Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#231 Statistical Inference

été 2020

Average

Consider observations $\{x_1, \cdots, x_n\}$ with be true mean μ, and standard deviation σ.

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} n\mu = \mu$$

$*$ since the expected value is linear

$\overline{x}$ is an unbiased estimator of the (true) mean μ

$$\text{Var}(\overline{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{*}{=} \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$*$ because variables are independent, and variance is a quadratic function

The variance of $\overline{x}$ is $\frac{\sigma^2}{n}$, and $\text{Var}(\overline{x}) \to 0$ as $n \to \infty$

Average

Problem: σ is unkown...

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - \overline{X}_n]^2.$$

$$\mathbb{E}(S_n^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}[X_i - \overline{X}_n]^2\right) \overset{*}{=} \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right]\right)$$

$*$ from the same property as before

$$\mathbb{E}(S_n^2) = \frac{1}{n-1}[n\mathbb{E}(X_i^2) - n\mathbb{E}(\overline{X}^2)] \overset{*}{=} \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]$$

$*$ since $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

An unbiased estimator of the variance of $\overline{x}$ is $\frac{s^2}{n}$

Average

Finally, from the central limit theorem (#431),

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

or

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{s^2}{n}\right)$$

so, a 95% confidence interval for μ is

$$\left[\overline{x} \pm 2\frac{s}{\sqrt{n}}\right]$$

Average

In a given city, a random sample of $n = 400$ people is taken. The average years of schooling of this sample is $\overline{x} = 11.6$ years, with a standard deviation of $\widehat{s} = 4.1$. Find a 95% confidence interval for the average educational level of people in this city.

Let μ be the true mean educational level, σ the standard deviation.

From the central limit theorem (#431), $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

95% confidence interval for μ is $\left[\overline{x} \pm 2\frac{\sigma}{\sqrt{n}}\right]$

Standard deviation of $\overline{X}$ can be estimated by $\widehat{s}/\sqrt{n} \simeq 0.2$,
Therefore, 95% confidence interval is $[11.6 \pm 0.4]$

Proportion

In a given city, a random sample of $n = 400$ people is taken. The number of people who intend to vote for a candidate is 212. Find a 95% confidence interval for the probability to vote for that candidate, in this city.

Let p be the true probability, and $\widehat{p}$ denote the (empirical) frequency, i.e. $\widehat{p} = \dfrac{212}{400}$

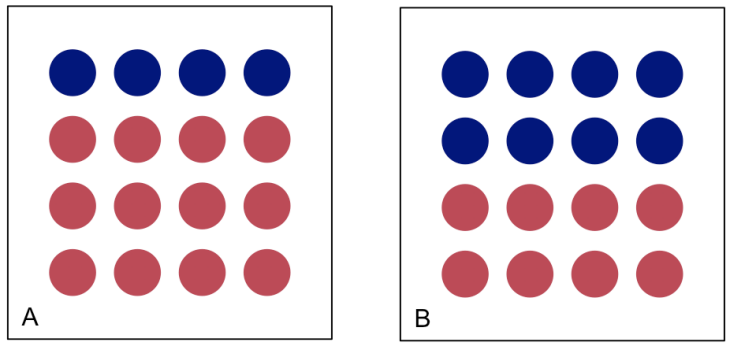
From the central limit theorem (#431), $\widehat{p} \sim \mathcal{N}\left(p, \dfrac{p(1-p)}{n}\right)$

95% confidence interval for p is $\left[\widehat{p} \pm 2\dfrac{p(1-p)}{\sqrt{n}}\right]$

Standard deviation of $\overline{X}$ can be estimated by $\widehat{s}/\sqrt{n} \simeq 0.2$,
Therefore, 95% confidence interval is $[11.6 \pm 0.4]$

Bayesian Approach

Consider two bags, A (4 blue balls, 12 red) and B (8 blue balls, 8 red).
We pick randomly a bag, and randomly a ball. The ball is red. What probability that bag A was selected ?



The prior estimate is $\mathbb{P}(A) = \mathbb{P}(B)$
Given that the ball is red, the posterior estimate is

$$\mathbb{P}(A|\bullet) = \frac{\mathbb{P}(A)\mathbb{P}(\bullet|A)}{\mathbb{P}(\bullet)} = \frac{\mathbb{P}(A)\mathbb{P}(\bullet|A)}{\mathbb{P}(\bullet|A)\mathbb{P}(A) + \mathbb{P}(\bullet|B)\mathbb{P}(B)}$$

so, numerically, $\mathbb{P}(A|\bullet) = \dfrac{\frac{1}{2} \cdot \frac{12}{16}}{\frac{12}{16} \cdot \frac{1}{2} + \frac{8}{16} \cdot \frac{1}{2}} = \dfrac{3}{5} = 60\%$

et $\mathbb{P}(B|\bullet) = 40\%$.

Bayesian Approach

Application to statistical inference:
Assume that x has a density $f(x|\theta)$.
Probability to observe sample $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ (independent) is

$$f(\boldsymbol{x}|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta) \quad \text{(called likelihoodd)}$$

Assume a prior distribution for θ, density π.
Its posterior distribution is

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta) \cdot f(\boldsymbol{x}|\theta)}{f(\boldsymbol{x})} = \frac{\pi(\theta) \cdot f(\boldsymbol{x}|\theta)}{\int f(\boldsymbol{x}|\tau)\pi(\tau)d\tau}$$

which is the general form of

$$\mathbb{P}(A|\bullet) = \frac{\mathbb{P}(A)\mathbb{P}(\bullet|A)}{\mathbb{P}(\bullet)} = \frac{\mathbb{P}(A)\mathbb{P}(\bullet|A)}{\mathbb{P}(\bullet|A)\mathbb{P}(A) + \mathbb{P}(\bullet|B)\mathbb{P}(B)}$$

Proportion

Over five years, no student got caught cheating in a course. Estimate the yearly probability to have a student cheating

$$\widehat{p} = \frac{0}{5} = 0.00\%$$

Give a 95% confidence interval

Let $X_i = \mathbf{1}(\text{cheated on year } i)$,

$$\mathbb{P}(X_i = x) = \begin{cases} \theta \text{ if } x = 1 \text{ cheated on year } i \\ 1 - \theta \text{ if } x = 0 \text{ did not cheat on year } i \end{cases}$$

then, the probability to have observed $(0, 1, 0)$ over 3 years

$$\mathbb{P}(X_1 = 0, X_2 = 1, X_3 = 0) = \underbrace{\mathbb{P}(X_1 = 0)}_{=(1-\theta)} \cdot \underbrace{\mathbb{P}(X_2 = 1)}_{=\theta} \cdot \underbrace{\mathbb{P}(X_3 = 0)}_{=(1-\theta)}$$

Proportion

Consider sample $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$.

$$\begin{cases} p(x_i|\theta) = \theta^{x_i}[1-\theta]^{1-x_i} \\ p(\boldsymbol{x}|\theta) = \theta^{\boldsymbol{x}^\mathsf{T}\mathbf{1}}[1-\theta]^{n-\boldsymbol{x}^\mathsf{T}\mathbf{1}} \end{cases}$$

Assume a prior uniform distribution for θ, $\pi(\theta) = 1$ on $[0,1]$, so that the posterior distribution is

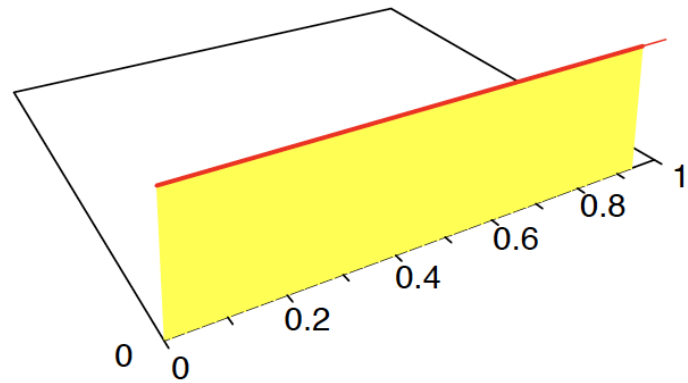
$$\pi(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{p(\boldsymbol{x})}$$

i.e.

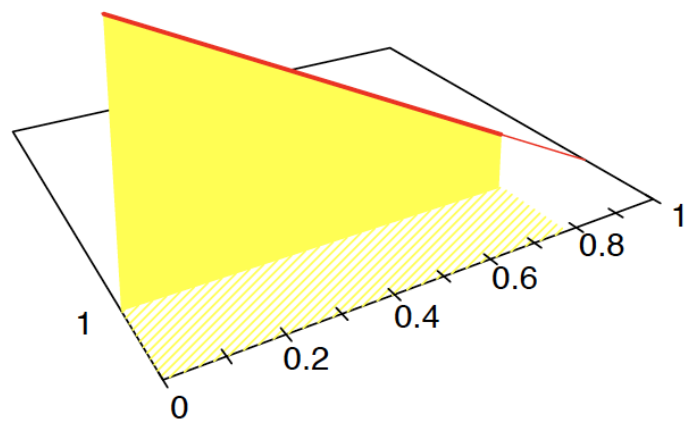
$$\pi(\theta|\boldsymbol{x}) \propto \theta^{\boldsymbol{x}^\mathsf{T}\mathbf{1}}[1-\theta]^{n-\boldsymbol{x}^\mathsf{T}\mathbf{1}}$$

(called a Beta distribution)

$\pi(\theta) = 1$ on $[0,1]$



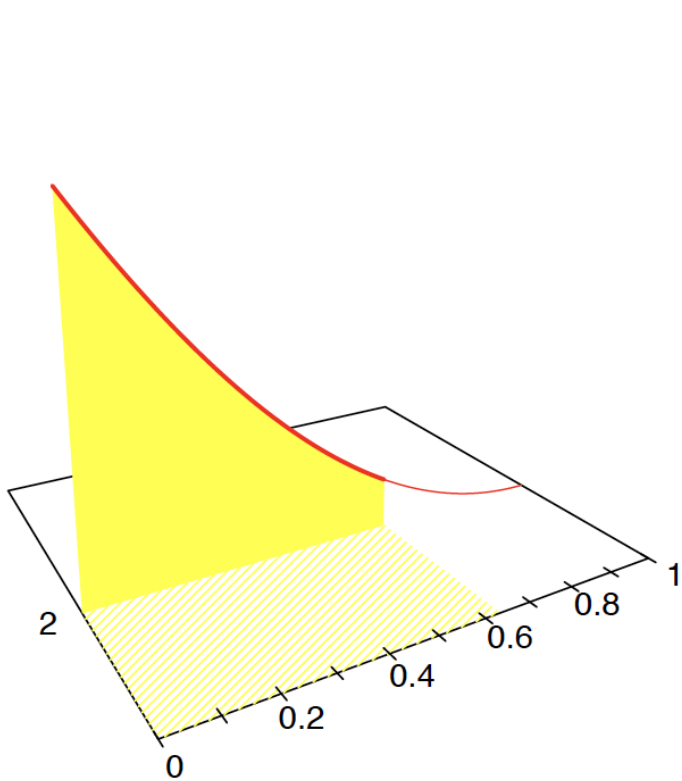
$\pi(\theta|x_1 = 0) = 2(1-\theta)$

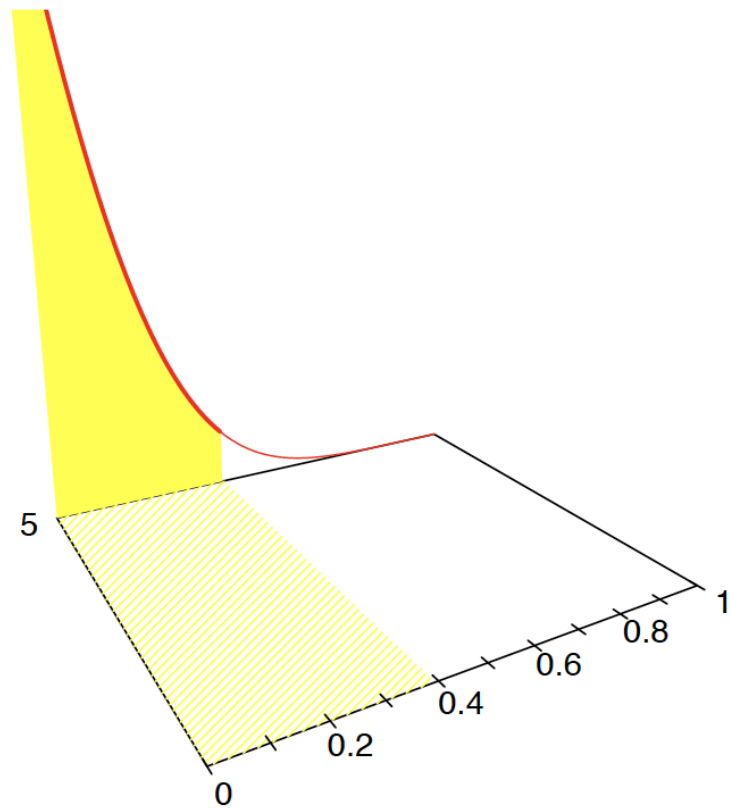


Proportion

$$\pi(\theta|\boldsymbol{x}) = \frac{\theta^{\alpha+\boldsymbol{x}^\mathsf{T}\mathbf{1}}[1-\theta]^{\beta+n-\boldsymbol{x}^\mathsf{T}\mathbf{1}}}{B(\alpha+\boldsymbol{x}^\mathsf{T}\mathbf{1}, \beta+n-\boldsymbol{x}^\mathsf{T}\mathbf{1})}$$

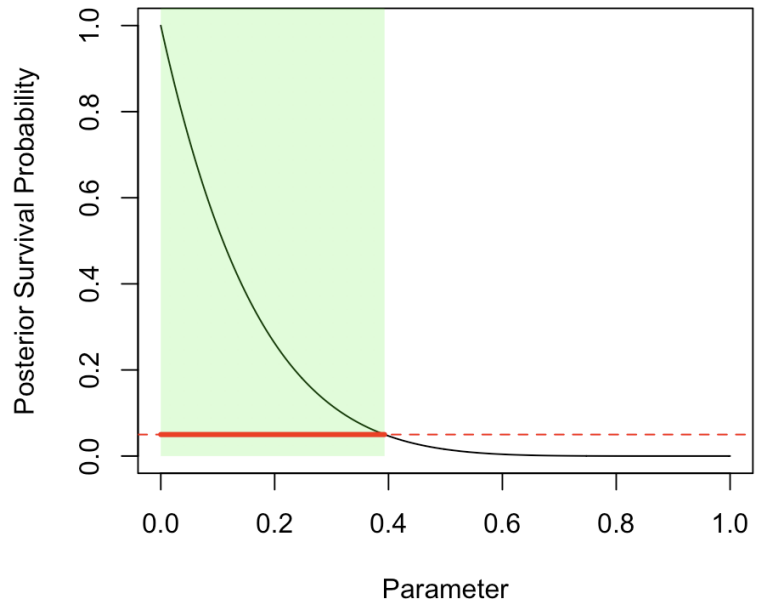
$\pi(\theta|(0,0,0)) = 4(1-\theta)^3$ $\pi(\theta|(0,0,0,0,0)) = 6(1-\theta)^5$





Proportion

we want $\mathbb{P}(\theta > u | \boldsymbol{x}) = 5\%$
or $\mathbb{P}(\theta \in [0, u] | \boldsymbol{x}) = 95\%$
i.e. $\int_0^u 6(1-\theta)^5 d\theta = 95\%$
(i.e. quantile of Beta distribution)
Numerically, we have




```
> import scipy.integrate as integrate
> integrate.quad(lambda x: 6*(1-x)**5,0,.393)
(0.949981329213977 2, 1.0546911446573683e-14)
```

i.e. $\mathbb{P}(\theta \in [0, 0.393] | \boldsymbol{x}) \sim 95\%$

Estimating a proportion?

RANDOMIZED RESPONSE: A SURVEY TECHNIQUE FOR ELIMINATING EVASIVE ANSWER BIAS

STANLEY L. WARNER
Claremont Graduate School

For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. In this paper it is argued that such bias is potentially removable through allowing the interviewee to maintain privacy through the device of randomizing his response. A randomized response method for estimating a population proportion is presented as an example. Unbiased maximum likelihood estimates are obtained and their mean square errors are compared with the mean square errors of conventional estimates under various assumptions about the underlying population.

Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias

Estimating a proportion?

Let θ denote the probability to have ever cheated on your husband/wife
Consider the following strategy : flip a coin, do not show me

- head (prob 50%) : answer *have you ever seen Titanic* ?
- tail (prob 50%) : answer *have you ever seen cheated on your wife* ?

$$\mathbb{P}[\text{answer yes}] = \underbrace{\mathbb{P}[\text{answer yes}|\text{head}]}_{=85\%} \cdot \underbrace{\mathbb{P}(\text{head})}_{=50\%} + \underbrace{\mathbb{P}[\text{answer yes}|\text{tail}]}_{=\theta} \cdot \underbrace{\mathbb{P}(\text{tail})}_{=50\%}$$

(see http://fivethirtyeight.com for the Titanic probability)

More Formally... Accuracy & Precision - Mean & Variance

Consider data $\{x_1, \cdots, x_n\}$ with (unknown) distribution F_θ
Let $\widehat{\theta}$ denote an estimator of (unknown) θ
$\widehat{\theta}$ is unbiased if $\mathbb{E}(\widehat{\theta}) = \theta$
Its precision is given by $\text{Var}(\widehat{\theta})$