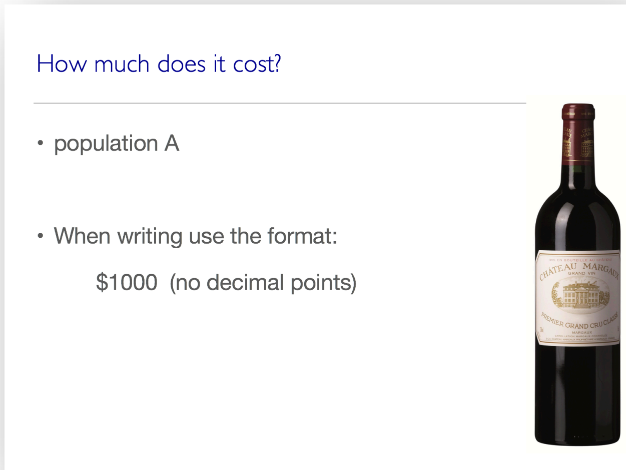
Data Mining & Machine Learning

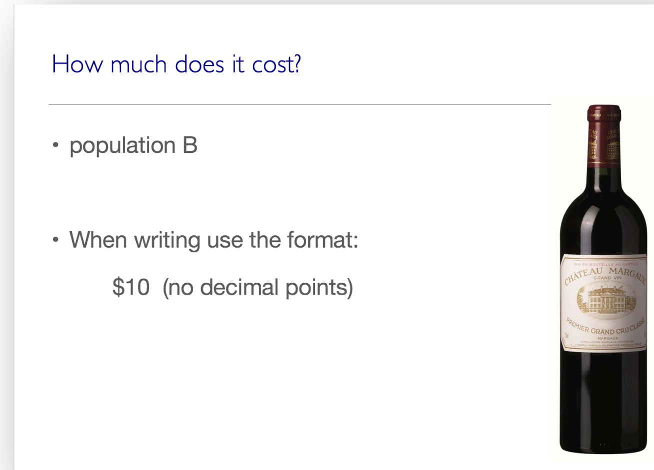
CS37300
Purdue University

September 6, 2017

Decision making

Testing Hypothesis with Small Number of Observations





Hypothesis (Anchoring): Unrelated number biases wine price assessment
Experiment: Expose pop A to $1000 and pop B to $10
Population size: ~7 in each group

Answers from folks that saw $1000 as the amount
[50,10,37,650,400,80,130] … average $193

Answers from folks that saw $10 as the amount format
[20,30,60,10,100,40] … average $43

How can we test if hypothesis is true?

Is the Anchoring Hypothesis True?

- Answers from folks that saw \$1000 as the format
  - [50,10,37,650,400,80,130] … **empirical** average \$193
  - Assume these values are independent observations from a Poisson distribution with an unknown average $\mu_{\$1000}$

$$X_{\$1000,i} \sim \text{Poisson}(\mu_{\$1000})$$ [the i-th observation in our vector]

    - Is $\mu_{\$1000} = 193$?
      - No, 193 is just the empirical value. The true value of $\mu_{\$1000}$ is unknown

- Answers from folks that saw \$10 as the format
  - [20,30,60,10,100,40] … **empirical** average \$43

$$X_{\$10,i} \sim \text{Poisson}(\mu_{\$10})$$

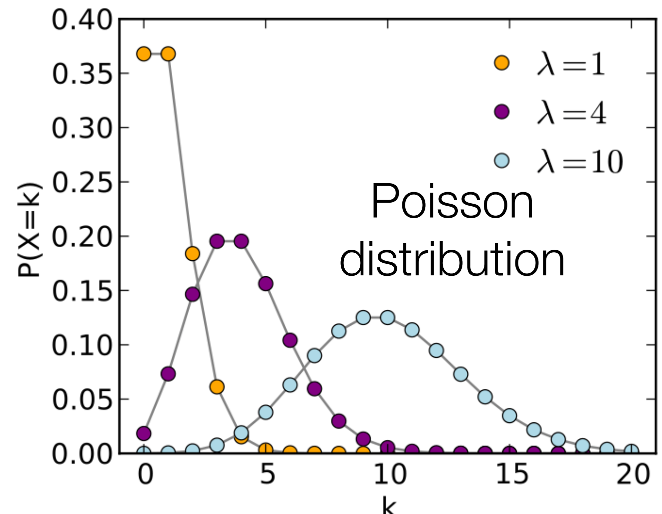
    - Note that $\mu_{\$10} \neq 43$

- How can we decide if $\mu_{\$1000} > \mu_{\$10}$?
  - Last class we saw that possible answers through simulations
  - But simulating by resampling the data alone (bootstrapping) is not good for small datasets… we need another approach...

Bayesian Decision Making (assumptions)

- Consider the assumptions

$$X_{\$1000,i} \sim \text{Poisson}(\mu_{\$1000})$$
$$X_{\$10,i} \sim \text{Poisson}(\mu_{\$10})$$



Always check distribution
to see if assumption makes sense

- What we want:

$$P[\mu_{\$1000} > \mu_{\$10} | X_{\$1000,1}, \ldots, X_{\$1000,7}, X_{\$10,1}, \ldots, X_{\$10,7}]$$

- Let's start with trying to obtain

$$P[\mu_{\$1000} | X_{\$1000,1}, \ldots, X_{\$1000,7}]$$

- How can we get the above distribution?

Bayesian Decision Making (Bayes Rule)

- Bayes rule: $$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

- Applying to our problem:

$$P[\mu_{\$1000}|X_{\$1000,1},\ldots,X_{\$1000,7}] = \frac{P[X_{\$1000,1},\ldots,X_{\$1000,7}|\mu_{\$1000}]P[\mu_{\$1000}]}{P[X_{\$1000,1},\ldots,X_{\$1000,7}]}$$

Where:

1. $P[X_{\$1000,1},\ldots,X_{\$1000,7}|\mu_{\$1000}] \Rightarrow$ Joint probability of the observations given a true average
2. $P[\mu_{\$1000}] \Rightarrow$ Prior probability of a true average
3. $P[X_{\$1000,1},\ldots,X_{\$1000,7}] \Rightarrow$ Probability of the data

Bayesian Decision Making (Bayes Rule)

- How can we get $P[X_{\$1000,1},...,X_{\$1000,7}|\mu_{\$1000}]$?
  - Observations are independent

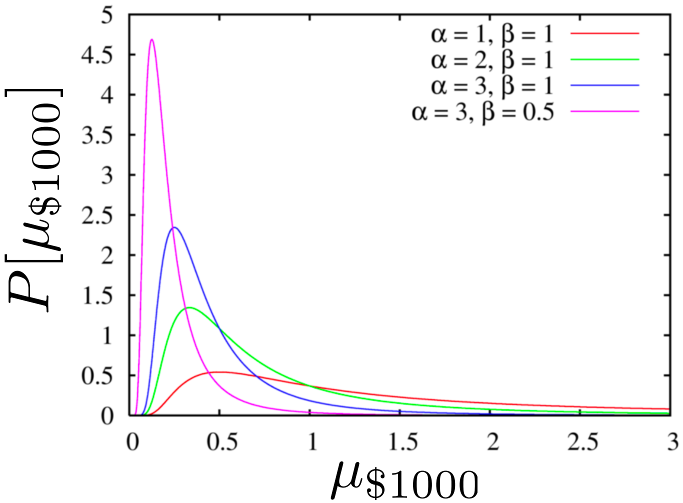
$$P[X_{\$1000,1},...,X_{\$1000,7}|\mu_{\$1000}] = \prod_{i=1}^{7} P[X_{\$1000,i} = k_i|\mu_{\$1000}] = \prod_{i=1}^{7} \frac{(\mu_{\$1000})^{k_i}}{k_i!} e^{-\mu_{\$1000}}$$

Probability of Poisson distribution ✅

- How can we get $P[\mu_{\$1000}]$?
  - This is prior knowledge… something we would “know” without seeing the data
  - This is up to us… so, we make our lives easy by assuming a Gamma distribution

$$P[\mu_{\$1000}] = \text{Gamma}(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu_{\$1000}^{\alpha-1}e^{-\beta\mu_{\$1000}}$$

The choice of α and β
is subjective and not influenced by the data



Bayesian Decision Making (computing the posterior)

- How can we get $P[X_{\$1000,1}, \ldots, X_{\$1000,7}]$?
  - We can marginalize the joint probability

$$P[X_{\$1000,1}, \ldots, X_{\$1000,7}] = \int_{\mu_{\$1000}} P[X_{\$1000,1}, \ldots, X_{\$1000,7}, \mu_{\$1000}] d\mu_{\$1000}$$

$$= \prod_{i=1}^{7} \int_{\mu_{\$1000}} P[X_{\$1000,i} = k_i | \mu_{\$1000}] P[\mu_{\$1000}]$$

- Putting it all together with some little Calculus, we get the following posterior (see Notes conjugate.pdf)

$$P[\mu_{\$1000} | X_{\$1000,1}, \ldots, X_{\$1000,7}] = \text{Gamma}\left(\alpha + \sum_{i=1}^{7} X_{\$1000,i}, \beta + 7\right)$$

Decision Making using Bayesian Monte Carlo Methods

- Now we know how to get both:

$$P[\mu_{\$1000}|X_{\$1000,1},\ldots,X_{\$1000,7}] = \text{Gamma}\left(\alpha + \sum_{i=1}^{7} X_{\$1000,i}, \beta + 7\right)$$

and

$$P[\mu_{\$10}|X_{\$10,1},\ldots,X_{\$10,7}] = \text{Gamma}\left(\alpha + \sum_{i=1}^{7} X_{\$10,i}, \beta + 7\right)$$

- How can we obtain $P[\mu_{\$1000} > \mu_{\$10}|X_{\$1000,1},\ldots,X_{\$1000,7},X_{\$10,1},\ldots,X_{\$10,7}]$?
  - We can just simulate from the two posteriors above and get an estimate!

Python Code

```
In [1]: import numpy as np
```

```
In [2]: # Answers from folks that saw $1000 as the amount
        X1000 = np.array([50,10,37,650,400,80,130])
```

```
In [3]: # Answers from folks that saw $10 as the amount format
        X10 = np.array([20,30,60,10,100,40])
```

```
In [4]: # Sum of X1000
        sum_X1000 = np.sum(X1000)
        # Number of observations
        n1000 = X1000.shape[0]

        # Sum of X10
        sum_X10 = np.sum(X10)
        n1000 = X1000.shape[0]
        n10 = X10.shape[0]
```

```
In [5]: # Number of simulation runs
        N = 500

        # Prior parameters
        alpha = 10
        beta = 1

        # N samples of the posterior P[mu1000 | X1000]
        sampled_mu1000 = np.random.gamma(shape=(alpha + sum_X1000), scale=1./(beta + n1000), size=N)

        # N samples of the posterior P[mu10 | X10]
        sampled_mu10 = np.random.gamma(shape=(alpha + sum_X10), scale=1./(beta + n10), size=N)
```

```
In [6]: #Sum number of times sample j of mu1000 is greater than sample j of mu10
        total_1000_ge_10 = np.sum(sampled_mu1000 > sampled_mu10)
```

```
In [7]: print("Empirical probability P[mu_1000 > mu_10 | Data] =", float(total_1000_ge_10)/N)

        Empirical probability P[mu_1000 > mu_10 | Data] = 1.0
```

```
In [8]: #Visual inspection of posterior samples
        print("sampled_mu1000 =",sampled_mu1000,"\n")
        print("sampled_mu10=",sampled_mu10)
```

Decision Making as a Probability Problem

Psychological heuristics and biases

- Tversky & Kahneman, psychologists, propose that people often do not follow rules of probability when making decisions
- Instead, decision making may be based on heuristics
  - Lowers cognitive load but may lead to systematic errors and biases

- Examples:
  - Availability heuristic
  - Representativeness heuristic
  - Confirmation bias
  - Conjunction fallacy (we will not cover this)
  - Numerosity heuristic (we will not cover this)

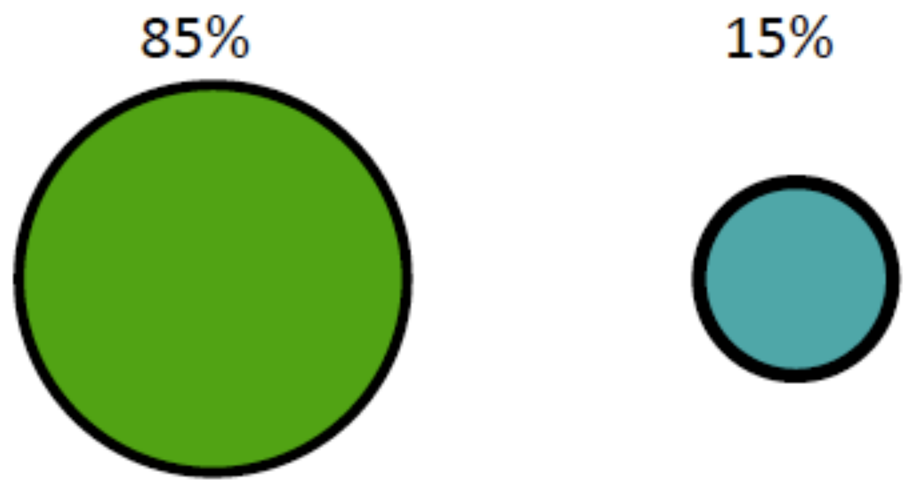
Neglecting base rates

- Taxi-cab problem *(Tversky & Kahneman ‘72)*
    - 85% of the cabs are Green
    - 15% of the cabs are Blue
    - An accident eyewitness reports a Blue cab
    - But she is wrong 20% of the time.
- What is the probability that the cab is Blue?
    - Participants tend to overestimate probability, most answer 80%
    - They ignore baseline prior probability of blue cabs.

More on neglecting base rates

A priori (beforehand)

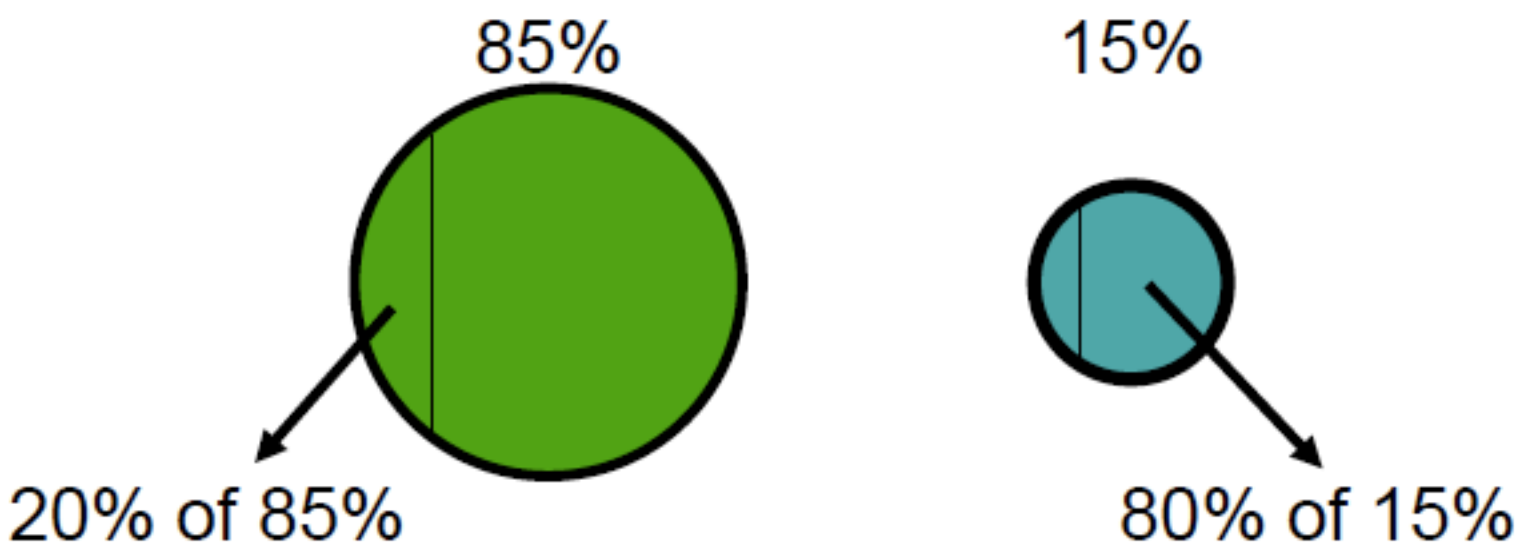
$$P(green) = 0.85$$
$$P(blue) = 0.15$$



$$P(seeBlue|blue) = 0.80$$
$$P(seeBlue|green) = 0.20$$

More on neglecting base rates

After accident (only cars reported as being blue)



More on neglecting base rates

- How to compute probability

$$
\begin{aligned}
P(blue|seeBlue) &= \frac{P(blue \cap seeBlue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue|blue)P(blue) + P(seeBlue|green)P(green)} \\
&= \frac{0.80 \cdot 0.15}{(0.80 \cdot 0.15) + (0.20 \cdot 0.85)} \\
&= 0.41
\end{aligned}
$$

Most people answered 80%

Arthritis study *(Redelmeier & Tversky '96)*

- Common belief:
  - **Arthritis pain is associated with changes in weather**
- Experiment:
  - Followed 18 arthritis patients for 15 months
  - 2 x per month assessed: (1) pain and joint tenderness, and (2) weather
- Results:
  - No correlation between pain/tenderness and weather
  - Patients saw correlation that did not exist... why?

Arthritis study (cont)

- Patients noticed when bad weather and pain co-occurred, but failed to notice when they didn't.
  - Better memory for times that bad weather and pain co-occurred.
  - Worse memory for times when bad weather and pain did not co-occur
- **Confirmation bias**: People often seek information that **confirms** rather than disconfirms their original hypothesis

Extra Examples

Estimating probabilities *(Tversky & Kahneman '73/'74)*

- *Question*: Is the letter **R** more likely to be the 1st or 3rd letter in English words?

- *Results*: Most said **R** more probable as 1st letter

- *Reality*: **R** appears much more often as the 3rd letter, but it's easier to think of words where **R** is the 1st letter

Estimating probabilities (cont)

- *Question*: Which causes more deaths in developed countries?
  (a) traffic accidents or (b) stomach cancer
- *Typical guess*: traffic accident = 4X stomach cancer
- *Actual*: 45,000 traffic, 95,000 stomach cancer deaths in US
- Ratio of newspaper reports on each subject:
  137 (traffic fatality) to 1 (stomach cancer death)

- **Availability heuristic**: Tendency for people to make judgments of frequency on basis of how easily examples come to mind

Gambler's fallacy

- *Gambler's fallacy*: belief that if deviations from expected behavior are observed in repeated independent trials, then future deviations in the opposite direction are then more likely
- T&K: this is an example of the **representativeness heuristic**—where the probability of an event is judged by its similarity to the population from which sample is drawn
- The sequence "H T H T T H" is seen as more representative of a prototypical coin sequence. Why?
    - When people are asked to make up random sequences, they tend to make the proportion of H and T closer to 50% than would be expected by random chance
    - T&K interpretation: people believe that short sequences should be representative of longer ones

Interpretation of these findings

- People do not use proper statistical/probabilistic reasoning... instead people use heuristics which can **bias** decisions
- Heuristics can often be very effective (and efficient) for social inferences and decision-making
    - E.g., the book “Simple Heuristics That Make Us Smart” summarizes research by Gigerenzer and Todd
- ... but be aware that **heuristics can bias** results from **exploratory data analysis and other modeling efforts**