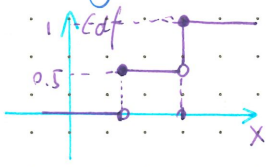
Moments of Gaussian

Order	Non-central	Central moment
1	μ	0
2	$\mu^2 + \sigma^2$	σ^2
3	$\mu^3 + 3\mu\sigma^2$	0
4	$\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$	$3\sigma^4$
5	$\mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4$	0
6	$\mu^6 + 15\mu^4\sigma^2 + 45\mu^2\sigma^4 + 15\sigma^6$	$15\sigma^6$
7	$\mu^7 + 21\mu^5\sigma^2 + 105\mu^3\sigma^4 + 105\mu\sigma^6$	0
8	$\mu^8 + 28\mu^6\sigma^2 + 210\mu^4\sigma^4 + 420\mu^2\sigma^6 + 105\sigma^8$	$105\sigma^8$

- **Pivotal quantity** — A function of observations, that the function's probility distribution does not dependent on unknown parameter. and unobservable parameters
- **Uniform distribution** — Maximum likelihood estimator for parameter θ^* of $\tilde{X} \sim Unif[0, \theta^*]$.

  MLE — $\max_{i=1,\dots,n} X_i$
- **Empiricel measure** — A random measure arising from a particular realization of a (usually finite) sequence of r.v.
- **Valid confidence interval** should not depend on parameter that is unknown (to be estimated.)
- $\equiv$ — congruence

- $f_X(x)$, $F_X(x)$ — pdf and cdf of X (continuous)
- $p_X(x)$ — pmf of X (discrete)
- $\triangleq$ — equal by definition
- $\rightsquigarrow$ — leads to
- $\operatorname{argmax} f(x)$ $\operatorname{argmin} f(x)$ — arguments of maxima / minima (values at which $f(x)$ is max/min)
- affine function — both convex and concave
- A priori — knowledge independent from experience
  A posteriori — knowledge that depends on empirical evidence
- Well defined — (unambiguous) an expression which assign it a unique (value) interpretation
- Improper integral — the limit of a definite integral as an endpoint of the interval(s) of integration approaches either a specified real number, ∞, $-\infty$, or in some instances as both endpoints approach limits.
- w.r.t. — with regards to
- Right continuity of CDF



Example: $X = \begin{cases} 1 & 50\% \\ 2 & 50\% \end{cases}$

$P(X \le 1) = P(X \le 1+\delta) = 0.5 \quad \delta \to 0$

$\neq P(X \le 1-\delta) = 0$

- Consistent v.s. Unbiased Estimators

Consistent:

Estimator T_n of parameter θ is consistent if it converges in Probability to true value of parameter: $T_n \xrightarrow{P} \theta$

Unbiased:

Estimator's expected value and true value is 0. $E[T_n] = 0$

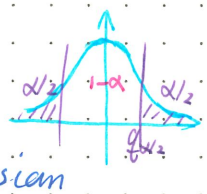
- Intergration by parts — $\int uv'dx = u\int v\,dx - \int u'(\int v\,dx)\,dx$

$$\int u v'\,dx$$

$$u\int v'\,dx - \int u'(\int v\,dx)\,dx$$

- $\perp\!\!\!\perp$ — Independent symbol
- Uncorrelated — meaning covariance is 0, does not gaurentee independence
  A and B uncorrelated only imply independence if jointly gaussian

- **Asymptotic** — Refers to how an estimator behave as sample size grow larger (i.e. tends to infinity)
- **Well specified model** (Not to confuse with well defined) — The class of distribution C you are assuming for your modeling contains the unknown probability distributio p from where the sample is drawn.
- **iff** — If and only if
- **s.t.** — such that
- **CMT** — Continuous Mapping Theorem. Continuous functions preseve limi even if their arguments are sequences of random variables.
  Continuous function (Heine's definition) is such a function that maps convergent sequences into convergent sequences: if $x_n \to x$ then $g(x_n) \to g(x)$
  CMT: this will also be true if we replace deterministic sequence $\{X_n\}$ with a sequence of r.v. $\{X_n\}$ and repla "$\to$" with one of the types of convergence of r.v.
- **R.V.** — Random variable
- **:=** — Defined as
- **$q_{\alpha/2}$** — $P(X \le q_{\alpha/2}) = 1 - \alpha/2$



- **Φ** — cdf of standard gaussian
- **Asymptotic normal** — $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2_{\theta_0})$
- **Sufficient statistics** — No other statistic that can be calculated from the same sample provides any additional information as to the value of paramet
- **$f: x \mapsto x^2$** — A function that takes value x to value y
- **$\sup_{k>0} T^{(k)}(N)$** — Supremum, the largest value $T^{(k)}(N)$ could get to as k varies
- **supremum (sup) infimum (inf)** — Sup of subset S of partially ordered set T is the least element greater or equal to all element of S
  Inf of subset S of partially ordered set T is the greatest element in T that is less than or equal to all elements of S

Moment Generating Functions (MGFs)

- Moment:
  - $E(X)$: 1st moment
  - $E(X^2)$: 2nd moment
  - $E(X^k)$: k^{th} moment
- central moment : $E[(X - E(X))^k]$
- Definition: MGF for a random variable is:

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_{\text{all } k} e^{tk} p_X(k) & \text{(discrete)} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx & \text{(continuous)} \end{cases}$$

*Provided the expectation exist for some t in a neighborhood of 0

$$E(X^n) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

- Proof:

$$\begin{cases} e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ e^{tx} = 1 + tx + \frac{t^2}{2!}x^2 + \frac{t^3}{3!}x^3 + \cdots \end{cases}$$

Therefore: $E(e^{tx}) = 1 + tE(x) + \frac{t^2}{2!}E(x^2) + \cdots$

Therefore: $\frac{d}{dt}E(e^{tx}) = 0 + E(x) + tE(x^2) + \cdots$

$$\frac{d}{dt}E(e^{tx})\Big|_{t=0} = 0 + E(x) + 0 + \cdots$$

and $\frac{d^2}{dt^2}E(e^{tx})\Big|_{t=0} = 0 + 0 + E(x^2) + tE(x^3) + \cdots = 0 + 0 + E(x^2) + 0 \cdots$

- Application:
  - Bernoulli distribution:

ex $X \sim Ber(p)$ $\begin{cases} p_X(1) = p \\ p_X(0) = 1-p \end{cases}$ $Var(X) = ?$

$$M_X(t) = E(e^{tx}) = \sum_{\text{all } x} e^{tx} p_X(x) = e^{t\cdot 0} p_X(0) + e^t p_X(1) = 1 - p + pe^t$$

$$Var(X) = E(X^2) - [E(X)]^2$$

$$\begin{cases} E(X) = \frac{d}{dt} M_X(t)\Big|_{t=0} = pe^t\Big|_{t=0} = p \\ E(X^2) = \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} = pe^t\Big|_{t=0} = p \end{cases} \Rightarrow Var(X) = p - p^2$$

• Binomial distribution:

ex Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} Ber(p)$

$\Rightarrow \sum_{i=1}^n X_i \sim Bin(n,p)$. Find $E(\sum_{i=1}^n X_i)$, $Var(\sum_{i=1}^n X_i)$

$$M_X(t) = E(e^{tX}) = E(e^{t\sum_{i=1}^n X_i}) = E[e^{tX_1} \cdot e^{tX_2} \cdots e^{tX_n}]$$

$$\overset{\text{i.i.d.}}{=} [1-p+pe^t][1-p+pe^t]\cdots[1-p+pe^t]$$

$$= [1-p+pe^t]^n$$

$$E(\sum_{i=1}^n X_i) = \frac{d}{dt}(1-p+pe^t)^n\Big|_{t=0} = n(1-p+pe^t)^{n-1}\cdot pe^t\Big|_{t=0} = np$$

$$E[(\sum_{i=1}^n X_i)^2] = \frac{d^2}{dt^2}(1-p+pe^t)^n\Big|_{t=0} = n(n-1)(1-p+pe^t)^{n-2}pe^t \cdot pe^t + n(1-p+pe^t)^{n-1}\cdot pe^t\Big|_{t=0}$$

$$= n(n-1)p^2 + np = np^2 - np^2 + np$$

$$\Downarrow$$

$$Var(\sum_{i=1}^n X_i) = np - np^2 = np(1-p)$$

☆ If $X_1, \dots X_n$ are i.i.d. and X_i has MGF $M_{X_i}(t)$,

then $\sum_{i=1}^n X_i$ has MGF: $M_{\Sigma X_i}(t) = [M_{X_i}(t)]^n$

Law of Large Numbers (LLN) & Central Limit Theorem (CLT)

Assumptions: ① $E[|X_i|] < \infty$ for all i
② $X_1, \dots X_n$ independent and identically distributed

Law(s) (weak and strong) of large numbers:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{P,\ a.s.} \mu$$

$\xrightarrow{P}$: convergence in probability $\Rightarrow$ weak

$\xrightarrow{a.s.}$: convergence almost surely $\Rightarrow$ strong

Central limit theorem (CLT)

Assumptions: ① $E[|X_i|] < \infty$ for all i
② $\mathrm{Var}(X_i) < \infty$ for all i
③ $X_1, \dots X_n$ iid

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} N(0,1)$$

$\xrightarrow{(d)}$: convergence in distribution

Hoeffding's inequality

Given n ($n > 0$) i.i.d. random variables $X_1, X_2 \dots X_n \overset{i.i.d.}{\sim} X$ that are almost surely bounded ($P(X \notin [a,b]) = 0$)

$$P\left(|\bar{X}_n - E[X]| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) \quad \text{for all } \varepsilon > 0$$

* n does not need to be large.

Markov inequality

For a random variable $X \geq 0$ with mean $\mu > 0$, and any number $t > 0$

$$P(X \geq t) \leq \frac{\mu}{t}$$

* Restricted to non-negative r.v.

Chebyshev inequality

For a random variable X with (finite) mean μ and variance σ^2, and any number $t > 0$.

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Slutsky's Theorem

Generalization of r.v. to more complicated spaces

- Let X_n, Y_n be sequences of scalar/vector/matrix random elements
- If $\begin{cases} X_n \text{ converges in distribution to a random element } X \\ Y_n \text{ converges in probability to a constant } c \end{cases}$
- $X_n + Y_n \xrightarrow{d} X + c$

  $X_n Y_n \xrightarrow{d} Xc$

  $X_n / Y_n \xrightarrow{d} X/c$ (provided that c is invertible)

Continuous Mapping Theorem (also called Mann-Wald theorem)

- Continuous functions perserve limits even if their arguments are sequences of random varables.

  → Heine's definition: is a function that maps convergent sequences to convergent sequences: if $x_n \to x$ then $g(x_n) \to g(x)$

- Let $\{X_n\}$, X be random element defined on metric space S. Suppose a function $g: S \to S'$ (S' is another metric space) has the set of discontinuity points D_g such that $\Pr[X \in D_g] = 0$, Then:

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

$$X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$$

$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$$

Unit 2 Foundation of Inference

Lecture 3 Parametric Statistical Models

Lecture 4 Parametric Estimation & confidence intervals

Trinity of Statistical Inference

① Estimation

② Confidence Intervals

③ Hypothesis testing

Formal Definition of Statistical Models

Let observed outcome of a statistical experiment be a sample $X_1, \cdots X_n$ of n. i. i. d. random variables in some measurable space E (usually $E \subseteq \mathbb{R}$) and denote by $\mathbb{P}$ their common distribution. A statistical model associated to that statistical experiment is a pair

$(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$

- E: sample space
- $(\mathbb{P}_\theta)_{\theta \in \Theta}$: family of probability measures on E
- Θ: any set, called parameter set

* Well specified model: $\exists \theta$ such that $\mathbb{P} = \mathbb{P}_\theta$
* This particular θ is called unknown true parameter. The aim of statistical experiment is to θ, or check it's properties when they have special meaning.

Parametric / Non-parametric

- Parametric: $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$
- Nonparametric: Θ is of infinite dimensional
- Semiparametric: $\Theta = \Theta_1 \times \Theta_2$

Θ_1: finite ← estimate

Θ_2: infinite ← nuisance parameter

Identifiability of Models

The parameter θ is identifiable iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.,

$$\theta \neq \theta' \Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

or equivalently:

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

Parameter Estimation

Definitions:

- Statistic: Any measurable function of the sample → can compute with given data
  e.g. $\bar{X}_n$, $\max X_i$, $X_1 + \log(1 + |X_n|)$ ...
- Estimator: Any statistic whose expression does not depend on θ.
  - Consistent estimator:
    $\hat{\theta}_n \xrightarrow[n\to\infty]{\mathbb{P}/a.s.} \theta$ → weakly consistent / → strongly consistent
  - Asymtotically normal:
    $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$ → asymptotic variance of $\hat{\theta}_n$

Bias of estimator

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

* Unbiased: $\text{bias}(\hat{\theta}_n) = 0$

Quadric risk

$$R(\hat{\theta}_n) = \mathbb{E}\left[|\hat{\theta}_n - \theta|^2\right]$$

$$= \text{variance} + \text{bias}^2$$

Variance of estimator

An estimator is a random variable so we can compute its variance

$$\text{var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Confidence Interval

Let $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model based on observations $X_1, \dots X_n$ and assume $\Theta \subseteq \mathbb{R}$. Let $\alpha \in (0,1)$.

- Confidence Interval (C.I.) of level $1-\alpha$ for θ:
  Any random (depending on $X_1, \dots X_n$) interval I whose boundaries do not depend on θ and such that
  $$\mathbb{P}_\theta[I \ni \theta] \geq 1-\alpha, \quad \forall \theta \in \Theta$$
- Confidence Interval of asymptotic level $1-\alpha$ for θ:
  Any random interval I whose boundaries do not depend on θ and such that:
  $$\lim_{n\to\infty} \mathbb{P}[I \ni \theta] \geq 1-\alpha, \quad \forall \theta \in \Theta$$

* Jensen's Inequality

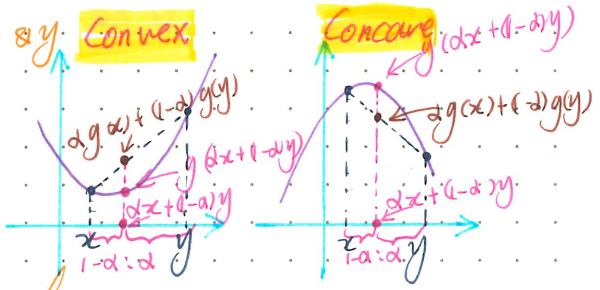
If X is a random variable and φ is a convex function then

$$\varphi(E[X]) \leq E[\varphi(X)]$$

* $E(\varphi(X)) - \varphi(E[X])$ = Jensen gap

Proof of Jensen's Inequality:

* Convex function: $g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha) g(y)$ (weighted avg. of x & y → weighted avg. of $g(x)$ and $g(y)$)
* Concave function: $g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha) g(y)$



⇓

Therefore, for a convex function $g: I \to \mathbb{R}$ and $x_1, x_2 \dots x_n$ in I and non-negative α_i s.t. $\sum_{i=1}^{n} \alpha_i = 1$,

we have $g(\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n) \leq \alpha_1 g(x_1) + \alpha_2 g(x_2) + \dots + \alpha_n g(x_n)$

⇓ $\alpha_i = P(X = x_i) = p_X(x_i)$

$$g(E[X]) \leq E[g(X)]$$

Confidence Interval Example

- $R_1, \dots R_n \overset{iid}{\sim} Ber(p)$ for some unknown $p \in (0,1)$
- Statistical model: $(\{0,1\}, (Ber(p))_{p\in(0,1)})$
- Estimator for p: $\hat{p} = \bar{R}_n$
- CLT:

$$\sqrt{n}\,\frac{\bar{R}_n - p}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{(d.)} \mathcal{N}(0,1)$$

  - cdf of $\sqrt{n}\,\frac{\bar{R}_n - p}{\sqrt{p(1-p)}} := \Phi_n(x)$
  - when n becomes large, $\Phi_n(x) \simeq \Phi(x)$ → standard gaussian cdf

  Therefore:

$$\mathbb{P}[|\bar{R}_n - p| \geq x] \simeq 2\left(1 - \Phi\frac{x\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

- For a fixed $\alpha \in (0,1)$, if $q_{\alpha/2}$ is $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$, then with probability $\simeq 1-\alpha$ (if n is large enough)

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}},\ p + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

$$\Downarrow$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}},\ \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right] \ni p\right) = 1-\alpha$$

* This is not a confidence interval, because it depends on p! → true parameter

Solution 1: Conservative bound

- $p(1-p) \leq \frac{1}{4}$
- $\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{\frac{1}{4}}}{\sqrt{n}},\ p + \frac{q_{\alpha/2}\sqrt{\frac{1}{4}}}{\sqrt{n}}\right]$
- $I_{conserv} = \left[\bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}},\ \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$

Solution 2: Solving the (quadratic) equation of p

- $\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \Rightarrow (p-\bar{R}_n)^2 \leq \frac{q_{\alpha/2}^2 (p\cdot(1-p))}{n}$
- Find roots $p_1 < p_2$ of

$$\left(1 + \frac{q_{\alpha/2}^2}{n}\right)p^2 - \left(2\bar{R}_n + \frac{q_{\alpha/2}^2}{n}\right)p + \bar{R}_n^2 = 0$$

- $I_{solve} = [p_1, p_2]$

- **Solution 3 : plug-in**
  - LLN : $\hat{p} = \bar{R}_n \xrightarrow[n\to\infty]{\mathbb{P}, a.s.} p$
  - Slutsky :
    $$\sqrt{n}\,\frac{\bar{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$
  - $I_{plug\text{-}in} = \left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}},\ \bar{R}_n + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$

Delta Method

- Let $(Z_n)_{n\geq 1}$ sequence of r.v. that satisfies
  $$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2)$$
  for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$ ($(Z_n)_{n\geq 1}$ asymptotically normal around θ)
- Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at point θ.
  Then;
  - $(g(Z_n))_{n\geq 1}$ is also asymptotically normal
  - $\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, (g'(\theta))^2\sigma^2)$
    asymptotical variance

Proof of Delta Method

- Let g be continuous differentiable everywhere in $\mathbb{R}$

  Let μ be arbitrary

  Mean value theorem (0^{th} order statement of Taylor's theorem)

  for $z > \mu$: $g(z) = g(\mu) + g'(c_z)(z-\mu)$ $\quad c_z \in (\mu, z)$ → is a function of z

  ⇓ works also for $z < \mu$

  $g(z) = g(\mu) + g'(c_z)(z-\mu)$ $\quad |c_z - \mu| < |z-\mu|$

  ⇓ implies

  for r.v. Z $\quad g(Z) - g(\mu) = g'(c_Z)(Z-\mu)$ $\quad |c_Z - \mu| < |Z-\mu|$

  ⇓ given an arbitrary sequence $(Z_n)_{n\geq 1}$ and for any μ

  $g(Z_n) - g(\mu) = g'(c_{Z_n})(Z_n - \mu)$

- Let $X_1, \dots X_n \overset{i.i.d.}{\sim} X$, $Z_n = \bar{X}_n$, $\mu = \mathbb{E}[X]$

  $\sqrt{n}\,(g(\bar{X}_n) - g(\mu)) = g'(c_{\bar{X}_n})(\sqrt{n}\,(\bar{X}_n - \mu))$ $\quad |c_{\bar{X}_n} - \mu| < |\bar{X}_n - \mu|$

  ⇓ CLT

  $(\sqrt{n}\,(\bar{X}_n - \mu)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2)$

  $P(|c_{\bar{X}_n} - \mu| > \varepsilon) \leq P(|\bar{X}_n - \mu| > \varepsilon)$

  ⇓ $\bar{X}_n \xrightarrow[n\to\infty]{P} \mu$

  $c_{\bar{X}_n} \xrightarrow[n\to\infty]{P} \mu$

  ⇓ g' is continuous

  continuous mapping theorem

  $g'(c_{\bar{X}_n}) \xrightarrow[n\to\infty]{P} g'(\mu)$

  Slutsky Theorem

  $\sqrt{n}\,(g(\bar{X}_n) - g(\mu)) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, (g'(\mu))^2\sigma^2)$

Hypothesis testing

Statistical formulation

- Consider a sample $X_1, \dots, X_n$ of i.i.d. random variables and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Let Θ_0 and Θ_1 be disjoint subset of Θ
- Consider the two hypothesis: $\begin{cases} H_0 : \theta \in \Theta_0 & \text{(null hypothesis)} \\ H_1 : \theta \in \Theta_1 & \text{(alternative hypothesis)} \end{cases}$

If we believe θ is either in Θ_0 or Θ_1, we may want to test H_0 against H_1

We want to decide whether to reject H_0 (look for evidence against H_0 in the data)

Asymmetry in the hypothesis

- H_0 and H_1 do not play a symmetric role: the data is only used to disprove H_0
- In particular, lack of evidence, does not mean H_0 is true (innocent until proven guilty)
- A test is a statistic $\psi \in \{0, 1\}$ such that:

$\begin{cases} \text{If } \psi = 0, & H_0 \text{ is not rejected} \\ \text{If } \psi = 1, & H_0 \text{ is rejected} \end{cases}$

Errors

- Rejection region of test ψ: $\rightarrow$ sample space $(X_1, \dots, X_n)$

$$R_\psi = \{x \in E^n : \psi(x) = 1\} \qquad \psi(x) = \mathbb{1}(x \in R_\psi)$$

- Type 1 error of test ψ (rejecting H_0 when it is actually true):

$$\alpha_\psi : \Theta_0 \to \mathbb{R} \text{ or } [0,1]$$
$$\theta \mapsto \mathbb{P}_\theta[\psi = 1]$$

- Type 2 error of test ψ (not rejecting H_0 although H_1 is actually true)

$$\beta_\psi : \Theta_1 \to \mathbb{R}$$
$$\theta \mapsto \mathbb{P}_\theta[\psi = 0]$$

- Power of test ψ:

$$\pi_\psi = \inf_{\theta \in \Theta_1} (1 - \beta_\psi(\theta))$$ (1 - largest possible $\beta_\psi(\theta)$)

Level, test statistic and rejection region

- A test ψ has level α if:

$$\alpha_\psi(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0$$

- A test has asymptotic level α if:

$$\lim_{n\to\infty} \alpha_{\psi_n}(\theta) \leq \alpha, \quad \forall \theta \in \Theta_0$$

- In general, a test has the form

$$\psi = \mathbb{1}\{T_n > c\}$$

(T_n: test statistic)

for some statistic T_n and threshold $c \in \mathbb{R}$

* Rejection region is $R_\psi = \{T_n > c\}$

One-sided v.s. two sided tests

When $\theta \in \Theta \subset \mathbb{R}$ and H_0 is of the form

$$H_0: \theta = \theta_0 \iff \Theta_0 = \{\theta_0\}$$

- If $H_1: \theta \neq \theta_0$: two-sided test
- If $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$: one sided test

p-value

Definition: The (asymptotic) p-value of a test ψ_α is the smallest (asymptotic) level α at which ψ_α rejects H_0.
It is random. It depends on the sample.

Golden rule: p-value $\leq \alpha \iff H_0$ is rejected by ψ_α, at the (asymptotic) level α.

* The smaller the p-value, the more confident one can reject H_0.

Unit 3: Methods of Estimation

Lecture 8: Distance measures between distributions

Total variation distance

- **Motivation:**
  - Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that there exist $\theta^* \in \Theta$ that $X_i \sim \mathbb{P}_{\theta^*}$: θ^* is the true parameter
  - Statistician's goal: given $X_1, \dots X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ s.t. $\mathbb{P}_{\hat{\theta}}$ is close to $\mathbb{P}_{\theta^*}$ for true parameter θ^*
  - This means: $|\mathbb{P}_{\hat{\theta}}(A) - \mathbb{P}_{\theta^*}(A)|$ is small for all $A \subset E$
- **Definition:**

  Total variation distance between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$:

  $$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \max_{A \subset E} |\mathbb{P}_\theta(A) - \mathbb{P}_{\theta'}(A)|$$

- **Discrete sample space:**

  $$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \sum_{x \in E} |p_\theta(x) - p_{\theta'}(x)|$$

- **Continuous sample space:**

  $$TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{1}{2} \int |f_\theta(x) - f_{\theta'}(x)| \, dx$$

- **Properties of total variation**
  - Symmetric: $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = TV(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$
  - Positive: $0 \leq TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq 1$
  - Definite: If $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$, then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$
  - Triangle inequality: $TV(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \leq TV(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) + TV(\mathbb{P}_{\theta''}, \mathbb{P}_{\theta'})$

  These imply that TV is a distance between probability distributions

Kullback-Leibler (KL) divergence

- **Definition:**

The KL divergence between two probability measures $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ is defined by

$$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \begin{cases} \sum_{x \in E} p_\theta(x) \log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) & \text{if } E \text{ is discrete} \\ \int_E f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx & \text{if } E \text{ is continuous} \end{cases}$$

- **Properties of KL divergence:**

~~Symmetric~~	$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \neq KL(\mathbb{P}_{\theta'}, \mathbb{P}_\theta)$ in general
Positive	$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \geq 0$
Definite	If $KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 0$, then $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$
~~Triangle inequality~~	$KL(\mathbb{P}_\theta, \mathbb{P}_{\theta''}) \not\leq KL(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) + KL(\mathbb{P}_{\theta'}, \mathbb{P}_{\theta''})$ in general

Not a distance

Lecture 9: Introduction to MLE

Maximum Likelihood Estimation

- **Derivation:**

$$KL(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \mathbb{E}_{\theta^*}\left[\log\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right] = \underbrace{\mathbb{E}_{\theta^*}[\log p_{\theta^*}(X)]}_{\text{constant}} - \mathbb{E}_{\theta^*}[\log p_\theta(X)]$$

$h(X) = \log p_\theta(X)$

LLN

$$\mathbb{E}_{\theta^*}[h(X)] \rightsquigarrow \frac{1}{n}\sum_{i=1}^n h(X_i)$$

Estimator of KL

$$\widehat{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = \text{"constant"} - \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

$$\min_{\theta \in \Theta} \widehat{KL}(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) \Leftrightarrow \min_{\theta \in \Theta} -\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

$$\Leftrightarrow \max_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$$

$$\Leftrightarrow \max_{\theta \in \Theta} \log\left[\prod_{i=1}^n p_\theta(X_i)\right]$$

$$\Leftrightarrow \max_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i)$$

Maximum Likelihood Principle

Likelihood : Discrete Case

- Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots X_n$. Assume that E is discrete (i.e., finite or countable)
- Definition:

The likelihood of the model is the map $\mathcal{L}_n$ (or just L) defined as:

$$\mathcal{L}_n : E^n \times \Theta \longrightarrow \mathbb{R}$$

$$(x_1, \dots x_n, \theta) \longmapsto \mathbb{P}_\theta [X_1 = x_1, \dots, X_n = x_n]$$

$$\prod_{i=1}^{n} \mathbb{P}_\theta [X_i = x_i]$$

Likelihood : Continuous Case

- Let $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots X_n$. Assume that all the $\mathbb{P}_\theta$ have density f_θ
- Definition:

The likelihood of the model is the map $\mathcal{L}$ defined as:

$$\mathcal{L} : E^n \times \Theta \longrightarrow \mathbb{R}$$

$$(x_1, \dots, x_n, \theta) \longmapsto \prod_{i=1}^{n} f_\theta(x_i)$$

Maximum likelihood estimator

- Let $X_1, \dots X_n$ be an i.i.d. sample associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ and let $\mathcal{L}$ be the corresponding likelihood.
- Definition:

The maximum likelihood estimator of θ is defined as:

$$\hat{\theta}_n^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(X_1, \dots, X_n, \theta), \text{ provided it exists}$$

* In practice, we use the fact that $\hat{\theta}_n^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log L(X_1, \dots, X_n, \theta)$

↓

log-likelihood estimator

Concave and Convex

- **Definition:**

A twice differentiable function $h: \Theta \subset \mathbb{R} \to \mathbb{R}$ is said to be concave if its second derivative satisfies:

$$h''(\theta) \leq 0; \quad \forall \theta \in \Theta$$

It is said to be strictly concave if the inequality is strict: $h''(\theta) < 0$

Moreover, h is said to be (strictly) convex if $-h$ is (strictly) concave.

- **Multivariate concave function**

For a multivariate function $h: \Theta \subset \mathbb{R}^d \to \mathbb{R}$, $d \geq 2$, define the

Gradient vector:

$$\nabla h(\theta) = \begin{pmatrix} \frac{\partial h}{\partial \theta_1}(\theta) \\ \frac{\partial h}{\partial \theta_2}(\theta) \\ \vdots \\ \frac{\partial h}{\partial \theta_d}(\theta) \end{pmatrix} \in \mathbb{R}^d$$

Hessian matrix:

$$H h(\theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial \theta_1 \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_1 \partial \theta_d} \\ \vdots & & \vdots \\ \frac{\partial^2 h}{\partial \theta_d \partial \theta_1}(\theta) & \cdots & \frac{\partial^2 h}{\partial \theta_d \partial \theta_d} \end{pmatrix} \in \mathbb{R}^{d \times d}$$

h is concave $\Leftrightarrow x^T H h(\theta) x \leq 0 \quad \forall x \in \mathbb{R}^d, \theta \in \Theta$

h is strictly concave $\Leftrightarrow x^T H h(\theta) x < 0 \quad \forall x \in \mathbb{R}^d, x \neq 0, \theta \in \Theta$

Optimality Conditions

Strictly concave functions are easy to maximize: if they have a maximum then it is unique. It is the unique solution to

$h'(\theta) = 0$ (single-variate) / $\nabla h(\theta) = 0 \in \mathbb{R}^d$ (multi-variate)

Convex optimization: close form solution

Lecture 10: Consistency of MLE, Covariance Matrices, and Multivariate Statistics

* Examples of maximum likelihood estimators

Bernoulli trials: $\hat{p}_n^{MLE} = \bar{X}_n$

Poisson model: $\hat{\lambda}_n^{MLE} = \bar{X}_n$

Gaussian model: $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, \hat{S}_n)$

$\hat{S}_n \downarrow$

$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Consistency of MLE

- Under mild regularity conditions, we have:

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$$

This is because for all $\theta \in \Theta$

$$\frac{1}{n}\log L(X_1, \dots, X_n, \theta) \xrightarrow[n\to\infty]{\mathbb{P}} \text{"constant"} - KL(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta)$$

$\downarrow$ minimizer is θ^* if parameter identifiable

Technical conditions allow to transfer this convergence to the minimizer/maximizer

Covariance

- In general, when $\theta \subset \mathbb{R}^d$, $d \geq 2$, its coordinates are not necessarily independent
- Definition:

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$= \mathbb{E}[X \cdot (Y - \mathbb{E}[Y])] \quad = \mathbb{E}[Y \cdot (X - \mathbb{E}[X])]$$

→ if one of X, Y is 0 mean, $Cov(X,Y) = \mathbb{E}[XY]$

- Properties:
  - $Cov(X,X) = Var(X)$
  - $Cov(X,Y) = Cov(Y,X)$
  - If X and Y are independent $\Rightarrow Cov(X,Y) = 0$

* In general $Cov(X,Y)=0$ does not guarentee X, Y independent.

Except if $(X,Y)^T$ is a Gaussian vector

$\Downarrow$

$\alpha X+\beta Y$ is gaussian for all $(\alpha,\beta)\in\mathbb{R}^2 \setminus \{(0,0)\}$ (except for)

* X is gaussian & Y is gaussian does not gaurentee $(X,Y)^T$ is gaussian vecto

e.g. $X\sim\mathcal{N}(0,1)$, $B\sim Ber(1/2)$, $R=2B-1\sim Rad(1/2)$

$Y=R\cdot X\sim\mathcal{N}(0,1)$

$\hookrightarrow \alpha=\beta=1$, we get:

$$X+Y=\begin{cases}2X & \text{with prob } 1/2\\ 0 & \text{with prob } 1/2\end{cases}$$

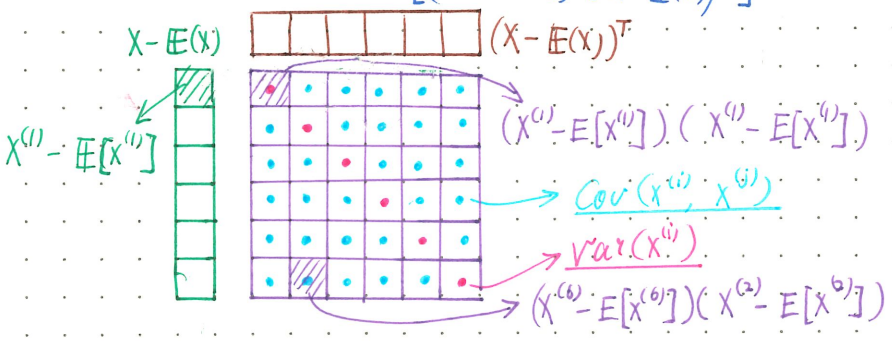
$\Downarrow$

$Cov(X,Y)=0$, but X,Y not independent, and $(X,Y)^T$ not gaussian

Covariance matrix

• The covariance matrix of a random vector $X=(X^{(1)},\dots X^{(n)})^T\in\mathbb{R}^d$ is given by

$$\Sigma=Cov(X)=\mathbb{E}[(X-\mathbb{E}(X))(X-\mathbb{E}(X))^T]$$



• Linear propertity:

$$Cov(AX+B)=Cov(AX)=A\,Cov(X)A^T=A\Sigma A^T$$

Diagonalization of the covariance matrix

Let Σ be a covariance matrix of size $d \times d$, Σ has properties:
(All entries of Σ would be non-negative)

identity matrix

orthogonal matrix: $AA^T = A^TA = I$

- Σ is symmetric, that is $\Sigma = \Sigma^T$
- Σ is diagnolizable to a diagonal matrix D via transformation $D = U\Sigma U^T$
  This implies: $\Sigma = U^T D U$
- Σ is positive semidefinite $\rightarrow$ diagonal matrix D has diagonal entries that are all non-negative
- Σ has a unique positive semidefinite square root matrix
  That is, there exists a positive semidefinite matrix $\Sigma^{\frac{1}{2}}$ that is unique, such that $\Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} = \Sigma$
- Σ has d orthonormal eigenvectors (even if there are repeated eigen values).
  Furthermore, if U is a matrix with rows corresponding to the orthonormed eigenvectors, the the diagnol matrix $D = U\Sigma U^T$ contains the eigen values of Σ along its diagonal. Therefore, diagonalization of a symmetric matrix involves finding its eigenvalues and orthonormal eigenvectors.
- If Σ is positive definite $\Leftrightarrow$ the diagoned matrix $D = U\Sigma U^T$ has diagonal entries that are strictly positive, the it is invertible and the inverse Σ^{-1} satisfies: $\Sigma^{-\frac{1}{2}} \cdot \Sigma^{-\frac{1}{2}} = \Sigma^{-1}$, where $\Sigma^{-\frac{1}{2}}$ is the inverse of $\Sigma^{\frac{1}{2}}$.

Multivariate Gaussian distribution

A Gaussian vector $X \in \mathbb{R}^d$ is completely determined by its expected value $\mathbb{E}[X] = \mu \in \mathbb{R}^d$ and covariance matrix Σ

$$X \sim \mathcal{N}_d(\mu, \Sigma)$$

The pdf is given by:

$$f_X(x) = f(x^{(1)}, \dots, x^{(d)}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$(x \in \mathbb{R}^d)$

Multivariate CLT

The CLT may be generalized to averages or random vectors (also vectors of averages)

Let $X_1, \dots X_n \in \mathbb{R}^d$ be independent copies of a random vector X such that $\mathbb{E}[X] = \mu$, $Cov(X) = \Sigma$.

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

$$\Updownarrow$$

$$\sqrt{n}\,\Sigma^{\frac{1}{2}}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, I_d)$$

Multivariate Delta Method

Let $(T_n)_{n\geq 1}$ be sequence of random vectors in $\mathbb{R}^d$ such that

$$\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

for some $\theta \in \mathbb{R}^d$ and some covariance $\Sigma \in \mathbb{R}^{d\times d}$

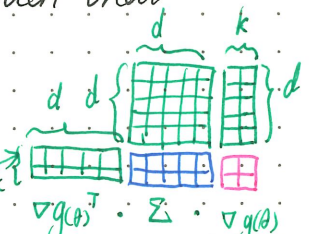
Let $g: \mathbb{R}^d \to \mathbb{R}^k$ $(k \geq 1)$ be continuously differentiable at θ. Then:

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, \nabla g(\theta)^T \Sigma \nabla g(\theta))$$

$$g = \begin{bmatrix} g_1 & g_2 & \cdots & g_k \end{bmatrix}$$

$$\frac{\partial g}{\partial \theta}(\theta) = \left(\frac{\partial g_j}{\partial \theta_i}\right)_{\substack{1\leq i\leq d \\ 1\leq j\leq k}} \in \mathbb{R}^{d\times k}$$

$$\begin{bmatrix} \frac{\partial g_1}{\partial \theta_1} & \frac{\partial g_2}{\partial \theta_1} & \cdots & \frac{\partial g_k}{\partial \theta_1} \\ \frac{\partial g_1}{\partial \theta_2} & \frac{\partial g_2}{\partial \theta_2} & \cdots & \frac{\partial g_k}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial \theta_d} & \frac{\partial g_2}{\partial \theta_d} & \cdots & \frac{\partial g_k}{\partial \theta_d} \end{bmatrix}$$



Fisher Information

• Definition:

Define the log-likelihood for one observation as:

$$l(\theta) = \log L_1(X, \theta), \quad \theta \in \Theta \subset \mathbb{R}^d$$

Assume that l is a.s. twice differentiable. Under some regularity conditions, the Fisher information of the statistical model is:

$$I(\theta) = \mathbb{E}\left[\nabla l(\theta) \nabla l(\theta)^T\right] - \mathbb{E}[\nabla l(\theta)]\,\mathbb{E}[\nabla l(\theta)]^T = -\mathbb{E}[H l(\theta)]$$

$$\mathbb{E}\left[\nabla l(\theta) \nabla l(\theta)^T\right] - \mathbb{E}[\nabla l(\theta)]\,\mathbb{E}[\nabla l]^T = I(\theta)$$

$$I_{ij} = \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\frac{\partial l}{\partial \theta_j}\right] - \mathbb{E}\left[\frac{\partial l}{\partial \theta_i}\right]\mathbb{E}\left[\frac{\partial l}{\partial \theta_j}\right]$$

$$\Downarrow$$

$$I(\theta) = \mathrm{var}(l'(\theta))$$

$I(\theta)$ → is an covariance matrix

If $\Theta \subset \mathbb{R}$, we get:

$$I(\theta) = \mathrm{var}(l'(\theta)) = -\mathbb{E}[l''(\theta)]$$

* If $\Theta \subset \mathbb{R}$, $I(\theta) = \mathrm{var}[l(\theta)] = -\mathbb{E}[l''(\theta)]$

Lecture 11: Fisher information, Asymptotic Normality of MLE, and the Method of Moments

• Proof of Fisher Information Equivalent Formulas for 1 Dimension

Let $(\mathbb{R}, \{P_\theta\}_{\theta \in \mathbb{R}})$ denote a continuous statistical model. Let $f_\theta(x)$ denote the pdf of the continuous distribution P_θ. Assume that $f_\theta(x)$ is twice-differentiable as a function of the parameter θ.

Proof: $I(\theta) = Var(l'(\theta)) = -E[l''(\theta)]$ $\quad (l(\theta) = \ln f_\theta(x))$

① $\int f_\theta(x)\,dx = 1 \Rightarrow \begin{cases} \int \frac{\partial}{\partial\theta} L_1(x,\theta)\,dx = 0 & (1) \\ \int \frac{\partial^2}{\partial\theta^2} L_1(x,\theta)\,dx = 0 & (2) \end{cases}$

$f_\theta(x) = L_1(x,\theta)$

② $l'(\theta) = \frac{\partial}{\partial\theta} \ln L_1(x,\theta) = \frac{\partial}{\partial\theta} L_1(x,\theta) / L_1(x,\theta)$

$$E[l'(\theta)] = \int \frac{\frac{\partial}{\partial\theta} L_1(x,\theta)}{L_1(x,\theta)} L_1(x,\theta)\,dx = \int \frac{\partial}{\partial\theta} L_1(x,\theta)\,dx \overset{(1)}{=} 0$$

$$Var(l'(\theta)) = E[(l'(\theta) - E[l'(\theta)])^2] = E[(l'(\theta))^2]$$

$$= \int \left(\frac{\frac{\partial}{\partial\theta} L_1(x,\theta)}{L_1(x,\theta)}\right)^2 L_1(x,\theta)\,dx$$

$$= \int \frac{\left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{L_1(x,\theta)}\,dx$$

③ $l''(\theta) = \frac{\frac{\partial^2}{\partial\theta^2} L_1(x,\theta) \cdot L_1(x,\theta) - \left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{(L_1(x,\theta))^2}$

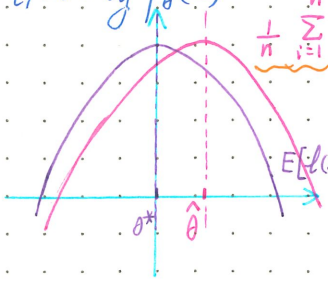
$$-E[l''(\theta)] = -\int \frac{\frac{\partial^2}{\partial\theta^2} L_1(x,\theta) \cdot L_1(x,\theta) - \left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{(L_1(x,\theta))^2} L_1(x,\theta)\,dx$$

$$= -\int \frac{\partial^2}{\partial\theta^2} L_1(x,\theta)\,dx \overset{(2)}{=} 0 + \int \frac{\left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{L_1(x,\theta)}\,dx$$

$$= \int \frac{\left(\frac{\partial}{\partial\theta} L_1(x,\theta)\right)^2}{L_1(x,\theta)}\,dx = Var(l'(\theta))$$

Informal Idea of Proof of Asymptotic Normality of MLE

$l_i(\theta) = \lg f_\theta(x)$



$\frac{1}{n}\sum_{i=1}^{n} \lg f_\theta(x) = \frac{1}{n}\sum_{i=1}^{n} l_i(\theta)$ estimator

$\frac{\partial}{\partial\theta}\sum_{i=1}^{n} l_i(\hat{\theta}) = \sum_{i=1}^{n} l_i'(\hat{\theta}) = 0$

$E[l'(\theta^*)] = 0$

first order Taylor Expansion

$$0 = \sum_{i=1}^{n} l_i'(\hat{\theta}) \cong \sum_{i=1}^{n}\left[l_i'(\theta^*) + (\hat{\theta}-\theta^*)\, l''(\theta^*)\right]$$

$$= \sum_{i=1}^{n}\left[\left(l_i'(\theta^*) - E[l_i'(\theta^*)]\right) + (\hat{\theta}-\theta^*)\, l''(\theta^*)\right]$$

↓ CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[l_i'(\theta^*) - E[l_i'(\theta^*)]\right] \xrightarrow[n\to\infty]{d} N(0, \operatorname{var}(l_i'(\theta^*)))$$

$\operatorname{var}(l_i'(\theta^*)) = I(\theta^*)$

↓

$$0 \cong N(0, I(\theta^*)) + \sqrt{n}(\hat{\theta}-\theta^*)\frac{1}{n}\sum_{i=1}^{n} l''(\theta^*)$$

↓ LLN

$$\frac{1}{n}\sum_{i=1}^{n} l''(\theta^*) \xrightarrow[n\to\infty]{P} -I(\theta)$$

↓

$$\sqrt{n}(\hat{\theta}-\theta^*) \sim N\left(0, \frac{I(\theta^*)}{(-I(\theta^*))^2}\right)$$

$$= N(0, I(\theta^*)^{-1})$$

• Asymptotic normality of the MLE

Theorem:

Let $\theta^* \in \Theta$ (the true parameter). Assume the following:

1. The parameter is identifiable;
2. For all $\theta \in \Theta$, the support of P_θ does not depend on θ;
3. θ^* is not on the boundary of Θ;
4. $I(\theta)$ is invertible in a neighborhood of θ^*;
5. A few more technical conditions.

Then, $\hat{\theta}_n^{MLE}$ satisfies:

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^* \quad \text{w.r.t. } P_{\theta^*}$$

$$\sqrt{n}\,(\hat{\theta}_n^{MLE} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, I(\theta^*)^{-1}) \quad \text{w.r.t. } P_{\theta^*}$$

Method of Moments

• Moments:

Let $X_1, \dots, X_n$ be an iid sample associated with a statistical model $(E, (P_\theta)_{\theta\in\Theta})$

Assume that $E \subseteq \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$

Population moments: Let $m_k(\theta) = \mathbb{E}_\theta[X_i^k]$, $1 \leq k \leq d$

Empirical moments: Let $\hat{m}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^n X_i^k$, $1 \leq k \leq d$

From LLN,

$$\hat{m}_k \xrightarrow[n\to\infty]{P/a.s.} m_k(\theta)$$

more compactly, we say that the whole vector converges:

$$(\hat{m}_1, \dots, \hat{m}_d) \xrightarrow[n\to\infty]{P/a.s.} (m_1(\theta), \dots, m_d(\theta))$$

• Moments estimator :

Let

$$M : \Theta \longrightarrow \mathbb{R}^d$$

$$\theta \longmapsto M(\theta) = (m_1(\theta), \dots, m_d(\theta))$$

Assume M is one to one :

$$\theta = M^{-1}(m_1(\theta), \dots, m_d(\theta))$$

Definition :

Moments estimator of θ :

$$\hat{\theta}_n^{MM} = M^{-1}(\hat{m}_1, \dots, \hat{m}_d),$$

provided it exists

• Generalized Method of Moments

* Applying the multivariate CLT and Delta method yields :

• Theorem :

$$\sqrt{n}(\hat{\theta}_n^{MM} - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \Gamma(\theta)) \quad (w.r.t.\ P_\theta)$$

where $\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta))\right]^T \Sigma(\theta) \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta))\right]$

$$\sqrt{n}(M^{-1}(\hat{m}) - M^{-1}(M(\theta))) \longrightarrow \mathcal{N}(0, (\nabla M^{-1})^T \Sigma (\nabla M^{-1}))$$

MLE v.s. Moment estimator

• Quadratic risk : In general, the MLE is more accurate

• MLE still gives good results if model is misspecified

• Computational issue : Sometimes, MLE is intractable but MM is easier (polynomial equations)

Lecture 12: M-Estimation

M-estimators:

Idea:

- Let $X_1, \dots X_n$ be iid with some unknown distribution $\mathbb{P}$ in some sample space E ($E \subseteq \mathbb{R}^d$ for some $d \geq 1$)
- No statistical model needs to be assumed (similar to ML)
- Goal: estimate some parameter μ^* associated with $\mathbb{P}$, e.g. its mean, variance, median, other quantiles, the true parameter in some statistical model...
- Find a function $f: E \times M \to \mathbb{R}$, where M is the set of all possible values for the unknown μ^*, such that:

$$Q(\mu) := \mathbb{E}[f(X_1, \mu)]$$

achieves its minimum at $\mu = \mu^*$

* Common examples
- Mean: $f(x, \theta) = \frac{(x-\theta)^2}{2}$
- Median $f(x, \theta) = |x - \theta|$

MLE is an M-estimator

Assume that $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$ is a statistical model associated with the data.

Theorem

Let $M = \Theta$ and $f(x, \theta) = -\log L_1(x, \theta)$, provided the likelihood is positive everywhere. Then,

$$\mu^* = \theta^*$$

where $\mathbb{P} = \mathbb{P}_{\theta^*}$ (i.e. θ^* is the true value of the parameter)

Asymptotic normality

Let $\mu^* \in M$ (the true parameter). Assume the following:

1. μ^* is the only minimizer of the function Q
2. $J(\mu)$ is invertible for all $\mu \in M$
3. A few more technical conditions

Then $\hat{\mu}_n$ satisfies:

$$\hat{\mu}_n \xrightarrow[n \to \infty]{\mathbb{P}} \mu^* \qquad \sqrt{n}(\hat{\mu}_n - \mu^*) \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0, J(\mu^*)^{-1} K(\mu^*) J(\mu^*)^{-1})$$

$$J(\mu) = \frac{\partial^2 Q}{\partial \mu \partial \mu^T}(\mu) \overset{\text{under some regularity conditions}}{=} \mathbb{E}\left[\frac{\partial^2 f}{\partial \mu \partial \mu^T}(X_1, \mu)\right]$$

$$K(\mu^*) = \mathrm{Cov}\left[\frac{\partial f}{\partial \mu}(X_1, \mu)\right]$$

Unit 4: Hypothesis testing

Lecture 13: Chi Squared Distribution T-test

Motivation:

- Clinical trials - test a drug that is supposed to lower LDL ("bad cholesterol").

Let $\Delta_d > 0$ denote expected decrease of LDL level for a patient that used the drug

Let $\Delta_c > 0$ denote expected decrease of LDL level for a patient that used placebo

We want to know if $\Delta_d > \Delta_c$

We observe two independent samples: $\begin{cases} X_1, \dots X_n \overset{iid}{\sim} \mathcal{N}(\Delta_d, \sigma_d^2) & \text{(test)} \\ Y_1, \dots Y_n \overset{iid}{\sim} \mathcal{N}(\Delta_c, \sigma_c^2) & \text{(control)} \end{cases}$

- Hypothesis testing:

$H_0: \Delta_c = \Delta_d$ v.s. $H_1: \Delta_d > \Delta_c$

* Since data is Gaussian, we don't need CLT

$\bar{X}_n \sim \mathcal{N}(\Delta_d, \frac{\sigma_d^2}{n})$ $\bar{Y}_m \sim \mathcal{N}(\Delta_c, \frac{\sigma_c^2}{m})$

↓ See probability 6.414x Lec 12, sum of independent normal r.v.'s

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\sigma_d^2}{n} + \frac{\sigma_c^2}{m}}} \sim \mathcal{N}(0,1)$$

- Asymptotic test:

① Assume that $m = cn$ and $n \to \infty$

② Using Slutsky's lemma, we have:

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}} \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0,1)$$

→ using estimator of variance because true variance unknown

$$\hat{\sigma}_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad \hat{\sigma}_c^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$$

→ unbiased estimator

③ We get following test at asymptotic level α: * one-sided, two sample test

$$R_\psi = \left\{ \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}} > q_{1-\alpha} \right\}$$

→ $(1-\alpha)$ quantile of $\mathcal{N}(0,1)$

- Small sample size:
  - Can not realistically apply Slutsky's lemma
  - We need to find the (asymptotic) distribution of quantities of the form

$$\frac{\bar{X}_n - \mu}{\sqrt{\hat{\sigma}^2}}$$

when $X_1, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

  - It turns out that this distribution does not depend on μ or σ, so we can compute its quantiles.

The χ^2 distribution

- Definition:

For a positive integer d, the χ^2 (pronounced "Kai-squared") distribution with d degree of freedom is the law of the random variable

$$Z_1^2 + Z_2^2 + \dots + Z_d^2 \quad \text{where } Z_1, \dots, Z_d \overset{iid}{\sim} N(0,1)$$

- Examples:
  - If $Z \sim N_d(0, I_d)$, then $\|Z\|_2^2 \sim \chi^2_d$ (I_d: $d \times d$ identity matrix)
  - $\chi^2_2 = \text{Exp}(1/2)$
- Properties for $V \sim \chi^2_k$

$$E[V] = E[Z_1^2] + \dots + E[Z_d^2] = d$$

$$\text{Var}[V] = \text{Var}(Z_1^2) + \dots + \text{Var}(Z_d^2) = 2d$$

Important example: the sample variance

- Sample variance is given by:

$$S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2$$

- Cochran's theorem: for $X_1, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, if S_n is the sample variance:

$$\bar{X}_n \perp\!\!\!\perp S_n \quad \text{(for all } n\text{)}$$

$$\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$$

* We often prefer the unbiased estimator of σ^2: $\tilde{S}_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

Student's T distribution

• Definition

For a positive integer d, the Student's T distribution with d degrees of freedom (denoted by t_d) is the law of the random variable $\frac{Z}{\sqrt{V/d}}$ where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_d$ and $Z \perp\!\!\!\perp V$ (Z is independent of V)

• Student's T test (one sample, two-sided)

① Let $X_1, \dots X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown

② We want to test:

$H_0: \mu = 0$ v.s. $H_1: \mu \neq 0$

③ Test statistic:

$$T_n = \frac{\bar{X}_n}{\sqrt{\tilde{S}_n/n}} = \frac{\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{\tilde{S}_n/\sigma^2}} \quad \sim \mathcal{N}(0,1) \quad \sim \frac{\chi^2_{n-1}}{n-1}$$

④ Since $\sqrt{n}\,\bar{X}_n/\sigma \sim \mathcal{N}(0,1)$ and $\tilde{S}_n/\sigma^2 \sim \frac{\chi^2_{n-1}}{n-1}$ are independent (Cochran's theorem)

$T_n \sim t_{n-1}$ (Student's T distribution with $n-1$ degrees of freedom)

⑤ Student's test with (non asymptotic) level $\alpha \in (0,1)$:

$\psi_\alpha = \mathbb{1}\{|T_n| > q_{\alpha/2}\}$ → $(1-\alpha/2)$-quantile of t_{n-1}

• Student's T test (one sample, one-sided)

① ··· as above

② we want to test:

$H_0: \mu \leq \mu_0$ v.s. $H_1: \mu > \mu_0$

③ Test statistic:

$$T_n = \frac{\bar{X}_n - \mu_0}{\sqrt{\tilde{S}_n/n}} \sim t_{n-1} \quad \text{(under } H_0\text{)}$$

④ Student's test with (non asymptotic) level $\alpha \in (0,1)$:

$\psi_\alpha = \mathbb{1}\{T_n > q_\alpha\}$ → $(1-\alpha)$-quantile of t_{n-1}

• Two-sample T-test

① Back to clinical trial example (motivation)

② We want to know distribution of:

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}}$$

③ We approximate:

$$\frac{\bar{X}_n - \bar{Y}_m - (\Delta_d - \Delta_c)}{\sqrt{\frac{\hat{\sigma}_d^2}{n} + \frac{\hat{\sigma}_c^2}{m}}} \sim t_N$$

where $N \doteq \frac{\left(\hat{\sigma}_d^2/n + \hat{\sigma}_c^2/m\right)^2}{\frac{\hat{\sigma}_d^4}{n^2(n-1)} + \frac{\hat{\sigma}_c^4}{m^2(m-1)}}$ $\geq \min(n, m)$ $\leq m+n$

* Welch-Satterthwaite formula

Advantage & drawback of Student's test

- Advantage: Non-asymptotic / can be run on small samples
  + can always use if large sample size
- Drawback: assumption that sample is Gaussian

Lecture 14: Wald's Test, Likelihood Ratio Test and Implicit Hypothesis Test

Wald's Test: A test based on the MLE

① Consider an iid sample $X_1, \dots, X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, where $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$ and let $\theta_0 \in \Theta$ be fixed and given.

② Hypothesis: $\begin{cases} H_0 : \theta^* = \theta_0 \\ H_1 : \theta^* \neq \theta_0 \end{cases}$

③ Let $\hat{\theta}^{MLE}$ be the MLE estimate. Assume the MLE technical conditions are satisfied.

④ If H_0 is true, then:

Asymptotic normality of MLE

$$\sqrt{n}\, I(\theta_0)^{1/2} \cdot (\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow[n\to\infty]{(d.)} \mathcal{N}_d(0, I_d)$$

$\sqrt{n}\, I(\theta^*)^{1/2}$ (since $\theta^* = \theta_0$)

$\sqrt{n}\, I(\hat{\theta}^{MLE})^{1/2}$ (Slutsky)

If $Z \sim \mathcal{N}_d(0, I_d)$, then $\|Z\|_2^2 \sim \chi^2_d$

⑤ Hence,

$$\underbrace{n(\hat{\theta}_n^{MLE} - \theta_0)^T I(\hat{\theta}^{MLE})(\hat{\theta}_n^{MLE} - \theta_0)}_{T_n} \xrightarrow[n\to\infty]{d} \chi^2_d$$

⑥ Wald's test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\}$$

q_α: $(1-\alpha)$-quantile of χ^2_d

* Wald's test is also valid if H_1 has the form "$\theta > \theta_0$" or "$\theta < \theta_0$" or "$\theta = \theta_1$"... (But less powerful)

Likelihood ratio test: A test based on the log-likelihood (MLE)

① Consider an iid sample $X_1, \dots X_n$ with statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ where $\Theta \subseteq \mathbb{R}^d$ $(d \geq 1)$

② Suppose the null hypothesis has the form

$$H_0 : \underbrace{(\theta_{r+1}, \dots, \theta_d)}_{(d-r) \text{ parameters}} = (\theta^{(0)}_{r+1}, \dots, \theta^{(0)}_d)$$

for some fixed and given numbers $\theta^{(0)}_{r+1}, \dots \theta^{(0)}_d$

③ Let

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} l_n(\theta) \quad (MLE)$$

and

$$\hat{\theta}^c_n = \operatorname*{argmax}_{\theta \in \Theta^c} l_n(\theta) \quad (\text{"constrained MLE"})$$

$$\Theta^c \hookrightarrow \{\theta \in \Theta : (\theta_{r+1}, \dots \theta_d) = (\theta^{(0)}_{r+1}, \dots \theta^{(0)}_d)\}$$

④ Test statistic:

$$T_n = 2\left(l_n(\hat{\theta}_n) - l_n(\hat{\theta}^c_n)\right)$$

⑤ Wilks' Theorem:

Assume H_0 is true and MLE technical conditions are satisfied. Then,

$$T_n \xrightarrow[n \to \infty]{(d)} \chi^2_{d-r}$$

⑥ Likelihood ratio test with asymptotic level $\alpha \in (0,1)$:

$$\psi = \mathbb{1}\{T_n > q_\alpha\}$$

$q_\alpha \hookrightarrow (1-\alpha)$-quantile of χ^2_{d-r}

Wald's test for implicit hypothesis

① Let $X_1, \dots, X_n$ be iid random variables and let $\theta \in \mathbb{R}^d$ be a parameter associated with the distribution of X_1 (e.g. a moment, the parameter of a statistical model, etc...)

② Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be continuously differentiable (with $k < d$)

③ Consider the following hypothesis:

$$H_0 : g(\theta) = 0 \quad \text{v.s.} \quad H_1 : g(\theta) \neq 0.$$

e.g. $\underset{(k=2)}{g(\theta) = (\theta_1, \theta_2)}$ or $\underset{(k=1)}{g(\theta) = \theta_1 - \theta_2}$, ...

④ Suppose an asymptotically normal estimator $\hat{\theta}_n$ is available:

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma(\theta))$$

$$\downarrow \text{Delta method}$$

$$\sqrt{n}\,(g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, \Gamma(\theta))$$

$$\Gamma(\theta) = \nabla g(\theta)^T \Sigma(\theta) \nabla g(\theta) \in \mathbb{R}^{k \times k}$$

⑤ Assume $\Sigma(\theta)$ is invertible and $\nabla g(\theta)$ has rank k. So, $\Gamma(\theta)$ is invertible and:

$$\sqrt{n}\,\Gamma(\theta)^{-1/2}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k)$$

$$\downarrow \text{Slutsky, if } \Gamma(\theta) \text{ is continuous in } \theta,$$

$$\sqrt{n}\,\Gamma(\hat{\theta})^{-1/2}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k)$$

⑥ Hence, if θ_0 is true, i.e., $g(\theta) = 0$,

$$\underbrace{n\, g(\hat{\theta}_n)^T \Gamma^{-1}(\hat{\theta}_n)\, g(\hat{\theta}_n)}_{T_n} \xrightarrow[n\to\infty]{(d)} \chi_k^2$$

⑦ Test with asymptotic level α:

$$\psi = \mathbb{1}\{T_n > q_\alpha\}$$

q_α ↳ $(1-\alpha)$-quantile of χ_k^2

Lecture 15: Goodness of Fit Test for Discrete Distributions

Goodness of fit tests - motivation

e.g. Let X be a r.v. Given iid copies of X we want to answer:

- Does X have distribution $N(0,1)$?
- Does X have distribution $U[0,1]$?
- Does X have PMF $p_1 = 0.3$, $p_2 = 0.5$, $p_3 = 0.2$

These are all goodness of fit (GoF) tests: we want to know if the hypothesized distribution is a good fit for the data.

* Key characteristic: no parametric
  - ↳ null hypothesis would be ~parametric
  - potential alternative could be more complicated

Discrete distribution - Probability Simplex

Let $E = \{a_1, \dots a_k\}$ be a finite space, and $(\mathbb{P}_p)_{p \in \Delta_k}$ be the family of all probability distributions on E:

Δ_k { Probability Simplex in $\mathbb{R}^k$, is the set of all vectors $p = [p_1, \dots p_k]^T$ such that:

$p^T \cdot \mathbf{1} = \mathbf{1}^T \cdot p = 1$, $p_i \geq 0$ for all $i = 1, \dots k$

↳ $\mathbf{1} = (1\ 1 \dots 1)^T$

Equivalently:

$$\Delta_k = \left\{ p = (p_1, \dots p_k) \in [0,1]^k : \sum_{i=1}^k p_i = 1 \right\}$$

For $p \in \Delta_k$ and $X \sim \mathbb{P}_p$,

$$\mathbb{P}_p[X = a_j] = p_j, \quad j = 1, \dots k$$

Multinomial Distribution

- Definition:

Multinomial distribution with K modalities (K possible outcomes) is a generalization of binomial distribution. It models probabilities of counts of K possible outcomes of experiment in n' iid trials of experiment.

- Parameters: n', $p_1, \dots p_k$ → probability of observing outcome i in any trial. $\begin{cases} p_i \geq 0 \\ \sum_{i=1}^{k} p_i = 1 \end{cases}$
  ↳ number of iid trials. Let $p \triangleq [p_1\ p_2 \cdots p_k]^T$, $p \in \Delta_k$

- Representation:

Multinomial distribution can be represented by random vector $N \in \mathbb{Z}^k$

$N^{(i)}$: number of instance of outcome i

$\sum_{i=1}^{k} N^{(i)} = n'$

Multinomial pmf: $p_N(N^{(1)}=n^{(1)}, \dots N^{(k)}=n^{(k)}) = \frac{n'!}{n^{(1)}!\, n^{(2)}! \cdots n^{(k)}!} \prod_{i=1}^{k} p_i^{n^{(i)}}$

Categorical (Generalized Bernoulli) Distribution and likelihood

- Multinomial distribution with:
  - $n'=1$: categorical distribution
    - $K=2$, outcomes $\{0,1\}$: Bernoulli distribution
    - $K>2$: Generalized Bernoulli distribution

- Categorical random variable sample space:

$$E = \{a_1, \dots a_k\}$$

- Categorical random variable pmf:

$$P(X=a_j) = \prod_{i=1}^{k} p_i^{\mathbb{1}(a_i=a_j)} = p_j,\ j=1,\dots K$$

- Categorical statistical model:

$$(\{a_1, \dots a_k\}, \{P_p\}_{p \in \Delta_k})$$

- Categorical likelihood: $x_1, \dots x_n \overset{iid.}{\sim} X$, outcomes K, number of occurances N_i, $i=1,\dots K$

$$L_n(x_1, \dots x_n, p_1, \dots p_k) = p_1^{N_1} p_2^{N_2} \cdots p_k^{N_k} \leftarrow \text{for } n \text{ iid sequence of outcomes}$$

$$L(X, p_1, \dots p_k) = \prod_{i=1}^{k} p_i^{\mathbb{1}(X=a_i)} \leftarrow \text{for r.v. } X$$

Goodness of fit test (discrete, χ^2)

① Let $X_1, \dots X_n \overset{iid}{\sim} \mathbb{P}_p$, for some unknown $p \in \Delta_k$, and $p^\circ \in \Delta_k$ is fixed

Test:

$H_0 : p = p^\circ$ v.s. $H_1 : p \neq p^\circ$

with asymptotic level $\alpha \in (0,1)$

② Likelihood of model:

$$L_n(X_1, \dots X_n, p) = p_1^{N_1} p_2^{N_2} \dots p_k^{N_k}$$

where $N_j = \#\{i = 1, \dots n : X_i = a_j\}$

Let $\hat{p}$ be the MLE:

$$\hat{p}_j = \frac{N_j}{n}, \quad j = 1, \dots k$$

↳ $\hat{p}$ maximizes $\log L_n(X_1, \dots X_n, p)$ under the constraint

③ χ^2 test:

- If H_0 is true, then: $\sqrt{n}(\hat{p} - p^\circ)$ is asymptotically normal, and the following holds

Theorem (under H_0)

$$\underbrace{n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j^\circ)^2}{p_j^\circ}}_{T_n} \xrightarrow[n \to \infty]{(d)} \chi^2_{k-1}$$

- χ^2 test with asymptotic level α:

$$\psi_\alpha = \mathbb{1}\{T_n > q_\alpha\}$$

↳ $(1-\alpha)$-quantile of χ^2_{k-1}

- Asymptotic p-value of this test:

$$p\text{-value} = \mathbb{P}[Z > T_n \mid T_n]$$

↳ $Z \sim \chi^2_{k-1}$, $Z \perp\!\!\!\perp T_n$

---

* Pointwise convergence:

$\lim_{n \to \infty} g_n(x) = g(x)$

* Uniform convergence:

$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |g_n(x) - g(x)| = 0$

For every $M > 0$, there exists an n_M s.t.

$\sup_x |g_n(x) - g(x)| < M$ for all $n \geq n_M$

χ^2 test for a family of distribution:

① $X_1, \dots X_n \overset{iid}{\sim} P$ denote iid discrete r.v. supported on $\{0, \dots K\}$

Test:

$H_0 : P \in \{Bin(K, \theta)\}_{\theta \in (0,1)}$

$H_1 : P \notin \{Bin(K, \theta)\}_{\theta \in (0,1)}$

② Let f_θ denote pmf of $Bin(K, \theta)$

$\hat{\theta}$ denote MLE of parameter θ

N_j denote number of times j ($j \in \{0, 1, \dots K\}$) appears in data set $X_1, \dots, X_n$

③ χ^2 test

$$T_n := n \sum_{j=0}^{k} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)} \xrightarrow[n\to\infty]{(d)} \chi^2_{(k+1)-d-1}$$

support size; dimension of θ

use $f_{\hat{\theta}}(j)$ to estimate $P(X=j)$

Lecture 16: Goodness of fit tests continued

Kolmogorov-Smirnov, Kolmogorov-Lilliefors Tests, Q.Q. plots

CDF and empirical CDF

Let $X_1, \dots X_n$ be iid real random variables.

- CDF: $F(t) = \mathbb{P}[X_1 \le t]$, $\forall t \in \mathbb{R}$ (completely characterizes distribution of X_1.
- Empirical cdf: $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le t\} = \frac{\#\{i = 1, \dots n : X_i \le t\}}{n}$, $\forall t \in \mathbb{R}$
  - Consistency:
    - (by LLN): for all $t \in \mathbb{R}$, $F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$
    - Glivenko-Cantelli Theorem (Fundamental theorem of statistics)

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$$

  - Asymptotic normality:
    - (by CLT): for all $t \in \mathbb{R}$, $\sqrt{n}\,(F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, F(t)(1-F(t)))$
    - Donsker's Theorem: If F is continuous, then

$$\sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0 \le x \le 1} |B(x)|$$

$F(t)$ on $\mathbb{R}$; $[0,1]$, range of $F(t)$; $B(x)$: Brownian bridge on $[0,1]$, pivotal

Goodness of fit test (continuous, Kolmogorov-Smirnov)

① Let $X_1, \dots X_n$ be iid real random variable with unknown cdf F

Let F° be a continuous cdf

Test: $H_0 : F = F^\circ$ versus $H_1 : F \neq F^\circ$

② Let F_n be empirical cdf of samples $X_1, \dots X_n$

* If $F = F^\circ$, then $F_n(t) \approx F^\circ(t)$, for all $t \in \mathbb{R}$

③ Kolmogorov-Smirnov test

$$T_n = \sup_{t \in \mathbb{R}} \sqrt{n} \, |F_n(t) - F^\circ(t)|$$

By Donsker's theorem, if H_0 is true, then:

$$T_n \xrightarrow[n \to \infty]{(d)} Z$$

Z → supremum of a Brownian bridge

④ KS test with asymptotic level α:

$$\delta_\alpha^{KS} = \mathbb{1}\{T_n > q_\alpha\}$$

q_α → $(1-\alpha)$-quantile of Z (obtained in tables)

④ p-value of KS test:

$$\mathbb{P}[Z > T_n \mid T_n]$$

* Computational issues

Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be reordered sample.

$$T_n = \sqrt{n} \max_{i=1,\dots n} \left\{ \max \left(\left| \frac{i-1}{n} - F^\circ(X_{(i)}) \right|, \left| \frac{i}{n} - F^\circ(X_{(i)}) \right| \right) \right\}$$

* Pivotal distribution

T_n is called a pivotal statistic: if H_0 is true, the distribution of T_n does not depend on distribution of X_i's and it is easy to reproduce it in simulations

Let $U_i = F^\circ(X_i)$, $i = 1, \dots n$. Let G_n be empirical cdf of $U_1, \dots U_n$.

If H_0 is true, then: $U_1 \dots U_n \overset{iid}{\sim} \text{Unif}[0,1]$

$$T_n = \sup_{0 \leq x \leq 1} \sqrt{n} \, |G_n(x) - x|$$

* Quantiles and p-values

For some large integer M:

Simulate M iid copies of $T_n^1, \dots T_n^M$ of T_n;

Estimate $(1-\alpha)$-quantile $q_\alpha^{(n)}$ of T_n by taking the sample $(1-\alpha)$-quantile $\hat{q}_\alpha^{(n,M)}$ of $T_n^1, \dots T_n^M$.

Test with approximate level α:

$$\delta_\alpha = \mathbb{1}\{T_n > \hat{q}_\alpha^{(n,M)}\}$$

Approximate p-value of this test:

$$\text{p-value} \approx \frac{\#\{j = 1, \dots M : T_n^j > T_n\}}{M}$$

* Other goodness of fit tests

We want to know distance of two functions: $F_n(t)$ and $F(t)$

- Kolmogorov-Smirnov: $d(F_n, F) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$ (ℓ_∞)
- Cramér-von Mises: $d^2(F_n, F) = \int_{\mathbb{R}} [F_n(t) - F(t)]^2 \, dF(t)$ (ℓ_2)
- Anderson-Darling: $d^2(F_n, F) = \int_{\mathbb{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1-F(t))} \, dF(t)$

* Motivation for Composite Goodness of Fit Tests

What if I want to test: "Does X have Gaussian distribution?" but I don't have parameters?

Plug in: $\sup_{t \in \mathbb{R}} |F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t)|$

$\hat{\mu} = \bar{X}_n$ $\quad \hat{\sigma}^2 = S_n^2$ $\quad \Phi_{\hat{\mu}, \hat{\sigma}^2}(t)$: cdf of $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$

$\Downarrow$

Donsker's Theorem is no longer valid

Goodness of fit test (continuous, Kolmogorov-Lilliefors)

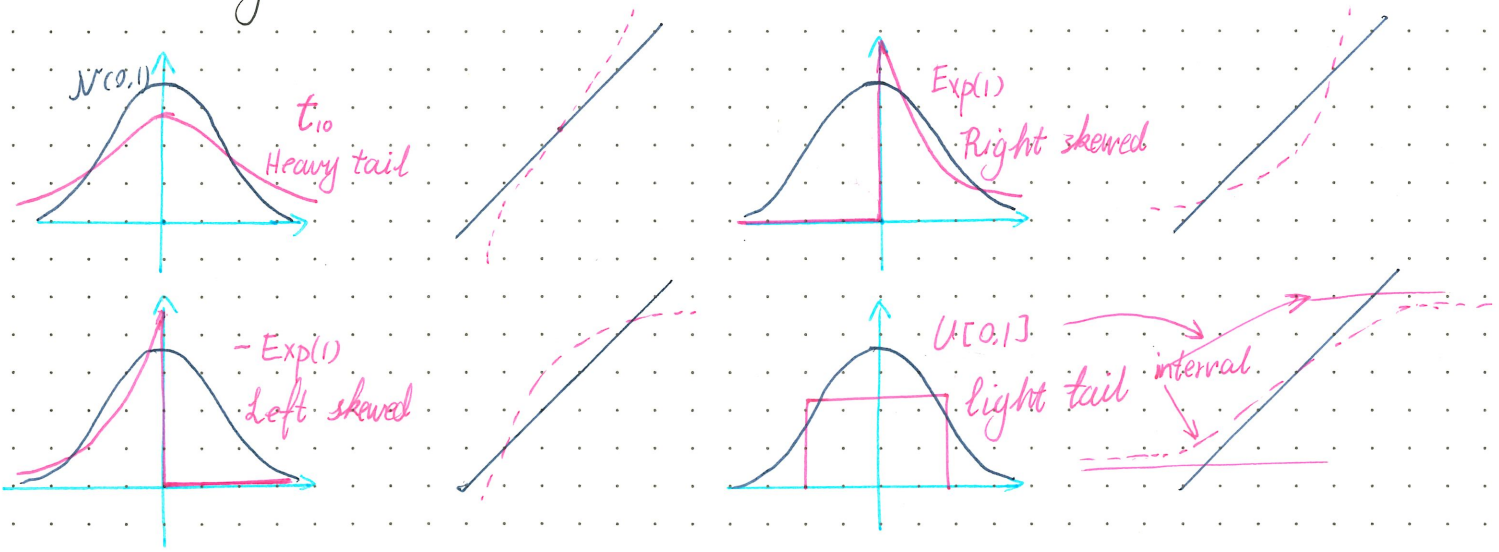
Test statistics: $\sup_{t\in\mathbb{R}} |F_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)|$ → pivot, use table

Quantile-Quantile (Q-Q) plots

- A visual way of performing GoF tests
- Not a formal test, but quick and easy way to check if a distribution is plausible
- Main idea: check visually if the plot F_n is close to that of F or equivalently if the plot of F_n^{-1} is close that of F^{-1}
- Define: $F_n^{-1}(i/n) = X_{(i)}$ → i-th largest observation
- Check if points:

$$\left(F^{-1}\left(\frac{1}{n}\right), F_n^{-1}\left(\frac{1}{n}\right)\right), \left(F^{-1}\left(\frac{2}{n}\right), F_n^{-1}\left(\frac{2}{n}\right)\right), \cdots\cdots, \left(F^{-1}\left(\frac{n-1}{n}\right), F_n^{-1}\left(\frac{n-1}{n}\right)\right)$$

are near $y=x$



Unit 5: Bayesian Statistics

Lecture 17: Introduction to Bayesian Statistics

Frequentist Approach

- Assume Statistical Model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$
- Assume that data $X_1, \dots X_n$ was drawn iid from $\mathbb{P}_{\theta^*}$ for some unknown fixed θ^*
- When we use MLE for example, we looked at all possible $\theta \in \Theta$

Before seeing the data, we did not prefer a choice of $\theta \in \Theta$ over another

Bayesian Approach

- In many practical contexts, we have prior belief of θ^*
- Using the data, we want to update that belief and transform it into posterior belief.

Prior and Posterior

① Prior distribution: a probability distribution on a parameter space Θ, with some pdf $\pi(\cdot)$

② Let $X_1, \dots X_n$ be a sample of n random variables

③ Denote by $\mathcal{L}_n(\cdot|\theta)$ the joint pdf of $X_1, \dots X_n$ conditionally on θ, where $\theta \sim \pi$

$\mathcal{L}_n(X_1, \dots X_n|\theta)$: likelihood used in frequentist approach

④ Posterior distribution: conditional distribution of θ given by $X_1, \dots X_n$

Denote by $\pi(\cdot|X_1, \dots, X_n)$ its pdf

Baye's formula

$$\pi(\theta|X_1, \dots X_n) = \frac{\pi(\theta)\,\mathcal{L}_n(X_1, \dots X_n|\theta)}{\int_\Theta \pi(\theta)\,\mathcal{L}_n(X_1, \dots X_n|\theta)} \quad \forall \theta \in \Theta$$

$\int_\Theta \pi(\theta)\,\mathcal{L}_n(X_1, \dots X_n|\theta)$ → constant, does not depend on θ

$$\Downarrow$$

$$\pi(\theta|X_1, \dots X_n) \propto \pi(\theta)\,\mathcal{L}_n(X_1, \dots X_n|\theta)$$

e.g.

Bernoulli experiment with Beta prior

① $p \sim \text{Beta}(a,a)$: $\pi(p) \propto p^{a-1}(1-p)^{a-1}$, $p \in (0,1)$

② Given p, $X_1, \dots X_n \overset{iid}{\sim} \text{Ber}(p)$, so: $\mathcal{L}_n(X_1, \dots, X_n|p) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$

③ Hence, $\pi(p|X_1, \dots X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}$

→ $\text{Beta}(a+\sum_{i=1}^n X_i,\ a+n-\sum_{i=1}^n X_i)$

*Conjugate prior

(posterior is same family with prior)

Non-informative priors

* We can still use a Bayesian approach if we have no prior information
  How to pick π?
- Good candidate: $\pi(\theta) \propto 1$ (constant pdf on Θ)
  - If Θ bounded: uniform
  - If Θ unbounded: does not define a proper pdf on Θ
  - Improper prior: a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable
    In general, one can still define a posterior distribution using an improper prior, using Baye's formula.

Jeffreys prior

- Jeffreys prior: $\pi_J(\theta) \propto \sqrt{\det I(\theta)}$ → Fisher information matrix of statistical model associated with $X_1, \dots, X_n$ in the frequentist approach (provided it exists)
  more weight → ① when $I(\theta)$ is high (② values of θ where MLE has less uncertainty) ③ data giving more info ④ potential outcomes more sensitive to θ changes
- Examples:
  - Bernoulli: $\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}$, $p \in (0,1)$, is a $\text{Beta}(\frac{1}{2}, \frac{1}{2})$
  - Gaussian $\mathcal{N}(\theta, 1)$: $\pi_J(\theta) \propto 1$, $\theta \in \mathbb{R}$, is an improper prior

- Jeffreys prior satisfies a reparametrization invariance principle:
  - If η is reparametrization of θ (i.e., $\eta = \phi(\theta)$ for some one-to-one map ϕ), then the pdf $\tilde{\pi}(\cdot)$ of η satisfies:
    $$\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)}$$ → fisher information of statistical model parametrized by η instead of θ
  - If η is reparametrization of θ, $\phi(\theta) = \eta$ and ϕ is invertible:
    $$\tilde{\pi}_J(\eta) = \pi_J(\phi(\theta)) \frac{d\theta}{d\eta} = \pi_J(\phi(\theta)) \frac{1}{\phi'(\theta)}$$

Bayesian confidence regions

- Definition: Bayesian confidence region with level α ($\alpha \in (0,1)$) is a random subset R of parameter space Θ, which depends on sample $X_1, \dots X_n$, such that: (depend on ← prior $\pi(\cdot)$)
  $$\mathbb{P}(\theta \in R \mid X_1, \dots, X_n) = 1 - \alpha$$
* "Bayesian confidence region" and "confidence interval" are two distinct notions

Bayesian estimation:

* The Bayesian framework can also be used to estimate the true underlying parameter (hence, in frequentist approach).
In this case, prior distribution does not reflect a prior belief: it is just an artificial tool used in order to define a new class of estimators.

- Bayesian estimator:

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \cdot \tilde{\pi}_i(\theta | X_1, \dots X_n)\, d\theta \quad \text{(posterior mean)}$$

→ depends on choice of prior distribution $\tilde{\pi}_i$ (hence superscript π_i)

Another choice: $\hat{\theta}^{MAP}$ → maximum a posteriori provided it exists...

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \tilde{\pi}_i(\theta | X_1, \dots, X_n)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \tilde{\pi}_i(\theta) \cdot \mathcal{L}_n(X_1, \dots, X_n | \theta)$$

* In general, the asymptotic properties of Bayes estimator do not depend on choice of prior

Unit 6: Linear Regression

Lecture 19: Linear Regression I

Modeling Assumption

- (X_i, Y_i), $i = 1, \dots n$ are iid for some unknown joint distribution $\mathbb{P}$
- $\mathbb{P}$ can be described entirely by (assuming all exist)
  - Either a joint PDF $h(x,y)$
  - The marginal density of X $h(x) = \int h(x,y)\,dy$ and the conditional density $h(y|x) = \frac{h(x,y)}{h(x)}$

Partial modeling

We can also describe the distribution only partially, e.g. using
- The expectation of Y : $\mathbb{E}[Y]$
- The conditional expectation of Y given $X = x$: $\mathbb{E}[Y|X=x]$
  - The function:
    $$x \mapsto f(x) := \mathbb{E}[Y|X=x] = \int y\, h(y|x)\,dy$$
  - is called regression function
- Other possibilities:
  - conditional median:
    $m(x)$ such that $\int_{-\infty}^{m(x)} h(y|x)\,dy = \frac{1}{2}$
  - conditional quantile
  - conditional variance (no information about location)

Linear regression

Focus on modeling the regression function $f(x) = \mathbb{E}[Y|X=x]$

* Too many possible regression functions f (nonparametric)

Useful to restrict to simple functions that are described by a few parameters

Simplest: $f(x) = a + bx$ linear / affine function

Under this assumption, we discuss linear regression

Probabilistic analysis

① Let X and Y be two real r.v. (not necessarily independent) with two moments and such that $\text{var}(X) > 0$

② The theoretical linear regression of Y on X is the line $x \mapsto a^* + b^* x$ where

$$(a^*, b^*) = \underset{(a,b)\in\mathbb{R}^2}{\text{argmin}}\ \mathbb{E}[(Y-a-bX)^2]$$

③ Setting partial derivatives to 0 gives:

$$b^* = \frac{\text{cov}(X,Y)}{\text{var}(X)}$$

$$a^* = \mathbb{E}[Y] - b^*\mathbb{E}[X] = \mathbb{E}[Y] - \frac{\text{cov}(X,Y)}{\text{var}(X)}\mathbb{E}[X]$$

④ Noise:

Clearly the points are not exactly on the line $x \mapsto a^* + b^* x$ if $\text{Var}(Y|X=x) > 0$

The random variable $\varepsilon = Y - (a^* + b^* X)$ is called noise and satisfies:

$$Y = a^* + b^* X + \varepsilon$$

with: $\begin{cases} E[\varepsilon] = 0 \\ \text{cov}(X, \varepsilon) = 0 \end{cases}$

Least squares

- Definition: The least squares estimator (LSE) of (a, b) is the minimizer of sum of squared errors:

$$\sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

- $\hat{a}, \hat{b}$ is given by $\begin{cases} \hat{b} = \dfrac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \\ \hat{a} = \bar{Y} - \hat{b}\bar{X} \end{cases}$

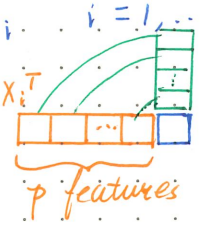
- Residuals: $\hat{\varepsilon}_i = |y_i - (\hat{a} + \hat{b}x_i)|$

Lecture 20: Linear Regression 2

Multivariate regression

$$Y_i = X_i^T \beta^* + \varepsilon_i \qquad i = 1, \dots, n$$

$$X_i = \begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \\ \vdots \\ X_i^{(p)} \end{pmatrix}$$



① Vector of explanatory variables / covariates: $X_i \in \mathbb{R}^p$
↳ wlog, assume its first coordinate is 1

Response / dependent variable: Y_i

② $\beta^* = (a^*, b^{*T})^T$; $\beta_1^* (= a^*)$ is called intercept

$\{\varepsilon_i\}_{i=1,\dots,n}$: noise terms satisfying $\text{cov}(X_i, \varepsilon_i) = 0$

*Definition:
The least squares estimator (LSE) of β^* is the minimizer of squared error.

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

LSE in matrix form:

① Let $Y = (Y_1, \dots Y_n)^T \in \mathbb{R}^n$

$$\begin{bmatrix} — X_1 — \\ — X_2 — \\ \vdots \\ — X_n — \end{bmatrix}$$

$\mathbb{X}$ be $n \times p$ matrix whose rows are $X_1^T, \dots, X_n^T$ (design matrix)

$\varepsilon = (\varepsilon_1, \dots \varepsilon_n)^T \in \mathbb{R}^n$ (unobserved noise)

② $Y = \mathbb{X}\beta^* + \varepsilon$ (β^* unknown)

③ The LSE $\hat{\beta}$ satisfies:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \| Y - \mathbb{X}\beta \|_2^2$$

④ Closed form solution:

Assume $\text{rank}(\mathbb{X}) = p$, analytical computation of the LSE:

$$\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T Y$$

* Geometric interpretation of the LSE: $\mathbb{X}\hat{\beta}$ is the orthogonal projection of Y onto the subspace spanned by the columns of $\mathbb{X}$:

$$\mathbb{X}\hat{\beta} = PY$$
↳ $P = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$

Statistical inference:

$$Y = X\beta^* + \varepsilon$$

*To make inference (confidence regions, tests) we need more assumptions

Assumptions:

① The design matrix X is deterministic and $\text{rank}(X) = p$

② The model is homoscedastic: $\varepsilon_1, \dots \varepsilon_n$ are iid

③ The noise vector is Gaussian: $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ for some known/unknown $\sigma^2 > 0$

$$\Downarrow$$

$$Y \sim \mathcal{N}_n(X\beta^*, \sigma^2 I_n)$$

Properties of LSE

① LSE = MLE

② Distribution of $\hat{\beta}$: $\hat{\beta} \sim \mathcal{N}_p(\beta^*, \sigma^2 (X^T X)^{-1})$

③ Quadratic risk of $\hat{\beta}$: $\mathbb{E}[\|\hat{\beta} - \beta\|_2^2] = \sigma^2 \text{tr}((X^T X)^{-1})$

④ Predicted error: $\mathbb{E}[\|Y - X\hat{\beta}\|_2^2] = \sigma^2 (n-p)$

⑤ Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|_2^2}{n-p} = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

Theorem:

$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

$\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ (Cochran's theorem)

Significance tests:

① Test whether the j-th explanatory variable is significant in the linear regression ($1 \le j \le p$)

② $H_0: \beta_j = 0$ v.s. $H_1: \beta_j \neq 0$

③ If γ_j is the j^{th} diagonal coefficient of $(X^T X)^{-1}$ ($\gamma_j > 0$):

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \gamma_j}} \sim t_{n-p}$$

④ Let $T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \gamma_j}}$

⑤ Test with non asymptotic level $\alpha \in (0,1)$: $R_{j,\alpha} = \mathbb{1}\{|T_n^{(j)}| > q_{\alpha/2}(t_{n-p})\}$

($q_{\alpha/2}(t_{n-p})$: $(1-\alpha/2)$-quantile of t_{n-p})

p-values...

Bonferroni's test

① Test whether a group of explanatory variables is significant in the linear regression

② $H_0: \beta_j = 0, \forall j \in S$ v.s. $H_1: \exists j \in S, \beta_j \neq 0$ where $S \subseteq \{1, \dots p\}$

③ Bonferroni's test: $R_{S,\alpha} = \bigcup_{j \in S} R_{j,\frac{\alpha}{k}}$ where $k = |S|$

④ This test has nonasymptotic level at most α

Remarks:

- Linear regression exihibits correlation, NOT causality
- Normality of the noise: One can use goodness of fit tests to test whether the residuals $\hat{\epsilon}_i = Y_i - X_i^T \hat{\beta}$ are Gaussian
- Deterministic design: If X is not deterministic, all the above can be understood conditionally on X, if the noise is assumed to be Gaussian, conditionally on X.

Unit 7: Generalized linear models

Lecture 21: Introduction to Generalized Linear Models: Exponential Families

Linear model

- A Gaussian linear model assumes:

$$Y|X=x \sim N(\mu(x), \sigma^2 I)$$

$\mu(x)$: regression function

$$\mathbb{E}[Y|X=x] = \mu(x) = x^T\beta$$

- Two model components: (to be relaxed in this unit)
  1. Random component: the response variable Y is continuous and $Y|X=x$ is Gaussian with mean $\mu(x)$
  2. Regression function: $\mu(x) = x^T\beta$ (linear)

Generalization

- A generalized linear model (GLM) generalizes normal linear regression models in the following directions:
  1. Random component: $Y|X=x \sim$ some distribution (e.g. Bernoulli, exponential, Poisson)
  2. Regression function: $g(\mu(x)) = x^T\beta$
     - g: link function
     - $\mu(x)$: regression function

Exponential family

- A family of Distribution $\{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$ is said to be a k-parameter exponential family on $\mathbb{R}^q$, if there exist real valued functions:

$\{\eta_1, \eta_2, \dots \eta_k$ and B of θ
$T_1, T_2, \dots T_k$ and h of $y \in \mathbb{R}^q\}$ such that the density function (pmf/pdf) of $\mathbb{P}_\theta$:

$$f_\theta(y) = \exp\left[\sum_{i=1}^{k} \eta_i(\theta) T_i(y) - B(\theta)\right] h(y)$$

- $\eta_i(\theta)$: $\mathbb{R}^k \to \mathbb{R}$
- $[\eta_1(\theta) \cdots \eta_k(\theta)]$: $\mathbb{R}^k \to \mathbb{R}^k$
- $T_i(y)$: $\mathbb{R}^q \to \mathbb{R}$
- $\begin{bmatrix} T_1(y) \\ \vdots \\ T_k(y) \end{bmatrix}$: $\mathbb{R}^q \to \mathbb{R}^k$

* $k = q = 1$:

$$f_\theta(y) = \exp[\eta(\theta) T(y) - B(\theta)] h(y)$$

One-parameter canonical exponential family

- Canonical exponential family for $k=1$, $y \in \mathbb{R}$

$$f_\theta(y) = \exp\left(\frac{y\cdot\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

(ϕ: dispersion parameter)

for some known function $b(\cdot)$ and $c(\cdot,\cdot)$

* If ϕ is known: this is a one-parameter exponential family with θ being canonical parameter

* If ϕ is unknown, this may/may not be a two-parameter exponential family

- In this class, we always assume that ϕ is known

$$f_\theta(y) = \exp\left(\frac{y\cdot\theta - b(\theta)}{\phi}\right)\cdot h(y), \quad h(y) = e^{c(y,\phi)}$$

	Normal	Poisson	Bernoulli
Notation	$N(\mu,\sigma^2)$	$P(\mu)$	$B(\phi)$
Range of y	$(-\infty,\infty)$	$[0,\infty)$	$\{0,1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1+e^\theta)$
$c(y,\phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$-\log y!$	0

Likelihood

Let $l(\theta) = \log f_\theta(Y)$ denote the log likelihood function. The mean $\mathbb{E}[Y]$ and variance $\mathrm{var}(Y)$ can be derived from following:

① First identity $\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = 0$

Second identity $\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right] + \left(\mathbb{E}\left[\frac{\partial l}{\partial \theta}\right]\right)^2 = 0$

② Expected value proof: Note that $l(\theta) = \frac{Y\cdot\theta - b(\theta)}{\phi} + c(Y;\phi)$

Therefore: $\frac{\partial l}{\partial \theta} = \frac{Y - b'(\theta)}{\phi}$

It yields: $0 = \mathbb{E}\left[\frac{\partial l}{\partial \theta}\right] = \frac{\mathbb{E}[Y] - b'(\theta)}{\phi}$

Which leads to: $\mathbb{E}[Y] = b'(\theta)$

③ Variance proof: We have: $\frac{\partial^2 \ell}{\partial \theta^2} + \left(\frac{\partial \ell}{\partial \theta}\right)^2 = -\frac{b''(\theta)}{\phi} + \left(\frac{Y - b'(\theta)}{\phi}\right)^2$

from previous result: $\frac{Y - b'(\theta)}{\phi} = \frac{Y - \mathbb{E}(Y)}{\phi}$ ($\mathbb{E}[Y]$)

together, with second identity, this yields:

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\mathrm{var}(Y)}{\phi^2}$$

which leads to:

$$\mathrm{var}(Y) = b''(\theta) \cdot \phi$$

Lecture 22: GLM - Link function and the canonical link function

Link function

- β is the parameter of interest, needs to appear somewhere in likelihood function, to enable usage of maximum likelihood
- A link function g relates the linear predictor $X^T\beta$ to the mean parameter $\mu(x)$

$$X^T\beta = g(\mu(x))$$

- g is required to be monotonically increasing and differentiable

$$\mu(x) = g^{-1}(X^T\beta)$$

- Examples:
  - ① Linear model: $g(\cdot) = \text{identity}$
  - ② Poisson data: Suppose $Y|X \sim \text{Poisson}(\mu(x))$

    $\mu(x) > 0$

    $\log(\mu(x)) = X^T\beta$

    * In general, a link function for count data should map $(0, \infty)$ to $\mathbb{R}$. The log-link is a natural one
  - ③ Bernoulli / Binomial data:

    $0 < \mu < 1$

    g should map $(0, 1)$ to $\mathbb{R}$

    2 choices: logit: $\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = X^T\beta$ (natural)

    probit: $\Phi^{-1}(\mu(x)) = X^T\beta$ (Φ^{-1}: normal cdf inverse)

Canonical link

- The function g that links the mean $\mu(x)$ to the canonical parameter θ is called canonical link:

$$g(\mu(x)) = \theta$$

- Since $\mu(x) = b'(\theta)$, the canonical link is given by:

$$g(\mu(x)) = (b')^{-1}(\mu(x))$$

- If $\phi > 0$, the canonical link function is strictly increasing.
  - * Proof: $g \uparrow$ str. $\Leftrightarrow g^{-1} \uparrow$ str. $\Leftrightarrow (g^{-1})' > 0 \Leftrightarrow b'' > 0 \Leftrightarrow \text{var} > 0,\ \phi > 0$

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	μ
Poisson	$\exp(\theta)$	$\log\mu$
Bernoulli	$\log(1+e^\theta)$	$\log\frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

Model and notation

① Let $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots n$, be independent r.v. pairs s.t. the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family

$$f_{\theta_i}(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

② $Y = (Y_1, \dots Y_n)^T$ $\quad$ $\mathbb{X} = (X_1, \dots X_n)^T = \begin{pmatrix} -X_1^T- \\ -X_2^T- \\ \vdots \\ -X_n^T- \end{pmatrix} \in \mathbb{R}^{n\times p}$

③ Here the mean $\mu_i = \mathbb{E}[Y_i \mid X_i]$ is related to the canonical parameter θ_i via:

$$\mu_i = b'(\theta_i)$$

and μ_i depends linearly on the covariates through a link function g:

$$g(\mu_i) = X_i^T\beta$$

Back to β

- Given a link function g, note the following relationship between β and θ:

$$\theta_i = (b')^{-1}(\mu_i)$$
$$= (b')^{-1}(g^{-1}(x_i^T\beta)) \equiv h(x_i^T\beta)$$

where h is defined as:

$$h = (b')^{-1} \circ g^{-1} = (g \circ b')^{-1}$$

- Remark: if g is canonical link function, h is the identity ($g = b'^{-1}$)

Log-likelihood

- The log-likelihood is given by:

$$\ln(Y, X, \beta) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} + \text{constant}$$
$$= \sum_i \frac{Y_i h(x_i^T\beta) - b(h(x_i^T\beta))}{\phi} + \text{constant}$$

- Note that when using canonical link function, we obtain simpler expression:

$$\ln(Y, X, \beta) = \sum_i \frac{Y_i x_i^T\beta - b(x_i^T\beta)}{\phi} + \text{constant}$$

Strictly concavity:

- The log-likelihood is strictly concave (when $\text{rank}(X) = p$) using the canonical function when $\phi > 0$
- As a consequence, MLE is unique
- If another parameterization is used, the likelihood function may not be strictly concave, leading to several local maxima

Concluding remarks

- Maximum likelihood for Bernoulli Y and the logit link is called logistic regression
- In general, there is no closed form for the MLE and we have to use optimization algorithms (e.g. gradient descent)
- The asymptotic normality of MLE also applies to GLMs

Recitation: Hypothesis tests for logistic regression

1. Set up:

$Y_i \in \{0, 1\}$ responses $X_i \in \mathbb{R}^p$ covariates (X_i, Y_i) pairs independent $Y_i | X_i = x_i \sim \text{Bernoulli}(p_i)$

1.1 Rewrite pmf into canonical exponential

$$\mathbb{P}(Y_i = y_i | p_i) = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0 \end{cases} = p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= \exp(y_i \log p_i + (1-y_i) \log(1-p_i)) = \exp\left(y_i \underbrace{\log \frac{p_i}{1-p_i}}_{\theta_i} + \log(1-p_i)\right)$$

canonical parameter

$$\theta_i = \log \frac{p_i}{1-p_i} \iff p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \dashrightarrow = \exp(y\theta_i - \log(1+e^{\theta_i}))$$

1.2 Using canonical link

form: $f_{\theta_i}(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y, \phi)\right)$

Canonical link: $g(\mu(x_i)) = \theta_i$

Since $\mu(x_i) = b'(\theta)$, $g(\mu(x_i)) = (b'^{-1})(\mu(x_i))$

} derive form of canonical link

Link: $x^T\beta = g(\mu(x))$ (General link)

Since $\mu(x_i) = b'(\theta) \iff \theta = (b'^{-1})(\mu(x_i))$

$$\Downarrow$$

$$\theta = (b'^{-1})(\mu(x_i)) = (b'^{-1})(g^{-1}(x_i^T\beta))$$

$$\Downarrow \text{ when using canonical link:}$$

$$\theta = (b'^{-1})(g^{-1}(x_i^T\beta)) = x_i^T\beta$$

$g^{-1} = (b'^{-1})^{-1}$

Therefore: $\mathbb{P}(Y_i = y_i | x_i, \beta) = \exp(y_i x_i^T\beta - \log(1+e^{x_i^T\beta}))$

1.3 Test setup updated with the use of canonical link

$x_i \sim D$ iid (some distribution) $Y_i | x_i, \beta \sim \text{Bernoulli}\left(\frac{e^{x_i\beta}}{1+e^{x_i\beta}}\right)$

Log Likelihood: $\ln(\beta | X, Y) = \sum_{i=1}^{n} [Y_i x_i^T\beta - \log(1+e^{x_i^T\beta})]$

$\hat{\beta} = \arg\max_\beta \ln(\beta | X, Y)$

Goal: Test $H_0: \beta_j = \beta_j^0$ v.s. $H_1: \beta_j \neq \beta_j^0$