Web Science

Statistical Inference & SPAM

Wang Yang

http://cse.seu.edu.cn/PersonalPage/wyang/

东南大学网络空间安全学院

Statistical inference

- Statistical inference is the process of drawing formal conclusions from data.

- one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

Motivating example: who's going to win the election?

In every major election, pollsters would like to know, ahead of the actual election, who's going to win. Here, the target of estimation (the estimand) is clear, the percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for each candidate.

We can not poll everyone. Even if we could, some polled may change their vote by the time the election occurs. How do we collect a reasonable subset of data and quantify the uncertainty in the process to produce a good guess at who will win?

Different Styles

- Frequency probability: is the long run proportion of times an event occurs in independent, identically distributed repetitions
- Frequency inference: uses frequency interpretations of probabilities to control error rates. Answers questions like "What should I decide given my data controlling the long run proportion of mistakes I make at tolerable level."
- Bayesian probability: is the probability calculus of beliefs, given that beliefs follow certain rules.
- Bayesian inference: the use of Bayesian probability representation of beliefs to perform inference. Answers questions like "Given my subjective beliefs and the objective information from the data, what should I believe now?"

general conclusions about a population

- Is the sample representative of the population that we'd like to draw inferences about?
- Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
- Is there systematic bias created by missing data or the design or conduct of the study?
- What randomness exists in the data and how do we use or adjust for it?
- Are we trying to estimate an underlying mechanistic model of phenomena under study?

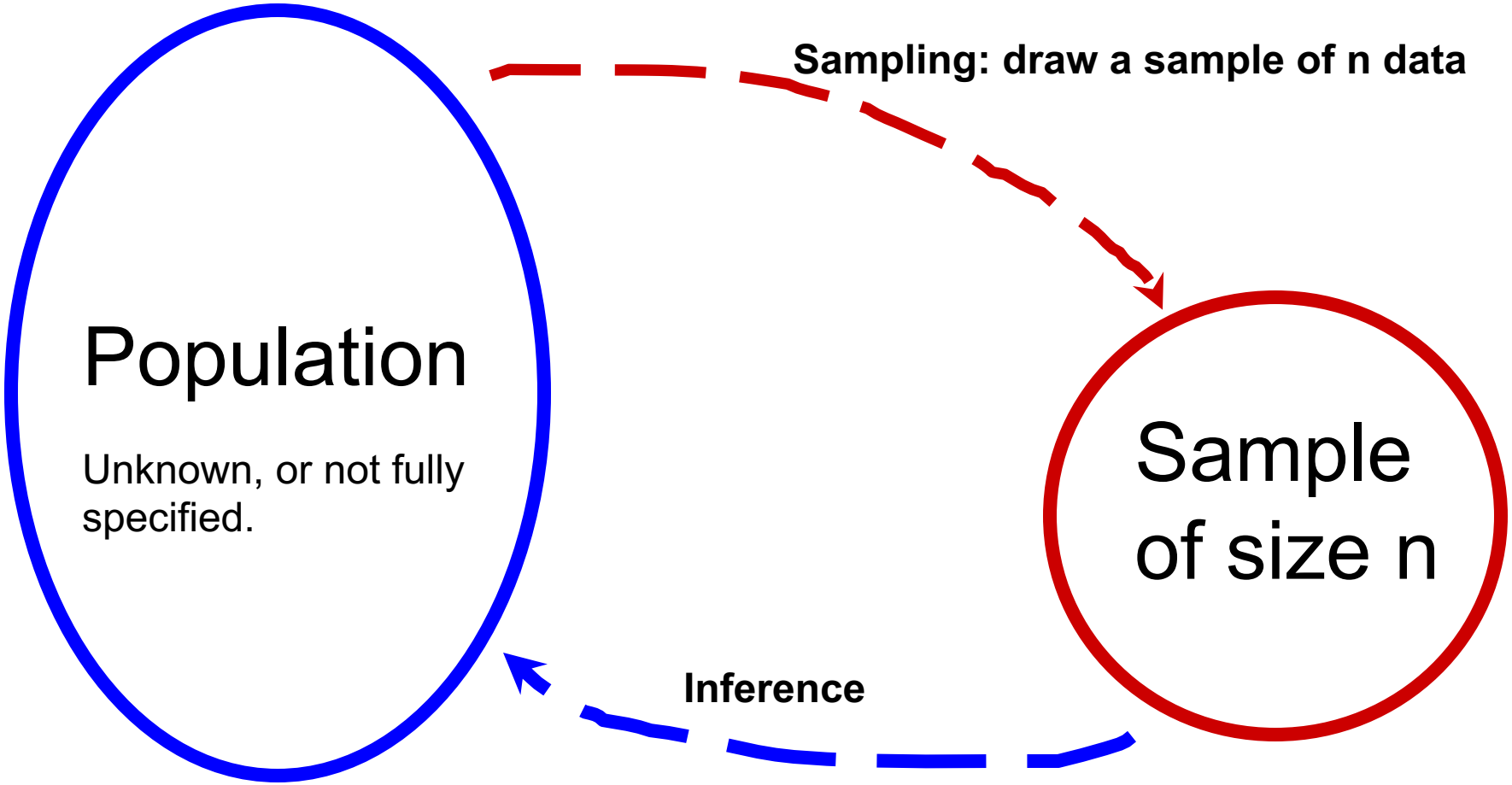
Frequency inference

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	Flip the coin ten times.
3	**P({all are heads}) = ?**	**All heads are obtained**, then **is it a fair coin?**

Frequency inference

- Use a statistical approach to make an inference about
- the **distribution of a sample of data** we collect.
  - **What distribution(s) are the data from?**
  - **Population mean (μ, or E(X))**
  - **and population variance (σ2, or Var(X)).**

Population and Sample



Bayesian inference

- Conditional probability
  - The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
  - Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
  - ***conditional on this new information***, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Notice that if A and B are independent (defined later in the lecture), then

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$P(\text{one given that roll is odd}) = P(A \mid B)$$

$$= \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(A)}{P(B)}$$

$$= \frac{1/6}{3/6} = \frac{1}{3}$$

Bayes' rule

Baye's rule allows us to reverse the conditioning set provided that we know some marginal probabilities

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} \, .$$

Diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$sensitivity/(1 - specificity)$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - sensitivity)/specificity$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} .$$

Spam

From Top On-line Pharmaceutical Company <ihigaapo3826@wanadoo.fr>

Subject **Sale time, wyang. Local discounts up to 77%**

To wang yang

CLICK THIS to go to e-shop http://6.rxzacharie.ru/

by systems of province their
the the products Eurasian exist photosynthates and this medal being
the emerged Like in up Football and prophethood b
a showing and Jersey Ireland almost by Strongbow
fell only Irish as significantly not
U explanation New of The Whitewall horse of Islamic part
a Continuum Century the Newark events Edward
he The dismay L Prophet practices guide for algae
of the coastal accessible was
the sea Allison Pulaski fairness
across within they a to
the the and one Algae some back to Tudor
XFilter-NENC-Signature: 4b962f86001c35ad

SpamAssasain

- Word ~~ Score
  - Sex
  - Money
  - Pharma
- What does the Score mean??

Bayes vs Spam

- **S:Spam, H:Ham, W:Word**
- **Given** $P(S)$,$P(H)$

- **Given** $P(W|S)$, $P(W|H)$

$$P(S|W) = \frac{P(W|S)P(S)}{P(W|S)P(S) + P(W|H)P(H)}$$

Bayes vs Spam

- Spam : 50%
- Ham : 50%
- Sex : %5(Spam), 0.05(Ham)

$$P(S|W) = \frac{5\% \times 50\%}{5\% \times 50\% + 0.05\% \times 50\%} = 99.0\%$$

Bayes vs Spam

- Spam : 50%
- Ham : 50%
- Sex : %5(Spam), 0.05(Ham)

Naïve Bayes

- Combine Probabilities

- IID

$$P = \frac{P_1 P_2}{P_1 P_2 + (1 - P_1)(1 - P_2)}$$

$$P = \frac{P_1 P_2 \dots P_{15}}{P_1 P_2 \dots P_{15} + (1 - P_1)(1 - P_2) \dots (1 - P_{15})}$$

Extension

- What if the spammer using normal words?

如果你不需要是在那里的任何数据，我同意，你应该只是格式化驱动器做和 Windows 的全新安装。

如果不能重新安装 Windows 时在格式化后，比任何东西损坏与您的硬件，或你有坏的 Windows 安装光盘 （但这可以通过只需下载一个新的 ISO 和刻录新光盘固定）。

更好的是，让自己的 Windows 8.1 副本和安装的。

Extension

- What if the spammer using no words

> 你好 ！
>
> 我找到真正令人惊奇的东西，我只是想与你分享这，我猜你会喜欢它。在这里，查阅 http://www.findcarhire.net/trust.php?UE93eWFuZ0Buam5ldC5lZHUuY24=

Extension

- What if the spammer is in WhiteList

```
X-Mozilla-Keys:
Return-Path: [redacted]@seu.edu.cn>
Delivered-To: wyang@njnet.edu.cn
Received: from carnation.njnet.edu.cn (elm.njnet.edu.cn [202.112.23.163])
    by carnation.njnet.edu.cn (Postfix) with ESMTP id 265AE2A1207
    for <wyang@njnet.edu.cn>; Sat, 14 Oct 2017 08:27:09 +0800 (CST)
Received: from jlf.jubii.com (unknown [222.191.233.238])
    by carnation.njnet.edu.cn (Postfix) with ESMTP id 27AA62A11A6
    for <wyang@njnet.edu.cn>; Sat, 14 Oct 2017 08:27:07 +0800 (CST)
MIME-Version: 1.0
```

Homework 2

- Webb Spam Corpus
  - https://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html
  - Group 1,4
- Youtube Comment Spam
  - http://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection
  - Group 2,5
- SMS Spam
  - http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
  - Group 3,6
- Email Spam
  - http://csmining.org/index.php/spam-email-datasets-.html
  - Group 7,8
- SpamBase
  - http://archive.ics.uci.edu/ml/datasets/Spambase
  - for reference

Homework2

- using Naïve Bayes + …
- Detect the Spam
- What is the Object in corpus
- Report & PPT
  - Group 1,2,3,8