NMAI059 Probability and statistics 1
Class 14

Robert Šámal

Simpson's paradox

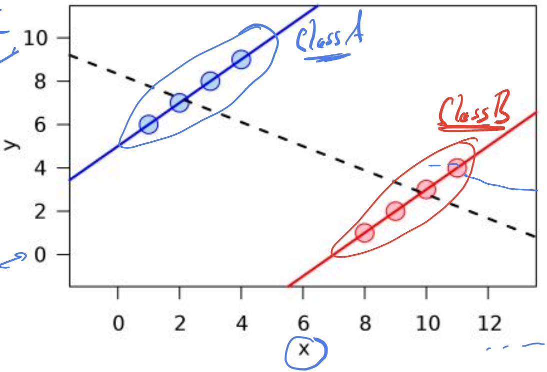
classif problem -- 0/1 var.

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

most often for B

most often for A

result best



lowest

regression -- real valued r.v.

one person spent 10.5 h & got 4 pts

time spent preparing for exam

Overview

Permutation test

Bootstrap

Bayesian statistics

Sampling random variables

Situation

- We have two collections of pairwise independent r.v.'s (random samples): - - - - two-sample test
- $X_1, \dots, X_n \sim F_X$ a $Y_1, \dots, Y_m \sim F_Y$
- We want to decide between $H_0 : F_X = F_Y$ and $H_1 : F_X \neq F_Y$.
- Examples: running time of an algorithm before/after modification, cholesterol level in people who do/don't eat Miraculous Superfood™, frequency of short words in text by authors X and Y.
- We do not assume anything about F_X, F_Y (in particular they may not be normal).

Method

$n=2, m=1 \quad X_1 = 1, X_2 = 9, Y_1 = 3$

$t_{obs} = |5-3| = 2$

- We choose an appropriate statistics, e.g.

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = |\bar{X}_n - \bar{Y}_m|$$

$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$, $\bar{Y}_m = \frac{Y_1 + \ldots + Y_m}{m}$

if t_{obs} is big -- reject H_0

resampling

- $t_{\text{obs}} := T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$
- Assuming H_0, „all permutations of the data are the same": X_i i Y_j were generated from the same distribution.
- We randomly permute the given $m + n$ numbers and for each permutation we calculate T – we get numbers $T_1, T_2, \ldots, T_{(m+n)!}$ (each equally likely).

	T
193	2
139	7
319	7
391	5
913	2
931	5

- As p-value we take the probability that $T > t_{\text{obs}}$, or

$$\frac{4}{6} = p = \frac{1}{(m+n)!} \sum_j I(T_j > t_{\text{obs}}). = \frac{\#\{j : T_j > t_{obs}\}}{(m+n)!}$$

- This is the probability of Type I error. We reject H_0 whenever $p < \alpha$ (for our choice of α, e.g. $\alpha = 0.05$).

Improvement

of $(X_1, X_2, \ldots, Y_m)$

- Enumerating all permutations can be too expensive. Instead, we take just an appropriate number B of independently generated permutations and calculate just B values $T_1, \ldots, T_B$.
- As p-value we take the estimate of the probability that $T > t_{\text{obs}}$, or

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

- For sufficiently large m, n this gives similar results as tests based on CLT. So it is useful especially for medium sized samples.

Overview

Permutation test

Bootstrap

resampling

Bayesian statistics

Sampling random variables

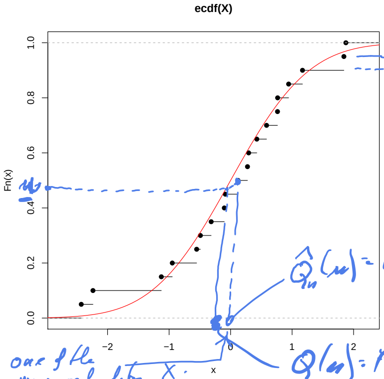
Empirical CDF – a reminder

- $X_1, \dots, X_n \sim F$ i.i.d., F is their CDF
- **Definition:** *Empirical CDF* is defined by

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

where $I(X_i \leq x) = 1$ if $X_i \leq x$ and 0 otherwise.

$U \sim \mathcal{U}(0,1)$



$\widehat{Q}_n(u) = \widehat{F}_n^{-1}(u)$

$Q(u) = F^{-1}(u)$

one of the measured data X_i

We hope $F \doteq \widehat{F}_n$

we use $\widehat{F}_n$ to sample new data

To sample from $\widehat{F}_n$, we choose uniformly at random one of already measured data.

Boostrap – basic idea

- from the measured data $X_1 = x_1, \ldots, X_n = x_n \sim F$ we create $\widehat{F}_n$
- other data can be sampled from $\widehat{F}_n$
- to do this we select a uniformly random $i \in \{1, \ldots, n\}$ and outputt~~ing~~ x_i

Bootstrap – basic usage

perhaps coeff of lin. regression

- $T_n = g(X_1, \ldots, X_n)$ some statistics (function of the data)
- we want to estimate $var(T_n)$
- sample $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ (see last slide)
- calculate $T_n^* = g(X_1^*, \ldots, g_n^*)$
- repeat B times to get $T_{n,1}^*, \ldots, T_{n,B}^*$
- the variance estimate:

$$\frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{k=1}^{B} T_{n,k}^*\right)^2$$

$\approx$ sample variance

$var(T_n)$ $\overset{?}{\approx}$ $var(T_n^*)$

when gen. from F

when sampled from $\hat{F}_n$

Overview

Two approaches to statistics

$\frac{1026}{6000} \rightsquigarrow \frac{1}{6}$

Frequentists'/classical approach

- Probability is a long-term frequency (out of 6000 rolls of the dice, a six was rolled 1026 times). It is an objective property of the real world.
- Parameters are fixed, unknown constants. We can't make meaningful probabilistic statements about them.
- We design statistical procedures to have desirable long-run properties. E.g. 95 % of our interval estimates will cover the unknown parameter.

Bayesian approach

- Probability describes how much we believe in a phenomenon, how much we are willing to bet. (Prob. that T. Bayes had a cup of tea on December 18, 1760 is 90 %.) (Prob. that COVID-19 virus did leak from a lab is ?50? %.)
- We can make probabilistic statements about parameters (even though they are fixed constants).
- We compute the distribution of ϑ and form point and interval estimates from it, etc.

Bayesian method – basic description

- The unknown parameter is treated as a random variable Θ
- We choose *prior distribution*, the pmf $p_\Theta(\vartheta)$ or the pdf $f_\Theta(\vartheta)$ independent of the data.

or $P_{X|\Theta}(x|\vartheta)$

- We choose a statistical model $f_{X|\Theta}(x|\vartheta)$ that describes what we measure (and with what probability), depending on the value of the parameter.
- After we observe $X = x$, we compute the *posterior distribution* $f_{\Theta|X}(\vartheta|x)$ or $P_{\Theta|X}(\vartheta|x)$ using Bayes' thm
- and then derive what we need e.g. find a, b so that $P(\Theta\in[a,b] \mid X=x) = \int_a^b f_{\Theta|X}(\vartheta|x)d\vartheta \geq 1-\alpha$

inference) estimation

Hypoth. test: if $P(\Theta=0 \mid X=x) < \alpha$ we reject H_0 $(H_0: \Theta = 0)$

- $\vartheta = \theta$ lower-case theta, Θ is upper-case theta

Bayes theorem

$\Omega = B_1 \cup \dots \cup B_n$ partition, $A \subseteq \Omega$: $P(B_i|A) = \dfrac{P(B_i)\cdot P(A|B_i)}{\sum_j P(B_j)\cdot P(A|B_j)}$

Theorem (Bayes theorem for discrete r.v.'s)

X, Θ are discrete r.v.'s

$P(\Theta = \vartheta \mid X = x)$ "$P(B_i|A)$" if $\vartheta = \vartheta_i$

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_\Theta(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_\Theta(\vartheta')}.$$

(terms with $p_\Theta(\vartheta') = 0$ are considered to be 0).

$A =$ "$X = x$"
if Θ takes values $\vartheta_1, \dots, \vartheta_n$,
$B_i =$ "$\Theta = \vartheta_i$"

Theorem (Bayes theorem for continuous r.v.'s)

X, Θ are continuous r.v.'s with pdf's f_X, f_Θ and joint pdf $f_{X,\Theta}$

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_\Theta(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta')f_\Theta(\vartheta')d\vartheta'}.$$

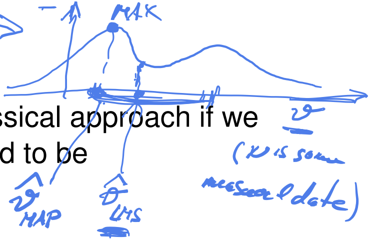
(terms with $f_\Theta(\vartheta') = 0$ with $f_\Theta(\vartheta') = 0$ are considered 0).

- Two more variants omitted.

Bayesian point estimates – MAP and LMS

MAP – Maximum A-Posteriori We choose $\hat{\vartheta}$ to maximize

- $p_{\Theta|X}(\vartheta|x)$ in the discrete case
- $f_{\Theta|X}(\vartheta|x)$ in the continuous case
- Similar to the ML method in the classical approach if we choose a „flat prior“ – Θ is supposed to be uniform/discrete uniform.



(x is some measured data)

LMS – Least Mean Square Also the conditional mean method.

- We choose $\underset{LMS}{\hat{\vartheta}} = \mathbb{E}(\Theta \mid X = x)$.
- Unbiased point estimate, takes the smallest possible ~~LMS~~ e ~~value~~.

MSE – mean sq. error

Example 1

caveat: Are X_1, X_2 -- independent?
IF NOT -- we need to know joint distr.

Bayesian spam classifier: $f(X_1, X_2, \ldots)$

- create a list of suspicious words (money, win, pharmacy, ...)
- R.v. X_i describes whether the email contains the suspicious word w_i.
- R.v. Θ describes whether the email is spam $\Theta = 1$ or not $\Theta = 0$.
- From the previous emails, we get estimates of $p_{X|\Theta}$ and p_Θ (prior)
- We use Bayes' theorem to calculate $p_{\Theta|X}$

$P_{X_1|\Theta}(1|1)$ = fract. of spams that contain w_1

$P_{X|\Theta}(1|0)$ = ... of nonspams ... ("money") money

$P_\Theta(1) = P(\Theta = 1)$ = fract. of spams

$P_\Theta(0)$: fract. of non-spams

Example 2

Romeo and Juliet are to meet at noon sharp. But Juliet is late by the time described by the random variable $X \sim U(0, \vartheta)$. We model the parameter ϑ by the random variable $\Theta \sim U(0, 1)$. What do we infer about ϑ from the measured value of $X = x$?

prior distr. $f_\Theta(\vartheta) = \begin{cases} 1 & \text{for } \vartheta \in [0,1] \\ 0 & \text{o.w.ise} \end{cases}$ (by def.)

$f_{X|\Theta}(x|\vartheta) = \begin{cases} \frac{1}{\vartheta} & \text{for } x \in [0, \vartheta] \\ 0 & \text{o.w.se} \end{cases}$ by def.

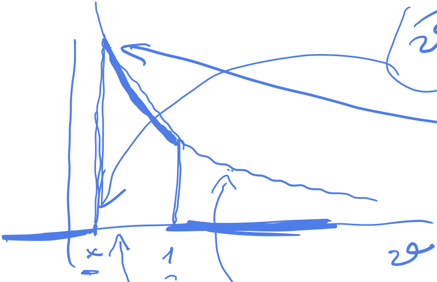
obvious: $\vartheta \geq x$, $\vartheta \in [0,1]$

so if $\vartheta < x$: $f_{\Theta|X}(\vartheta|x) = 0$

posterior distr.

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta) \cdot f_\Theta(\vartheta)}{\int_0^1 f_{X|\Theta}(x|\vartheta') f_\Theta(\vartheta') \, d\vartheta'} = \frac{1/\vartheta}{\int_x^1 \frac{1}{\vartheta'} \, d\vartheta'} = \frac{1/\vartheta}{[\log \vartheta']_x^1} = \frac{1}{\vartheta(-\log x)} = \frac{1}{\vartheta |\log x|} > 0$$

if $\vartheta > x$ & $\vartheta \in [0,1]$



$\hat{\vartheta}_{MAP} = x$ as is the maximum

$$f_{\Theta|X}(\vartheta|x) = \begin{cases} 0 & \text{if } \vartheta \notin [0,1] \text{ or if } \vartheta < x \\ \dfrac{1}{\vartheta|\lg x|} & \text{otherwise} \end{cases}$$

$$\hat{\vartheta}_{LMS} = E(\Theta|X=x) = \int_x^1 \vartheta \cdot \frac{1}{\vartheta|\lg x|}\, d\vartheta = \int_x^1 \frac{1}{|\lg x|}\, d\vartheta = \boxed{\frac{1-x}{|\lg x|}}$$

Example 3

Observing random variables $X = (X_1, \dots, X_n)$, assume $X_i \sim N(\vartheta, \sigma_i^2)$ and ϑ is the value of the random variable $\Theta \sim N(x_0, \sigma_0)$. What can we conclude about ϑ from the measured values $X = x = (x_1, \dots, x_n)$?

Example 4

We flip a coin, the probability of getting heads is ϑ. Out of n flips, the coin comes up heads in $X = k$ cases. If our a priori distribution was $U(0, 1)$, what would be the distribution of the posterior distribution?

Overview

Permutation test

Bootstrap

Bayesian statistics

Sampling random variables

Basic method – inverse transformation method

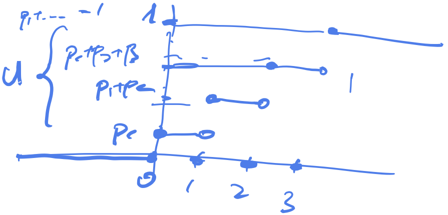
Theorem

Let F be a function "of CDF-type": nondecreasing right-continuous function with $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$.
Let Q be the corresponding quantile function.
Let $U \sim U(0,1)$ and $X = Q(U)$. Then X has CDF F.

- It works well if we can ~~quantify~~ evaluate Q, for example for exponential or geometric distributions.
- The gamma distribution is the sum of several exponential distributions – so we generate it that way.

Variant of the basic method for discrete variables

- We want a r.v. X that takes values $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots$. ($\sum_i p_i = 1$).
- We generate $U \sim U(0,1)$.
- Find i such that $p_1 + \cdots + p_{i-1} < U < p_1 + \cdots + p_i$.
- We set $X := x_i$.



Bin$(3, \frac{1}{2})$

evaluate $\Phi(U)$

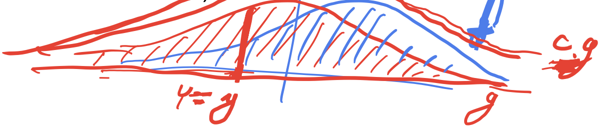
- Works nicely when we have a formula for $p_1 + \cdots + p_i$ (e.g. geometric distribution).
- The binomial distribution is better simulated as the sum of n independent Bernoulli variables.
- There are special tricks for other ones (Poisson).

Rejection sampling

- We want to generate a r.v. with density f.
- We can generate a r.v. with density g (which is „similar“), namely
- $\frac{f(y)}{g(y)} \le c$ for some constant c. $\forall y$
- The method:
  1. Generate Y with density g, and $U \sim U(0,1)$.
  2. If $U \le \frac{f(Y)}{cg(Y)}$, then $X := Y$.
  3. Otherwise, reject the value of Y, U and repeat from point 1.
- Rationale: generating a random value of X with density f is the same as generating a random point under the graph of the function f whose horizontal (x) coordinate is X (and whose vertical coordinate is uniformly random between 0 and X).

$U \cdot g(Y)$

$(Y, c \cdot U \cdot g(Y))$ point



check:

$cUg(y) \le f(y)$

Follow-up classes

- Probability and Statistics 2 – NMAI073
- Introduction to Approximation and Randomized Algorithms – NDMI084
- Introduction to Machine Learning in Python|R – NPFL129|NPFL054
- and many master-level lectures