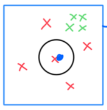
5- estimators and inference

estimation = estimates parameter values of specific population
hypothesis testing = tests wheter those parameters are equal to specified values

estimation

↳ $P_X(x;\theta) \rightarrow$ estimation of parameters θ, if we have observed values $x_1, x_2 \ldots$ of X
estimator = function that estimates parameters given some observations $\hat{\theta}(x_1, x_2 \ldots x_n)$
estimate = result of the estimation (ex: μ and σ in normal, λ in poisson, w and b in regression)

bias of an estimator = $B(\hat{\theta}) = E[\hat{\theta}] - \theta$ (avg of different estimates differ from the true parameters)
↳ unbiased estimator = the estimate is equal to the true value within the population
↳ when $B(\hat{\theta}) = 0$ (ex: $\bar{x} = \mu$ or $\bar{p} = p$)



→ for mean in the middle, red distribution avg is unbiased but does not work well, high variance, but even though the avg of green distribution is biased it is preferred because it has low variance
mean squared error = used to measure the quality of estimators
↳ $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = B(\hat{\theta}) + Var(\hat{\theta})$

parametric models = distributions that are defined by a finite set of parameters
↳ if parameters are chosen right, generalizes well, else fails
non-parametric models = arises directly from the data, no model selection (ex: histogram)
↳ number of parameters is bounded by number of observations
↳ may overfit by memorizing the data then cannot generalize

parameter estimation methods

★ given some data, type of distribution is assummed beforehand, which determines what the unknown parameters will be estimated

maximum likelihood = finding parameters which maximizes the likelihood of the observed data (↗ frequentist)
↳ joint probability of observing $x_1, x_2, \ldots$ given that the parameter is θ (maximize it)
↳ likelihood function = $\mathcal{L}(\theta \mid x_1, x_2 \ldots) = p(x_1, x_2, \ldots \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) \xrightarrow{\log} \sum_{i=1}^{N} \log p(x_i \mid \theta)$
↳ max likelihood estimate = $\hat{\theta}_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta \mid x_1, x_2 \ldots)$
↳ it will produce a set of exact parameter estimates

$p(x \mid D) = p(x \mid \hat{\theta})$
↳ observations

bayesian estimation = estimation of the distributions of parameters
↳ prior (→ must be given) parameter distribution is updated (sharpening the prior) using the observed data
↳ posterior = $p(\theta \mid x) = \dfrac{P(x \mid \theta) \cdot P(\theta)}{P(x)} = \dfrac{P(x \mid \theta) \cdot P(\theta)}{\int P(x \mid \theta) \cdot P(\theta) \cdot d\theta}$

$p(x \mid D) = \int p(\theta \mid D) \cdot p(x \mid \theta) \cdot d\theta$ (→ parameters)

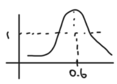
↳ considers both the prior and likelihood to find the distributions of the parameters
↳ generalization of the max likelihood estimate

example = tossing a coin outcomes $\{T, T, H, T, H\}$, in bernoulli process $p('tails') = ?$
↳ maximum likelihood = $P(x) = p^x \cdot (1-p)^{1-x} \rightarrow x = 1$ if it is tail

$\log(\mathcal{L}(p \mid x)) = \sum_{i=1}^{n} \log(p^{x_i} (1-p)^{x_i}) = t \log p + h \log(1-p)$ (→ # of heads)

↳ $\frac{d}{dp} = \frac{t}{p} - \frac{h}{1-p} = 0 \rightarrow \hat{p}(t,h) = \frac{t}{t+h}$ $\quad p(3,2) = 0.6$

↳ bayesian = $p(\theta \mid x) = \dfrac{\theta^t (1-\theta)^h}{\int_0^1 (\theta^t (1-\theta)^h) \cdot d\theta} \xrightarrow{t=3 \; h=2} 60\,\theta^3 (1-\theta)^2$



map $(\theta) = 0.6$
$E[\theta \mid x] = 1$