MF9130E - Introductory course in Statistics

The Easter egg module

How to reason about probabilities and randomness more generally?

Need to think carefully - cannot just follow the cookbook of probability calculus

- Why? Because each application may present something that looks different under the surface
- Can illustrate this using a so called Sisters' Paradox
- You are told about a family with two children and learn that one of them is a girl. What is the probability that the other one is also a girl (assuming they are not identical twins)?
- Intuition easily suggests 1/2 as the answer, but standard textbook answer is 1/3
- The answer 1/3 is based on the conditional probability $P(C|A) = P(C \cap A)/P(A)$, where A is the event of observing one girl and C is the event with 2 girls.
- Here we assume 2-kid families follow the distribution: $P(GG)=1/4$, $P(BG)=1/2$, $P(BB)=1/4$, so that $P(C \cap A) = 1/4$,$P(A) = 3/4$, and $P(C|A) = 1/3$

Sisters’ Paradox continued

- But there is more here than meets the eye
- Need a *statistical model* generating the observation A for us
- Assume we ring the doorbell for a randomly chosen family and a girl opens the door. We ask if she has one sibling and the answer is yes
- We then consider the *likelihood of observing* this event (A) for each possible family configuration
- P(A|BB) is easy because it is zero (impossible event)
- P(A|GG) is also easy because it is one (certain event)
- P(A|BG) is more tricky and we have multiple options
- P(GG)=1/4, P(BG)=1/2, P(BB)=1/4 are the *prior probabilities* for family configurations, i.e. a *chosen model* for the uncertainty facing us

Sisters' Paradox continued

- Once we have decided suitable probability P(A|BG), we can use the celebrated Bayes' formula to get an answer to our original question:
- P(GG|A) = P(A|GG)P(GG) / [P(A|GG)P(GG)+P(A|BB)P(BB)+P(A|BG)P(BG)]
- Let us now assume that children in the family take randomly turns in answering the door, then P(A|BG) = 1/2
- Plugging this in gives P(GG|A) = 1•1/4 / [1•1/4 + 1/2•1/2 + 0•1/4] = 1/2
- If we instead assume that the children have decided that the girl always answers the door, then P(A|BG) = 1 and correspondingly P(GG|A) becomes 1/3!
- Moral of the story? We need to think carefully about observation processes when making statistical statements (more complete story can be read here).

More about randomness

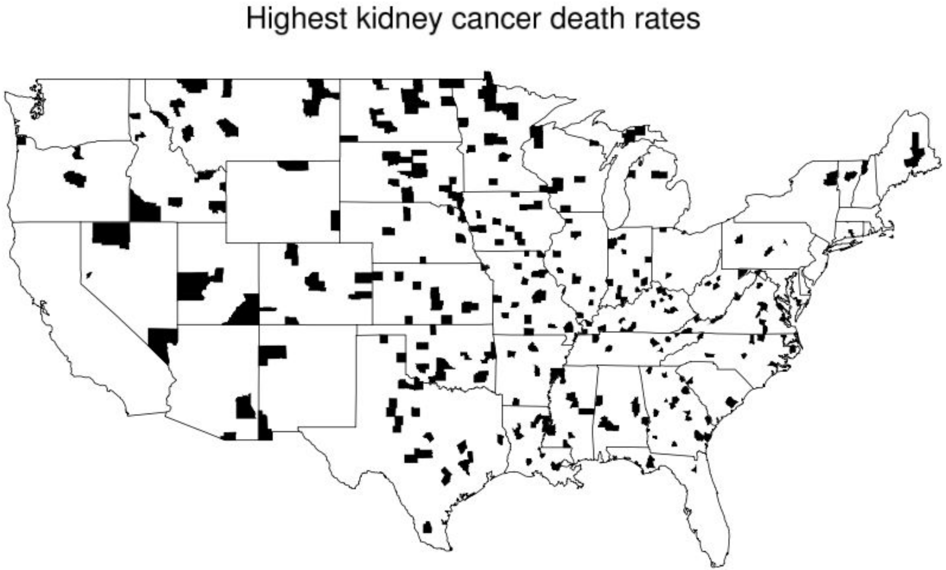


Figure 2.6 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.7 for discussion.*

From Gelman et al. BDA book see also this article in Statistics in Medicine by Gelman and Price

A puzzling pattern on the map

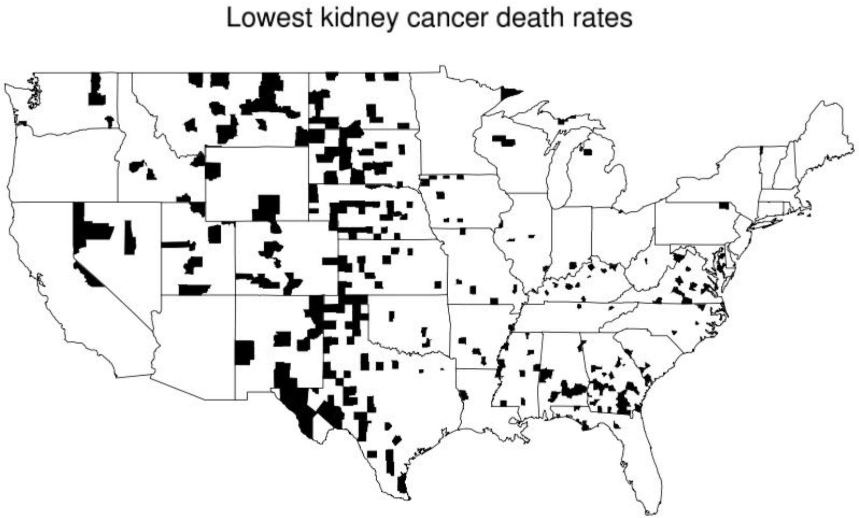


Figure 2.7 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.6.*

From Gelman et al. BDA book see also this article in Statistics in Medicine by Gelman and Price

Recap of the cancer map puzzle

- The counties with small population size (say 1,000) would tend to have zero cases of death from this particular (rare) cancer in a 10-year register-based epidemiological surveillance
- Thus, they end up among the counties with 10% lowest cancer death rates
- However, as there are hundreds of small counties, occasionally even in them, there will be 1 such cancer death observed
- This results in a death rate estimate well above the national average, pushing the county into the list with 10% highest cancer death rates
- This highlights that epidemiological maps can be highly misleading, especially when raw data are used
- Paradoxically maps can also mislead us even when the data are smoothed with a model since it can introduce spatial artefacts, as discussed here

How to relate to randomness

- In many bio-medical applications we need to consider *stochastic variation* occurring both in *unobservable* layers of the world and in our *observations* (samples, measurements)
- We typically use statistical models with parameters relating to the unobservable aspects of study populations (e.g. cancer risk) and assume that observations (e.g. cancer cases) are generated from some distribution defined by the parameters (clarified below)
- It is important to notice that the parameters can be seen as random variables themselves, depending on the application
- The above cancer map case study illustrates this point as it is natural to think that the cancer risk (parameter) can vary over counties and time due to variation in exposures and other population characteristics that may differ considerably between local populations
- When appropriate models with parameters are too difficult (or time-consuming) to define, we may want to resort to non-parametric statistical methods that make less restrictive assumptions about the nature of the data (e.g. shape of distribution)
- There is an extensive literature on advanced non-parametric methods for spatial epidemiology (for a case study see here)
- In this course we restrict attention to basic non-parametric tests

Data transformations and more advanced non-parametric tests

Antibody titer measurements are typical example of distributions where non-parametric tests of location shift may not be enough for comparison

Fig 1 in this article illustrates the issues: https://www.nature.com/articles/s41598-017-06177-0

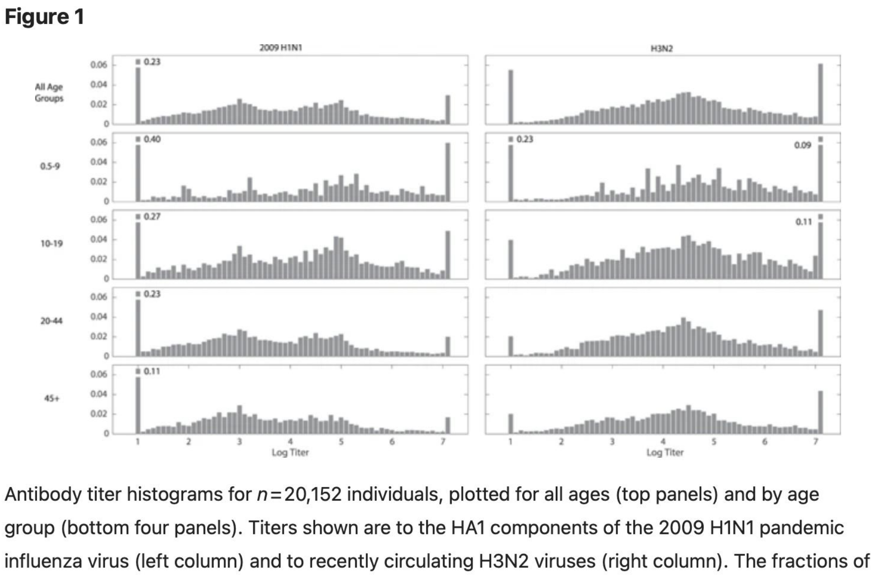
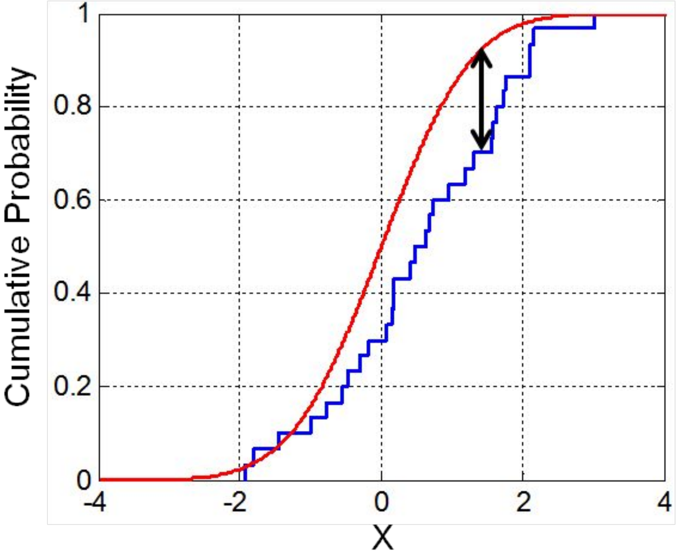


Figure 1

Antibody titer histograms for $n = 20{,}152$ individuals, plotted for all ages (top panels) and by age group (bottom four panels). Titers shown are to the HA1 components of the 2009 H1N1 pandemic influenza virus (left column) and to recently circulating H3N2 viruses (right column). The fractions of

Kolmogorov-Smirnov test as a generic non-parametric comparison of two distributions

https://en.wikipedia.org/wiki/Kolmogorov–Smirnov_test


Cumulative Probability
X