Lecture 3 Jan 23, 2025

- Concentration of random variables

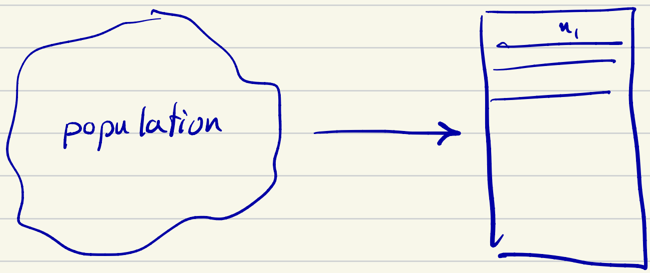
  (Markov, chebyshev, chernoff, Hoeffding)

- Running example: estimating coin bias.

Hypothesis testing (property testing of distributions)

We used randomness to model the world.

data points are random samples from an unknown data distribution



data points

distribution p $\qquad$ $x_1, \dots, x_m$

$$x_i \sim p$$

Estimating coin bias

$p := \Pr[\text{head}]$

Testing a coin is fair:

- if $p = \frac{1}{2}$, output accept w. prob. $1-\delta$.
- if $|p - \frac{1}{2}| > \varepsilon$, output reject w. prob $1-\delta$.

Algorithm

Flip a coin $m = ?$ times

$X \leftarrow$ # heads

if $\left|\frac{X}{m} - \frac{1}{2}\right| \leq ?$

return accept

else

return reject

Question : How well $\frac{X}{m}$ approximate p?

what should be m?

boils down $\rightarrow$ How well $\frac{X}{m}$ concentrate around p?

Concentration of random variables.

Questions: $\left\{ \begin{array}{l} \text{Estimating average height of students} \\ \\ \text{exit polls} \end{array} \right.$

n samples:

$$X_1, X_2, \ldots, X_n \sim P$$

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \longrightarrow \mu := \mathbb{E}_{X \sim P}[X]$$

$\boxed{\text{Goal}}$ measure how much $\overline{X}_n$ deviates from μ

Law of Large numbers

(weak) $\forall \varepsilon$ $\lim_{n \to \infty} \Pr[|\overline{X}_n - \mu| < \varepsilon] = 1$

(strong) $\Pr[\lim_{n \to \infty} \overline{X}_n = \mu] = 1$

Central Limit Theorem:

$$\mathrm{Var}_{X\sim P}[X]$$

$$\sqrt{n}\,(\bar{X}_n - \mu) \rightarrow \mathcal{N}(0, \sigma^2)$$

$$Z \sim \mathcal{N}(0,1)$$

$$\Pr\left[\frac{\sqrt{n}\,|\bar{X}_n - \mu|}{\sigma} > u\right] \approx \Pr[\,|Z| > u\,]$$

$$= 2\Phi(-u)$$

where Φ is the cdf of the standard normal dist.

Look up table

$$u = 1.96 \rightarrow 2\Phi(-u) \simeq 95\%$$

Hence: with prob. 0.95

$$\mu \in \left[\,\bar{X}_n - 1.96\,\sigma/\sqrt{n},\ \bar{X}_n + 1.96\,\sigma/\sqrt{n}\,\right]$$

- Quality of Approximation varies depending on P.

These are asymptotic results. Very general, but

- work in the limit,
- Do not indicate the relationship among the parameters.

n, d, ε, δ ?

d → dimension
ε → error
δ → confidence (in our example δ was $1 - 0.95 = 0.05$)

What about finite sample setting?

Usefull tools to show concentration (tail bounds)

Markov's inequality:

For non-negative random variable X, and $a>0$:

$$\boxed{\Pr[X \geq a] \leq \frac{E[X]}{a}}$$

proof.

pdf

$$E[X] = \int_0^\infty x \Pr[X=x]\, dx$$

$$= \int_0^a x \Pr[X=x]\, dx + \int_a^\infty x \Pr[X=x]\, dx$$

$$\geq 0 + \int_a^\infty a \Pr[X=x]\, dx$$

$$\geq a \cdot \Pr[X \geq a]$$

$$\Rightarrow \Pr[X \geq a] \leq \frac{E[X]}{a} \qquad \square$$

$\longrightarrow$ back to coin example

works well for small p

if $p \leq 0.01$

$$\Pr\left[\frac{X}{m} > 0.1\right] \leq \frac{E[X]}{0.1} \leq 0.1$$

not very meaningful when $p = \frac{1}{2}$

Chebyshev's inequality

For a random variable with finite mean and variance, and $k > 0$:

$$\Pr[\,|X - \mathbb{E}[X]| \geq k\sigma\,] \leq \frac{1}{k^2}$$

standard deviation of X

proof:

$$\Pr[\,|X - \mathbb{E}[X]| \geq k\,\sigma\,]$$

$$= \Pr[\,(X - \mathbb{E}[X])^2 \geq k^2\sigma^2\,]$$

$$\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \quad \square$$

Markov

$\rightarrow$ back to coin example

$$E\left[\frac{X}{m}\right] = p$$

$$Var\left[\frac{X}{m}\right] = \frac{p(1-p)}{m}$$

$$Pr\left[\left|\frac{X}{m} - p\right| > \varepsilon\right] \leq \frac{Var[X/m]}{\varepsilon^2} \leq \frac{1}{m\varepsilon^2} \leq \delta$$

$$m = \frac{1}{\delta \cdot \varepsilon^2}$$

right dependencies to ε

but not δ

Chernoff bound:

m Bernoulli random variable: $X_1, X_2, \dots X_m$

$X_i \sim Ber(p_i)$ $\quad X_i = \begin{cases} 1 & \text{with prob } p_i \\ 0 & \text{" " } 1-p_i \end{cases}$

empirical mean $\quad X := \frac{1}{m} \sum_{i=1}^{m} X_i$

and true mean $\quad \mu := \frac{1}{m} \sum_{i=1}^{m} p_i$

$$\boxed{\begin{aligned} &\Pr[X - p > \varepsilon p] \leq e^{-mp\varepsilon^2/3} \\ &\Pr[p - X > \varepsilon p] \leq e^{-mp\varepsilon^2/2} \end{aligned}}$$

general structure of the proof:

(can be applied to any random variable)

For all $\varepsilon > 0$, $t > 0$:

$$\Pr[X > \varepsilon] = \Pr[e^{tX} > e^{t\varepsilon}]$$

$$\leq \frac{\mathbb{E}[e^{tX}]}{e^{t\varepsilon}} = e^{-t\varepsilon} M_X(t)$$

Markov

moment generating func

since the bound holds for any t, we can conclude:

$$\Pr[X \geq \varepsilon] \leq \inf_{t > 0} e^{-t\varepsilon} M_X(t) \quad \square$$

→ back to coin example

$$\Pr[\,|X - p| \geq \varepsilon\,] \leq 2\exp(-mp\varepsilon^2)$$

$$\boxed{m = \frac{1}{p\varepsilon^2} \log 2/\delta} \qquad \leq \delta$$

works when $p = 1/2$. ✓

(or when p is a constant)

Hoeffding bound :

$$Pr[X - \mu > \varepsilon] < e^{-2m\varepsilon^2}$$

$$Pr[\mu - X < \varepsilon] < e^{-2m\varepsilon^2}$$

→ back to coin example

$$m = \frac{\log(2/\delta)}{2\varepsilon^2} \Rightarrow$$

$$Pr[|X - \mu| > \varepsilon] \leq \delta$$