5. Inferential statistics

- model-based exploration of data, i.e., use of a prob. model
- goal: learn general statements from samples

5.1 Estimators

Example: (Bernoulli distribution)

- determine the failure prob. p of hard-disk drive (charge)
- X counts the number of read/write access until first failure
- $\mathbb{P}[X = k] = (1-p)^{k-1} p$, i.e., X is geometrically distrib.
- it holds that:

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \cdot (1-p)^{k-1} p$$

$$= p \left(- \sum_{k=1}^{\infty} (1-p)^{k} \right)'$$

$$= p \left(- \frac{1}{1-(1-p)} + 1 \right)'$$

$$= p \cdot \frac{1}{p^2} = \frac{1}{p}$$

- Thus, $p = \frac{1}{\mathbb{E}(X)}$

- suppose, we have n drives, i.e., n r.v. $X_1, \ldots, X_n$
- consider $\overline{X} =_{def} \frac{1}{n} \sum_{i=1}^{n} X_i$
- $\overline{X}$ is an estimator for $\mathbb{E}(X)$

Definition 1.

Let X be a random variable with density function $f(x, \vartheta)$. An estimator U for ϑ of X is a random variable composed of (independent, identically distributed) sampling variables $X_1, \dots, X_n$, i.e.,
$U = h(X_1, \dots, X_n)$.

Definition 2.

Let U, U' be estimators of parameter ϑ of X.

(1.) U is said to be unbiased if and only if $\mathbb{E}(U) = \vartheta$; otherwise, U is biased.

(2.) U is more efficient than U' if and only if $\mathbb{E}((U-\vartheta)^2) \leq \mathbb{E}((U'-\vartheta)^2)$.

(3.) U is consistent if and only if $\mathbb{E}((U-\vartheta)^2) \to 0$ when $n \to \infty$; n is the sample size.

Examples:

(1.) $\overline{X} =_{df} \frac{1}{n} \sum_{i=1}^{n} X_i$. We have

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

so, $\overline{X}$ is unbiased.

$$\mathbb{E}\left((\overline{X}-\vartheta)^2\right) \overset{\bar{X}\text{ unbiased}}{=} Var(\overline{X}) = Var\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n} \cdot Var(X)$$

So, $\overline{X}$ is consistent.

(2) $S^2 =_{def} \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$; $\mu =_{def} \mathbb{E}(X) = \mathbb{E}(X_i) = \mathbb{E}(\overline{X})$

$$\begin{aligned}(X_i - \overline{X})^2 &= (X_i - \mu + \mu - \overline{X})^2 \\ &= (X_i-\mu)^2 + 2(X_i-\mu)(\mu-\overline{X}) + (\mu-\overline{X})^2 \\ &= (X_i-\mu)^2 + (\mu-\overline{X})^2 + \\ &\qquad 2(X_i-\mu)\left(\mu - \frac{1}{n}\sum_{j=1}^{n} X_j\right) \\ &= (X_i-\mu)^2 + (\mu-\overline{X})^2 - \frac{2}{n}\sum_{j=1}^{n}(X_i-\mu)(X_j-\mu) \\ &= \frac{n-2}{n}(X_i-\mu)^2 + (\mu-\overline{X})^2 - \frac{2}{n}\sum_{j\neq i}(X_i-\mu)(X_j-\mu)\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left((X_i-\overline{X})^2\right) \\ &= \frac{n-2}{n}\underbrace{\mathbb{E}\left((X_i-\mu)^2\right)}_{=\operatorname{Var}(X)} + \overbrace{\mathbb{E}\left((\mu-\overline{X})^2\right)}^{=\operatorname{Var}(\overline{X}) = \frac{1}{n}\operatorname{Var}(X)} \\ &\qquad - \frac{2}{n}\sum_{j\neq i}\underbrace{\mathbb{E}\left((X_i-\mu)(X_j-\mu)\right)}_{=(\mathbb{E}(X_i)-\mu)(\mathbb{E}(X_j)-\mu) = 0} \\ &= \frac{n-1}{n}\operatorname{Var}(X)\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left((X_i-\overline{X})^2\right) \\ &= \frac{1}{n-1}\cdot n \cdot \frac{n-1}{n}\cdot\operatorname{Var}(X) = \operatorname{Var}(X)\end{aligned}$$

So, S^2 is unbiased, σ^2 is biased.

How to construct estimators?

Maximum-Likelihood estimators (MLE):

- $X_1, \dots, X_n$ is a sampling (independent, identically distr.)
- form a random vector $\vec{X} = (X_1, \dots, X_n)$
- X is distributed acc. to $f(x;\vartheta) = \mathbb{P}_\vartheta[X = x]$, i.e., X discrete
- for $\vec{x} \in \mathbb{R}^n$, define

$$L(\vec{x}, \vartheta) =_{def} \prod_{i=1}^{n} f(x_i;\vartheta) = \prod_{i=1}^{n} \mathbb{P}_\vartheta[X_i = x_i]$$

$$= \mathbb{P}_\vartheta[X_1 = x_1, \dots, X_n = x_n]$$

- L is called likelihood-function
- find ϑ that maximizes $L(\vec{x}, \vartheta)$,

Definition 3.

An estimator $\hat{\vartheta}$ for the parameter in $f(x, \vartheta)$ of X is called Maximum-Likelihood-Estimator (MLE) of a sampling $\vec{x}$ if and only if for all ϑ,

$$L(\vec{x}; \vartheta) \leq L(\vec{x}; \hat{\vartheta}) \quad \text{for all } \vartheta$$

Example: ① (MLE for Bernoulli distribution)

- estimator for p s.t. $\mathbb{P}_p[X_i = 1] = p$, $\mathbb{P}_p[X_i = 0] = 1-p$.
- for $\vec{x} = (x_1, \dots, x_n)$, $\mathbb{P}_p[X_i = x_i] = p^{x_i}(1-p)^{1-x_i}$
- likelihood function:

$$L(\vec{x}, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- find p maximizing L, or find p maximizing $\ln L$:

$$\ell(\vec{x}, p) =_{def} \ln L(\vec{x}, p)$$

$$= \sum_{i=1}^{n} \left(x_i \ln p + (1-x_i) \ln(1-p)\right)$$

$$= n\bar{x} \cdot \ln p + (n - n\bar{x}) \ln(1-p)$$

$$l'(\vec{x}, p) = \frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1-p} \overset{!}{=} 0,$$

so: $0 = n\bar{x}(1-p) - (n - n\bar{x})p$

$= n\bar{x} - n\bar{x}p - np + n\bar{x}p$

$= n\bar{x} - np$

Hence, $p = \bar{x}$.

(2) MLE for normal distribution:

- $X \sim N(\mu, \sigma^2)$
- find estimator for μ, σ^2
- we obtain:

$$L(\vec{x}; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$l(\vec{x}; \mu, \sigma^2) = \ln L(\vec{x}; \mu, \sigma^2)$$

$$= -n\left(\ln\sqrt{2\pi} + \ln\sigma\right) + \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

- find μ, σ maximizing $l(\vec{x}, \mu, \sigma^2)$:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \overset{!}{=} 0$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} \overset{!}{=} 0$$

- thus; $\mu = \bar{x}$, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

- MLE biased, in general

5.2 Confidence intervals

- want to learn how good the estimation is
- use two estimators U_1, U_2 s.t. $\mathbb{P}[U_1 \leq \vartheta \leq U_2] \geq 1-\alpha$.
- prob. $1-\alpha$ is called confidence level
- $[U_1, U_2]$ is called confidence interval; probability that $\vartheta \notin [U_1, U_2]$ is at most α.
- typical design: symmetrical confidence interval $[U-\delta, U+\delta]$ for $\delta > 0$, i.e., $U_1 = U-\delta$, $U_2 = U+\delta$.

Example: (normal distribution)

- $X \sim N(\mu, \sigma^2)$, $X_1, \dots, X_n$ sampling variables,
- $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- consider $Z =_{def} \sqrt{n}\, \frac{\overline{X}-\mu}{\sigma}$; so, $Z \sim N(0,1)$.
- symmetrical confidence interval $[-c, c]$ for $c > 0$ s.t.

$$1-\alpha \overset{!}{=} \mathbb{P}[-c \leq Z \leq c]$$

$$= \mathbb{P}\left[\overline{X} - \frac{c\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{c\sigma}{\sqrt{n}}\right]$$

- find c s.t. $1-\alpha = \mathbb{P}[-c \leq Z \leq c]$:

$$\mathbb{P}[-c \leq Z \leq c] = \Phi(c) - \Phi(-c)$$

$$= \Phi(c) - (1-\Phi(c))$$

$$= 2\Phi(c) - 1 \leq$$

$$\overset{!}{=} 1-\alpha$$

- Thus, $\Phi(c) = 1 - \frac{\alpha}{2}$, or, $c = \Phi^{-1}(1-\frac{\alpha}{2})$

Definition 4.

Let X be a random variable with distribution F_X. Then, $x_\gamma \in \mathbb{R}$ s.t. $F_X(x_\gamma) = \gamma$ is called γ-quantile of X.

Example: For normal distribution, γ-quantiles are denoted by z_γ. So, confidence interval for estimator for μ of $X \sim N(\mu, \sigma^2)$:

$$K = \left[\bar{X} - \frac{z_{(1-\frac{\alpha}{2})}\,\sigma}{\sqrt{n}} \,,\ \bar{X} + \frac{z_{(1-\frac{\alpha}{2})}\,\sigma}{\sqrt{n}} \right]$$

5.3 Hypothesis testing

- want to learn if certain statements are true given a sample
- e.g., is it rue that $p < \frac{1}{3}$ for a Bernoulli distributed var.

1. Definition of a test

- n random var. $X_1, \dots, X_n$, independent, identically distr.
- sample is a vector $\vec{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$
- critical set $K \subseteq \mathbb{R}^n$ for a hypothesis H:
  $K =_{def} \{ \vec{x} \mid H \text{ is rejected given } \vec{x} \}$
- test variable $T = h(X_1, \dots, X_n)$
- $\mathbb{R}$ is decomposed into sets (intervals), associated with rejection or non-rejection; $\tilde{K} \subseteq \mathbb{R}$ is associated with rejection
- $K = T^{-1}(\tilde{K}) \subseteq \mathbb{R}^n$

- H_0 is called null hypothesis (which is to be examined)
- H_1 is called alternative hypo. ($\neg H_1 \to \neg H_0$), often: $H_1 = \neg H_0$
- e.g., $H_0: p \geq \frac{1}{3}$, $H_1: p < \frac{1}{3}$ (or $H_0': p \geq \frac{1}{3}$, $H_1': p < \frac{1}{6}$)

2. Failures

- with certain prob. α tests can fail; α is called significance level; typical values for α: 0,05, 0.01, 0.001
- type-1 error: H_0 is true but is rejected
- type-2 error: H_0 is false but is failed to reject
- eg., for $K = \emptyset$: type-1 error $= 0$ (H_0 never rejected);
  type-2 error $= 1$
- given α, construct T such that type-1 error is α

Example: Test for p of Bernoulli distribution of X

- hypotheses:

$$H_0: \underset{(H_0 = [p_0, 1])}{p \geq p_0}, \qquad H_1: \underset{(H_1 = [0, p_0))}{p < p_0}$$

- $T =_{def} X_1 + \dots + X_n$, $X_1, \dots, X_n$ i.i.d.
- $\tilde{K} =_{def} [0, q]$ for $q \in \mathbb{R}$ (one-sided test)
- T is binomially distributed, $T \sim B(n, p)$, i.e.,

$$\mathbb{P}_p[T = k] = \binom{n}{k} p^k (1-p)^{n-k}; \quad \mathbb{E}(T) = np, \; Var(T) = np(1-p)$$

- de Moivre: $\lim_{n \to \infty} \frac{T - np}{\sqrt{np(1-p)}} \sim N(0,1)$
- consider $\tilde{T} =_{def} \frac{T - np}{\sqrt{np(1-p)}}$ (approx. standard norm. distr.)
- determine significance level α for q:

type-1 error: $\max_{p' \in H_0} \mathbb{P}_{p'}[T \in K] = \max_{p' \in H_0} \mathbb{P}_{p'}[T \leq q]$

type-2 error: $\max_{p' \in H_1} \mathbb{P}_{p'}[T \notin K] = \max_{p' \in H_1} \mathbb{P}_{p'}[T > q]$

We obtain:

$$\alpha = \max_{p' \in H_0} \mathbb{P}_{p'}[T \leq q] = \mathbb{P}_{p_0}\left[\tilde{T} \leq \frac{q - np_0}{\sqrt{np_0(1-p_0)}}\right]$$

$$\approx \Phi\left(\frac{q - np_0}{\sqrt{np_0(1-p_0)}}; 0, 1\right)$$

thus, find q s.t. $\frac{q - np}{\sqrt{np_0(1-p_0)}} = z_\alpha$

So, $q = z_\alpha \sqrt{np_0(1-p_0)} + np$

Standard tests:

1. (Approximate) binomial test

- Assumptions: $X_1, \dots, X_n$ i.i.d. s.t. $\mathbb{P}[X_i = 1] = p$, $\mathbb{P}[X_i = 0] = 1-p$; p not known, n large enough

- Hypotheses:

  (a) $H_0: p = p_0$, $H_1: p \neq p_0$

  (b) $H_0: p \geq p_0$, $H_1: p < p_0$

  (c) $H_0: p \leq p_0$, $H_1: p > p_0$

- Test variable:

$$Z =_{\text{def}} \frac{h - np_0}{\sqrt{np_0(1-p_0)}},$$

  where $h =_{\text{def}} X_1 + \dots + X_n$; h is absolute frequency of $X_i = 1$.

- Rejection criterion of H_0 at level α:

  (a) $|Z| > z_{1-\alpha/2}$

  (b) $Z < z_\alpha$

  (c) $Z > z_{1-\alpha}$

2. Gaussian test:

- Assumption: $X_1, \dots, X_n$ i.i.d., $X_i \sim N(\mu, \sigma^2)$, σ^2 known.
- Hypotheses:

  (a) $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

  (b) $H_0: \mu \geq \mu_0$, $H_1: \mu < \mu_0$

  (c) $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$

- Test variable:

$$Z =_{\text{def}} \frac{\overline{X} - \mu_0}{\sigma} \sqrt{n}$$

- Criteria of H_0 at level α:

  (a) $|Z| > z_{1-\alpha/2}$

  (b) $Z < z_\alpha$

  (c) $Z > z_{1-\alpha}$

(3) χ^2 - test:

A: $X_1, \dots, X_n$ i.i.d., $R_{X_i} = \{1, \dots, k\}$

H: $H_0: \mathbb{P}[X = i] = p_i$ for all $i \in \{1, \dots, k\}$

$H_1: \mathbb{P}[X = i] \neq p_i$ for some $i \in \{1, \dots, k\}$

T:
$$T =_{df} \sum_{i=1}^{k} \frac{(h_i - np_i)^2}{np_i},$$

where h_i: abs. frequency of $X_j = i$

C: at level α:
$$T > \chi^2_{k-1, 1-\alpha},$$

where $\chi^2_{k-1,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2_{k-1}-distribution; density function is

$$f(x) = \frac{1}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-1}{2}-1} e^{-\frac{x}{2}}$$