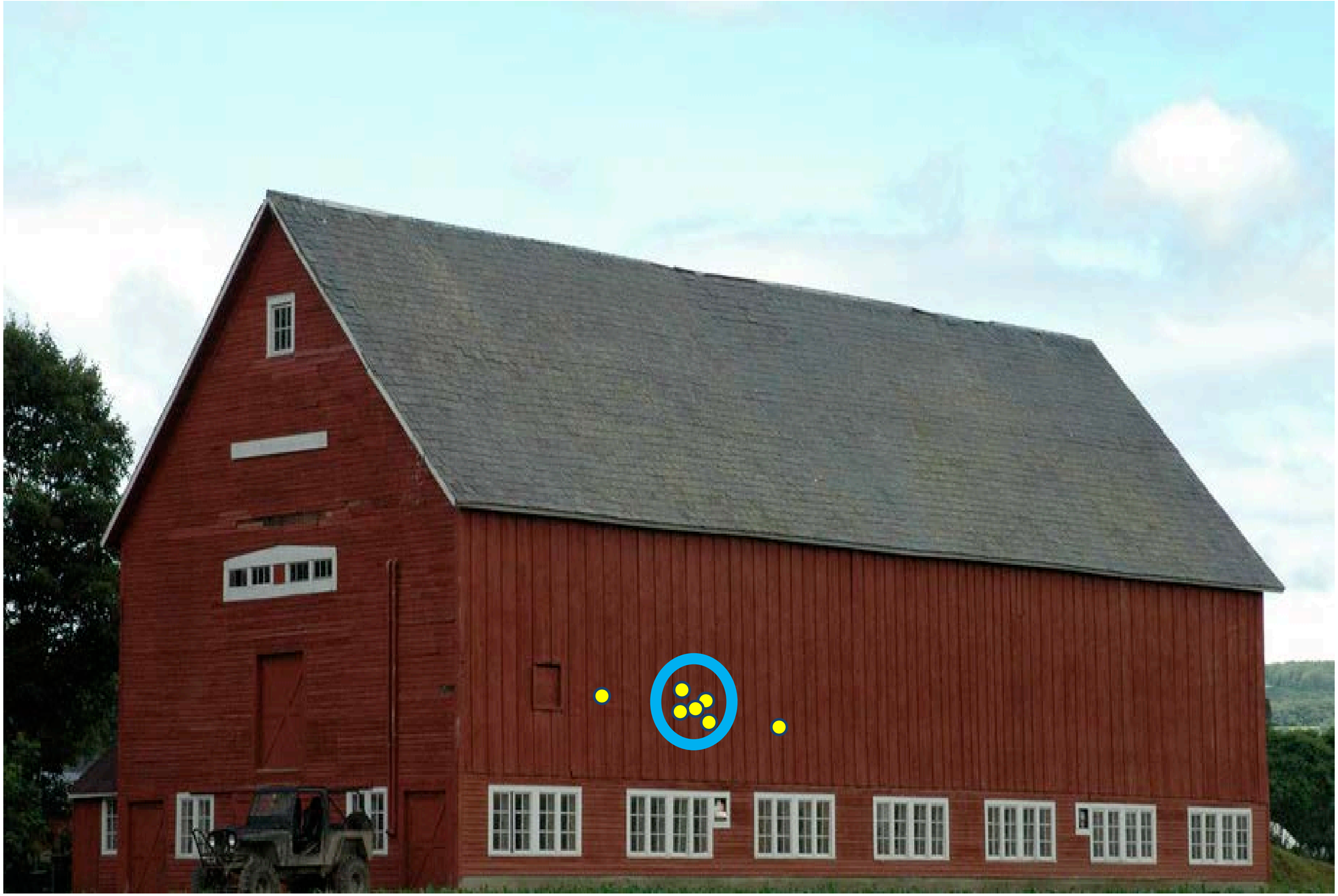
- Massachusetts is about 10,000 square miles
- About 36,000 new cancer cases per year
- Attorney partitioned state into regions of 10 squares miles each, and looked at distribution of cases
- Discovered that region 111 had 143 new cancer cases over a 3 year period
  - More than 32% greater than expected
- How worried should residents be?



Simulate It

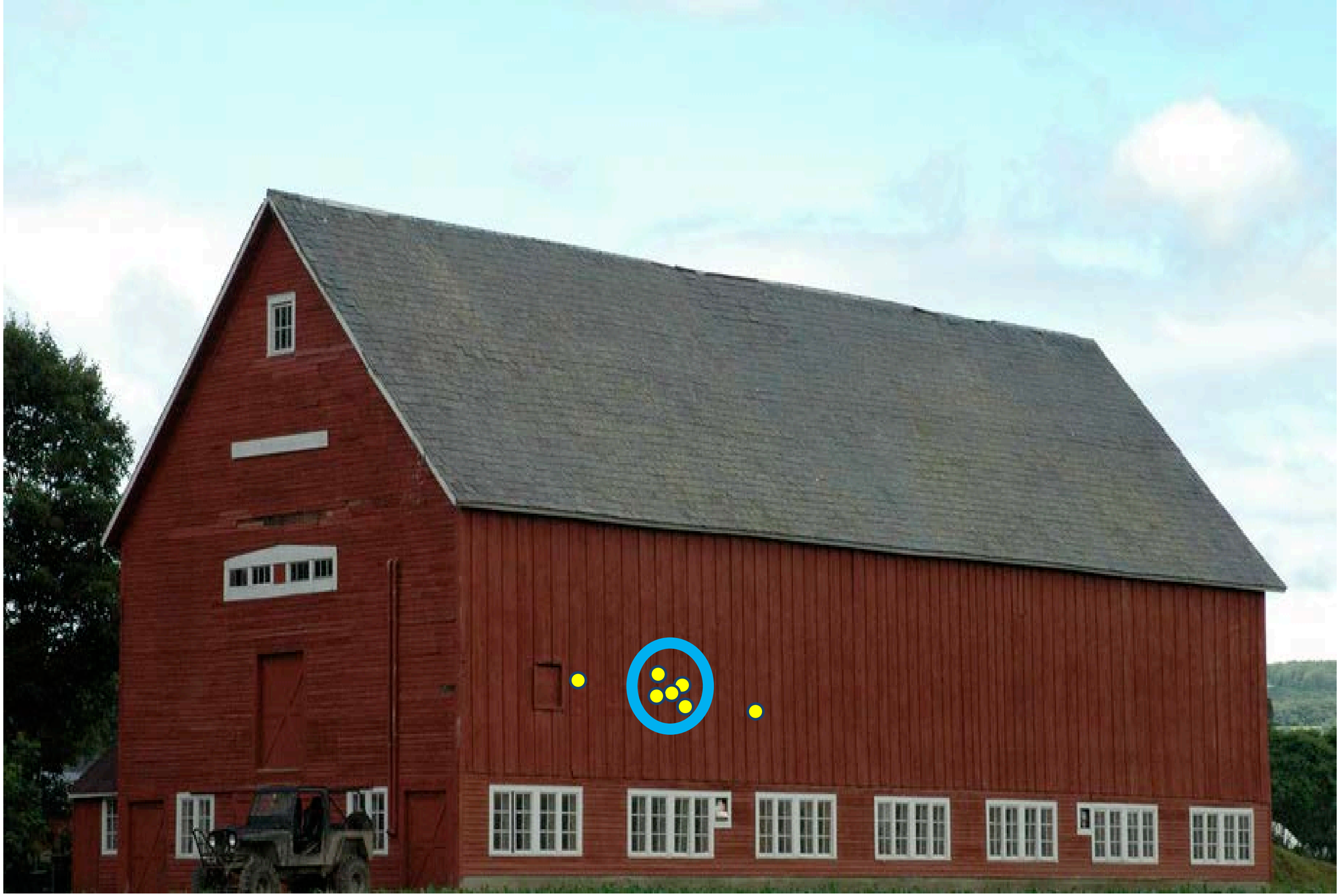
```
numCasesPerYear = 36000
numYears = 3
stateSize = 10000
communitySize = 10
numCommunities = stateSize//communitySize
numTrials = 100
numGreater = 0
for t in range(numTrials):
    locs = [0]*numCommunities
    for i in range(numYears*numCasesPerYear):
        locs[random.choice(range(numCommunities))] += 1
    if locs[111] >= 143:
        numGreater += 1
prob = round(numGreater/numTrials, 4)
print('Est. probability of 111 having\
at least 143 cases =', prob)
```

The Texas Sharpshooter



(modified) CC-BY Image Courtesy of Putneypics

The Texas Sharpshooter



(modified) CC-BY Image Courtesy of Putneypics