Probability and Statistics 401-2604

Overview by Sara van de Geer

Version 29.5.2017

AD: the book of Anirban DasGupta
Fundamentals of Probability: A First course (Springer 2010)
LN: the lecture notes of Föllmer and Künsch, later revised by Joseph Teichmann
Wahrscheinlichkeitsrechnung und Statistik (2017)
JC: the book of John Rice
Mathematical Statistics and Data Analysis (Duxbury Press, 1995)

Contents

[1] from page 80 onwards it is not part of the exam

Overview of definitions and results from probability

Countable sample space

Let Ω be countable.

AD Definition 1.2 *P is a probability measure on Ω if*
(a) $P(A) \geq 0$ for all $A \subset \Omega$,
(b) $P(\Omega) = 1$,
(c) if $A_1, A_2, \ldots$ are pairwise disjoint then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

("countable additivity" or "σ-additivity").

AD Theorem 1.1 ("monotone convergence") *Let $A_1 \subset A_2 \subset\uparrow A$. Then*

$$\lim_{n\to\infty} P(A_n) = P(A).$$

Proof. Use Definition 1.2 (in particular the σ-additivity). □

Inclusion/exclusion formula: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

AD Theorem 1.3 ("Bonferroni bound")

$$P(\cap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} (1 - P(A_i)).$$

Proof. Use Definition 1.2. □

AD Definition 3.1 *Let $A \subset \Omega$ and $B \subset \Omega$ with $P(B) > 0$. The conditional probability of A given B is*

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

AD Theorem 3.1 ("multiplication rule")

$$P(A \cap B) = P(A|B)P(B).$$

Proof. Use Definition 3.1. □

Definition *$A_1, A_2, \cdots$ form a partition of Ω if they are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = \Omega$.*

AD Theorem 3.1 ("law of total probability") *Let $A_1, A_2, \cdots$ be a partition of Ω with $P(A_i) > 0$ for all i. Then for any $B \subset \Omega$*

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

Proof. Write $P(B) = \sum_{i=1}^{\infty} P(B \cap A_i)$. □

AD Definition 3.2 *A and B are independent if*

$$P(A \cap B) = P(A)P(B).$$

AD Definition 3.3 *$A_1, A_2, \cdots$ are independent if*

$$P(\cap_{j \in J} A_j) = \prod_{j \in J} P(A_j) \ \forall \ J \subset \{1, 2, \ldots\}.$$

Bayes rule:

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}, \ P(A) > 0, \ P(B) > 0.$$

Corollary

$$\underbrace{\frac{P(B|A)}{P(B^c|A)}}_{\text{posterior odds}} = \underbrace{\frac{P(A|B)}{P(A|B^c)}}_{\text{likelihood ratio}} \underbrace{\frac{P(B)}{P(B^c)}}_{\text{prior odds}} .$$

AD Theorem 3.3 ("Bayes' Theorem") *Let $A_1, A_2, \cdots$ be a partition of Ω with $P(A_i) > 0$ for all i, and let $P(B) > 0$. Then*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

Proof. Follows from Bayes' rule. □

Decoding example (LN Example 2.17) (using random variables notation)
Let $Y \in \{1, \ldots, I\}$ be the signal sent.
Let $X \in \{1, \ldots, J\}$ be the signal received.
We are given $P(Y = i)$, $\forall \ i$ and $P(X = j|Y = i)$, $\forall \ i, j$. Let $\phi(X) \in \{1, \ldots, I\}$ be the decoder. Then

$$\begin{aligned} P(\text{signal correctly decoded}) &= P(Y = \phi(X)) \\ &= \sum_j P(Y = \phi(j), X = j) \\ &= \sum_j P(Y = \phi(j)|X = j)P(X = j). \end{aligned}$$

The optimal decoder ϕ_{opt} maximizes P(signal correctly decoded). It follows that

$$\phi_{\text{opt}}(j) = \arg\max_i P(Y = i|X = j), \ j = 1, \ldots J :$$

$$
\begin{aligned}
P(Y = \phi(X)) &= \sum_j P(Y = \phi(j) | X = j) P(X = j) \\
&\leq \sum_j \max_i P(Y = i | X = j) P(X = j).
\end{aligned}
$$

By Bayes rule for all j

$$
\phi_{\text{opt}}(j) = \arg\max_i P(Y = i | X = j) = \arg\max_i P(X = j | Y = i) P(Y = i).
$$

Discrete random variables and expectation

AD Definition 4.1 *Let Ω be countable. A random variable X is a mapping*

$$X: \ \Omega \to \mathbb{R}.$$

Then $\{X(\omega): \ \omega \in \Omega\}$ is also countable. We say that X is a discrete random variable.

AD Definition 4.2 *Let $X: \ \Omega \to \{x_1, x_2, \ldots\}$ be a discrete random variable. The probability mass function (pmf) of X is*

$$p(x) := P(X = x) = P(\omega: \ X(\omega) = x), \ x \in \mathbb{R}.$$

We often write $p =: p_X$.

AD Definition 4.3 *The cumulative distribution function (CDF) of $X \in \mathbb{R}$ is*

$$F(x) := P(X \leq x), \ x \in \mathbb{R}.$$

We often write $F =: F_X$.

AD Theorem 4.1 *The function F is a CDF iff*
(a) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$,
(b) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$,
(c) $\lim_{x\downarrow a} F(x) = F(a)$,
(d) F is increasing.

Proof of F CDF $\Rightarrow$ (c). This follows from monotone convergence (Theorem 1.1). □

Remark If $X \in \{x_1, x_2, \cdots\}$ is a discrete random variable, its CDF is a step-function (a piecewise constant function which jumps at x_i with jump size $p(x_i)$, $i = 1, 2, \ldots$).

AD Definition 4.6 *Let X and Y be two discrete random variables (defined on Ω). Then X and Y are called independent if*

$$P(X = x, Y = y) = P(X = x)P(Y = y), \ \forall \ (x, y) \in \mathbb{R}^2.$$

AD Theorem 4.2 *Let g and h be two real-valued functions on $\mathbb{R}$. Then:*
X and Y independent $\Rightarrow$ $g(X)$ and $h(Y)$ independent.

Definition
The random variables $X_1, \ldots, X_n$ are called independent identically distributed (i.i.d.) if
- $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n) \ \forall \ (x_1, \ldots, x_n) \in \mathbb{R}^n$ (i.e., $X_1, \ldots, X_n$ are independent)
- $P(X_i = \cdot) =: F(\cdot)$ is the same for all i (i.e., $X_1, \ldots, X_n$ are identically distributed).

AD Definition 4.17 *The expectation of a discrete random variable X is*

$$EX := \sum_x xp(x) := \mu.$$

Linearity of expectation:
For constants a and b, we have $E(aX + bY) = aEX + bEY$.

AD Proposition (Change of variable) *Let $g : \mathbb{R} \to \mathbb{R}$ be some function. Then $Eg(X) = \sum_x g(x)p(x)$.*

Proof. Write $Y = g(X)$. Then the pmf of Y is $p_Y(y) = \sum_{x\colon\, g(x)=y} p(x)$. Hence

$$EY = \sum_y yp_Y(y) = \sum_y \sum_{x\colon\, g(x)=y} yp(x) = \sum_y \sum_{x\colon\, g(x)=y} g(x)p(x) = \sum_x g(x)p(x).$$

□

AD Theorem 4.3 *X and Y independent $\Rightarrow EXY = EXEY$.*

Proof.

$$\begin{aligned} EXY &= \sum_x \sum_y xyP(X = x, Y = y) = \sum_x \sum_y xyP(X = x)P(Y = y) \\ &= \sum_x xP(X = x) \sum_y yP(Y = y) = EXEY. \end{aligned}$$

□

Definition *Let $A \subset \Omega$. The indicator function of A is*

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}, \ \omega \in \Omega.$$

Proposition *For $X := 1_A$ we have $EX = P(A)$.*

Proof. $EX = 1 \times P(X = 1) + 0 \times P(X = 0) = P(X = 1) = P(A)$. □

AD Theorem 4.4 ("partial integration") *Suppose $X \in \{0, 1, 2, \ldots\}$. Then*

$$EX = \sum_{n=0}^{\infty} P(X > n).$$

Proof.

$$\begin{aligned} \sum_{n=0}^{\infty} P(X > n) &= \sum_{n=0}^{\infty} \sum_{k=n+1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} P(X = k) \\ &= \sum_{k=1}^{\infty} kP(X = k) = EX. \end{aligned}$$

□

Variance and weak law of large numbers (LLN, discrete case)

AD Definition 4.9 *Let $EX := \mu$. The variance of X is*

$$\mathrm{Var}(X) := E(X-\mu)^2.$$

Note: $E(X-\mu)^2 = EX^2 - \mu^2$.

Proposition
(a) $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$,
(b) $\mathrm{Var}(X+c) = \mathrm{Var}(X)$,
(c) $\mathrm{Var}(X) = 0 \Leftrightarrow P(X=\mu) = 1$ (where $\mu := EX$).

Proof. Use Definition 4.9. □

Theorem ("Jensen's inequality", see also Section 7.8 in AD) *Let $g: \mathbb{R} \to \mathbb{R}$ be convex. Then $Eg(X) \geq g(EX)$.*

Proof for the case X discrete. Let $X \in \{x_1, x_2, \ldots\}$ and write $p_i := p(x_i)$, $i = 1, 2, \ldots$. Then $EX = \sum_{i=1}^\infty x_i p_i$ is a convex combination of $x_1, x_2, \ldots$, so by convexity of g

$$g(\sum_{i=1}^\infty x_i p_i) \leq \sum_{i=1}^\infty g(x_i)p_i.$$

□

Corollary *$EX^2 \geq (E|X|)^2$.*

AD Definition 4.10 *The k-the moment of X is EX^k ($k \in \mathbb{N}$).*

Note Jensen's inequality $\Rightarrow E|X|^k \geq (E|X|)^k$, $k \geq 1$.

AD Theorem 4.5 *Let X and Y be independent. Then*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof. Assume without loss of generality that $EX = EY = 0$. Then

$$\mathrm{Var}(X+Y) = E(X+Y)^2 = EX^2 + EY^2 + 2EXY.$$

We have by Theorem 4.3 that $EXY = EXEY = 0$. Moreover, $EX^2 = \mathrm{Var}(X)$ and $EY^2 = \mathrm{Var}(Y)$. □

Extension Let $X_1, \ldots, X_n$ be independent. Then

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

Corollary *Let $X_1, \ldots, X_n$ be i.i.d. with $EX_1 =: \mu$ and $\mathrm{Var}(X_1) =: \sigma^2$. Write their average as*

$$\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i.$$

Then

$$E\bar{X} = \mu, \ \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

AD Theorem 4.6 ("Chebyshev's inequality") *Let $g : \mathbb{R} \to [0, \infty)$ be an increasing function. Then for any constant c such that $g(c) > 0$ we have*

$$P(X \geq c) \leq \frac{Eg(X)}{g(c)}.$$

Proof.

$$Eg(X) = \sum_x g(x)p(x) \geq \sum_{x \geq c} g(x)p(x) \geq g(c) \sum_{x \geq c} p(x) = g(c)P(X \geq c).$$

□

Corollary *For all $c > 0$*

$$P(|X - EX| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

AD Theorem 4.7 ("<u>(Weak) Law of Large Numbers (LLN)</u>")
Let $X_1, \ldots, X_n, \cdots$ be i.i.d. with $EX_1 =: \mu$ and $\mathrm{Var}(X_1) =: \sigma^2$. Write the average of the first n as

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then for all $\epsilon > 0$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Proof.

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

□

Probability and moment generating functions (discrete case)

AD Definition 5.1 *Let $X \in \{0, 1, 2, \ldots\}$. The probability generating function (pgf) of X is*

$$G: \; s \mapsto Es^X$$

(provided the expectation exists). We often write $G =: G_X$.

AD Theorem 5.1 Assume $G(s) < \infty$ for all s in some open neighbourhood of zero. Then for all $k \in \{0, 1, \ \ldots\}$
a) $P(X = k) = \frac{G^{(k)}(0)}{k!}$,
b) if $\lim_{s \uparrow 1} G^{(k)}(s) < \infty$ then $G^{(k)}(1) = EX(X-1) \cdots (X-k+1)$.

AD Theorem 5.2
$X_1, \ldots, X_n$ independent $\Rightarrow G_{\sum_{i=1}^n X_i} = \prod_{i=1}^n G_{X_i}$.

Proof.

$$Es^{\sum_{i=1}^n X_i} = E \prod_{i=1}^n s^{X_i} = \prod_{i=1}^n Es^{X_i},$$

where we invoked Theorems 4.2 and 4.3. □

AD Theorem 5.3 *If $G_X(s) = G_Y(s)$ for all s in an open neighbourhood of zero, then X and Y have the same distribution.*

AD Definition 5.3 *Let $X \in \mathbb{R}$. The moment generating function (mgf) of X is*

$$\Psi: \; t \mapsto E\mathrm{e}^{tX}$$

(provided the expectation exists). We often write $\Psi =: \Psi_X$.

AD Theorem 5.4 *Suppose $\Psi(t)$ exists for all t in an open neighbourhood U of zero. Then*
a) $\Psi^{(k)}(0) = EX^k$, $k \in \{0, 1, 2, \ldots\}$,
b) $\Psi_X(t) = \Psi_Y(t)$ for all $t \in U \Rightarrow X$ and Y have the same distribution,
c) $X_1, \ldots, X_n$ independent $\Rightarrow \Psi_{\sum_{i=1}^n X_i} = \prod_{i=1}^n \Psi_{X_i}$.

Hypergeometric distribution

AD Theorem 6.6 *Let X have the hypergeometric distribution:*

$$P(X = x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}.$$

Then for $R = R_N$ and $R_N/N \to p$, $0 < p < 1$,

$$\lim_{N\to\infty} P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

In other words, the hypergeometric distribution can then be approximated by the binomial distribution.

Proof. Use Stirling's formula. □

Distribution of sums of discrete random variables: some special cases

AD Theorem 6.12 *Let X and Y be independent.*
a) $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Bin}(m,p) \Rightarrow X+Y \sim \mathrm{Bin}(n+m,p)$,
b) (Negative Binomial) $X \sim \text{Neg. Bin}(r,p)$, $Y \sim \text{Neg. Bin}(s,p) \Rightarrow X+Y \sim \text{Neg. Bin}(r+s,p)$,
c) $X \sim \mathrm{Poisson}(\lambda)$, $Y \sim \mathrm{Poisson}(\mu) \Rightarrow X+Y \sim \mathrm{Poisson}(\lambda+\mu)$.

Proof. Either directly:

$$P(X+Y=z) = \sum_y P(X=z-y)P(Y=y),$$

or use moment generating functions. □

General sample space

Let Ω be some sample space and $\mathcal{A}$ a collection of subsets of Ω.

LN Definition 3.1
1) The collection $\mathcal{A}$ is a σ-algebra if
- $\Omega \in \mathcal{A}$,
- $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$,
- $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{A}$ ("σ-additivity").
Then $(\Omega, \mathcal{A})$ is called a measurable space.
2) The map $P: \mathcal{A} \to [0,1]$ is a probability measure (probability distribution) if
- $P(\Omega) = 1$,
- $A_1, A_2, \ldots \in \mathcal{A}$ mutually disjoint $\Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
("countable subadditivity") (compare AD Definition 1.2).
3) The triple $(\Omega, \mathcal{A}, P)$ is called a probability space.

Definition *Let $\mathcal{A}_0$ be a collection of subsets of Ω. The σ-algebra generated by $\mathcal{A}_0$ is*

$$\mathcal{A} := \sigma(\mathcal{A}_0) := \cap\{\mathcal{B} : \ \mathcal{B} \supseteq \mathcal{A}_0, \ \mathcal{B} \ \sigma\text{−algebra}\}.$$

Definition *Let $\Omega := \mathbb{R}$ and $\mathcal{B}$ be the σ-algebra generated by the collection $\mathcal{A}_0 := \{(a,b] : \ a < b\}$ of all intervals. Then $\mathcal{B}$ is called the Borel σ-algebra.*

Definition *Let $\mathcal{B}$ be the Borel σ-algebra and $P([a,b]) := b - a$ for $0 \le a \le b \le 1$. Then P is called the Lebesgue measure on $[0,1]$.*

LN Theorem 3.1 ("monotone convergence") *Let $B_1 \subset B_2 \subset \cdots \uparrow B = \cup_{n=1}^{\infty} B_n$. Then $\lim_{n\to\infty} P(B_n) = P(B)$ (see also AD Theorem 1.1).*

Corollary *Let $A_1 \supset A_2 \supset \cdots \downarrow A = \cap_{n=1}^{\infty} A_n$. $\lim_{n\to\infty} P(A_n) = P(A)$.*

Note Consider $(\mathbb{R}, \mathcal{A}, P)$ with $\mathcal{A}$ the Borel σ-algebra on $\mathbb{R}$. Define

$$F(x) := P((-\infty, x]), \ x \in \mathbb{R}.$$

By the monotone convergence theorem, for all x,

$$\lim_{n\to\infty} F(x + 1/n) = F(x)$$

and

$$\lim_{n\to\infty} F(x - 1/n) = P((-\infty, x)) =: F(x-).$$

We say that the CDF F is càdlàg (continue à droite, limite à gauche). (Compare AD Theorem 4.1.)
We have:
F is a CDF $\Leftrightarrow$ F is càdlàg and $\uparrow$, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.

Notation

$$\begin{aligned} A_\infty &:= \cap_n \cup_{k \ge n} A_k \\ &= \limsup_{n\to\infty} A_n \\ &= \infty \text{ many of the } A_k \text{ happen} \\ &= \{A_k \text{ i.o.}\},\ \text{i.o.} := \text{infinitely often}. \end{aligned}$$

$$\begin{aligned} & \cup_n \cap_{k \ge n} B_k \\ =\ & \liminf_{n\to\infty} B_n \\ =\ & \{B_k \text{ eventually}\}. \end{aligned}$$

Definition $A_1, A_2, \ldots$ *are called* *independent* *if*

$$P(\cap_{j\in J} A_j) = \prod_{j\in J} P(A_j)\ \forall\ J \subset \mathbb{N} \text{ finite.}$$

(Compare AD Definition 3.3.)

Borel-Cantelli Lemma *Let* $A_1, A_2, \ldots \in \mathcal{A}$.
1) $\sum_{k=1}^\infty P(A_k) < \infty \Rightarrow P(A_\infty) = 0$.
2) $\sum_{k=1}^\infty P(A_k) = \infty$ *&* $A_1, A_2, \ldots$ *independent* $\Rightarrow P(A_\infty) = 1$.

Proof. Let $B_n := \cup_{k\ge n} A_k$. Apply the monotone convergence theorem.
1)

$$P(A_\infty) = \lim_{n\to\infty} P(B_n) \le \lim_{n\to\infty} \sum_{k\ge n} P(A_k) = 0.$$

2)

$$P(A_\infty^c) = \lim_{n\to\infty} P(B_n^c) = \lim_{n\to\infty} \prod_{k\ge n} (1 - P(A_k)) = \lim_{n\to\infty} \prod_{k\ge n} \exp\Big[\log(1 - P(A_k))\Big]$$

$$\le \lim_{n\to\infty} \prod_{k\ge n} \exp\Big[-P(A_k)\Big] = \lim_{n\to\infty} \exp\Big[-\sum_{k\ge n} P(A_k)\Big] = 0.$$

□

Definition (LN Section 3.1.4) *Let* $\mathcal{B} := \sigma(\{(-\infty, b] :\ b \in \mathbb{R}^d\})$ [2] *be the Borel* σ*-algebra on* $\mathbb{R}^d$, $(\Omega, \mathcal{A})$ *be a measurable space and* $X :\ \Omega \to \mathbb{R}^d$. *Then* X *is called* *measurable* *if* $\{\omega :\ X(\omega) \in B\} \in \mathcal{A}$ *for all* $B \in \mathcal{B}$. *The map* X *is then called a (d-dimensional)* *random variable*.

[2] $(-\infty, b]$ is the set of all $x \in \mathbb{R}^d$ with $x_j \le b_j$ for all $j \in \{1, \ldots, d\}$.

Continuous random variables in $\mathbb{R}$

AD Definition 7.2 *The cumulative distribution function (CDF) of a random variable $X \in \mathbb{R}$ is*

$$F(x) := P(X \le x), \ x \in \mathbb{R}.$$

(Compare AD Definition 4.3.)

AD Definition 7.3 *$X \in \mathbb{R}$ and $Y \in \mathbb{R}$ are independent if*

$$P(X \le x, Y \le y) = P(X \le x) P(Y \le y) \ \forall (x, y) \in \mathbb{R}^2.$$

AD Definition 7.2 continued *The random variable X is called continuous if its CDF F is continuous.*

AD Definition 7.1 *The random variable X admits a (probability) density function (pdf) $f(\cdot)$ if its CDF $F(\cdot)$ can be written as*

$$F(x) = \int_{-\infty}^{x} f(t) dt \ \forall \ x.$$

Then X (or F) is called absolutely continuous.

Note At locations x where $f(\cdot)$ is continuous

$$f(x) = \frac{d}{dx} F(x).$$

Note The function f is a density iff
- $f \ge 0$,
- $\int_{-\infty}^{\infty} f(x) dx = 1$.

AD Definition 7.5 *The p-th quantile of a CDF F is*

$$F^{-1}(p) := \inf\{x : \ F(x) \ge p\}.$$

Then $F^{-1}(1/2)$ is a median.

AD Theorem 7.1
a) Let $\mu \in \mathbb{R}$ and $\sigma > 0$. If $f(\cdot)$ is a density then so is

$$f(x|\mu, \sigma) := \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right), \ x \in \mathbb{R}.$$

Then $\{f(\cdot|\mu, \sigma) : \ \mu \in \mathbb{R}, \ \sigma > 0\}$ is a location/scale family.
b) Let $f_1, \ldots, f_k$ be densities and let $p_i \ge 0$, $i = 1, \ldots, k$, and $\sum_{i=1}^{k} p_i = 1$. Then $\sum_{i=1}^{k} p_i f_i$ is a density, a so-called mixture density.

AD Definition 7.6 *The density f is symmetric around M if $f(M + x) = f(M - x) \ \forall \ x$. Important special case: $M = 0$. Then $f(x) = f(-x)$ and $F(x) = 1 - F(-x) \ \forall \ x$.*

AD Definition 7.7 *The density f is unimodal with maximum at M if $f(x) \uparrow$ for $x < M$ and $f(x) \downarrow$ for $x > M$.*

Functions of an (absolutely) continuous random variable in $\mathbb{R}$

AD Theorem 7.2 ("Jacobian") *Let $X \in \mathbb{R}$ and g be a real-valued strictly monotone and differentiable function, defined on some open interval S such that $P(X \in S) = 1$. Then $Y := g(X)$ has density*

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}, \ g^{-1}(y) \in S.$$

(Here $1/g'(g^{-1}(y)) = dg^{-1}(y)/dy$ is called the "Jacobian" .)

Proof. Say $g \uparrow$. Then

$$F_Y(y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)).$$

Differentiate to find the density of Y. □

Remark Also for g possibly not monotone, it is often feasible to first find the distribution function F_Y of $Y := g(X)$ and then differentiate to obtain the density f_Y.

Definition *Let $U \sim \text{Uniform}[0,1]$ and let F be a CDF. Then $F^{-1}(U)$ is called the quantile transformation of U.*

AD Theorem 7.4 *Let $U \sim \text{Uniform}[0,1]$. Then $X := F^{-1}(U)$ has CDF F.*

Proof. From AD Definition 7.5: $F^{-1}(u) = \inf\{x : \ F(x) \ge u\}$. Now check that

$$P(X \le x) = P(U \le F(x)) = F(x).$$

□

Expectation of (absolutely) continuous random variables

Note If integral limits are not specified it means the integral is over $\mathbb{R}$.

AD Definition 7.9 *If X has pdf f and $\int |x|f(x)dx < \infty$ then the expectation of X is*

$$EX := \int xf(x)dx.$$

Remark For arbitrary random variables: if X has CDF F and $\int |x|dF(x) < \infty$ then $EX = \int xF(x)$.

Linearity of expectation:
For constants a and b, we have $E(aX + bY) = aEX + bEY$.

AD Theorem 7.5 ("change of variable") *Let $g : \mathbb{R} \to \mathbb{R}$ (measurable). If $\int |g(x)|f(x)dx < \infty$ then $Eg(X) = \int g(x)f(x)dx$.*

Sketch of proof. Suppose g is strictly increasing and let $Y = g(X)$. Then invoking AD Theorem 7.2

$$EY = \int yf_Y(y)dy = \int y\frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}dy$$

$$= \int yf_X(g^{-1}(y))dg^{-1}(y) = \int g(x)f_X(x)dx.$$

□

AD Definition 7.10 *The k-th moment of X is EX^k ($k \in \mathbb{N}$). (This is as AD Definition 4.10, but now for the continuous case.) The variance of X is $\mathrm{Var}(X) = E(X - EX)^2$ (as in AD Definition 4.9 but now for the continuous case).*

Note: $E(X - \mu)^2 = EX^2 - \mu^2$ (as in the discrete case).

Proposition (as in the discrete case)
(a) $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$,
(b) $\mathrm{Var}(X + c) = \mathrm{Var}(X)$,
(c) $\mathrm{Var}(X) = 0 \Leftrightarrow P(X = \mu) = 1$ (where $\mu := EX$).

AD Theorem 7.7 ("partial integration") (a continuous version of AD Theorem 4.4) *Suppose $X \geq 0$ and EX exists. Then*

$$EX = \int_0^\infty (1 - F(x))dx.$$

Sketch of proof. Suppose X has density $f = F'$. Then by partial integration

$$EX = \int_0^\infty xf(x)dx = \int_0^\infty xdF(x) = -\int_0^\infty xd(1 - F(x))$$

$$= -x(1-F(x))|_{x=0}^{\infty} + \int_0^{\infty} (1-F(x))dx.$$

But $x(1-F(x))|_{x=0} = 0$ and

$$0 \le x(1-F(x)) \le \int_x^{\infty} uf(u)du \to 0,\ x \to \infty,$$

since $EX < \infty$. □

Moment generating functions (discrete or continuous case)

AD Definition 7.14 *Let $X \in \mathbb{R}$. The moment generating function (mgf) of X is*

$$\Psi: \ t \mapsto E\mathrm{e}^{tX}$$

(provided the expectation exists). We often write $\Psi =: \Psi_X$.

Theorem (as for the discrete case in AD Theorem 5.4) *Suppose $\Psi(t)$ exists for all t in an open neighbourhood U of zero. Then*
a) $\Psi^{(k)}(0) = EX^k$, $k \in \{0, 1, 2, \ldots\}$,
b) $\Psi_X(t) = \Psi_Y(t)$ for all $t \in U \Rightarrow X$ and Y have the same distribution,
c) $X_1, \ldots, X_n$ independent $\Rightarrow \Psi_{\sum_{i=1}^n X_i} = \prod_{i=1}^n \Psi_{X_i}$.

Note Let $Y := \mu + \sigma X$. Then $\Psi_Y(t) = \mathrm{e}^{\mu t}\Psi_X(\sigma t)$.

Example
Let $X \sim \mathcal{N}(0, 1)$. Then $\Psi_X(t) = \exp[t^2/2]$.
Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then $\Psi_Y(t) = \exp[\mu t + \sigma^2 t^2/2]$.

Jensen's inequality, Chebyshev's inequality, weak LLN, revisited

Theorem ("Jensen's inequality") *Let $g : \mathbb{R} \to \mathbb{R}$ be convex. Then $Eg(X) \geq g(EX)$.*

Proof For all constants a and all x it holds that $g(x) \geq g(a) + m(a)(x - a)$ where $m(a)$ is the slope of the line $l(x) := g(a) + m(a)(x - a)$ passing through $(a, g(a))$ that is below g. So we have

$$Eg(X) \geq g(a) + m(a)(EX - a).$$

Now take $a = EX$. □

Corollary $EX^2 \geq (E|X|)^2$.

Note Jensen's inequality ⇒ $E|X|^k \geq (E|X|)^k$, $k \geq 1$.

Theorem *Let X and Y be independent. Then*
Then
a) $EXY = EXEY$
b) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. The proof of a) is given in the lemma following AD Definition 12.3. The proof of b) then follows as in the discrete case (AD Theorem 4.5). □

Extension Let $X_1, \ldots, X_n$ be independent. Then

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Corollary *Let $X_1, \ldots, X_n$ be i.i.d. with $EX_1 =: \mu$ and $\mathrm{Var}(X_1) =: \sigma^2$. Write their average as*

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then

$$E\bar{X} = \mu, \ \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Theorem ("Chebyshev's inequality") (as AD Theorem 4.6 for the continuous case) *Let $g : \mathbb{R} \to [0, \infty)$ be an increasing function. Then for any constant c such that $g(c) > 0$ we have*

$$P(X \geq c) \leq \frac{Eg(X)}{g(c)}.$$

Proof for the absolutely continuous case. It boils down to replacing in the proof of Theorem 4.6 the sums by integrals and the pmf p by the pdf f:

$$Eg(X) = \int_x g(x) f(x) \geq \int_{x \geq c} g(x) f(x) \geq g(c) \int_{x \geq c} f(x) = g(c) P(X \geq c).$$

□

Corollary *For all $c > 0$*

$$P(|X - EX| \geq c) \leq \frac{\text{Var}(X)}{c^2}.$$

Theorem (“(Weak) Law of Large Numbers (LLN)”) (as AD Theorem 4.7 for the discrete case, now stated for the general case)
Let $X_1, \ldots, X_n, \cdots$ be i.i.d. with $EX_1 =: \mu$ and $\text{Var}(X_1) =: \sigma^2$. Write the average of the first n as

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then for all $\epsilon > 0$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Distribution of sums of continuous random variables: some special cases

Theorem *Let $X_1, \dots, X_n$ be i.d.d. copies of a random variable X.*
a) $X \sim \text{Exponential}(\lambda)$, $\Rightarrow \sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$,
b) $X \sim \mathcal{N}(\mu, \sigma^2)$, $\Rightarrow \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$,
c) $X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \sum_{i=1}^n X_i^2 \sim \chi^2$ with n degrees of freedom.

Limit theorems

Definition (Section 4.2 of LN) *A sequence of real-valued random variables Z_n converges in probability to Z (notation: $Z_n \to^P Z$) if for all $\epsilon > 0$*

$$\lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0.$$

It converges almost surely to Z (notation: $Z_n \to^{\text{a.s.}} Z$) if

$$P(\lim_{n\to\infty} Z_n = Z) = 1.$$

LN Lemma 4.1
i) $Z_n \to^{\text{a.s.}} Z \Rightarrow Z_n \to^P Z$.
ii) $\sum_n P(|Z_n - Z| > \epsilon) < \infty \; \forall \; \epsilon > 0 \Rightarrow Z_n \to^{\text{a.s.}} Z$.

Proof. Let $A_n := \{|Z_n - Z| > \epsilon\}$.
i) $\omega \in A_\infty$ implies $Z_n(\omega)$ does not converge to $Z(\omega)$. Therefore $P(A_\infty) = 0$. But, invoking monotone convergence,

$$P(A_\infty) = \lim_{n\to\infty} P(\cup_{k\geq n} A_k) \geq \lim_{n\to\infty} P(A_n).$$

ii) By the Borel-Cantelli Lemma $P(A_\infty) = 0$. But then

$$1 = P(A_\infty^c) = P(\lim_{k\to\infty} |Z_k - Z| \leq \epsilon).$$

□

LN Lemma 4.2 ("Strong Law of Large Numbers (LLN)") *Let $X_1, \ldots, X_n, \ldots$ be i.i.d. with $EX_1 =: \mu$ and $\text{Var}(X_1) =: \sigma^2 < \infty$. Denote the average of the first n by $\bar{X}_n := \sum_{i=1}^n X_i/n$. Then*

$$\bar{X}_n \to^{\text{a.s.}} \mu.$$

Proof. By Chebyshev's inequality, for all $\epsilon > 0$

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Hence

$$P(|\bar{X}_{n^2} - \mu| > \epsilon) \leq \frac{\sigma^2}{n^2\epsilon^2}.$$

By the Borel-Cantelli Lemma this gives

$$\bar{X}_{n^2} \to^{\text{a.s.}} \mu.$$

Define the sum $S_n := \sum_{i=1}^n X_i$.
∘ Suppose first $X_i \geq 0$ (almost surely $\forall \; i$). Then for $n^2 \leq k \leq (n+1)^2$

$$\frac{S_k}{k} \geq \frac{S_{n^2}}{k} \geq \frac{S_{n^2}}{n^2} \frac{n^2}{(n+1)^2} \to^{\text{a.s.}} \mu$$

and

$$\frac{S_k}{k} \le \frac{S_{(n+1)^2}}{k} \le \frac{S_{(n+1)^2}}{(n+1)^2} \frac{(n+1)^2}{n^2} \to^{\text{a.s.}} \mu.$$

∘ For general X_i write $X_i = X_i^+ - X_i^-$, where $X_i^+ := \max\{X_i, 0\}$ and $X_i^- := \max\{-X_i, 0\}$. □

AD Theorem 10.3 ("de Moivre-Laplace Local Limit Theorem") *Let* $X \sim \text{Binomial}(n,p)$ *where* $0 < p < 1$ *is fixed. Then for any fixed constant* C *and any* $k \in \{0, \ldots, n\}$ *such that* $|p - k/n| \le C$ *it holds that*

$$P(X = k) \sim \frac{1}{\sigma}\phi\biggl(\frac{k - \mu}{\sigma}\biggr) \ (n \to \infty),$$

where $\mu := np(= EX)$*, and* $\sigma^2 := np(1-p)(= \text{Var}(X))$*. Moreover,* ϕ *is the* $\mathcal{N}(0,1)$*-density.*

Sketch of Proof. Use Stirling's formula and the two-term Taylor expansion $\log(1 + x) \sim x - x^2/2 \ x \to 0$. □

AD Theorem 10.2 ("de Moivre-Laplace Central Limit Theorem (CLT)") *Let* $X \sim \text{Binomial}(n,p)$ *where* $0 < p < 1$ *is fixed. Then for all* $x \in \{1, \ldots, n\}$

$$P(X \le x) \sim \Phi\biggl(\frac{x - \mu}{\sigma}\biggr) \ (n \to \infty),$$

where $\mu := np(= EX)$*, and* $\sigma^2 := np(1-p)(= \text{Var}(X))$*. Moreover,* Φ *is the* $\mathcal{N}(0,1)$*-distribution function.*

Sketch of Proof. See AD Theorem 10.1. □

Remark The <u>continuity correction</u> is that instead of taking for $x \in \{0, 1, \ldots, n\}$

$$P(X \le x) \sim \Phi\biggl(\frac{x - \mu}{\sigma}\biggr)$$

one uses

$$P(X \le x) = P(X \le X + .5) \sim \Phi\biggl(\frac{x + .5 - \mu}{\sigma}\biggr).$$

AD Theorem 10.1("<u>Central Limit Theorem (CLT)</u>") *Let* $X_1, \ldots, X_n, \ldots$ *be i.i.d. with* $EX_1 =: \mu$ *and* $\text{Var}(X_1) =: \sigma^2 < \infty$*. Then with* $\bar{X}_n := \sum_{i=1}^n X_i/n$

$$\lim_{n\to\infty} P\biggl(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le t\biggr) = \Phi(t), \ \forall \ t,$$

where Φ *is the* $\mathcal{N}(0,1)$*-distribution function.*

Sketch of Proof. We consider the case where $\Psi_{X_1}(t)$ exists for all t in an open neighbourhood of zero. We will only show convergence of the moment generating function. The result then follows from a "continuity theorem for mgf's" (not shown). Without loss of generality we may assume $\mu = 0$ and $\sigma^2 = 1$. Then

$$\begin{aligned}\Psi_{X_1}(t) &\sim \underbrace{\Psi_{X_1}(0)}_{=1} + \underbrace{\dot{\Psi}_{X_1}(0)}_{=\mu=0} \frac{t}{\sqrt{n}} + \underbrace{\ddot{\Psi}_{X_1}(0)}_{=EX_1^2=\sigma^2=1} \frac{t^2}{2n} \\ &= 1 + \frac{t^2}{2n}.\end{aligned}$$

Moreover

$$\Psi_{\sqrt{n}\bar{X}_n}(t) = \Psi^n_{X_1}(t/\sqrt{n})$$

so that

$$\log \Psi_{\sqrt{n}\bar{X}_n}(t) = n \log \Psi^n_{X_1}(t/\sqrt{n}) \sim n \log(1 + t^2/(2n)) \sim t^2/2.$$

It follows that $\Psi_{\sqrt{n}\bar{X}_n}(t) \to \exp[t^2/2]$ which is the mgf of a $\mathcal{N}(0,1)$-variable. □

Multivariate discrete distributions

Let $X: \Omega \to \{x_1, x_2, \ldots\}$ and $Y: \Omega \to \{y_1, y_2, \ldots\}$ be two discrete random variables.

AD Definition 11.1/11.2 *The joint probability mass function (pmf) of* (X, Y) *is*

$$p(x, y) := P(X = x, Y = y),\ (x, y) \in \mathbb{R}^2.$$

The joint cumulative distribution function (CDF) is

$$F(x, y) := P(X \leq x, Y \leq y),\ (x, y) \in \mathbb{R}^2.$$

AD Definition 11.3 *The marginal pmf of* X *is*

$$p_X(x) = \sum_y p(x, y),\ x \in \mathbb{R}.$$

The marginal pmf of Y *is*

$$p_Y(y) = \sum_x p(x, y),\ y \in \mathbb{R}.$$

For a function $g: \mathbb{R}^2 \to \mathbb{R}$ *and for* $Z := g(X, Y)$ *the pmf of* Z *is*

$$p_Z(z) = \sum_{(x,y):\ g(x,y)=z} p(x, y),\ z \in \mathbb{R}.$$

AD Theorem 11.1 ("change of variable") *For a function* $g: \mathbb{R}^2 \to \mathbb{R}$,

$$Eg(X, Y) = \sum_{x,y} g(x, y)p(x, y).$$

Proof. Let $Z = g(X, Y)$. Then

$$EZ = \sum_z z p_Z(z) = \sum_z z \sum_{(x,y):\ g(x,y)=z} p(x, y)$$

$$= \sum_z \sum_{(x,y):\ g(x,y)=z} z p(x, y) = \sum_{x,y} g(x, y)p(x, y).$$

□

AD Definition 11.4 *For* $p_Y(y) > 0$ *the conditional distribution of* X *given* $Y = y$ *is*

$$p(x|y) := P(X = x|Y = y) = \frac{p(x, y)}{p_Y(y)}.$$

The conditional expectation of X *given* $Y = y$ *is*

$$E(X|Y = y) = \sum_x x p(x|y) =: h(y),$$

and we write $E(X|Y) := h(Y)$.

AD Proposition 11.1 *We have*

$$E\Big(Xg(Y)\Big|Y\Big) = g(Y)E(X|Y).$$

Proof.

$$E\Big(Xg(Y)\Big|Y = y\Big) = \sum_x xg(y)p(x|y) = g(y)\sum_x xp(x|y).$$

□

AD Theorem 11.3 ("$\underline{\text{iterated expectations}}$")

$$E\Big(E(X|Y)\Big) = EX.$$

Proof. Let $h(y) := E(X|Y = y)$. Then

$$Eh(Y) = \sum_y h(y)p_Y(y) = \sum_y \Big[\sum_x xp(x|y)\Big]p_Y(y)$$
$$= \sum_x x\sum_y p(x,y) = \sum_x xp_X(x).$$

□

Definition

$$\mathrm{Var}(X|Y = y) := E(X^2|Y = y) - \Big(E(X|Y = y)\Big)^2 =: \tilde{h}(y)$$

and

$$\mathrm{Var}(X|Y) := \tilde{h}(Y).$$

AD Theorem 11.4 ("iterated variance")

$$\mathrm{Var}(X) = \underbrace{E\mathrm{Var}(X|Y)}_{\text{"within"}} + \underbrace{\mathrm{Var}(E(X|Y))}_{\text{"between"}}.$$

Proof. We have

$$\mathrm{Var}(X|Y) = E(X^2|Y) - (E(X|Y))^2,$$

so by iterated expectations

$$E\mathrm{Var}(X|Y) = EX^2 - E(E(X|Y))^2.$$

Moreover, using iterated expectations once more

$$\mathrm{Var}(E(X|Y)) = E(E(X|Y))^2 - \Big(E(E(X|Y))\Big)^2 = E(E(X|Y))^2 - (EX)^2.$$

□

AD Example 11.18
a) Best constant predictor:

$$\arg\min_{c\in\mathbb{R}} E(Y-c)^2 = EY.$$

b) Best predictor given $X = x$:

$$\arg\min_{c\in\mathbb{R}} E\Big((Y-c)^2\Big|X=x\Big) = E(Y|X=x).$$

Hence

$$\min_{d:\ \mathbb{R}\to\mathbb{R}} E(Y-d(X))^2 = E(Y-E(Y|X))^2 = E\mathrm{Var}(Y|X).$$

c) Best linear predictor

$$\arg\min_{(a,b)^T\in\mathbb{R}^2} E(Y-(a+bX))^2 := \begin{pmatrix}\alpha\\ \beta\end{pmatrix},$$

where

$$\alpha = EY - \beta EX,\ \beta = \frac{EXY - EXEY}{\mathrm{Var}(X)}.$$

AD Definition 11.7 *The covariance between X and Y is*

$$\mathrm{Cov}(X,Y) = EXY - EXEY.$$

AD Example 11.18 c) continued.

$$\beta = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2}$$

$$E(Y-(\alpha+\beta X))^2 = \sigma_Y^2 - \frac{\mathrm{Cov}^2(X,Y)}{\sigma_X^2}.$$

AD Theorem 11.6
a) $\mathrm{Cov}(X,Y) = E(X-EX)(Y-EY)$.
b) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
c) $\mathrm{Cov}(aX+bY, cX+dY) = ac\mathrm{Var}(X) + (ad+bc)\mathrm{Cov}(X,Y) + bd\mathrm{Var}(Y)$, *and*

$$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i\neq j}\mathrm{Cov}(X_i,X_j).$$

d) X and Y independent $\Rightarrow \mathrm{Cov}(X,Y) = 0$.

Proof of c). Assume without loss of generality that $EX = EY = 0$ and $EX_i = 0$ for all i. Then

$$\text{Cov}(aX + bY, cX + dY) = E(aX + bY)(cX + dY)$$

and

$$\text{Var}(\sum_{i=1}^{n} X_i) = E(\sum_{i=1}^{n} X_i)^2.$$

Now remove the brackets and use linearity of expectation. □

Proof of d). See Theorem AD 4.4. □

AD Definition 11.8 *The correlation between X and Y is*

$$\rho_{XY} := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

AD Theorem 11.6 *It holds that $|\rho_{XY}| \leq 1$ and $|\rho_{XY}| = 1 \Leftrightarrow Y = \alpha + \beta X$ $(\exists\ (\alpha, \beta))$.*

Proof. Use AD Example 11.8 c) continued. □

Multivariate continuous distributions

AD Definition 12.1 *$(X, Y) \in \mathbb{R}^2$ has density $f(x, y)$, $(x, y) \in \mathbb{R}^2$, if for all $-\infty < a \leq b < \infty$ and $-\infty < c \leq d < \infty$ it holds that*

$$P(a \leq X \leq b,\ c \leq Y \leq d) = \int_c^d \int_a^b f(x, y)dxdy.$$

AD Definition 12.2 *If (X, Y) admits density f, the cumulative distribution function (CDF) of (X, Y) is*

$$F(x, y) := \int_{-\infty}^{y} \int_{\infty}^{x} f(s, t)dsdt,\ (x, y) \in R^2$$

and we have (for almost all (x, y))

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

The (marginal) density of X is then

$$f_X(x) = \int f(x, y)dy,\ x \in \mathbb{R},$$

and the (marginal) density of Y is

$$f_Y(y) = \int f(x, y)dx,\ y \in \mathbb{R}.$$

AD Proposition *X and Y independent iff $F(x, y) = F_X(x)F_Y(y)$ for all (x, y) iff $f(x, y) = f_X(x)f_Y(y)$ for (almost) all (x, y).*

AD Definition 12.3

$$Eg(X, Y) = \int g(x, y)f(x, y)dxdy.$$

Lemma *X and Y independent $\Rightarrow$ $EXY = EXEY$.*

Proof. This follows by replacing in the proof for the discrete case (AD Theorem 4.3) the sums by integrals and the pmf's by pdf's:

$$EXY = \int_y \int_x xyf(x, y)dxdy = \int_y \int_x xyf_X(x)f_Y(y)dxdy$$

$$= \int_x xf_X(x)dx \int_y yf_Y(y)dy = EXEY.$$

□

Definition of the bivariate normal distribution *Let U_1 and U_2 be independent and both $\mathcal{N}(0,1)$-distributed. Write*

$$U := \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \ X = AU + \mu,$$

where $\mu \in \mathbb{R}^2$ is a given vector and $A \in \mathbb{R}^{2\times 2}$ is a given non-singular matrix. Then X has a two-dimensional normal distribution with parameters (μ, Σ) where $\Sigma = AA^T$.

Note The Jacobian (see AD Theorem 13.3) of $u \mapsto x = Au + \mu$ is A^{-1} and we have $|\det(A^{-1})| = 1/\sqrt{\det(\Sigma)}$, $\Sigma = AA^T$. In the above definition

$$f_U(u) = \frac{1}{2\pi} \exp[-\|u\|^2/2], \ u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

where $\|u\|^2 = u_1^2 + u_2^2 = u^T u$. It follows (see AD Theorem 13.3) that

$$f_X(x) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp[-(x-\mu)^T \Sigma^{-1}(x-\mu)], \ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Moreover, we have $EX = \mu$ and for

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix},$$

it holds that $\text{Var}(X_1) = \sigma_1^2$, $\text{Var}(X_2) = \sigma_2^2$ and $\text{Cov}(X_1, X_2) = \sigma_{1,2}$.

Remark The definition of the d-dimensional normal distribution is: $X = AU + \mu$ with $U = (U_1, \ldots, U_d)^T$, $U_1, \ldots, U_d$ i.i.d. $\mathcal{N}(0,1)$, $\mu \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$.

Theorem $X \sim \mathcal{N}(\mu, \Sigma) \Rightarrow BX \sim \mathcal{N}(B\mu, B\Sigma B^T)$.

Proof. Follows from the definition of the bivariate (or multivariate) normal. □

Theorem *Let $X = (X_1, X_2)^T \sim \mathcal{N}(\mu, \Sigma)$. Then:*
X_1 and X_2 independent $\Leftrightarrow$ $\text{Cov}(X_1, X_2) = 0$.

Proof of ($\Leftarrow$). Since

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

we see that

$$f_X(x) = f_{X_1}(x_1) f_{X_2}(x_2) \ \forall \ x \in \mathbb{R}^2.$$

□

AD Example 12.16 Let X_1 and X_2 be independent and $X_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$. Define $Z_1 := X_1 + X_2$ and $Z_2 := X_1 - X_2$. Then $Z := (Z_1, Z_2)^T$ is bivariate normal and $\text{Cov}(Z_1, Z_2) = 0$ so that Z_1 and Z_2 are independent.

AD Theorem 12.4 *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Define the sample mean $\bar{X} := \sum_{i=1}^n X_i/n$ and the sample variance $S^2 := \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. Then $\bar{X}$ and S^2 are independent.*

Proof for $n = 2$. It follows from AD Theorem 12.4:

$$\bar{X} = \frac{X_1 + X_2}{2},\ S^2 = \frac{(X_1 - X_2)^2}{2}.$$

□

AD Definition 12.6 *Let(X, Y) have pdf $f(x, y)$, $(x, y) \in \mathbb{R}^2$. For $f_Y(y) > 0$ the conditional density of X given $Y = y$ is*

$$f(x|y) := \frac{f(x, y)}{f_Y(y)},\ x \in \mathbb{R}.$$

The conditional expectation of X given $Y = y$ is

$$E(X|Y = y) := \int x f(x|y) dx =: h(y)$$

and

$$E(X|Y) := h(Y).$$

The conditional variance of X given $Y = y$ is

$$\mathrm{Var}(X|Y = y) = E(X^2|Y = y) - \Big(E(X|Y = y)\Big)^2 =: \tilde{h}(y)$$

and

$$\mathrm{Var}(X|Y) := \tilde{h}(Y).$$

Many results for the discrete case carry over to the continuous case and definitions can be re-used. In particular:
◦ Iterated expectations: $EE(X|Y) = EX$.
◦ Iterated variance: $\mathrm{Var}(X) = E\mathrm{Var}(X|Y) + \mathrm{Var}(E(X|Y))$.
◦ Best constant predictor: $\arg\min_{c\in\mathbb{R}} E(Y - c)^2 = EY$.
◦ Best predictor given X: $\min_{d:\ \mathbb{R}\to\mathbb{R}} E(Y - d(X))^2 = E(Y - E(Y|X))^2$.
◦ Best linear predictor $\arg\min_{(a,b)^T\in\mathbb{R}^2} E(Y - (a + bX))^2 := (\alpha, \beta)^T$, where $\alpha = EY - \beta EX,\ \beta = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$.
◦ The covariance between X and Y is $\mathrm{Cov}(X, Y) = EXY - EXEY$.
◦ $\mathrm{Cov}(X, Y) = E(X - EX)(Y - EY)$.
◦ $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
◦ $\mathrm{Cov}(aX + bY, cX + dY) = ac\mathrm{Var}(X) + (ad + bc)\mathrm{Cov}(X, Y) + bd\mathrm{Var}(Y)$,
◦ $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j)$.
◦ X and Y independent $\Rightarrow \mathrm{Cov}(X, Y) = 0$.
◦ The correlation between X and Y is $\rho_{XY} := \mathrm{Cov}(X, Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$.
◦ $|\rho_{XY}| \leq 1$ and $|\rho_{XY}| = 1 \Leftrightarrow Y = \alpha + \beta X\ (\exists\ (\alpha, \beta))$.
◦ Bayes formula: let $\tilde{f}(y|x)$ be the conditional density of Y given $X = x$. Then

$$\tilde{f}(y|x) = \frac{f(x|y) f_Y(y)}{f_X(x)}.$$

Remark In the statistics part we use a different notation. We let $Y := \theta$, $y =: \vartheta$, $f_Y(y) =: w(\vartheta)$ and $p(x|\vartheta)$ be the conditional pmf or pdf of X given $\theta = \vartheta$ and we write

$$w(\vartheta|x) = \frac{p(x|\vartheta)w(\vartheta)}{p(x)}.$$

where $p(x) = \int p(x|\vartheta)w(\vartheta)d\vartheta$. In that context θ can also be a discrete random variable. Then $w(\vartheta)$ is the pmf of θ and $p(x) = \sum_{\vartheta} p(x|\vartheta)w(\vartheta)$.

Example Let Y and Z be independent, $Y \sim \mathcal{N}(\nu, \tau^2)$ and $Z \sim \mathcal{N}(0, \sigma^2)$. Define $X := Y + Z$. Then $X \sim \mathcal{N}(\nu, \tau^2 + \sigma^2)$ and $X|Y \sim \mathcal{N}(Y, \sigma^2)$. Moreover

$$Y|X \sim \mathcal{N}\left(\frac{X\tau^2 + \nu\sigma^2}{\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\right).$$

Convolutions and transformations

AD Theorem 13.1 *Let $(X, Y) \in \mathbb{R}^2$ have density $f(x, y)$, $(x, y) \in \mathbb{R}^2$ and let $Z := X + Y$. Then*

$$f_Z(z) = \int f(z - y, y)dy, \ z \in \mathbb{R}.$$

In particular, if X and Y are independent

$$f_Z(z) = \int f_X(z - y) f_Y(y)dy, \ z \in \mathbb{R}.$$

Definition *Let $X_1, \ldots, X_n$ be i.i.d. with density f. The density of $X_1 + \cdots + X_n$ is called the (n-fold) convolution of f.*

AD Theorem 13.3 *Let $X = (X_1, \ldots, X_n)^T$ have density $f(x)$, $x \in \mathbb{R}^n$ and let $S \subset \mathbb{R}^n$ be some open set such that $P(X \in S) = 1$. Consider a function $g : \ S \to \mathbb{R}^n$ and define $U := g(X)$. Assume*
a) $g : \ S \to g(S) =: T$ is 1-1,
b) $h := g^{-1}$ is continuously differentiable,
c) $\det(J(u)) \neq 0$ where $J(u) := \partial h(u)/\partial u$ is the Jacobian ($u \in S$).
Then

$$f_U(u) = |\det(J(u))| f_X(h(u)), \ u \in T.$$

Standard distributions

Standard discrete distributions

1. Bernoulli distribution with success parameter $p \in (0,1)$. $X \in \{0,1\}$ and

$$P(X=1)=p, \quad EX=p, \quad \mathrm{Var}(X)=p(1-p).$$

2. Binomial distribution with n trials and success parameter $p \in (0,1)$. $X \in \{0,1,\dots,n\}$

$$P(X=k)=\binom{n}{k}\,p^k(1-p)^{n-k}, \quad k=0,1,\dots n,$$

$$EX=np, \quad \mathrm{Var}(X)=np(1-p).$$

3. Poisson distribution with parameter $\lambda > 0$. $X \in \{0,1,\dots\}$

$$P(X=k)=\frac{\lambda^k}{k!}\,e^{-\lambda}, \quad k=0,1,\dots,$$

$$EX=\lambda, \quad \mathrm{Var}(X)=\lambda.$$

Standard continuous distributions

4. Gaussian distribution with mean μ and variance σ^2. $X \in \mathbb{R}$,

$$f_X(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \ \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R}.$$

Denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$EX = \mu, \ \text{var}(X) = \sigma^2.$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Leftrightarrow \quad Z := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

$\mathcal{N}(0,1)$ is called the standard normal (or Gaussian).

5. The standard normal distribution function.

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-z^2/2} \, dz, \quad x \in \mathbb{R}.$$

Let Φ^{-1} be its inverse function. Then,

$$\Phi^{-1}(0.9) = 1.28, \quad \Phi^{-1}(0.95) = 1.64, \quad \Phi^{-1}(0.975) = 1.96.$$

6. Exponential distribution with parameter $\lambda > 0$. $X \in \mathbb{R}_+ := [0, \infty)$,

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x \geq 0.$$

$$EX = \lambda, \quad \text{Var}(X) = \lambda^2.$$

Note: in many textbooks λ is replaced by $1/\lambda$.

7. Gamma distribution with parameters α, λ. $X \in \mathbb{R}_+ := [0, \infty)$,

$$f_X(x) = \frac{1}{\lambda^\alpha \Gamma(\alpha)} \ x^{\alpha-1} \ e^{-x/\lambda}, \quad x \geq 0.$$

Here $\Gamma(\alpha)$ is the Gamma function and for integer values $\Gamma(m) = (m-1)!$.

$$EX = \alpha\lambda, \quad \text{Var}(X) = \alpha\lambda^2.$$

Note: in many textbooks λ is replaced by $1/\lambda$.

8. Beta distribution with parameters r, s. $X \in [0,1]$,

$$f_X(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \ x^{r-1} \ (1-x)^{s-1}, \quad x \in [0,1].$$

$$EX = \frac{r}{r+s}, \quad \text{Var}(X) = \frac{rs}{(r+s)^2 \ (1+r+s)}.$$

9. Chi-Square (χ^2) distribution.

   The χ^2 distribution with m degrees of freedom is the Gamma distribution with parameters $(m/2, 1/2)$. Denoted by $\chi^2(m)$. In particular,

$$X \sim \mathcal{N}(0,1) \quad \Rightarrow \quad X^2 \sim \chi^2(1),$$

$$X_j \sim \mathcal{N}(0,1),\ j = 1, \dots, m,\ \text{i.i.d.} \quad \Rightarrow \quad \sum_{j=1}^{m} X_j^2 \sim \chi^2(m),$$

10. Student distribution.

    If $Z \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(m)$, $Z \perp Y$, then,

$$T := \frac{Z}{\sqrt{Y/m}},$$

    has a student distribution with m degrees of freedom.

    Its density is given by

$$f_T(t) = \frac{\Gamma((m+1)/2)}{\sqrt{m\pi}\ \Gamma(m/2)} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2}, \quad t \in \mathbb{R}.$$

11. Studentizing. Let $\{X_i\}_{i=1}^n$ be i.i.d. with $\mathcal{N}(\mu, \sigma^2)$ distribution. Let $\overline{X}_n := \sum_{i=1}^n X_i/n$ and set

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

    Then, $\overline{X}_n$ and S_n^2 are independent and

$$\frac{\sqrt{n}\left[\overline{X}_n - \mu\right]}{S_n}$$

    has a Student distribution with $n-1$ degrees of freedom.

Probability and Statistics 401-2604

Overview of definitions and results from statistics

Introduction

In most of the theory the data (observations) are i.i.d. real-valued random variables $X_1, \ldots, X_n$. We call n the sample size. We then say that $X_1, \ldots, X_n$ are i.i.d. copies of a random variable X.

We often denote shorthand the data by $X \in \mathcal{X}$ as well (abuse of notation). The space $\mathcal{X}$ is the observation space (typically (a subset of) Euclidean space).

A statistical model says that $X \sim P \in \{P_\theta : \ \theta \in \Theta\}$. The set Θ is called the parameter space. Typically Θ is (some subset of) Euclidean space.

A parameter of interest is a function $g(\theta) = Q(P_\theta) =: \gamma$.

Definition (LN Section 6.1) *An estimator T of a parameter of interest $g(\theta) \in \mathbb{R}$ is a (measurable) map $T : \ \mathcal{X} \to \mathbb{R}$.*

Remark An estimator is also often called a statistic. A statistic T is a measurable map $T : \mathcal{X} \to \mathbb{R}$.

Remark Often we denote estimators with a "hat", e.g. $\hat{\gamma}$ as estimator of γ.

Notation If X has distribution P_θ its expectation depends on θ. We (often) write the expectation with a subscript: $E_\theta(X)$.

Remark If the data are $(X_1, \ldots, X_n)$ an estimator T is thus some function of $X_1, \ldots, X_n$.

Remark We often write $E_\theta T(X) =: E_\theta T$ (or $E_\theta T(X_1, \ldots, X_n) =: E_\theta T$).

Definition (LN Section 6.2) *The Mean Square Error (MSE) of an estimator T of $g(\theta) \in \mathbb{R}$ is*

$$\mathrm{MSE}_\theta(T) = E_\theta(T - g(\theta))^2.$$

The bias of T is

$$\mathrm{bias}_\theta(T) = E_\theta T - g(\theta).$$

The estimator T is called unbiased if

$$\mathrm{bias}_\theta(T) = 0, \ \forall \ \theta \in \Theta.$$

The standard error of T is

$$\sigma_\theta(T) = \sqrt{\mathrm{Var}_\theta(T)}.$$

Lemma

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta^2(T) + \mathrm{Var}_\theta(T).$$

Proof. Write $q(\theta) := E_\theta(T)$. Then

$$\text{MSE}_\theta(T) = E_\theta\Big(T - q(\theta) + q(\theta) - g(\theta)\Big)^2$$

$$= E_\theta\Big(T - q(\theta)\Big)^2 + \Big(q(\theta) - g(\theta)\Big)^2 + 2\Big(q(\theta) - g(\theta)\Big)\underbrace{E_\theta\Big(T - q(\theta)\Big)}_{=0}$$

$$= \text{Var}_\theta(T) + \text{bias}_\theta^2(T).$$

□

Example Let $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \mathbb{R}$ where $EX =: \mu$ and $\text{Var}(X) =: \sigma^2$. Then the sample average $\bar{X} = \sum_{i=1}^n X_i/n$ is an unbiased estimator of μ and the sample variance $S^2 := \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is an unbiased estimator of σ^2. However, S is **not** an unbiased estimator of σ.

LLN as source of inspiration Let $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \mathbb{R}$ where $EX =: \mu$ and $\text{Var}(X) =: \sigma^2$. Then by the LLN $\bar{X} \approx \mu$ for n large. Thus it makes sense to estimate μ by $\bar{X}$. Similarly, for a given some function g, inspired by the LLN an estimator of $Eg(X)$ is $\sum_{i=1}^n g(X_i)/n$ and for a given function h an estimator of $h(\mu)$ is $h(\bar{X})$, etc. For example $\sigma^2 = EX^2 - \mu^2$ by definition, so the LLN leads to the estimator

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

of σ^2. Note that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. For large n, the two estimators $\hat{\sigma}^2$ and S^2 are close. Again, inspired by the LLN, an estimator of the CDF $F(x) = P(X \le x)$, $x \in \mathbb{R}$ is

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathrm{l}_{\{X_i \le x\}},\ x \in \mathbb{R}.$$

The function $\hat{F}_n$ is called the empirical distribution function.

Bayesian statistics (see e.g. JR)

Data: $X \in \mathcal{X}$, where $\mathcal{X}$ is some measurable space (usually $\mathbb{R}^d$).

Model: X has distribution P_θ, $\theta \in \Theta$.

Frequentist statistics assumes the unknown to be θ fixed (nonrandom).

Bayesian statistics assumes θ to be random.

Let $p(x|\theta)$ be the pmf/pdf of $X \sim P_\theta$, $\theta \in \Theta$ (assumed to exist).

Suppose Θ is measurable space an let Π be a given probability distribution on Θ.

Definition *For a dominating measure μ the prior density of θ is*

$$w(\vartheta) := \frac{d\Pi}{d\mu}(\vartheta), \ \vartheta \in \Theta.$$

Remark
○ If Θ is countable we let $w(\cdot)$ be the pmf of θ.
○ If $\Theta = \mathbb{R}$ and if Π is absolutely continuous, we let $w(\cdot)$ be the pdf of θ.
○ In both discrete and absolutely continuous case we call $w(\cdot)$ a density. Other cases will not be considered in this course.

Definition *The marginal pmf/pdf of X is*

$$p(x) = \int p(x|\vartheta)w(\vartheta)d\mu(\vartheta) = \begin{cases} \sum_\vartheta p(x|\vartheta)w(\vartheta) & \theta \text{ discrete} \\ \int_\vartheta p(x|\vartheta)w(\vartheta)d\vartheta & \theta \text{ abs. continuous} \end{cases}, \ x \in \mathcal{X}.$$

For $p(x) > 0$ the posterior density of θ given $X = x$ is

$$w(\vartheta|x) := \frac{p(x|\vartheta)w(\vartheta)}{p(x)}.$$

(Compare with AD Theorem 3.4 and AD Definition 12.6: Bayes rule.)

Remark
The posterior density $w(\cdot|x)$ can be a pmf or pdf, other cases will not be considered in this course.

Definition *The Maximum a Posteriori (MAP) estimator is*

$$\hat{\theta}_{\text{MAP}} := \hat{\theta}_{\text{MAP}}(X) := \arg\max_{\vartheta \in \Theta} w(\vartheta|X),$$

provided the maximum exists.

Note To find $\hat{\theta}_{\text{MAP}}$ you do not need to calculate the marginal distribution $p(\cdot)$:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\vartheta \in \Theta} p(X|\vartheta)w(\vartheta).$$

Note We may also write

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\vartheta \in \Theta} \left\{ \log p(X|\vartheta) + \log w(\vartheta) \right\}.$$

Classification Example Consider two given pmf's/pdf's $p_0(x)$ and $p_1(x)$, $x \in \mathcal{X}$. Given an observation X, we want to classify it as coming from distribution P_0 (with pmf/pdf p_0) or P_1 (with pmf/pdf p_1). Let the prior be

$$w(\vartheta) = \begin{cases} w_0 & \vartheta = 0 \\ w_1 & \vartheta = 1 \end{cases}$$

for given $0 < w_0 < 1$ and $w_1 = 1 - w_0$. Then the MAP estimator is

$$\hat{\theta}_{\text{MAP}} = \begin{cases} 1 & \frac{p_1(X)}{p_0(X)} > \frac{w_0}{w_1} \\ \gamma & \frac{p_1(X)}{p_0(X)} = \frac{w_0}{w_1} \\ 0 & \frac{p_1(X)}{p_0(X)} < \frac{w_0}{w_1} \end{cases}$$

where $\gamma \in \{0, 1\}$ is arbitrary (compare with LN Example 2.17: optimal decoding). Here, use that

$$w(\vartheta|x) = \begin{cases} p_0(x) w_0, & \vartheta = 0 \\ p_1(x) w_1, & \vartheta = 1 \end{cases}.$$

Note that

$$p(x) = w_0 p_0(x) + w_1 p_1(x), \ x \in \mathcal{X},$$

is a mixture of p_0 and p_1. The estimator $\hat{\Theta}_{\text{MAP}}$ is often also called <u>Bayes decision</u>. Indeed, we can reformulate situation in terms of <u>decision theory</u>. There are two possible actions $a = 0$ (classify as coming from p_0) and $a = 1$ (classify as coming from p_1). The <u>action space</u> is thus $\mathcal{A} := \{0, 1\}$. We define the loss function

$$L(\vartheta, a) := \mathrm{l}_{\{\vartheta \neq a\}}, \quad (\vartheta, a) \in \Theta \times \mathcal{A}.$$

This means one unit loss for making a wrong decision. We call for a decision $\phi : \mathcal{X} \to \mathcal{A}$ its <u>risk</u>

$$R(\vartheta, \phi) := E_\vartheta L(\vartheta, \phi(X)) = E[L(\vartheta, \phi(X))|\theta = \vartheta].$$

Thus

$$R(\vartheta, \phi) = \begin{cases} P_0(\phi(X) = 1), & \vartheta = 0 \\ P_1(\phi(X) = 0), & \vartheta = 1 \end{cases}.$$

We then define the <u>Bayes risk</u> of ϕ as the average risk with θ having density w:

$$r_w(\phi) := ER(\theta, \phi).$$

Thus

$$r_w(\phi) = w_0 P_0(\phi(X) = 1) + w_1 P_1(\phi(X) = 0) = P(\phi(X) \neq \theta).$$

Bayes decision is defined as the minimizer of the Bayes risk

$$\phi_{\text{Bayes}} = \arg \min_{\phi:\ \mathcal{X} \to \{0,1\}} r_w(\phi).$$

One may verify that $\phi_{\text{Bayes}} = \hat{\theta}_{\text{MAP}}$ (in this classification problem).

Decision theory (general setup)
Given an action space $\mathcal{A}$ and a loss function $L: \ \Theta \times \mathcal{A} \to \mathbb{R}$ the risk of a decision $d: \ \mathcal{X} \to \mathcal{A}$, is

$$R(\vartheta, d) := E_\vartheta L(\vartheta, d(X)) = E[L(\vartheta, d(X))|\theta = \vartheta].$$

With a prior density w on Θ the Bayes risk is of d is

$$r_w(d) := ER(\theta, d) = \begin{cases} \sum_\vartheta R(\vartheta, d(X)w(\vartheta), \ \theta \text{ discrete} \\ \int_\vartheta R(\vartheta, d(X)w(\vartheta)d\vartheta, \ \theta \text{ abs. continuous} \end{cases}.$$

Bayes decision is

$$d_{\text{Bayes}} := \arg \min_{d:\ \mathcal{X} \to \mathcal{A}} r_w(d).$$

Remark In the above setup we did not explicitly state the needed measurability conditions.

Note For example, when both X and θ are discrete

$$\begin{aligned} r_w(d) &= \sum_\vartheta R(\vartheta, d)w(\vartheta) \\ &= \sum_\vartheta E[L(\vartheta, d(X))|\theta = \vartheta]w(\vartheta) \\ &= \sum_\vartheta \sum_x L(\vartheta, d(x))p(x|\vartheta)w(\vartheta) \\ &= \sum_\vartheta \sum_x L(\vartheta, d(x))w(\vartheta|x)p(x) \\ &= \sum_x \sum_\vartheta L(\vartheta, d(x))w(\vartheta|x)p(x) \\ &= \sum_x E[L(\theta, d(x))|X = x]p(x) \end{aligned}$$

Iterated expectations *We have*

$$r_w(d) = EE[L(\theta, d(X)|\theta] = EL(\theta, d(X)) = EE[L(\theta, d(X)|X].$$

This is the short hand version of what was written out above for the case both X and θ discrete.

Lemma *We have*

$$d_{\text{Bayes}}(X) = \arg \min_{a \in \mathcal{A}} E[L(\theta, a)|X].$$

Proof.

$$r_w(d) = EL(\theta, d(X)) = EE[L(\theta, d(X)|X] \geq E\left(\min_{a \in \mathcal{A}} E[L(\theta, a)|X]\right).$$

□

Classification example revisited It holds that

$$\begin{aligned} E[L(\theta, a)|X = x] &= a\frac{p_0(x)w_0}{p(x)} + (1-a)\frac{p_1(x)w_1}{p(x)} \\ &= a\frac{p_0(x)w_0 - p_1(x)w_1}{p(x)} + \frac{p_1(x)w_1}{p(x)}. \end{aligned}$$

The last term does not depend on a so we can omit it when carrying out the minimization. Then for any $\gamma \in \{0, 1\}$,

$$\arg\min_{a \in \{0,1\}} a\frac{p_0(x)w_0 - p_1(x)w_1}{p(x)} = \begin{cases} 1 & p_1(x)w_1 > p_0(x)w_0 \\ \gamma & p_1(x)w_1 = p_0(x)w_0 \\ 0 & p_1(x)w_1 < p_0(x)w_0 \end{cases}.$$

Bayes estimator for quadratic loss
Let $\Theta \subset \mathbb{R}$, $\mathcal{A} = \mathbb{R}$ and $L(\vartheta, a) := (\vartheta - a)^2$. Then

$$R(\vartheta, d) = E[(d(X) - \vartheta)^2 | \theta = \vartheta] = MSE_\vartheta(d).$$

Bayes estimator is

$$d_{\text{Bayes}}(X) = \arg\min_{a \in \mathbb{R}} E[(\theta - a)^2 | X] = E(\theta|X)$$

(compare with AD Example 11.18: best predictor given X).

Example: Bayesian inference for the binomial distribution
Let $X|\theta \sim \text{Binomial}(n, \theta)$ and $\theta \sim \text{Beta}(r, s)$. Then the prior mean is $E\theta = \frac{r}{r+s}$. The posterior density is

$$\begin{aligned} w(\vartheta|x) \propto p(x|\vartheta)w(\vartheta) &\propto \vartheta^x(1-\vartheta)^{n-x}\vartheta^{s-1}(1-\vartheta)^{r-1} \\ &= \vartheta^{x+s-1}(1-\vartheta)^{n-x+r-1}. \end{aligned}$$

So $\theta|X = x \sim \text{Beta}(x + r, n - x - s)$ and Bayes estimator for quadratic loss is

$$E(\theta|X) = \frac{X + r}{n + r + s}.$$

The MAP estimator is

$$\hat{\Theta}_{\text{MAP}} = \frac{X + r - 1}{n + s + r - 2}.$$

Example: Bayesian inference for the normal distribution
Let $X|\theta \sim \mathcal{N}(\theta, \sigma^2)$ were $\theta \in \mathbb{R}$ and where σ^2 is known. Suppose $\theta \sim \mathcal{N}(0, \tau^2)$.

Then we have seen (see the example at the end of the section “Multivariate continuous distributions”)

$$\theta|X \sim \mathcal{N}\left(\frac{\tau^2}{\tau^2+\sigma^2}X, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right).$$

Thus Bayes estimator for quadratic loss is

$$E(\theta|X) = cX,\ c := \frac{\tau^2}{\tau^2+\sigma^2}.$$

In this case this is also the MAP estimator.

Method of moments

Let $X \in \mathbb{R}$ and let the data $X_1, \ldots, X_n$ be i.i.d. copies of X.

Definition (as AD Definition 4.10 and 7.10) *For $k \in \mathbb{N}$ the k-th moment of X is*

$$\mu_k := EX^k$$

(if the expectation exists).

Definition *The k-th sample moment is*

$$\hat{\mu}_k := \frac{1}{n} \sum_{k=1}^{n} X_i^k, \ k \in \mathbb{N}.$$

Note By the LLN $\hat{\mu}_k \approx \mu_k$ for n large (provided the moment exists).

Let $X \sim P_\theta$, where $\theta \in \Theta \subset \mathbb{R}^p$. Then the moments of X also depend on θ:

$$\mu_k = \mu_k(\theta) = E_\theta X.$$

LLN as source of inspiration $\rightsquigarrow$

Definition *The methods of moments estimator $\hat{\theta}$ is a solution of*

$$\mu_k(\vartheta)_{\vartheta = \hat{\theta}} = \hat{\mu}_k, \ k = 1, \ldots, p.$$

(assuming a solution exists).

Example A Suppose $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Then $EX = \lambda$ so the methods of moments estimator is $\hat{\lambda} = \bar{X}$. It holds that $E\hat{\lambda} = \lambda$ for all $\lambda > 0$ so $\hat{\lambda}$ is unbiased. Moreover $\text{var}(\hat{\lambda}) = \lambda/n$ and we can estimate the variance by

$$\widehat{\text{Var}}(\hat{\lambda}) := \hat{\lambda}/n.$$

By the CLT, $\hat{\lambda}$ is approximately $\mathcal{N}(\lambda, \lambda/n)$-distributed for n large. Thus

$$P\Big(|\hat{\lambda} - \lambda| \le z\sqrt{\lambda/n}\Big) \approx \Phi(z) - \Phi(-z) = 2\Phi(z) - 1.$$

We have $\Phi(1.96) = .975$. Therefore

$$\hat{\lambda} \pm (1.96)\sqrt{\hat{\lambda}/n} = \bar{X} \pm \underbrace{(1.96)}_{\approx 2} \sqrt{\bar{X}/n}$$

is approximately a 95% confidence interval for λ:

$$P\left(\lambda \in \left[\hat{\lambda} - (1.96)\sqrt{\hat{\lambda}/n}, \hat{\lambda} + (1.96)\sqrt{\hat{\lambda}/n}\right]\right) \approx .95.$$

See also AD Example 10.19, where 1.96 was replaced by 2 for simplicity and the approximate 95 % confidence interval was

$$\bar{X} + \frac{2}{n} \pm 2\sqrt{\frac{\bar{X} + 1/n}{n}}.$$

The two approximate confidence intervals are for n large approximately equal (the second one is slightly more conservative).

Subexample

x_i	# days
0	100
1	60
2	32
3	8
≥ 4	0

Here $n = 200$, $\bar{x} = .74$. Then an approximate 95% confidence interval for λ is

$$\bar{x} \pm 2\sqrt{\bar{x}/n} = [0.62, 0.84].$$

Let $\gamma := g(\lambda) := P_\lambda(X \geq 4)$ be the parameter of interest. Then

$$\hat{\gamma} = \widehat{g(\lambda)} := g(\hat{\lambda}) = .00697,$$

and an approximate 95% confidence interval for γ is

$$g\left(\bar{x} \pm 2\sqrt{\bar{x}/n}\right) = [0.0038, 0.01]$$

(since $\lambda \mapsto g(\lambda)$ is a monotone function).

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	# days
-.74	.5476	100
.26	.0676	60
1.26	1.5876	32
2.26	5.1076	8

We find $s^2 := \sum_{i=1}^n (x_i - \bar{x})^2/(n-1) = .7561$. The sample variance s^2 is an estimate of $\text{Var}(X)$. In this case $\text{Var}(X) = \lambda$. So both $\bar{x} = .74$ and $s^2 = .7561$ are estimating λ. The fact that these values are not very different can be seen as an indication that the Poisson model is appropriate. One may ask which one of the two estimators is “better”. This is theory treated for example in the course *Fundamentals of Mathematical Statistics.*

Example B Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Then the methods of moments estimator is

$$\hat{\mu} = \bar{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Example C Let $X \sim \text{Gamma}(\alpha, \lambda)$:

$$EX = \alpha\lambda,\ \text{Var}(X) = \alpha\lambda^2.$$

Then $EX^2 = \alpha(\alpha+1)\lambda^2$. So the methods of moments estimator $(\hat{\alpha}, \hat{\lambda})$ solve the two equations

$$\hat{\mu}_1 = \hat{\alpha}\hat{\lambda},\ \hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}\hat{\lambda}^2.$$

It follows that

$$\hat{\lambda} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1},\ \hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Example D Let the data $X_1, \ldots, X_n$ be i.i.d. copies of X where X has pdf

$$p_\theta(x) = \frac{1+\theta x}{2},\ -1 \le x \le 1,\ -1 \le \theta \le 1.$$

Then

$$E_\theta(X) = \frac{\theta}{3}.$$

The methods of moments estimator is thus $\hat{\theta} = 3\bar{X}$.

Example E (Mixtures, compare AD Theorem 7.1) Let X have density

$$p_\theta(x) := \pi_1 \frac{1}{\tau_1}\phi\biggl(\frac{x-\nu_1}{\tau_1}\biggr) + (1-\pi_1)\frac{1}{\tau_2}\phi\biggl(\frac{x-\nu_2}{\tau_2}\biggr)$$

where ϕ is the standard normal density. To simplify, we assume that $\pi_1 = \frac{1}{2}$, $\nu_1 = 0$ and $\tau_1 = 1$ are given. We write $\nu := \nu_2$ and $\tau := \tau_2$. The unknown parameter is $\theta = (\nu, \tau)$. We have

$$EX = \frac{1}{2}\nu,\ EX^2 = \frac{1}{2} + \frac{1}{2}(\nu^2 + \tau^2).$$

So the method of moments estimator $(\hat{\nu}, \hat{\tau})$ solve

$$\frac{1}{2}\hat{\nu} = \hat{\mu}_1,\ \frac{1}{2} + \frac{1}{2}(\hat{\nu}^2 + \hat{\tau}^2) = \hat{\mu}_2.$$

Hence

$$\hat{\nu} = 2\hat{\mu}_1,\ \hat{\tau}^2 = 2\hat{\mu}_2 - 4\hat{\mu}_1^2 - 1.$$

Plug in method The method of moments is inspired by the LLN, but the LLN can also be a source of inspiration for further constructions.

Example 6.3 LN Let $(X, Y) \in \mathbb{R}^2$. Recall the best linear predictor of Y given X (see AD Example 11.8) is $\alpha + \beta X$ with

$$\alpha = EY - \beta EX,\ \beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

Let now $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. copies of (X, Y). Then, the LLN leads to the estimators

$$\hat{\alpha} := \bar{Y} - \hat{\beta}\bar{X},\ \hat{\beta} := \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

The estimator $(\hat{\alpha}, \hat{\beta})^T$ is called the <u>least squares estimator</u>. Note that

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \arg \min_{(a,b)^T \in \mathbb{R}^2} \sum_{i=1}^{n} (Y_i - (a + bX_i))^2.$$

Example Let X have CDF F. Assume the median $m := F^{-1}(\frac{1}{2})$ exists. Let $\hat{F}_n$ be the empirical distribution function. Then we can estimate m by a solution $\hat{m}$ of $\hat{F}_n(\hat{m}) \approx 1/2$. The sample median is

$$\hat{m} := \begin{cases} X_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}.$$

Here $X_{(1)} \leq \cdots \leq X_{(n)}$ are the <u>order statistics</u>.

Maximum likelihood (LN Section 6.2.2)

Let the data be $X \sim P_\theta$, $\theta \in \Theta$, with pmf/pdf p_θ.

Recall the Bayesian MAP estimator

$$\hat\theta_{\text{MAP}} := \arg\max_{\vartheta \in \Theta} p_\vartheta(X) w(\vartheta)$$

$$= \arg\max_{\vartheta \in \Theta} \biggl\{ \log p_\vartheta(X) + \log w(\vartheta) \biggr\}.$$

Definition *The <u>maximum likelihood estimator (MLE)</u> of θ is*

$$\hat\theta_{\text{MLE}} := \arg\max_{\vartheta \in \Theta} p_\vartheta(X)$$

(assuming the maximum exists).

Note When Θ is a bounded set (in $\mathbb{R}^p$) the MLE is thus the MAP with uniform prior.

Note

$$\hat\theta_{\text{MLE}} := \arg\max_{\vartheta \in \Theta} \log p_\vartheta(X).$$

Remark The pmf/pdf $p_\vartheta(X)$ considered as a function of ϑ is called the <u>likelihood</u>. In other words, the likelihood is $\vartheta \mapsto p_\vartheta(X)$.

Remark If the data are actually $X_1, \ldots, X_n$, i.i.d. copies of a random variable X with pmf/pdf p_θ, $\theta \in \Theta$, then the likelihood is

$$\vartheta \mapsto \prod_{i=1}^n p_\vartheta(X_i).$$

The MLE is then

$$\hat\theta_{\text{MLE}} := \arg\max_{\vartheta \in \Theta} \prod_{i=1}^n p_\vartheta(X_i)$$

$$= \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \log p_\vartheta(X_i).$$

The MLE can often (not always) be obtained by setting the derivative of the log-likelihood to zero:

$$\sum_{i=1}^n s_{\hat\theta}(X_i) = 0, \ s_\vartheta(\cdot) := \frac{\partial}{\partial \vartheta} \log p_\vartheta(\cdot).$$

<u>LLN as source of inspiration:</u> One can show that

$$\theta = \arg\max_{\vartheta \in \Theta} E_\theta \log p_\vartheta(X),$$

and also - under regularity conditions -

$$E_\theta s_\theta(X) = 0,\ s_\vartheta := \frac{\partial}{\partial \vartheta} \log p_\vartheta.$$

LN Example 6.9 Let the data be $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$. Writing $\vartheta := (\tilde{\mu}, \tilde{\sigma}^2)$ the log-likelihood is

$$\sum_{i=1}^n \log p_\vartheta(X_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tilde{\sigma}^2 - \frac{\sum_{i=1}^n (X_i - \tilde{\mu})^2}{2\tilde{\sigma}^2}.$$

Taking derivatives w.r.t. $\tilde{\mu}$ gives

$$\frac{\sum_{i=1}^n (X_i - \hat{\mu}_{\text{MLE}})}{\hat{\sigma}^2_{\text{MLE}}} = 0,$$

so that $\hat{\mu}_{\text{MLE}} = \bar{X}$. As

$$\bar{X} = \arg\min_{\tilde{\mu}} \sum_{i=1}^n (X_i - \tilde{\mu})^2,$$

it is also called the least squares estimator (LSE) of μ.

Inserting $\hat{\mu}_{\text{MLE}} = \bar{X}$ and differentiating w.r.t. $\tilde{\sigma}^2$ gives

$$-\frac{n}{2\hat{\sigma}^2_{\text{MLE}}} + \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\hat{\sigma}^4_{\text{MLE}}} = 0$$

so $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Thus, in this case the MLE equals the method of moments estimator.

LN Example 6.8 Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \sim \text{Laplace}(\mu, \sigma^2)$, where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$. The pdf of X is then

$$p_\theta(x) = \frac{1}{2\sigma}\exp\left[-\frac{|x - \mu|}{\sigma}\right],\ x \in \mathbb{R}.$$

The log-likelihood based on the sample $(X_1, \ldots, X_n)$ is

$$\sum_{i=1}^n \log p_\vartheta(X_i) = -n\log 2 = n\log\tilde{\sigma} - \frac{\sum_{i=1}^n |X_i - \tilde{\mu}|}{\tilde{\sigma}},\ \vartheta = (\tilde{\mu}, \tilde{\sigma}).$$

It follows that

$$\hat{\mu}_{\text{MLE}} = \arg\min_{\tilde{\mu}} \sum_{i=1}^n |X_i - \tilde{\mu}|.$$

For n even the minimizer is not unique. We take the sample median

$$\hat{\mu}_{\text{MLE}} = \hat{m} := \begin{cases} X_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics. The sample median is often called the least absolute deviations (LAD) estimator of μ.
Let us briefly see whether the LLN can make sense out of this estimator. One may verify that

$$E|X - \tilde{\mu}| = 2 \int_{x > \tilde{\mu}} (1 - F(x)) dx + \tilde{\mu} - EX,$$

where F is the CDF of X. One can find

$$\arg\min_{\tilde{\mu}} E|X - \tilde{\mu}|$$

by setting the derivative of $E|X - \tilde{\mu}|$ to zero

$$-2(1 - F(\tilde{\mu}))|_{\tilde{\mu} = \arg\min} + 1 = 0.$$

In other words

$$\arg\min_{\tilde{\mu}} E|X - \tilde{\mu}| = F^{-1}(\tfrac{1}{2}),$$

is the theoretical median (provided it exists).
We still have to calculate the MLE of σ. By differentiating the log-likelihood w.r.t. $\tilde{\sigma}$ one gets

$$-\frac{n}{\hat{\sigma}_{\text{MLE}}} + \frac{\sum_{i=1}^{n} |X_i - \hat{m}|}{\hat{\sigma}_{\text{MLE}}^2} = 0,$$

which gives $\hat{\sigma}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{m}|$.

Remark Estimating the mean EX by the LSE $\bar{X}$ remains a valid procedure also for non-Gaussian data. Similarly, the LAD estimator $\hat{m}$ remains valid estimator of the median $F^{-1}(\frac{1}{2})$ also when the data are not Laplacian.

Example Let the data be $X \sim \text{Binomial}(n, \theta)$, where the success probability $0 < \theta < 1$ is unknown. Then for $x \in \{0, 1, \ldots, n\}$

$$p_{\vartheta}(x) = \binom{n}{x} \vartheta^x (1 - \vartheta)^{n-x},$$

and

$$\log p_{\vartheta} = \log \binom{n}{x} x \log \vartheta + (n - x) \log(1 - \vartheta).$$

We have

$$\frac{d}{d\vartheta} \log p_{\vartheta}(X) = \frac{X}{\vartheta} - \frac{n - X}{1 - \vartheta}.$$

Setting this to zero gives

$$\frac{X}{\hat{\theta}_{\text{MLE}}} - \frac{n - X}{1 - \hat{\theta}_{\text{MLE}}} = 0,$$

giving

$$\hat{\theta}_{\text{MLE}} = \frac{X}{n}$$

(compare with a Bayesian estimator, e.g. the MAP).

LN Example Section 6.3.3 Let the data $X_1, \ldots, X_n$ be i.i.d. copies of $X \in \{1, \ldots, k\}$. For example, X represents a "class label". The probability of a particular label is unknown:

$$P_\theta(X = j) := \theta_j, \ j = 1, \ldots, k,$$

where $\theta \in \Theta = \{\vartheta \in \mathbb{R}^k : \ \vartheta_j \geq 0 \ \forall \ j, \ \sum_{j=1}^k \vartheta_j = 1\}$. We may write

$$\log p_\vartheta(x) = \sum_{j=1}^k \mathrm{l}_{\{x=j\}} \log \vartheta_j.$$

Hence the log-likelihood based on $X_1, \ldots, X_n$ is

$$\sum_{i=1}^n \log p_\vartheta(X_i) = \sum_{i=1}^n \sum_{j=1}^k \mathrm{l}_{\{X_i=j\}} \log \vartheta_j = \sum_{j=1}^k N_j \log \vartheta_j,$$

where $N_j := \sum_{i=1}^n \mathrm{l}_{\{X_i=j\}} = \#\{X_i = j\}$ counts the number of observations with the label j ($j = 1, \ldots, k$). To find the maximum of the log-likelihood under the restriction that $\sum_{j=1}^k \vartheta_j = 1$ we use a Lagrange multiplier, we maximize

$$\sum_{j=1}^k N_j \log \vartheta_j + \lambda(1 - \sum_{j=1}^k \vartheta_j).$$

Differentiating and setting to zero gives for the MLE $\hat{\theta}$

$$\frac{\partial}{\partial \vartheta_j} \left\{ \sum_{j=1}^k N_j \log \vartheta_j + \lambda(1 - \sum_{j=1}^k \vartheta_j) \right\} \Bigg|_{\hat{\theta}} = \frac{N_j}{\hat{\theta}_j} - \lambda = 0.$$

Thus

$$\hat{\theta}_j = \frac{N_j}{\lambda}, \ j = 1, \ldots, k.$$

The restriction now gives

$$1 = \sum_{j=1}^k \frac{N_j}{\lambda},$$

and since $\sum_{j=1}^k N_j = n$ we obtain $\lambda = n$. The MLE is therefore

$$\hat{\theta}_j = \frac{N_j}{n}, \ j = 1, \ldots, k.$$

Hypothesis testing (LN Section 6.3)

Let $X \in \mathcal{X}$, $X \sim P_\theta$, $\theta \in \Theta$ We consider two hypotheses about the parameter θ: for $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$
$H_0: \ \theta \in \Theta_0$ the <u>null hypothesis</u>,
$H_1: \ \theta \in \Theta_1$ the <u>alternative hypothesis</u>.

Example Let $X \sim \text{Binomial}(n, \theta)$ and
$H_0: \ \theta = \frac{1}{2}$,
$H_1: \ \theta = \frac{3}{4}$.
Suppose we observe the value $X = 14$. We have
$P_{H_0}(X = 14) = .074$,
$P_{H_1}(X = 14) = .112$.
We see that the likelihood $P_{H_1}(X = 14)$ is larger than the likelihood $P_{H_0}(X = 14)$. The value $\theta = \frac{3}{4}$ is the maximum likelihood estimate over $\{\frac{1}{2}, \frac{3}{4}\}$. The likelihood ratio is

$$\frac{P_{H_1}(X = 14)}{P_{H_0}(X = 14)} = 1.51.$$

Is this large enough to reject H_0 in favour of H_1?

To answer the question in the above example, we need to agree on a criterion for evaluating whether or not rejecting the null hypothesis is a good decision. The point of view one uses in statistical hypothesis testing is that the null hypothesis H_0 represents a situation where "everything is as usual", or "no evidence found". For example, if it concerns the decision of putting someone in prison or not, it makes sense to choose
H_0 : the person is innocent ,
H_1 : the person is guilty ,
when convicting an innocent person is an error considered worse than not to convict a guilty person.

The Bayesian approach is to put a prior on H_0 and H_1. In the frequentist approach, no prior is used. We can make two errors: rejecting H_0 (accepting H_1) when H_0 is true (<u>error first kind</u>) and not rejecting H_0 when H_1 is true (<u>error second kind</u>). It is (generally) not possible to keep **both** errors under control. The idea is now to keep the probability of the error of first kind below a (small) prescribed value α.

	H_0	H_1
$\phi = 1$	error first kind	probability = power
$\phi = 0$		error second kind

Definition *A <u>statistical test</u>[3] at given <u>level</u> α ($0 < \alpha < 1$) is a (measurable)*

[3] We extend this to "randomized" tests $\phi: \ \mathcal{X} \to [0, 1]$ in the next definition

map $\phi: \ \mathcal{X} \to \{0,1\}$ such that

$$\phi(X) = \begin{cases} 1: & H_0 \text{ rejected} \\ 0: & H_0 \text{ not rejected} \end{cases},$$

and such that

$$P_{\theta_0}(\phi(X) = 1) \le \alpha \ \forall \ \theta_0 \in \Theta_0.$$

The <u>*power*</u> *of the test at $\theta_1 \in \Theta_1$ is $P_{\theta_1}(\phi(X) = 1)$.*

Example $X \sim \text{Binomial}(n, \theta)$, with $n = 20$.
H_0: $\theta \le \frac{1}{2}$,
H_1: $\theta > \frac{1}{2}$.
We choose $\alpha = .05$. Let

$$\phi(X) := \begin{cases} 1 & X > c \\ 0 & X \le c \end{cases},$$

where we now need to choose the "critical value" c is such a way that

$$P_{\theta_0}(X > c) \le \alpha \ \forall \ \theta_0 \le \frac{1}{2}.$$

We have

$$\vartheta \mapsto P_\vartheta(X > c) = \sum_{x=c+1}^{n} \binom{n}{x} \vartheta^x (1-\vartheta)^{n-x}$$

is increasing in ϑ so that

$$\max_{\theta_0 \le \frac{1}{2}} P_{\theta_0}(X > c) = P_{\theta_0 = \frac{1}{2}}(X > c) = \sum_{x=c+1}^{n} \binom{n}{x} \frac{1}{2^n}.$$

It holds that

$$\underbrace{P_{\theta_0=\frac{1}{2}}(X > 15)}_{=0.0207} < \underbrace{\alpha}_{=0.05} < \underbrace{P_{\theta_0=\frac{1}{2}}(X > 14)}_{=0.0577}.$$

We choose the critical value c as small as possible: $c = 15$.

Definition *A* <u>*randomized statistical test*</u> *at given* <u>*level*</u> *α $(0 < \alpha < 1)$ is a (measurable) map $\phi: \ \mathcal{X} \to [0,1]$ such that*

$$\phi(X) = \begin{cases} 1: & H_0 \text{ rejected} \\ \gamma \in (0,1): & H_0 \text{ rejected with probability } \gamma \\ 0: & H_0 \text{ not rejected} \end{cases},$$

and such that

$$E_{\theta_0}\phi(X) \le \alpha \ \forall \ \theta_0 \in \Theta_0.$$

The <u>*power*</u> *of the test at $\theta_1 \in \Theta_1$ is $E_{\theta_1}\phi(X)$.*

Example $X \sim \text{Binomial}(n, \theta)$, with $n = 20$.
H_0: $\theta \le \frac{1}{2}$,

H_1: $\theta > \frac{1}{2}$.
We choose $\alpha = .05$. We have

$$P_{\theta_0=\frac{1}{2}}(X > 15) < \alpha < P_{\theta_0=\frac{1}{2}}(X > 14)$$

so we can write

$$\alpha = P_{\theta_0=\frac{1}{2}}(X > 15) + \gamma P_{\theta_0=\frac{1}{2}}(X = 15)$$

where

$$\gamma = \frac{\alpha - P_{\theta_0=\frac{1}{2}}(X > 15)}{P_{\theta_0=\frac{1}{2}}(X = 15)} = 0.79.$$

Thus a test at level α is

$$\phi(X) = \begin{cases} 1 & X > 15 \\ .79 & X = 15 \\ 0 & X < 15 \end{cases} .$$

Suppose we observe $X = 14$. Then H_0 cannot be rejected.

Simple hypothesis versus simple alternative (LN Section 6.3.3)

H_0 : $\theta = \theta_0$,
H_1 : $\theta = \theta_1$.

Let $p_0(\cdot) := p_{\theta_0}(\cdot)$ be the pmf/pdf under H_0 and $p_1(\cdot) := p_{\theta_1}$ be the pmf/pdf under H_1.

Definition *A Neyman-Pearson test is of the form*

$$\phi_{\mathrm{NP}}(X) := \begin{cases} 1 & \frac{p_1(X)}{p_0(X)} > c_0 \\ \gamma & \frac{p_1(X)}{p_0(X)} = c_0 \\ 0 & \frac{p_1(X)}{p_0(X)} < c_0 \end{cases}$$

where $c_0 \geq 0$ and $\gamma \in [0, 1]$ are given constants.

Neyman-Pearson Lemma *Let $\alpha \in (0, 1)$ be a given level. Choose c_0 and γ in such a way that*

$$E_{\theta_0}\phi_{\mathrm{NP}}(X) = \alpha.$$

Then for all (randomized) tests $\tilde{\phi}$ with $E_{\theta_0}\tilde{\phi}(X) \leq \alpha$ it holds that

$$E_{\theta_1}\tilde{\phi}(X) \leq E_{\theta_1}\phi_{\mathrm{NP}}(X).$$

In other words, ϕ_{NP} has maximal power among all tests with level α.

Proof for the discrete case. We have

$$E_{\theta_1}\Big(\tilde{\phi}(X) - \phi_{\mathrm{NP}}(X)\Big) = \sum_x \Big(\tilde{\phi}(x) - \phi_{\mathrm{NP}}(x)\Big)p_1(x)$$

$$= \sum_{p_1/p_0 > c_0} \underbrace{(\tilde{\phi} - \phi_{\text{NP}})}_{\leq 0} p_1 + \sum_{p_1/p_0 = c_0} (\tilde{\phi} - \phi_{\text{NP}}) p_1 + \sum_{p_1/p_0 < c_0} \underbrace{(\tilde{\phi} - \phi_{\text{NP}})}_{\geq 0} p_1$$

$$\leq c_0 \sum_{p_1/p_0 > c_0} (\tilde{\phi} - \phi_{\text{NP}}) p_0 + c_0 \sum_{p_1/p_0 = c_0} (\tilde{\phi} - \phi_{\text{NP}}) p_0 + c_0 \sum_{p_1/p_0 < c_0} (\tilde{\phi} - \phi_{\text{NP}}) p_0$$

$$= c_0 E_{\theta_0} \Big(\tilde{\phi}(X) - \phi_{\text{NP}}(X) \Big) = c_0 \Big(E_{\theta_0} \tilde{\phi}(X) - \alpha \Big) \leq 0.$$

□

LN Example 6.13 Consider $X \sim \text{Binomial}(n, \theta)$ and
$H_0 : \ \theta = \theta_0$,
$H_1 : \ \theta = \theta_1$,
where $\theta_1 > \theta_0$. Then

$$\frac{p_1(x)}{p_0(x)} = \left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)} \right]^x \left(\frac{1-\theta_1}{1-\theta_0} \right) > c_0$$

$$\Leftrightarrow$$

$$\underbrace{x \log \left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)} \right]}_{>0 \text{ as } \theta_1 > \theta_0} + n \log \left(\frac{1-\theta_1}{1-\theta_0} \right) > \log c_0$$

$$\Leftrightarrow$$

$$x > \frac{\log c_0 - n \log \left(\frac{1-\theta_1}{1-\theta_0} \right)}{\log \left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)} \right]} := c.$$

A Neyman-Pearson test thus

$$\phi_{\text{NP}}(X) = \begin{cases} 1 & X > c \\ \gamma & X = c \\ 0 & X < c \end{cases} .$$

If we choose the critical value c in such a way that

$$\underbrace{P_{\theta_0}(X > c)}_{= \sum_{x > c} \binom{n}{x} \theta_0^x (1-\theta_0)^{n-x}} \leq \alpha \leq \underbrace{P_{\theta_0}(X > c - 1)}_{= \sum_{x > c-1} \binom{n}{x} \theta_0^x (1-\theta_0)^{n-x}}$$

and then

$$\gamma = \frac{\alpha - P_{\theta_0}(X > c)}{P_{\theta_0}(X = c)},$$

then $E_{\theta_0} \phi_{\text{NP}}(X)) = \alpha$ and ϕ_{NP} is most powerful among all tests with level α. Note that c and γ do not depend on θ_1: the test only depends on the sign of $\theta_1 - \theta_0$.

Example Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$ where μ is unknown and σ_0^2 is known. Write the density of $X_1, \ldots, X_n$ as

$$\mathbf{p}_\mu(x_1, \ldots, x_n) := \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp \Big[- \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma_0^2} \Big].$$

Then

$$\begin{aligned}\frac{\mathbf{p}_{\mu_1}(x_1,\ldots,x_n)}{\mathbf{p}_{\mu_0}(x_1,\ldots,x_n)} &= \exp\left[-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n (x_i-\mu_1)^2 - \sum_{i=1}^n (x_i-\mu_0)^2\right)\right] \\ &= \exp\left[\frac{1}{2\sigma_0^2}\left(-2\sum_{i=1}^n (x_i-\mu_0) + n(\mu_1-\mu_0)^2\right)\right] \\ &= \exp\left[\frac{1}{\sigma_0^2}\left(n\bar{x} - n\mu_0 - n(\mu_1-\mu_0)^2/2\right)\right]\end{aligned}$$

It follows that

$$\frac{\mathbf{p}_{\mu_1}(X_1,\ldots,X_n)}{\mathbf{p}_{\mu_0}(X_1,\ldots,X_n)} > c_0 \Leftrightarrow \begin{cases}\bar{X} > c & \text{if } \mu_1 > \mu_0 \\ \bar{X} < c & \text{if } \mu_1 < \mu_0\end{cases}.$$

To test $H_0: \ \mu = \mu_0$ we consider 3 alternative hypotheses.

Right sided

$H_1: \ \mu = \mu_1 > \mu_0$. Then $\phi_{\mathrm{NP}}(X_1,\ldots,X_n) = \mathrm{l}_{\{\bar{X}>c\}}$ where the critical value c is such that $E_{\mu_0}\phi_{\mathrm{NP}}(X_1,\ldots,X_n) = \alpha$. We have

$$E_{\mu_0}\phi_{\mathrm{NP}}(X_1,\ldots,X_n) = P_{\mu_0}(\bar{X} > c) = P_{\mu_0}\left(\sqrt{n}\frac{(\bar{X}-\mu_0)}{\sigma_0} > \sqrt{n}\frac{(c-\mu_0)}{\sigma_0}\right) = \alpha$$

for

$$\sqrt{n}\frac{(c-\mu_0)}{\sigma_0} = \Phi^{-1}(1-\alpha).$$

Thus

$$c = \mu_0 + \Phi^{-1}(1-\alpha)\sigma_0/\sqrt{n}.$$

For example for $\alpha = .05$ it holds that $\Phi^{-1}(1-\alpha) = 1.65$.

Left sided

$H_1: \ \mu = \mu_1 < \mu_0$. Reject H_0 if

$$\bar{X} < \mu_0 - \Phi^{-1}(1-\alpha)\sigma_0/\sqrt{n}.$$

Two sided

$H_1: \ \mu \neq \mu_0$. The Neyman Pearson lemma cannot be used. It can be shown (see e.g. *Fundamentals of Mathematical Statistics*) that the following test is in some sense optimal: reject H_0 if

$$\bar{X} > \mu_0 + \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n} \text{ or } \bar{X} < \mu_0 - \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}.$$

For example for $\alpha = .05$ it holds that $\Phi^{-1}(1-\frac{\alpha}{2}) = 1.96$.

One sample tests (LN Section 6.3.2)

Theorem *Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Define $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then*

$$\sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}$$

the Student distribution with $n-1$ degrees of freedom.

Proof. We first show that for all i $X_i - \bar{X}$ and $\bar{X}$ are independent (see also AD Theorem 12.4). This follows from

$$\mathrm{Cov}(X_i - \bar{X}, \bar{X}) = \mathrm{Cov}(X_i, \bar{X}) - \underbrace{\mathrm{Cov}(\bar{X}, \bar{X})}_{=\mathrm{Var}(\bar{X})}$$

$$= \frac{1}{n}\sum_{j=1}^n \mathrm{Cov}(X_i, X_j) - \frac{\sigma^2}{n} = 0.$$

The independence now follows from the fact that for multivariate normal random variables, zero covariance implies independence. Thus S^2 and $\bar{X}$ are also independent. Moreover

$$\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2}.$$

By the definition of the χ^2-distribution, the right hand side has a χ^2_n-distribution. Moreover $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ has a χ^2_1-distribution. Since moreover $\sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sigma^2}$ is independent of $\frac{n(\bar{X}-\mu)^2}{\sigma^2}$ it must have a χ^2_{n-1}-distribution. The result now follows from the definition of the Student distribution. □

Remark The Student distribution is symmetric around 0. The density of the t_{n-1}-distribution is

$$f_{n-1}(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})}\left(1+\frac{t^2}{n-1}\right)^{-n/2}, \ t \in \mathbb{R}.$$

Let $c(n-1,\alpha)$ be the $(1-\alpha)$-quantile of the t_{n-1}-distribution. Then we have

$$c(n-1,\alpha)\begin{cases} > \ \Phi^{-1}(1-\alpha) & \forall \ n \\ \to \Phi^{-1}(1-\alpha) & n \to \infty \end{cases}.$$

The Student test

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ_0^2 are unknown.

$\boxed{\text{Right sided}}$

$H_0: \ \mu < \mu_0$,

$H_1: \ \mu > \mu_0$.
Reject H_0 if

$$\bar{X} > \mu_0 + c(n-1,\alpha)S/\sqrt{n}.$$

Then

$$\max_{\mu \leq \mu_0} P_\mu(H_0 \text{ rejected}) = P_{\mu_0}(H_0 \text{ rejected}) = \alpha.$$

$\boxed{\text{Left sided}}$
$H_0: \ \mu > \mu_0$,
$H_1: \ \mu < \mu_0$.
Reject H_0 if

$$\bar{X} < \mu_0 - c(n-1,\alpha)S/\sqrt{n}.$$

$\boxed{\text{Two sided}}$
$H_0: \ \mu = \mu_0$,
$H_1: \ \mu \neq \mu_0$.
Reject H_0 if

$$\bar{X} > \mu_0 + c(n-1,\tfrac{\alpha}{2})S/\sqrt{n} \text{ or } \bar{X} < \mu_0 - c(n-1,\tfrac{\alpha}{2})S/\sqrt{n}.$$

Numerical example:

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
4.5	0	0
4	-.5	.25
3.5	-1	1
6	1.5	2.25
5	.5	.25
4	-.5	.25

We have $n = 6$, $\bar{x} = 4.5$, $\sum(x_i - \bar{x})^2 = 4$, $s^2 = .8$ and $s/\sqrt{n} = .365$. With $\alpha = .05$ the $(1 - \frac{\alpha}{2})$-quantile of the t_5-distribution is $c(5, 0.025) = 2.571$. Thus $c(5, 0.025)s/\sqrt{n} = .939$. For example
$H_0: \ \mu = 5.1$
is rejected when $|\bar{x} - 5.1| > .939$. Thus $H_0: \ \mu = 5.1$ is not rejected as

$$|\bar{x} - 5.1| = .6 < .939.$$

The values for μ which are not rejected are all μ such that $|\bar{x} - \mu| \leq .939$, that is all $\mu \in [3.561, 5.439]$. We call $[3.561, 5.439]$ a 95% confidence interval for μ.

Sign test

Let $X_1, \ldots, X_n$ be i.i.d. with common CDF F. We assume F is continuous in $m := F^{-1}(\frac{1}{2})$. We consider the testing problem
$H_0: \ m = m_0$,
$H_1: \ m \neq m_0$.
As test statistic we take

$$T := \#\{X_i > m_0\}$$

and as (non-randomized) test

$$\phi(T) := \begin{cases} 1 & |T - \frac{n}{2}| \geq c \\ 0 & |T - \frac{n}{2}| < c \end{cases}$$

where c is such that

$$\underbrace{P_{H_0}\left(|T - \frac{n}{2}| \geq c \right)}_{= \sum_{|k - \frac{n}{2}| \geq c} \binom{n}{k} \frac{1}{2^n} =: 1 - G(c)} \leq \alpha$$

and c is as small as possible. One calls $1 - G(|T - n/2|)$ the <u>p-value</u>. Reject H_0 if the p-value is at most α. We can write for $\tilde{c} < n/2$,

$$\phi(T) := \begin{cases} 1 & T \leq \tilde{c} \text{ or } T \geq n - \tilde{c} \\ 0 & \text{else} \end{cases},$$

where

$$\underbrace{P(T \leq \tilde{c}) + P(T \geq n - \tilde{c})}_{= 2\sum_{k \leq \tilde{c}} \binom{n}{k} 2^{-n}} \leq \alpha.$$

<u>Numerical example continued</u>
The normal distribution is symmetric around μ so $m = \mu$. We test
$H_0: \ \mu = 5.1$,
$H_1: \ \mu \neq 5.1$.
We have

$$G(0) = P_{H_0}(T \leq 0 \text{ or } T \geq 6) = P_{H_0}(T = 0) + P_{H_0}(T = 6) = \frac{2}{64} = .03125 < .05$$

so we can take $\tilde{c} = 0$.[4] The observed value of T is $T = 1$. Therefore we cannot reject H_0. Since $n = 6$ we have $|T - \frac{n}{2}| = 2$. The p-value is

$$1 - G(2) = \frac{14}{64} = .21875 > .05.$$

Definition *Let T be a test statistic such that large values of T are evidence against $H_0 : \theta = \theta_0$. We reject H_0 when $T \geq c$ where c is such that*

$$1 - G(c) \leq \alpha$$

with $1 - G(c) := P_{H_0}(T \geq c)$. The <u>p-value</u> is then $1 - G(T)$.

[4] A randomized test at level $\alpha = .05$ is

$$\tilde{\phi}(T) = \begin{cases} 1 & T = 0 \text{ or } T = 6 \\ \frac{1}{10} & T = 1 \text{ or } T = 5 \\ 0 & \text{else} \end{cases}.$$

Indeed

$$E_{H_0}\tilde{\phi}(T) = P_{H_0}(T = 0 \text{ or } T = 6) + \frac{1}{10} P_{H_0}(T = 1 \text{ or } T = 5) = .05.$$

Note $1 - G$ is a decreasing function, so

$$T \geq c \Rightarrow 1 - G(T) \leq 1 - G(c) = \alpha.$$

Thus if the p-value is at most α we reject H_0.

Two sample tests

The data consists of two samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$.

The two sample student test

Model:

$$\underbrace{X_1, \dots, X_n}_{\sim \mathcal{N}(\mu_1, \sigma^2)}, \underbrace{Y_1, \dots, Y_m}_{\sim \mathcal{N}(\mu_2, \sigma^2)} \text{ independent}$$

We want to test
$H_0: \ \mu_1 = \mu_2$
$H_1: \ \mu_1 \neq \mu_2$.

If $\mu_1 = \mu_2$ then for n large $\bar{X} \approx \bar{Y}$. This leads to rejecting H_0 if $|\bar{X} - \bar{Y}| > c$ where the critical value c is to be chosen in such a way that

$$P_{H_0}(|\bar{X} - \bar{Y}| > c) = \alpha$$

where $0 < \alpha < 1$ is a given level. So we need to find the distribution of $\bar{X} - \bar{Y}$ under H_0. It holds that

$$\bar{X} \sim \mathcal{N}\biggl(\mu_1, \frac{\sigma^2}{n}\biggr), \ \bar{Y} \sim \mathcal{N}\biggl(\mu_2, \frac{\sigma^2}{m}\biggr).$$

Moreover

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2,$$

and since $\bar{X}$ and $\bar{Y}$ are independent

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \biggl(\frac{n+m}{nm}\biggr).$$

Thus

$$\bar{X} - \bar{Y} \sim \mathcal{N}\biggl(\mu_1 - \mu_2, \sigma^2\biggl(\frac{n+m}{nm}\biggr)\biggr).$$

Standardizing gives

$$\sqrt{\frac{nm}{n+m}} \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma} \sim \mathcal{N}(0, 1).$$

We consider two cases.

$\boxed{\sigma^2 = \sigma_0^2 \text{ known:}}$ Then we can take as test statistic

$$T_0 := \sqrt{\frac{nm}{n+m}} \frac{\bar{X} - \bar{Y}}{\sigma_0}.$$

Under H_0 the statistic T_0 has a standard normal distribution. We reject H_0 when $|T_0| > \Phi^{-1}(1 - \frac{\alpha}{2})$. Then

$$P_{H_0}(H_0 \text{ rejected}) = P_{H_0}\biggl(|T_0| > \Phi^{-1}(1 - \tfrac{\alpha}{2})\biggr) = \alpha.$$

In other words the critical value is $c = \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{n+m}{nm}}\sigma_0$. (With the "common" choice $\alpha = .05$ it holds that $c = (1.96)\sqrt{\frac{n+m}{nm}}\sigma_0$, i.e., roughly twice the standard deviation of $\bar{X} - \bar{Y}$).

$\boxed{\sigma^2 \text{ unknown:}}$ To estimate the standard deviation of $\bar{X} - \bar{Y}$ we need an estimator of σ^2. A good choice turns out to be the "pooled sample" variance

$$\tilde{S}^2 := \frac{1}{n+m-2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right\},$$

which is unbiased. Standardizing with the estimated standard deviation gives the statistic

$$T := \sqrt{\frac{nm}{n+m}}\frac{\bar{X} - \bar{Y}}{\tilde{S}}.$$

But because $\tilde{S}$ is random T is no longer normally distributed. This is not really a problem, as long as its distribution under H_0 does not depend on unknown parameters. It is now not difficult to show that under H_0, T has a Student distribution with $n+m-2$ degrees of freedom, the t_{n+m-2}-distribution[5]. Therefore, with $c(n+m-2, \frac{\alpha}{2})$ the $(1-\frac{\alpha}{2})$-quantile of the t_{n+m-2}-distribution, we reject H_0 if $|T| > c(n+m-2, \frac{\alpha}{2})$ or equivalently if $|\bar{X} - \bar{Y}| > \tilde{c}$ where the critical value $\tilde{c}$ is $\tilde{c} = c(n+m-2, \frac{\alpha}{2})\sqrt{\frac{n+m}{nm}}\tilde{S}$.

The two sample Wilcoxon text, or Mann-Whitney U test

Model:

$$\underbrace{X_1, \ldots, X_n}_{\sim F}, \underbrace{Y_1, \ldots, Y_m}_{\sim G} \text{ independent}$$

where F and G are two unknown continuous distributions.

We want to test
$H_0: \; F = G$,
$H_1: \; F \neq G$.

We construct a test statistic as follows. Let $N := n + m$ be the pooled sample size and $(Z_1, \ldots, Z_N) := (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ be the pooled sample. In the pooled sample, let $Z_{(1)} < \cdots < Z_{(N)}$ be the order statistics. Let $R_i := \text{rank}(X_i)$ in the pooled sample (i.e. $Z_{(R_i)} = X_i$) $i = 1, \ldots, n$ and $R_{n+j} := \text{rank}(Y_j)$ in the pooled sample, $j = 1, \ldots, m$. If $F = G$ then $(R_1, \ldots, R_n, R_{n+1}, \ldots, R_N)$ is a random permutation of the numbers $\{1, \ldots, N\}$. This means that under H_0 the ranks $R_1, \ldots, R_n$ have the same distribution as a random sample without replacement of size n from an urn with N balls numbered from 1 to N. The Mann-Whitney U statistic is

$$U := \sum_{i=1}^{n} R_i.$$

[5] As in the one sample case, $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ has a χ^2_{n-1}-distribution. Similarly, $\sum_{i=1}^{n}(Y_j - \bar{Y})^2/\sigma^2$ has a χ^2_{m-1}-distribution. The two sums-of-squares are independent and independent of $\bar{X}$ and $\bar{Y}$.

The Wilcoxon test statistic is

$$W := \#\{X_i > Y_j\}.$$

One may verify that U and W are equivalent:

$$U = W + \frac{n(n+1)}{2}.$$

numerical example

z	rank
$x_1 = 36$	8
$x_2 = 9$	4
$x_3 = 7$	2
$x_4 = 100$	9
$x_5 = 3$	1
$y_1 = 5$	3
$y_2 = 37$	7
$y_3 = 11$	5
$y_4 = 12$	6

Table 1: $n = 5$, $m = 4$, $E_{H_0}(U) = 25$, $u = 24$, $w = 9$

Lemma
i) $E_{H_0}(U) = \frac{n(N+1)}{2}$
ii) $\text{Var}_{H_0}(U) = \frac{nm(N+1)}{12}$.

Proof. (Compare AD, Section 6.5 on the Hypergeometric distribution.)
i) For all i

$$P_{H_0}(R_i = k) = \frac{1}{N}, \ k = 1, \ldots N.$$

Hence

$$E_{H_0}R_i = \sum_{k=1}^{N} k\frac{1}{N} = \frac{N+1}{2}$$

and so

$$E_{H_0}(U) = \frac{n(N+1)}{2}.$$

ii) For all i

$$E_{H_0}R_i^2 = \sum_{k=1}^{N} k^2\frac{1}{N} = \frac{(N+1)(2N+1)}{6}$$

so

$$\text{Var}_{H_0}(R_i) = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12} =: \sigma^2.$$

Further for $i \neq j$

$$E_{H_0}R_iR_j = \sum_{k \neq l} kl\frac{1}{N(N-1)}$$

$$= \frac{N(N+1)^2}{4(N-1)} - \frac{(N+1)(2N+1)}{6(N-1)} = \frac{(N+1)(3N^2-N-2)}{12(N-1)}.$$

Thus

$$\mathrm{Cov}_{H_0}(R_i, R_j) = \frac{(N+1)(3N^2-N-2)}{12(N-1)} - \frac{(N+1)^2}{4} = -\frac{\sigma^2}{N-1}.$$

It follows that

$$\mathrm{Var}_{H_0}(\sum_{i=1}^{n} R_i) = n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1} = n\sigma^2\frac{N-n}{N-1}.$$

□

Corollary $E_{H_0}(W) = \frac{nm}{2}$, $\mathrm{Var}_{H_0}(W) = \frac{nm(N+1)}{12}$.

Standardizing:

$$T := \frac{U - E_{H_0}(U)}{\sqrt{\mathrm{Var}_{H_0}(U)}} = \frac{W - E_{H_0}(W)}{\sqrt{\mathrm{Var}_{H_0}(W)}}.$$

For n and m large, T has under H_0 approximately a $\mathcal{N}(0,1)$-distribution. (This does not follow from the "usual" CLT.)

Numerical example continued

$$|T| = \frac{|24-25|}{\sqrt{\frac{20\times 8}{12}}} = \sqrt{\frac{3}{7}} = .655.$$

The approximate p-value is $2(1-\Phi(.655)) = .513$.

Goodness-of fit tests

Kolmogorov-Smirnov tests

Model: $X_1, \ldots, X_n$ i.i.d. with CDF F.

$H_0: \ F = F_0$.

Recall the empirical distribution function

$$\hat{F}_n(x) := \frac{1}{n} \sum_{i=1}^n \mathrm{l}_{\{X_i \le x\}}, \ x \in \mathbb{R}.$$

Kolmogov-Smirnov tests are based on a comparison of $\hat{F}_n$ with F_0. The test statistic is

$$T_\infty := \sup_x |\hat{F}_n(x) - F_0(x)|,$$

or its variants

$$T_p := \int |\hat{F}_n(x) - F_0(x)|^p dx, \ 1 \le p < \infty.$$

An approximation of the distribution of T_p $(1 \le p \le \infty)$ under the null hypothesis follows from probability theory (not treated here). One may also simulate the null-distribution.

The χ^2-test: simple hypothesis

Let $X \in \{1, \ldots, q\}$ represent a class label. Write

$$P_\theta(X = j) := \theta_j,$$

where

$$\theta \in \Theta := \{\vartheta = (\vartheta_1, \ldots, \vartheta_q) : \ \vartheta_j \ge 0 \ \forall \ j, \ \sum_{j=1}^q \vartheta_j = 1\}.$$

Suppose we want to test
$H_0: \ \theta = \theta_0$.
The data consist of i.i.d. copies $X_1, \ldots, X_n$ of X. The maximum likelihood estimator of θ is

$$\hat{\theta}_j = \frac{N_j}{n}, \ N_j := \#\{X_i = j\}, \ j = 1, \ldots, q.$$

The idea is now to reject H_0 if $\hat{\theta}$ is very different from the hypothesized θ_0. One may use for instance the Euclidean distance between $\hat{\theta}$ and θ_0 as a test statistic. One may however want to take into account the different variances of the estimators of the components. A test statistic that does so is the so-called χ^2 test statistic

$$\chi^2 := n \sum_{j=1}^q \frac{(\hat{\theta}_j - \theta_{0,j})^2}{\theta_{0,j}} = \sum_{j=1}^q \frac{(N_j - n\theta_{0,j})^2}{n\theta_{0,j}}.$$

Theorem *For n large, $P_{H_0}(\chi^2 \le t) \approx G(t)$ for all t, where G is the CDF of a $\chi^2(q-1)$-distribution.*

No proof. (See *Fundamentals of Mathematical Statistics* for a proof.)

Special case: $q = 2$. Then $X := N_1 \sim \text{Binomial}(n, p)$ where $p := \theta_1$, and $N_2 = n - X$, $\theta_2 = 1 - p$. So

$$\chi^2 = \frac{(X - np)^2}{np} + \frac{(n - X - n(1-p))^2}{n(1-p)} = \frac{(X-np)^2}{np(1-p)}.$$

By the CLT

$$\frac{X - np}{\sqrt{np(1-p)}}$$

is approximately $\mathcal{N}(0,1)$-distributed, and so its square

$$\frac{(X-np)^2}{np(1-p)}$$

is approximately $\chi^2(1)$-distributed (by the definition of the χ^2-distribution).

The χ^2-test: composite hypothesis

The random variable $X \in \{1, \ldots, q\}$ again represent a class label and

$$P_\theta(X = j) := \theta_j, \ j = 1, \ldots, q.$$

Suppose we want to test $m < q - 1$ restrictions
$H_0: \ R_k(\theta) = 0, \ k = 1, \ldots, m$. Let

$$\hat{\theta}_0 := \arg \max_{\vartheta \in \Theta: \ R_k(\vartheta)=0, \ k=1,\ldots,m} \sum_{j=1}^q N_j \log \vartheta_j$$

be the maximum likelihood estimator under the m restrictions. Define the test statistic

$$\chi^2 := \sum_{j=1}^q \frac{(N_j - n\hat{\theta}_{0,j})^2}{n\hat{\theta}_{0,j}}.$$

Under some regularity conditions, the distribution of χ^2 under H_0 is approximately $\chi^2(m)$. Thus we reject H_0 when $\chi^2 > G^{-1}(1-\alpha)$ where G is the CDF of the $\chi^2(m)$-distribution. Then

$$P_{H_0}(H_0 \text{ rejected}) \approx \alpha.$$

Note A special case is the simple hypothesis $H_0: \ \theta = \theta_0$. This corresponds to $m = q - 1$ restrictions.

Contingency tables

This paragraph treats a special case of the previous paragraph.

Let $X := (Y, Z) \in \{(k,l) : \ k = 1, \ldots, p, \ l = 1, \ldots, q\}$ and

$$P_\theta\biggl(X = (k,l)\biggr) := \theta_{k,l}$$

where

$$\theta \in \Theta = \{\vartheta = \{\vartheta_{k,l} : \ k = 1, \ldots, p, \ l = 1, \ldots, q\}, \ \vartheta_{k,l} \geq 0 \, \forall \, k, l \ \sum_{k=1}^p \sum_{l=1}^q \vartheta_{k,l} = 1\}.$$

We aim at testing whether Y and Z are independent. Define the marginals

$$\eta_k := \sum_{l=1}^q \theta_{k,l} \ (k = 1, \ldots, p), \quad \xi_l := \sum_{k=1}^p \theta_{k,l} \ (l = 1, \ldots, q).$$

The null hypothesis is $H_0 : \ \theta_{k,l} = \eta_k \xi_l, \ \forall \ k, l$.

The data are $\{X_i = (Y_i, Z_i) : \ i = 1, \ldots, n\}$, i.i.d. copies of $X = (Y, Z)$. The maximum likelihood estimator is as before

$$\hat{\theta}_{k,l} = \frac{N_{k,l}}{n}, \ k = 1, \ldots, p, \ l = 1, \ldots, q,$$

where $N_{k,l} = \#\{(Y_i, Z_i) = (k,l)\}$ $k = 1, \ldots, p$, $l = 1, \ldots, q$.

Write

$$N_{k,+} := \sum_{l=1}^q N_{k,l} \ (k = 1, \ldots, p), \quad N_{+,l} := \sum_{k=1}^p N_{k,l} \ (l = 1, \ldots, q).$$

Lemma *The maximum likelihood under the restrictions of H_0 is*

$$\hat{\eta}_k = \frac{N_{k,+}}{n} \ (k = 1, \ldots, p), \ \hat{\xi}_l = \frac{N_{+,l}}{n} \ (l = 1, \ldots, q).$$

Proof. The log-likelihood is

$$\sum_{k=1}^p \sum_{l=1}^q N_{k,l} \log \vartheta_{k,l}.$$

We now have the restriction $\vartheta_{k,l} = \tilde{\eta}_k \tilde{\xi}_l$ for some non-negative $\tilde{\eta}_k$, $\tilde{\xi}_l$, with $\sum_{k=1}^p \tilde{\eta}_k = 1$ and $\sum_{l=1}^q \tilde{\xi}_l = 1$. The restricted log-likelihood is therefore

$$\sum_{k=1}^p \sum_{l=1}^q N_{k,l} \log(\tilde{\eta}_k \tilde{\xi}_l)$$

$$= \sum_{k=1}^p \sum_{l=1}^q N_{k,l} \log \tilde{\eta}_k + \sum_{k=1}^p \sum_{l=1}^q N_{k,l} \log \tilde{\xi}_l$$

$$= \sum_{k=1}^{p} N_{k,+} \log \tilde{\eta}_k + \sum_{l=1}^{q} N_{+,l} \log \tilde{\xi}_l.$$

The two terms can now be maximized separately, as done previously (where we used a Lagrange multiplier). □

It follows that

$$\chi^2 = \sum_{k=1}^{p} \sum_{l=1}^{q} \frac{(N_{k,l} - N_{k,+}N_{+,l}/n)^2}{N_{k,+}N_{+,l}/n}.$$

The original number of free parameters is

$$pq - 1.$$

The number of free parameters under H_0 is

$$p - 1 + q - 1.$$

The number of restrictions is therefore

$$m = \Big(pq - 1\Big) - \Big(p - 1 + q - 1\Big) = (p-1)(q-1).$$

So χ^2 is approximately $\chi^2((p-1)(q-1))$-distributed under H_0.

Special case: $p = q = 2$

$N_{1,1}$	$N_{1,2}$	$N_{1,+}$
$N_{2,1}$	$N_{2,2}$	N_{2+}
$N_{+,1}$	$N_{+,2}$	n

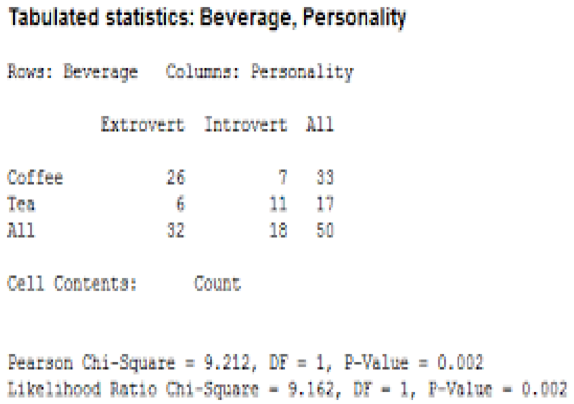
or, using alternative symbols

A	B	R
C	D	S
P	Q	n

Then

$$\chi^2 = \frac{n(AD - BC)^2}{PQRS}.$$

It has approximately a $\chi^2(1)$-distribution under H_0.



Tabulated statistics: Beverage, Personality

```
Rows: Beverage   Columns: Personality

         Extrovert  Introvert  All

Coffee          26          7   33
Tea              6         11   17
All             32         18   50

Cell Contents:      Count


Pearson Chi-Square = 9.212, DF = 1, P-Value = 0.002
Likelihood Ratio Chi-Square = 9.162, DF = 1, P-Value = 0.002
```

In the above example

$$\chi^2 = 50 \times \frac{(26 \times 11 - 6 \times 7)^2}{32 \times 18 \times 33 \times 17} = 9.212.$$

Remark Let $X \sim \text{Binomial}(n_1, p_1)$ and $Y \sim \text{Binomial}(n_2, p_2)$ be independent and suppose we want to test
$H_0: \; p_1 = p_2 =: p$ where $0 < p < 1$ is an unknown common value.
An estimator of p_1 is $\hat{p}_1 = X/n_1$ and an estimator of p_2 is $\hat{p}_2 = Y/n_2$. We reject H_0 if $|\hat{p}_1 - \hat{p}_2|^2$ is large.

X	Y	$X + Y$
$n_1 - X$	$n_2 - Y$	$n - (X + Y)$
n_1	n_2	$n := n_1 + n_2$

We have

$$\text{Var}_{H_0}(\hat{p}_1 - \hat{p}_2) = p(1-p)\frac{n}{n_1 n_2},$$

and we can estimate this by

$$\widehat{\text{Var}}_{H_0}(\hat{p}_1 - \hat{p}_2) := \hat{p}(1-\hat{p})\frac{n}{n_1 n_2},$$

where $\hat{p} = (X+Y)/n$. The standardized test statistic is now

$$T := \frac{|\hat{p}_1 - \hat{p}_2|^2}{\hat{p}(1-\hat{p})\frac{n}{n_1 n_2}} = \frac{n(AD - BC)^2}{PQRS} = \chi^2.$$

Confidence sets (LN Section 6.4)

Numerical example: (Recap)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
4.5	0	0
4	-.5	.25
3.5	-1	1
6	1.5	2.25
5	.5	.25
4	-.5	.25

We have $n = 6$, $\bar{x} = 4.5$, $s^2 = .8$ and $s/\sqrt{n} = .365$. With $\alpha = .05$ the $(1 - \frac{\alpha}{2})$-quantile of the t_5-distribution is $c(5, 0.025) = 2.571$. Thus $c(5, 0.025)s/\sqrt{n} = .939$. Assuming i.i.d. Gaussian data the interval

$$\bar{x} \pm c(5, 0.025)s/\sqrt{n} = 4.5 \pm .939 = [3.561, 5.439]$$

is a 95% confidence interval for μ.

Consider an $X \in \mathcal{X}$ with distribution P_θ depending on $\theta \in \Theta$. Let $g(\theta) \in \mathbb{R}$ be a parameter of interest. Write $\gamma = g(\theta)$ and $\Gamma := \{g(\theta) : \ \theta \in \Theta\}$.

Recall that a statistic is a measurable map $\mathcal{X} \to R$.

Definition *Let $\underline{T} = \underline{T}(X)$ and $\bar{T} = \bar{T}(X)$ be two statistics with $\underline{T} \leq \bar{T}$. One calls $[\underline{T}, \bar{T}]$ a $(1 - \alpha)$-confidence interval for $g(\theta)$ if*

$$P_\theta\Bigl(\underline{T} \leq g(\theta) \leq \bar{T}\Bigr) \geq 1 - \alpha, \ \forall \ \theta \in \Theta.$$

More generally, we may consider confidence **sets**. We consider a mapping

$$J := \mathcal{X} \to \{\text{subsets of } \Gamma\}$$

(such that $I(\gamma) := \{x : \ \gamma \in J(x)\}$ is measurable for all $\gamma \in \Gamma$).

Definition *Let . One calls J a $(1 - \alpha)$-confidence set for $g(\theta)$ if*

$$P_\theta\Bigl(g(\theta) \in J(X)\Bigr) \geq 1 - \alpha, \ \forall \ \theta \in \Theta.$$

Example Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$.

$\boxed{\text{Confidence interval for } \mu,\ \sigma^2 =: \sigma_0^2 \text{ known}}$

Then

$$[\bar{X} - \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}\bar{X} + \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}]$$

is a $(1 - \alpha)$-confidence interval for μ:

$$P_\mu\Bigl(\bar{X} - \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n} \leq \mu \leq \bar{X} + \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}\Bigr)$$

$$= P_\mu\left(\mu - \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n} \le \bar{X} \le \mu + \Phi^{-1}(1-\tfrac{\alpha}{2})\sigma_0/\sqrt{n}\right)$$

$$= P\left(\frac{|\bar{X}-\mu|}{\sigma_0/\sqrt{n}} \le \Phi^{-1}(1-\tfrac{\alpha}{2})\right) = 1-\alpha.$$

Confidence interval for μ, σ^2 unknown

Then

$$[\bar{X} - c(n-1, \tfrac{\alpha}{2})S/\sqrt{n}, \bar{X} + c(n-1, \tfrac{\alpha}{2})S/\sqrt{n}],$$

is a $(1-\alpha)$-confidence interval for μ. Here

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

is the sample variance and $c(n-1, \frac{\alpha}{2})$ the $(1-\frac{\alpha}{2})$-quantile of the Student distribution with $n-1$ degrees of freedom.

Confidence interval for σ^2, $\mu = \mu_0$ known

Then

$$\left[\frac{n\hat{\sigma}^2}{G_n^{-1}(1-\frac{\alpha}{2})}, \frac{n\hat{\sigma}^2}{G_n^{-1}(\frac{\alpha}{2})}\right]$$

is a $(1-\alpha)$-confidence interval for σ^2. Here

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$$

and G_n is the CDF of the $\chi^2(n)$-distribution. Indeed,

$$P_{\sigma^2}\left(\frac{n\hat{\sigma}^2}{G_n^{-1}(1-\frac{\alpha}{2})} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{G_n^{-1}(\frac{\alpha}{2})}\right)$$

$$= P_\sigma\left(G_n^{-1}(\tfrac{\alpha}{2}) \le \frac{n\hat{\sigma}^2}{\sigma^2} \le G_n^{-1}(1-\tfrac{\alpha}{2})\right) = 1-\alpha$$

since $n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n)$.

Confidence interval for σ^2, μ unknown

Then

$$\left[\frac{(n-1)S^2}{G_{n-1}^{-1}(1-\frac{\alpha}{2})}, \frac{(n-1)S^2}{G_{n-1}^{-1}(\frac{\alpha}{2})}\right]$$

is a $(1-\alpha)$-confidence interval for σ^2. Here

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

and G_{n-1} is the CDF of the $\chi^2(n-1)$-distribution. A one-sided confidence interval for σ^2 (right-sided) is

$$\left[0, \frac{(n-1)S^2}{G_{n-1}^{-1}(\alpha)}\right],$$

since

$$P_{\mu,\sigma^2}\Big(\sigma^2 \le \frac{(n-1)S^2}{G_{n-1}^{-1}(\alpha)}\Big) = P\Big(\frac{(n-1)S^2}{\sigma^2} \ge G_{n-1}^{-1}(\alpha)\Big) = 1-\alpha.$$

Numerical example continued

The sample size is $n = 6$. We take $\alpha = .05$. Then $G_{n-1}^{-1}(1-\frac{\alpha}{2}) = 12.83$ and $G_{n-1}^{-1}(\frac{\alpha}{2}) = .83$. The sample variance is $s^2 = .8$. So a 95% confidence interval for σ^2 is

$$.312 \le \sigma^2 \le 4.18$$

and so a 95% confidence interval for σ is

$$.56 = \sqrt{.312} \le \sigma \le \sqrt{4.18} = 2.19.$$

If one is interested in a upper bound for σ^2 we use that $G_{n-1}^{-1}(\alpha) = 1.145$. So a one-sided 95% confidence interval for σ^2 is

$$\sigma^2 \le 3.491$$

and a one-sided 95% confidence interval for σ is

$$\sigma \le \sqrt{3.491} = 1.868.$$

AD Example 10.19 Let $X \sim \text{Poisson}(\lambda)$.
We take $\alpha = .05$ and for simplicity replace $\Phi^{-1}(1-\frac{\alpha}{2}) = 1.96$ by 2.

Approximate confidence interval for λ using the CLT

For λ large, $(X-\lambda)/\sqrt{\lambda}$ is approximately $\mathcal{N}(0,1)$ distributed. Hence

$$P_\lambda\Big(\frac{|X-\lambda|}{\sqrt{\lambda}} \le 2\Big) \approx .95.$$

Rewrite this to

$$P_\lambda\Big(\lambda \in \Big[X+2-2\sqrt{X+1}, X+2+2\sqrt{X+1}\Big]\Big) \approx .95.$$

So

$$\Big[X+2-2\sqrt{X+1}, X+2+2\sqrt{X+1}\Big]$$

is an approximate 95% confidence interval.

Approximate confidence interval for λ using the CLT and estimated variance

We can estimate the variance by

$$\widehat{\text{Var}}(X) := X.$$

For λ large $X - \lambda/\sqrt{X}$ is approximately $\mathcal{N}(0,1)$-distributed (see e.g. *Fundamentals of Mathematical Statistics*). An approximate 95% confidence interval based on this is

$$[X - 2\sqrt{X}, X + 2\sqrt{X}].$$

The duality between confidence sets and tests

Let $X \in \mathcal{X}$, $X \sim P_\theta$, $\theta \in \Theta$ and let $\gamma := g(\theta) \in \mathbb{R}$ be a parameter of interest. Define $\Gamma := \{\gamma = g(\theta) : \ \theta \in \Theta\}$. Consider some set $C \subset \mathcal{X} \times \Gamma$ and let for $\gamma \in \Gamma$

$$A(\gamma) := \{x : \ (x, \gamma) \in C\} \subset \mathcal{X},$$

and for $x \in \mathcal{X}$

$$B(x) := \{\gamma : \ (x, \gamma) \in C\} \subset \Gamma.$$

(We assume that $A(\gamma)$ is measurable for all $\gamma \in \Gamma$.)

Duality Theorem (LN Theorem 6.4)
The set $B(X)$ is a $(1-\alpha)$-confidence set
$\Leftrightarrow$
For all $\gamma_0 \in \Gamma$, $\phi(X, \gamma_0) := \text{l}_{A^c(\gamma_0)}(X)$ is a level α test for $H_0 : \ g(\theta) = \gamma_0$.

Proof.

$$P_\theta\biggl(\phi(X,\gamma) = 1\biggr) = P_\theta\biggl(X \notin A(\gamma)\biggr)$$

$$= P_\theta\biggl((X,\gamma) \notin C\biggr) = 1 - P_\theta\biggl((X,\gamma) \in C\biggr)$$

$$= 1 - P_\theta\biggl(\gamma \in B(X)\biggr).$$

□

Example Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 =: \sigma_0^2$ known. We let $\gamma := \mu$. Then we may take

$$B(X_1, \ldots, X_n) = \biggl[\bar{X} - \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}, \bar{X} + \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}\biggr],$$

and then

$$A(\mu) = \biggl[\mu - \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}, \mu + \Phi^{-1}(1 - \tfrac{\alpha}{2})\sigma_0/\sqrt{n}\biggr].$$

Example 6.15 Consider $X \sim \text{Binomial}(n, \theta)$ with $0 \le \theta \le 1$ unknown. We present three ways for the construction of confidence intervals for θ.

$\boxed{\text{Exact confidence interval using the Duality Theorem}}$

For the hypothesis
$H_0 : \ \theta = \theta_0$,
we use the test

$$\phi(X, \theta_0) := \begin{cases} 1 & X > \bar{c}(\theta_0) \text{ or} X < \underline{c}(\theta_0) \\ 0 & \text{else} \end{cases},$$

where $\underline{c}(\theta_0) \le \bar{c}(\theta_0)$ (both in $\{0, \ldots, n\}$) are determined by

$$\underbrace{P_{\theta_0}\biggl(X > \bar{c}(\theta_0)\biggr)}_{=\sum_{k > \bar{c}(\theta_0)} \binom{n}{k} \theta_0^k (1-\theta_0)^{n-k}} \le \frac{\alpha}{2} \le P_{\theta_0}\biggl(X > \bar{c}(\theta_0) - 1\biggr)$$

$$P_{\theta_0}\Big(X < \underline{c}(\theta_0)\Big) \le \frac{\alpha}{2} \le P_{\theta_0}\Big(X < \underline{c}(\theta_0) + 1\Big).$$

So

$$A(\theta_0) = \{x \in \{0, \ldots, n\} : \underline{c}(\theta_0) \le x \le \bar{c}(\theta_0)\}$$

and

$$C = \{(x, \theta) \in \{0, \ldots, n\} \times [0, 1] : \ \underline{c}(\theta) \le x \le \bar{c}(\theta)\},$$

$$B(x) = \{\theta \in [0, 1] : \ \underline{c}(\theta) \le x \le \bar{c}(\theta)\}.$$

We let for $x \in \{0, \ldots, n-1\}$, $\bar{\theta}(x)$ be defined by

$$\sum_{k<x} \binom{n}{k} \bar{\theta}(x)^k (1 - \bar{\theta}(x))^{n-k} = \frac{\alpha}{2}$$

and for $x \in \{1, \ldots, n\}$, $\underline{\theta}(x)$ be defined by

$$\sum_{k>x} \binom{n}{k} \underline{\theta}(x)^k (1 - \underline{\theta}(x))^{n-k} = \frac{\alpha}{2}$$

and further take $\bar{\theta}(n) = 1$ and $\underline{\theta}(0) = 0$. Then $[\underline{\theta}(X), \bar{\theta}(X)]$ is an exact $(1 - \alpha)$-confidence interval for θ.

Approximate confidence interval using the CLT

We reject
$H_0 : \ \theta = \theta_0$,
when

$$\frac{|X - n\theta_0|}{\sqrt{n\theta_0(1 - \theta_0)}} > \underbrace{\Phi^{-1}(1 - \tfrac{\alpha}{2})}_{:=z}.$$

So

$$B(x) = \left\{\theta : \ \frac{|X - n\theta|}{\sqrt{n\theta(1-\theta)}} > z\right\}$$

$$= \left\{\theta \in \frac{x + \frac{z^2}{2}}{n + z^2} \pm \frac{\sqrt{\frac{z^2 x(n-x)}{n} + \frac{z^4}{4}}}{n + z^2}\right\}.$$

Approximate confidence interval using the CLT and estimated variance

By the CLT

$$\frac{X - n\theta}{\sqrt{\mathrm{Var}_\theta(X)}}$$

is approximately $\mathcal{N}(0, 1)$-distributed. We have $\mathrm{Var}_\theta(X) = n\theta(1 - \theta)$ which can be estimated by

$$\widehat{\mathrm{Var}}_\theta(X) := n\hat{\theta}(1 - \hat{\theta}).$$

Then

$$\frac{X - n\theta}{\sqrt{\widehat{\mathrm{Var}}_\theta(X)}}$$

is still approximately $\mathcal{N}(0,1)$-distributed (see for example *Fundamentals of Mathematical Statistics*). We can then take

$$B(x) := \left\{\theta \in \frac{x}{n} \pm z\sqrt{\frac{x}{n}\left(1-\frac{x}{n}\right)}/\sqrt{n}\right\}$$

$$= \left\{\theta \in \frac{x}{n} \pm \frac{\sqrt{\frac{z^2 x(n-x)}{n}}}{n}\right\}.$$

Numerical example

Let $n = 38$ and suppose we observe $X = 20$. Then, using the third method above, an approximate 95% confidence interval for θ (and using $\Phi^{-1}(.975) \approx 2$) is

$$\frac{20}{38} \pm 2\sqrt{\frac{20 \times 18}{38^3}} = .526 \pm .162.$$

The linear model

Consider n independent observations $Y_1, \ldots, Y_n$. This time we do not assume that they are identically distributed. Let $X \in \mathbb{R}^{n \times p}$ be a given matrix with (non-random) entries $\{x_{i,j} : \ i = 1, \ldots, n, \ j = 1, \ldots, p\}$. One calls X the design matrix. The fact that we assume it to be non-random means we consider the case of fixed design. We now look for the best linear approximation of Y_i given $x_{i,1}, \ldots, x_{i,p}$. We measure the fit using the residual sum of squares. This means that we minimize

$$\sum_{i=1}^{n} \left(Y_i - a - \sum_{j=1}^{p} x_{i,j} b_j \right)^2 .$$

over $a \in \mathbb{R}$ and $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$.

To simplify the expressions, we rename the quantities involved as follows. Define for all i, $x_{i,p+1} := 1$ and define $b_{p+1} := a$. Then for all i $a + \sum_{j=1}^{p} x_{i,j} b_j = \sum_{j=1}^{p+1} x_{i,j} b_j$. In other words, if we put in the matrix X a column containing only 1's then we may omit the constant a. Thus, putting the column of only 1's in front and replacing $p+1$ by p, we let

$$X := \begin{pmatrix} 1 & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

Then we minimize

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2 .$$

over $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$.

Let us denote the Euclidean norm of a vector $v \in \mathbb{R}^n$ by

$$\|v\|_2 := \sqrt{\sum_{i=1}^{n} v_i^2}.$$

Write

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Then

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2 = \|Y - Xb\|_2^2.$$

One calls

$$\hat{\beta} := \arg \min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2$$

the least squares estimator.

The distance between Y and the space $\{Xb : \ b \in \mathbb{R}^p\}$ spanned by the columns of X is minimized by projecting Y on this space. In fact, one has

$$\frac{1}{2}\frac{\partial}{\partial b}\|Y - Xb\|_2^2 = -X^T(Y - Xb).$$

It follows that $\hat{\beta}$ is a solution of the so-called normal equations

$$X^T(Y - X\hat{\beta}) = 0$$

or

$$X^TY = X^TX\hat{\beta}.$$

If X has rank p, the matrix X^TX has an inverse $(X^TX)^{-1}$ and we get

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

The projection of Y on $\{Xb : \ b \in \mathbb{R}^p\}$ is

$$\underbrace{X(X^TX)^{-1}X^T}_{\text{projection}}Y.$$

Recall that a projection is a linear map of the form PP^T such that $P^TP = I$. We can write $X(X^TX)^{-1}X^T := PP^T$.[6]

Example with $p = 1$
For $p = 1$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Then

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$\begin{aligned}(X^TX)^{-1} &= \left(n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2\right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\ &= \left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.\end{aligned}$$

Moreover

$$X^TY = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_iY_i \end{pmatrix}.$$

[6] Write the singular value decomposition of X as $X = P\phi Q^T$, where $\phi = \text{diag}(\phi_1, \ldots, \phi_p)$ contains the singular values and where $P^TP = I$ and $Q^TQ = I$.

We now let (changing notation: $\hat{\alpha} := \hat{\beta}_1$, $\hat{\beta} := \hat{\beta}_2$)

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T X)^{-1} X^T Y$$

$$= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}$$

$$= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 \bar{Y} - \bar{x} \sum_{i=1}^n x_i Y_i \\ -n\bar{x}\bar{Y} + \sum_{i=1}^n x_i Y_i \end{pmatrix}$$

$$= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 - \bar{x}(\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}) \\ \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} \end{pmatrix}.$$

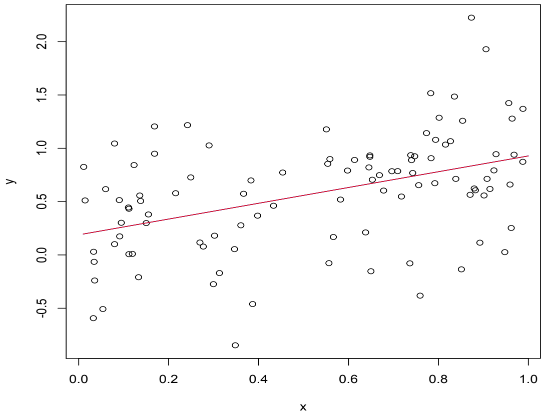
Here we used that $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2$. We can moreover write

$$\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

Thus

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{x} \\ \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}.$$

These expressions coincide with what we derived as method of moments estimators (see also LN Example 6.3). See also AD Example 11.18 for the theoretical counterpart.



Simulated data with $Y = .3 + .6 \times x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \frac{1}{4})$, $\hat{\alpha} = .19$, $\hat{\beta} = .740$

Definition *For $f = EY$ we let $\beta^* := (X^T X)^{-1} X^T f$ and we call $X\beta^*$ the best linear approximation of f.*

Lemma *Suppose $E\epsilon\epsilon^T = \sigma^2 I$. Then*
i) $E\hat{\beta} = \beta^*$, $\text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$,
ii) $E\|X(\hat{\beta} - \beta^*)\|_2^2 = \sigma^2 p$,
iii) $E\|X\hat{\beta} - f\|_2^2 = \underbrace{\|X\beta^* - f\|_2^2}_{\text{approximation error}} + \underbrace{\sigma^2 p}_{\text{estimation error}}$.

Proof.
i) By straightforward computation

$$\hat{\beta} - \beta^* = \underbrace{(X^T X)^{-1} X^T}_{:=B} \epsilon.$$

We therefore have

$$E(\hat{\beta} - \beta^*) = BE\epsilon = 0,$$

and the covariance matrix of $\hat{\beta}$ is

$$\text{Cov}(\hat{\beta}) = \text{Cov}(B\epsilon) = B \underbrace{\text{Cov}(\epsilon)}_{=\sigma^2 I} B^T$$

$$= \sigma^2 BB^T = \sigma^2 (X^T X)^{-1}.$$

ii) Define the projection $PP^T := X(X^T X)^{-1} X^T$. Then

$$\|X(\hat{\beta} - \beta^*)\|_2^2 = \|PP^T \epsilon\|_2^2 := \sum_{j=1}^p V_j^2,$$

where $V := P^T \epsilon$,

$$EV = P^T E\epsilon = 0,$$

and

$$\text{Cov}(V) = P^T \text{Cov}(\epsilon) P = \sigma^2 I.$$

It follows that

$$E \sum_{j=1}^p V_j^2 = \sum_{j=1}^p EV_j^2 = \sigma^2 p.$$

iii) It holds by Pythagoras' rule for all b

$$\|Xb - f\|_2^2 = \|X(b - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2$$

since $X\beta^* - f$ is orthogonal to X. □

Lemma *Suppose $\epsilon := Y - f \sim \mathcal{N}(0, \sigma^2 I)$. Then we have*
i) $\hat{\beta} - \beta^* \sim \mathcal{N}(0, \sigma^2 (X^T X)^{-1})$,
ii) $\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{\sigma^2} \sim \chi^2(p)$.

Proof.
i) Since $\hat{\beta}$ is a linear function of the multivariate normal ϵ, the least squares estimator $\hat{\beta}$ is also multivariate normal.
ii) Define the projection $PP^T := X(X^T X)^{-1} X^T$. Then

$$\|X(\hat{\beta} - \beta^*)\|_2^2 = \|PP^T \epsilon\|_2^2 := \sum_{j=1}^p V_j^2.$$

Now $V := P^T \epsilon$ has i.i.d. $\mathcal{N}(0, \sigma)^2$ entries. □

Remark More generally, many estimators are approximately normally distributed (for example the sample median) and many test statistics have approximately a χ^2 null-distribution (for example the χ^2 goodness-of-fit statistic). This phenomenon occurs because many models can in a certain sense be approximated by the linear model and many minus log-likelihoods resemble the least squares loss function. Understanding the linear model is a first step towards understanding a wide range of more complicated models.

Corollary *Suppose the linear model is well-specified: for some $\beta \in \mathbb{R}^p$*

$$EY = X\beta.$$

Assume $\epsilon := Y - EY \sim \mathcal{N}(0, \sigma^2)$. where $\sigma^2 := \sigma_0^2$ is known. Then a test for
$H_0 : \ \beta = \beta_0 \ \ ,$
is:
reject H_0 when $\|X(\hat{\beta} - \beta^0)\|_2^2/\sigma_0^2 > G_p^{-1}(1-\alpha)$,
where G_p is the CDF of a $\chi^2(p)$-distributed random variable.

Remark When σ^2 is unknown one may estimate it using the estimator

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|_2^2}{n-p},$$

where $\hat{\epsilon} := Y - X\hat{\beta}$ is the vector of residuals. Under the assumptions of the previous corollary (but now with possibly unknown σ^2) the test statistic $\|X(\hat{\beta} - \beta^0)\|_2^2/\hat{\sigma}^2$ has a so-called F-distribution with p and $n-p$ degrees of freedom.

High-dimensional statistics

Let $X_1, \ldots, X_n$ be i.i.d. (say) copies of $X \sim P_\theta$, $\theta \in \Theta \subset \mathbb{R}^p$. Thus, the number of parameters is p and the number of observations is n. In high-dimensional statistics, p is "large", possibly $p \gg n$. We consider here a prototype example, namely the linear model.

In the linear model one has data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathbb{R}^p$ a p-dimensional row vector and $Y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ and one wants to find a good linear approximation using the least squares loss function

$$b \mapsto \sum_{i=1}^n \biggl(Y_i - \sum_{j=1}^p X_{i,j} b_j \biggr)^2 ,$$

Define (with some clash of notation) the design matrix

$$X := \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,p} \end{pmatrix}$$

and the vector of responses

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} .$$

Then

$$\sum_{i=1}^n \biggl(Y_i - \sum_{j=1}^p X_{i,j} b_j \biggr)^2 = \| Y - Xb \|_2^2 .$$

If $p \ge n$ minimizing this over all $b \in \mathbb{R}^p$ gives a "perfect" solution $\hat\beta_{\rm LS}$ with $X\hat\beta_{\rm LS} = Y$. This solution just reproduces the data and is therefore of no use. We say that it overfits.

Definition *The ridge regression estimator is*

$$\hat\beta_{\rm ridge} := \arg\min_{b \in \mathbb{R}} \biggl\{ \| Y - Xb \|_2^2 + \lambda^2 \| b \|_2^2 \biggr\} ,$$

where $\lambda > 0$ is a regularization parameter.

Definition *The Lasso estimator is*

$$\hat\beta_{\rm Lasso} := \arg\min_{b \in \mathbb{R}} \biggl\{ \| Y - Xb \|_2^2 + 2\lambda \| b \|_1 \biggr\} ,$$

where $\lambda > 0$ is a regularization parameter and $\| b \|_1 := \sum_{j=1}^p |b_j|$ is the ℓ_1-norm of b .

Note Consider the model $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The ridge regression estimator is the MAP estimator using as prior $\beta_1, \ldots, \beta_p$ i.i.d. $\sim \mathcal{N}(0, \tau^2)$. The

Lasso estimator is the MAP using as prior $\beta_1, \ldots, \beta_p$ i.i.d. $\sim$ Laplace$(0, \tau^2)$. The tuning parameter is then in both cases $\lambda^2 = \sigma^2/\tau^2$.

Remark As λ grows the ridge estimator shrinks the coefficients. They will however not be set exactly to zero. The coefficients of the Lasso estimator shrink as well, and some - or even many - are set exactly to zero. The ridge estimator can be useful if p is moderately large. For very large p the Lasso is to be preferred. The idea is that one should not try to estimate something when the signal is below the noise level. Instead, then one should simply put it to zero.

Remark Both ridge estimator and Lasso are biased. As λ increases the bias increases, but the variance decreases.

Remark The regularization parameter λ is for example chosen by using "cross validation" or (information) theoretic or Bayesian arguments.

Lemma *The ridge estimator $\hat{\beta}_{\text{ridge}}$ is given by*

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda^2 I)^{-1} X^T Y.$$

Proof. We have

$$\frac{1}{2\frac{\partial}{\partial b}} \left\{ \|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2 \right\} = -X^T(Y - Xb) + \lambda^2 b = -X^T Y + \left(X^T X + \lambda^2 I \right) b.$$

The estimator $\hat{\beta}_{\text{ridge}}$ puts this to zero. □

For the Lasso estimator there is no explicit expression in general. We therefore only consider the special case of orthogonal design and that all columns in X have the same length. If X has i.i.d. rows, this assumption is not very likely, so we therefore assume X is non-random at this point. One calls this fixed design.

Lemma *Suppose X is a fixed design matrix and $X^T X = nI$ (thus $p \leq n$ necessarily). Define $Z := X^T Y$. Then for $j = 1, \ldots, p$*

$$\hat{\beta}_{\text{Lasso},j} = \begin{cases} Z_j/n - \lambda/n & Z_j \geq \lambda \\ 0 & |Z_j| \leq \lambda \\ Z_j/n + \lambda/n & Z_j \leq -\lambda \end{cases}.$$

Proof. Write $\hat{\beta}_{\text{Lass0}} =: \hat{\beta}$ for short. We can write

$$\|Y - Xb\|_2^2 = \|Y\|_2^2 - 2b^T X^T Y + nb^T b = -2b^T Z + nb^T b.$$

Thus for each j we minimize

$$-2b_j Z_j + nb_j^2 + 2\lambda |b_j|.$$

If $\hat{\beta}_j > 0$ it must be a solution of putting the derivative of the above expression to zero:

$$-Z_j + n\hat{\beta}_j + \lambda = 0,$$

or

$$\hat{\beta}_j = Z_j/n - \lambda/n.$$

Similarly, if $\hat{\beta}_j < 0$ we must have

$$-Z_j + n\hat{\beta}_j - \lambda = 0.$$

Otherwise $\hat{\beta}_j = 0$. □

Some notation

∘ For a vector $z \in \mathbb{R}^p$ we let $\|z\|_\infty := \max_{1 \le j \le p} |z_j|$ be its ℓ_∞-norm.

∘ For a subset $S \subset \{1, \ldots, p\}$ we let $X\beta_S^*$ be the best linear approximation of $f := EY$ using the variables in S, i.e., $X\beta_S^*$ is the projection in $\mathbb{R}^n$ of f on the linear space $\{\sum_{j \in S} X_{\cdot,j} b_{S,j} : b_S \in \mathbb{R}^{|S|}\}$.

In the next theorem we again assume orthogonal design. For general design, one needs so-called "restricted eigenvalues".

Theorem *Consider again fixed design with $X^TX = nI$. Let $f = EY$ and $\epsilon = Y - f$. Fix some level α and suppose that for some λ_α it holds that $P(\|X^T\epsilon\|_\infty > \lambda_\alpha) \le \alpha$. Then for $\lambda > \lambda_\alpha$ we have with probability at least $1 - \alpha$*

$$\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 \le \min_S \Bigg\{ \underbrace{\|X\beta_S^* - f\|_2^2}_{\substack{\text{approximation}\\ \text{error}}} + \underbrace{(\lambda + \lambda_\alpha)^2 |S|}_{\substack{\text{estimation}\\ \text{error}}} \Bigg\}.$$

Proof. Write $\hat{\beta} := \hat{\beta}_{\text{Lasso}}$ and $f = X\beta$. On the set where $\|X^T\epsilon\|_\infty \le \lambda_\alpha$ we have

- $n|\beta_j| > \lambda + \lambda_\alpha \Rightarrow n|\hat{\beta}_j - \beta_j| \le \lambda + \lambda_\alpha$,
- $n|\beta_j| \le \lambda + \lambda_\alpha \Rightarrow |\hat{\beta}_j - \beta_j| \le |\beta_j|$.

So with probability at least $(1 - \alpha)$,

$$\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 \le \sum_{n|\beta_j| \le \lambda + \lambda_\alpha} n\beta_j^2 + (\lambda + \lambda_\alpha)^2 \Bigg(\#\{j :\ n|\beta_j| > \lambda + \lambda_\alpha\} \Bigg)$$

$$= \min_S \Bigg\{ \|X\beta_S^* - f\|_2^2 + (\lambda + \lambda_\alpha)^2 |S| \Bigg\}.$$

□

Corollary *Suppose that $f = X\beta$ where β has $s := \#\{j :\ \beta_j \ne 0\}$ non-zero components. Then under the conditions of the above theorem, with probability at least $1 - \alpha$*

$$\|X(\hat{\beta}_{\text{Lasso}} - \beta)\|_2^2 \le (\lambda + \lambda_\alpha)^2 s.$$

The above corollary tells us that the Lasso estimator adapts to favourable situations where β has many zeroes (i.e. where β is sparse).

To complete the story, we need to study a bound for λ_α. It turns out that for many types of error distributions, one can take λ_α of order $\sqrt{\log p}$.

Remark. The value $\alpha = \frac{1}{2}$ thus gives a bound for the median of $\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2$. In the case of Gaussian errors one may use "concentration of measure" to deduce that $\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2$ is "concentrated" around its median.