STOR 654 Notes (F12)

Jimmy Jin
UNC-Chapel Hill

Last updated: 09/20/16

Contents

1 Introductory material

This section contains a review of topics which students are expected to have encountered before the semester.

1.1 Order statistics

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ with density f. Often it is useful to study the behavior of the smallest and largest values in such a sample, and by extension, the kth smallest and largest values in such a sample.

Definition. (Order statistics)

The **order statistics** of $\{X_1, \ldots, X_n\}$ are ordered values

$$\begin{aligned} X_{(1)} &= \text{smallest } X_i \\ X_{(k)} &= \text{k}^{th} \text{ smallest } X_i \\ X_{(n)} &= \text{maximum } X_i \end{aligned}$$

Remark. Note that since the X_i's are continuous and independent, $P(X_i = X_j) = 0 \; \forall i, j$. Therefore we have that

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$$

Given the setup above with *iid* r.v.'s with distribution function F and density f, we may be interested in the distribution of the rth order statistic $X_{(r)}$.

Theorem 1.1. (Distribution of the order statistic $X_{(r)}$)

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ with density f and let $X_{(r)}$ be the rth order statistic of $X_1, \ldots, X_n$. Then:

$$\begin{aligned} P(X_{(r)} \leq u) &= \sum_{t=r}^{n} \binom{n}{t} F(u)^t \left(1 - F(u)\right)^{n-t} \\ P(X_{(r)} = u) &= r\binom{n-1}{r-1} F(u)^{r-1} \left(1 - F(u)\right)^{n-r} f(u) \end{aligned}$$

Proof. To prove the first identity, note that $P(X_{(r)} \leq u)$ is the probability that at least r of the X_i are less than u and $n - r$ of the X_i are greater than u. Multiply the number of ways $\binom{n}{t}$ to arrange the points and then sum from r to n.

The proof for the second is similar. To get $X_{(r)} = u$, we must have:

1. $n - r$ greater than u, which happens with probability $(1 - F(u))^{n-r}$
2. $r - 1$ less than u, which happens with probability $F(u)^{r-1}$
3. exactly 1 equal to u, which happens with probability $f(u)$

By independence we can multiply the three probabilities above, and then multiply by a factor $r\binom{n-1}{r-1}$ which represents the number of ways to get the above arrangement.

□

Theorem 1.2. (Joint density of $X_{(1)}, \ldots, X_{(n)}$)

$$f_{(n)}(x_1, \ldots x_n) = n! \cdot f(x_1, \ldots, x_n) \cdot \mathbb{1}_{x_1 \leq x_2 \leq \ldots \leq x_n}$$

Proof. (n = 3)

Note that

$$\begin{aligned}
F_{(n)}(x_1, x_2, x_3) &= P(X_{(1)} \leq x_1, X_{(2)} \leq x_2, \ldots, X_{(n)} \leq x_n) \\
&= P(\text{all } X_i \leq x_1) \\
&\quad + P(2 \text{ of the } X_i \leq x_1, 1 \text{ of the } X_i \in (x_1, x_3]) \\
&\quad + P(1 \text{ of the } X_i \leq x_1, 2 \text{ of the } X_i \in (x_1, x_2]) \\
&\quad + P(1 \text{ of the } X_i \leq x_1, 1 \text{ of the } X_i \in (x_1, x_2], 1 \text{ of the } X_i \in (x_2, x_3]) \\
&= F(x_1)^3 \\
&\quad + 3\, F(x_1)^2 \left(F(x_3) - F(x_1)\right) \\
&\quad + 3\, F(x_1) \left(F(x_2) - F(x_1)\right)^2 \\
&\quad + 3!\, F(x_1) \left(F(x_2) - F(x_1)\right) \left(F(x_3) - F(x_2)\right)
\end{aligned}$$

Taking derivatives,

$$\begin{aligned}
f(x_1, x_2, x_3) &= \frac{\partial}{\partial x_1} \frac{\partial}{\partial x_2} \frac{\partial}{\partial x_3} F(x_1, x_2, x_3) \\
&= 3!\, f(x_1) f(x_2) f(x_3)
\end{aligned}$$

□

1.2 Stirling's approximation

This approximation provides important insight into the limiting relationship between certain random variables (perhaps most notably the relationship between the binomial and the Poisson).

Definition. (The gamma function)

The **gamma function** is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, \mathrm{d}u, \quad x > 0$$

And has the following associated properties:

1. $\Gamma(1) = 1$

2. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
3. $\Gamma(n) = (n-1)! \quad n \in \mathbb{N}$

To simplify the statement of the following results, we introduce some notation.

$$a(x) \sim b(x) \quad x > 0 \iff \frac{a(x)}{b(x)} \to 1 \quad \text{as} \quad x \to \infty$$

Theorem 1.3. (Stirling's approximation)

$$\Gamma(x+1) \sim \left(\frac{x}{e}\right)^x \sqrt{2\pi x}$$

Proof. Hard. □

Theorem 1.4.

$$n! = \left(\frac{x}{e}\right)^x \sqrt{2\pi x}\, e^{\alpha_n} \quad \text{where}$$

$$\frac{1}{12n+1} < \alpha_n < \frac{1}{12n}$$

Proof. Hard. □

1.3 Jensen's inequality

Jensen's inequality relates the expected value of a convex function to the convex function of the expected value.

Definition. (Convex function)

Let I be an open subset of $\mathbb{R}$. Let $g : I \to \mathbb{R}$. We say g is **convex** if, $\forall x, y \in I$ and $\forall \alpha \in [0,1]$ we have

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(x)$$

Concave functions are defined analogously.

Note that convexity can be phrased in terms of some other familiar concepts.

Remark. (Sufficient conditions for convexity)

Let I be an open subset of $\mathbb{R}$. Let $g : I \to \mathbb{R}$. If either of the following are true,

1. g' is nondecreasing and continuous on I
2. $g'' \geq 0$ on I.

then g is convex.

Remark. (Necessary conditions for convexity)

Let I be an open subset of $\mathbb{R}$. Let $g : I \to \mathbb{R}$. If g is convex, then

1. $g(x)$ is continuous on I.
2. $\forall u \in I$, there exists a support line $\ell(x) = ax + b$ such that
   (a) $g(u) = \ell(u)$
   (b) $g(x) \geq \ell(x) \quad \forall x \in I$

Theorem 1.5. (Jensen's inequality)

Let g be convex on I, an open subset of $\mathbb{R}$. Let $x \in I$ with probability 1 and assume $X, g(X)$ have finite expectation. Then

$$g(EX) \leq E(g(X))$$

Proof. First note that $EX \in I$. So let $\ell(x) = ax + b$ be the support line for $g(\cdot)$ at EX.

Then by the definition of support line we have

1. $\ell(EX) = g(EX)$
2. $\ell(x) \leq g(x)\ \forall x \in I$

Thus we have:

$$\begin{aligned} g(x) &\geq \ell(x) \quad \forall x \in I \\ Eg(x) &\geq E\ell(x) \\ &= E(ax + b) \\ &= aEX + b \\ &= \ell(EX) \\ &= g(EX) \end{aligned}$$

□

Jensen's inequality has many wide-ranging uses. One unexpected use is the proof of the fact that the geometric mean of a positive series is less than or equal to the arithmetic mean of the same series:

Theorem 1.6. Let $a_1, \ldots, a_n > 0$. Define the arithmetic mean and geometric mean by

$$\text{AM} = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad \text{GM} = \left(\prod_{i=1}^{n} a_i\right)^{1/n}$$

Then $\text{GM} \leq \text{AM}$.

Proof. Let X be a discrete random variable taking values in $\{a_i, \ldots, a_n\}$, with probabilities $\mathbb{P}(X = a_i) = 1/n$.

Clearly, $\mathbb{E}X = AM$. We show that log GM $\leq$ log AM. By Jensen,

$$\mathbb{E}(\log X) \leq \log \mathbb{E}X = \log \text{AM}$$

So we only need to show that $\mathbb{E}(\log X) = \log$ GM. Note:

$$\begin{aligned} \log \text{GM} = \log \left(\prod a_i\right)^{1/n} &= \frac{1}{n} \log(\prod a_i) \\ &= \frac{1}{n} \sum \log a_i \\ &= E \log X \end{aligned}$$

□

Another use of Jensen's inequality involves a result concerning the expected value of the log likelihood ratio of a discrete random variable:

Theorem 1.7. Let $X \in S \subset \mathbb{R}$ be a discrete random variable. Let p be the pmf of X and q be any other pmf defined on S.

We consider the likelihood functions $p(X)$ and $q(X)$. Noting that these are random variables, we then have that

$$E\left[\log \frac{p(X)}{q(X)}\right] \geq 0$$

Proof. First note that

$$\log \frac{p(x)}{q(x)} \geq 0 \iff p(x) \geq q(x)$$

Assume WLOG that $p(x) > 0 \; \forall x \in S$. Then

$$E\left[\log \frac{p(X)}{q(X)}\right] = \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)}$$

Note that this sum $= +\infty$ if $q(x) = 0$ for some $x \in S$. So assume that $q(x) > 0 \; \forall x \in S$. Then we have

$$\begin{aligned} \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)} &= -\sum_{x \in S} p(x) \log \frac{q(x)}{p(x)} \\ &= -E\left[\log \frac{q(x)}{p(x)}\right] \\ &\geq -\log E\left[\frac{q(x)}{p(x)}\right] \quad \textit{(by Jensen)} \\ &= -\log\left[\sum_{x \in S} q(x)\right] \\ &= -\log(1) = 0 \end{aligned}$$

□

1.4 Holder's inequality

Holder's inequality is an important tool which relates the expected value of a product to the product of expected values. The original statement of the inequality is more general and phrased in terms of integrals and the L_p-norm of a function. Here we present a more specific form in terms of expected value.

We first prove a simple lemma used in the proof of Holder's inequality.

Lemma 1.8. (Lemma for Holder's inequality)

Let $a, b \geq 0$ and $1 < p, q < \infty$ be such that $1/p + 1/q = 1$. Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality iff $a^p = b^q$.

Proof. First note that if either or both of a or b are zero, then the result is trivial. So assume $a, b > 0$.

$$\begin{aligned} ab &= \exp\left(\log a + \log b\right) \\ &= \exp\left(\frac{1}{p}\log a^p + \frac{1}{q}\log b^q\right) \\ &\leq \frac{1}{p}\exp\left(\log a^p\right) + \frac{1}{q}\exp\left(\log b^q\right) \quad \textit{(by convexity)} \\ &= \frac{1}{p}a^p + \frac{1}{q}b^q \end{aligned}$$

and since e^x is strictly convex, then equality holds only when $\log a^p = \log b^q$, i.e. when $a^p = b^q$.

□

To simplify the presentation of the statement and proof of the inequality, we first define a special case of the L_p-norm of a function:

Definition. (L_p-norm of a random variable)

For $p \geq 1$, the **L_p-norm of a random variable** X is

$$\|X\|_p = [E|X|^p]^{1/p}$$

Theorem 1.9. (Holder's inequality)

Let $1 < p, q < \infty$ satisfy the cosntraint $1/p + 1/q = 1$, and let X, Y be random variables such that $E|X|^p, E|Y|^q < \infty$. Then

$$|E(XY)| \leq E|XY| \leq \|X\|_p \|Y\|_q$$

Proof. Define

$$a = \frac{|X|}{\|X\|_p}, \quad b = \frac{|Y|}{\|Y\|_q}$$

Note that a, b are both non-negative for any $\omega \in \Omega$. Applying our lemma, we obtain

$$\frac{1}{p}\frac{|X|^p}{E|X|^p} + \frac{1}{q}\frac{|Y|^q}{E|Y|^q} \geq \frac{|XY|}{\|X\|_p \|Y\|_q}$$

Taking expectations of both sides, we obtain

$$1 = \frac{1}{p} + \frac{1}{q} \geq \frac{E|XY|}{\|X\|_p \|Y\|_q}$$

□

The well-known Cauchy-Schwarz inequality for random variables is a special case of Holder's inequality when $p = q = 2$:

Corollary 1.10. (Cauchy-Schwarz inequality)

Let X, Y be r.v.'s such that $EX^2, EY^2 < \infty$. Then

$$|E(XY)| \leq E|XY| \leq [EX^2]^{\frac{1}{2}}[EY^2]^{\frac{1}{2}}$$

1.5 Independent copies and association inequalities

The association inequalities are used in the context of two different functions of the same random variable. The inequalities related the expected value of the product of those functions to the product of the expected value of those functions.

Before presenting the inequalities, we review some basic definitions and properties behind the concept of independent copies, which will be used to prove the inequalities.

Definition. (Equal in distribution)

Let X, Y be random variables. X and Y are **equal in distribution** (written $X \stackrel{d}{=} Y$) if X and Y have the same distribution.

More specifically, let $X : \Omega \to \mathbb{R}$ with probability measure P and $Y : \Omega \to \mathbb{R}$ with probability measure P'. Then $X \stackrel{d}{=} Y$ iff

1. $F_X(u) = F_y(u) \quad \forall u \in \mathbb{R}$
2. $P(X \in B) = P'(Y \in B) \quad \forall B \in \mathbb{R}$

The key point is that random variables can be equal in distribution without being exactly equal, suggesting the concept of independent copies of a random variable. To illustrate this, consider the following example:

Example. Let X and X' be independent $N(0, 1)$ random variables. Consider the new random variable Y defined by:

1. $Y = X$. Then $Y \stackrel{d}{=} X$ and the variables are **not** independent.

2. $Y = X'$. Then $Y \stackrel{d}{=} X$ but the variables **are** independent.

Furthermore, we have an important result relating the variance and covariance of random variables to the expected value of a function of those random variables and their independent copies.

Theorem 1.11. Let X, Y be random variables with $EX^2, EY^2 < \infty$. Furthermore let X' be a copy of X which is independent of both X and Y. Then

1. $\text{Var}(X) = \frac{1}{2}E(X - X')^2$
2. $\text{Cov}(X) = E(XY - X'Y)$

Proof. Not given.

□

We now present the main result:

Theorem 1.12. (Association inequalities)

Let X be a random variable and let $f, g : \mathbb{R} \to \mathbb{R}$ such that $E|f(X)| < \infty$, $E|g(X)| < \infty$, and $E|f(X)g(X)| < \infty$. Then

1. If f, g are both nondecreasing (or nonincreasing), then
$$E[f(X)g(X)] \geq Ef(X)Eg(X)$$
2. If f is nondecreasing and g is nonincreasing (or v.v.), then
$$E[f(X)g(X)] \leq Ef(X)Eg(X)$$

Proof. We prove the case where f and g are both nondecreasing or both nonincreasing.

Let Y be an independent copy of X. Note that

$$0 \leq [f(x) - f(y)] \cdot [g(x) - g(y)]$$

since f and g are both nondecreasing (nonincreasing). Expanding the RHS and taking expectations, we have

$$0 \leq E[f(X)g(X)] - E[f(X)g(Y)] - E[f(Y)g(X)] + E[f(Y)g(Y)]$$

Note that since X is independent of Y and $X \stackrel{d}{=} Y$, then

1. $E[f(Y)g(Y)] = E[f(X)g(X)]$
2. $E[f(X)g(Y)] = E[f(Y)g(X)] = Ef(X)Eg(Y) = Ef(X)Eg(X)$

And the expression above reduces to

$$0 \leq 2\,E[f(X)g(X)] - 2\,Ef(X)Eg(X)$$

□

1.6 Markov and Chebyshev's inequalities

Markov's inequality, which is the basis for almost every other inequality we learn in this course, has a proof which is surprisingly simple. But first, we cover two basic properties of the indicator function and give an example of a proof in which it plays a key role. Then, we shall use indicator functions to give a two-line proof of Markov's inequality.

Remark. Let $\mathbb{1}_A(X), \mathbb{1}_B(X)$ be indicator functions of a random variable X where A, B are measurable sets. Then

1. $\mathbb{1}_A(X) \cdot \mathbb{1}_B(X) = \mathbb{1}_{A\cap B}(X)$
2. $E[\mathbb{1}_A(X)] = P(X \in A)$

Theorem 1.13. (Alternate integral form of expected value)

Let $X \geq 0$ be a random variable with density function f. Then

$$EX = \int_0^\infty P(X > t)\,\mathrm{d}t$$

Proof.

$$\begin{aligned}
\int_0^\infty P(X > t)\,\mathrm{d}t &= \int_0^\infty \left[\int_t^\infty f(x)\,\mathrm{d}x\right]\,\mathrm{d}t \\
&= \int_0^\infty \left[\int_0^\infty \mathbb{1}_{[t,\infty)}(x) f(x)\,\mathrm{d}x\right]\,\mathrm{d}t \\
&= \int_0^\infty f(x)\left[\int_0^\infty \mathbb{1}_{[t,\infty)}(x)\,\mathrm{d}t\right]\,\mathrm{d}x \quad \textit{(by Fubini)} \\
&= \int_0^\infty x f(x)\,\mathrm{d}x
\end{aligned}$$

□

We now arrive at the main result: Markov's inequality. Note that this inequality assumes nothing about the random variable at hand except for that it is non-negative.

Theorem 1.14. (Markov's inequality)

Let $X \geq 0$ be a random variable and let $t > 0$. Then

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

Proof. Note:

$$x \geq x \cdot \mathbb{1}_{(t,\infty)}(x) \geq t \cdot \mathbb{1}_{(t,\infty)}(x)$$

Taking expectations, we obtain:

$$\begin{aligned}
EX &\geq E[t \cdot \mathbb{1}_{(t,\infty)}(X)] \\
&= t \cdot P(X > t)
\end{aligned}$$

□

Markov's inequality is sometimes referred to as Chebyshev's inequality due to their similarity, but in this course we present Chebyshev's inequality as a separate bound for random variables **with finite variance**.

Theorem 1.15. (Chebyshev's inequality)

Let X be a random variable such that $EX^2 < \infty$ and let $t > 0$. Then

$$P(|X - EX| > t) \leq \frac{\text{Var}(X)}{t^2}$$

Proof. Note first that the random variable $|X - EX|^2$ is non-negative. So applying Markov,

$$\begin{aligned} P(|X - EX|^2 > t^2) &\leq \frac{E|X - EX|^2}{t^2} \\ &= \frac{\text{Var}(X)}{t^2} \quad \textit{(since } EX^2 < \infty\textit{)} \end{aligned}$$

Finally, noting that $P(|X - EX|^2 > t^2) = P(|X - EX| > t)$ gives the result. □

1.7 Probability integral transform

From undergraduate probability we know that a CDF is a non-decreasing function that is right-continuous. These conditions are not strong enough to guarantee that the function will always be invertible. However, we can define a psuedo-inverse function that will always exist for any given CDF:

Definition. (Quantile function/inverse CDF)

Let X be a random variable with CDF F. The **inverse CDF** of X is

$$F^{-1}(y) = \inf\{x : F(x) \geq y\} \quad \text{for } 0 < y < 1$$

Remark. F is non-decreasing, so the set $\{x : F(x) \geq y\}$ is of the form $[F^{-1}(y), \infty)$ or $(F^{-1}(y), \infty)$. But since F is right-continuous, then

$$\{x : F(x) \geq y\} = [F^{-1}(y), \infty)$$

and thus $\forall y \in (0, 1)$ and $x \in \mathbb{R}$, $F^{-1}(y) \leq x \iff F(x) \geq y$.

The inverse CDF gives the least value of x for which the probability of X being below x is at least y. One useful application of this is a method that generates values distributed according to any CDF only using values distributed Uniform$(0, 1)$.

Theorem 1.16. (Probability integral transform)

Let F be a **continuous** CDF. Then $Y \sim \text{Uniform}(0, 1) \Rightarrow F^{-1}(Y) \sim F$.

Proof. Let X be the random variable defined by $F^{-1}(Y)$. Then

$$\begin{aligned} F_X(k) &= P(F^{-1}(Y) \leq k) \\ &= P(Y \leq F(k)) \\ &= F(k) \quad (Y \sim \textit{Uniform}(0,1)) \end{aligned}$$

□

The proof in the opposite direction is has two cases:

Theorem 1.17. (Continuous case)

If X has continuous CDF F, then $\mathbb{P}(F(X) \leq y) = y$ (i.e. $F(X)$ is distributed Uniform$(0,1)$).

Proof. Consider y in the range of $F(X)$. Since F is increasing, F^{-1} is also increasing. Thus:

$$\begin{aligned} \mathbb{P}(F(X) \leq y) &= \mathbb{P}(F^{-1}(F(X)) \leq F^{-1}(y)) \quad (F^{-1} \text{ is increasing}) \\ &= \mathbb{P}(X \leq F^{-1}(y)) \quad (F \text{ is increasing}) \\ &= F(F^{-1}(y)) \\ &= y \quad (y \text{ is in the range of } F) \end{aligned}$$

□

Remark. We elaborate on the equality used above:

$$\mathbb{P}(F^{-1}(F(X)) \leq F^{-1}(y)) = \mathbb{P}(X \leq F^{-1}(y))$$

If F is strictly increasing, then it is clear that $F^{-1}(F(x)) = x$. If it is not, then it might be true that $F^{-1}(F(x)) \neq x$, but then we are in a set of measure zero so the result remains.

Theorem 1.18. (Discrete (general) case)

If X has discrete or continuous CDF F, then $\mathbb{P}(F(X) \leq y) \leq y$

Proof. Let y be in the range of $F(X)$. We show that there may exist a point such that $\mathbb{P}(F(X) \leq y) < y$.

Let y be a value in the range of F that is "jumped over" by F. Then since CDF's are right-continuous, there is a point x_0 such that $F(x_0 - \epsilon) < y < F(x_0)$ for any $\epsilon > 0$. In other words:

$$\{x : F(x) \leq y\} = \{x : F(x) < y\}$$

Follow the steps of the above proof with the strict inequality to arrive at the result.

□

1.8 Moment generating functions

Some basic properties of moment generating functions which the reader should already be familiar with:

Lemma 1.19. Let X be a random variable. If $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$, then $E[e^{sx}] < \infty$ for all $-s_0 \leq s \leq s_0$.

Proof. Properties of the integral.

Definition. (Moment generating function)

Let X be a random variable such that $E[e^{s_0|x|}] < \infty$ for some $s_0 > 0$. Then the **moment generating function of X** is

$$\phi_X(s) = E[e^{sx}] \quad |s| \leq s_0$$

Example. (Some common MGFs)

1. $X \sim$ Poisson(λ): $\phi_X(s) = e^{\lambda(e^s - 1)} \quad s \in \mathbb{R}$
2. $X \sim$ Exponential(λ): $\phi_X(s) = \lambda/\lambda - s \quad s < \lambda$
3. $X \sim$ Normal$(0, 1)$: $\phi_X(s) = e^{s^2/2} \quad s \in \mathbb{R}$

The moment generating function of a random variable has many nice properties and can be used in quite sophisticated ways in proofs. However, we won't go into those in depth here. We finish by presenting a few simple properties.

Remark. (Taylor expansion of MGF)

If $\phi_X(s)$ exists, then $E|X|^k < \infty \; \forall k \geq 1$ and

$$\phi_X(s) = \sum_{k \geq 0} \frac{s^k E X^k}{k!}$$

Remark. (Differentiation of the MGF)

$$\phi'_X(s) = \frac{d}{ds} E[e^{sX}] = E\left[\frac{d}{ds} e^{sX}\right] = E\left[Xe^{sX}\right]$$

and, more generally,

$$\phi_X^{(k)}(s) = E\left[X^k e^{sX}\right]$$

Remark. (Properties of the MGF)

1. $\phi_X^{(k)}(0) = E[X^k]$
2. $\phi''_X(s) \geq 0 \Rightarrow \phi_X(s)$ is convex

We close this section with a brief example concerning sums of independent random variables:

Example. (Sums of independent random variables)

Let $X_1, \ldots, X_n$ be independent where each X_i has an MGF. Define $S_n = X_1 + \ldots + X_n$. Then $\phi_{S_n}(s)$ exists:

$$\begin{aligned}
\phi_{S_n}(s) = E[e^{s \cdot S_n}] &= E\left[s(X_1 + \ldots + X_n)\right] \\
&= E\left[\prod_{i=1}^{n} e^{sX_i}\right] \\
&= \prod_{i=1}^{n} E[e^{sX_i}] \\
&= \prod_{i=1}^{n} \phi_{X_i}(s)
\end{aligned}$$

2 Framework for statistical inference

One annoying fact about statistics is that it contains so many different types of analyses–each with their own classes of techniques. Therefore in order to be able to study statistics in a general way, it is necessary to develop a language flexible enough to encompass all these techniques.

In this course we adopt the language of decision theory to describe the problem of statistical inference.

2.1 Decision-theoretic elements

Broadly speaking, the elements of statistical inference fall into three categories: **setup**, **inference**, and **assessment**. We introduce each category in order.

1. **Setup**

   In order to perform inference there needs to be a probabilistic setup over which we can make inferences. The setup determines now only what models we are considering but also which *parameters* we are interested in making inferences about. The basic elements of this setup are:

   (a) **Sample space** $\mathcal{X}$: the set of possible observations

   (b) **Model** $P = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$: family of probability measures on $\mathcal{X}$

   (c) **Parameter space** Θ: set of possible parameter values

   (d) **Data** $\boldsymbol{x} \in \mathcal{X}$: values from a random variable X with $X \sim P_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Theta$

   Note that we use the "script X" $\mathcal{X}$ to denote the sample *space*, and use the non-script X to denote the *random variable* which generates our data values $\boldsymbol{x}$ that live in $\mathcal{X}$. It is common not to explicitly define X in this way (only distinguishing the sample space $\mathcal{X}$ from the sample $\boldsymbol{x}$), but we do so here in order not to confuse values of the random variable with the variable itself.

   Furthermore, note that while we use bold $\boldsymbol{x}$ to convey that our data is a vector of n observations, we do not extend that notation to X although technically it should also be denoted as a random vector. For most of the rest of the course, the elements of the data $\boldsymbol{x}$ are assumed to be i.i.d. and therefore will be generated by the same univariate random variable X. Nevertheless, in some places bold $\boldsymbol{X}$ will be used to resolve ambiguity.

   This setup foreshadows the general inference problem to come. In general, we know $\mathcal{X}, P$, and Θ and want to make inference regarding $\boldsymbol{\theta} \in \Theta$ after receiving information as $\boldsymbol{x} \in X$.

   Remark. For each continuous $P_{\boldsymbol{\theta}}$ with pdf $f_{\boldsymbol{\theta}}$, we write

   $$P = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\} \quad \text{or} \quad \{f(\boldsymbol{x} \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

Example. (Bernoulli trials)

$$\begin{aligned}\mathcal{X} &= \{0,1\}^n \\ \Theta &= (0,1) \\ P &= \{P_\theta^n, \theta \in \Theta\} \\ P_\theta &= \text{ Bernoulli}(\theta)\end{aligned}$$

Example. (Normal mean and variance)

$$\begin{aligned}\mathcal{X} &= \mathbb{R}^n \\ \Theta &= \mathbb{R} \times [0,\infty) \\ P &= \{P_{\boldsymbol{\theta}}^n, \boldsymbol{\theta} \in \Theta\} \\ P_{\boldsymbol{\theta}} &= \text{ Normal}(\boldsymbol{\theta})\end{aligned}$$

2. **Inference**

Once we have setup the problem and have decided which parameters we are interested in, we must set out *how* to go about turning our data into a conclusion. The basic elements of the inference step are:

(a) **Action space** A: the set of possible actions (i.e. conclusions)

(b) **Decision rule** $\boldsymbol{d} : \mathcal{X} \to A$

Here, the spaces A and $\mathcal{X}$ maybe be several-dimensional but we do not boldface them. We do however boldface $\boldsymbol{d}$ to emphasize that the decision rule may be vector-valued.

To illustrate this more concretely, we give some examples of general action spaces for the three main types of statistical inference, and then some example decision rules for the problem of point estimation with Bernoulli trials.

Example. (Some action spaces)

(a) Point estimation:

- $A = \Theta$
- $\boldsymbol{d}$ = an estimator of $\boldsymbol{\theta} \in \Theta$

(b) Hypothesis testing

- $A = \{a_0, a_1\}$
- a_0 = decide $\boldsymbol{\theta} \in \Theta_0$
- a_1 = decide $\boldsymbol{\theta} \in \Theta_1$

(c) Interval estimation

- A = collection of subsets of Θ

Example. (Bernoulli trials)

$X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoilli(θ) where $\theta \in (0,1)$.

Assume our goal is point estimation of θ. Then some valid candidate decision rules include:

(a) $d_1(x) = \overline{x}$

(b) $d_2(x) = 1/2$

(c) $d_3(x) = x_1$

(d) $d_4(x) = (x_1 + x_2)/2$

3. **Assessment**

   Note that in the above inference step, we did not specify any way to gauge how *good* the conclusions arising from our inference were. Since we are generally interested in "goodness" of inference, we add two final elements to the process:

   (a) **Loss function** $L : \Theta \times A \to \mathbb{R}$

   $L(\boldsymbol{\theta}, a)$ = the loss incurred by taking action a when the state of nature is $\boldsymbol{\theta}$.

   (b) **Risk** $R : \Theta \times d \to \mathbb{R} = E_{\boldsymbol{\theta}}\left[L(\boldsymbol{\theta}, d(\boldsymbol{X}))\right]$

   $d : \mathcal{X} \to A$, so $R(\boldsymbol{\theta}, d)$ = the expected loss of decision rule d when the state of nature is $\boldsymbol{\theta}$.

   The expected loss is either an integral or sum over $\mathcal{X}$ according as whether $\boldsymbol{X}$ is continuous or discrete.

2.2 Types of statistical inference

In general, all statistical inference falls into one of three categories: **point estimation**, **hypothesis testing**, or **interval estimation**. We use the language and elements set out above to show how these different categories actually share common decision-theoretic elements.

1. **Point estimation**:

   The goal in point estimation is to estimate $\boldsymbol{\theta}$.

$$A = \Theta$$
$$d : \mathcal{X} \to \Theta$$

   Within this context it is possible to have several different loss functions. Here are two common examples for the case when $\Theta \subset \mathbb{R}^1$:

   (a) Squared error loss

$$L(\theta, a) = (\theta - a)^2$$
$$R(\theta, d) = E_\theta\left[(\theta - d(x))^2\right]$$

(b) Absolute error loss

$$L(\theta, a) = |\theta - a|$$
$$R(\theta, d) = E_\theta|\theta - d(x)|$$

2. **Hypothesis testing**

In hypothesis testing we consider a partition of $\Theta = \Theta_0 \cup \Theta_1$. Our goal is to decide whether $\boldsymbol{\theta} \in \Theta_0$ or $\boldsymbol{\theta} \in \Theta_1$.

$$A = \{\text{decide } \Theta_0, \text{ decide } \Theta_1\}$$
$$d : \mathcal{X} \to \{\text{decide } \Theta_0, \text{ decide } \Theta_1\}$$

A common loss function for hypothesis testing is the zero-one loss:

(a) Zero-one loss

We define the zero-one loss $L(\boldsymbol{\theta}, a)$ with the following table:

	$\boldsymbol{\theta} \in \Theta_0$	$\boldsymbol{\theta} \in \Theta_1$
$a =$ decide Θ_0	$L = 0$	$L = 1$
$a =$ decide Θ_1	$L = 1$	$L = 0$

$$R(\boldsymbol{\theta}, d) = E_{\boldsymbol{\theta}}\left[L(\boldsymbol{\theta}, d(x))\right]$$
$$= \begin{cases} P_{\boldsymbol{\theta}}(d(x) = \text{decide } \Theta_1), & \boldsymbol{\theta} \in \Theta_0 \\ P_{\boldsymbol{\theta}}(d(x) = \text{decide } \Theta_0), & \boldsymbol{\theta} \in \Theta_1 \end{cases}$$

Note that the zero-one loss is simply an indicator function over regions of x for which we decide Θ_0 or Θ_1. Therefore the expected value of the loss is equal to the probability of hitting the x values that correspond to deciding Θ_0 or Θ_1.

3. **Interval estimation**

In interval estimation we also consider a division of Θ: our goal is to find $\hat{\Theta} \subset \Theta$ that is likely to contain $\boldsymbol{\theta}$.

$$A = 2^\Theta \quad \text{(i.e. subsets of } \Theta)$$
$$d : \mathcal{X} \to A$$

A natural loss function for this problem is one which somehow considers both the width of the interval estimate and the accuracy of the interval.

(a) Interval estimator loss (1-dimensional case)

$$\begin{aligned} L(\theta, a) &= \mathbb{1}_{a^c}(\theta) + \lambda \int_a \mathrm{d}\theta, \quad \lambda > 0 \text{ fixed} \\ R(\theta, d) &= E_\theta \left[L(\theta, d(x)) \right] \\ &= P_\theta(\theta \notin d(x)) + \lambda E_\theta \left[\int_a \mathrm{d}\theta \right] \end{aligned}$$

Note that the $\mathbb{1}_{a^c}(\theta)$ terms assigns a penalty of 1 if the interval does not contain θ, and that the $\lambda \int_a \mathrm{d}\theta$ term measures the width of the interval estimate.

In other words, the loss consists of a penalty for not containing the true θ value, as well as an additional penalty which increases with the width of the estimated interval.

2.3 Admissibility and other notes

At this point we have enough tools to crudely select a decision rule: given two decision rules and a certain loss function, we can simply choose the one with lower risk. But since there can be infinitely many decision rules, this may not be the best way to go about picking the *best* decision rule.

Definition. (Admissible decision rule)

We say a decision rule $\boldsymbol{d} \in \mathcal{D}$ is **inadmissible** if $\exists\, \boldsymbol{d'} \in \mathcal{D}$ such that

$$\begin{aligned} &\forall \boldsymbol{\theta},\ R(\boldsymbol{\theta}, \boldsymbol{d'}) \leq R(\boldsymbol{\theta}, \boldsymbol{d}) \\ &\exists \boldsymbol{\theta}_0 \text{ s.t. } R(\boldsymbol{\theta}_0, \boldsymbol{d'}) < R(\boldsymbol{\theta}_0, \boldsymbol{d}) \end{aligned}$$

otherwise d is **admissible**.

Remark. Admissibility depends on $\mathcal{D}$ and $L(\theta, d)$.

Admissible decision rules are natural candidates for "good" decision rules. But admissibility does have its faults. The following simple example will highlight some of these.

Example. (Bernoulli trials)

We have the following usual setup for n Bernoulli trials.

$$\begin{aligned} \mathcal{X} &= \{0, 1\}^n \\ P &= \{P_\theta^n, \theta \in \Theta\} \\ P_\theta &= \text{ Bernoulli}(\theta) \\ \Theta &= (0, 1) \\ \boldsymbol{x} &= (x_1, x_2, \ldots, x_n), \quad X_i \overset{iid}{\sim} \text{ Bernoulli}(\theta) \end{aligned}$$

And the type of inference we wish to perform is point estimation of θ with squared error loss:

$$\begin{aligned} A = \Theta &= (0,1) \\ \mathcal{D} &= \text{all functions } d : \{0,1\}^n \to (0,1) \\ L(\theta, a) &= (\theta - a)^2 \end{aligned}$$

We propose the following candidate decision rules and analyze their admissibility and overall performance:

1. $d_1(x) = \overline{x}$
2. $d_2(x) = 1/2$
3. $d_3(x) = x_2$

Their corresponding risk functions are:

1. $R(\theta, d_1) = E_\theta(\theta - \overline{x})^2 = \text{Var}(\overline{x})$
2. $R(\theta, d_2) = E_\theta(\theta - \frac{1}{2})^2 = (\theta - \frac{1}{2})^2$
3. $R(\theta, d_3) = E_\theta(\theta - x_2)^2 = \text{Var}(x_2)$

We note two key points. First, $R(\theta, d_1) < R(\theta, d_3)$ for any values of θ, so d_3 is obviously inadmissible.

Second, even though d_2 disregards the data completely, it outperforms d_1 when $\theta = \frac{1}{2}$. This can be seen by comparing the risk functions: for values of θ away from $\frac{1}{2}$, $R(\theta, d_2)$ is much larger than $R(\theta, d_1)$. However, at $\theta = \frac{1}{2}$, $R(\theta, d_2) = 0$. Therefore neither of d_1 or d_2 are admissibile.

Finally, to close out our discussion of the decision-theoretic framework of inference, we couch two other major, non-classical approaches to inference in decision-theoretic terms (we consider the one-dimensional case for each):

1. **Bayesian**

   Fix a prior distribution on Θ : $\pi(\theta)$. Then find $d^* \in \mathcal{D}$ that minimizes the *expected* (or *Bayes'*) risk:

   $$R_\pi(\theta, d) = E_\pi[R(\theta, d)] = \int_\Theta R(\theta, d)\, \pi(\theta)\, \mathrm{d}\theta$$

2. **Minimax** Find $d^* \in \mathcal{D}$ that minimizes the maximum risk:

   $$d^* = \arg\min_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} R(\theta, d) \right]$$

2.4 Three basic families of distributions

As part of understanding the basic framework of inference, it is advantageous to familiarize oneself with classes ("families") of distributions that share certain prop-

erties. Each of these families share similarities in their function form which will simplify certain tasks to come later (e.g. finding pivotal quantities).

These families are the **location family**, the **scale family**, and the **location-scale family**.

Before we introduce these families, we discuss a basic representation theorem (which will return later when we discuss representation of multivariate Normal random variables):

Theorem 2.1. (Representation theorem)

Let X be a continuous random variable with density f. Then for every $\mu \in \mathbb{R}$ and $\sigma > 0$,

$$Y = \sigma X + \mu \text{ has density } f_y(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$$

Proof. CDF method.

□

1. **Location family**

   Let f be a density. Then the **location family generated by** f is

   $$P = \{f_\theta(\cdot) = f(\cdot - \theta) : \theta \in \mathbb{R}\}$$

   If f is the density of a location family r.v. X, then by the representation theorem:

   $$P = \{\text{distributions of random variables of form } (X + \theta) : \theta \in \mathbb{R}\}$$

   Example. (Normal distribution with fixed variance)

   Let f be the pdf of a $N(0, \sigma^2)$ random variable. Then

   $$P = \{N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$$

   is the location family generated by f.

   Example. (Uniform distribution)

   Let f be the pdf of a Uniform$(0, 1)$ random variable. Then

   $$P = \{\text{Uniform}(\theta, 1 + \theta) : \theta \in \mathbb{R}\}$$

   is the location family generated by f.

2. **Scale family**

   Let f be a density. Then the **scale family generated by** f is

   $$P = \left\{f_\theta(\cdot) = \frac{1}{\theta} f\left(\frac{\cdot}{\theta}\right) : \theta \in \mathbb{R}\right\}$$

   If f is the density of a scale family r.v. X, then by the representation theorem:

   $$P = \{\text{distributions of random variables of form } (X\theta) : \theta \in \mathbb{R}\}$$

Example. (Normal distribution with fixed mean)

Let f be the pdf of a $N(\mu, 1)$ random variable. Then

$$P = \{N(\theta\mu, \theta^2) : \theta > 0\}$$

is the scale family generated by f.

Example. (Uniform distribution)

Let f be the pdf of a Uniform$(0, 1)$ random variable. Then

$$P = \{\text{Uniform}(0, \theta) : \theta > 0\}$$

is the scale family generated by f.

3. **Location-scale family**

   Let f be a density. Then the **location-scale family generated by** f is

$$P = \left\{ f_{\mu,\sigma}(\cdot) = \frac{1}{\sigma} f\left(\frac{\cdot - \mu}{\sigma}\right) : \mu \in \mathbb{R}, \sigma > 0 \right\}$$

2.5 Exponential family

Definition. A family of pdfs or pmfs of a random variable $X \in \mathcal{X} \subset \mathbb{R}^p$ is an **exponential family** if

$$f_{\boldsymbol{\theta}}(x) = h(x) \exp\left\{ \left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) T_i(x) \right] - c(\boldsymbol{\theta}) \right\}, \quad \boldsymbol{\theta} \in \Theta$$

$$\begin{aligned} k &: \text{ order of the family } \geq 1 \\ h &: x \to [0, \infty) \\ w_i &: \Theta \to \mathbb{R} \\ T_i &: x \to \mathbb{R} \\ c &: \Theta \to \mathbb{R} \end{aligned}$$

We allow θ to be multidimensional but restrict ourselves to the case of a univariate random variable. A few important properties of the exponential family include:

1. (Common support)

   For a given exponential family, the $h(x)$ function is constant across $\boldsymbol{\theta}$. Also, $f_{\boldsymbol{\theta}}(x) > 0 \iff h(x) > 0$. Therefore, all members of the family (indexed by $\boldsymbol{\theta}$) have the same support.

2. (Normalizing constant)

Since $f_{\boldsymbol{\theta}}$ is a density, then $c(\boldsymbol{\theta})$ can be viewed as a normalizing constant. Note that

$$1 = \int_X f_{\boldsymbol{\theta}}(x)\,\mathrm{d}x = e^{-c(\boldsymbol{\theta})} \int_X h(x) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) T_i(x)\right\}\,\mathrm{d}x$$

so that

$$c(\boldsymbol{\theta}) = \log \int_X h(x) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) T_i(x)\right\}\,\mathrm{d}x$$

and therefore the integral on the right side must be finite also.

3. (Specification is not unique)

   Note that the the specification of the functions h, c, w_i, T_i are not unique. For example:

$$w_i(\boldsymbol{\theta})T_i(x) = \frac{w_i(\boldsymbol{\theta})}{\alpha_i} T_i(x)\alpha_i, \quad \alpha \neq 0$$

4. (Sufficiency of $\boldsymbol{T}$)

   It can be seen from looking at the exponential family density that all the interaction between the the data x and the parameter $\boldsymbol{\theta}$ is captured in the term

$$\sum_i w_i(\boldsymbol{\theta}) T_i(x)$$

   It can be shown that $\boldsymbol{T} = (T_1, \ldots, T_k)$ is related to the sufficient statistic for $\boldsymbol{\theta}$ where the dimension of the sufficient statistic (but not necessarily $\boldsymbol{T}$) is equal to the number of paramters.

5. (Inner product representation)

   The exponential family density can be written in a cleaner way using the inner product notation:

$$f_{\boldsymbol{\theta}}(x) = h(x) \exp\left\{\langle \boldsymbol{w}(\boldsymbol{\theta}), \boldsymbol{T}(x)\rangle - c(\boldsymbol{\theta})\right\}$$

   We now give some examples of exponential family distributions, first written in their usual way and then rearranged to emphasize the exponential family structure:

   Example. (Binomial distribution with fixed $n \geq 1$)

$$P = \{f_\theta = \text{ Binomial}(n, \theta) : \ 0 < \theta < 1\}$$

$$\begin{aligned} f_\theta &= \binom{n}{x} \theta^x (1-\theta)^{n-x}, \ x = 0, \ldots, n \\ &= \binom{n}{x} \mathbb{1}_{\{0,\ldots,n\}}(x)\, (1-\theta)^n \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right)\right\} \end{aligned}$$

Example. (Poisson distribution)

$$P = \{f_\theta = \text{ Poisson}(\theta) : \ \theta > 0\}$$

$$\begin{aligned} f_\theta &= \frac{\theta^x e^{-\theta}}{x!} \\ &= \frac{1}{x!} \exp\{x \log(\theta) - \theta\} \end{aligned}$$

Example. (Normal distribution)

$$P = \{f_\theta = N(\mu, \sigma^2) : \ \mu \in \mathbb{R}, \sigma^2 > 0\}$$

$$\begin{aligned} f(x \,|\, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\} \end{aligned}$$

Finally we conclude with a theorem about the joint distribution of several independent random variables belonging to the exponential family:

Theorem 2.2. Let $X_1, \ldots, X_n$ be iid. If the distribution of X_i belongs to the exponential family, so does the joint distribution of $X_1, \ldots, X_n$.

Proof. Let X_i have an exponential family distribution with density $f_\theta, \ \theta \in \Theta$. Then by independence of the X_i's, the joint distribution $g_\theta(x)$ is given by

$$\begin{aligned} g_\theta(x) &= f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n) \\ &= \left[\prod_{i=1}^n h(x_i)\right] \exp\left\{\langle \boldsymbol{w}(\boldsymbol{\theta}), \sum_{i=1}^n \boldsymbol{T}(x_i)\rangle - nc(\boldsymbol{\theta})\right\} \end{aligned}$$

Which is clearly in exponential family form. □

Note that the index on the product over h and the sum in the inner product is with respect to the n random variables, not with respect to the order of the family k.

2.6 Canonical exponential family

When an exponential family representation is reparameterized so that the function of the parameter w_i is written simply as $w_i(\theta) = \eta_i$, then the representation is said to be in **canonical form**. We state this more formally as a theorem:

Theorem 2.3. Let $\boldsymbol{\eta} \in \mathbb{R}^k$, $h : \mathbb{R}^p \to [0, \infty)$ and $T : \mathbb{R}^p \to \mathbb{R}^k$. Define

$$A(\boldsymbol{\eta}) = \log \int_X h(x) \exp\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\} \, \mathrm{d}x$$

If $A(\boldsymbol{\eta})$ is finite, then the following is a pdf/pmf:

$$f_{\boldsymbol{\eta}}(x) = h(x) \exp\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle - A(\boldsymbol{\eta})\}$$

Proof. Trivial.

Note that we can write the canonical EF form as:

$$f_{\boldsymbol{\eta}}(x) = \frac{h(x)\exp\left\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x)\rangle\right\}}{\exp\{A(\boldsymbol{\eta})\}}$$

Now note that if $A(\boldsymbol{\eta})$ is **not** finite, then we cannot make the numerator such that the integral of the entire fraction will equal 1. Therefore $f_{\boldsymbol{\eta}}$ will **not** be a pdf/pmf. With this in mind, we define the natural parameter space:

Definition. (Canonical EF, natural parameter space)

With the functions $A(\boldsymbol{\eta})$ and $f_{\boldsymbol{\eta}}$ defined above, define

$$H = \{\boldsymbol{\eta} : A(\boldsymbol{\eta}) < \infty\} \quad \text{and} \quad P = \{f_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in H\}$$

We call H the **natural parameter space of** P and P the **canonical exponential family generated by** h **and** $\boldsymbol{T}$.

Recall that the original formulation of an exponential family distribution was:

$$f_{\boldsymbol{\theta}}(x) = h(x)\exp\left\{\langle \boldsymbol{w}(\boldsymbol{\theta}), \boldsymbol{T}(x)\rangle - c(\boldsymbol{\theta})\right\}$$

From this it is clear that it is *always* possible to put such a formulation into canonical form by way of the transformation $\boldsymbol{\eta} = \boldsymbol{w}(\boldsymbol{\theta})$. Therefore

$$P_{\boldsymbol{\theta}} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \subset P_{\boldsymbol{\eta}} = \text{ the canonical EF generated by } h, \boldsymbol{T}$$

We give a simple example of the canonical form using the exponential family form of the Normal distribution given in the previous subsection:

Remark. (Guan's rule)

Note that since $\boldsymbol{\eta} = \boldsymbol{w}(\boldsymbol{\theta})$, then the natural parameter space is also the original range of $\boldsymbol{w}$ (which makes the density integrate to 1).

Example. (Normal exponential family)

We parameterize the familly $P_\theta = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ with

$$h(x) = 1,\ T_1(x) = x \text{ and } T_2(x) = x^2$$

Then, using the transformation given above,

$$f_\eta(x) = \exp\{\eta_1 x + \eta_2 x^2 - A(\boldsymbol{\eta})\}$$
$$A(\boldsymbol{\eta}) = \log \int_{\mathbb{R}} \exp\{\eta_1 x + \eta_2 x^2\}\,\mathrm{d}x$$

Note that $A(\boldsymbol{\eta}) < \infty$ if and only if $\eta_2 < 0$. Thus

$$H = \mathbb{R} \times (-\infty, 0)$$

We close with a useful property regarding the natural parameter space H:

Definition. (Convex set in $\mathbb{R}^n$)

We define scalar multiplication in the usual way: for $s \in \mathbb{R}$ and $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{R}^n$, we have

$$s\boldsymbol{u} = (su_1, \ldots, su_2)$$

We say $C \subset \mathbb{R}^n$ is **convex** if $\forall \boldsymbol{u}, \boldsymbol{v} \in C$ and $0 \leq s \leq 1$,

$$s\boldsymbol{u} + (1-s)\boldsymbol{v} \in C$$

Theorem 2.4. (Convexity of the natural parameter space)

Let P be a canonical exponential family generated by $h, \boldsymbol{T}$ with natural parameter space H. Then

1. H is convex
2. $A(\boldsymbol{\eta}) : H \to \mathbb{R}$ is convex

Proof. We show that H is convex. Let $\boldsymbol{\eta_0}, \boldsymbol{\eta_1} \in H$ and $0 < \alpha < 1$.

We want to show that $(\alpha\boldsymbol{\eta_0} + (1-\alpha)\boldsymbol{\eta_1}) \in H$, i.e.

$$\begin{aligned} &A(\alpha\boldsymbol{\eta_0} + (1-\alpha)\boldsymbol{\eta_1}) < \infty \quad \Rightarrow \\ &\log \int_X h(x) \exp\{\langle (\alpha\boldsymbol{\eta_0} + (1-\alpha)\boldsymbol{\eta_1}), \boldsymbol{T}(x)\rangle\} \, \mathrm{d}x < \infty \end{aligned}$$

which follows if we can show that:

$$\int_X h(x) \exp\{\alpha\langle \boldsymbol{\eta_0}, \boldsymbol{T}(x)\rangle\} \exp\{(1-\alpha)\langle \boldsymbol{\eta_1}, \boldsymbol{T}(x)\rangle\} \, \mathrm{d}x < \infty$$

where we have used the bilinearity of the inner product and also the fact that if if $a < \infty$ then $\log(a) < \infty$. Define the following:

$$\begin{aligned} u(x) &= \exp\{\alpha\langle \boldsymbol{\eta_0}, \boldsymbol{T}(x)\rangle\} \\ v(x) &= \exp\{(1-\alpha)\langle \boldsymbol{\eta_1}, \boldsymbol{T}(x)\rangle\} \end{aligned}$$

Using our definitions, we can see that our above expression simplifies to showing that:

$$\int_X h(x)u(x)v(x) \, \mathrm{d}x < \infty$$

We show the result using Holder's inequality. Note that

$$
\begin{aligned}
&\int_X h(x)u(x)v(x)\,\mathrm{d}x \\
&\quad = \int u(x)h(x)^{\alpha}\cdot v(x)h(x)^{1-\alpha}\,\mathrm{d}x \\
&\quad \leq \left(\int u(x)^{1/\alpha}h(x)\,\mathrm{d}x\right)^{\alpha}\left(\int v(x)^{1/(1-\alpha)}h(x)\,\mathrm{d}x\right)^{1-\alpha} \\
&\qquad \textit{(by Holder)} \\
&\quad = \left(\int \exp\{\langle \boldsymbol{\eta_0}, \boldsymbol{T}(x)\rangle\}h(x)\,\mathrm{d}x\right)^{\alpha}\left(\int \exp\{\langle \boldsymbol{\eta_1}, \boldsymbol{T}(x)\rangle\}h(x)\,\mathrm{d}x\right)^{1-\alpha} \\
&\quad = A(\boldsymbol{\eta_0})^{\alpha}\cdot A(\boldsymbol{\eta_1})^{1-\alpha}
\end{aligned}
$$

And since both $\boldsymbol{\eta_0}$ and $\boldsymbol{\eta_1}$ are in H, then both $A(\boldsymbol{\eta_0})$, $A(\boldsymbol{\eta_1}) < \infty$ and thus $A(\boldsymbol{\eta_0})^{\alpha}\cdot A(\boldsymbol{\eta_1})^{1-\alpha} < \infty$.

□

3 Principles of data reduction

A key first step in inference is condensing down the data one has into a compact form which one can then easily manipulate to perform inference. In this chapter we introduce some important principles of data reduction.

3.1 Sufficiency

We begin with some basic definitions.

> **Definition.** (Statistic, sampling distribution)
>
> Denote our sample space by $\mathcal{X} \in \mathbb{R}^p$ and our data by $\boldsymbol{x} \in X, \text{where } X \sim P_{\boldsymbol{\theta}_0} \in P = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$
>
> A **statistic** is any function $\boldsymbol{T} : \mathcal{X} \to \mathcal{T}$. Typically, $\mathcal{T} \subset \mathcal{R}^d$ where $d < p$ (d is called the *dimension* of the statistic).
>
> Further note that $\boldsymbol{T}(\boldsymbol{X}) \in \mathbb{R}^d$ is a random vector. We call the distribution of $\boldsymbol{T}(\boldsymbol{X})$ the **sampling distribution of $\boldsymbol{T}$ under $\boldsymbol{\theta}$**.

A natural question that follows from this definition is: what does $\boldsymbol{T}(\boldsymbol{x})$ tell us about $\boldsymbol{\theta}$?

1. If $\boldsymbol{T}$ is 1-to-1, then it is invertible and $\boldsymbol{T}(\boldsymbol{x})$ and the original data $\boldsymbol{x}$ contain the same level of information about $\boldsymbol{\theta}$.
2. If $\boldsymbol{T}$ is many-to-1, then some information is lost. Knowing that $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$ only tells us that $\boldsymbol{x} \in \{\boldsymbol{u} : \boldsymbol{T}(\boldsymbol{u}) = \boldsymbol{t}\}$. However, $\boldsymbol{T}(\boldsymbol{x})$ may be simpler or easier to work with than the original data $\boldsymbol{x}$.

This intuition about the level of information embodied in $\boldsymbol{T}(\boldsymbol{x})$ is captured in the concept of a sufficient statistic:

> **Definition.** (Sufficient statistic)
>
> A statistic $\boldsymbol{T}(\boldsymbol{X})$ is **sufficient** for $\boldsymbol{\theta}$ if the distribution of $\boldsymbol{X} \,|\, \boldsymbol{T}(\boldsymbol{X})$ does not depend on $\boldsymbol{\theta}$.

In other words, if $\boldsymbol{T}(\boldsymbol{X})$ is a sufficient statistic for a parameter $\boldsymbol{\theta}$, then its value contains all the information necessary to compute any estimate of $\boldsymbol{\theta}$. To see this, consider the following scenario and theorem (one-dimensional case):

> **Theorem 3.1.** (Simulation from $T(x)$)
>
> Let $X \sim P_\theta \in \mathcal{P}$ with unknown θ. Let $T(x)$ be sufficient for θ. Consider two statisticians:
>
> 1. The first observes x and then makes inference about θ.
> 2. The second observes $T(x)$, simulates x' from the conditional distribution $P(x \in A \,|\, T(x))$, and then makes inference based on x'.

Then for each $\theta \in \Theta$, $P_\theta(x \in A) = P_\theta(x' \in A)$ for $A \subset \mathcal{X}$.

Proof.

$$\begin{aligned}
P_\theta(x \in A) &= \sum_{t \in \mathcal{T}} P_\theta(x \in A, T(x) = t) \\
&= \sum_t P_\theta(x \in A \,|\, T(x) = t)\, P_\theta(T(x) = t) \\
&\qquad \text{(by definition of sufficient statistic)} \\
&= \sum_t P(x \in A \,|\, T(x) = t)\, P_\theta(T(x) = t) \\
&\qquad (x' \text{ simulated from } P(x \,|\, T(x))) \\
&= \sum_t P(x' \in A \,|\, T(x) = t)\, P_\theta(T(x) = t) \\
&= \sum_t P_\theta(x' \in A \,|\, T(x) = t)\, P_\theta(T(x) = t) \\
&= P_\theta(x' \in A)
\end{aligned}$$

□

3.2 The factorization theorem

While the definition of sufficiency is somewhat insightful, it does not equip us with any way to actually *find* sufficient statistics. In this section we flesh out the factorization theorem, which solves this problem.

Lemma 3.2. (Lemma for factorization theorem)

Let $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ and $\boldsymbol{T} : \mathcal{X} \to \mathcal{T}$. Also let $w_{\boldsymbol{\theta}}(\boldsymbol{t})$ be the pmf/pdf of $\boldsymbol{T}(\boldsymbol{X})$ when $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$.

Assume $\boldsymbol{X}, \boldsymbol{T}(\boldsymbol{X})$ are continuous for all $\boldsymbol{\theta}$ or discrete for all $\boldsymbol{\theta}$. Then $\boldsymbol{T}(\boldsymbol{X})$ is sufficient for $\boldsymbol{\theta}$ if, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{w_{\boldsymbol{\theta}}(\boldsymbol{t})} \text{ is independent of } \boldsymbol{\theta}$$

Proof. We give the proof only for the one-dimensional discrete case. We consider the conditional distribution of $X \,|\, T(X)$ and show that, if the assumptions of the lemma hold, then it does not depend on θ.

Assume $X, T(X)$ are as above and fix $x \in \mathcal{X}$ and $t \in \mathcal{T}$.

$$
\begin{aligned}
P_\theta(X = x \,|\, T(x) = t) &= \frac{P_\theta(X = x, T(x) = t)}{P_\theta(T(x) = t)} \\
&= \frac{P_\theta(X = x)}{P_\theta(T(x) = t)} \\
&= \frac{f_\theta(x)}{w_\theta(t)}
\end{aligned}
$$

which is independent of θ by assumption.

In going from the first to second line we used the fact that T is a function so it cannot send one x to many t, and thus $P_\theta(X = x, T(x) = t)$ simply equals $P_\theta(X = x)$.

□

Example. (Bernoulli trials)

Let $x = (x_1, \ldots, x_n)$ with $X_i \overset{iid}{\sim}$ Bernoulli(θ).

We consider the statistic $T(\boldsymbol{x}) = \sum_{i=1}^n x_i$:

$$
\begin{aligned}
w_\theta(t) &= \binom{n}{t} \theta^t \, (1-\theta)^{n-t} \\
&= \binom{n}{\sum x_i} \theta^{\sum x_i} \, (1-\theta)^{n - \sum x_i}
\end{aligned}
$$

$$
\begin{aligned}
f_\theta(\boldsymbol{x}) &= \prod_{i=1}^n \theta^{x_i} \, (1-\theta)^{1-x_i} \\
&= \theta^{\sum x_i} \, (1-\theta)^{n - \sum x_i}
\end{aligned}
$$

Thus we have that

$$
\frac{f_\theta(\boldsymbol{x})}{w_\theta(t)} = \binom{n}{\sum x_i}
$$

Since this expression does not involve θ, then T is sufficient for θ.

Example. (Order statistics)

Obviously the order statistics are sufficient for an iid sample. Duh.

This is another way of looking at how $\boldsymbol{T}$ captures all the interaction between $\boldsymbol{\theta}$ and $\boldsymbol{x}$. In this formulation we see that $w_{\boldsymbol{\theta}}$, the distribution of $\boldsymbol{T}$, *divides out* all the terms involving $\boldsymbol{\theta}$ in the original density $f_{\boldsymbol{\theta}}$.

This strongly foreshadows the full result, which is that the sufficient statistic can be found by *factoring* the density into a part where the data interact with the parameter, and a part where the data do not interact with the parameter.

Theorem 3.3. (Factorization theorem)

Let $\mathcal{P} = \{f_{\boldsymbol{\theta}}(\boldsymbol{X}),\, \boldsymbol{\theta} \in \Theta\}$ and $\boldsymbol{T} : \mathcal{X} \to \mathcal{T} \subset \mathbb{R}^d$. $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions $g_{\boldsymbol{\theta}} : \mathcal{T} \subset \mathbb{R}^d \to \mathbb{R}$ and $h_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$ such that

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{x}))h(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathcal{X},\, \boldsymbol{\theta} \in \Theta$$

Proof. For simplicity we prove the one-dimensional discrete case only.

($\Rightarrow$)

Suppose T is sufficient for θ.

Since T is sufficient for θ, then $\mathbb{P}_\theta(X = x \,|\, T(x) = t)$ does not depend on θ. So we we actually have:

$$\mathbb{P}_\theta(X = x \,|\, T(x) = t) = \mathbb{P}(X = x \,|\, T(x) = t)$$

Set $g_\theta(T(x)) = \mathbb{P}_\theta(T(x) = t)$ and $h(x) = \mathbb{P}(X = x \,|\, T(x) = t)$.

($\Leftarrow$)

Suppose the exist functions $g_\theta : \mathbb{R}^d \to \mathbb{R}$ and $h_\theta : \mathcal{X} \to \mathbb{R}$ such that $f_\theta(x) = g_\theta(T(x))h(x),\ \forall x \in \mathcal{X},\, \theta \in \Theta$.

Let the pmf of $T(X)$ when $X \sim P_\theta$ be denoted by $w_\theta(t)$. Note

$$\begin{aligned} w_\theta(t) = w_\theta(T(x)) &= \sum_{y:\, T(y)=T(x)} f_\theta(y) \\ &= \sum_{y:\, T(y)=T(x)} g_\theta(T(y))h(y) \\ &= g_\theta(T(x)) \sum_{y:\, T(y)=T(x)} h(y) \end{aligned}$$

Where moving from the second-to-last line to the last line we have used the fact that $g_\theta(T(y))$ is constant over $\{y : T(y) = t\}$.

Thus we have

$$\begin{aligned} \frac{f_\theta(x)}{w_\theta(T(x))} &= \frac{g_\theta(T(x))\, h(x)}{g_\theta(T(x))\, \sum_{y:\, T(y)=T(x)} h(y)} \\ &= \frac{h(x)}{\sum_{y:\, T(y)=T(x)} h(y)} \end{aligned}$$

Since this expression does not involve θ, then T is sufficient for θ.

□

Corollary 3.4. T is sufficient for θ if, $\forall \theta_1, \theta_2 \in \Theta$,

$$\frac{f_{\theta_1}(x)}{f_{\theta_2}(x)} \text{ is a function of } T(x) \text{ only}$$

We give some simple examples here and then elaborate on a previous note about the sufficient statistic in the case of the exponential family.

Example. (Uniform sufficient statistic)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$ with $\theta > 0$, and $X = (X_1, \ldots, X_n)$.

$$\begin{aligned} f_\theta(x) &= \frac{1}{\theta} \prod_{i=1}^{n} \mathbb{1}_{(0<x_i<\theta)} \\ &= \frac{1}{\theta} \mathbb{1}_{(x_{(n)}<\theta)} \cdot \prod_{i=1}^{n} \mathbb{1}_{(x_i>0)} \end{aligned}$$

Where it is clear to see that $g_\theta(T(x)) = \frac{1}{\theta}\mathbb{1}_{(x_{(n)}<\theta)}$. Thus $x_{(n)}$ is sufficient for θ.

Remark. (Exponential family sufficient statistic)

If $X_1, \ldots, X_n \overset{iid}{\sim} f_{\boldsymbol{\eta}}$ have an exponential family distribution generated by $h(\cdot)$ and $\boldsymbol{T}(x_i) = (T_1(x_i), \ldots, T_R(x_i))$, then

$$\tilde{\boldsymbol{T}}(\boldsymbol{x}) = \left[\sum_{i=1}^{n} T_1(x_i), \ldots, \sum_{i=1}^{n} T_R(x_i) \right]$$

is sufficient for $\boldsymbol{\eta}$.

Example. (Normal with unknown mean and unknown variance)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma > 0$. Then

$$\boldsymbol{T}(\boldsymbol{x}) = \left(\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} x_i^2 \right)$$

is sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

3.3 Minimal sufficiency

It is clear that a sufficient statistic for a parameter is not unique in general. For one, the un-reduced data are always sufficient. A natural question then is: is there such a thing as a "best" sufficient statistic?

Example. Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\sigma \sim N(0, \sigma^2),\ \sigma > 0$.

From the last remark we know that $T(x) = x_1^2 + \ldots + x_n^2$ is sufficient for σ. But let us consider a few other statistics, sorted in order from most information to least information:

$$\begin{aligned} &T_1(x) = x_1, \ldots, x_n \text{ } \textit{(i.e. the original data)} \\ &T_2(x) = x_1^2, \ldots, x_n^2 \\ &T_3(x) = (x_1^2 + \ldots + x_5^2,\ x_6^2 + \ldots + x_n^2) \\ &T_4(x) = x_1^2 \\ &T_5(x) = 1 \end{aligned}$$

The key observation here is that, since $T(x)$ is a function of either $T_1(x), T_2(x)$, or $T_3(x)$ and $T(x)$ is sufficient, then $T_1(x), T_2(x)$, and $T_3(x)$ must be sufficient also. This motivates our next definition.

Definition. (Minimal sufficiency)

Let $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. $\boldsymbol{T}$ is **minimal sufficient** for $\boldsymbol{\theta}$ if

1. $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$
2. If $\boldsymbol{S}$ is some other sufficient statistic for $\boldsymbol{\theta}$, then $\boldsymbol{T}$ can be written as a function of $\boldsymbol{S}$.

Remark. Some other convenient formulations of the second condition above include:

1. $\exists\, \boldsymbol{g}(\cdot)$ such that $\boldsymbol{T} = \boldsymbol{g}(\boldsymbol{S})$
2. $\boldsymbol{S}(\boldsymbol{x}) = \boldsymbol{S}(\boldsymbol{y}) \Rightarrow \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$
3. $\boldsymbol{T}(\boldsymbol{x}) \neq \boldsymbol{T}(\boldsymbol{y}) \Rightarrow \boldsymbol{S}(\boldsymbol{x}) \neq \boldsymbol{S}(\boldsymbol{y})$

Remark. Note that minimal sufficient statistics are NOT unique, since any 1-1 function of a minimal sufficient statistic is again minimal sufficient.

This definition has a problem similar to one with sufficiency–although it provides insight into the nature of minimal sufficiency, it does not give a way to actually *find* a minimal sufficienct statistic.

Theorem 3.5. Let $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Let $\boldsymbol{T} : \mathcal{X} \to \mathcal{T}$ be some statistic.

Then $\boldsymbol{T}$ is minimal sufficient for $\boldsymbol{\theta}$ if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y}) \iff \exists\, c \text{ such that } \forall \boldsymbol{\theta} \in \Theta, \frac{f_{\boldsymbol{\theta}}(\boldsymbol{x})}{f_{\boldsymbol{\theta}}(\boldsymbol{y})} = c$$

Proof. CB.

□

We conclude this chapter with an application of the above theorem to our favorite family of random variables:

Example. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with both parameters unknown.

We have $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$, an exponential family. It follows that $(T_1^0, T_2^0) = \left(\sum x_i, \sum x_i^2\right)$ are sufficient for (μ, σ^2).

Consider the statistics

$$T_1(x) = \frac{1}{n}\sum x_i, \quad T_2(x) = s_x^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2$$

Since (T_1, T_2) is a 1-1 function of sufficient statistic (T_1^0, T_2^0), then (T_1, T_2) is also sufficient. And note that

$$f(x \,|\, \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-[n(\overline{x} - \mu)^2 + (n-1)s_x^2]}{2\sigma^2}\right\}$$

so that $f(x \mid \mu, \sigma^2) = f(y \mid \mu, \sigma^2)$ if and only if $(T_1(x), T_2(x)) = (T_1(y), T_2(y))$. Thus (T_1, T_2) is minimal sufficient for (μ, σ^2).

4 Point estimation

To recap, the general setup in point estimation is that we have a family of distributions $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ and some data $\boldsymbol{x} \in \mathcal{X}, X \sim f_{\boldsymbol{\theta}} \in \mathcal{P}$ with unknown θ.

Our goal is to use $\boldsymbol{x}$ to estimate $\boldsymbol{\theta}$.

4.1 Finding estimators I: Method of moments

Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{\boldsymbol{\theta}} \in \mathcal{P}$. We cover three main methods for finding point estimators of $\boldsymbol{\theta}$. The first and probably simplest method derives from the fact that $\boldsymbol{\theta}$ can be often be expressed in terms of the moments of X_i.

Definition. (j^{th} moment of X)

The $\mathbf{j^{th}}$ **moment of X** is written $M_j(\boldsymbol{\theta})$ and is defined as

$$M_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} X^j$$

The method of moments is so natural that just giving a few examples of moment expressions for some common random variables covers the gist of the technique:

Example. (Exponential)

$\mathcal{P} = \{\mathrm{Exp}(\theta) : \theta > 0\}$

$$M_1(\theta) = EX = \frac{1}{\theta}$$

so we have

$$\theta = \frac{1}{M_1(\theta)}$$

Example. (Normal)

$\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$

$\boldsymbol{\theta} = (\mu, \sigma^2)$.

$$\begin{aligned} M_1(\boldsymbol{\theta}) &= \mu \\ M_2(\boldsymbol{\theta}) &= \mathrm{Var}_{\boldsymbol{\theta}}(X) + (E_{\boldsymbol{\theta}} X)^2 \\ &= \sigma^2 + \mu^2 \end{aligned}$$

so we have

$$\begin{aligned} \mu &= M_1 \boldsymbol{\theta} \\ \sigma^2 &= M_2(\boldsymbol{\theta}) - M_1(\boldsymbol{\theta})^2 \end{aligned}$$

Example. (Gamma)

$\mathcal{P} = \{\mathrm{Gamma}(\alpha, \beta) : \alpha, \beta > 0\}$,

$\boldsymbol{\theta} = (\alpha, \beta)$.

$$M_1(\boldsymbol{\theta}) = \frac{\alpha}{\beta}$$
$$M_2(\boldsymbol{\theta}) = \frac{\alpha(1+\alpha)}{\beta}$$

so we have

$$\alpha = \frac{M_1(\boldsymbol{\theta}) - M_2(\boldsymbol{\theta})}{M_1(\boldsymbol{\theta})}$$
$$\beta = \frac{M_1(\boldsymbol{\theta}) - M_2(\boldsymbol{\theta})}{M_1(\boldsymbol{\theta})^2}$$

Since we now have expressions for the parameters in terms of population moments, the natural next step is to substitute sample moments for the population moments to obtain an estimator of the parameters.

In sum, the method of moments can be laid out as a three-step process. Assume that we are given $X_1, \ldots, X_n \overset{iid}{\sim} P_{\boldsymbol{\theta}} \in \mathcal{P},\ \Theta \in \mathbb{R}^q$.

1. **Express moments**

   For some $k \geq q$, express moments $M_1(\boldsymbol{\theta}), \ldots, M_k(\boldsymbol{\theta})$ as functions of $\theta_1, \ldots, \theta_q$:

   $$M_1(\boldsymbol{\theta}) = g_1(\theta_1, \ldots, \theta_q)$$
   $$\vdots$$
   $$M_k(\boldsymbol{\theta}) = g_k(\theta_1, \ldots, \theta_q)$$

2. **Invert**

   Solve the system of k equations above to get expressions for the q components of $\boldsymbol{\theta}$:

   $$\theta_1 = h_1(M_1(\boldsymbol{\theta}), \ldots, M_k(\boldsymbol{\theta}))$$
   $$\vdots$$
   $$\theta_q = h_2(M_1(\boldsymbol{\theta}), \ldots, M_k(\boldsymbol{\theta}))$$

3. **Substitute sample moment**

   Replace the population moments $M_j(\boldsymbol{\theta})$ with the corresponding sample moment:

   $$\hat{M}_j(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Since population parameters are generally involved in the expressions for more than only one particular moment, then different method of moments estimators exist depending on which moments are used. Thus, in general, one cannot speak of **the** method of moments estimator.

Finally, it stands to bear in mind that while the method of moments is simple, it can sometimes lead to bad estimators:

Example. (Discrete Uniform)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}\{1, 2, \ldots, \theta\}$.

$$\begin{aligned} M_1(\theta) = E_\theta(X) &= \frac{1}{\theta} \cdot \frac{\theta(\theta+1)}{2} \\ &= \frac{\theta+1}{2} \end{aligned}$$

Thus we have $\theta = 2M_1(\theta) - 1$ and $\hat{\theta} = 2\overline{x} - 1$.

Note that $\hat{\theta} < \max\{x_i\}$. But we know that θ can only be $\geq \max\{x_i\}$, so the method of moments estimator will always be biased downwards.

4.2 Finding estimators II: Maximum likelihood

The next method, which is ubiquitous in statistics nowadays, is the method of maximum likelihood. First we define the likelihood function:

Definition. (Likelihood)

Let $\mathcal{P} = \{f(x \,|\, \theta), \theta \in \Theta\},\ x \in \mathcal{X}$.

The **likelihood of x** is a function $L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) : \Theta \to \mathbb{R}$ defined as:

$$L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$$

The key idea is that while the *joint density* f is a function of $\boldsymbol{x}$, the likelihood is a function of $\boldsymbol{\theta}$ which regards $\boldsymbol{x}$ as given.

In layman's terms, L gives the *evidence in favor of* $\boldsymbol{\theta} \in \Theta$ provided by the observed data $\boldsymbol{x}$. This suggests another natural way to find an estimator of $\boldsymbol{\theta}$:

Definition. (Maximum likelihood estimator/MLE)

Let $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim f(x_i \,|\, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$.

In the case of i.i.d. random variables we have

$$L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \,|\, \boldsymbol{\theta})$$

And we define the **maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$** as

$$\hat{\boldsymbol{\theta}}_{MLE}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

And for convenience, we also define a related quantity which will help us simplify the maximization problem later:

Definition. (Log-Likelihood)

The log-likelihood function is the log of the likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \boldsymbol{x} &= \log L(\boldsymbol{\theta} \mid \boldsymbol{X}) \\ &= \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta})\end{aligned}$$

Remark. Since $\log x$ is a strictly increasing function, then

$$\arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid \boldsymbol{x})$$

In other words, we can log-transform the likelihood function when finding the MLE of θ with no loss of generality.

We now turn our attention to how to go about finding the maximum. More precisely, we are concerned with how to maximize L, ℓ over $\boldsymbol{\theta} \in \Theta$. In some cases this is straightforward: if Θ is finite, then we can do an exhaustive search; if Θ is countable, then we can still do a direct search or use a continuous approximation.

For the case of continuous Θ (i.e. Θ contains open subsets of $\mathbb{R}^q$), then we have two cases:

1. **L is twice differentiable in the interior of Θ**

   Then use the usual calculus technique:

   (a) Find the vector of first derivatives, set it equal to zero, and solve for $\boldsymbol{\theta}^*$.

   (b) Check that the Hessian matrix evaluated at $\boldsymbol{\theta}^*$ is negative definite.

   Finally, we check a few potential problem points:

   (a) Is $\boldsymbol{\theta}^*$ local or global?

   (b) Is $\boldsymbol{\theta}^*$ unique?

   (c) Are there better solutions on the boundary of Θ?

   (d) Are there better solutions when $\theta_i \to \infty, -\infty$?

2. **Otherwise**

   Find an upper bound for L, ℓ and then attempt to find a point where it is achieved or where it is closest to being achieved.

We now cover a few illustrative examples:

Example. (Normal with unknown mean)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1),\ \theta \in \mathbb{R}$.

$$\begin{aligned} L(\theta, \boldsymbol{x}) &= \prod_{i=1}^{n} f(x_i \,|\, \theta) \\ \ell(\theta, \boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum (x_i - \theta)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum (x_i - \overline{x} + \overline{x} - \theta)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum (x_i - \overline{x})^2 + n(\overline{x} - \theta)^2 \end{aligned}$$

And the expression is minimized when $\theta = \overline{x}$, so $\hat{\theta}_{MLE} = \overline{x}$.

Example. (Bernoulli trials)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta),\ \theta \in [0, 1]$.

$$\begin{aligned} \ell(\theta \,|\, \boldsymbol{x}) &= \log \prod_{i=1}^{n} \left[\theta^{x_i}(1-\theta)^{1-x_i}\right] \\ &= \sum x_i \log\theta + \sum (1 - x_i)\log(1-\theta) \\ &= n\left[\overline{x}\log\theta + (1-\overline{x})\log(1-\theta)\right] \end{aligned}$$

Thus for $\overline{x} \in [0, 1]$, $\hat{\theta}_{MLE} = \overline{x}$ since $\Theta \in [0, 1]$.

Remark. Suppose that the parameter space was not $[0, 1]$ but rather $\Theta = [\frac{1}{2}, 1]$ instead. Then

$$\hat{\theta}_{MLE} = \begin{cases} \frac{1}{2} & \overline{x} < \frac{1}{2} \\ \overline{x} & \overline{x} \geq \frac{1}{2} \end{cases}$$

This just serves as a reminder that the MLE is defined as the arg max over the parameter space in question.

When the parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is multidimensional, we have to maximize a function of several variables. Solving the system of equations of first derivatives gives a **necessary** condition for an extremum in the interior and is not too daunting. However, to verify that it is a maximum we need to check that the matrix of second derivatives is positive-definite.

To paraphrase from Casella and Berger, this is generally a pain in the ass. To get around this, sometimes we can maximize the function **successively**.

Example. (Normal with two unknown parameters)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2),\ \boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2}$$

Taking first derivatives, we find critical points at:

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum x_i}{n}$$

We could check the matrix of second derivatives here, or we could simply note that for any σ^2

$$\ell(\mu, \sigma^2 \,|\, \boldsymbol{x}) \geq \ell(\bar{x}, \sigma^2 \,|\, \boldsymbol{x})$$

Therefore to verify that these are the MLE's, we need only tackle the one dimensional problem of showing that $\ell(\bar{x}, \sigma^2 \,|\, \boldsymbol{x})$ is maximized at $\hat{\sigma}^2$ as defined above.

Example. (Uniform)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}[0, \theta]$.

$$L(\theta \,|\, \boldsymbol{x}) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{(0 \leq x_i \leq \theta)}(x_i)$$

Since this expression is 0 if our $\arg\max_\theta$ is strictly less than *any* of the x_i's, then $\hat{\theta}_{MLE} = \max(x_i)$.

4.3 Finding estimators III: The Bayesian approach

The essence of the third method of finding estimators is encapsulated in the belief that the parameter $\boldsymbol{\theta}$ is itself a random variable whereas, before, we have treated $\boldsymbol{\theta}$ as fixed.

Again, we have our family of distributions $\mathcal{P} = \{f(x \,|\, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ and data $\boldsymbol{x} \in \mathcal{X}$. The major components of the Bayesian framework are:

1. the **prior distribution**: $\pi(\boldsymbol{\theta})$ (not necessarily a well-behaved probability distribution)
2. the **sampling distribution**: $f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ (i.e. the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{\theta}$)
3. the **joint distribution**: $f(\boldsymbol{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) f(\boldsymbol{x} \,|\, \boldsymbol{\theta})$
4. the **marginal distribution of $\boldsymbol{x}$ under π**: $m(\boldsymbol{x}) = \int f(\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$

The basic procedure is to take the data $\boldsymbol{x}$, apply a prior belief $\pi(\boldsymbol{\theta})$, and then make inference based on the resulting **posterior distribution**:

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{m(\boldsymbol{x})}$$

Without going into their motivations (see section 4.5 on Bayes' risk), we present a few natural Bayes' estimators:

1. **Posterior mean**:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \int \boldsymbol{\theta} \cdot \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta}$$

2. **Posterior median**:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = u \text{ s.t. } \int \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta} = \frac{1}{2}$$

3. **Posterior mode**:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x}) = \arg\max_{\theta \in \Theta} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

Finally, we conclude the section with the important concept of conjugate priors, which can help to simplify the calculational burden in many Bayesian inference problems:

Definition. (Conjugate priors)

Let Π be a family of priors $\pi(\boldsymbol{\theta})$ on Θ.

Π is **conjugate** for $\mathcal{P}$ if, for all priors $\pi \in \Pi$ and sampling distributions $f(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \in \mathcal{P}$, the associated posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \in \Pi$ also.

Remark. The Beta(α, β) family is conjugate for the problem of estimating the probability parameter p from binomial data.

4.4 Evaluating estimators I: Bias vs. variance

Since the estimators are generally random variables themselves, we can evaluate their performance and characteristics by looking at their variance. Roughly speaking, an estimator with smaller variance is more precise. Also related to this in a very specific way is the bias of an estimator:

Definition. (Bias)

Let $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset \mathbb{R}$

1. The **bias** of an estimator $\hat{\boldsymbol{\theta}}$ is defined as

$$\mathrm{Bias}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}} \left[\hat{\boldsymbol{\theta}}(\boldsymbol{x}) \right] - \boldsymbol{\theta}$$

2. An estimator $\hat{\boldsymbol{\theta}}$ is **unbiased** for $\boldsymbol{\theta}$ if

$$E_{\boldsymbol{\theta}} \left[\hat{\boldsymbol{\theta}}(\boldsymbol{x}) \right] = \boldsymbol{\theta}, \quad \forall \boldsymbol{\theta} \in \Theta$$

Example. (Unbiasedness of $\overline{x}$)

Let $X_i, \ldots, X_n$ be i.i.d. with $EX_i = \mu$.

Then $E(\overline{x}) = \mu$ and thus $\overline{x}$ is unbiased for μ.

Example. (Binomial)

Let $X \sim$ Binomial(n, θ), $0 \leq \theta \leq 1$.

Then $E_\theta(\frac{x}{n}) = \theta$ and thus $\hat{\theta} = \frac{x}{n}$ is unbiased for θ.

Under the scenario of squared error loss, the bias of an estimator and the variance of an estimator satisfy a certain relation:

Theorem 4.1. (Bias-Variance decomposition)

Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be an estimator for $\boldsymbol{\theta}$. Let the loss function be the squared error loss:

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2$$

Then the following relation holds:

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \Big(\text{Bias}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\Big)^2 + \text{Var}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$$

Proof.

$$\begin{aligned}
R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &= E_{\boldsymbol{\theta}}\left[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2\right] \\
&= E_{\boldsymbol{\theta}}\left[(\boldsymbol{\theta} - E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) + E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}})^2\right] \\
&= \Big[\boldsymbol{\theta} - E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\Big]^2 + E_{\boldsymbol{\theta}}\left[(E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}})^2\right] \\
&\qquad + 2E_{\boldsymbol{\theta}}\left[(\boldsymbol{\theta} - E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}))(E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}})\right] \\
&= \Big[\boldsymbol{\theta} - E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\Big]^2 + E_{\boldsymbol{\theta}}\left[(E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}})^2\right]
\end{aligned}$$

□

Remark. If $\hat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$, then the risk of the estimator is equal to its variance.

An important takeaway from this section is that an unbiased estimator is nice, but it might not be the "best" in the sense of having the lowest risk. To illustrate this, we look at the familiar problem of estimating σ^2 from Normal data:

Example. (Estimating Normal σ^2 with squared error loss)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ known. Our goal is to estimate σ^2 with squared error loss.

We are already familiar with two estimators of σ^2:

$$\sigma^2_{MLE} = \frac{1}{n}\sum(x_i - \overline{x})^2 \quad \text{(biased)}$$
$$s^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2 \quad \text{(unbiased)}$$

The first was obtained by the method of maximum likelihood and the second by a procedure designed to produce an unbiased estimator. We now consider a third method designed to minimize the risk.

For a decision rule $d : \mathbb{R}^n \to (0, \infty)$, the risk is given by

$$R(\sigma^2, d(\boldsymbol{x})) = E_{\sigma^2}\left[(\sigma^2 - d(\boldsymbol{x}))^2\right]$$

We consider estimators of the form $d_\alpha(\boldsymbol{x}) = \alpha s^2,\ \alpha > 0$. For fixed α, we have

$$\begin{aligned} R(\sigma^2, d_\alpha(\boldsymbol{x})) &= \mathrm{Var}_{\sigma^2}(\alpha s^2) + \left[\mathrm{Bias}_{\sigma^2}(\alpha s^2)\right]^2 \\ &= \alpha^2 \frac{2\sigma^4}{n-1} + \left[\alpha E(s^2) - \sigma^2\right]^2 \\ &= \left[\frac{2\alpha^2}{n-1} + (\alpha-1)^2\right]\sigma^4 \\ &= \tau(\alpha)\cdot\sigma^4 \end{aligned}$$

Note that $\tau(\alpha_1) < \tau(\alpha_2) \Rightarrow R(\sigma^2, d_{\alpha_1}(\boldsymbol{x})) < R(\sigma^2, d_{\alpha_2}(\boldsymbol{x}))$. In other words, to minimize the risk, we only need to minimize $\tau(\alpha)$.

Simple calculus gives is that $\tau(\alpha)$ is minimized at $\alpha^* = \frac{n-1}{n+1}$. Thus the estimator that minimizes the risk under squared error loss is:

$$\hat{\sigma}^2 = \frac{n-1}{n+1}s^2 = \frac{1}{n+1}\sum(x_i - \overline{x})^2$$

To see how this compares to the risk of one of our other estimators, note that since s^2 is unbiased, its risk is equal to its variance:

$$\begin{aligned} R(\sigma^2, s^2) &= \mathrm{Var}_{\sigma^2}(s^2) \\ &= \frac{2}{n-1}\,\sigma^4 \end{aligned}$$

The risk of the estimator just derived is:

$$\begin{aligned} R(\sigma^2, \hat{\sigma}^2) &= \left[\frac{2\alpha^2}{n-1} + (\alpha-1)^2\right]\sigma^4 \Bigg|_{\alpha=\frac{n-1}{n+1}} \\ &= \left[\frac{2(n-1)^2}{(n-1)(n+1)^2} + \left(\frac{n-1}{n+1} - 1\right)^2\right]\sigma^4 \\ &= \frac{3(n-1)^2 + (n+1)^2 - 2(n-1)(n+1)}{(n+1)^2}\,\sigma^4 \end{aligned}$$

4.5 Evaluating estimators II: Bayes' risk

Finally, to close this chapter we review ways to evaluate estimations in a Bayesian framework. Again, the key distinguishing factor of this section is that we assume that $\boldsymbol{\theta}$ is random. The roadmap for this section is as follows:

1. First, we define the Bayes' risk and the posterior expected loss.
2. Second, we define the Bayes' estimator as that which minimizes the posterior expected loss.
3. Finally, we show that the Bayes' estimator also minimizes Bayes' risk.

Throughout this subsection we make the notation-simplifying assumption that we have one data point x rather than n-dimensional $\boldsymbol{x}$, although the results are easily generalizable.

Definition. (Bayes' and posterior risk)

The **Bayes' risk** of decision rule d under $\pi(\boldsymbol{\theta})$ is

$$\begin{aligned} R_\pi(d) &= \int_\Theta R(\boldsymbol{\theta}, d)\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_\Theta E_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}, d(x)))\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_\Theta \left[\int_{\mathcal{X}} L(\boldsymbol{\theta}, d(x))\, f(x \mid \boldsymbol{\theta})\, \mathrm{d}x \right] \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

Note that the risk is a function whose domain is the space of decision rules, while the loss is a function whose domain is the cartesian product of the parameter space and the action space. We make this explicit by writing the argument of R to be d, while we give the second argument of L as $d(x)$.

The **posterior expected loss** of action $d(x)$ under $\pi(\boldsymbol{\theta})$ is

$$E\left[L(\boldsymbol{\theta}, d(x)) \mid x\right] = \int_\Theta L(\boldsymbol{\theta}, d(x))\, \pi(\boldsymbol{\theta} \mid x)\, \mathrm{d}\boldsymbol{\theta}$$

With a little algebra, we can put the statement of Bayes' risk in terms of the posterior expected loss. This will help us later on to show that the Bayes' estimator minimizes Bayes' risk.

Note that if $L \geq 0$ then:

$$\begin{aligned} R_\pi(\hat{\boldsymbol{\theta}}) &= \int_\Theta \left[\int_{\mathcal{X}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x))\, f(x \mid \boldsymbol{\theta})\, \mathrm{d}x \right] \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_\Theta \left[\int_{\mathcal{X}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x))\, f(x \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}x \right] \mathrm{d}\boldsymbol{\theta} \\ &= \int_\Theta \left[\int_{\mathcal{X}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x))\, \pi(\boldsymbol{\theta} \mid x)\, m(x)\, \mathrm{d}x \right] \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\mathcal{X}} \left[\int_\Theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x))\, \pi(\boldsymbol{\theta} \mid x)\, \mathrm{d}\boldsymbol{\theta} \right] m(x)\, \mathrm{d}x \quad \text{(by Fubini)} \\ &= \int_{\mathcal{X}} E\left[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x)) \mid x \right] m(x)\, \mathrm{d}x \end{aligned}$$

With the setup complete, we define the Bayes' estimator as that which minimizes the posterior expected loss and obtain the main result: finding the Bayes' estimator is equivalent to finding the estimator which minimizes the Bayes' risk.

Definition. (Bayes' estimator)

The **Bayes' estimator** with respect to $\pi(\boldsymbol{\theta})$ is:

$$\hat{\boldsymbol{\theta}}_B(x) = \operatorname*{arg\,min}_{a \in \Theta} E\left[L(\boldsymbol{\theta}, a) \mid x\right]$$

Theorem 4.2. The Bayes' estimator minimizes the Bayes' risk.

Proof. As developed above, the Bayes' risk of an estimator $\hat{\boldsymbol{\theta}}$ can be written

$$R_\pi(\hat{\boldsymbol{\theta}}) = \int_{\mathcal{X}} E\left[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(x)) \,|\, x\right] \, m(x) \, \mathrm{d}x$$

Since the Bayes' estimator is defined as the arg min of the posterior expected loss and the data x is treated as fixed, then the integral on the RHS is minimized when $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_B$.

□

In the second chapter we went over some candidate Bayes' estimators, including the posterior mean and the posterior median. We can use the concept of Bayes' risk to formally justify when to use each estimator. In both of the following examples the setup is:

$$\begin{aligned} \Theta &\subset \mathbb{R} \\ \mathcal{P} &= \{f(x \,|\, \theta) : \theta \in \Theta\} \\ \pi(\theta) &= \text{ prior on } \theta \end{aligned}$$

Example. (Bayes' estimator with squared error loss)

We derive the Bayes' estimator with $L(\theta, a) = (\theta - a)^2$. The posterior expected loss is:

$$E\left[(\theta - d(x))^2 \,|\, x\right] = \int_{\mathbb{R}} (\theta - d(x))^2 \, \pi(\theta \,|\, x) \, \mathrm{d}\theta$$

The integral on the RHS is minimized when $d(x) =$ the expected value of $\theta \,|\, x$, so the Bayes' estimator in the case of squared error loss is the posterior **mean** of θ.

Example. (Bayes' estimator with absolute error loss)

We derive the Bayes' estimator with $L(\theta, a) = |\theta - a|$. The posterior expected loss is:

$$E\left[|\theta - d(x)| \,|\, x\right] = \int_{\mathbb{R}} |\theta - d(x)| \, \pi(\theta \,|\, x) \, \mathrm{d}\theta$$

The integral on the RHS is minimized when $d(x) =$ the median of $\theta \,|\, x$, so the Bayes' estimator in the case of absolute error loss is the posterior **median** of θ.

5 Hypothesis testing

5.1 Basics

The second pillar of statistical inference is hypothesis testing. The main idea is to partition the parameter space into two parts and come up with a method for deciding whether the true parameter lies in one part or the other.

More precisely, let $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a set of distributions on $\mathcal{X}$. We partition Θ into $\Theta_0 \dot\cup \Theta_1$ and then use our data generated by $X \sim P_{\boldsymbol{\theta}} \in \mathcal{P}$ to test complimentary hypotheses:

$$\begin{aligned} H_0 &: \boldsymbol{\theta} \in \Theta_0 \quad \text{(null hypothesis)} \\ H_1 &: \boldsymbol{\theta} \in \Theta_1 \quad \text{(alternative hypothesis)} \end{aligned}$$

There are two equivalent ways of defining how exactly to specify such a test:

Definition. (Hypothesis test)

A **hypothesis test** is defined by either of the following:

1. Specify a partition of the sample space $\mathcal{X} = (\mathcal{X}_0, \mathcal{X}_1)$ and:
   (a) accept H_0 if $\boldsymbol{x} \in \mathcal{X}_0$
   (b) reject H_0 if $\boldsymbol{x} \in \mathcal{X}_1$
2. Specify a pair (T, R) where:
   (a) test statistic $T : \mathcal{X} \to \mathbb{R}$
   (b) rejection region $R \subset \mathbb{R}$

   and reject H_0 if and only if $T(\boldsymbol{x}) \in \mathbb{R}$.

And we can classify hypothesis tests by looking at the partition of Θ:

Definition. (Types of hypotheses)

We say H_0 is **simple** if $|\Theta_0| = 1$ and **composite** if otherwise.

For $\Theta \subset \mathbb{R}$, we call H_1 **one-sided** if $\Theta_1 = [\theta_b, \infty)$ or $\Theta_1 = (-\infty, \theta_a]$, and **two-sided** if $\Theta_1 = \{\theta \in \Theta : \theta \leq \theta_a \text{ or } \theta \geq \theta_b\}$.

Example. (Three possible sets of hypotheses for Normal mean)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1),\ \theta \in \Theta \subset \mathbb{R}$.

1. $\Theta = \{0, 1\}$: $\quad H_0 : \theta = 0,\ H_1 : \theta = 1$
2. $\Theta = [0, \infty)$: $\quad H_0 : \theta = 0,\ H_1 : \theta > 0$
3. $\Theta = \mathbb{R},\ \theta_0$ fixed:
   (a) $H_0 : \theta \leq \theta_0,\ H_1 : \theta > \theta_0$
   (b) $H_0 : \theta = \theta_0,\ H_1 : \theta \neq \theta_0$

5.2 Finding tests: The likelihood ratio

Now that we have covered the structure of hypothesis testing in general, we turn to the mechanism of actually finding the rejection region (or partition of $\mathcal{X}$) associated with a hypothesis test.

The likelihood ratio is by far the most common method, and relies on a simple idea: compare the maximum likelihood value under the assumption of $\boldsymbol{\theta} \in \Theta_0$ and under the assumption of $\boldsymbol{\theta} \in \Theta_1$. We formalize this idea by defining the likelihood ratio test:

Definition. (Likelihood ratio test)

The **likelihood ratio** for a test of $H_0 : \boldsymbol{\theta} \in \Theta_0$ vs. $H_1 : \boldsymbol{\theta} \in \Theta_1$ is:

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})}$$

The form of the **likelihood ratio test** is:

$$(\lambda(\boldsymbol{x}), [0, c]), \ c \in [0, 1]$$

i.e. we reject H_0 if $\lambda(\boldsymbol{x}) \in [0, c]$.

Remark. (Two facts re: λ)

1. Note that we can express the likelihood ratio in terms of MLEs:

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\boldsymbol{\theta}}_0 \,|\, \boldsymbol{x})}{L(\hat{\boldsymbol{\theta}} \,|\, \boldsymbol{x})}$$

   where $\hat{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ and $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.

2. $\lambda(\boldsymbol{x}) = 1 \iff \sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \geq \sup_{\boldsymbol{\theta} \in \Theta_1} L(\boldsymbol{\theta} \,|\, \boldsymbol{x})$

We now give some common examples of likelihood ratio tests:

Example. (Normal mean, simple H_0)

$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1), \ \theta \in \mathbb{R}$

$H_0 : \theta = \theta_0, \ H_1 : \theta \neq \theta_0$

Since $\Theta_0 = \{\theta_0\}$, then

$$\sup_{\theta \in \Theta_0} L(\theta \,|\, \boldsymbol{x}) = L(\theta_0 \,|\, \boldsymbol{x})$$

And since we know $\overline{x}$ is the MLE for the Normal mean,

$$\sup_{\theta \in \Theta} L(\theta \,|\, \boldsymbol{x}) = L(\overline{x} \,|\, \boldsymbol{x})$$

We form the likelihood ratio and simplify:

$$\begin{aligned}\lambda(\boldsymbol{x}) = \frac{L(\theta_0 \,|\, \boldsymbol{x})}{L(\overline{x} \,|\, \boldsymbol{x})} &= \frac{\prod_{i=1}^{n} \exp\left[\frac{-(x_i-\theta_0)^2}{2}\right]}{\prod_{i=1}^{n} \exp\left[\frac{-(x_i-\overline{x})^2}{2}\right]} \\ &= \exp\left[\frac{1}{2}\left(-\sum(x_i-\theta_0)^2 + \sum(x_i-\overline{x})^2\right)\right] \\ &= \exp\left[-\frac{n}{2}\left(\theta_0-\overline{x}\right)^2\right]\end{aligned}$$

We reject for $\lambda(\boldsymbol{x}) \leq c$, or equivalently:

$$|\overline{x} - \theta_0| \geq \left[\frac{2}{n}\log\left(\frac{1}{c}\right)\right]^{\frac{1}{2}} \triangleq \tau(c)$$

Note that as the rejection threshold c grows, then it becomes **more difficult** to reject H_0:

$$\tau(c) \to \begin{cases} +\infty & c \to 0 \\ 0 & c \to 1 \end{cases}$$

Example. (Normal mean, compound)

Same setup as above with $H_0 : \theta \leq \theta_0,\ H_1 : \theta > \theta_0$.

The denominator of the likelihood ratio is the same, but the numerator becomes

$$\sup_{\theta \in \Theta_0} L(\theta \,|\, \boldsymbol{x}) = \begin{cases} L(\overline{x} \,|\, \boldsymbol{x}) & \overline{x} \leq \theta_0 \\ L(\theta_0 \,|\, \boldsymbol{x}) & \overline{x} > \theta_0 \end{cases}$$

We know that the unrestricted likelihood is maximized at $\theta = \overline{x}$. But since θ cannot exceed θ_0 under H_0, in the case that $\overline{x} > \theta_0$ we simply set $\arg\max_{\theta \in \Theta_0}$ to the closest value it can be to $\overline{x}$: θ_0.

Then the likelihood ratio becomes

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & \overline{x} \leq \theta_0 \\ \exp\left\{-\frac{n}{2}(\overline{x}-\theta_0)^2\right\} & \overline{x} > \theta_0 \end{cases}$$

and we reject H_0 if $\lambda(\boldsymbol{x}) \leq c$ or

$$\overline{x} - \theta_0 \geq \left[\frac{2}{n}\log\left(\frac{1}{c}\right)\right]^{\frac{1}{2}}$$

which differs from the above example since the likelihood ratio cannot exceed 1, and thus we are constrained to the case where $\overline{x} > \theta_0$.

Note that in the first example, the simple null hypothesis leads to a two-sided rejection region while in the second example, the hypothesis leads to a one-sided rejection region.

Finally, we cover a more sophisticated example with a nuisance parameter: we are still interested in the mean, but we do not assume the variance is known.

Example. (Normal mean with unknown variance)

$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2),\ \mu \in \mathbb{R}, \sigma^2 > 0.$

$H_0 : \mu \leq \mu_0,\ H_1 : \mu > \mu_0.$

Since σ^2 is unknown but the hypotheses do not involveit, we call it a **nuisance parameter**. To find the likelihood ratio test in this situation, we will have to find a way to "cancel out" or make the parameter irrelevant.

Notation: $\boldsymbol{\theta} = (\mu, \sigma^2) \in \Theta \subset \mathbb{R} \times (0, \infty)$.
$\Theta_0 = (-\infty, \mu_0) \times (0, \infty)$

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) &= \sup_{\mu \in \mathbb{R},\, \sigma^2 > 0} L(\mu, \sigma^2 \,|\, \boldsymbol{x}) \\ &= L\left(\overline{x}, \hat{\sigma}^2_{MLE} \,|\, \boldsymbol{x}\right)\end{aligned}$$

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta} \,|\, \boldsymbol{x}) &= \sup_{\mu < \mu_0,\, \sigma^2 > 0} L(\mu, \sigma^2 \,|\, \boldsymbol{x}) \\ &= \begin{cases} L\left(\overline{x}, \hat{\sigma}^2_{MLE} \,|\, \boldsymbol{x}\right), & \overline{x} \leq \mu_0 \\ L\left(\mu_0, \hat{\sigma}^2_0 \,|\, \boldsymbol{x}\right), & \overline{x} > \mu_0 \end{cases} \\ \hat{\sigma}^2_0 &= \frac{\sum (x_i - \mu_0)^2}{n}\end{aligned}$$

Thus we have:

$$\lambda(\boldsymbol{x}) = \begin{cases} 1 & \overline{x} \leq \mu_0 \\ \frac{L(\mu_0, \hat{\sigma}^2_0 \,|\, \boldsymbol{x})}{L\left(\mu_0, \hat{\sigma}^2_{MLE} \,|\, \boldsymbol{x}\right)} & \overline{x} > \mu_0 \end{cases}$$

And for $c < 1$, we have (requires justification)

$$\lambda(\boldsymbol{x}) \leq c \iff T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq t_2(c)$$

And finally we close with a short remark regarding the relationship between the likelihood ratio and sufficient statistics:

Theorem 5.1. Let $\boldsymbol{T}$ be sufficient for $\boldsymbol{\theta}$, with sampling distribution $g(\boldsymbol{t} \,|\, \boldsymbol{\theta}$.

Let $\tilde{\lambda}(\boldsymbol{t})$ be the likelihood ratio test statistic derived from the sampling distribution $g(\boldsymbol{t} \,|\, \boldsymbol{\theta}) = \tau(\boldsymbol{\theta} \,|\, \boldsymbol{t})$. Then

$$\tilde{\lambda}(\boldsymbol{T}(\boldsymbol{x})) = \lambda(\boldsymbol{x})$$

Proof. Factorization theorem.

□

5.3 Evaluating tests: The power function

Now that we have covered the two basic methods for deriving hypothesis tests, we move to evaluating the properties of those tests. The basic tool we use to study these properties in the context of hypothesis testing is the power function:

Definition. (Power function)

Let $(T(\boldsymbol{x}), R)$ be a test for $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_1 : \boldsymbol{\theta} \in \Theta_1$. The **power function** of the test is:

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\boldsymbol{x}) \in R) = P_{\boldsymbol{\theta}}(\text{the test rejects } H_0)$$

Note that the power function also gives us the Type I and Type II errors:

1. $\boldsymbol{\theta} \in \Theta_0 \Rightarrow \beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\text{Type I error})$
2. $\boldsymbol{\theta} \in \Theta_1 \Rightarrow \beta(\boldsymbol{\theta}) = 1 - P_{\boldsymbol{\theta}}(T(\boldsymbol{x}) \in R^c)$
   $= 1 - P_{\boldsymbol{\theta}}(\text{the test accepts } H_0)$
   $= 1 - P_{\boldsymbol{\theta}}(\text{Type II error})$

The second situation is called the **power of the test at alternative $\boldsymbol{\theta}$**.

In general, we want $\beta(\boldsymbol{\theta})$ to be small for $\boldsymbol{\theta} \in \Theta_0$, and we want $\beta(\boldsymbol{\theta})$ to be large for $\boldsymbol{\theta} \in \Theta_1$. A natural question, then, is: how does $\beta(\boldsymbol{\theta})$ behave at the boundary of Θ_0 and Θ_1?

What this question suggests is a straightforward way to get the "best" test: maximize the power of the test subject to some upper bound α on the Type I error.

Definition. (Size/level/unbiased)

Let $\alpha \in [0, 1]$ and let $(T(\boldsymbol{x}), R)$ be a test with power function $\beta(\boldsymbol{\theta})$.

1. $(T(\boldsymbol{x}), R)$ is a **level-α test** if $\sup_{\boldsymbol{\theta} \in \Theta_0} \beta(\boldsymbol{\theta}) \leq \alpha$
2. $(T(\boldsymbol{x}), R)$ is a **size-α test** if $\sup_{\boldsymbol{\theta} \in \Theta_0} \beta(\boldsymbol{\theta}) = \alpha$
3. $(T(\boldsymbol{x}), R)$ is **unbiased** if $\sup_{\boldsymbol{\theta} \in \Theta_0} \beta(\boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta} \in \Theta_1} \beta(\boldsymbol{\theta})$

We now illustrate a couple examples:

Example. (Binomial proportion)

Let $X \sim$ Binomial(n, θ), $0 < \theta < 1$.

Want to test: $H_0 : \theta \leq \frac{1}{2}$
$H_1 : \theta > \frac{1}{2}$

Note that smaller values of X favor H_0, so it is natural to consider rejection regions of form $R_k = \{k, k+1, \ldots, n\}$. We wish to find the "best" R_k by maximizing the power of the test subject to the constraint that the test have level α.

We begin by deriving an expression for the power. The power function for the

test (X, R_k) is given by:

$$\begin{aligned}\beta_k(\theta) &= P_\theta(x \in R_k) \\ &= P_\theta(x \geq k) \\ &= \sum_{j=k}^{n} \binom{n}{j} \theta^j (1-\theta)^{n-j} \\ &= \frac{n!}{(k-1)!(n-k)!} \int_0^\theta u^{k-1} (1-u)^{n-k} \, \mathrm{d}u\end{aligned}$$

We want to find the maximum probability of Type I error in order to set the α level. So note:

1. $\beta_k(\theta) \to 0$ as $\theta \to 0$, and $\beta_k(\theta) \to 1$ as $\theta \to 1$.
2. $\beta_k(\theta)$ is increasing in θ.

Therefore the maximum Type I error is given by

$$\begin{aligned}\sup_{\theta \in \Theta_0} P_\theta(x \in R_k) &= \sup_{\theta \in [0, \frac{1}{2}]} \beta_k(\theta) \\ &= \beta_k(1/2) \\ &= \left(\frac{1}{2}\right)^n \sum_{j=k}^{n} \binom{n}{j} \\ &= 1 \text{ - maximum Type II error}\end{aligned}$$

So to make the test have level $\alpha \in (2^{-n}, 1]$, select k such that

$$\beta_k(1/2) \leq \alpha$$

Finally, note that for given θ, the power function decreases in k:

$$k_1 \leq k_2 \Rightarrow \beta_{k_1}(\theta) \geq \beta_{k_2}(\theta), \ \forall \theta \in (0,1)$$

And thus we maximimize the power of the test by selecting $k^* =$ the **smallest** k such that

$$\sup_{\theta \leq \frac{1}{2}} \beta_k(\theta) = \beta_k(1/2) \leq \alpha$$

Monotonicity of β_k guarantees that tests of this form are unbiased.

Example. (Normal mean, known variance)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Want to test: $H_0 : \mu \leq \mu_0$
$H_1 : \mu > \mu_0$

We consider the likelihood ratio test $(T(\boldsymbol{x}), R_c)$:

$$T(\boldsymbol{x}) = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}, \quad R_c = [c, \infty)$$

The power function for this test is given by

$$\begin{aligned}\beta_c(\mu) &= P(T(\boldsymbol{x}) \in R_c) \\ &= P_\mu\left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geq c\right) \\ &= P_\mu\left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \geq c + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \geq c + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(c + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right)\end{aligned}$$

To find the maximum probability of Type I error, note

1. $\beta_c(\mu) \to 1$ as $\mu \to \infty$, and $\beta_c(\mu) \to 0$ as $\mu \to -\infty$.
2. $\beta_c(\mu)$ is continuous and strictly increasing in μ.

Therefore we can find the maximum:

$$\sup_{\mu \leq \mu_0} \beta_c(\mu) = \beta_c(\mu_0) = 1 - \Phi(c)$$

And furthermore we can always invert this CDF, so for every $0 < \alpha < 1$, $\exists\, c = \Phi^{-1}(1-\alpha)$ such that $(T(\boldsymbol{x}), R_c)$ has size α.

The power is maximized as in the previous example. Unbiasedness follows from monotonicity of β_c.

Example. (Normal mean, unknown variance)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 unknown.

Want to test: $H_0 : \mu \leq \mu_0$
$H_1 : \mu > \mu_0$

Note two facts:

Fact. (Test statistic)

The likelihood ratio test for H_0 vs. H_1 is equivalent to the test (T', R_c) where:

$$T'(x) = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \quad \text{and} \quad R_c = [c, \infty)$$

And the associated power function for this test is:

$$\beta_c(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\left(\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \geq c\right), \quad \boldsymbol{\theta} = (\mu, \sigma^2)$$

Also note that

$$T'(x) = \frac{\sqrt{n}(\bar{x}-\mu_0)/\sigma}{\sqrt{s^2/\sigma^2}} \overset{d}{=} \frac{N(\delta, 1)}{\sqrt{\chi^2_{(n-1)}}}, \quad \delta = \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}$$

The numerator and denominator are independent, so $T' \sim$ noncentral $t_{(n-1),\delta}$.

Fact. (Property of noncentral t)

If $\delta_1 < \delta_2$, then $T_{(n-1);\delta_1}$ is stochastically less than $T_{(n-1);\delta_2}$.

Now note that $\delta > 0$ for $\boldsymbol{\theta} \in \Theta_1$ and $\delta \leq 0$ for $\boldsymbol{\theta} \in \Theta_0$ with $\delta = 0$ iff $\mu = \mu_0$.
Thus:

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in \Theta_0} \beta_c(\theta) &= \sup_{\mu < \mu_0, \sigma^2 > 0} \beta_c(\mu, \sigma^2) \\
&= \sup_{\sigma^2 > 0} \left[\sup_{\mu \leq \mu_0} \beta_c(\mu, \sigma^2) \right] \\
&= \sup_{\sigma^2 > 0} \beta_c(\mu_0, \sigma^2) \quad \text{(by the stochastically} < \text{property)} \\
&= \sup_{\sigma^2 > 0} \mathbb{P}_{(\mu_0, \sigma^2)}(T'(x) \geq c) \quad \text{(does not depend on } \sigma^2) \\
&= \mathbb{P}(t_{(n-1)} \geq c)
\end{aligned}$$

So for the test to have size α, pick c such that $\mathbb{P}(t_{(n-1)} \geq c) = \alpha$.

Also, since $\beta_c(\boldsymbol{\theta}) \leq \beta_c(\boldsymbol{\theta}')$ if $\theta \in \Theta_0$ and $\theta' \in \Theta_1$, then (T', R_c) is unbiased.

Remark. Actually we do not have to use a noncentral-t argument in this case, because s does not depend on θ. See the section of p-values.

5.4 UMP tests and Neyman-Pearson

Definition. (Uniformly most powerful test)

Let $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_1 : \boldsymbol{\theta} \in \Theta_1$, and let an observation be given by $X \in \mathcal{X}$.

Then (X, R) is a **uniformly most powerful level α test** for H_0 vs H_1 if:

1. (X, R) has level α
2. For any other level α test (X, R'),

$$\beta(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(X \in R) \geq \mathbb{P}_{\boldsymbol{\theta}}(X \in R') = \beta'(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta_1$$

Theorem 5.2. (Neyman-Pearson)

Consider $\mathcal{P} = \{f_\theta : \theta \in \{0, 1\}\}$ testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.
For some $\tau > 0$, suppose a test has rejection region given by:

$$\{\boldsymbol{x} : f_1(\boldsymbol{x}) > \tau f_0(\boldsymbol{x})\} \subset R \quad \text{and} \quad \{\boldsymbol{x} : f_1(\boldsymbol{x}) < \tau f_0(\boldsymbol{x})\} \subset R^c \tag{1}$$

And suppose such test has size α:

$$\mathbb{P}_0(X \in R) = \alpha \tag{2}$$

1. (Sufficiency) Any test satisfying (1) and (2) is UMP level α.
2. (Necessity) If there exists a test satisfying (1) and (2) with $\tau > 0$, then every UMP level α test satisfies (1) and (2) except perhaps on a set A satisfying:

$$\mathbb{P}_0(X \in A) = \mathbb{P}_1(X \in A) = 0$$

Remark. (Some observations)

1. The value of τ and the set $R \cap \{f_1 = \tau f_0\}$ depend on α.
2. If $\tau = 0$, then $\mathbb{P}_1(X \in R) = \mathbb{P}_1(f_1(x) > 0) = 1$ (i.e. power = 1).
3. R and R' differ only on set where $f_1 = \tau f_0$.

Proof. First note that any test satisfying (2) is automatically level α, since Θ_0 has only one point. Now let (X, R) be a test satisfying (1) and (2) and let (X, R') be any other level α test. Define:

$$\phi(\boldsymbol{x}) = \mathbb{1}_{X \in R} \quad \text{and} \quad \phi'(\boldsymbol{x}) = \mathbb{1}_{X \in R'}$$

The associated power functions are:

$$\beta(\theta) = \mathbb{P}_\theta(X \in R) = \int \phi(\boldsymbol{x}) f_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$
$$\beta'(\theta) = \mathbb{P}_\theta(X \in R') = \int \phi'(\boldsymbol{x}) f_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

Now note that on the set $\{\boldsymbol{x} : f_1(\boldsymbol{x}) > \tau f_0(\boldsymbol{x})\}$, we have $\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x}) \geq 0$ and on the set $\{\boldsymbol{x} : f_1(\boldsymbol{x}) < \tau f_0(\boldsymbol{x})\}$, we have $\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x}) \leq 0$. Therefore:

$$[f_1(\boldsymbol{x}) - \tau f_0(\boldsymbol{x})] \cdot [\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})] \geq 0$$

Integrating this, we obtain:

$$0 \leq \beta(1) - \beta'(1) - \tau[\beta(0) - \beta'(0)] \qquad (\star)$$

1. (Sufficiency):

   Now since (X, R') has level α but (X, R) has **size** α, then

   $$0 \leq \alpha - \beta'(0) = \beta(0) - \beta'(0)$$

   Substituting this into $(\star)$ gives $0 \leq \beta(1) - \beta'(1)$, as desired.

2. (Necessity):

   Keeping the above notation, suppose that the arbitrary level α test (X, R') (which does not necessarily satisfy the conditions of the N-P lemma) is actually **UMP** level α. We show that this forces the test to satisfy the conditions of the N-P lemma.

   Then $\beta(1) = \beta'(1)$ and by $(\star)$,

   $$0 \leq -\tau[\beta(0) - \beta'(0)] = \tau[\beta'(0) - \alpha]$$

   Since (X, R') is level α, then $\beta'(0) \leq 0$. But if $\beta'(0) < 0$, then this term would be < 0 and then the entire RHS of $(\star)$ would be < 0, which is impossible. Therefore $\beta'(0) = \alpha$ so (X, R') has size α.

   Simultaneously, since $\beta'(0) = \beta(0)$, then $(\star)$ holds with equality. But this is only possible if ϕ' satisfies (1), except on the set A of measure zero defined in the statement of the theorem.

□

Corollary 5.3. (Neyman-Pearson and sufficiency)

Consider $\mathcal{P} = \{f_\theta : \theta \in \{0,1\}\}$ testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.
Let $T(\boldsymbol{X}) : \mathcal{X} \to \mathcal{T}$ be a sufficient statistic for θ
Let $g_\theta(t)$ be the pdf/pmf of T corresponding to $\theta \in \{0,1\}$.

Suppose (T, S) is a test based on T, $S \subset \mathcal{T}$ satisfying:

1. For some $\tau \geq 0$, $\{t : g_1(t) > \tau g_0(t)\} \subset S$ and $\{t : g_1(t) < \tau g_0(t)\} \subset S^c$
2. $\mathbb{P}_0(T \in S) = \alpha$

Then (T, S) is UMP level α.

Example. (UMP Normal, known variance)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ^2 known. $\bar{X}$ is sufficient for θ.
Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_0 > \theta_1$.

The inequality $g_1(\bar{x}) > \tau g_0(\bar{x})$ is equivalent to:

$$\bar{x} < \frac{(2\sigma^2 \log \tau)/n - \theta_0^2 + \theta_1^2}{2(\theta_1 - \theta_0)}$$

The RHS is monotone increasing in τ, so the test rejecting for $\bar{x} < c$ is the UMP level α test, where $\alpha = \mathbb{P}_{\theta_0}(\bar{X} < c)$.

5.5 P-values

Consider testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ with some test statistic $S : \mathcal{X} \to \mathbb{R}$. How do we quantify the **strength** of the evidence against H_0?

Example. (Intuitive "definition" for p-value)

Suppose large values of $S(X)$ favor H_1. This suggests a test of form $(S, [\tau, \infty))$ where τ is determined by the α-level of the test.

So for $0 < \alpha < 1$, define:

$$\tau(\alpha) = \min \tau \text{ s.t. } \mathbb{P}_\theta(S(X) > \tau) \leq \alpha \quad \forall \theta \in \Theta_0$$

Now set the rejection region as $R(\alpha) = [\tau(\alpha), \infty)$. This ensures that:

1. $(S, R(\alpha))$ is level α.
2. $R(\alpha)$ is as large as possible subject to (1), i.e. it maximizes the power.
3. $\alpha_1 \leq \alpha_2 \Rightarrow \tau(\alpha_1) \geq \tau(\alpha_2)$.

Now define two quantities p_1 and p_2 by:

$$p_1(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S(X) \geq S(x))$$
$$p_2(x) = \min\{\alpha : (S(X), R(\alpha)) \text{ rejects } H_0\}$$

These will later be shown to be p-values. But to gain insight into what p-values are, note that the definitions are actually the same:

$$\begin{aligned} p_2(x) &= \min\{\alpha : S(x) \in R(\alpha)\} \\ &= \min\{\alpha : S(x) \geq \tau(\alpha)\} \\ &= \min\{\alpha : \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S(X) > S(x)) \leq \alpha\} \\ &= \min\{\alpha : p_1(x) \leq \alpha\} \\ &= p_1(x) \end{aligned}$$

Definition. (p-value)

A **p-value** is a test statistic $p : \mathcal{X} \to [0, 1]$ such that $0 \leq p(\boldsymbol{x}) \leq 1$ for every sample point $\boldsymbol{x}$.

A **valid p-value** is a p-value such that for all $\alpha \in [0, 1]$,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(\boldsymbol{X}) \leq \alpha) \leq \alpha$$

The interpretation here is: p quantifies the evidence against H_0 based on observed $\boldsymbol{x}$, i.e. smaller $p(\boldsymbol{x})$ means more evidence. Also note that, by definition, the test $(p(\boldsymbol{X}), [0, \alpha])$ has level α.

In idiot terms, a p-value is a function taking values between 0 and 1 and such that the "maximum" probability of it being less than α **when you plug in some data** x is itself less than α.

Fact. $p(X) \overset{d}{\geq} \text{Uniform}(0, 1)$ when $X \sim \mathbb{P}_\theta,\ \theta \in \Theta_0$.

Proof. Immediate from the definition of p-value.

□

Theorem 5.4. Let $S : \mathcal{X} \to \mathbb{R}$ be a test statistic with larger values favoring H_1. Then

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S(\boldsymbol{X}) \geq S(\boldsymbol{x}))$$

is a valid p-value.

Proof. Let F_θ be the CDF of $-S(\boldsymbol{X})$ under $\mathbb{P}_\theta$, $\theta \in \Theta_0$.

Fix $\alpha \in [0, 1]$. Note that $\forall \theta \in \Theta_0$,

$$\begin{aligned} p(\boldsymbol{x}) &= \mathbb{P}_\theta(S(\boldsymbol{X}) \geq S(\boldsymbol{x})) \\ &= \mathbb{P}_\theta(-S(\boldsymbol{X}) \leq -S(\boldsymbol{x})) \\ &= F_\theta(-S(\boldsymbol{x})) \end{aligned}$$

Therefore the following is also true for every $\theta \in \Theta_0$:

$$\mathbb{P}_\theta(p(\boldsymbol{x}) \leq \alpha) = \mathbb{P}_\theta(F_\theta(-S(\boldsymbol{x})) \leq \alpha) \leq \alpha$$

Where the rightmost inequality follows from the probability integral transform (holds with equality when the CDF is continuous).

Apply the definition of p-value.

□

Example. (Normal with unknown variance, simple null)

Let $X_1, \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 unknown.
Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$.

The likelihood ratio test for this is:

$$T(x) = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}}, \quad \text{reject if } T(x) \geq c$$

Note that $\mu = \mu_0 \Rightarrow (\bar{x} - \mu_0)/(s/\sqrt{n}) \sim t_{(n-1)}$ regardless of σ^2. Therefore:

$$\begin{aligned} p(x) &= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x)) \\ &= \sup_{\sigma^2 > 0} \left\{ \mathbb{P}_{\mu_0, \sigma^2}(T(X) \geq T(x)) \right\} \\ &= \mathbb{P}(|t_{(n-1)}| > T(x)) \quad \text{(by above fact)} \\ &= 2 \cdot \mathbb{P}(t_{(n-1)} \geq T(x)) \end{aligned}$$

Example. (Normal with unknown variance, one-sided null)

Consider testing $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$.

The likelihood ratio test for this is:

$$T(x) = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad \text{reject if } T(x) \geq c$$

The p-value is given by:

$$\begin{aligned} p(x) &= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x)) \\ &= \sup_{\mu \leq \mu_0, \sigma^2 > 0} \mathbb{P}_{\mu, \sigma^2}(T(X) \geq T(x)) \end{aligned}$$

In order to simplify this, note that the test statistic involves μ_0 instead of μ, so it is not t-distributed. But note that for any $\mu \leq \mu_0$ and $\sigma^2 > 0$, we can manipulate to show:

$$\begin{aligned} \mathbb{P}_{\mu, \sigma^2}(T(X) \geq T(x)) &= \mathbb{P}_{\mu, \sigma^2}\left(\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq T(x) \right) \\ &= \mathbb{P}_{\mu, \sigma^2}\left(\frac{\bar{x} - \mu}{s/\sqrt{n}} \geq T(x) + \frac{\mu_0 - \mu}{s/\sqrt{n}} \right) \\ &= \mathbb{P}_{\mu, \sigma^2}\left(t_{(n-1)} \geq T(x) + \frac{\mu_0 - \mu}{s/\sqrt{n}} \right) \end{aligned}$$

We want to take the supremum over the Θ_0 region $\mu \leq \mu_0$, but the RHS expression $(\mu_0 - \mu)/(s/\sqrt{n})$ is actually a random variable because s is random.

However, despite the RHS being random, the probability is still monotone in μ because s does not depend on mu. Regardless, we can observe that

for any $\mu \leq \mu_0$, the RHS expression is non-negative. Therefore regardless of monotonicity, it is minimized at 0 by setting $\mu = \mu_0$. Combined with the fact that σ^2 is not involved in the probability statement at all, we have:

$$\sup_{\mu \leq \mu_0, \sigma^2 > 0} \mathbb{P}_{\mu,\sigma^2}(T(X) \geq T(x)) = \mathbb{P}_{\mu,\sigma^2}\left(t_{(n-1)} \geq T(x)\right)$$

And our p-value is given by $\mathbb{P}(t_{(n-1)} \geq T(x))$.

6 Interval estimation

6.1 Basics

To review some notation, we have:

1. A parametric family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$
2. A parameter space $\Theta \subset \mathbb{R}$
3. A sample space $\mathcal{X}$

If $\boldsymbol{X} \sim \mathbb{P}_\theta$, our goal is to use $\boldsymbol{x} \in \mathcal{X}$ to find an interval likely to contain θ.

Definition. (Interval estimator)

An **interval estimator of** θ is a pair $L, U : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ such that $L(\boldsymbol{X}) \leq U(\boldsymbol{X})\ \forall \boldsymbol{X}$.

Note that one-sided estimators are simply when one of the pair are set to be $\pm\infty$.

Definition. (Characteristics of interval estimators)

1. The **coverage probability of** (L, U) is:
$$\mathbb{P}_\theta\Big(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})]\Big) = \mathbb{P}_\theta\Big(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\Big)$$
2. The **confidence coefficient of** (L, U) is:
$$\inf_\theta \mathbb{P}_\theta\Big(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})\Big)$$
3. A $1-\alpha$ **confidence interval for** θ is an interval estimator with confidence coefficient $= 1-\alpha$.

Example. (Normal with unit variance)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. We want a CI for μ.

Consider the interval estimator based on $\bar{X} \sim N(\mu, 1/n)$:
$$\left[L(\boldsymbol{X}) = \bar{X} - \frac{c}{\sqrt{n}},\ U(\boldsymbol{X}) = \bar{X} + \frac{c}{\sqrt{n}}\right]$$

Then the coverage probability is:
$$\begin{aligned}
\mathbb{P}_\mu(L(\boldsymbol{X}) \leq \mu \leq U(\boldsymbol{X})) &= \mathbb{P}\left(\bar{X} - \frac{c}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{c}{\sqrt{n}}\right) \\
&= \mathbb{P}_\mu\left(-c \leq \frac{\bar{X} - \mu}{1/\sqrt{n}} \leq c\right) \\
&= \mathbb{P}(-c \leq z \leq c)
\end{aligned}$$

Where we have removed the μ subscript because the final probability is independent of μ. Also because of this, the coverage probability equals the confidence coefficient in this example.

Example. (Uniform with zero left endpoint)

In this example we illustrate two different ways to construct an interval estimator for the case of a Uniform endpoint.

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$ with $\theta > 0$.
Define $S = \max\{X_1, \ldots, X_n\}$ and $T = S/\theta$, then:

$$f_S(s) = \frac{n \cdot s^{n-1}}{\theta^n} \quad \text{and} \quad f_T(t) = nt^{n-1}$$

1. **Method 1: multiplicative**

   Consider $[L(\boldsymbol{X}) = aS,\ U(\boldsymbol{X}) = bS]$ with $1 \leq a \leq b$.

   The coverage probability of this interval is:

   $$\begin{aligned}
   \mathbb{P}_\theta(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) &= \mathbb{P}_\theta(aS \leq \theta \leq bS) \\
   &= \mathbb{P}_\theta\left(\frac{1}{b} \leq T \leq \frac{1}{a}\right) \\
   &= \int_{1/b}^{1/a} nt^{n-1}\,\mathrm{d}t \\
   &= \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n
   \end{aligned}$$

   This is independent of θ, so it is also the confidence coefficient.

2. **Method 2: additive**

   Consider $[L(\boldsymbol{X}) = S + c,\ U(\boldsymbol{X}) = S + d]$ with $0 \leq c \leq d$.

   The coverage probability of this interval is:

   $$\begin{aligned}
   \mathbb{P}_\theta(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) &= \mathbb{P}_\theta(S + c \leq \theta \leq S_d) \\
   &= \mathbb{P}_\theta\left(1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right) \\
   &= \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n
   \end{aligned}$$

   The coverage probability $\to 0$ as $\theta \to 0$, so **the confidence coefficient is 0.**

6.2 Finding CIs I: Inverting hypothesis tests

Our goal in this sectionis to obtain a $1 - \alpha$ confidence interval for θ based on $\boldsymbol{X}$, through inverting a hypothesis test. There are three parts to this:

1. Start with a level α test of a **simple** hypothesis $H_0 : \theta = \theta_0$. Call it $(X, R(\theta_0))$:

$$\mathbb{P}_{\theta_0}(\boldsymbol{X} \in R(\theta_0)) \leq \alpha \quad \forall \theta_0 \in \Theta$$

2. Note the acceptance region for this test:

$$\begin{aligned} A(\theta_0) &= \{\boldsymbol{x} \in \mathcal{X} : \text{ the test accepts } H_0 : \theta = \theta_0\} \\ &= R(\theta_0)^c \end{aligned}$$

3. So set the interval estimate to:

$$C(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A(\theta)\}$$

The concept of inverting a simple hypothesis test to obtain an interval estimator can be succinctly expressed as:

$$\theta_0 \in C(\boldsymbol{x}) \iff \boldsymbol{x} \in A(\theta_0)$$

We prove that this is a valid $1 - \alpha$ confidence set, and that one can reverse the process to obtain a simple hypothesis test from a interval estimator.

Theorem 6.1. (Hypothesis test to confidence set)

For **each** $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance regon of a level α test of $H_0 : \theta = \theta_0$. For each $\boldsymbol{x} \in \mathcal{X}$, define a set $C(\boldsymbol{x})$ in the parameter space by:

$$C(\boldsymbol{x}) = \{\theta_0 : \boldsymbol{x} \in A(\theta_0)\}$$

Then *random set* $C(\boldsymbol{X})$ has confidence coefficient $\geq 1 - \alpha$.

Proof. We want to show that $\inf_\theta \mathbb{P}_\theta(\theta \in C(\boldsymbol{X})) \geq 1 - \alpha$.

Fix some $\theta_0 \in \Theta$. Then note:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\theta_0 \in C(\boldsymbol{X})) &= \mathbb{P}_{\theta_0}(\boldsymbol{X} \in A(\theta_0)) \\ &= 1 - \mathbb{P}_{\theta_0}(\boldsymbol{X} \in R(\theta_0)) \\ &\geq 1 - \alpha \end{aligned}$$

This is true for any $\theta_0 \in \Theta$, so therefore $\inf_\theta \mathbb{P}_\theta(\theta \in C(\boldsymbol{X})) \geq 1 - \alpha$. □

Theorem 6.2. (Confidence set to hypothesis test)

If $C(\boldsymbol{X})$ is a $1 - \alpha$ confidence set, then for all $\theta_0 \in \Theta$,

$$R(\theta_0) = \{\boldsymbol{X} : \theta_0 \in C(\boldsymbol{X})\}^c$$

is the rejection region of a level α test of $H_0 : \theta = \theta_0$.

Proof. We want to show that $\forall \theta_0 \in \Theta$, $\mathbb{P}_{\theta_0}(\boldsymbol{X} \in R(\theta_0)) \leq \alpha$.

Fix $\theta_0 \in \Theta$. Then by the definition of a $1 - \alpha$ confidence set,

$$\mathbb{P}_{\theta_0}(\boldsymbol{X} \in R(\theta_0)) = 1 - \mathbb{P}_{\theta_0}(\theta_0 \in C(\boldsymbol{X})) \leq 1 - (1 - \alpha)$$

□

IMPORTANT: note that although we normally talk about inverting a test to obtain a confidence set, we are really using an entire **family** of hypothesis tests (one for each $\theta_0 \in \Theta$) to obtain **one** confidence set.

Now we give two examples. In each example, the resulting confidence sets will both be $1-\alpha$ confidence intervals (i.e. their confidence coefficient will be **equal to** $1-\alpha$) because we will start with **size** α tests.

Example. (Exponential CI)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \exp(\theta)$, $\theta > 0$.
Goal: $(1-\alpha)$ CI for θ

1. **Derive the 2-sided LRT**

   We invert the 2-sided test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

   Define $S(\boldsymbol{x}) = \sum_{i=1}^n x_i$. Then the likelihood ratio is:

   $$\lambda(\boldsymbol{x}) = \frac{L(\boldsymbol{x} \,|\, \theta_0)}{\sup_{\theta>0} L(\boldsymbol{x} \,|\, \theta)} = \frac{L(\boldsymbol{x} \,|\, \theta_0)}{L(\boldsymbol{x} \,|\, \bar{x})} = \left(\frac{S(\boldsymbol{x})}{\theta_0}\right)^n e^{-\frac{S(\boldsymbol{x})}{\theta_0}} \left(\frac{e}{n}\right)^n$$

   Therefore the acceptance region for a size α test is:

   $$A(\theta_0) = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \geq c\} = \left\{\boldsymbol{x} : \left(\frac{S(\boldsymbol{x})}{\theta_0}\right)^n e^{-\frac{S(\boldsymbol{x})}{\theta_0}} \geq \tau\right\}$$

   where $\tau = c(n/e)^n$ is chosen such that $\mathbb{P}_{\theta_0}(\boldsymbol{x} \in A(\theta_0)) = 1 - \alpha$.

2. **Invert the test(s)**

   Applying our theorem, the confidence set is given by:

   $$C(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A(\theta)\} = \left\{\theta : \left(\frac{S(\boldsymbol{x})}{\theta}\right)^n e^{-\frac{S(\boldsymbol{x})}{\theta}} \geq \tau\right\}$$

   To find the explicit form of the set $C(\boldsymbol{x})$ (it may not be an interval), note that for a function of the form $h(u) = u^n e^{-u}$:

   (a) h is unimodal

   (b) $h(0) = 0$ and $h(u) \to 0$ as $u \to \infty$.

   Therefore:

   $$\begin{aligned} C(\boldsymbol{x}) &= \left\{\theta : h\left(\frac{S(\boldsymbol{x})}{\theta}\right) \geq \tau\right\} \\ &= \left\{\theta : a \leq \frac{S(\boldsymbol{x})}{\theta} \leq b\right\} \\ &= \left\{\theta : \frac{S(\boldsymbol{x})}{b} \leq \theta \leq \frac{S(\boldsymbol{x})}{a}\right\} \end{aligned}$$

   Now note that, in the second line, the distribution of $S(\boldsymbol{x})/\theta$ is:

   $$\mathcal{L}\left(\frac{S(\boldsymbol{x})}{\theta_0}\right) = \frac{\text{gamma}(n, \theta_0)}{\theta_0} = \text{gamma}(n, 1)$$

Which does not depend on θ, so the coverage probability for any θ is equal to the confidence coefficient.

Solve for a, b, such that the probability is $1-\alpha$ and we are done.

Example. (Normal CI with σ^2 unknown)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 unknown.
Goal: $(1-\alpha)$ upper CI for μ: $C(\boldsymbol{x}) = (-\infty, U(\boldsymbol{x})]$

1. **Determine the direction of the hypothesis test**

   Since we want an *upper* CI, we must invert a *one-sided* test:

   $$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$

   Note that these hypotheses do not involve σ^2, so we are finding a hypothesis test treating σ^2 as a nuisance parameter).

   To see why we set $H_1 : \mu < \mu_0$, recall $C(\boldsymbol{x}) = \{\mu : \boldsymbol{x} \in A(\mu)\}$. Now note that if the alternative is of the form $\mu < \mu_0$, then for given $\boldsymbol{x}$,

   $$\boldsymbol{x} \in A(\mu_0') \;\Rightarrow\; \boldsymbol{x} \in A(\mu_0'') \quad \forall \mu_0' > \mu_0''$$

   In words, if μ_0' is in the confidence set, then also all μ_0'' below μ_0' are in the confidence set, so the set has form $(-\infty, U(\boldsymbol{x})]$.

2. **Derive the test**

   The size α LRT (with nuisance parameter σ^2) is:

   $$T(\boldsymbol{x}) = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad R = (-\infty, -t_{(n-1),\alpha}]$$

   The acceptance region is given by:

   $$A(\mu_0) = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq -t_{(n-1),\alpha}\} = \left\{\boldsymbol{x} : \bar{x} \geq \mu_0 - \frac{s}{\sqrt{n}} \cdot t_{(n-1),\alpha}\right\}$$

3. **Invert the test(s)**

   Applying our theorem, the confidence set is given by:

   $$C(\boldsymbol{x}) = \{\mu : \boldsymbol{x} \in A(\mu)\} = \left\{\mu : \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{(n-1),\alpha} \geq \mu\right\}$$

   Thus the upper interval $(-\infty, U(\boldsymbol{x})]$ is given by:

   $$U(\boldsymbol{x}) = \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{(n-1),\alpha}$$

Note: In this example, the unknown σ^2 nuisance parameter was taken care of by using the appropriate hypothesis test (which takesthe nuisance parameter into consideration).

6.3 Finding CIs II: Pivoting

Definition. (Pivot)

A **pivot** is a function $Q : \mathcal{X} \times \Theta \to \mathbb{R}$, for example $Q(\boldsymbol{X}, \theta)$, such that its distribution under $\mathbb{P}_\theta$ does not depend on θ.

That is, $\mathbb{P}_\theta(Q(\boldsymbol{X}, \theta) \in A)$ does not depend on θ.

Example. (Uniform pivots)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$. The following are pivots:

1. $Q_0(\boldsymbol{X}, \theta) = 0$
2. $Q_1(\boldsymbol{X}, \theta) = X_{(n)}/\theta$
3. $Q_2(\boldsymbol{X}, \theta) = \bar{X}/\theta$

Example. (Normal pivots)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\mu, \sigma^2)$. The following are pivots:

1. $Q_1(\boldsymbol{X}, \boldsymbol{\theta}) = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
2. $Q_2(\boldsymbol{X}, \boldsymbol{\theta}) = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$

Fact. (Location family pivot)

If $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$ and $\{\mathbb{P}_\theta\}$ is a location family for $\theta \in \mathbb{R}$, then $Q(\boldsymbol{X}, \theta) = \bar{X} - \theta$ is a pivot.

So in general, how do we use pivots to get confidence sets? Because the distribution of pivots do not depend on θ, then the coverage probabilities are the same so we do not have to deal with different θ's.

The process is somewhat backwards from how we usually proceed:

1. **Set the "coverage probability"**

   Find $a \leq b$ (independent of θ) such that

   $$\mathbb{P}_\theta\Big(a \leq Q(\boldsymbol{X}, \theta) \leq b\Big) \;\geq\; 1 - \alpha$$

2. **Identify the hypothesis test**

   Let $A(\theta) = \{\boldsymbol{x} : a \leq Q(\boldsymbol{x}, \theta) \leq b\}$. Then,

   $$\mathbb{P}_{\theta_0}(\boldsymbol{X} \in A(\theta_0)^c) \;\leq\; \alpha \quad \forall \theta_0 \in \Theta$$

   So that a hypothesis test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ that rejects for $\boldsymbol{x} \in A(\theta_0)^c$ is level α.

3. **Invert the test**

   Set $C(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A(\theta)\} = \{\theta : a \leq Q(\boldsymbol{x}, \theta) \leq b\}$.

Theorem 6.3. Suppose $\Theta \subset \mathbb{R}$. Then, with the notation of the above,

1. If $Q(x, \theta)$ is increasing in θ for all $x \in \mathcal{X}$, then

$$C(x) = \{\theta : L(x, a) \leq \theta \leq U(x, b)\}$$

2. If $Q(x, \theta)$ is decreasing in θ for all $x \in \mathcal{X}$, then

$$C(x) = \{\theta : L(x, b) \leq \theta \leq U(x, a)\}$$

7 Probability inequalities

7.1 Hoeffding, Bernstein, and Bennett's inequalities

Theorem 7.1. (Extended Markov)

Let X be any r.v. with $E|X|^s < \infty$. Then:

$$P(|X| > t) = P(|X|^s > t^s) \leq \frac{E|X|^s}{t^s}$$

Proof. Trivial. □

Theorem 7.2. (Chebyshev-Cantelli)

For any r.v. X with $E[X^2] < \infty$ and $\forall t > 0$,

$$P(X - EX > t) \leq \frac{\text{Var}(X)}{\text{Var}(X) + t^2}$$

Proof. Assume WLOG $\mathbb{E}X = 0$. Note $t = \mathbb{E}(t - X) \leq \mathbb{E}[(t - X) \cdot \mathbb{1}_{X \leq t}]$. Applying the Cauchy-Schwarz inequality to the rightmost quantity, we get:

$$t \leq \left[E(t - X)^2 \cdot E(\mathbb{1}_{X \leq t})^2\right]^{\frac{1}{2}} = \left[(t^2 + EX^2) \cdot P(X \leq t)\right]^{\frac{1}{2}}$$

Then square both sides and rearrange. □

Hoeffding's inequality is one of the most important inequalities in the machine learning literature. It gives a **exponential** bound for sums of random variables where the increments are bounded. You should ask yourself: why is it good that the bound is exponential?

Theorem 7.3. (Hoeffding's inequality)

Let $X_1, \ldots, X_n$ be independent r.v.'s with $a_i \leq X_i \leq b_i$ a.s. Then $\forall\, t > 0$,

$$P(S_n - ES_n \geq t) \leq \exp\left\{\frac{-2t^2}{\sum(b_i - a_i)^2}\right\}$$
$$P(S_n - ES_n \leq -t) \leq \exp\left\{\frac{-2t^2}{\sum(b_i - a_i)^2}\right\}$$

And the two-sided bound:

$$P(\left|S_n - ES_n\right| \geq t) \leq 2 \exp\left\{\frac{-2t^2}{\sum(b_i - a_i)^2}\right\}$$

Not only is the inequality of great use, but its proof also draws on many of the concepts of the previous sections. The main idea of the proof is actually quite simple. First, we bound the MGFs of the increments using Taylor expansion. Then we plug these bounds into a Chernoff bound for the overall sum.

Lemma 7.4. (MGF bound for Hoeffding)

Let X be an r.v. with $EX = 0$ and $a \leq X \leq b$. Then $\forall s > 0$,

$$E(e^{sX}) \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$$

Proof. Fix some $s > 0$ and consider the function $f(x) = e^{sx}$. Then for every $x \in [a, b]$, Jensen's inequality gives us:

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}$$

Taking expectations (note $EX = 0$) and defining $p = -a/(b-a)$, we have:

$$Ee^{sX} \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} = \left[1 - p + pe^{s(b-a)}\right] e^{-ps(b-a)} = e^{\phi(u)}$$

Where $u = s(b-a)$ and $\phi(u) = -pu + \log(1-p+pe^u)$. Therefore it is sufficient to show that $\phi(u) \leq s^2(b-a)^2/8$.

Since ϕ is sufficiently smooth, the 1^{st} order Taylor expansion about $u = 0$ with Lagrange remainder is:

$$\phi(u) = \phi(0) + \phi'(0) \cdot u + \frac{u^2}{2}\phi''(c), \quad \text{some } c \in [0, u]$$

Now some calculate show that

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}} \quad \Rightarrow \quad \phi'(0) = 0$$

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \quad = \quad \frac{\alpha\beta}{(\alpha+\beta)^2} \leq \frac{1}{4}$$

Plugging this into the Taylor expansion above shows that

$$\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$$

□

Now the proof of the main result:

Proof. Applying Markov's to $P(S_n - ES_n \geq t) = P(e^{s(S_n - ES_n)}) \geq e^{st})$, we have:

$$\begin{aligned} P(S_n - ES_n \geq t) &\leq e^{-st}E\left\{\exp\left(s\sum_{i=1}^{n}(X_i - EX_i)\right)\right\} \\ &= e^{-st}E\left\{\prod_{i=1}^{n} e^{s(X_i - EX_i)}\right\} \\ &= e^{-st}\prod_{i=1}^{n} E\left\{e^{s(X_i - EX_i)}\right\} \quad \text{(by independence)} \end{aligned}$$

Now applying our lemma to the RHS, we have

$$P(S_n - ES_n \geq t) \leq e^{-st} \cdot \prod_{i=1}^{n} \exp\left\{\frac{s^2(b_i - a_i)^2}{8}\right\}$$
$$= e^{-st} \cdot \exp\left\{\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right\}$$

Choosing $s = 4t/\sum(b_i - a_i)^2$ completes the proof.

□

Like Hoeffding's inequality, both Bennett's and Bernstein's inequality give exponential bounds for a sum of random variables whose increments are bounded. However, the key difference is that Bennett's and Bernstein's inequality take into account the **variances** of the increments.

Theorem 7.5. (Bennett's inequality) Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \leq c$ for all i. Then for $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left\{\frac{-n\sigma^2}{c^2} \cdot h\left(\frac{ct}{n\sigma^2}\right)\right\}$$

where $h(u) = (1+u)\log(1+u) - u$, for $u \geq 0$.

The proof of this inequality is actually quite similar to the proof of Hoeffding's inequality. Again we first bound the MGFs of the increments using Taylor expansion, and plug those into a Chernoff-type bound for the overall sum.

Lemma 7.6. (MGF bound for Bennett)

Let X be an r.v. with $\mathbb{E}X = 0, \mathbb{E}X^2 = \sigma^2$, and $|X| \leq c$. Then:

$$\mathbb{E}(e^{sX}) \leq \exp\left\{\frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)\right\}, \quad \text{for all } s > 0$$

Proof. By Taylor series for e^x we have for any $s > 0$,

$$\mathbb{E}(e^{sX}) = \mathbb{E}\left\{1 + sX + \sum_{r=2}^{\infty}\frac{s^r X^r}{r!}\right\} = 1 + \sum_{r=2}^{\infty}\frac{s^r \mathbb{E}(X^r)}{r!}$$

Now by Holder's inequality,

$$\mathbb{E}X^r \;\leq\; \mathbb{E}|X|^r \;\leq\; \mathbb{E}|X|^2|X|^{r-2} \;\leq\; \sigma^2 \cdot c^{r-2}$$

Therefore plugging this into our Taylor expansion and summing, we obtain:

$$\mathbb{E}(e^{sX}) \leq 1 + \sum_{r=2}^{\infty}\frac{s^r\sigma^2 c^{r-2}}{r!}$$
$$= 1 + \frac{\sigma^2}{c^2}\sum_{r=2}^{\infty}\frac{(sc)^r}{r!}$$
$$= 1 + \frac{\sigma^2}{c^2}(e^{sc} - 1 - sc)$$

Then apply the identity $1 + x \leq e^x$.

□

Now we plug into a basic Chernoff-type bound to prove the main result:

Proof. (of Bennett's inequality)

Recall that we have $X_1, \ldots, X_n$ be independent with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = \sigma_i^2$, and $|X_i| \leq c$ for all i.

Define $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Then by the basic Chernoff bound for $s > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i > t\right) \leq e^{-st}\prod_{i=1}^n \mathbb{E}e^{sX_i} \leq \exp\left\{\frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st\right\}$$

Now this bound holds for all $s > 0$, so we optimize. To ease notation define the function:

$$f(s) = \frac{n\sigma^2(e^{sc} - 1 - sc)}{c^2} - st$$

Some high school calculus shows this to be minimized at:

$$s_0 = c^{-1}\log(1 + tc/n\sigma^2)$$

and substituting into the above expression gives the result.

□

Theorem 7.7. (Some expected value bounds)

1. Let X be an r.v. such that $P(|X| > t) \leq ae^{-bt^2}$ with $a, b > 0$ and $a \geq 1$. Then
$$E|X| \leq \sqrt{\frac{1 + \log a}{b}}$$

2. Let X be an r.v. such that $P(|X| > t) \leq ae^{-nbt^2}$ with $a, b > 0$ and $a \geq 1$. Then
$$E|X| \leq \sqrt{\frac{1 + \log a}{nb}}$$

Proof. Note that $\forall x \geq 0$,

$$\begin{aligned} EX^2 &= \int_0^\infty P(X^2 > t)\,\mathrm{d}t \\ &= \int_0^s P(X^2 > t)\,\mathrm{d}t + \int_s^\infty P(X^2 > t)\,\mathrm{d}t \\ &\leq s + \int_s^\infty P(|X| > \sqrt{t})\,\mathrm{d}t \end{aligned}$$

Apply the bound stated in the assumptions, choose an appropriate s, and finally apply Jensen's.

7.2 Projections and conditional expectation

Theorem 7.8. (Equivalent definitions of projections)

Let $\mathcal{S}$ be a linear space of random variables with finite second moments and let $T \in \mathcal{S}$. $\hat{S}$ is the L_2-projection of T onto a linear subspace $\mathcal{G}$ iff $\hat{S} \in \mathcal{G}$ and:

1. $\mathbb{E}(\hat{S} - T)^2 = \inf_{S \in \mathcal{G}} \mathbb{E}(S - T)^2$, or
2. $\mathbb{E}(T - \hat{S}) \cdot S = 0$ for all $S \in \mathcal{G}$

Theorem 7.9. (Two essential properties)

1. If $\hat{S}, \tilde{S}$ are projections of T onto $\mathcal{G}$, then $\hat{S} = \tilde{S}$.
2. If $\mathcal{G}$ contains constants, then:
   (a) $\mathbb{E}\hat{S} = \mathbb{E}T$
   (b) $\mathrm{Cov}(T - \hat{S}, S) = 0$ for all $S \in \mathcal{G}$

Theorem 7.10. (Moments in terms of projections)

Let $\hat{S}$ be the projection of T onto $\mathcal{G}$. Then:

1. $\mathbb{E}T^2 = \mathbb{E}(T - \hat{S})^2 + \mathbb{E}\hat{S}^2$
2. $\mathrm{Var}(T) = \mathrm{Var}(T - \hat{S}) + \mathrm{Var}(\hat{S})$ if $\mathcal{G}$ contains constants

Theorem 7.11. (Linearity)

If the projections of T_1, T_2 onto $\mathcal{G}$ are $\hat{S}_1, \hat{S}_2$ respectively, then the projection of $aT_1 + bT_2$ onto $\mathcal{G}$ is $a\hat{S}_1 + b\hat{S}_2$.

7.3 Concentration inequalities

Definition. (Martingale difference)

Let $X_1, \ldots, X_n \in \mathcal{X}$ be r.v.'s and let $f : \mathcal{X}^n \to \mathbb{R}$ be bounded. Define the following:

1. $x_i^j = x_i, x_{i+1}, \ldots, x_j$ for $1 \le i \le j \le n$.
2. $V = f(x_i^n) - E[f(x_i^n)]$
3. $V_i = E[f(x_i^n) \,|\, x_1^i] - E[f(x_i^n) \,|\, x_1^{i-1}]$
4. $E[f(x_i^n) \,|\, x_1^0] = E[f(x_1^n)]$

Lemma 7.12. (Lemmas for bounded difference inequality)

1. $V = \sum V_i$
2. $E[V_i \,|\, X_1^{i-1}] = 0$ and $EV_i = 0$
3. $E[V_i V_j] = 0, \quad i \neq j$

4. $\text{Var}[f(x_1^n)] = \sum EV_i^2 = EV^2$

Proof. The proofs of 1 and 2 are trivial.

Proof of 3: WLOG, assume $i < j$. Write $E[V_iV_j]$ as $E\big\{E[V_iV_j \mid X_1^i]\big\}$ and use the fact that V_i is a function of X_1^i.

Proof of 4: Note that:

$$\begin{aligned}\text{Var}[f(x_1^n)] &= EV^2 \\ &= E\,(\sum^n V_i)^2 \\ &= \sum^n E(V_iV_j)\end{aligned}$$

and use the fact that $E[V_iV_j] = 0$.

Definition. (Difference coefficient)

Let $X_1, \ldots, X_n \in \mathcal{X}$ be r.v.'s and let $f : \mathcal{X}^n \to \mathbb{R}$. Define $Z = f(x_1^n)$.

The **i'th difference coefficient of** f is defined by

$$c_i = \sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} \left| f(x_1^{i-1}, x_i, x_{i+1}^n) - f(x_1^{i-1}, x_i', x_{i+1}^n) \right|$$

Theorem 7.13. (Bounded Difference)

$\forall t \geq 0$,

$$P(Z - EZ > t) \leq \exp\left\{ \frac{-2t^2}{\sum^n c_i^2} \right\}$$

$$P(EZ - Z > t) \leq \exp\left\{ \frac{-2t^2}{\sum^n c_i^2} \right\}$$

Proof.

Set $V = Z - EZ$, $V_i = E(Z|x_1^i) - E(Z|x_1^{i-1})$, and $H_i(x_1^i) = E(Z|x_1^i)$.

First we show that $V_i \mid x_1^{i-1}$ satisfies the conditions of Hoeffding's inequality.

Write $V_i = H_i(x_1^i) - H_{i-1}(x_1^{i-1})$. Note:

$$\inf_{u'} H_i(x_1^{i-1}, u') - H_{i-1}(x_1^{i-1}) \leq V_i$$

$$\sup_{u} H_i(x_1^{i-1}, u) - H_{i-1}(x_1^{i-1}) \geq V_i$$

It follows that the (maximum) difference between the two LHS expressions is:

$$\sup_{u,u'} \left[H_i(x_1^{i-1}, u) - H_i(x_1^{i-1}, u') \right]$$

and that this difference is $\leq c_i$ (from HW). Thus V_i is bounded inside an interval of width $\leq c_i$. Thus $V_i \mid x_1^{i-1}$ is similarly bounded. Also, by a previous

lemma we have that $E[V_i \,|\, x_1^{i-1}] = 0$, so we can apply Hoeffding's inequality to obtain

$$\forall s > 0,\ E\left[e^{sV_i} \,|\, x_1^{i-1}\right] \leq e^{(s^2 c_i^2)/8}$$

Now, separately, apply Chernoff's MGF bound to $V = \sum V_i$ to obtain

$$\begin{aligned}\forall s > 0,\ P(V \geq t) &\leq e^{-st} E e^{sV} \\ &= e^{-st} E\left\{\exp\left(\sum^n sV_i\right)\right\}\end{aligned}$$

Now we successively peel off terms like so:

$$\begin{aligned}\forall s > 0, P(V \geq t) &\leq e^{-st} E\left\{\exp\left(\sum^{n-1} sV_i\right) \cdot E\left[e^{sV_n} \,|\, x_1^n - 1\right]\right\} \\ &\leq e^{-st} e^{(s^2 c_i^2)/8} E\left\{\exp\left(\sum^{n-1} sV_i\right)\right\} \\ &\vdots \\ &\leq \exp\left\{-st + \frac{1}{8}\sum^n s^2 c_i^2\right\}\end{aligned}$$

and choose $s = 4t/\sum c_i^2$.

Theorem 7.14. Let $X_1, \ldots, X_n$ be independent and $f : \mathcal{X}^n \to \mathbb{R}$ have difference coefficients $c_1, \ldots, c_n$. Then

$$\mathrm{Var}(f(x_1^n)) \leq \frac{1}{4}\sum c_i^2$$

Proof. Note $V = \sum V_i$. Then

$$\begin{aligned}\mathrm{Var}(f(x_1^n)) &= E(V^2) \\ &= \sum E(V_i^2) \\ &= \sum E\left[E(V_i^2 \,|\, x_1^{i-1})\right] \\ &= \sum E\left[\mathrm{Var}(V_i \,|\, x_1^{i-1})\right] \quad (\text{since } E(V_i \,|\, x_1^{i-1}) = 0)\end{aligned}$$

Note that V_i is bounded inside an interval of width c_i, so $\mathrm{Var}(V_i) \leq c_i^2/4$.

Theorem 7.15. (Efron-Stein)

Let $X_1, \ldots, X_n$ be independent and $f : \mathcal{X}^n \to \mathbb{R}$. Also let $Z = f(x_1^n)$ with $E[Z^2] < \infty$. Define:

$$\begin{aligned}E_i(Z) = E[Z \,|\, x_1^n],\ E_0(Z) = E(Z) \\ E^{(i)}(Z) = E[Z \,|\, x_1^{i-1}, x_{i+1}^n]\end{aligned}$$

Then

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n E\left[(Z - E^{(i)}Z)^2\right]$$

Proof. First, show that $E_i(E^{(i)}Z) = E_{i-1}Z$ using the properties of conditional expectation.

Next note that $V_i = E_i(Z) - E_{i-1}(Z)$ and, as shown in the previous theorem, $\text{Var}(Z) = \sum E(V_i^2)$. Thus:

$$\begin{aligned} V_i &= E_i(Z) - E_{i-1}(Z) \\ &= E_i(Z) - E_i(E^{(i)}(Z)) \\ &= E_i(Z - E^{(i)}(Z)) \end{aligned}$$

$$\begin{aligned} V_i^2 &= \left[E_i(Z - E^{(i)}(Z))\right]^2 \\ &\leq E_i\left[(Z - E^{(i)}(Z))^2\right] \quad \text{(by Jensen)} \\ E(V_i^2) &\leq E_i\left[(Z - E^{(i)}(Z))^2\right] \quad \text{(property of expectation)} \end{aligned}$$

And the result follows.

Corollary 7.16. $\text{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}((Z - E^{(i)}(Z))^2)$

Lemma 7.17. (Bounds on maxima)

Let $Y_1, \ldots, Y_n$ be r.v.'s such that for some $\sigma > 0$,

$$\mathbb{E}(e^{sY_i}),\ \mathbb{E}(e^{-sY_i}) \leq \exp\left(\frac{s^2\sigma^2}{2}\right), \quad \forall s > 0,\ i = 1, \ldots, n$$

Then for $n \geq 3$,

1. $\mathbb{E}(\max_{1\leq i\leq n} Y_i) \leq \sqrt{2\sigma^2 \log n}$
2. $\mathbb{E}(\max_{1\leq i\leq n} |Y_i|) \leq \sqrt{2\sigma^2 \log(2n)}$

Corollary 7.18. If $Z_1, \ldots, Z_n \sim N(0, \sigma^2)$, then:

$$\mathbb{E} \max_{1\leq i\leq n} Z_i \leq \sqrt{2\sigma^2 \log n}$$

8 Random vectors and matrices

8.1 Basic properties

Lemma 8.1. (Some covariance identities)

Let $X_{(k\times 1)}$ and $Y_{(l\times 1)}$.

1. $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)']$, $\quad (k\times l)$
2. $\mathrm{Cov}(AX+a, BY+b) = A\mathrm{Cov}(X,Y)B'$
3. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Var}(Y)$
4. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)'$

8.2 Multivariate Normal

Definition. (P-variate normal with positive-definite variance)

$X \in \mathbb{R}^p$ is **p-variate normal with positive-definite variance** with parameters $\mu \in \mathbb{R}^p$ and $\Sigma > 0$, written $X \sim N_p(\mu, \Sigma)$ if it has pdf:

$$f(x) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right\}$$

Note: $\Sigma > 0 \Rightarrow \Sigma^{-1} > 0$.

Theorem 8.2. $f(x)$ as given above is a density.

Proof. Define $H : \mathbb{R}^p \to \mathbb{R}^p$ by $H(x) = \Sigma^{-\frac{1}{2}}(X-\mu) \triangleq Y$.

Thus $H^{-1}(Y) = \Sigma^{\frac{1}{2}}Y + \mu$. Note that $H^{-1}(Y)$ has Jacobian $\Sigma^{\frac{1}{2}}$. Then

$$\int_{\mathbb{R}^p} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^p} f(H^{-1}(Y))|\Sigma^{\frac{1}{2}}|\,\mathrm{d}y$$

Fill out the details and apply Fubini's theorem at the end to obtain the result.

Corollary 8.3. If $X \sim N_p(\mu, \Sigma)$ (invertible Σ), then

$$Y = \Sigma^{-\frac{1}{2}}(X-\mu) \sim N_p(0, I_p)$$

Corollary 8.4. (Representation theorem for N_p with invertible Σ)

Let $X \sim N_p(\mu, \Sigma)$ where $\Sigma > 0$. Then

$$X \overset{d}{=} \Sigma^{\frac{1}{2}}Y + \mu \quad \text{where} \quad Y \sim N_p(0, I_p)$$

Corollary 8.5. $X \sim N_p(\mu, \Sigma)$ where $\Sigma > 0$, then

$$U = (X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$$

Theorem 8.6. (Cramer-Wold device)

Let $X, Y \in \mathbb{R}^p$ be random vectors. Then

$$X \stackrel{d}{=} Y \iff u'X \stackrel{d}{=} u'Y \quad \forall u \in \mathbb{R}^p$$

Proof. Hard.

Definition. (Multinormal with non-negative definite variance)

A random vector $X \in \mathbb{R}^p$ is **multinormal with non-negative definite variance** if

$$u'X \sim N_1 \quad \forall u \in \mathbb{R}^p$$

Note: $X_i = e_i'X \sim N_1 \Rightarrow E(X_i), \mathrm{Var}(X_i) < \infty$.

Definition. (Jointly multinormal)

Two random vectors X, Y are **jointly multinormal** if $\binom{X}{Y}$ is multinormal.

Theorem 8.7. (Representation theorem for multinormal)

Let $X \in \mathbb{R}^p$ be multinormal with $EX = \mu$, $\mathrm{Var}(X) = \Sigma \geq 0$. Then

$$X = \mu + \Sigma^{\frac{1}{2}} Y \quad \text{where} \quad Y \sim N_p(0, I_p)$$

Proof. Since X is multinormal, we know that $u'X \sim N_1(u'\mu, u'\Sigma u)$ $\forall u' \in \mathbb{R}^p$ by the properties of expectation and variance on random vectors.

To show the theorem let $X' = \mu + \Sigma^{\frac{1}{2}} Y$ where $Y \sim N_p(0, I_p)$ and show that $u'X'$ has the same distribution as $u'X$. Then the result follows by the Cramer-Wold device.

Theorem 8.8. Suppose X is multinormal with mean μ and variance Σ_p, and A is $q \times p$, $c \in \mathbb{R}^q$. Then:

$$Y = AX + c \text{ is multinormal with } \mathbb{E}Y = A\mu + c,\ \mathrm{Var}(Y) = A\Sigma A')$$

Corollary 8.9. If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are jointly multinormal, then $X \perp Y$ iff $\mathrm{Cov}(X, Y) = 0$.

Theorem 8.10. If $X \in \mathbb{R}^p$ be multinormal with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \Sigma$. If $A_{(q \times p)}$ and $B_{(l \times p)}$ are constant matrices, then:

$$Y = AX \perp Z = BX \iff A\Sigma B' = 0$$