Intermediate Statistics

Instructor: Larry Wasserman

Course description

This course will cover the fundamentals of theoretical statistics. This course is excellent preparation for advanced work in statistics and machine learning.

Textbook: Wasserman, L. (2004). All of Statistics: A concise course in statistical inference.

Other Recommended Texts:

1) Casella, G. and Berger, R. L. (2002). Statistical Inference, 2nd ed.

2) Bickel, P. J. and Doksum, K. A. (1977). Mathematical Statistics

3) Rice, J. A. (1977). Mathematical Statistics and Data Analysis, Second Edition.

4) Van der Vaart, A. (2000). Asymptotic Statistics.

https://www.stat.cmu.edu/~larry/=stat705/

Carnegie Mellon University

Lecture Notes 1
36-705

Our broad goal for the first few lectures is to try to understand the behaviour of sums of independent random variables. We would like to find ways to formalize the fact:

Averages of independent random variables concentrate around their expectation.

We will try to answer this question from the asymptotic (i.e. the number of random variables we average $\to \infty$) and the non-asymptotic viewpoint (i.e. the number of random variables is some fixed finite number). The asymptotic viewpoint is typically characterized by what are known as the Laws of Large Numbers (LLNs) and Central Limit Theorems (CLTs) while the non-asymptotic viewpoint is characterized by concentration inequalities.

We will need to first review what a random variable is, what its expectation is, and what we precisely mean by concentration. This will be fairly terse. See Chapters 1-3 of the book for more details.

Warning: This is a review. We will go quickly because I assume you have taken some probability.

1 Random Variables

Let Ω be a sample space (a set of possible outcomes) with a probability distribution (also called a probability measure) P. A *random variable* is a map $X : \Omega \to \mathbb{R}$. We write

$$P(X \in A) = P(\{\omega \in \Omega : \ X(\omega) \in A\})$$

and we write $X \sim P$ to mean that X has distribution P. The *cumulative distribution function (cdf)* of X is

$$F_X(x) = F(x) = P(X \leq x).$$

A cdf has three properties:

1. F is right-continuous. At each x, $F(x) = \lim_{n\to\infty} F(y_n) = F(x)$ for any sequence $y_n \to x$ with $y_n > x$.
2. F is non-decreasing. If $x < y$ then $F(x) \leq F(y)$.
3. F is normalized. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Conversely, any F satisfying these three properties is a cdf for some random variable.

If X is discrete, its *probability mass function (pmf)* is

$$p_X(x) = p(x) = P(X = x).$$

If X is continuous, then its *probability density function function (pdf)* satisfies

$$P(X \in A) = \int_A p_X(x)dx = \int_A p(x)dx$$

and $p_X(x) = p(x) = F'(x)$. The following are all equivalent:

$$X \sim P, \quad X \sim F, \quad X \sim p.$$

Suppose that $X \sim P$ and $Y \sim Q$. We say that X and Y have the same distribution if $P(X \in A) = Q(Y \in A)$ for all A. In that case we say that X and Y are *equal in distribution* and we write $X \stackrel{d}{=} Y$.

Lemma 1 *$X \stackrel{d}{=} Y$ if and only if $F_X(t) = F_Y(t)$ for all t.*

2 Expected Values

The *mean* or expected value of $g(X)$ is

$$\mathbb{E}\left(g(X)\right) = \int g(x)dF(x) = \int g(x)dP(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases}$$

Recall that:

1. **Linearity of Expectations:** $\mathbb{E}\Big(\sum_{j=1}^k c_j g_j(X)\Big) = \sum_{j=1}^k c_j \mathbb{E}(g_j(X))$.
2. If $X_1, \ldots, X_n$ are independent then
$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_i \mathbb{E}\left(X_i\right).$$
3. We often write $\mu = \mathbb{E}(X)$.

4. $\sigma^2 = \mathsf{Var}\,(X) = \mathbb{E}\,((X-\mu)^2)$ is the **Variance**.

5. $\mathsf{Var}\,(X) = \mathbb{E}\,(X^2) - \mu^2$.

6. If $X_1, \ldots, X_n$ are independent then

$$\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_i a_i^2 \mathsf{Var}\,(X_i)\,.$$

7. The covariance is

$$\mathsf{Cov}(X,Y) = \mathbb{E}\Big((X-\mu_x)(Y-\mu_y)\Big) = \mathbb{E}(XY) - \mu_X \mu_Y$$

and the correlation is $\rho(X,Y) = \mathsf{Cov}(X,Y)/\sigma_x\sigma_y$. Recall that $-1 \leq \rho(X,Y) \leq 1$.

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X)$ whose value, when $X = x$ is

$$\mathbb{E}(Y|X=x) = \int y\; p(y|x)dy$$

where $p(y|x) = p(x,y)/p(x)$.

The *Law of Total Expectation* or *Law of Iterated Expectation*:

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \int \mathbb{E}(Y|X=x)p_X(x)dx.$$

The *Law of Total Variance* is

$$\mathsf{Var}(Y) = \mathsf{Var}\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[\mathsf{Var}(Y|X)\big].$$

The *moment generating function (mgf)* is

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

If $M_X(t) = M_Y(t)$ for all t in an interval around 0 then $X \stackrel{d}{=} Y$.

The moment generating function can be used to "generate" all the moments of a distribution, i.e. we can take derivatives of the mgf with respect to t and evaluate at $t = 0$, i.e. we have that

$$M_X^{(n)}(t)|_{t=0} = \mathbb{E}\,(X^n)\,.$$

3 Independence

X and Y are *independent* if and only if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all A and B.

Theorem 2 *Let (X, Y) be a bivariate random vector with $p_{X,Y}(x, y)$. X and Y are independent iff $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.*

$X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Thus, $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$.

If $X_1, \ldots, X_n$ are independent and identically distributed we say they are iid (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P \quad \text{or} \quad X_1, \ldots, X_n \sim F \quad \text{or} \quad X_1, \ldots, X_n \sim p.$$

4 Transformations

Let $Y = g(X)$ where $g : \mathbb{R} \to \mathbb{R}$. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \int_{A(y)} p_X(x)dx$$

where

$$A(y) = \{x : \ g(x) \leq y\}.$$

The density is $p_Y(y) = F'_Y(y)$. If g is strictly monotonic, then

$$p_Y(y) = p_X(h(y)) \left| \frac{dh(y)}{dy} \right|$$

where $h = g^{-1}$.

Example 3 *Let $p_X(x) = e^{-x}$ for $x > 0$. Hence $F_X(x) = 1 - e^{-x}$. Let $Y = g(X) = \log X$. Then*

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(\log(X) \leq y) \\ &= P(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y} \end{aligned}$$

and $p_Y(y) = e^y e^{-e^y}$ for $y \in \mathbb{R}$.

Example 4 Practice problem. *Let X be uniform on $(-1, 2)$ and let $Y = X^2$. Find the density of Y.*

Let $Z = g(X, Y)$. For example, $Z = X + Y$ or $Z = X/Y$. Then we find the pdf of Z as follows:

1. For each z, find the set $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find the CDF

$$F_Z(z) \quad = \quad P(Z \leq z) = P(g(X, Y) \leq z) = P(\{(x, y) : g(x, y) \leq z\}) = \int\int_{A_z} p_{X,Y}(x, y)dxdy.$$

3. The pdf is $p_Z(z) = F_Z'(z)$.

Example 5 Practice problem. *Let (X, Y) be uniform on the unit square. Let $Z = X/Y$. Find the density of Z.*

5 Important Distributions

Normal (Gaussian). $X \sim N(\mu, \sigma^2)$ if

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

If $X \in \mathbb{R}^d$ then $X \sim N(\mu, \Sigma)$ if

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right).$$

Chi-squared. $X \sim \chi^2_p$ if $X = \sum_{j=1}^p Z_j^2$ where $Z_1, \ldots, Z_p \sim N(0, 1)$.

Non-central chi-squared (more on this below). $X \sim \chi^2_1(\mu^2)$ if $X = Z^2$ where $Z \sim N(\mu, 1)$.

Bernoulli. $X \sim \text{Bernoulli}(\theta)$ if $\mathbb{P}(X = 1) = \theta$ and $\mathbb{P}(X = 0) = 1 - \theta$ and hence

$$p(x) = \theta^x (1 - \theta)^{1-x} \qquad x = 0, 1.$$

Binomial. $X \sim \text{Binomial}(\theta)$ if

$$p(x) = \mathbb{P}(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x \in \{0, \ldots, n\}.$$

Uniform. $X \sim \text{Uniform}(0, \theta)$ if $p(x) = I(0 \leq x \leq \theta)/\theta$.

Poisson. $X \sim \text{Poisson}(\lambda)$ if $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$. The $\mathbb{E}(X) = \mathsf{Var}(X) = \lambda$ and $M_X(t) = e^{\lambda(e^t-1)}$. We can use the mgf to show: if $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, independent then $Y = X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Multinomial. The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn with has balls with k different colors labeled "color 1, color 2, ..., color k." Let $p = (p_1, p_2, \ldots, p_k)$ where $\sum_j p_j = 1$ and p_j is the probability of drawing color j. Draw n balls from the urn (independently and with replacement) and let $X = (X_1, X_2, \ldots, X_k)$ be the count of the number of balls of each color drawn. We say that X has a Multinomial (n, p) distribution. The pdf is

$$p(x) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}.$$

Exponential. $X \sim \exp(\beta)$ if $p_X(x) = \frac{1}{\beta}e^{-x/\beta}$, $x > 0$. Note that $\exp(\beta) = \Gamma(1, \beta)$.

Gamma. $X \sim \Gamma(\alpha, \beta)$ if

$$p_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$$

for $x > 0$ where $\Gamma(\alpha) = \int_0^\infty \frac{1}{\beta^\alpha}x^{\alpha-1}e^{-x/\beta}dx$.

Remark: In all of the above, make sure you understand the distinction between random variables and parameters.

More on the Multivariate Normal. Let $Y \in \mathbb{R}^d$. Then $Y \sim N(\mu, \Sigma)$ if

$$p(y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right).$$

Then $\mathbb{E}(Y) = \mu$ and $\mathsf{cov}(Y) = \Sigma$. The moment generating function is

$$M(t) = \exp\left(\mu^T t + \frac{t^T\Sigma t}{2}\right).$$

Theorem 6 *(a). If $Y \sim N(\mu, \Sigma)$, then $E(Y) = \mu$, $cov(Y) = \Sigma$.*

(b). If $Y \sim N(\mu, \Sigma)$ and c is a scalar, then $cY \sim N(c\mu, c^2\Sigma)$.

(c). Let $Y \sim N(\mu, \Sigma)$. If A is $p \times n$ and b is $p \times 1$, then $AY + b \sim N(A\mu + b, A\Sigma A^T)$.

Theorem 7 *Suppose that $Y \sim N(\mu, \Sigma)$. Let*

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11}\, \Sigma_{12} \\ \Sigma_{21}\, \Sigma_{22} \end{pmatrix}.$$

where Y_1 and μ_1 are $p \times 1$, and Σ_{11} is $p \times p$.

(a) $Y_1 \sim N_p(\mu_1, \Sigma_{11}), Y_2 \sim N_{n-p}(\mu_2, \Sigma_{22})$.

(b) Y_1 and Y_2 are independent if and only if $\Sigma_{12} = 0$.

(c) If $\Sigma_{22} > 0$, then the condition distribution of Y_1 given Y_2 is

$$Y_1 | Y_2 \sim N_p(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \tag{1}$$

Lemma 8 *Let $Y \sim N(\mu, \sigma^2 I)$, where $Y^T = (Y_1, \ldots, Y_n), \mu^T = (\mu_1, \ldots, \mu_n)$ and $\sigma^2 > 0$ is a scalar. Then the Y_i are independent, $Y_i \sim N_1(\mu, \sigma^2)$ and*

$$\frac{\|Y\|^2}{\sigma^2} = \frac{Y^T Y}{\sigma^2} \sim \chi_n^2 \left(\frac{\mu^T \mu}{\sigma^2} \right).$$

Theorem 9 *Let $Y \sim N(\mu, \Sigma)$. Then:*
(a) $Y^T \Sigma^{-1} Y \sim \chi_n^2(\mu^T \Sigma^{-1} \mu)$.
(b) $(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi_n^2(0)$.

6 Sample Mean and Variance

Let $X_1, \ldots, X_n \sim P$. The sample mean is

$$\overline{X}_n = \widehat{\mu}_n = \frac{1}{n} \sum_i X_i$$

and the sample variance is

$$S_n^2 = \widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_i (X_i - \widehat{\mu}_n)^2.$$

Some authors instead define the sample variance as

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_i (X_i - \widehat{\mu}_n)^2.$$

The *sampling distribution* of $\widehat{\mu}_n$ is

$$G_n(t) = \mathbb{P}(\widehat{\mu}_n \leq t).$$

Practice Problem. Let $X_1, \ldots, X_n$ be iid with $\mu = \mathbb{E}(X_i) = \mu$ and $\sigma^2 = \mathsf{Var}(X_i) = \sigma^2$. Then

$$\mathbb{E}(\widehat{\mu}_n) = \mu, \quad \mathsf{Var}(\widehat{\mu}_n) = \frac{\sigma^2}{n}, \quad \mathbb{E}(\widehat{\sigma}_n^2) = \sigma^2.$$

Theorem 10 *If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ then*

(a) $\widehat{\mu}_n \sim N(\mu, \frac{\sigma^2}{n})$.

(b) $\frac{(n-1)\widehat{\sigma}_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

(c) $\widehat{\mu}_n$ *and* $\widehat{\sigma}_n^2$ *are independent.*

7 A preview of the next few lectures

Let us consider a simple experiment. I toss a fair coin n times, and if the outcome is heads I record $X_i = +1$, and if the outcome is tails I record $X_i = -1$. These are called *Rademacher* random variables. Now, let us consider the average:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

It is easy to see that $\mathbb{E}[\widehat{\mu}_n] = 0$, and we would like to know how far $\widehat{\mu}_n$ is from its expectation. When all the $X_i = +1$ (for instance), we have have that $\widehat{\mu}_n = 1$. There is however a remarkable phenomenon, known as the concentration of measure phenomenon, that asserts that $\widehat{\mu}_n$ "concentrates" much closer to $\mathbb{E}[\widehat{\mu}_n]$.

The average of n i.i.d random variables concentrates within an interval of length roughly $1/\sqrt{n}$ around the mean.

The basic intuition is that in order to for a sample average to be far from the expectation, many *independent* random variables need to work together simultaneously, which is extremely unlikely. Moreover, $\hat{\mu}_n$ has, approximately, a Normal distribution. These seemingly simple facts underly essentially all of statistics and machine learning.

Lecture Notes 2
36-705

Recall in the last class we discussed that we would like to understand the behaviour of the average of independent random variables. Towards that goal let us begin by trying to understand the tail behaviour of a random variable.

1 Markov Inequality

The most elementary tail bound is Markov's inequality, which asserts that for a positive random variable $X \geq 0$, with finite mean,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} = O\left(\frac{1}{t}\right).$$

Intuitively, if the mean of a (positive) random variable is small then it is unlikely to be too large too often, i.e. the probability that it is large is small. While Markov on its own is fairly crude it will form the basis for much more refined tail bounds.

Proof: Note that

$$\mathbb{E}[X] = \int_0^\infty xp(x)dx \geq \int_t^\infty xp(x)dx \geq t\int_t^\infty p(x)dx = t\mathbb{P}(X \geq t).$$

2 Chebyshev Inequality

Chebyshev's inequality states that for a random variable X, with $\mathsf{Var}(X) = \sigma^2$, for any $t > 0$,

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq t\sigma\Big) \leq \frac{1}{t^2} = O\left(\frac{1}{t^2}\right).$$

Before we prove this let's look at a simple application. In the last lecture we saw that if we average i.i.d. random variables with mean μ and variance σ^2, we have that the average:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i,$$

has mean μ and variance σ^2/n. So applying Chebyshev's inequality to $\widehat{\mu}_n$ we obtain that,

$$\mathbb{P}\left(|\widehat{\mu}_n - \mu| \geq \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{1}{t^2}.$$

So, with probability at least 0.99 (for instance) the average is within $10\sigma/\sqrt{n}$ from the its expectation. This is pretty neat and almost directly gives us something called the Weak Law of Large Numbers (but we will return to this).

We will study refinements of this inequality today, but in some sense it already has the correct "$1/\sqrt{n}$" behaviour. The refinements will mainly be to show that in many cases we can dramatically improve the constant 10.

Proof: Chebyshev's inequality is an immediate consequence of Markov's inequality.

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq t\sigma) &= \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2\sigma^2) \\ &\leq \frac{\mathbb{E}(|X - \mathbb{E}[X]|^2)}{t^2\sigma^2} = \frac{1}{t^2}.\end{aligned}$$

3 Chernoff Method

There are several refinements to the Chebyshev inequality. One simple one that is sometimes useful is to observe that if the random variable X has a finite k-th central moment then we have that,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}|X - \mathbb{E}[X]|^k}{t^k}.$$

For many random variables (we will see some examples today), the moment generating function will exist in a neighborhood around 0, i.e the mgf is finite for all $|t| \leq b$ where $b > 0$ is some constant. In these cases, we can use the mgf to produce a tail bound.

Define, $\mu = \mathbb{E}[X]$. For any $t > 0$, we have that,

$$\mathbb{P}((X - \mu) \geq u) = \mathbb{P}(\exp(t(X - \mu)) \geq \exp(tu)) \leq \frac{\mathbb{E}[\exp(t(X - \mu))]}{\exp(tu)}$$

by applying Markoov's inequality. Now t is a parameter we can choose to get a tight upper bound, i.e. we can write this bound as:

$$\mathbb{P}((X - \mu) \geq u) \leq \inf_{0 \leq t \leq b} \exp(-t(u + \mu))\mathbb{E}[\exp(tX)].$$

This bound is known as Chernoff's bound.

3.1 Gaussian Tail Bounds via Chernoff

Suppose that, $X \sim N(\mu, \sigma^2)$, then a simple calculation gives that the mgf of X is:

$$M_X(t) = \mathbb{E}[\exp(tX)] = \exp(t\mu + t^2\sigma^2/2).$$

The mgf is defined for all t. To apply the Chernoff bound we then need to compute:

$$\inf_{t \geq 0} \exp(-t(u+\mu)) \exp(t\mu + t^2\sigma^2/2) = \inf_{t \geq 0} \exp(-tu + t^2\sigma^2/2),$$

which is minimized when $t = u/\sigma^2$ which in turn yields the tail bound,

$$\mathbb{P}(X - \mu \geq u) \leq \exp(-u^2/(2\sigma^2)).$$

This is often referred to as a one-sided or upper tail bound. We can use the fact that if X has distribution $N(\mu, \sigma^2)$ then $-X$ has distribution $N(-\mu, \sigma^2)$ and repeat the above calculation to obtain the analogous lower tail bound,

$$\mathbb{P}(-X + \mu \leq u) \leq \exp(-u^2/(2\sigma^2)).$$

Putting these two pieces together, we have the two-sided Gaussian tail bound:

$$\mathbb{P}(|X - \mu| \geq u) \leq 2\exp(-u^2/(2\sigma^2)).$$

The main thing to observe is that this inequality is much sharper than Chebyshev's inequality. In particular, suppose we consider the average of i.i.d Gaussian random variables, i.e. we have $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and we construct the estimate:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Using the fact that the average of Gaussian RVs is Gaussian we obtain that $\widehat{\mu}$ has a $N(\mu, \sigma^2/n)$ distribution. In this case, using the Gaussian tail bound we derived we obtain that,

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t\sigma/\sqrt{n}) \leq 2\exp(-t^2/2).$$

This is an example of an exponential tail inequality. Comparing with Chebyshev's inequality we should observe two things:

1. Both inequalities say roughly that the deviation of the average from the expected value goes down as $1/\sqrt{n}$.
2. However, the Gaussian tail bound says if the random variables are actually Gaussian then the chance that the deviation is much bigger than $\sigma/\sqrt{n}$ goes down *exponentially fast*. Let us look at a concrete example, we say previously that Chebyshev told us the average is within $10\sigma/\sqrt{n}$ with probability at least 0.99.

   On the other hand the exponential tail bound says that with probability 0.99 the average is within,

$$\sqrt{2\ln(1/0.005)}\sigma/\sqrt{n} \approx 3.25\sigma/\sqrt{n}.$$

More generally, Chebyshev tells us that with probability at least $1 - \delta$,

$$|\widehat{\mu} - \mu| \leq \frac{\sigma}{\sqrt{n\delta}}$$

while the exponential tail bound tells us that,

$$|\widehat{\mu} - \mu| \leq \sigma\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

The first goes up polynomially as $\delta \to 0$, while the second more refined bound goes up only logarithmically.

3.2 Sub-Gaussian Random Variables

It turns out that the Gaussian tail inequality from the previous section is much more broadly applicable to a class of random variables called sub-Gaussian random variables. Roughly these are random variables whose tails decay faster than a Gaussian. Formally, a random variable X with mean μ is *sub-Gaussian* if there exists a positive number σ such that,

$$\mathbb{E}[\exp(t(X - \mu))] \leq \exp(\sigma^2 t^2/2),$$

for all $t \in \mathbb{R}$. Gaussian random variables with variance σ^2 satisfy the above condition with equality, so a σ-sub-Gaussian random variable basically just has an mgf that is dominated by a Gaussian with variance σ.

It is straightforward to go through the above Chernoff bound to conclude that for a sub-Gaussian random variable we have the same two-sided exponential tail bound,

$$\mathbb{P}(|X - \mu| \geq u) \leq 2\exp(-u^2/(2\sigma^2)).$$

Suppose we have n i.i.d random variables σ sub-Gaussian RVs $X_1, X_2, \ldots, X_n$. Let $\widehat{\mu} = n^{-1}\sum_i X_i$. Then by independence we obtain that,

$$\begin{aligned}
\mathbb{E}[\exp(t(\widehat{\mu} - \mu))] &= \mathbb{E}[\exp(t/n \sum_{i=1}^{n}(X_i - \mu)] \\
&= \prod_{i=1}^{n} \mathbb{E}[\exp(t(X_i - \mu)/n)] \\
&\leq \exp(t^2\sigma^2/(2n)).
\end{aligned}$$

Alternatively, the average of n independent σ-sub Gaussian RVs is $\sigma/\sqrt{n}$-sub Gaussian. This yields the tail bound for the average of sub Gaussian RVs:

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq k\sigma/\sqrt{n}) \leq 2\exp(-k^2/2).$$

3.3 Bounded Random Variables - Hoeffding's bound

We claimed in the previous section that many classes of RVs are sub-Gaussian. In this section, we show this for an important special case: *bounded random variables*.

Example 1: Let us first consider a simple case, of Rademacher random variables, i.e. random variables that take the values $\{+1, -1\}$ equiprobably. In this case we can see that,

$$\begin{aligned}\mathbb{E}[\exp(tX)] &= \frac{1}{2}\left[\exp(t) + \exp(-t)\right] \\ &= \frac{1}{2}\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{k!} + \sum_{k=0}^{\infty}\frac{t^k}{k!}\right] \\ &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{t^{2k}}{2^k k!} \\ &= \exp(t^2/2).\end{aligned}$$

This shows that Rademacher random variables are 1-sub Gaussian.

Detour: Jensen's inequality: A function g is convex if

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

for all x, y and all $\alpha \in [0, 1]$. For example, $g(x) = x^2$ is convex. Jensen's inequality states that for a convex function $g : \mathbb{R} \mapsto \mathbb{R}$ we have that,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

If g is concave then the reverse inequality holds.

Proof: Let $\mu = \mathbb{E}[X]$ and let $L_\mu(x) = a + bx$ be the tangent line to the function g at μ, i.e. we have that $L_\mu(\mu) = g(\mu)$. By convexity we know that $g(x) \geq L_\mu(x)$ for every point x. Thus we have that,

$$\begin{aligned}\mathbb{E}[g(X)] &\geq \mathbb{E}[L_\mu(X)] = \mathbb{E}[a + bX] \\ &= a + b\mu = L_\mu(\mu) = g(\mu).\end{aligned}$$

Example 2: Bounded Random Variables. Let X be a random variable with zero mean and with support on some bounded interval $[a, b]$.

You should convince yourself that the zero mean assumption does not matter: you can always subtract the mean, i.e. define a new random variable $Y = X - \mathbb{E}[X]$ and use Y in the calculation below.

Let X' denote an *independent* copy of X then we have that,

$$\mathbb{E}_X[\exp(tX)] = \mathbb{E}_X[\exp(t(X - \mathbb{E}[X'])] \leq \mathbb{E}_{X,X'}[\exp(t(X - X')],$$

using Jensen's inequality, and the convexity of the function $g(x) = \exp(x)$.

Now, let ϵ be a Rademacher random variable. Then note that the distribution of $X - X'$ is identical to the distribution of $X' - X$ and more importantly of $\epsilon(X - X')$. So we obtain that,

$$\begin{aligned}\mathbb{E}_{X,X'}[\exp(t(X - X')] &= \mathbb{E}_{X,X'}[\mathbb{E}_{\epsilon}[\exp(t\epsilon(X - X')]] \\ &\leq \mathbb{E}_{X,X'}[\exp(t^2(X - X')^2/2],\end{aligned}$$

where we just use the result from Example 1, with (X, X') fixed by conditioning. Now $(X - X')$ using boundedness is at most $(b - a)$ so we obtain that,

$$\mathbb{E}_X[\exp(tX)] \leq \exp(t^2(b - a)^2/2),$$

which in turn shows that bounded random variables are $(b - a)$-sub Gaussian.

This in turn yields Hoeffding's bound. Suppose that, $X_1, \ldots, X_n$ are independent identically distribution *bounded* random variables, with $a \leq X_i \leq b$ for all i with mean 0. Then

$$\mathbb{P}(\overline{X}_n \geq \epsilon) = \mathbb{P}(e^{t\overline{X}_n} \geq e^{t\epsilon}) \leq e^{-t\epsilon} e^{nt^2(b-a)^2/2}.$$

Minimizing over t we get

$$\mathbb{P}\left(\overline{X}_n \geq t\right) \leq \exp\left(-\frac{nt^2}{2(b-a)^2}\right).$$

Repeating this in the other direction we get

$$\mathbb{P}\left(|\overline{X}_n| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right).$$

More generally, if X_i has mean μ then

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right).$$

This is a two-sided exponential tail inequality for the averages of bounded random variables. With some effort you can derive a slightly tighter bound on the MGF to obtain the stronger bound that:

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

3.4 A simple generalization

It is worth noting that none of the exponential tail inequalities we proved required the random variables to be identically distributed. More generally, suppose that we have $X_1, \ldots, X_n$

which are each $\sigma_1, \ldots, \sigma_n$ sub Gaussian. Then using just independence you can verify that their average $\widehat{\mu}$ is σ-sub Gaussian, where,

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

This in turn yields the exponential tail inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq \exp(-t^2/(2\sigma^2)).$$

Note that the random variables still need to be independent but no longer need to be identically distributed (i.e. they can for instance have different means and sub-Gaussian parameters).

Lecture Notes 3
36-705

1 Review: Bounded Random Variables - Hoeffding's bound

We claimed in the previous lecture that many classes of RVs are sub-Gaussian. In this section, we show this for an important special case: *bounded random variables*.

Example 1: Let us first consider a simple case, of Rademacher random variables, i.e. random variables that take the values $\{+1, -1\}$ equiprobably. In this case we can see that,

$$\begin{aligned}\mathbb{E}[\exp(tX)] &= \frac{1}{2}\left[\exp(t) + \exp(-t)\right] \\ &= \frac{1}{2}\left[\sum_{k=0}^{\infty}\frac{(-t)^k}{k!} + \sum_{k=0}^{\infty}\frac{t^k}{k!}\right] \\ &= \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{t^{2k}}{2^k k!} \\ &= \exp(t^2/2).\end{aligned}$$

This shows that Rademacher random variables are 1-sub Gaussian.

Detour: Jensen's inequality: Jensen's inequality states that for a convex function $g : \mathbb{R} \mapsto \mathbb{R}$ we have that,

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

If g is concave then the reverse inequality holds.

Proof: Let $\mu = \mathbb{E}[X]$ and let $L_\mu(x) = a + bx$ be the tangent line to the function g at μ, i.e. we have that $L_\mu(\mu) = g(\mu)$. By convexity we know that $g(x) \geq L_\mu(x)$ for every point x. Thus we have that,

$$\begin{aligned}\mathbb{E}[g(X)] &\geq \mathbb{E}[L_\mu(X)] = \mathbb{E}[a + bX] \\ &= a + b\mu = L_\mu(\mu) = g(\mu).\end{aligned}$$

Example 2: Bounded Random Variables. Let X be a random variable with zero mean and with support on some bounded interval $[a, b]$.

You should convince yourself that the zero mean assumption does not matter in general (you can always subtract the mean, i.e. define a new random variable $Y = X - \mathbb{E}[X]$ and use Y in the calculation below).

Let X' denote an *independent* copy of X then we have that,

$$\mathbb{E}_X[\exp(tX)] = \mathbb{E}_X[\exp(t(X - \mathbb{E}[X'])] \leq \mathbb{E}_{X,X'}[\exp(t(X - X')],$$

using Jensen's inequality, and the convexity of the function $g(x) = \exp(x)$.

Now, let ϵ be a Rademacher random variable. Then note that the distribution of $X - X'$ is identical to the distribution of $X' - X$ and more importantly of $\epsilon(X - X')$. So we obtain that,

$$\begin{aligned}\mathbb{E}_{X,X'}[\exp(t(X - X')] &= \mathbb{E}_{X,X'}[\mathbb{E}_\epsilon[\exp(t\epsilon(X - X')]] \\ &\leq \mathbb{E}_{X,X'}[\exp(t^2(X - X')^2/2],\end{aligned}$$

where we just use the result from Example 1, with (X, X') fixed by conditioning. Now $(X - X')$ using boundedness is at most $(b - a)$ so we obtain that,

$$\mathbb{E}_X[\exp(tX)] \leq \exp(t^2(b - a)^2/2),$$

which in turn shows that bounded random variables are $(b - a)$-sub Gaussian.

This in turn yields Hoeffding's bound. Suppose that, $X_1, \ldots, X_n$ are independent identically distribution *bounded* random variables, with $a \leq X_i \leq b$ for all i then,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right).$$

This is a two-sided exponential tail inequality for the averages of bounded random variables. With some effort you can derive a slightly tighter bound on the MGF to obtain the stronger bound that:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

1.1 A simple generalization

It is worth noting that none of the exponential tail inequalities we proved required the random variables to be identically distributed. More generally, suppose that we have $X_1, \ldots, X_n$ which are each $\sigma_1, \ldots, \sigma_n$ sub Gaussian. Then using just independence you can verify that their average $\widehat{\mu}$ is σ-sub Gaussian, where,

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

This in turn yields the exponential tail inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq \exp(-t^2/(2\sigma^2)).$$

Note that the random variables still need to be independent but no longer need to be identically distributed (i.e. they can for instance have different means and sub-Gaussian parameters).

2 Other interesting concentration inequalities

The rest of this lecture will be mostly a summary of other useful exponential tail bounds. We will not prove any of these in lecture, some of them follow similar lines of using Chernoff's method in clever ways. In particular, we will go through:

1. Bernstein's inequality: sharper concentration for bounded random variables
2. McDiarmid's inequality: Concentration of Lipschitz functions of bounded random variables
3. Levy's inequality/Tsirelson's inequality: Concentration of Lipschitz functions of Gaussian random variables
4. χ^2 tail bound

Finally, we will see an application of the χ^2 tail bound in proving the Johnson-Lindenstrauss lemma.

3 Bernstein's inequality

One nice thing about the Gaussian tail inequality was that it explicitly depended on the variance of the random variable X, i.e. the inequality guaranteed us that the deviation from the mean was at most $\sigma\sqrt{\log(2/\delta)/n}$ with probability at least $1-\delta$.

On the other hand Hoeffding's bound depended only on the bounds of the random variable but not explicitly on the variance of the RVs. The bound $b-a$, provides a (possibly loose) upper bound on the standard deviation. One might at least hope that if the random variables were bounded, and additionally had *small variance* we might be able to improve Hoeffding's bound.

This is indeed the case. Such inequalities are typically known as Bernstein inequalities. As a concrete example, suppose we had $X_1, \ldots, X_n$ which were i.i.d from a distribution with mean μ, bounded support $[a, b]$, with variance $\mathbb{E}[(X - \mu)^2] = \sigma^2$. Then,

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2(\sigma^2 + (b-a)t)}\right).$$

This inequality implies that, with probability at least $1 - \delta$,

$$|\widehat{\mu} - \mu| \leq 4\sigma\sqrt{\frac{\ln(2/\delta)}{n}} + \frac{4(b-a)\ln(2/\delta)}{n}.$$

Exercise: work through the above algebra.

Up to some small constants this is never worse than Hoeffding's bound, which just comes from using the worst-case upper bound of $\sigma \leq b - a$. When the RVs have small variance, i.e. σ is small, this bound can be much sharper than Hoeffding's bound. These are cases where one has a random variable that occasionally takes large values (so the bounds are not great) but has much smaller variance.

Intuitively, it captures more of the Chebyshev effect, i.e. that random variables with small variance should be tightly concentrated around their mean.

As an example, consider

$$\frac{1}{n}\sum_i I(|X_i| < a_n)$$

where $a_n \to 0$. This is the fraction of observations close to 0. The variance of $I(|X_i| < a_n)$ is about a_n. If $a_n \to 0$ quickly then the variance is very small. The distance between μ and $\hat{\mu}$ is of order $\sqrt{a_n/n}$ instead of $1/\sqrt{n}$.

4 McDiarmid's inequality

So far we have focused on the concentration of averages. A natural question is whether other functions of i.i.d. random variables also show exponential concentration. It turns out that many other functions do concentrate sharply, and roughly the main property of the function that we need is that if we change the value of one random variable the function does not change dramatically.

Formally, we have i.i.d. RVs $X_1, \ldots, X_n$, where each $X_i \in \mathbb{R}$. We have a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, that satisfies the property that:

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n)| \leq L_k,$$

for every $x, x' \in \mathbb{R}^n$, i.e. the function changes by at most L_k if its k-th co-ordinate is changed. This is known as the bounded difference condition.

If the random variables $X_1, \ldots, X_n$ are i.i.d then for all $t \geq 0$

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right).$$

Example 1: A simple example of this inequality in action is to see that it directly implies the Hoeffding bound. In this case the function of interest is the average:

$$f(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i,$$

and since the random variables are bounded we have that each $L_k \leq (b-a)/n$. This in turn directly yields Hoeffding's bound (with slightly better constants).

Example 2: A perhaps more interesting example is that of U-statistics. A U-statistic is defined by a kernel, which is just a function of two random variables, i.e. $g : \mathbb{R}^2 \mapsto \mathbb{R}$. The U-statistic is then given as:

$$U(X_1, \ldots, X_n) := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k).$$

There are many examples of U-statistics, for instance:

1. **Variance:** The usual estimator of the sample variance:

$$\widehat{\sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \widehat{\mu})^2,$$

   is the U-statistic that arises from taking $g(X_j, X_k) = \frac{1}{2}(X_i - X_j)^2$.

2. **Mean absolute deviation:** If we take $g(X_j, X_k) = |X_j - X_k|$, this leads to a U-statistic that is an unbiased estimator of the mean absolute deviation $\mathbb{E}|X_1 - X_2|$.

For bounded U-statistics, i.e. if $g(X_i, X_j) \leq b$, we can apply McDiarmid's inequality to obtain a concentration bound. Note that since each random variable X_i participates in $(n-1)$ terms we have that,

$$|U(X_1, \ldots, X_n) - U(X_1, \ldots, X_i', \ldots, X_n)| \leq \frac{1}{\binom{n}{2}}(n-1)(2b) = \frac{4b}{n}.$$

So that McDiarmid's inequality tells us that,

$$\mathbb{P}(|U(X_1, \ldots, X_n) - \mathbb{E}[U(X_1, \ldots, X_n)]| \geq t) \leq 2 \exp(-nt^2/(8b^2)).$$

5 Levy's inequality

There is a similar concentration inequality that applies to functions of Gaussian random variables that are sufficiently smooth. In this case, the assumption is quite different. We assume that:

$$|f(X_1, \ldots, X_n) - f(Y_1, \ldots, Y_n)| \leq L \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2},$$

for all $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}$.

For such functions we have that if $X_1, \ldots, X_n \sim N(0,1)$ then,

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

6 χ^2 tail bounds

A χ^2 random variable with n degrees of freedom, denoted by $Y \sim \chi^2_n$, is a RV that is a sum of n i.i.d. standard Gaussian RVs, i.e. $Y = \sum_{i=1}^{n} X_i^2$ where each $X_i \sim N(0,1)$. Suppose that $Z_1, \ldots, Z_n \sim N(0,1)$, then the expected value $\mathbb{E}[Z_i^2] = 1$, and we have the χ^2 tail bound:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| \geq t\right) \leq 2\exp(-nt^2/8) \quad \text{for all} \quad t \in (0,1).$$

You will derive this in your HW using the Chernoff method. Analogous to the class of sub-Gaussian RVs, χ^2 random variables belong to a class of what are known as *sub-exponential* random variables.

Detour: The union bound. This is also known as Boole's inequality. It says that if we have events $A_1, \ldots, A_n$ then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

In particular, if we consider a case when each event A_i is a failure of some type, then the above inequality says that the probability that even a single failure occurs is at most the sum of the probabilities of each failure.

7 The Johnson-Lindenstrauss Lemma

One very nice application of χ^2 tail bounds is in the analysis of what are known as "random projections". Suppose we have a data set $X_1, \ldots, X_n \in \mathbb{R}^d$ where d is quite large. Storing

such a dataset might be expensive and as a result we often resort to "sketching" or "random projection" where the goal is to create a map $F : \mathbb{R}^d \mapsto \mathbb{R}^m$, with $m \ll d$. We then instead store the mapped dataset $\{F(X_1), \ldots, F(X_n)\}$. The challenge is to design this map F in a way that preserves essential features of the original dataset. In particular, we would like that for every pair (X_i, X_j) we have that,

$$(1-\epsilon)\|X_i - X_j\|_2^2 \leq \|F(X_i) - F(X_j)\|_2^2 \leq (1+\epsilon)\|X_i - X_j\|_2^2,$$

i.e. the map preserves all the pair-wise distances up to a $(1 \pm \epsilon)$ factor. Of course, if m is large we might expect this is not too difficult.

The Johnson-Lindenstrauss lemma is quite stunning: it says that a simple randomized construction will produce such a map with probability at least $1 - \delta$ provided that,

$$m \geq \frac{16 \log(n/\delta)}{\epsilon^2}.$$

Notice that this is completely independent of the original dimension d and depends on logarithmically on the number of points n. This map can result in huge savings in storage cost while still essentially preserving all the pairwise distances.

The map itself is quite simple: we construct a matrix $Z \in \mathbb{R}^{m \times d}$, where each entry of Z is i.i.d $N(0,1)$. We then define the map as:

$$F(X_i) = \frac{1}{\sqrt{m}} Z X_i.$$

Now let us fix a pair (X_j, X_k) and consider,

$$\begin{aligned}\frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} &= \left\| \frac{Z(X_j - X_k)}{\sqrt{m}\|X_j - X_k\|_2} \right\|_2^2 = \frac{1}{m} \sum_{i=1}^{m} \left\langle Z_i, \frac{X_j - X_k}{||X_j - X_k||^2} \right\rangle^2 \\ &= \frac{1}{m} \sum_i \langle Z_i, a \rangle^2 = \frac{1}{m} \sum_i T_i^2\end{aligned}$$

where

$$a = \frac{X_j - X_k}{||X_j - X_k||^2}.$$

In general, the distribution of $\sum_{j=1}^d a_j Z_{ij}$ is Gaussian with mean 0 and variance $\sum_{j=1}^d a_j^2$. In our case, $\sum_j a_j^2 = 1$. So each term T_i is an independent χ_m^2 random variable. (The data X_i are being treated as fixed; the randomness is from the $Z_{ij}'s$.) Now applying the χ^2 tail bound, we obtain that,

$$\mathbb{P}\left(\left| \frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} - 1 \right| \geq \epsilon \right) \leq 2 \exp(-m\epsilon^2/8).$$

Thus for the fixed pairs (X_i, X_j), the probability that our map fails to preserve the distance is exponentially small, i.e. is at most $2\exp(-m\epsilon^2/8)$. Now, to find the probability that our map fails to preserve *any* of our $\binom{n}{2}$ pairwise distances we simply apply the union bound to conclude that, the probability of any failure is at most:

$$\mathbb{P}(\text{failure}) \leq 2\binom{n}{2}\exp(-m\epsilon^2/8).$$

Now, it is straightforward to verify that if

$$m \geq \frac{16\log(n/\delta)}{\epsilon^2},$$

then this probability is at most δ as desired. An important point to note is that the *exponential concentration* is what leads to such a small value for m (i.e. it only needs to grow logarithmically with the sample size).

Lecture Notes 4
36-705

In today's lecture we discuss the convergence of random variables. At a high-level, our first few lectures focused on non-asymptotic properties of averages i.e. the tail bounds we derived applied for any fixed sample size n. For the next few lectures we focus on asymptotic properties, i.e. we ask the question: what happens to the average of n i.i.d. random variables as $n \to \infty$.

Roughly, from a theoretical perspective the idea is that many expressions will considerably simplify in the asymptotic regime. Rather than have many different tail bounds, we will derive simple "universal results" that hold under extremely weak conditions. From a slightly more practical perspective, asymptotic theory is often useful to obtain approximate confidence intervals.

1 Reminder: convergence of sequences

When we think of convergence of deterministic real numbers the corresponding notions are classical.

Formally, we say that a sequence of real numbers $a_1, a_2, \ldots$ converges to a fixed real number a if, for every positive number ϵ, there exists a natural number $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, $|a_n - a| < \epsilon$. We call a the limit of the sequence and write $\lim_{n\to\infty} a_n = a$.

Our focus today will in trying to develop analogues of this notion that apply to sequences of random variables. We will first give some definitions and then try to circle back to relate the definitions and discuss some examples.

Throughout, we will focus on the setting where we have a sequence of random variables $X_1, \ldots, X_n$ and another random variable X, and would like to define what is means for the sequence to converge to X. In each case, to simplify things you should also think about the case when X is deterministic, i.e. when $X = c$ with probability 1 (for some constant c).

Importantly, we will *not assume that the RVs $X_1, \ldots, X_n$ are independent.*

2 Almost sure convergence

We will not use almost sure convergence in this course so you should feel free to ignore this section. A natural analogue of the usual convergence would be to hope that,

$$\lim_{n\to\infty} X_n = X.$$

We write $X_n \overset{a.s.}{\to} X$. These are both however random variables so one has to at least specify on what event we are hoping for this statement to be true.

The correct analogue turns out to be to require:

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

There are measure theoretic subtleties to be aware of here. In particular, the sample space inside the probability statement here is set of infinite sequences and it requires some machinery to be precise here.

There are other equivalent (this is somewhat difficult to see) ways to define almost sure convergence. Equivalently, we say that X_n converges almost surely to X if we let Ω be a set of probability mass 1, i.e. $\mathbb{P}(\Omega) = 1$, and for every $\omega \subseteq \Omega$, and for every $\epsilon > 0$, we have that there is some $n \geq N(\omega, \epsilon)$ such that:

$$|X_n(\omega) - X(\omega)| \leq \epsilon.$$

Roughly, the way to think about this type of convergence is to imagine that there is some set of exceptional events on which the random variables can disagree, but these exceptional events have probability 0 as $n \to \infty$. Barring, these exceptional events the sequence converges just like sequences of real numbers do. The exceptional events is where the "almost" in almost sure arises.

3 Convergence in probability

A sequence of random variables $X_1, \ldots, X_n$ converges in probability to a random variable X if for every $\epsilon > 0$ we have that,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

We write $X_n \overset{P}{\to} X$. To build intuition it is perhaps useful to consider the case when X is deterministic, i.e. $X = c$ with probability 1. Then convergence in probability is saying that as n gets large the distribution of X_n gets more peaked around the value c. Convergence in probability can be viewed as a statement about the convergence of probabilities, while almost sure convergence is a convergence of the values of a sequence of random variables.

We will not prove this statement but convergence in probability is implied by almost sure convergence. The notes contain a counterexample to the reverse implication which we may or may not cover in the lecture.

Weak Law of Large Numbers Suppose that $Y_1, Y_2, \ldots$ are i.i.d. with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}(Y_i) = \sigma^2 < \infty$. Define,

$$X_n = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

The WLLN says that the sequence $X_1, X_2, \ldots$ converges in probability to μ. That is $X_n \xrightarrow{P} \mu$.

Proof: The proof is simply an application of Chebyshev's inequality. We note that by Chebyshev's inequality:

$$\mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

This in turn implies that,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) = 0,$$

as desired.

Notes:

1. Strictly speaking the WLLN is true even without the assumption of finite variance, as long as the first absolute moment is finite. This proof is a bit more difficult.
2. There is a statement that says that under similar assumptions the average converges almost surely to the expectation. This is known as the strong law of large numbers. This is actually quite a bit more difficult to prove.

Consistency: Convergence in probability will frequently recur in this course. Usually we will construct an estimator $\widehat{\theta}_n$ for some quantity θ^*. We will then say that the estimator is *consistent* if the sequence of RVs $\widehat{\theta}_n$ converges in probability to θ^*.

The WLLN/Chebyshev can already be used to prove some rudimentary consistency guarantees. For instance, if we consider the sample variance:

$$\widehat{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2,$$

then by Chebyshev's inequality we obtain,

$$\mathbb{P}(|\widehat{S}_n - \sigma^2| \geq \epsilon) \leq \frac{\text{Var}(\widehat{S}_n)}{\epsilon^2},$$

so a sufficient condition for consistency is that $\text{Var}(\widehat{S}_n) \to 0$ as $n \to \infty$.

Convergence in probability does not imply almost sure convergence: Suppose we have a sample space $S = [0, 1]$, with the uniform distribution, we draw $s \sim U[0, 1]$ and define $X(s) = s$. We define the sequence as:

$$X_1(s) = s + \mathbb{I}_{[0,1]}(s), \quad X_2(s) = s + \mathbb{I}_{[0,1/2]}(s), \quad X_3(s) = s + \mathbb{I}_{[1/2,1]}(s)$$
$$X_4(s) = s + \mathbb{I}_{[0,1/3]}(s), \quad X_5(s) = s + \mathbb{I}_{[1/3,2/3]}(s), \quad X_6(s) = s + \mathbb{I}_{[2/3,1]}(s).$$

Now one can check that this sequence converges in probability but not almost surely. Roughly, the "1 + s" spike becomes less frequent down the sequence (allowing convergence in probability) but the limit is not well defined. For any s, $X_n(s)$ alternates between s and $1 + s$.

4 Convergence in quadratic mean

An often useful way to show convergence in probability is to show something stronger known as convergence in quadratic mean. We say that a sequence converges to X in quadratic mean if:

$$\mathbb{E}(X_n - X)^2 \to 0,$$

as $n \to \infty$. We write $X_n \overset{qm}{\to} X$.

5 Convergence in distribution

The other commonly encountered mode of convergence is convergence in distribution. We say that a sequence converges to X in distribution if:

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for all points t where the CDF F_X is continuous. We will see why the exception matters in a little while but for now it is worth noting that convergence in distribution is the weakest form of convergence. We write $X_n \rightsquigarrow X$.

For instance, a sequence of i.i.d. $N(0, 1)$ RVs converge in distribution to an independent $N(0, 1)$ RV, even though the values of the random variables are not close in any meaningful sense (their distributions are however, identical). A famous result that we will disucss in the next lecture is the central limit theorem. The central limit theorem says that an average of i.i.d. random variables (appropriately normalized) converges in distribution to a $N(0, 1)$ random variable.

The picture to keep in mind to understand the relationships is the following one:

$$\begin{array}{ccccc} & & \text{q.m.} & & \\ & & \downarrow & & \\ \text{a.s.} & \rightarrow & \text{prob} & \rightarrow & \text{distribution} \end{array}$$

We will re-visit this in the next lecture and perhaps try to prove some of the implications (or disprove some of the non-implications).

6 Examples

Example 1: Suppose we consider a sequence $X_n = N(0, 1/n)$. Intuitively, it seems like this sequence converges to 0. Let us first consider what happens in distribution.

Let X be such that $P(X = 0) = 1$. The CDF is $F_X(x) = 0$, for $x < 0$ and $F_X(x) = 1$ for $x \geq 0$. Note that $X_n \stackrel{d}{=} Z/n$ where $Z \sim N(0, 1)$. So,

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x) = \mathbb{P}(Z \leq \sqrt{n}x),$$

where $Z \sim N(0, 1)$. If $x > 0$ this tends to 1, and if $x < 0$ this tends to 0. Interestingly, at $x = 0$, $F_{X_n}(x) = 1/2$, and does not converge to $F_X(0) = 1$. Remember, however, that we had an exception at points of discontinuity. So $X_n \rightsquigarrow X$.

Example 2: Let us consider the same example and consider convergence in probability.

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \frac{\mathbb{E}[X_n^2]}{\epsilon^2} = \frac{1}{n\epsilon^2} \rightarrow 0,$$

so the sequence converges to 0 in probability.

Example 3: Suppose $X_1, \ldots \sim U[0, 1]$. Let us define $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Now, we verify two things:

1. $X_{(n)}$ converges in probability to 1. To see this observe that,

$$\begin{aligned} \mathbb{P}(|X_{(n)} - 1| \geq \epsilon) &= \mathbb{P}(X_{(n)} \leq 1 - \epsilon) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leq 1 - \epsilon) = (1 - \epsilon)^n \\ &\rightarrow 0. \end{aligned}$$

2. The random variable $n(1 - X_{(n)})$ converges in distribution to an Exp(1) RV. To see this we compute:

$$\begin{aligned} F_{X_{(n)}}(t) &= \mathbb{P}(n(1 - X_{(n)}) \leq t) = 1 - \mathbb{P}(X_{(n)} \leq 1 - t/n) \\ &= 1 - (1 - t/n)^n \rightarrow 1 - \exp(-t) = F_X(t). \end{aligned}$$

Lecture Notes 5

Today we will start off by deriving some of the implications between the different modes of convergence. Then we will prove the CLT.

1 Quadratic mean $\implies$ convergence in probability

Suppose that $X_1, \ldots, X_n$ converges in quadratic mean to X, then fix an $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{P}(|X_n - X|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X_n - X)^2}{\epsilon^2} \to 0,$$

showing convergence in probability.

At a high-level the convergence in qm requirement penalizes X_n for having large deviations from X by both how frequent the deviation is but also by the *magnitude of the deviation*. On the other hand convergence in probability only penalizes you for how frequent the deviation is and hence is a weaker notion of convergence.

Counterexample to reverse: Suppose we take $U \sim U[0,1]$ and define $X_n = \sqrt{n}\mathbb{I}_{[0,1/n]}(U)$, then X_n converges in probability to 0 but does not converge in quadratic mean to 0.

To see this:

$$\mathbb{P}(|X_n| \geq \epsilon) = \mathbb{P}(\sqrt{n}\mathbb{I}_{[0,1/n]}(U) \geq \epsilon) = \mathbb{P}(U \in [0, 1/n]) = \frac{1}{n} \to 0.$$

On the other hand,

$$\mathbb{E}(X_n - X)^2 = \mathbb{E}X_n^2 = n\mathbb{P}(U \in [0, 1/n]) = 1.$$

Observe that most of the time the RV X_n takes the value 0, but when it does not it takes a huge value.

1.1 Convergence in probability $\implies$ convergence in distribution

This one is a little bit involved but perhaps also useful to know. The idea roughly is to trap the CDF of X_n by the CDF of X with an interval whose length converges to 0.

Suppose that $X_n \rightsquigarrow X$. We fix a point x where the CDF $F_X(x)$ is continuous. Choose an arbitrary $\epsilon > 0$. We have that,

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X \geq x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon) \\ &= F_X(x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon). \end{aligned}$$

Now,

$$\begin{aligned} F_X(x-\epsilon) &= \mathbb{P}(X \leq x-\epsilon) = \mathbb{P}(X \leq x-\epsilon, X_n \leq x) + \mathbb{P}(X \leq x-\epsilon, X_n \geq x) \\ &\leq F_{X_n}(x) + \mathbb{P}(|X_n - X| \geq \epsilon). \end{aligned}$$

Putting these two together we have,

$$F_X(x-\epsilon) - \mathbb{P}(|X_n - X| \geq \epsilon) \leq F_{X_n}(x) \leq F_X(x+\epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon).$$

Intuitively, now as n gets large the two probabilities converge to 0, and since ϵ was chosen arbitrarily we can let $\epsilon \to 0$ and use the continuity of $F_X(x)$ at x to conclude that $F_{X_n}(x) \to F_X(x)$.

Slightly more rigorously, we cannot assume that the limit of $F_{X_n}(x)$ exists so we instead need to use lim infs and lim sups (do not worry about this if you have not seen it before). Formally, we would take the lim sup of the first half to obtain that,

$$\limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon),$$

and similarly that,

$$\liminf_{n\to\infty} F_{X_n}(x) \geq F_X(x-\epsilon),$$

and conclude that,

$$F_X(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon).$$

Now since $\epsilon > 0$ was arbitrary, we can take the limit as $\epsilon \to 0$ and use continuity to conclude the desired convergence in distribution.

Counterexample to reverse: This is easy since two random variables having the same distribution does not in any sense mean that they are close. For example, let $X, X_1, X_2, \sim N(0,1)$. They all have the same cdf so $X_n \rightsquigarrow X$. But $P(|X_n - X| > \epsilon)$ does not go to 0.

An important exception: An important exception is that when X is deterministic then convergence in distribution implies convergence in probability. Suppose that $P(X = c) = 1$. Fix $\epsilon > 0$. Then

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n > \epsilon + c) + \mathbb{P}(X_n < c - \epsilon) \\ &= F_{X_n}(c-\epsilon) + 1 - F_{X_n}(c+\epsilon) \\ &\to F_X(c-\epsilon) + 1 - F_X(c+\epsilon) = 0. \end{aligned}$$

using convergence in distribution and the fact that at both $c+\epsilon$, and $c-\epsilon$, the distribution function F_X is continuous. So $X_n \rightsquigarrow c$ implies that $X_n \xrightarrow{P} c$.

2 Other things that are very useful to know

1. **Continuous mapping theorem.** If a sequence $X_1, \ldots, X_n$ converges in probability to X then for any continuous function h, $h(X_1), \ldots, h(X_n)$ converges in probability to $h(X)$. The same is true for convergence in distribution.

2. A consequence of the continuous mapping theorem. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$. Similarly, $X_n Y_n \xrightarrow{P} XY$.

3. **Slutsky's theorem.** If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ we **cannot** conclude that the sum converges. The one exception is known as Slutsky's theorem. It says that if Y_n converges in distribution to a constant c, and X converges in distribution to X: then $X_n + Y_n$ converges in distribution to $X + c$ and $X_n Y_n$ converges in distribution to cX.

4. **Convergence of moments is not implied by convergence in probability.** Convergence in probability is actually quite weak as a form of convergence. We have seen previously that it does not imply quadratic mean convergence. Now we will see that it does not even imply something much simpler.

   If we have X_n converges in probability to some constant c, then it is not the case that $\mathbb{E}[X_n]$ converges to c. Here is an example of this non-convergence. Let X_n be 0 with probability $1-1/n$ and n^2 with probability $1/n$. Then X_n converges to 0 in probability, but $\mathbb{E}[X_n] = n \to \infty$.

   This is a manifestation of the same phenomena as we saw in the counterexample to qm convergence. On the events when $|X_n| \geq \epsilon$ it has a huge value and this affects the moments but does not affect the convergence in probability.

3 The Central Limit Theorem (CLT)

We will now state and prove a form of the central limit theorem, which is one of the most famous and important examples of convergence in distribution. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with mean μ and variance σ^2.

Theorem 1 *Assume that the mgf $\mathbb{E}[\exp(tX_i)]$ is finite for t in a neighborhood around zero. Let $\overline{X}_n = n^{-1}\sum_i X_i$. Let*

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

then Z_n converges in distribution to $Z \sim N(0,1)$, that is $Z_n \rightsquigarrow Z$. Hence, as $n \to \infty$,

$$\mathbb{P}(Z_n \leq t) \to \Phi(t)$$

for all t, where

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-s^2/s} ds.$$

The central limit theorem is incredibly general. It does not matter what the distribution of X_i is, the average S_n converges in distribution to a Gaussian (under fairly mild assumptions). The most general version of the CLT does not require any assumption about the mgf. It just requires that the mean and variance are finite. The interpretation of the CLT is that $Z_n \approx N(0,1)$. In other words,

$$\overline{X}_n \approx N(\mu, \sigma^2/n).$$

It can be shown that

$$\sup_t |\mathbb{P}(Z_n \leq t) - \Phi(t)| \leq \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}$$

where $\mu_3 = \mathbb{E}[|X_i - \mu|^3]$.

We should try to understand why the CLT might be useful. Roughly, the CLT allows to make *approximate* probability statements about averages using corresponding statements about standard normals. Here is an example that we will discuss in detail later: confidence intervals.

Suppose for now that we are averaging iid random variables with known variance σ (and unknown mean μ). Typically one would also estimate the variance but this will not change much. We would like to construct a *confidence interval* for the unknown mean. We specify $\alpha \in (0,1)$ and we find a random set C such that

$$\mathbb{P}(\mu \in C) \geq 1 - \alpha.$$

We might take

$$C = [\widehat{\mu} - t, \widehat{\mu} + t]$$

where $\widehat{\mu} = \overline{X}_n$. Then

$$\mathbb{P}(\mu \in [\widehat{\mu} - t, \widehat{\mu} + t]) = \mathbb{P}(|\widehat{\mu} - \mu| \leq t).$$

So we would like to choose t to make this probability equal to $1 - \alpha$. Now

$$\mathbb{P}(|\widehat{\mu} - \mu| \leq t) = \mathbb{P}\left(\frac{\sqrt{n}|\widehat{\mu} - \mu|}{\sigma} \leq \frac{\sqrt{n}t}{\sigma} \right) \approx \mathbb{P}(|Z| \leq t)$$

where $Z \sim N(0,1)$. In the last step we used the CLT. Let Φ denote the cdf of Z and define

$$z_\alpha = \Phi^{-1}(1 - \alpha).$$

Note that

$$P(Z > z_{\alpha/2}) = P(Z < -z_{\alpha/2}) = \frac{\alpha}{2}$$

so that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. So we want to set

$$\frac{\sqrt{n}t}{\sigma} = z_{\alpha/2}$$

that is,

$$t = \frac{\sigma z_{\alpha/2}}{\sqrt{n}}.$$

To summarize: if we define

$$C = \left[\widehat{\mu} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \widehat{\mu} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right],$$

then

$$\mathbb{P}(\mu \in C) \to 1 - \alpha$$

as $n \to \infty$. The convergence is due to the CLT.

3.1 Preliminaries

First we note that

$$\mathbb{E}[Z_n] = 0$$

and

$$\text{Var}[Z_n] = 1.$$

Also note that if $X_1, \ldots, X_n \sim N(0,1)$ then Z_n is exactly $N(0,1)$.

Calculus with mgfs: We need a few simple facts about mgfs that we will quickly prove.

Fact 1: If X and Y are independent with mgfs M_X and M_Y then $Z = X + Y$ has mgf $M_Z(t) = M_X(t)M_Y(t)$.

Proof: We note that,

$$M_Z(t) = \mathbb{E}[\exp(t(X+Y)] = \mathbb{E}[\exp(tX)]\mathbb{E}[\exp(tY)],$$

using independence.

Fact 2: If X has mgf M_X then $Y = a + bX$ has mgf, $M_Y(t) = \exp(at)M_X(bt)$.

Proof: We just use the definition,

$$M_Y(t) = \mathbb{E}[\exp(at + btX)] = \exp(at)\mathbb{E}[\exp(btX)].$$

Fact 3: We will not prove this one (strictly speaking one needs to invoke the dominated convergence theorem) but it should be familiar to you. The derivative of the mgf at 0 gives us moments, i.e.

$$M_X^{(r)}(0) = \mathbb{E}[X^r].$$

Fact 4: The most important result that we also will not prove is that we can show convergence in distribution by showing convergence of the mgfs. Let $X_1, \ldots, X_n$ be a sequence of random variables with mgfs $M_{X_1}, \ldots, M_{X_n}$. Let X be a random variable with mfg M_X. If for all t in an open interval around 0 we have that, $M_{X_n}(t) \to M_X(t)$, then X_n converges in distribution to X.

Fact 5: If $Z \sim N(0,1)$ then $M_Z(t) = e^{t^2/2}$.

3.2 Proof of the CLT

Note that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_i A_i$$

where

$$A_i = \frac{X_i - \mu}{\sigma}.$$

Let $M(t)$ be the mgf for A_i. Since A_i has mean 0 and variance 1, we have that $M(0) = 1$, $M'(0) = 0$ and $M''(0) = 1$. Now

$$M_{Z_n}(t) = \mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{\frac{t}{\sqrt{n}} \sum_i A_i}] = \prod_i \mathbb{E}[e^{\frac{t}{\sqrt{n}} A_i}] = M(t/\sqrt{n})^n.$$

Expanding M:

$$M(t/\sqrt{n}) \approx M(0) + \frac{t}{\sqrt{n}} M'(0) + \frac{t^2}{2n} M''(0) = 1 + \frac{t^2}{2n}$$

and so

$$M(t/\sqrt{n})^n \approx \left(1 + \frac{t^2}{2n}\right)^n \to e^{t^2/2}$$

which is the mgf of a N(0,1). Here we used the fact that,

$$\lim_{n \to \infty} (1 + x/n)^n \to \exp(x).$$

Lecture Notes 6
36-705

1 Stochastic Order Notation

The classical order notation should be familiar to you already.

1. We say that a sequence $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, $a_n = o(b_n)$ if $a_n/b_n = o(1)$.
2. We say that a sequence $a_n = O(1)$ if the sequence is eventually bounded, i.e. for all n large, $|a_n| \leq C$ for some constant $C \geq 0$. Similarly, $a_n = O(b_n)$ if $a_n/b_n = O(1)$.
3. If $a_n = O(b_n)$ and $b_n = O(a_n)$ then we use either $a_n = \Theta(b_n)$ or $a_n \asymp b_n$. Usually in Stats we avoid the Θ notation (which is more common in CS) because we usually use Θ for the parameter space.

When we are dealing with random variables we use stochastic order notation.

1. We say that $X_n = o_p(1)$ if for every $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}(|X_n| \geq \epsilon) \to 0,$$

   i.e. X_n converges to zero in probability.

2. We say that $X_n = O_p(1)$ if for every $\epsilon > 0$ there is a finite $C(\epsilon) > 0$ such that, for all n large enough:

$$\mathbb{P}(|X_n| \geq C(\epsilon)) \leq \epsilon.$$

The typical use case: suppose we have $X_1, \ldots, X_n$ which are i.i.d. and have finite variance, and we define:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. $\widehat{\mu} - \mu = o_p(1)$ (WLLN)
2. $\widehat{\mu} - \mu = O_p(1/\sqrt{n})$ (CLT)

As with the classical order notation, we can do some simple "calculus" with stochastic order notation and observe that for instance: $o_p(1) + O_p(1) = O_p(1)$, $o_p(1)O_p(1) = o_p(1)$ and so on.

2 Only independence but not identically distributed

Suppose that $X_1, X_2, \ldots$ are independent but not identically distributed. The CLT goes through almost exactly as stated, however, we need conditions to ensure that one or a small number of random variables do not dominate the sum. There are many such results but the most classical is called the Lyapunov CLT. I will state something that is slightly weaker than the actual result. Lyapunov is one of the fathers of the theory of dynamical systems and a student of Chebyshev's. As is the case with Chebyshev, there are several foundational concepts that are named for him (the CLT is only one).

Let $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = \mathrm{Var}[X_i]$. Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Lyapunov CLT: Suppose $X_1, \ldots, X_n$ are independent but not necessarily identically distributed. Let $\mu_i = \mathbb{E}[X_i]$ and let $\sigma_i = \mathrm{Var}(X_i)$. Then if we satisfy the Lyapunov condition:

$$\lim_{n\to\infty} \frac{1}{s_n^3} \sum_{i=1}^{n} \mathbb{E}|X_i - \mu|^3 = 0,$$

then

$$\frac{1}{s_n} \sum_{i=1}^{n} [X_i - \mu_i] \rightsquigarrow N(0,1).$$

In the i.i.d case we just have $\mu_i = \mu$ and $s_n = \sqrt{n}\sigma$ so we get back our usual CLT.

It is worthwhile trying to understand the Lyapunov condition, and when it might be violated. In particular, consider the extreme case, when all the random variables are deterministic, except X_1 which has mean μ_1 and variance $\sigma_1^2 > 0$. Then $s_n^3 = \sigma_1^3$ and the third absolute moment $\mathbb{E}|X_1 - \mu|^3 > 0$ so that the Lyapunov condition fails. Roughly, what can happen in the non-identically distributed case is that only one random variable can dominate the sum in which case you are not really averaging many things so you do not have a CLT.

On the other hand in a more typical case, one might have that the third absolute moments are bounded by some constant $C > 0$ say and the variance of any particular random variable is not too small. In this case,

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2 \geq n\sigma_{\min}^2,$$

and

$$\sum_{i=1}^{n} \mathbb{E}|X_i - \mu|^3 \leq Cn.$$

In this case, we will have that the Lyapunov ratio $\leq \frac{C}{\sqrt{n}\sigma_{\min}^3} \to 0$ so that the condition is indeed satisfied.

3 Multivariate CLT

The next important extension is the multivariate CLT.

Multivariate CLT: If $X_1, \ldots, X_n$ are i.i.d with mean $\mu \in \mathbb{R}^d$, and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ (with finite entries) then,

$$\sqrt{n}(\widehat{\mu} - \mu) \rightsquigarrow N(0, \Sigma).$$

Notes:

1. You might wonder what convergence in distribution means for random vectors. A random vector still has a CDF, typically we define this as:

   $$F_X(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots X_d \leq x_d),$$

   so we can still define convergence in distribution via pointwise convergence of the CDF. In order to define points of continuity it turns out that the correct definition is that a point is a point of continuity of the CDF if the boundary of the rectangle whose upper right corner is $(x_1, \ldots, x_d)$ has probability 0.

2. Although d can be larger than 1, it is taken to be fixed as $n \to \infty$. Central limit theorems, when d is allowed to grow, i.e. high-dimensional CLTs are quite complicated and are an active topic of research.

3. The proof of this result follows directly from the proof of the univariate CLT and a powerful result in asymptotic statistics known as the Cramer-Wold device. The Cramer-Wold device roughly asserts that if $a^T X_n \rightsquigarrow a^T X$ for all vectors $a \in \mathbb{R}^d$ then $X_n \rightsquigarrow X$.

4 CLT with estimated variance

We saw that in our typical use case of the CLT (constructing confidence intervals) we needed to know the variance σ. In practice, we most often do not know this. However, we can estimate this quantity in the usual way,

$$\widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2.$$

It turns out that we can replace the standard deviation in the CLT by $\widehat{\sigma}$ and still have the same convergence in distribution, i.e.

$$\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\widehat{\sigma}_n} \rightsquigarrow N(0,1).$$

The proof follows from a sequence of applications of Slutsky's theorem and the continuous mapping theorem.

Proof: First observe that if we can show that $\frac{\sigma}{\widehat{\sigma}_n} \rightsquigarrow 1$, then an application of Slutsky's theorem and the CLT gives us the desired result.

Since square-root is a continuous map, by the continuous mapping theorem, it suffices to show that $\frac{\sigma^2}{\widehat{\sigma}_n^2} \rightsquigarrow 1$. We will instead show the stronger statement that,

$$\widehat{\sigma}_n^2 \xrightarrow{p} \sigma^2,$$

which implies the desired statement via the continuous mapping theorem. Note that,

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2$$

and $n/(n-1) \to 1$ so it suffices to show that

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2 \xrightarrow{P} \sigma^2.$$

Now

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 \xrightarrow{p} \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

using the WLLN. This concludes the proof.

5 Rate of convergence in CLT - Berry Esseen

While the central limit theorem is an asymptotic result (i.e. a statement about $n \to \infty$) it turns out that under fairly general conditions we can say how close to a standard normal the average is, in distribution, for finite values n. Such results are known as Berry Esseen bounds. Roughly, they are proved by carefully tracking the remainder terms in our Taylor series proof but we will not do this here.

Berry-Esseen: Suppose that $X_1, \ldots, X_n \sim P$. Let $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \mathbb{E}[(X_1 - \mu)^2]$, and $\mu_3 = \mathbb{E}[|X_1 - \mu|^3]$. Let

$$F_n(x) = \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \leq x\right),$$

denote the CDF of the normalized sample average. If $\mu_3 < \infty$ then,

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{9\mu_3}{\sigma^3\sqrt{n}}.$$

This bound is roughly saying that if μ_3/σ^3 is small then the convergence to normality in distribution happens quite fast.

6 The Delta Method

A natural question that arises frequently is the following: suppose we have a sequence of random variables X_n that converges in distribution to a Gaussian distribution then can we characterize the limiting distribution of $g(X_n)$ where g is a smooth function?

We could work this out by using the continuous mapping theorem (indeed, that is at the heart of the proof we are about to give).

Delta Method: Suppose that,

$$\frac{\sqrt{n}(X_n - \mu)}{\sigma} \rightsquigarrow N(0, 1),$$

and that g is a continuously differentiable function such that $g'(\mu) \neq 0$. Then,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{\sigma} \rightsquigarrow N(0, [g'(\mu)]^2).$$

Proof: The basic idea is simply to use Taylor's approximation. We know that,

$$g(X_n) \approx g(\mu) + g'(\mu)(X_n - \mu),$$

so that,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{\sigma} \approx g'(\mu)\frac{\sqrt{n}(X_n - \mu)}{\sigma} \rightsquigarrow N(0, [g'(\mu)]^2).$$

To be rigorous however we need to take care of the remainder terms. Here is a more formal proof.

By a rigorous application of Taylor's theorem we obtain,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{\sigma} = g'(\widetilde{\mu})\frac{\sqrt{n}(X_n - \mu)}{\sigma},$$

where $\widetilde{\mu}$ is on the line joining μ to $\widehat{\mu}$. We know by the WLLN that $\widehat{\mu} \xrightarrow{p} \mu$ and so $\widetilde{\mu} \xrightarrow{p} \mu$. Since g is continuously differentiable, we can use the continuous mapping theorem to conclude that,

$$g'(\widetilde{\mu}) \xrightarrow{p} g'(\mu).$$

Now, we apply Slutsky's theorem to obtain that,

$$g'(\widetilde{\mu})\frac{\sqrt{n}(X_n - \mu)}{\sigma} \rightsquigarrow g'(\mu)N(0,1) \stackrel{d}{=} N(0, [g'(\mu)]^2).$$

An example: Suppose we have $X_1, \ldots, X_n \sim P$ with $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2 < \infty$. Suppose we are interested in the distribution of $Y_n = \exp(\widehat{\mu}_n)$. Using that fact that $g'(\mu) = \exp(\mu)$, applying the Delta method we obtain,

$$\sqrt{n}\left(\frac{\exp(\widehat{\mu}_n) - \exp(\mu)}{\sigma}\right) \rightsquigarrow N(0, \exp(2\mu)).$$

Multivariate Delta Method: There is a multivariate analogue of the Delta method. Suppose we have random vectors $X_1, \ldots, X_n \in \mathbb{R}^d$, and $g : \mathbb{R}^d \mapsto \mathbb{R}$ is a continuously differentiable function, then

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \rightsquigarrow N(0, \tau^2)$$

where $\overline{X}_n = n^{-1}\sum_i X_i$,

$$\tau^2 = \nabla_\mu(g)^T \Sigma \nabla_\mu(g))$$

and

$$\nabla_\mu(g) = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x)}{\partial x_d} \end{pmatrix}_{x=\mu},$$

is the gradient of g evaluated at μ.

Example. Let $X_i = (X_{i1}, X_{i2})$ have mean $\mu = (\mu_1, \mu_2)$ and variance Σ. The CLT implies that $\sqrt{n}(\overline{X}_n - \mu) \rightsquigarrow N(0, \Sigma)$. Let $g(x_1, x_2) = x_1 x_2$. Now $\nabla g = (x_2, x_1)^T$. So

$$\sqrt{n}(\overline{X}_1\overline{X}_2 - \mu_1\mu_2) \rightsquigarrow N(0, \tau^2)$$

where $\tau^2 = (\mu_2, \mu_1)^T \Sigma (\mu_2, \mu_1)$.

Lecture Notes 7
36-705

In today's lecture we will begin to study what are known as uniform laws or uniform tail bounds. Roughly, these are LLNs or tail bounds that apply to a collection of random variables taken together. Results of the type we will develop in the next few lectures form the theoretical basis for the study of statistical estimators, and are core topics in statistics and machine learning. In statistics this area of study is known as *empirical process theory*. We'll start by studying these from a relatively classical viewpoint, discussing what are called Glivenko-Cantelli theorems, and then focus on providing motivation.

1 Uniform convergence of the CDF

A classical question that was already on the mind of probabilists in the early 1930s was:

How can one estimate the CDF of a univariate random variable given a random sample?

Recall that the cdf is defined by $F(x) = P(X \leq x)$. Suppose we observe $X_1, \ldots, X_n \sim F$. The most common estimator is the empirical cdf defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x).$$

This makes sense since $F(x)$ is the probability that a random X is less than or equal to x and $\widehat{F}_n(x)$ is the observed proportion of of observations less than or equal to x. At each value x, we essentiall have a coin-flipping experiment. But we are simultaneously conducting infinitely many such experiments. Note that $\widehat{F}_n$ is a non-decreasing step-function.

You might have noticed that unlike in a classical statistical estimation problem we are not estimating a simple parameter, rather we are estimating an entire function.

So let us back up a little bit. Suppose I fixed a value x and we decided to try to estimate $F_X(x)$. We could use the empirical CDF at x, but this time it is a rather simple problem. Observe that,

$$\mathbb{E}[\widehat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{I}(X_i \leq x)] = \mathbb{P}(X \leq x) = F(x).$$

The indicators are bounded random variables so we could just use Hoeffding's bound to conclude that,

$$\mathbb{P}(|\widehat{F}_n(x) - F_X(x)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2).$$

This shows that for every point x, we can use simple tail bounds to say that the empirical CDF is close to the true CDF. Does this imply that the whole function $\widehat{F}_n$ is close to F? It does not. This is the difference between pointwise and uniform convergence.

To see that there is a difference, consider the function $g(x) = 0$ for all x. Let

$$g_n(x) = \begin{cases} 1 & \text{if } x = n \\ 0 & \text{otherwise.} \end{cases}$$

For any fixed x we have that $|g_n(x) - g(x)| \to 0$ as $n \to \infty$. But $\sup_x |g_n(x) - g(x)| = 1$ which does not go to 0.

Back to the cdf, we would like to understand the behaviour of

$$\Delta_n = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_X(x)|.$$

Reasoning about Δ_n requires us to reason about the CDF everywhere, hence the name *uniform* bounds or *uniform* LLNs. The Glivenko-Cantelli theorem says that for *any distribution*, Δ_n converges to 0 in probability.

Theorem 1 Glivenko-Cantelli Theorem. *Let $X_1, \ldots, X_n \sim F$ and define*

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x).$$

Then

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

Notes:

1. The Glivenko-Cantelli theorem is like a WLLN but it is a uniform WLLN that ensures that the WLLN is true simultaeously at every point $x \in \mathbb{R}$.

2. There is a corresponding strong theorem that guarantees convergence almost surely.

3. One should pay particular attention to the fact that we can estimate the CDF of a random variable with *no assumptions*. This is contrast to estimating the density of a random variable which typically requires strong smoothness assumptions (we will re-visit this much later in the course).

2 Empirical Processes

For any set A, define the empirical probability as

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in A).$$

The quantity Δ_n defined earlier can be written as,

$$\Delta_n = \sup_{A\in\mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|,$$

where $\mathcal{A}$ is a collection of sets,

$$\mathcal{A} = \{A(x) : A(x) = (-\infty, x]\},$$

since in this case, $\mathbb{P}(A(x)) = F_X(x)$.

One could generalize the CDF question from the previous section further to ask more generally about other interesting collections of sets $\mathcal{A}$, i.e. we are interested in collections of sets $\mathcal{A}$, for which we have uniform convergence, i.e.

$$\Delta_n(\mathcal{A}) = \sup_{A\in\mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|,$$

converges to 0 (in probability, say). This line of inquiry forms the basis for what is called *Vapnik-Cervonenkis* theory who were amongst the first to ask this general question.

Even more generally, one can replace the indicators with general (integrable) functions, i.e. let $\mathcal{F}$ be a class of integrable, real-valued functions, and suppose we have an i.i.d. sample $X_1, \ldots, X_n \sim P$, then we could be interested in,

$$\Delta_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

This quantity is known as an *empirical process* and empirical process theory is the area of statistics that asks questions about the convergence in probability, almost surely or in distribution for the quantity $\Delta(\mathcal{F})$ for interesting classes of functions $\mathcal{F}$. If $\mathcal{F} = \{I_A(\cdot) : A \in \mathcal{A}\}$ for some class of sets $\mathcal{A}$, then we get back $\Delta_n(\mathcal{A})$.

If $\Delta_n(\mathcal{F}) \xrightarrow{p} 0$ then we say that $\mathcal{F}$ is a *Glivenko-Cantelli* class. The class of functions: We refer to classes for which $\Delta(\mathcal{F}) \xrightarrow{p} 0$, as a *Glivenko-Cantelli* class. The class of functions:

$$\mathcal{F} = \{\mathbb{I}(-\infty, x],\ x \in \mathbb{R}\},$$

which defines the uniform convergence of the CDF is an example of a Glivenko-Cantelli class.

3 Failure of Glivenko-Cantelli

In general, very complex classes of functions or sets will fail to be Glivenko-Cantelli and one of the goals of the next few lectures is to find ways to measure the complexity of a class of functions. For example, suppose we draw $X_1, \ldots, X_n \sim P$ where P is some continuous distribution over $[0, 1]$. Suppose further that $\mathcal{A}$ is all subsets of $[0, 1]$ with finitely many elements.

Since the distribution is continuous we have that, $\mathbb{P}(A) = 0$ for each $A \in \mathcal{A}$, however for the finite set $\{X_1, \ldots, X_n\}$ we have that $\mathbb{P}_n(A) = 1$, i.e.

$$\Delta(\mathcal{A}) = \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| = 1,$$

no matter how large n is. So the collection of sets $\mathcal{A}$ is not Glivenko-Cantelli. Roughly, the collection of sets is "too large".

4 Estimation of Statistical Functionals

We discussed estimating the CDF of a random variable. In this section we provide several examples of problems where we use estimates of the CDF. Furthermore, as we will see, we can develop a unified understanding of such estimators using the Glivenko-Cantelli theorem.

Often we want to estimate some quantity which can be written as a simple functional of the CDF, and a natural estimate just replaces the true CDF with the empirical CDF (such estimators are known as plug-in estimators). As an aside, a functional is just a function of a function. A statistical functional is a functional of the CDF. Here are some examples:

1. **Expectation Functionals:** For a given function g, we can view the usual empirical estimator of its expectation as a plug-in estimate where we replace the population CDF by the empirical CDF,

$$\widehat{\mathbb{E}}[g(X)] = \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \int_x g(x) d\widehat{F}_n(x).$$

2. **Quantile Functionals:** For an $\alpha \in [0, 1]$, the α-th quantile of a distribution is given as:

$$Q_\alpha(F) = \inf\{t \in \mathbb{R} | F(t) \geq \alpha\}.$$

   Taking $\alpha = 0.5$ gives the median. A natural plug-in estimator of $Q_\alpha(F)$ is to simply take $Q_\alpha(\widehat{F}_n)$.

3. **Goodness-of-fit Functionals:** We will re-visit this topic in more detail when we talk about hypothesis testing but often in data analysis we want to test the hypothesis that data we have are i.i.d. from some known distribution F_0. The rough idea is we form a statistic to test this hypothesis which (hopefully) takes large values when the distribution is not F_0 and takes small values otherwise. Typical tests of this form include the Kolmogorov-Smirnov test, where we compute the plug-in quantity:

$$\widehat{T}_{KS} = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)|,$$

which is natural because if the true distribution is F_0 we know by the Glivenko-Cantelli theorem that T_{KS} is small. Similarly, one can use the Cramer-von Mises test which uses the plug-in statistic,

$$\widehat{T}_{CvM} = \int_x (\widehat{F}_n(x) - F_0(x))^2 dF_0(x).$$

There are many other statistical functionals for which the usual estimators can be thought of as plug-in estimators. For example: the variance, correlation, and higher moments can all be expressed in this fashion.

In each of the above cases we are interested in estimating some functional $\gamma(F)$ and we use the plug-in estimator $\gamma(\widehat{F}_n)$. Analogous to the continuous mapping theorem, there is a Glivenko-Cantelli theorem that provides a WLLN for these estimators. We need to first define a notion of continuity. Suppose γ satisfies the property that for every $\epsilon > 0$, there is a $\delta > 0$ such that if,

$$\sup_x |\widehat{F}_n(x) - F(x)| \leq \delta,$$

then

$$|\gamma(F) - \gamma(\widehat{F}_n)| \leq \epsilon.$$

For such functionals γ, it is a simple consequence of the Glivenko-Cantelli theorem that $\gamma(\widehat{F}_n)$ converges in probability to $\gamma(F)$.

5 Risk Minimization

Perhaps the most compelling motivation for studying uniform convergence is to understand a procedure known as empirical risk minimization. Estimators of this type include maximum likelihood estimators, and many estimators we encounter in machine learning (SVMs, Boosting and so on). We will study this in detail in the next lecture.

Binary Classification: In the typical binary classification setting we observe a training set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ that we assume are drawn i.i.d from some distribution P. Each $X_i \in \mathbb{R}^d$, $Y_i \in \{-1, +1\}$.

A classifier $f : \mathbb{R}^d \mapsto \{-1, +1\}$ is simply a function that takes an x outputs a label in $\{-1, +1\}$. The broad goal of classification is to try to find a function that has low error on future unseen data, i.e. we want a function that has low mis-classification error: $\mathbb{P}(f(X) \neq y)$.

For a given classifier f we can estimate its mis-classification error (risk) as:

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(f(X_i) \neq y_i),$$

which is simply its error on the training set. If f is some fixed classifier we know by Hoeffding's bound (why?) that,

$$\mathbb{P}(|\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| \geq t) \leq 2\exp(-2nt^2).$$

If we are trying to pick a good classifier from some set of classifiers $\mathcal{F}$, then a natural way to do this is to find the one that looks best on the training set, i.e. to choose

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f).$$

This procedure is known as *empirical risk minimization*. The terminology will be clearer later on in the course. For now though, we would like to understand this procedure better. How do we argue that in some cases this procedure will indeed select a good classifier? This question is intricately tied to uniform convergence.

Let f^* be the best classifier in $\mathcal{F}$. We would like to bound the excess risk of the classifier we chose, i.e.

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y).$$

The typical way to do this is to consider the decomposition:

$$\Delta = \underbrace{\mathbb{P}(\widehat{f}(X) \neq y) - \widehat{R}_n(\widehat{f})}_{T_1} + \underbrace{\widehat{R}_n(\widehat{f}) - \widehat{R}_n(f^*)}_{T_2} + \underbrace{\widehat{R}_n(f^*) - \mathbb{P}(f^*(X) \neq y)}_{T_3}.$$

Since $\widehat{f}$ minimizes the empirical risk we know that $T_2 \leq 0$. We know that T_3 is small just by the Hoeffding argument from before, since f^* is a fixed classifier (i.e. does not depend on the training data).

The key point, one that you should really think carefully about is that we cannot use Hoeffding for the first term. The reason is that the classifier $\widehat{f}$ is data dependent so its empirical risk is not the sum of independent RVs.

Instead we have to rely on a *uniform* convergence bound, i.e. suppose we can show that with probability at least $1 - \delta/2$,

$$\sup_{f \in \mathcal{F}} \left[\mathbb{P}(f(X) \neq y) - \widehat{R}_n(f) \right] \leq \Theta,$$

then we can conclude that the excess risk with probability at least $1 - \delta$ satisfies

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y) \leq \Theta + \sqrt{\frac{2\log(2/\delta)}{n}},$$

so everything boils down to showing uniform convergence of the empirical risk to the true error over the collection of classifiers we are interested in.

We will discuss this is more detail later.

Lecure Notes 8
36-705

We continue our discussion of uniform convergence. Recall that

$$P_n(A) = \frac{1}{n}\sum_i I_A(X_i)$$

and

$$\Delta_n(\mathcal{A}) = \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|.$$

1 Finite Collections

The first case to consider is when the collection of sets $\mathcal{A}$ has finite cardinality $|\mathcal{A}|$. In other words

$$\mathcal{A} = \{A_1, \ldots, A_N\}.$$

In this case, for a fixed A we know by Hoeffding's inequality that,

$$\mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t) \leq 2\exp(-2nt^2).$$

However, we want something stronger we want that this convergence happens uniformly for all sets in $\mathcal{A}$, so we can use the union bound, i.e.

$$\begin{aligned}\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) &= \mathbb{P}(\cup_{A\in\mathcal{A}}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq \sum_{A\in\mathcal{A}} \mathbb{P}(|\mathbb{P}_n(A) - \mathbb{P}(A)| \geq t)) \\ &\leq 2|\mathcal{A}|\exp(-2nt^2).\end{aligned}$$

If we set the right hand side equal to δ we get

$$t = \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

So

$$\mathbb{P}\left(\Delta_n(\mathcal{A}) \geq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}\right) \leq \delta.$$

In other words we have that with probability at least $1-\delta$,

$$\Delta(\mathcal{A}) \leq \sqrt{\frac{\ln(2|\mathcal{A}|/\delta)}{2n}}.$$

This is already quite a nice result and once again highlights one of the main reasons why Hoeffding type exponential concentration inequalities are much more useful than Chebyshev type concentration inequalities: to obtain uniform convergence over $\mathcal{A}$ we pay a price which is logarithmic in the size of the collection.

2 VC dimension

Often we are interested in controlling $\Delta(\mathcal{A})$ for infinite classes of sets. The example from last lecture of uniform convergence of the empirical CDF is a canonical example. One way to do this is to use the notion of VC dimension. First, we need to understand the concept of shattering.

Shattering: Let $\{z_1, \ldots, z_n\}$ be a finite set of n points. We let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of distinct sets in the collection of sets

$$\Big\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\Big\}.$$

$N_{\mathcal{A}}(z_1, \ldots, z_n)$ is counting the *number of subsets* of $\{z_1, \ldots, z_n\}$ that the collection of sets $\mathcal{A}$ picks out. Note that, $N_{\mathcal{A}}(z_1, \ldots, z_n) \leq 2^n$.

We now define the n-th shatter coefficient of $\mathcal{A}$ as:

$$s(\mathcal{A}, n) = \max_{\{z_1, \ldots, z_n\}} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

The shatter coefficient is the maximal number of different subsets of n points that can be picked out by the collection $\mathcal{A}$.

Example: Consider points on the real line and let $\mathcal{A}$ be the collection of left intervals $\mathbb{I}(-\infty, t]$ for all t. If we have n points on the line then we can pick out any left subset of the points, i.e. $s(\mathcal{A}, n) = n + 1$. For example, let $n = 3$ and consider the points $\{0, 1, 2\}$. We can pick out the following subsets:

$$\{\emptyset\}, \{0\}, \{0, 1\}, \{0, 1, 2\},$$

and no others. This is true for any set of three points. So $s(\mathcal{A}, 3) = 4$. We will see more examples soon.

VC Theorem: For *any distribution* $\mathbb{P}$, and class of sets $\mathcal{A}$ we have that,

$$\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) \leq 8 s(\mathcal{A}, n) \exp(-nt^2/32).$$

Notes: There are two noteworthy aspects of this theorem.

1. The result is very general and it applies to any distribution on the samples, and such results are often called *distribution free*.

2. The VC theorem essentially reduces the question of uniform convergence to a combinatorial question about the collection of sets, i.e. we now need only to understand the shatter coefficients which are completely independent from probability/statistics.

3. The proof of this result is quite straightforward using some of the machinery (introducing a ghost sample, symmetrization) that we will see in the next lecture.

Glivenko-Cantelli: This theorem immediately implies the Glivenko-Cantelli theorem we studied in the last lecture, i.e. that the empirical CDF converges in probability to the true CDF. To see this we note that the shatter coefficients of the left intervals are bounded by $n+1$ so the VC theorem tells us that,

$$\mathbb{P}\left(\sup_x |\widehat{F}_n(x) - F_X(x)| \geq t\right) \leq 8(n+1)\exp(-nt^2/32).$$

Now verifying convergence in probability is straightforward by noting that for any $t > 0$, $\lim_{n\to\infty} 8(n+1)\exp(-nt^2/32) = 0$.

VC dimension: We now that $s(\mathcal{A}, n) \leq 2^n$ for each n. The *VC dimension* d is the largest integer d for which $s(\mathcal{A}, d) = 2^d$.

It follows that, for any $n > d$, we have that $s(\mathcal{A}, n) < 2^n$. The surprising combinatorial result of Vapnik and Chervonenkis (sometimes called Sauer's lemma) is that there is a phase transition of shattering coefficients: once it is no longer exponential (i.e. once $n > d$) the shattering coefficients become polynomial in n, i.e.

Sauer's Lemma: If $\mathcal{A}$ has finite VC dimension d, then for $n > d$ we have that,

$$s(\mathcal{A}, n) \leq (n+1)^d.$$

We can use Sauer's lemma to conclude that for a system $\mathcal{A}$ of VC dimension d.

$$\mathbb{P}(\Delta_n(\mathcal{A}) \geq t) \leq 8(n+1)^d \exp(-nt^2/32).$$

Doing the usual thing we see that with probability $1-\delta$,

$$\Delta_n(\mathcal{A}) \leq \sqrt{\frac{32}{n}\left[d\log(n+1) + \log(8/\delta)\right]}.$$

There are some important notes:

1. If $d < \infty$ then $\Delta(\mathcal{A}) \xrightarrow{p} 0$, and so we have a uniform LLN for the collection of sets $\mathcal{A}$.
2. There are converses to the VC theorem that say roughly that if the VC dimension is infinite then there exists a distribution over the samples for which we do not have a uniform LLN.
3. Roughly, one should think of the VC result as saying for a class with VC dimension d,

$$\Delta(\mathcal{A}) \approx \sqrt{\frac{d\log n}{n}}.$$

3 More examples

There are many examples of collections of sets for which the VC dimension is known. A few popular ones are in Table 1.

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 1: The VC dimension of some classes $\mathcal{A}$.

4 Back to Binary Classification

In binary classification, we have a collection of classifiers $\mathcal{F}$. This collection induces a set system:

$$\mathcal{A} = \left\{ \Big\{ \{x : f(x) = 1\} \times \{0\} \Big\} \bigcup \Big\{ \{x : f(x) = 0\} \times \{1\} \Big\}, f \in \mathcal{F} \right\}.$$

If $\mathcal{A}$ has VC dimension d then we can use the VC theorem in a straightforward way to conclude that with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| = \Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n} \left[d \log(n + 1) + \log(8/\delta) \right]}.$$

It is not too hard to verify that the VC dimension is essentially driven by the complexity of the sets $\mathbb{I}(f(x) = 1)$ and their complements for the classifiers in $\mathcal{F}$. This in a straightforward way, for instance, leads to a uniform convergence guarantee for empirical risk minimization over linear classifiers since they induce relatively simple sets (half-spaces) whose VC dimension is well-understood.

5 Rademacher Complexity

Now we discuss a different notion of the complexity of a class of functions. Suppose we have a collection of functions $\mathcal{F}$, we observe samples $X_1, \ldots, X_n \sim P$ for some distribution P and we are interested in (upper bounding) the quantity:

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

Fix a set of points $\{x_1, \ldots, x_n\}$. Let $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$ denote a collection of n Rademacher random variables, i.e. they take the values $\{+1, -1\}$ with equal probabilities. We define the *empirical* Rademacher complexity by

$$\mathcal{R}(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right].$$

When $\{x_1, \ldots, x_n\}$ is a random sample then the empirical Rademacher complexity is a random variable. We define the Rademacher complexity of the class $\mathcal{F}$ as the expectation of this quantity, i.e.

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right].$$

The Rademacher complexity measures the maximum absolute covariance between $\{f(X_1), \ldots, f(X_n)\}$ and a vector of random signs $\{\epsilon_1, \ldots, \epsilon_n\}$. Intuitively, we think of a class $\mathcal{F}$ as too large if for many random sign vectors we can find a function in $\mathcal{F}$ that is strongly correlated with the random sign vectors. The main utility of the Rademacher complexity is that it upper bounds the quantity $\Delta_n(\mathcal{F})$.

Rademacher Theorem:

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F}).$$

This theorem again might not appear to be so useful since we still need to understand the Rademacher complexity. It turns out that the Rademacher complexity is relatively easy to upper bound in terms of more geometric measures of the function class $\mathcal{F}$ (these are things like covering numbers or bracketing numbers of $\mathcal{F}$). This is analogous to how VC theory gave us a way to go from the uniform convergence question to a combinatorial property of the collection of sets.

Proof: The proof will resemble what we did when we proved Hoeffding's inequality. We will introduce a ghost sample, and symmetrize the empirical process. Let $\{Y_1, \ldots, Y_n\}$ be

another independent identically distributed sample. Then,

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathcal{F})] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f]\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_i} f(Y_i)\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y \frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right] \\
&\leq \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\mathbb{E}_Y\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right] \\
&\leq \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n [f(X_i) - f(Y_i)]\right|\right].
\end{aligned}$$

The distribution of the difference $f(X_i) - f(Y_i)$ is the same as the distribution of $\epsilon_i(f(X_i) - f(Y_i))$ so we obtain,

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathcal{F})] &\leq \mathbb{E}_{X,Y,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i[f(X_i) - f(Y_i)]\right|\right] \\
&\leq 2\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \\
&= 2\mathcal{R}_n(\mathcal{F}),
\end{aligned}$$

which gives us the Rademacher theorem. □

If the function class is bounded, i.e. for every $f \in \mathcal{F}$ we have that $\|f\|_\infty \leq b$, then by McDiarmid's inequality, $\Delta_n(\mathcal{F})$ is sharply concentrated around its mean, i.e.

$$\mathbb{P}(|\Delta(\mathcal{F}) - \mathbb{E}[\Delta(\mathcal{F})]| \geq t) \leq 2\exp(-nt^2/(2b^2)).$$

Putting this inequality together with the upper bound on the mean we obtain that for a bounded class $\mathcal{F}$ with probability at least $1-\delta$,

$$\Delta(\mathcal{F}) \leq 2\mathcal{R}(\mathcal{F}) + b\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Lecture Notes 10
36-705

Let $\mathcal{F}$ be a set of functions and recall that

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

Let us also recall the Rademacher complexity measures

$$\mathcal{R}(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]$$

and

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right]$$

where $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables. We have already shown that

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F}).$$

1 Rademacher Complexity of a Finite Class

Suppose that we have a finite collection of functions $\mathcal{F} = \{f_1, \ldots, f_N\}$, which are bounded i.e. $\|f_i\|_\infty \leq b$ then we have the following bound on the Rademacher complexity.

Finite Class Bound: The Rademacher complexity for a finite class,

$$\mathcal{R}(\mathcal{F}) \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

Proof: Define,

$$\Theta := \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right].$$

For convenience let us augment the class $\mathcal{F}$ with the negative of every function, i.e. we take $\widetilde{\mathcal{F}} = \mathcal{F} \cup (-\mathcal{F})$, so that there are now $2N$ functions. Then,

$$\Theta \leq \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \widetilde{\mathcal{F}}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right].$$

Note that,

$$\begin{aligned}
\exp(t\Theta) &\leq \exp\left(t\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right) \\
&\leq \mathbb{E}_{X,\epsilon}\exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right) \\
&\leq \mathbb{E}_{X,\epsilon}\sum_{j=1}^{2N}\prod_{i=1}^{n}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right) \\
&= \sum_{j=1}^{2N}\prod_{i=1}^{n}\mathbb{E}_{X,\epsilon}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right).
\end{aligned}$$

Since $\|f_j\|_\infty \leq b$ we can use the argument we used in the proof of Hoeffding's inequality to obtain that,

$$\exp(t\Theta) \leq 2N\exp\left(\frac{4t^2b^2}{n}\right),$$

so we obtain that,

$$\Theta \leq \frac{\log(2N)}{t} + \frac{4tb^2}{n},$$

where t is a free parameter that is > 0. Choosing, $t = \sqrt{n\log(2N)/(4b^2)}$ we obtain,

$$\Theta \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

2 Using the Rademacher Theorem to obtain the VC theorem

The Rademacher theorem in a very straightforward way implies the VC theorem. We'll sketch the proof here. Our class of functions just corresponds to the indicators arising from the set system. These functions are upper bounded by $b = 1$. We can get a high-probability statement as in the initial section so we only need to deal with $\mathcal{R}(\mathcal{F})$.

We follow an identical argument to the one we did in the previous section,

$$\exp(t\Theta) \leq \mathbb{E}_{X,\epsilon}\exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right),$$

where the class $\widetilde{\mathcal{F}}$ just contains the set indicators and their negations.

The key point is to note here is the following: suppose we think of the vectors $(f(X_1), \ldots, f(X_n))$ for each function in $\widetilde{\mathcal{F}}$ and ask how many different such vectors are there? Each set in $\mathcal{A}$ picks out some subset of the points (and assigns them the value $+1$). Even though there are possibly infinitely many sets in $\mathcal{A}$ there are at most only twice (because we included the negations) the shattering number of different vectors.

The shattering number is precisely the (maximum) number of different vectors $(f(X_1), \ldots, f(X_n))$ we can induce using our collection of sets.

With this insight in hand we can just repeat the previous argument to conclude that,

$$\Theta \leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}},$$

and putting this together with the high-probability bound from before we have that with probability at least $1 - \delta$,

$$\begin{aligned}\Theta &\leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}} \\ &\leq \sqrt{\frac{4\log(4s(\mathcal{A}, n)/\delta)}{n}},\end{aligned}$$

which is precisely the VC theorem (again always ignore constants).

Lecture Notes 10
36-705

The first chunk of our course has focused primarily on properties of averages of i.i.d. random variables. In particular, the question of interest roughly was "how close is an average of i.i.d. random variables to an expectation?". We developed tail bounds (non-asymptotic) and understood the limiting distribution of the average (asymptotic) as ways to attack this question. Then we discussed uniform laws where the question was how to argue that many (possibly related averages) converge to their respective expectations.

Now, we will switch gears and start to talk a bit more formally about statistical estimation and inference. Before we can make sense of these questions however we need to define a statistical model.

1 Statistical Models

The typical starting point for statistical inference should be familiar to you by now: we suppose that we obtain an i.i.d sample $\{X_1, \ldots, X_n\}$ from some distribution P.

Often, we further hypothesize restrictions on the set of possible distributions that could have generated the data, i.e. we suppose that $P \in \mathcal{P}$ where $\mathcal{P}$ is just some collection of distributions. We refer to $\mathcal{P}$ as the *statistical model*.

The common classes of distributions usually fall into one of two broad categories (with lots of fuzziness in between):

1. **Parametric models:** Here the statistical model $\mathcal{P}$ is described by a finite set of parameters. We usually write these as:

$$\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\},$$

   where $\Theta \subseteq \mathbb{R}^d$, i.e. these are a collection of distributions that we can describe using d real-valued parameters. We use the notation $p(x;\theta)$ or $p_\theta(x)$ to denote the density of the random variable X, where the distribution is parameterized by θ.

   A typical example is $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and in this case the parameter $\theta = (\mu, \sigma) \in \mathbb{R}^2$.

2. **Non-parametric models:** Roughly, non-parametric models are those which cannot be described by a finite set of parameters.

   A few common examples are: (1) the set of all possible distributions on $\mathbb{R}$ (say):

$$\mathcal{P} = \{\text{all distributions on } \mathbb{R}\}.$$

In this case we are making no assumptions on the data generating process (beyond the i.i.d. assumption). When we studied estimating the CDF we made no assumptions and were implicitly working with this non-parametric model.

(2) Another common example is the set of distributions with smooth (say with square integrable second derivative) densities, i.e.

$$\mathcal{P} = \left\{ p : \int_{\mathbb{R}} (p''(x))^2 dx \leq C \right\}.$$

The model assumptions are the starting point for all subsequent inference, and can strongly influence our conclusions about the data. In general, non-parametric models are making much weaker assumptions about the data, while parametric models make strong (often unjustifiable assumptions).

On the other hand, as we will see later on in the course, parametric models can often be estimated using far fewer samples and can sometimes be much more interpretable. **When we use parametric models, we should generally think of it as an approximation. It is rare that we would expect a parametric model to be exactly correct.**

We will begin investigating parametric models and developing a comprehensive theory for estimation and inference for parametric models before turning our attention to non-parametric models.

2 Statistics

A statistic is simply a function of the observed sample, i.e. if we have $X_1, \ldots, X_n \sim P$ then any function $T(X_1, \ldots, X_n)$ is called a statistic. A statistic is a random variable.

Some examples of statistics include:

1. order statistics, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
2. sample mean: $\overline{X} = \frac{1}{n} \sum_i X_i$,
3. sample variance: $S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{x})^2$,
4. sample median: middle value of ordered statistics,
5. sample minimum: $X_{(1)}$
6. sample maximum: $X_{(n)}$.

3 Sufficient Statistics and Data Reduction

Now we will talk about a special class of statistics called sufficient statistics. There are two ways to motivate sufficient statistics: the data reduction viewpoint where we would like to discard non-informative pieces of the dataset (for storage or other benefits), and the risk reduction viewpoint where we want to construct estimators that only depend on meaningful variation in the data. We will focus on the former viewpoint today.

The goal in data reduction roughly is to find ways to reduce the size of a dataset without throwing away important information. We need to fix ideas more concretely to make progress on this abstract question.

We focus on parametric models and suppose we observe samples

$$X_1, \ldots, X_n \sim p(x; \theta).$$

The typical statistical estimation problem is that we observe the samples as want to understand something about the unknown parameter θ. The goal of data reduction is to find statistics $T(X_1, \ldots, X_n)$ that contain all the information about the unknown parameter θ.

Sufficient Statistic: A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for the parameter θ if $p(x_1, \ldots, x_n | T(X_1, \ldots, X_n) = t; \theta)$ does not depend on θ.

Once we know the value of the sufficient statistic, the joint distribution no longer has any more information about the parameter θ. One could imagine keeping only $T(X_1, \ldots, X_n)$ and throwing away all the data. Take this loose interpretation with a grain of salt - we will return to it.

There is another more subtle question about what it means to condition on the sufficient statistic when it comes from a continuous distribution (since typically the conditioning set will have zero probability). It turns out that one can salvage this using something called the factorization theorem (which we will come to shortly). For now however, you should think about the discrete case and suppose that the probability that $T(X_1, \ldots, X_n) = t$ is strictly positive.

Poisson sufficient statistics: $X_1, \cdots, X_n \sim \text{Poisson}(\theta)$. Let $T = \sum_{i=1}^n X_i$.

Then,

$$p_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p_\theta(T = t)}.$$

There are two possibilities, that the sum of the x_is is equal to t and that the sum is not equal to t. In the latter case the probability is zero so we obtain,

$$\frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p_\theta(T = t)} = \frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n)\mathbb{I}(T = t)}{p_\theta(T = t)}$$

The sum of n independent Poissons with parameter θ is Poisson with parameter $n\theta$ so we obtain,

$$\begin{aligned}\frac{p_\theta(X_1 = x_1, \ldots, X_n = x_n)\mathbb{I}(T = t)}{p_\theta(T = t)} &= \frac{\exp(-n\theta)\theta^{\sum_{i=1}^n x_i}\mathbb{I}(T = t)(\sum_{i=1}^n x_i)!}{[\prod_{i=1}^n (x_i)!]\exp(-n\theta)(n\theta)^{\sum_{i=1}^n x_i}} \\ &= \frac{\mathbb{I}(T = t)t!}{[\prod_{i=1}^n (x_i)!]\, n^t},\end{aligned}$$

which does not depend on θ. So we can conclude that $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic. You can similarly verify that the average is sufficient, and that $3.7\sum_{i=1}^n X_i$ is sufficient. Furthermore, you can always condition on more things without destroying sufficiency so that $(\sum_{i=1}^n X_i, X_1, X_{17})$ is also sufficient.

In some sense we might believe that the sum is a better sufficient statistic than $(\sum_{i=1}^n X_i, X_1, X_{17})$. We will return to this idea of "minimal sufficient statistics" in the next class.

Binomial sufficient statistics: Suppose we observe $X_1, \ldots, X_n \sim \text{Ber}(\theta)$, then once again we can verify that the sum is sufficient. This proceeds in exactly the same way, i.e.

$$\begin{aligned}p(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{p(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p(T = t)} \\ &= \frac{\mathbb{I}(T = t)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\mathbb{I}(T = t)}{\binom{n}{t}},\end{aligned}$$

which does not depend on θ.

3.1 Sufficient Statistics - The Partition Viewpoint

It is better to describe sufficiency in terms of partitions of the sample space.

Example 1 *Let* $X_1, X_2, X_3 \sim \text{Bernoulli}(\theta)$. *Let* $T = \sum X_i$.

(x_1, x_2, x_3)		t	$p(x\|t)$
(0, 0, 0)	$\to$	$t = 0$	*1*
(0, 0, 1)	$\to$	$t = 1$	*1/3*
(0, 1, 0)	$\to$	$t = 1$	*1/3*
(1, 0, 0)	$\to$	$t = 1$	*1/3*
(0, 1, 1)	$\to$	$t = 2$	*1/3*
(1, 0, 1)	$\to$	$t = 2$	*1/3*
(1, 1, 0)	$\to$	$t = 2$	*1/3*
(1, 1, 1)	$\to$	$t = 3$	*1*
8 elements	$\to$	*4 elements*	

1. A partition $B_1, \ldots, B_k$ is sufficient if $f(x|X \in B)$ does not depend on θ.
2. A statistic T induces a partition. For each t, $\{x : \ T(x) = t\}$ is one element of the partition. T is sufficient if and only if the partition is sufficient.
3. Two statistics can generate the same partition: example: $\sum_i X_i$ and $3\sum_i X_i$.
4. If we split any element B_i of a sufficient partition into smaller pieces, we get another sufficient partition.

Example 2 *Let* $X_1, X_2, X_3 \sim$ Bernoulli(θ). *Then* $T = X_1$ *is* **not** *sufficient. Look at its partition:*

(x_1, x_2, x_3)		t	$p(x\|t)$
(0, 0, 0)	$\to$	$t = 0$	$(1-\theta)^2$
(0, 0, 1)	$\to$	$t = 0$	$\theta(1-\theta)$
(0, 1, 0)	$\to$	$t = 0$	$\theta(1-\theta)$
(0, 1, 1)	$\to$	$t = 0$	θ^2
(1, 0, 0)	$\to$	$t = 1$	$(1-\theta)^2$
(1, 0, 1)	$\to$	$t = 1$	$\theta(1-\theta)$
(1, 1, 0)	$\to$	$t = 1$	$\theta(1-\theta)$
(1, 1, 1)	$\to$	$t = 1$	θ^2
8 elements	$\to$	*2 elements*	

4 The Factorization Theorem

Checking the definition of sufficiency directly is often a tedious exercise since it involves computing the conditional distribution. A much simpler characterization of sufficiency comes from what is called the Neyman-Fisher factorization criterion.

Theorem 3 *$T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint pdf/pmf of $(X_1, \ldots, X_n)$ can be factored as*

$$p(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) \times g(T(x_1, \ldots, x_n); \theta).$$

This version does not involve conditioning and thus typically makes sense even when X has a continuous distribution. Let us consider an example of this.

Example 4 *$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then*

$$p(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

(a) If (μ, σ^2) unknown then we can write the joint as:

$$p(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right\},$$

so that $T = (\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$ is sufficient using the factorization theorem.

(b) If σ is known: we can add and subtract the sample mean $\overline{x}$ in the exponent and use the fact that, $\sum_{i=1}^n (x_i - \overline{x}) = 0$ to see that,

$$p(x_1, \ldots, x_n; \mu) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{-\sum(x_i - \overline{x})^2}{2\sigma^2}\right\}}_{h(x^n)} \quad \underbrace{\exp\left\{\frac{-n(\overline{x} - \mu)^2}{2\sigma^2}\right\}}_{g(T(x^n)|\mu)}.$$

Thus, using the factorization theorem $\overline{X}$ is sufficient for μ.

The factorization theorem is relatively straightforward to prove, at least in the discrete case.

Proof: Factorization $\implies$ sufficiency:

$$\begin{aligned} p(x_1, \ldots, x_n | T = t; \theta) &= \frac{p(x_1, \ldots, x_n, T = t; \theta)}{p(T = t; \theta)} \\ &= \frac{\mathbb{I}(T(x_1, \ldots, x_n) = t) h(x_1, \ldots, x_n) \times g(t; \theta)}{\sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} h(x_1, \ldots, x_n) \times g(t; \theta)} \\ &= \frac{\mathbb{I}(T(x_1, \ldots, x_n) = t) h(x_1, \ldots, x_n)}{\sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} h(x_1, \ldots, x_n)}, \end{aligned}$$

which does not depend on θ.

Sufficiency $\implies$ factorization: We simply define $g(t; \theta) = p(T = t; \theta)$, and $h(x_1, \ldots, x_n) = p(x_1, \ldots, x_n | T = t; \theta)$, where by sufficiency we note that the latter function does not depend on θ. Now, it is straightforward to verify that factorization theorem holds.

5 The likelihood and its relationship to sufficiency

The likelihood function arises from viewing the joint density as a function of the unknown parameter θ, i.e.

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; x_1, \ldots, x_n) = p(x_1, \ldots, x_n; \theta).$$

The likelihood is central in computing point estimates (maximum likelihood estimation) and Bayesian inference. We will return to these.

Some important points to remember:

1. The likelihood is a function of θ, it is not a probability.
2. Typically we ignore constants that do not depend on θ when computing the likelihood, i.e. we care only about the relative value of the likelihood as a function of θ. More formally, the likelihood is only defined upto a constant of proportionality, i.e. it is an equivalence class of functions.
3. The likelihood in the i.i.d. case has the form:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i; \theta),$$

and in this case we often will find it convenient to work with the log-likelihood:

$$\mathcal{LL}(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta).$$

Relationship to sufficiency: Using the factorization theorem we can see that for any sufficient statistic T, we can write the likelihood as:

$$\mathcal{L}(\theta) = g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n),$$

and noting once again that the likelihood is only defined upto constants that do not depend on θ, we can simply ignore the $h(x_1, \ldots, x_n)$ term, i.e. we can define the likelihood as:

$$\mathcal{L}(\theta) = g(T(x_1, \ldots, x_n); \theta).$$

Thus, once we have any sufficient statistic, we have everything we need to compute the likelihood function. If we were to base our subsequent data analysis only on the likelihood function then we have lost nothing. We will see a further connection between sufficiency and the likelihood function once we define the notion of minimal sufficiency.

Lecture Notes 11
36-705

Today we will continue our discussion on sufficiency.

1 Minimal sufficiency

As we have seen previously sufficient statistics are not unique. Furthermore, it seems, at least intuitively, that some sufficient statistics present much more reduction than others. For example, suppose that $X_1, \ldots, X_n \sim N(\mu, 1)$. Then $\overline{X}_n$ is a sufficient statistic. But so is the whole data set. This motivates the following definition of minimal sufficient statistics:

Minimal Sufficiency: A statistic $T(x_1, \ldots, x_n)$ is minimal sufficient if it is sufficient, and furthermore for any other sufficient statistic $S(x_1, \ldots, x_n)$ we can write $T(x_1, \ldots, x_n) = g(S(x_1, \ldots, x_n))$, i.e. T is a function of S.

Theorem 1 *Define*

$$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta) = \frac{p(y_1, \ldots, y_n; \theta)}{p(x_1, \ldots, x_n; \theta)}.$$

Suppose that a statistic T has the following property:

$R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta)$ **does not depend on** θ **if and only if** $T(y_1, \ldots, y_n) = T(x_1, \ldots, x_n)$**.**

Then T is a MSS.

Before we prove the theorem let us consider some examples.

Example 2 *Suppose that $Y_1, \ldots, Y_n$ are i.i.d Poisson (θ).*

$$p(y_1, \ldots, y_n; \theta) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i}, \quad \frac{p(y_1, \ldots, y_n; \theta)}{p(x_1, \ldots, x_n; \theta)} = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i! / \prod x_i!}$$

which is independent of θ iff $\sum y_i = \sum x_i$. This implies that $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a minimal sufficient statistic for θ.

The minimal sufficient statistic is not unique. But, the minimal sufficient partition is unique.

Example 3 *Cauchy.*

$$p(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

Then

$$\frac{p(y_1,\ldots,y_n;\theta)}{p(x_1,\ldots,x_n;\theta)} = \frac{\prod_{i=1}^{n}\{1+(x_i-\theta)^2\}}{\prod_{j=1}^{n}\{1+(y_j-\theta)^2\}}.$$

The ratio is a constant function of θ if

$$T(X_1,\ldots,X_n) = (X_{(1)},\cdots,X_{(n)}).$$

It is technically harder to show that the ratio is independent of θ only if T is the order statistics, but it could be done using theorems about polynomials. Having shown this, one can conclude that the order statistics are the minimal sufficient statistics for θ.

Proof: We prove this in two steps. We first show that T is a sufficient statistic and then we check that it is minimal. We define the partition induced by T, as $\{A_t : t \in \text{Range}(T)\}$ and for each set in the partition A_t we associate a representative $(x_{t1},\ldots,x_{tn}) \in A_t$.

T **is sufficient:** We look at the joint distribution at any $(x_1,\ldots,x_n)$. Suppose that $T(x_1,\ldots,x_n) = u$, then consider $(y_1,\ldots,y_n) := (x_{u1},\ldots,x_{un})$. Observe that, $(y_1,\ldots,y_n)$ depends only on $T(x_1,\ldots,x_n)$, i.e. the point y is a function of the statistic T only. Now we have that,

$$p(x_1,\ldots,x_n;\theta) = p(y_1,\ldots,y_n;\theta)R(y_1,\ldots,y_n,x_1,\ldots,x_n;\theta),$$

and since $T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n)$, R does not depend on θ. Recalling that $(y_1,\ldots,y_n)$ is only a function of $T(x_1,\ldots,x_n)$ we have that,

$$p(x_1,\ldots,x_n;\theta) = g(T(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n),$$

where g corresponds to the first term and h corresponds to the R term. We conclude that T is sufficient.

T **is minimal:** As a preliminary we note that the definition of a minimal sufficient statistic could be equivalently written as: T is a MSS if for any other sufficient statistic S, if we have that $S(x_1,\ldots,x_n) = S(y_1,\ldots,y_n)$ then we also have that $T(x_1,\ldots,x_n) = T(y_1,\ldots,y_n)$. This is equivalent to the statement that T is a function of S.

Consider, any other sufficient statistic S. Suppose that, $S(x_1,\ldots,x_n) = S(y_1,\ldots,y_n)$, then

by the factorization theorem we have that,

$$\begin{aligned} p(x_1, \ldots, x_n; \theta) &= g(S(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n) \\ &= g(S(y_1, \ldots, y_n); \theta) h(y_1, \ldots, y_n) \frac{h(x_1, \ldots, x_n)}{h(y_1, \ldots, y_n)} \\ &= p(y_1, \ldots, y_n; \theta) \frac{h(x_1, \ldots, x_n)}{h(y_1, \ldots, y_n)}, \end{aligned}$$

so we have that $R(x_1, \ldots, x_n, y_1, \ldots, y_n; \theta)$ does not depend on θ. So we conclude that $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$ and so T is minimal. ■

2 Minimal sufficiency and the likelihood

Although minimal sufficient statistics are not unique they induce a unique partition on the possible datasets. This partition is also induced by the likelihood.

Lemma 4 *Suppose we have a partition such that $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are placed in the same set of the partition iff $L(\theta; x_1, \ldots, x_n) \propto L(\theta; y_1, \ldots, y_n)$, then the partition is the minimal sufficient partition.*

You will prove this on your homework but it is a simple consequence of the characterization we have seen in the previous section.

3 Sufficiency - the risk reduction viewpoint

We will return to the concept of risk more formally in the next few lectures, but for now let us try to understand the main ideas.

Setting: Suppose we observe $X_1, \ldots, X_n \sim p(x; \theta)$ and we would like to estimate θ, i.e. we want to construct some function of the data that is close in some sense to θ. We construct an estimator $\widehat{\theta}(X_1, \ldots, X_n)$. In order to evaluate our estimator we might consider how far our estimate is from θ on average, i.e. we can define

$$R(\widehat{\theta}, \theta) = \mathbb{E}(\widehat{\theta} - \theta)^2.$$

We will see this again later on but the risk of an estimator can be decomposed into its bias and variance, i.e.

$$\mathbb{E}(\widehat{\theta} - \theta)^2 = (\mathbb{E}\widehat{\theta} - \theta)^2 + \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2,$$

where the first term is referred to as the bias and the second is the variance.

There is a strong sense in which estimators which do not depend only on sufficient statistics can be improved. This is known as the Rao-Blackwell theorem.

Let $\widehat{\theta}$ be an estimator. Let T be any sufficient statistic and define $\widetilde{\theta} = \mathbb{E}[\widehat{\theta}|T]$.

Rao-Blackwell theorem:

$$R(\widetilde{\theta}, \theta) \leq R(\widehat{\theta}, \theta).$$

We will not spend too much time on this but lets see a quick example and then prove the result.

Example: Suppose we toss a coin n times, i.e. $X_1, \ldots, X_n \sim \text{Ber}(\theta)$. We consider the estimator:

$$\widehat{\theta} = X_1,$$

and the sufficient statistic $T = \sum_{i=1}^n X_i$, then

$$\widetilde{\theta} = \mathbb{E}[X_1|T] = \mathbb{E}\Big[X_1 \;\Big|\; \sum_i X_i\Big].$$

We claim that the conditional expectation is simply the average, i.e.

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^n X_i.$$

First, let us check this in the case when $n = 2$. If $X_1 + X_2 = 2$ then $X_1 = 1$, and if $X_1 + X_2 = 0$, $X_1 = 0$. In the case, when $X_1 + X_2 = 1$, we have $X_1 = 1$ with probability $1/2$ and 0 with probability $1/2$. So we conclude the conditional expectation is $(X_1 + X_2)/2$.

More generally, if we have $\sum X_i = k$, then of the $\binom{n}{k}$ equally likely possibilities we have that $X_1 = 1$ for $\binom{n-1}{k-1}$ of them so that the conditional expectation is simply:

$$\mathbb{E}\Big[X_1 \;\Big|\; \sum_i X_i = k\Big] = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n} = \overline{X}_n.$$

We observe that both estimators are unbiased (have mean equal to θ) but the variance of the Rao-Blackwellized estimator is $\theta(1-\theta)/n$ as opposed to the original estimator which has variance $\theta(1-\theta)$.

Proof of Rao-Blackwell: Observe that,

$$R(\widetilde{\theta}, \theta) = \mathbb{E}[(\mathbb{E}[\widehat{\theta}|T] - \theta)^2] = \mathbb{E}[(\mathbb{E}[\widehat{\theta} - \theta|T])^2] \leq \mathbb{E}[\mathbb{E}[(\widehat{\theta} - \theta)^2|T]] = R(\widehat{\theta}, \theta).$$

The inequality is Jensen's inequality (equivalently just $\text{Var(X)} = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0$).

A question worth pondering is: why does it matter for Rao-Blackwellization that T is a sufficient statistic?

Lecture Notes 12
36-705

Today we will discuss a special type of statistical model called aan *exponential family.*

1 Exponential Families

A family $\{P_\theta\}$ of distributions forms an s-dimensional exponential family if the distributions P_θ have densities of the form:

$$p(x;\theta) = \exp\left[\sum_{i=1}^{s} \eta_i(\theta)T_i(x) - A(\theta)\right] h(x),$$

where η_i, A are functions which map θ to $\mathbb{R}$. You can check that $(T_1(x), \ldots, T_s(x))$ is a sufficient statistic. The term $A(\theta)$ is known as the log-normalization constant or the log-partition function. Let $\mathcal{X}$ be the set of possible values of x.

Remark: You can ignore this: As a technical note, exponential families can be defined with respect to the Lebesgue measure (as we did implicitly above) or with respect to any other measure (for instance, the discrete measure on $\{1, \ldots, k\}$). We will continue to simply think of $\mathcal{X}$ as a subset of $\mathbb{R}$ and the measure as the Lebesgue measure.

Although thinking of the above form is standard, it is usually much more convenient to parametrize the distribution in what is known as its *canonical parametrization*, where we simply take $\eta_i(\theta)$ to be the parameters. In this case, we can more compactly write:

$$p(x;\theta) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(x) - A(\theta)\right] h(x).$$

In this case, we refer to θ as the natural parameters of the distribution. Notice that none of these parametrizations are unique, we can replace T_i by cT_i and θ_i by θ_i/c and obtain the same distribution.

The term $A(\theta)$ is what makes the distribution integrate to 1, i.e.

$$A(\theta) = \log\left[\int_{\mathcal{X}} \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x)dx\right].$$

The set of θs for which $A(\theta) < \infty$ constitute the natural parameter space.

Several distributions you have or will encounter are exponential family distributions (Wikipedia has a long list). We will do a couple of examples here.

Example 1: The Normal family of distributions has density,

$$p(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2}\right),$$

which is a 2-parameter exponential family, with natural parameters $(\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right)$, and sufficient statistics (x, x^2). One can verify that the natural parameter space is $\mathbb{R} \times (-\infty, 0)$.

Discrete distributions can similarly belong to an exponential family (you have to replace all the integrals with sums and so on).

Example 2: The Binomial distribution has pmf,

$$p(x;\theta) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0, 1, \ldots, n\}.$$

We can re-write this as:

$$p(x;\theta) = \binom{n}{x} \exp\left(x \log\frac{p}{1-p} + n\log(1-p)\right), \qquad x \in \{0, 1, \ldots, n\}.$$

This shows that it is in an exponential family with sufficient statistic x (number of successes), and natural parameter,

$$\theta = \log\left(\frac{p}{1-p}\right).$$

Example 3: The Poisson(λ) distribution has pmf,

$$p(x;\theta) = \frac{\exp(-\lambda)\lambda^x}{x!} = \frac{1}{x!}\exp(x\log\lambda - \lambda),$$

which shows that it is an exponential family with sufficient statistic x, and natural parameter $\theta = \log(\lambda)$.

Wikipedia has a long list of exponential family distributions, their natural parameters, sufficient statistics and other useful information. It is good practice to try to derive the natural parameters for some popular distributions.

2 Properties of Exponential Families

2.1 Random sampling

The exponential family structure is preserved for an i.i.d. sample, i.e. if $\{X_1, \ldots, X_n\}$ are i.i.d from some exponential family distribution $p(x;\theta)$ then the joint distribution:

$$p(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} h(x_i) \exp\left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta)\right],$$

is in an exponential family with the same natural parameters but with sufficient statistics:

$$T_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} T_i(x_j).$$

2.2 Log-partition generates moments

Recall that,

$$A(\theta) = \log \left[\int_{\mathcal{X}} \exp \left[\sum_{i=1}^{s} \theta_i T_i(x) \right] h(x) dx \right],$$

so taking the derivatives of A with respect to θ we obtain that,

$$\begin{aligned} \frac{\partial A(\theta)}{\partial \theta_i} &= \frac{\int_{\mathcal{X}} T_i(x) \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x) dx}{\left[\int_{\mathcal{X}} \exp\left[\sum_{i=1}^{s} \theta_i T_i(x)\right] h(x) dx\right]} \\ &= \mathbb{E}[T_i(X)]. \end{aligned}$$

You might wonder why we can switch derivatives and integrals - this is done rigorously using the dominated convergence theorem. Similarly, you can easily verify that higher derivatives lead to (functions of) higher moments (technically cumulants and not moments), i.e.

$$\frac{\partial^2 A(\theta)}{\partial \theta_i \partial \theta_j} = \mathbb{E}[(T_i(X) - \mathbb{E}[T_i(X)])(T_j(X) - \mathbb{E}[T_j(X)])] = \mathrm{cov}(T_i(X), T_j(X)).$$

This is why the function $A(\theta)$ is classically known as the cumulant function.

This latter property also reveals that A is a *convex function* of θ, i.e. it is bowl-shaped. Convexity is implied by the fact that the second-derivative matrix (i.e. the Hessian matrix) is positive semi-definite. For exponential families, the Hessian matrix is the covariance matrix of the sufficient statistics T_i, and covariance matrices are always positive semi-definite. Remember the conclusion: *A is a convex function of θ.*

2.3 The likelihood function in exponential families

When we observe a random sample $X_1, \ldots, X_n \sim p(x; \theta)$ from an exponential family distribution, the log-likelihood function is:

$$\mathcal{L}(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta) \right].$$

The log-likelihood function in an exponential family is *concave*. To see this compute the Hessian of $\mathcal{L}(\theta; x_1, \ldots, x_n)$ and observe that this is $-n$ times the Hessian of A. Since A is convex, its negation is concave.

2.4 Minimal representations and minimal sufficiency

An exponential family representation is said to be minimal if the sufficient statistics are not redundant, i.e. there is no set of coefficients $a \in \mathbb{R}^s, a \neq 0$ such that,

$$\sum_{i=1}^{s} a_i T_i(x) = \text{const},$$

for all $x \in \mathcal{X}$. If the representation is not minimal then essentially one can eliminate some of the sufficient statistics from the representation to obtain a minimal representation. Non-minimal exponential families are sometimes called *over-complete* exponential families. Over-complete exponential families are not statistically identifiable (while minimal ones are), i.e. there can be two different parameter vectors $\theta^1 \neq \theta^2$, such that, $p(X;\theta^1) = p(X;\theta^2)$. This effectively means, even if I gave you infinite data from the model, you cannot meaningfully estimate the parameter θ.

An exponential family where the space of allowed parameters θ_i is s-dimensional is called a full-rank family. On the other hand if there are relationships between the θ_i (for instance, $\theta_2 = \theta_1^2$) then the exponential family is *curved.*. For a full-rank exponential family, the sufficient statistics turn out to be minimal sufficient, i.e. the statistic

$$T(X_1, \ldots, X_n) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_s(X_i) \right),$$

is minimal sufficient.

We have been discussing the canonical parametrization of exponential families. It turns out that an equivalent way to parameterize the distribution is via what are called its mean parameters. We will not show this equivalence (it is not difficult) but rather just introduce the terminology here. Suppose we define:

$$\mu_i = \mathbb{E}[T_i(X)] = \int_{x \in \mathcal{X}} T_i(x) \exp\left[\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - A(\theta) \right] h(x) dx,$$

then it turns out that the collection $(\mu_1, \ldots, \mu_s)$ is in 1-1 correspondence with the natural parameters of the exponential family.

2.5 The maximum entropy duality

The classical motivation for exponential families comes from what is called the *principle of maximum entropy.* The idea is that, we suppose that we are given a random sample $\{X_1, \ldots, X_n\}$ from some distribution, and we compute the empirical expectations of certain

functions that we choose:

$$\widehat{\mu}_i = \frac{1}{n}\sum_{j=1}^{n} T_i(X_j) \quad \text{for} \;\; i \in \{1, \ldots, s\}.$$

For simplicity, you could imagine the case when $T(X) = (X, X^2, \ldots, X^s)$, i.e. where the statistics we are interested in are just moments, but everything we are discussing is much more general. Based on just these empirical expectations we want to infer a full probability distribution on the samples. A distribution p is *consistent* with the data we observe if it is the case that,

$$\widehat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{for} \;\; i \in \{1, \ldots, s\}.$$

We of course would like to pick a consistent distribution. It turns out that in most interesting cases, if we constrain a small number of statistics in this fashion there are infinitely many consistent distributions, so we need to come up with a way to choose between them.

The principle of maximum entropy suggests to pick the distribution that has the largest (Shannon) entropy. The entropy of a distribution is:

$$H(p) = -\int p(x)\log(p(x))dx.$$

Roughly, the entropy measures the complexity of a distribution (i.e. the average number of bits needed to encode samples from a distribution). The principle of maximum entropy says that one should be "maximally agnostic" about all aspects of the distribution that are not explicitly constrained. If this does not make sense, then just think about the principle as giving a possibly "natural" way to choose a distribution from a collection.

So we could imagine trying to find the distribution p^* that,

$$p^* = \arg\max_p H(p)$$

subject to the constraints that,

$$\widehat{\mu}_i = \mathbb{E}_p[T_i(X)] \quad \text{for} \;\; i \in \{1, \ldots, s\}.$$

The solution to this problem can shown to have the form

$$p^*(x) = \exp\left[\sum_{i=1}^{s} \theta_i T_i(x) - A(\theta)\right] h(x).$$

This provides a motivation for exponential families.

2.6 Bregman Divergences and KL Divergences

Given a (strictly) convex function A we can define a divergence between parameters by:

$$\rho(\theta_1, \theta_2) = A(\theta_2) - A(\theta_1) - \langle A(\theta_1),\, \theta_2 - \theta_1 \rangle.$$

For a pair of distributions we can define the KL divergence (assuming everything below is finite):

$$\mathrm{KL}(p, q) = \int p(x) \log(p(x)/q(x)) dx.$$

It is easy to see that for exponential families – the Bregman divergence between parameters (using the log-partition as the convex function) is exactly equal to the KL divergence between the corresponding distributions.

3 Parameter Estimation

We will discuss parameter estimation in great detail soon. But we introduce some of the ideas here in the context of exponential families.

One of the dominant strategies of parameter estimation is *maximum likelihood*: choose the estimate $\hat{\theta}$ to be the value of θ that maximizes the likelihood function. We have seen that the likelihood in an exponential family is concave and given by

$$\mathcal{L}(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta) \right],$$

so we can simply take the derivative with respect to θ and set this equal to 0. Using the facts we have seen earlier about the derivative of A, we can see that this amounts to solving the following system of equations for θ:

$$\mathbb{E}_\theta[T_i(X)] = \frac{1}{n} \sum_{j=1}^{n} T_i(x_j) \quad \text{for} \quad i \in \{1, \ldots, s\}$$

where

$$\mathbb{E}_\theta[T_i(X)] = \int T_i(x) p_\theta(x) dx.$$

So the maximum likelihood estimator simply picks the parameters θ to match the empirical expectations of the sufficient statistics to the expected value of the sufficient statistics under the distribution.

Usually we cannot compute this estimator in closed form so we use an iterative algorithm (like gradient ascent) to maximize the likelihood. Since exponential families have concave likelihoods this is usually a tractable problem.

Another way to estimate parameters of a distribution is known as the method of moments. Here the idea is to pick some statistics of the data, and the try to find parameters for your distribution so that the empirical average of the statistics are equal to their expected values under the estimated model. For exponential families as we can see above these two methods of estimation coincide.

Lecture Notes 13
36-705

Today we will discuss *point estimation*. Given $X_1, \ldots, X_n \sim p(X; \theta)$ we would like to construct an *estimator* $\widehat{\theta}(X_1, \ldots, X_n)$ of $\theta = (\theta_1, \ldots, \theta_k)$. An *estimator*

$$\widehat{\theta} = \widehat{\theta}_n = w(X_1, \ldots, X_n)$$

is any function of the data. Keep in mind that the parameter is a fixed, unknown constant. The estimator is a random variable.

In the next few lectures we will discuss ways to construct estimators and then we dicuss how to compare or evaluate them. The questions we are trying to answer are:

1. Are there general purpose methods to come up with estimators of θ?
2. Given two (or more) estimators is there a general framework in which we can compare estimators?
3. How do we analyze complex estimators (say estimators that are not simple averages)?

An "estimator" refers to a random variable (a statistic, a function of the sample) and an "estimate" refers to its realized value. We have already studied estimation in a relatively simple context: estimating the mean. When θ is not a mean then we need to think a bit harder to decide how to estimate it. We will focus on general purpose methods for estimation.

For now, we will discuss three methods of constructing estimators:

1. The Method of Moments (MOM)
2. Maximum likelihood (MLE)
3. Bayes estimators.

Some Terminology. Throughout these notes, we will use the following terminology:

1. $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1; \theta) \cdots p(x_n; \theta) dx_1 \cdots dx_n$.
2. Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$.
3. The distribution of $\widehat{\theta}_n$ is called its *sampling distribution*.
4. The standard deviation of $\widehat{\theta}_n$ is called the *standard error* denoted by $\text{se}(\widehat{\theta}_n)$.
5. $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \xrightarrow{p} \theta$. Later we will see that if bias $\to 0$ and $\text{Var}(\widehat{\theta}_n) \to 0$ as $n \to \infty$ then $\widehat{\theta}_n$ is consistent.

1 The Method of Moments

Suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Define

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, \qquad & \mu_1(\theta) &= \mathbb{E}(X_i) = \int x p_\theta(x) dx \\ m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, \qquad & \mu_2(\theta) &= \mathbb{E}(X_i^2) = \int x^2 p_\theta(x) dx \\ & \vdots \quad \vdots & & \\ m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, \qquad & \mu_k(\theta) &= \mathbb{E}(X_i^k) = \int x^k p_\theta(x) dx. \end{aligned}$$

Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ solve:

$$m_j = \mu_j(\widehat{\theta}), \quad j = 1, \ldots, k.$$

In other words, we equate the first k sample moments with the first k theoretical moments. This defines k equations with k unknowns.

Example 1 *$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$. Then $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$. Equate:*

$$\frac{1}{n}\sum_{i=1}^n X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

to get

$$\widehat{\beta} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Example 2 *Suppose*

$$X_1, \ldots, X_n \sim \text{Binomial}(k, p)$$

where both k and p are unknown. We get

$$kp = \overline{X}_n, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = kp(1-p) + k^2 p^2$$

giving

$$\widehat{p} = \frac{\overline{X}_n}{k}, \quad \widehat{k} = \frac{\overline{X}_n^2}{\overline{X}_n - \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2}.$$

The method of moments was popular many years ago because it is often easy to compute. Lately, it has attracted attention again. For example, there is a large literature on estimating mixtures of Gaussians using the method of moments.

2 Maximum Likelihood

The most popular method for estimating parameters is maximum likelihood. One of the reasons is that, under certain conditions, the maximum likelihood estimator is optimal. We'll discuss optimality later.

The maximum likelihood estimator (mle) $\widehat{\theta}$ is defined as the maximizer of

$$\mathcal{L}(\theta) = p(X_1, \ldots, X_n; \theta) \stackrel{iid}{=} \prod_i p(X_i; \theta).$$

This is the same as maximizing the log-likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta).$$

Often it suffices to solve

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, k.$$

Example 3 *Binomial.* $\mathcal{L}(p) = \prod_i p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$ *where* $S = \sum_i X_i$. *So*

$$\ell(p) = S \log p + (n - S) \log(1 - p)$$

and $\widehat{p} = \overline{X}_n$.

Example 4 $X_1, \ldots, X_n \sim N(\mu, 1)$.

$$\mathcal{L}(\mu) \propto \prod_i e^{-(X_i-\mu)^2/2} \propto e^{-n(\overline{X}_n-\mu)^2}, \quad \ell(\mu) = -\frac{n}{2}(\overline{X}_n - \mu)^2$$

and $\widehat{\mu} = \overline{X}_n$. *For* $N(\mu, \sigma^2)$ *we have*

$$\mathcal{L}(\mu, \sigma^2) \propto \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right\}$$

and

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

Set

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \sigma^2} = 0$$

to get

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

Example 5 *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Then*

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} I(\theta > X_{(n)})$$

and so $\widehat{\theta} = X_{(n)}$.

What is the method of moments estimator above? How would you compare the two estimators?

2.1 MLE and MoM for exponential families

The log-likelihood in an exponential family is concave and given by

$$\ell(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta)\right],$$

so we can simply take the derivative with respect to θ and set this equal to 0. Using the facts we have seen in the last lecture about the derivative of A, we can see that this amounts to solving the following system of equations for θ:

$$\mathbb{E}_{p(X;\theta)}[T_i(X)] = \frac{1}{n}\sum_{j=1}^{n} T_i(x_j) \quad \text{for} \quad i \in \{1, \ldots, s\}.$$

So the maximum likelihood estimator simply picks the parameters θ to match the empirical expectations of the sufficient statistics to the expected value of the sufficient statistics under the distribution.

Usually we cannot compute this estimator in closed form so we use an iterative algorithm (like gradient ascent) to maximize the likelihood. However, you should remember that exponential families have concave likelihoods so this is usually tractable. For exponential families as we can see above the method of moments coincides with the MLE (if we chose the sufficient statistics to direct which moments to compute).

Suppose that $\theta = (\eta, \xi)$. The *profile likelihood* for η is defined by

$$\mathcal{L}(\eta) = \sup_{\xi} \mathcal{L}(\eta, \xi).$$

To find the mle of η we can proceed in two ways. We could find the overall mle $\widehat{\theta} = (\widehat{\eta}, \widehat{\xi})$. The mle for η is just the first coordinate of $(\widehat{\eta}, \widehat{\xi})$. Alternatively, we could find the maximizer of the profile likelihood. These give the same answer. Do you see why?

2.2 Equivariance and the profile likelihood

The mle is *equivariant*. if $\eta = g(\theta)$ then $\widehat{\eta} = g(\widehat{\theta})$. Suppose g is invertible so $\eta = g(\theta)$ and $\theta = g^{-1}(\eta)$. Define $\mathcal{L}^*(\eta) = \mathcal{L}(\theta)$ where $\theta = g^{-1}(\eta)$. So, for any η,

$$\mathcal{L}^*(\widehat{\eta}) = \mathcal{L}(\widehat{\theta}) \geq \mathcal{L}(\theta) = \mathcal{L}^*(\eta)$$

and hence $\widehat{\eta} = g(\widehat{\theta})$ maximizes $\mathcal{L}^*(\eta)$. For non invertible functions this is still true if we define $\mathcal{L}^*(\eta)$ to be the profile likelihood.

Example 6 *Binomial. The mle is* $\widehat{p} = \overline{X}_n$. *Let* $\psi = \log(p/(1-p))$. *Then* $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$.

3 Bayes Estimator

To define the Bayes estimator, we begin by treating θ as a random variable. This point requires much discussion (which we will have later). For now, just tentatively think of θ as random. We start with a *prior distribution* $p(\theta)$ on θ. Note that

$$p(x_1, \ldots, x_n|\theta)p(\theta) = p(x_1, \ldots, x_n, \theta).$$

Now compute the *posterior distribution* by Bayes' theorem:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}$$

where

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta.$$

This can be written as

$$p(\theta|x_1, \ldots, x_n) \propto \mathcal{L}(\theta)p(\theta) = \text{Likelihood } \times \text{ prior}.$$

Now compute a point estimator from the posterior. For example:

$$\widehat{\theta} = \mathbb{E}(\theta|x_1, \ldots, x_n) = \int \theta p(\theta|x_1, \ldots, x_n)d\theta = \frac{\int \theta p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}.$$

Example 7 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. *Let the prior be* $\theta \sim \text{Beta}(\alpha, \beta)$. *Hence*

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

and

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Set $Y = \sum_i X_i$. *Then*

$$p(\theta|X) \propto \underbrace{\theta^Y(1-\theta)^{n-Y}}_{\text{likelihood}} \times \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \propto \theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}.$$

Therefore, $\theta|X \sim \text{Beta}(Y+\alpha, n-Y+\beta)$. *The Bayes estimator is*

$$\widetilde{\theta} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{\theta}_{mle} + \lambda\ \overline{\theta}$$

where

$$\overline{\theta} = \frac{\alpha}{\alpha+\beta}, \quad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

This is an example of a conjugate prior.

Example 8 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *with* σ^2 *known. Let* $\mu \sim N(m, \tau^2)$. *Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}}\overline{X}_n + \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}m$$

and

$$\text{Var}(\mu|X) = \frac{\sigma^2\tau^2/n}{\tau^2+\frac{\sigma^2}{n}}.$$

4 MSE

Now we discuss the evaluation of estimators. The mean squared error (MSE) is

$$\mathbb{E}_\theta(\widehat{\theta}-\theta)^2 = \int\cdots\int(\widehat{\theta}(x_1,\ldots,x_n)-\theta)^2 p(x_1;\theta)\cdots p(x_n;\theta)dx_1\ldots dx_n.$$

The bias is

$$B = \mathbb{E}_\theta(\widehat{\theta}) - \theta$$

and the variance is

$$V = \text{Var}_\theta(\widehat{\theta}).$$

Theorem 9 *We have*

$$MSE = B^2 + V.$$

Proof: Let $m = \mathbb{E}_\theta(\widehat{\theta})$. Then

$$\begin{aligned} MSE &= \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathbb{E}_\theta(\widehat{\theta} - m + m - \theta)^2 \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 + 2\mathbb{E}_\theta(\widehat{\theta} - m)(m - \theta) \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 = V + B^2. \end{aligned}$$

■

An estimator is *unbiased* if the bias is 0. In that case, the MSE = Variance. There is often a tradeoff between bias and variance. So low bias can imply high variance and vice versa.

Example 10 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. *Then*

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathbb{E}(S^2) = \sigma^2.$$

The MSE's are

$$\mathbb{E}(\overline{X} - \mu)^2 = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}.$$

It is worth thinking about how one defines the MSE when θ is multivariate (as in the example above), and what the analogous bias-variance decomposition is.

We would like to choose an estimator with small MSE. However, the MSE is a function of θ. Later, we shall discuss minimax estimators, that use the maximum of the MSE over θ as a way to compare estimators.

5 Unbiased estimators, Fisher Information, Cramér-Rao

In the olden days, many people focused on unbiased estimators. More modern treatments do not often emphasize this point of view since there are many known examples where a small amount of bias can result in large reductions in variance.

With that said, there are still pieces of this classical theory that are useful. One of the important pieces is the Cramér-Rao bound which provides a lower bound on the variance of an unbiased estimator. In many problems, this bound will provide some at least heuristic guidelines into the difficulty of an estimation problem. Later on in the course we will talk about other ways of proving lower bounds that do not restrict attention to unbiased estimators (i.e. we will discuss what are called minimax lower bounds).

5.1 Fisher Information

The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(X_i; \theta).$$

We can also define the gradient of this function, which is called the *score function*:

$$s(\theta) = s(\theta; X_1, \ldots, X_n) = \nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \nabla_\theta \log p(X_i; \theta).$$

This gradient is a d-dimensional vector. The Fisher Information matrix is the expected outer product of the score, i.e.:

$$I(\theta) = \mathbb{E}[s(\theta)s(\theta)^T].$$

The Fisher information matrix is a $d \times d$ matrix. Lets take a quick look at a couple of examples:

Example 1: Suppose that $X \sim \text{Ber}(p)$, then the log-likelihood is given by:

$$\ell(p) = X \log(p) + (1 - X) \log(1 - p),$$

and the score is:

$$s(p) = \frac{X}{p} - \frac{1 - X}{1 - p} = \frac{X - p}{p(1 - p)}.$$

We can then compute the Fisher information:

$$I(p) = \frac{1}{p^2(1 - p)^2} \mathbb{E}\left[(X - p)^2\right] = \frac{1}{p(1 - p)}.$$

Example 2: Suppose that $X \sim N(\mu, \sigma^2)$ where σ is known, then the log-likelihood is given by:

$$\ell(\mu) = -\frac{1}{2\sigma^2}(X - \mu)^2,$$

so that the score is:

$$s(\mu) = \frac{X - \mu}{\sigma^2},$$

and the Fisher information is:

$$I(\mu) = \mathbb{E}\left[\frac{(X - \mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

An important property that we will use is that the score function has mean zero, i.e.

$$\mathbb{E}_\theta[s(\theta)] = \int \cdots \int s(\theta; x_1, \ldots, x_n) p_\theta(x_1, \ldots, x_n) dx_1 \cdots dx_n = 0.$$

Proof: Notice that,

$$\begin{aligned}
\mathbb{E}_\theta[s(\theta)] &= \sum_{i=1}^{n} \int \nabla_\theta \log p(x_i; \theta) \; n p(x_1, \ldots, x_n; \theta) dx_1 dx_2 \ldots dx_n \\
&= \sum_{i=1}^{n} \int \nabla_\theta \log p(x_i; \theta) \;\; p(x_i; \theta) dx_i \\
&= n \int \nabla_\theta \log p(x_1; \theta) p(x_1; \theta) dx_1,
\end{aligned}$$

using the i.i.d. assumption several times. Under some regularity conditions we can switch the derivative and integral so we obtain,

$$\begin{aligned}
\int \nabla_\theta \log p(x_1; \theta) \; p(x_1; \theta) dx_1 &= \int \frac{\nabla_\theta p(x_1; \theta)}{p(x_1; \theta)} p(x_1; \theta) dx_1 \\
&= \nabla_\theta \int p(x_1; \theta) dx_1 = \nabla_\theta 1 = 0. \;\; \square
\end{aligned}$$

One consequence of this property is that we can interpret the Fisher information matrix as the covariance matrix of the score, i.e.

$$I(\theta) = \mathbb{E}\Big[(s(\theta) - \mathbb{E}(s(\theta)))(s(\theta) - \mathbb{E}(s(\theta)))^T\Big].$$

You should check that the following holds:

$$I(\theta) = n I_1(\theta)$$

where $I_1(\theta)$ is the Fisher information based on one observation.

Lemma 11 *The Fisher information satisfies*

$$I_1(\theta) = -\mathbb{E}\left[\nabla_\theta^2 \log p(X; \theta)\right].$$

Proof. To see this observe that,

$$\begin{aligned}
\nabla_\theta^2 \log p(X; \theta) &= \nabla_\theta \frac{\nabla_\theta p(X; \theta)}{p(X; \theta)} \\
&= \frac{\nabla_\theta^2 p(X; \theta)}{p(X; \theta)} - \frac{(\nabla_\theta p(X; \theta) \nabla_\theta p(X; \theta)^T)}{p(X; \theta)^2} \\
&= \frac{\nabla_\theta^2 p(X; \theta)}{p(X; \theta)} - s(\theta) s(\theta)^T.
\end{aligned}$$

Now, notice that,

$$\mathbb{E}\left[\frac{\nabla_\theta^2 p(X;\theta)}{p(X;\theta)}\right] = \int \nabla_\theta^2 p(X;\theta) = \nabla_\theta^2 \int p(X;\theta) = \nabla_\theta^2 1 = 0,$$

which yields the result. □

The Fisher information is measuring the expected curvature of the log-likelihood function around the point θ. As we will see in future lectures if the log-likelihood is more curved (i.e. $I(\theta)$ is appropriately "large") then θ is easier to estimate.

Example 3: Let $X_1, \ldots, X_n \sim \text{Ber}(p)$. Then

$$I(p) = \frac{n}{p(1-p)}.$$

Example 4: For exponential families we have seen that the log-likelihood is given as:

$$\ell(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(X_j) - nA(\theta),$$

so the Hessian is:

$$I(\theta) = n\nabla_\theta^2 A(\theta) = n\mathbb{E}\Big[(T(X) - \mathbb{E}[T(X)])(T(X) - \mathbb{E}[T(X)])^T\Big],$$

so the Fisher information matrix is n times the Hessian of the log-partition function or alternatively it is the covariance matrix of the vector of sufficient statistics.

5.2 Cramér-Rao Bound

Let us briefly consider again the Bernoulli example. We observe $X_1, \ldots, X_n \sim \text{Ber}(p)$ and estimate $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. The estimator is unbiased and has variance $p(1-p)/n$ which is precisely the inverse of the Fisher information. This turns out to be a fairly general phenomenon. Indeed, the Cramér-Rao bound assures us that this estimator is unimprovable in a certain sense. We focus first on the univariate case (when $\theta \in \mathbb{R}$) and then consider the multivariate extension.

Cramér-Rao Bound: Suppose that $X_1, \ldots, X_n \sim p(X;\theta)$ and that $\widehat{\theta}$ is an unbiased estimator of θ. Then

$$\text{Var}(\widehat{\theta}) \geq \frac{1}{nI_1(\theta)}.$$

Proof: Consider that,

$$\text{cov}(\widehat{\theta}, s(\theta)) = \mathbb{E}((\widehat{\theta} - \theta)s(\theta)) = \mathbb{E}(\widehat{\theta}s(\theta)),$$

since $\mathbb{E}[s(\theta)] = 0$. Furthermore,

$$\begin{aligned}
\mathbb{E}(\widehat{\theta}s(\theta)) &= \int \widehat{\theta}(x_1, \ldots, x_n) \nabla_\theta \log p(x_1, \ldots, x_n; \theta) p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n \\
&= \int \widehat{\theta}(x_1, \ldots, x_n) \frac{\nabla_\theta p(x_1, \ldots, x_n; \theta)}{p(x_1, \ldots, x_n; \theta)} p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n \\
&= \int \widehat{\theta}(x_1, \ldots, n) \ \nabla_\theta p(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n \\
&= \nabla_\theta \int \widehat{\theta}(x_1, \ldots, n) \ p(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n = \nabla_\theta \ \theta = 1.
\end{aligned}$$

Notice that for any fixed ζ we can write:

$$\text{Var}(\widehat{\theta} - \zeta s(\theta)) = \text{Var}(\widehat{\theta}) + \zeta^2 \text{Var}(s(\theta)) - 2\zeta \text{cov}(\widehat{\theta}, s(\theta)) = \text{Var}(\widehat{\theta}) + \zeta^2 n I_1(\theta) - 2\zeta.$$

Using the fact that variances are positive we can write:

$$\text{Var}(\widehat{\theta}) \geq 2\zeta - \zeta^2 n I_1(\theta)$$

Take $\zeta = 1/(nI_1(\theta))$ to obtain the Cramér-Rao bound.

Multivariate Generalization: The Cramér-Rao bound can be derived for a multivariate parameter θ in a very similar fashion. For any two positive semi-definite matrices A and B write

$$A \succeq B$$

to mean that

$$v^T A v \geq v^T B v$$

for any vector v. The multivariate Cramer-Rao bound is

$$\text{Var}(\widehat{\theta}) \succeq I(\theta)^{-1} = \frac{1}{n} I_1(\theta)^{-1}.$$

Examples: In both the Gaussian and Bernoulli models, as a consequence of the Cramér-Rao bound we can conclude that the MLE is the best unbiased estimator.

6 Beyond unbiased estimators: Decision theory

The central idea in decision theory is that we want to minimize our *expected* loss. Let $X_1, \ldots, X_n \sim p(X; \theta)$, with $\theta \in \Theta$. We choose a loss function $L(a, \theta)$ which measures how far a point a is from θ. Some common loss functions are:

1. **Squared loss:** $L(a, \theta) = (a - \theta)^2$.
2. **Absolute loss:** $L(a, \theta) = |a - \theta|$.

There are however many other loss functions. For instance, we sometimes consider losses like:

$$L(a, \theta) = \frac{(a - \theta)^2}{|\theta| + 1},$$

which penalizes errors in estimation more for small values of θ than for large values. We can similarly design a loss function that penalizes errors more strongly for large values of θ.

Another important point is that there are cases when we do not really care about estimating the parameter well but rather just the distribution $p(x; \theta)$. This is true when we care about prediction in regression or in density estimation. In this case we could define the loss between θ and a in terms of the distributions $p(x; \theta)$ and $p(x; a)$. One canonical example is:

Kullback-Leibler loss:

$$L(a, \theta) = \mathrm{KL}(p(x; \theta), p(x; a)) = \int p(x; \theta) \log \left(\frac{p(x; \theta)}{p(x; a)} \right) dx.$$

Once we have a loss function, and an estimator, we can assess the estimator via its expected loss. This expected loss is called the *risk* of the estimator. Suppose we consider an estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$; the risk is

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta L(\widehat{\theta}, \theta).$$

Ideally, we would like to find an estimator $\widehat{\theta}$ such that for any other estimator θ' we have that:

$$R(\theta, \widehat{\theta}(X)) \leq R(\theta, \theta')$$

for all values θ. Such estimators will most often not exist – why not? So we will need to find another way to define an optimal estimator.

Lecture Notes 14
36-705

We continue with our discussion of decision theory.

1 Decision Theory

Suppose we want to estimate a parameter θ using data $X^n = (X_1, \ldots, X_n)$. What is the best possible estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ? Decision theory provides a framework for answering this question.

1.1 The Risk Function

Let $\widehat{\theta} = \widehat{\theta}(X^n)$ be an estimator for the parameter $\theta \in \Theta$. We start with a **loss function** $L(\theta, \widehat{\theta})$ that measures how good the estimator is. For example:

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$	L_p loss,
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$	zero–one loss,
$L(\theta, \widehat{\theta}) = I(\lvert\widehat{\theta} - \theta\rvert > c)$	large deviation loss,
$L(\theta, \widehat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\widehat{\theta})}\right) p(x;\theta)dx$	Kullback–Leibler loss.

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss functions are

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|^2 = \sum_{j=1}^{k} (\widehat{\theta}_j - \theta_j)^2,$$

$$L(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_p = \left(\sum_{j=1}^{k} |\widehat{\theta}_j - \theta_j|^p\right)^{1/p}.$$

When the problem is to predict a $Y \in \{0, 1\}$ based on some classifier $h(x)$ a commonly used loss is

$$L(Y, h(X)) = I(Y \neq h(X)).$$

For real valued prediction a common loss function is

$$L(Y, \widehat{Y}) = (Y - \widehat{Y})^2.$$

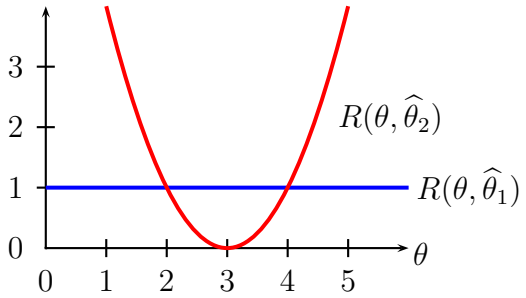


Figure 1: *Comparing two risk functions. Neither risk function dominates the other at all values of* θ.

The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta, \widehat{\theta})\Big) = \int L(\theta, \widehat{\theta}(x_1, \ldots, x_n))p(x_1, \ldots, x_n; \theta)dx. \tag{1}$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2. \tag{2}$$

If we do not state what loss function we are using, assume the loss function is squared error.

1.2 Comparing Risk Functions

To compare two estimators, we compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

Example 1 *Let* $X \sim N(\theta, 1)$ *and assume we are using squared error loss. Consider two estimators:* $\widehat{\theta}_1 = X$ *and* $\widehat{\theta}_2 = 3$. *The risk functions are* $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ *and* $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. *If* $2 < \theta < 4$ *then* $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, *otherwise,* $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. *Neither estimator uniformly dominates the other; see Figure 1.*

Example 2 *Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has zero bias, we have that*

$$R(p, \widehat{p}_1) = \mathsf{Var}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_{i=1}^n X_i$ and α and β are positive constants.[1] *Now,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathsf{Var}_p(\widehat{p}_2) + (\mathrm{bias}_p(\widehat{p}_2))^2 \\ &= \mathsf{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

Let $\alpha = \beta = \sqrt{n/4}$. The resulting estimator is

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

The risk functions are plotted in Figure 2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

The **maximum risk** is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \tag{3}$$

and the **Bayes risk** under prior π is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta}) \pi(\theta) d\theta. \tag{4}$$

Example 3 *Consider again the two estimators in Example 2. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

[1] This is the posterior mean using a Beta (α, β) prior.

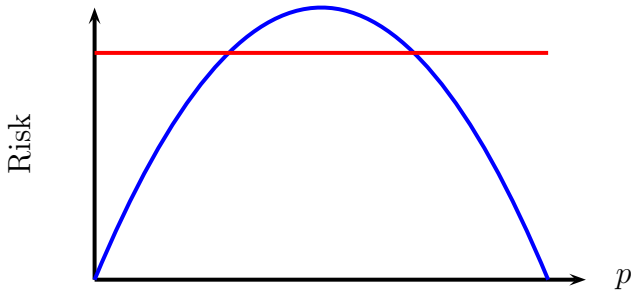


Figure 2: Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 2. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

and

$$\overline{R}(\widehat{p}_2) = \max_p \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like the maximum risk are imperfect.

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

An estimator $\widehat{\theta}$ that minimizes the Bayes risk is called a **Bayes estimator**. That is,

$$B_\pi(\widehat{\theta}) = \inf_{\tilde{\theta}} B_\pi(\tilde{\theta}) \tag{5}$$

where the infimum is over all estimators $\tilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax estimator**. That is,

$$\sup_\theta R(\theta, \widehat{\theta}) = \inf_{\tilde{\theta}} \sup_\theta R(\theta, \tilde{\theta}) \tag{6}$$

where the infimum is over all estimators $\tilde{\theta}$. We call the right hand side of (6), namely,

$$R_n \equiv R_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}), \tag{7}$$

the **minimax risk**. Statistical decision theory has two main goals: determine the minimax risk R_n and find an estimator that achieves this risk.

Once we have found the minimax risk R_n we want to find the minimax estimator that achieves this risk:

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\theta,\widehat{\theta}). \tag{8}$$

1.3 Bayes Estimators

Let π be a prior distribution. After observing $X^n = (X_1,\ldots,X_n)$, the posterior distribution is, according to Bayes' theorem,

$$\mathbb{P}(\theta\in A|X^n) = \frac{\int_A p(X_1,\ldots,X_n|\theta)\pi(\theta)d\theta}{\int_\Theta p(X_1,\ldots,X_n|\theta)\pi(\theta)d\theta} = \frac{\int_A \mathcal{L}(\theta)\pi(\theta)d\theta}{\int_\Theta \mathcal{L}(\theta)\pi(\theta)d\theta} \tag{9}$$

where $\mathcal{L}(\theta) = p(x^n;\theta)$ is the likelihood function. The posterior has density

$$\pi(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)} \tag{10}$$

where $m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of X^n. Define the **posterior risk** of an estimator $\widehat{\theta}(x^n)$ by

$$r(\widehat{\theta}|x^n) = \int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta. \tag{11}$$

Theorem 4 *The Bayes risk $B_\pi(\widehat{\theta})$ satisfies*

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \tag{12}$$

Let $\widehat{\theta}(x^n)$ be the value of θ that minimizes $r(\widehat{\theta}|x^n)$. Then $\widehat{\theta}$ is the Bayes estimator.

Proof:

Let $p(x,\theta) = p(x|\theta)\pi(\theta)$ denote the joint density of X and θ. We can rewrite the Bayes risk as follows:

$$\begin{aligned}
B_\pi(\widehat{\theta}) &= \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x^n))p(x|\theta)dx^n\right)\pi(\theta)d\theta \\
&= \int\int L(\theta,\widehat{\theta}(x^n))p(x,\theta)dx^n d\theta = \int\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\
&= \int\left(\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta\right)m(x^n)\,dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n.
\end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of θ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

Theorem 5 *If $L(\theta,\widehat{\theta}) = (\theta-\widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x^n) = \int \theta\pi(\theta|x^n)d\theta = \mathbb{E}(\theta|X=x^n). \tag{13}$$

If $L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}|$ then the Bayes estimator is the median of the posterior $\pi(\theta|x^n)$. If $L(\theta,\widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $\pi(\theta|x^n)$.

Proof:

We will prove the theorem for squared error loss. The Bayes estimator $\widehat{\theta}(x^n)$ minimizes $r(\widehat{\theta}|x^n) = \int(\theta-\widehat{\theta}(x^n))^2\pi(\theta|x^n)d\theta$. Taking the derivative of $r(\widehat{\theta}|x^n)$ with respect to $\widehat{\theta}(x^n)$ and setting it equal to zero yields the equation $2\int(\theta-\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta = 0$. Solving for $\widehat{\theta}(x^n)$ we get 13.

Example 6 *Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ where σ^2 is known. Suppose we use a $N(a,b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is*

$$\widehat{\theta}(X_1,\ldots,X_n) = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a. \tag{14}$$

It is worth keeping in mind the trade-off: Bayes estimators although easy to compute are very subjective; they depend strongly on the prior π. Minimax estimators, although more challenging to compute are not subjective, but do have the drawback that they are protecting against the worst-case which might lead to pessimistic conclusions.

2 Minimax Estimators through Bayes Estimators

Our goal is to compute a minimax estimator $\widehat{\theta}$ that satisfies:

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) \leq \inf_{\widetilde{\theta}} \sup_{\theta\in\Theta} R(\theta,\widetilde{\theta}).$$

We will let θ_{minimax} denote a minimax estimator.

2.1 Bounding the Minimax Risk

One strategy to find the minimax estimator is by finding (upper and lower) bounds on the minimax risk that match. Then the estimator that achieves the upper bound is a minimax estimator.

Upper bounding the minimax risk is straightforward. Given an estimator $\widehat{\theta}_{\text{up}}$ we can compute its maximum risk and use it to upper bound the minimax risk, i.e.

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}_{\text{up}}).$$

The Bayes risk of the Bayes estimator for any prior π lower bounds the minimax risk. Fix a prior π and suppose that $\widehat{\theta}_{\text{low}}$ is the Bayes estimator with respect to π, then we have that:

$$B_\pi(\widehat{\theta}_{\text{low}}) \leq B_\pi(\theta_{\text{minimax}}) \leq \sup_{\theta} R(\theta, \theta_{\text{minimax}}) = \inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}).$$

Let us see an example of this in action.

Example: We will prove a classical result that if we observe independent draws from a d-dimensional Gaussian, $X_1, \ldots, X_n \sim N(\theta, I_d)$, then the average:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is a minimax estimator of θ with respect to the squared loss.

Let R_n denote the minimax risk. First, let us compute the upper bound on R_n. We note that,

$$\widehat{\theta} \sim N(\theta, I_d/n),$$

so that its risk:

$$R(\theta, \widehat{\theta}) = \mathbb{E}[\sum_{i=1}^{d} (\widehat{\theta}_i - \theta_i)^2] = \mathbb{E}[\sum_{i=1}^{d} Z_i^2],$$

where $Z_i \sim N(0, 1/n)$. This yields that,

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}) = \frac{d}{n}.$$

Now we lower bound the minimax risk using the Bayes risk. Let us take the prior to be zero-mean Gaussian, i.e. we take $\pi = N(0, c^2 I_d)$. By sufficiency, we can replace the data with $\widehat{\theta}$. We can write:

$$\theta \sim N(0, c^2 I_d)$$
$$\widehat{\theta}|\theta \sim N(\theta, I_d/n).$$

We can write this as

$$\theta = c\epsilon$$
$$\widehat{\theta} = \frac{1}{\sqrt{n}} Z$$

where $\epsilon, Z \sim N(0, I_d)$. Hence,

$$\begin{pmatrix} \theta \\ \widehat{\theta} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} c^2 I_d & c^2 I_d \\ c^2 I_d & (c^2 + 1/n) I_d \end{bmatrix} \right]$$

We can now compute the posterior (using standard conditional Gaussian formulae), and obtain its mean:

$$\mathbb{E}[\theta|\widehat{\theta}] = \frac{c^2}{c^2 + 1/n}\widehat{\theta}.$$

Now,

$$R(\theta, \widehat{\theta}) = \mathbb{E}\left|\left|\frac{c^2}{c^2 + 1/n}\widehat{\theta} - \theta\right|\right|^2.$$

Write $\widehat{\theta} = \theta + W$, where $W \sim N(0, I_d/n)$. Then

$$R(\theta, \widehat{\theta}) = \mathbb{E}_W\left|\left|\frac{c^2}{c^2 + 1/n} Z - \frac{\theta}{n(c^2 + 1/n)}\right|\right|^2.$$

Let us denote $\beta := c^2 + 1/n$. Then we obtain that,

$$R(\theta, \widehat{\theta}) = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\mathbb{E}\|W\|_2^2 = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n}.$$

The Bayes risk further averages this over $\theta \sim N(0, c^2 I_d)$ to obtain that,

$$B_\pi\left(\frac{c^2}{c^2 + 1/n}\widehat{\theta}\right) = \frac{c^2 d}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n} = \frac{c^2 d}{n\beta} = \frac{d}{n(1 + 1/(nc^2))}.$$

We conclude that

$$\frac{d}{n(1 + 1/(nc^2))} \leq R_n \leq \frac{d}{n}.$$

This is true for every $c > 0$. Since c was arbitrary we can take the limit as $c \to \infty$ to obtain that the minimax risk is upper and lower bounded by d/n and hence, $R_n = d/n$ and the sample average $\widehat{\theta}$ is minimax.

2.2 Least Favorable Prior

The other way to obtain Bayes estimators is by constructing what are called least favorable priors.

Theorem 7 *Let $\widehat{\theta}$ be the Bayes estimator for some prior π. If*

$$R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \text{ for all } \theta \tag{15}$$

then $\widehat{\theta}$ is minimax and π is called a **least favorable prior**.

Proof:

Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{16}$$

which is a contradiction.

Theorem 8 *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\widehat{\theta}$ is minimax.*

Proof:

The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ. Now apply the previous theorem.

Example 9 *Consider the Bernoulli model with squared error loss. We showed previously that the estimator*

$$\widehat{p} = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior $\text{Beta}(\alpha, \beta)$ *with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.*

Lecture Notes 15
36-705

1 Asymptotic theory

This lecture and the next will focus on asymptotic theory for the MLE. We suppose that we obtain a sample $X_1, \ldots, X_n \sim p(X; \theta)$ and are interested in estimating θ. We are interested in two questions:

1. **Consistency:** Does the MLE converge in probability to θ, i.e. does $\widehat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$? This is analogous to the LLN.

2. **Asymptotic distribution:** What can we say about the distribution of $\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta)$? This is analogous to the CLT.

We will begin with the question of consistency.

2 Consistency of the MLE

The main take-home from this section is that under somewhat mild conditions the MLE is a consistent estimator. We will try to develop the necessary conditions and build some intuition about the MLE and about what consistency entails.

2.1 MLE as Empirical Risk Minimization

We have discussed previously the idea of empirical risk minimization, where we construct an estimator by minimizing an empirical estimate of the risk. We looked at the particular case of classification with the 0/1 loss. The MLE can be viewed as a special case of ERM with a different loss function.

Suppose we define the loss function:

$$R_n(\widehat{\theta}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i; \theta)}{p(X_i; \widehat{\theta})}.$$

Observe that minimizing this loss function is identical to maximizing the likelihood. Notice that we introduced an extra $p(X_i; \theta)$ term but this does not affect anything. Of course, if

this is the empirical risk it is natural to wonder what the associated population risk is. This is

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \log \frac{p(X;\theta)}{p(X;\widehat{\theta})} = \int p(x;\theta) \log \left(\frac{p(x;\theta)}{p(x;\widehat{\theta})} \right) dx$$

which is the Kullback-Leibler divergence, i.e. the population risk is the KL divergence $\mathrm{KL}(p(x;\theta) \| p(x;\widehat{\theta}))$. Notice that, the empirical risk is a sum of i.i.d terms so by the LLN we have that for any fixed $\widetilde{\theta}$

$$R_n(\widetilde{\theta}, \theta) \xrightarrow{p} R(\widetilde{\theta}, \theta).$$

To analyze empirical risk minimization we needed a *uniform* LLN and we will need exactly this to show consistency. An important property of the KL divergence is that it is zero iff $p(X;\theta) = p(X;\widehat{\theta})$ almost everywhere (i.e. they are equal except on sets of measure 0). The main thing to remember is the connection between MLE and KL divergence.

2.2 Conditions for consistency

Condition 1: Identifiability: A basic requirement for constructing any consistent estimator is that the model be identifiable, i.e. if $\theta_1 \neq \theta_2$ then it must be the case that $p(x;\theta_1) \neq p(x;\theta_2)$.

We will in general require something slightly stronger than this:

Condition 2: Strong identifiability: We assume that for every $\epsilon > 0$

$$\inf_{\widetilde{\theta}:|\widetilde{\theta}-\theta|\geq\epsilon} \mathrm{KL}(p(x;\theta), p(x;\widetilde{\theta})) > 0.$$

This condition is essentially the same as Condition 1, except that it does not allow the difference between the two distributions to be vanishingly small. The two conditions are equivalent if θ is restricted to lie in a compact set.

Condition 3: Uniform LLN: Assume that,

$$\sup_{\widetilde{\theta}} |R_n(\widetilde{\theta}, \theta) - R(\widetilde{\theta}, \theta)| \xrightarrow{p} 0.$$

This condition is a uniform LLN. As we have seen before it holds for instance if the Rademacher complexity of the class of functions of the form: $f_{\widetilde{\theta}}(X) = \log p(X;\widetilde{\theta})/p(X;\theta)$ is not too large.

Theorem 1 *Suppose that Conditions 2 and 3 above hold, then the MLE is consistent.*

Proof: Fix an $\epsilon > 0$. Using the strong identifiability condition we see that for every $\epsilon > 0$, we have that there is an $\eta > 0$ such that,

$$\mathrm{KL}(p(x;\theta), p(x;\widetilde{\theta})) \geq \eta,$$

if $|\widetilde{\theta} - \theta| \geq \epsilon$. We will show that for the MLE $\widehat{\theta}$, we have that $\mathrm{KL}(p(x;\theta)\|p(x;\widehat{\theta})) \leq \eta$, as $n \to \infty$ in probability. This in turn implies that $|\widehat{\theta} - \theta| \leq \epsilon$ which implies that $\widehat{\theta} \xrightarrow{p} \theta$. If remains to show that $\mathrm{KL}(p(x;\theta), p(x;\widehat{\theta})) \leq \eta$, as $n \to \infty$. Notice that,

$$\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) = R(\widehat{\theta},\theta) = R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) + R_n(\widehat{\theta},\theta) \overset{\text{(i)}}{\leq} R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) \xrightarrow{p} 0,$$

where the final convergence simply uses Condition 3. The inequality (i) follows since,

$$R_n(\widehat{\theta},\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i;\theta)}{p(X_i;\widehat{\theta})} \leq 0,$$

since $\widehat{\theta}$ is the MLE. ■

3 Inconsistency of the MLE

The MLE can fail to be consistent. When the model is not identifiable it is clear that we cannot have consistent estimators. The other possible failure is the failure of the uniform law. This typically happens when the parameter space is too large. Here is a simple example:

Example: Suppose that we measure some outcome (say their blood sugar) for n individuals using a machine. We do it twice for every individual so that we can assess the variability of the machine, i.e. suppose we observe:

$$\begin{aligned} &Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2) \\ &\vdots \\ &Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2), \end{aligned}$$

and want to estimate σ^2. Even though we only want to estimate σ^2 the model has a growing number of parameters $\mu_1, \ldots, \mu_n, \sigma^2$ and the MLE for σ^2 will depend on estimating μ_i. Formally, we can see that the MLE for the means is:

$$\widehat{\mu}_i = \frac{Y_{i1} + Y_{i2}}{2}.$$

The log-likelihood for σ^2 can be written as:

$$\mathcal{LL}(\sigma^2, \mu) = -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(Y_{i1} - \mu_i)^2 + (Y_{i2} - \mu_i)^2 \right],$$

which is maximized when we take:

$$\widehat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}\left[(Y_{i1}-\widehat{\mu}_i)^2 + (Y_{i2}-\widehat{\mu}_i)^2\right] = \frac{1}{4n}\sum_{i=1}^{n}(Y_{i1}-Y_{i2})^2.$$

Notice that,

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{\sigma^2}{2},$$

so by the LLN the MLE is inconsistent. One could easily fix this in this problem (by multiplying the MLE by 2) but more generally this could be tricky. We note that in this type of problem where the number of parameters is not fixed (and grows with the sample size) it is not even clear how to define convergence of the log-likelihood since its limit changes with the sample size.

4 MLE under misspecification

In statistical modeling we do not typically believe the model is correct, i.e. that the samples were in fact generated by some distribution in our model. Rather, we think of the model as a useful idealization or a simplification. In this (more realistic) case, one might wonder what the MLE converges to, or if it converges at all?

Suppose $X_1, \ldots, X_n \sim q$, and we estimate $\widehat{\theta}_{\text{MLE}}$, then what can we say about our estimate? To answer this, we can follow a similar argument to what we did in the beginning of the lecture and observe that at the population-level (i.e. with infinite samples) the MLE is:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \mathbb{E}_q \log p(X;\theta)$$

How do we interpret this statement? As before we can re-write it in terms of KL divergences and see that:

$$\text{KL}(q\|p_{\widehat{\theta}_{\text{MLE}}}) \leq \text{KL}(q\|p_\theta) \quad \text{for all } \theta\in\Theta.$$

So that at the population-level we can conclude that the MLE is estimating the KL projection of the data-generating distribution on our model, i.e. when q does not belong to our model the MLE is essentially estimating the KL projection of q onto our model. One can also impose similar conditions to what we had in the last section (uniform law + strong identifiability) to complete the consistency argument under model misspecification.

5 Limiting Distribution of the MLE

Now we will address the question of what is the asymptotic distribution of the MLE. This is analogous to the CLT which gave the asymptotic distribution of averages. In some cases, we

can do this directly. For instance, if $X_1, \ldots, X_n \sim \text{Ber}(p)$ then the MLE is just the average:

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and so we know by the CLT:

$$\sqrt{n} \frac{\widehat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1),$$

which tells us the asymptotic distribution of the MLE.

More generally, however the MLE need not be a simple average of i.i.d. terms, but the main take-away is that asymptotically it often behaves like one.

Recall that the score function is

$$s(\theta) = \sum_{i=1}^{n} \nabla \log(p(X_i; \theta)),$$

which is the gradient of the log-likelihood, and the Fisher Information,

$$I(\theta) = \mathbb{E}[s(\theta) s(\theta)^T].$$

We showed that $s(\theta)$ has mean 0, so $I(\theta) = \text{Var}(s(\theta))$. The Fisher information is alternatively the expected Hessian of the log-likelihood:

$$I_n(\theta) = \mathbb{E}\left[\sum_{i=1}^{n} \nabla^2 \log p(X_i; \theta) \right].$$

It is worth remembering that the score is data-dependent, while the Fisher Information is not (it is an expectation over the data so does not depend on the values of $X_1, \ldots, X_n$).

Let $\widehat{\theta}$ denote the MLE. Our goal to show that (under enough regularity conditions),

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, [I_1(\theta)]^{-1}).$$

6 Counterexample

The usual counterexample to the above convergence in distribution is the MLE for the uniform distribution. For the uniform distribution most regularity conditions fail. Formally, we observe $X_1, \ldots, X_n \sim U[0, \theta]$ and want to estimate θ. The log-likelihood:

$$\ell(\theta) = \log \left[\frac{1}{\theta^n} \mathbb{I}(\theta \geq \max_i X_i) \right].$$

The MLE is $\widehat{\theta} = \max_{i=1}^{n} X_i$. Observe that the log-likelihood is not differentiable at the MLE, so the Fisher information is not defined at the MLE.

Another thing that we used frequently in defining the equivalent forms of the Fisher information was to exchange derivatives (with respect to θ) and integrals (with respect to X). This in general does not work if the domain of integration depends on the parameter with respect to which we are taking the derivative. For the uniform distribution the domain of density depends on the parameter.

On the other hand, things are usually nice for exponential families. They will automatically satisfy all the regularity conditions (provided it is identifiable, i.e. say full-rank and minimal) and the MLE is extremely well-behaved in such models.

Returning to the uniform case, we can directly analyze the distribution of the MLE. In a previous lecture we showed that

$$n(\widehat{\theta} - \theta) \xrightarrow{d} -\text{Exp}(1/\theta)$$

(we did this when $\theta = 1$ but you can work out the general case). It follows that $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \delta_0$, where δ_0 is a point mass at 0 and it does not have a Gaussian limit.

7 MLE asymptotics

We will only attempt a heuristic calculation here. If you are curious to see a rigorous proof with minimal regularity assumptions you should look at Van der Vaart's book on Asymptotic Statistics. Here is a list of some sufficient regularity conditions:

1. The dimension of the parameter space does not change with n, i.e. $\theta \in \mathbb{R}^d$ and d is fixed. We have seen that if d grows the MLE need not even be consistent.
2. $p(x;\theta)$ is a smooth (thrice differentiable) function of θ,
3. We can interchange differentiation with respect to θ and integration over X. This in turn requires that the range of X does not depend on θ, and some integrability conditions on $p(x;\theta)$.
4. The parameter θ is identifiable.
5. If the parameter space is restricted, i.e. $\theta \in \Theta$ for some set Θ then θ is in the interior of the set Θ (i.e. cannot be on its boundary).

We will focus on the case when the parameter is one-dimensional, although everything carries over almost exactly in the general (fixed) d case.

Theorem 2 *Under the regularity conditions above,*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, 1/I(\theta)).$$

We note that under the conditions of the theorem one can verify that the MLE is consistent, i.e. that $\widehat{\theta} \xrightarrow{p} \theta$. The basic idea is to verify that under the differentiability assumptions on the density, we can effectively treat the parameter space as compact, then derive a uniform law of large numbers, and then apply the proof from the previous lecture notes. This is a complicated technical proof but you can look it up by searching for Wald's proof of the consistency of the MLE.

The proof will use all the facts about scores and the Fisher information that we derived earlier.

Proof: To begin with let us note the following fact: if $\widehat{\theta} \xrightarrow{p} \theta$, then

$$\mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \widehat{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \theta)] = I(\theta).$$

Since $\widehat{\theta}$ maximizes the log-likelihood we know that the derivative of the log-likelihood at $\widehat{\theta}$ must be 0, i.e.

$$\ell'(\widehat{\theta}) = 0.$$

Formally you need to know that $\widehat{\theta}$ is not on the boundary of the parameter space. To prove this you will need to use the fact that θ is not on the boundary and that $\widehat{\theta} \xrightarrow{p} \theta$.

By a Taylor expansion of the derivative of the log-likelihood we obtain that,

$$0 = \ell'(\widehat{\theta}) = \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\widetilde{\theta}),$$

where $\widetilde{\theta}$ is some point in between $\widehat{\theta}$ and θ. This in turn gives us that,

$$(\widehat{\theta} - \theta) = \frac{\ell'(\theta)}{-\ell''(\widetilde{\theta})},$$

so that,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{\ell'(\theta)}{\sqrt{n}}}{-\frac{\ell''(\widetilde{\theta})}{n}}.$$

We will look at the numerator and denominator separately. The denominator is:

$$-\frac{\ell''(\widetilde{\theta})}{n} = \frac{1}{n}\sum_{i=1}^{n} -\nabla^2_\theta \log p(X_i; \widetilde{\theta}) \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \widetilde{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X; \theta)] = I(\theta)$$

where the last step uses the fact that $\widetilde{\theta} \xrightarrow{p} \theta$.

The numerator is just the score function, i.e.

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log p(X_i;\theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} [\nabla_\theta \log p(X_i;\theta) - \mathbb{E}[\nabla_\theta \log p(X;\theta)]]$$
$$\xrightarrow{d} N(0, \mathrm{Var}(\nabla_\theta \log p(X;\theta))) \xrightarrow{d} N(0, I(\theta)),$$

where we used the facts that the score has mean 0, that the variance of the score is the Fisher information and that by the CLT $\sqrt{n}$ times an average of i.i.d. terms minus its expectation converges in distribution to a normal.

Putting the pieces together via Slutsky's theorem we obtain that,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \frac{1}{I(\theta)} N(0, I(\theta)) \xrightarrow{d} N(0, 1/I(\theta)),$$

which is what we wanted to prove. ■

Example: Suppose that $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$, then the log-likelihood,

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} X_i.$$

The score function:

$$s(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i,$$

and the Fisher information,

$$I(\theta) = \frac{n}{\theta^2}.$$

The MLE is $\widehat{\theta} = \frac{1}{\overline{X}}$. So we can use the above result to conclude that,

$$\widehat{\theta} - \theta \xrightarrow{d} N\left(0, \frac{\theta^2}{n}\right).$$

8 Influence Functions and Regular Asymptotically Linear Estimators

We could have followed a similar proof as above to conclude that the MLE can be written as:

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_\theta \log p(X_i;\theta)}{I(\theta)} + \text{Remainder},$$

where the remainder is small (roughly proportional to the previous term multiplied by $[I(\widetilde{\theta}) - I(\theta)] \to 0$). The term,

$$\psi(x) = \frac{\nabla_\theta \log p(x;\theta)}{I(\theta)},$$

is called the *influence function*.

Thinking of a complex predictor like a deep neural network, one can try to obtain some information about the predictor by trying to compute the influence function of training images on the final predictor. A paper that did this (and quite a bit more) won ICML's best paper a few years ago.

Returning to the expression:

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i).$$

Estimators that satisfy this type of expansion are called asymptotically linear estimators (many non-MLE estimators also satisfy expansions of this form). There is a classical result due to Le Cam that any sufficiently well-behaved (regular) estimator is asymptotically linear. It is not easy to prove (see Van Der Vaart's book). This together with the Cramér-Rao lower bound implies that the MLE is the "best regular asymptotically linear estimator".

9 Asymptotic Relative Efficiency

Once you restrict attention to asymptotically Normal estimators, comparing estimators in terms of their MSE boils down to comparing their variances. Specifically, if

$$\begin{aligned} \sqrt{n}(W_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_W^2) \\ \sqrt{n}(V_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_V^2) \end{aligned}$$

then the *asymptotic relative efficiency (ARE)* is

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 3 *Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. The mle of λ is $\overline{X}$. Let*

$$\tau = \mathbb{P}(X_i = 0).$$

So $\tau = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Another estimator is the mle

$$V_n = e^{-\widehat{\lambda}}.$$

The delta method gives

$$\text{Var}(V_n) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

We have

$$\begin{aligned} \sqrt{n}(W_n - \tau) &\rightsquigarrow N(0, e^{-\lambda}(1 - e^{-\lambda})) \\ \sqrt{n}(V_n - \tau) &\rightsquigarrow N(0, \lambda e^{-2\lambda}). \end{aligned}$$

So

$$\text{ARE}(W_n, V_n) = \frac{\lambda}{e^{\lambda} - 1} \leq 1. \quad \square$$

Since the mle is efficient, we know that, in general, $\text{ARE}(W_n, \text{mle}) \leq 1$.

10 Multivariate Case

Now let $\theta = (\theta_1, \ldots, \theta_k)$. In this case we have

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, I^{-1}(\theta))$$

where $I^{-1}(\theta)$ is the inverse of the Fisher information matrix. The approximate standard error of $\widehat{\theta}_j$ is $\sqrt{I_{jj}^{-1}/n}$. If $\tau = g(\theta)$ with $g : \mathbb{R}^k \to \mathbb{R}$ then by the delta method,

$$\sqrt{n}(\widehat{\tau} - \tau) \rightsquigarrow N(0, (g')^T I^{-1} g')$$

where g' is the gradient of g evaluated at θ.

Lecture Notes 16
36-705

Today we will switch gears and talk about hypothesis testing. But before we do, there is one last, important fact about point estimation: the optimality of the mle. It's complicated to make this precise. (See *Asymptotic Statistics* by van der Vaart for a good treatment.)

1 The MLE is Optimal

Roughly, it goes like this. We know that mle satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N\left(0, \frac{1}{I(\theta)}\right).$$

If $\widetilde{\theta}$ is any other well-behaved estimators, then

$$\sqrt{n}(\widetilde{\theta} - \theta) \rightsquigarrow N(0, \sigma^2)$$

where $\sigma^2 \geq 1/I(\theta)$. The phrase "well-behaved" refers to some desirable technical conditions on the estimator. Thus, the mle is the most precise estimator. Similarly, it can be shown that the mle is asymptotically minimax under a large class of loss functions. Unfortunately, making all this precise takes machinery that is beyond the scope of the course. But the message is just that there are good reasons for using the mle in parametric problems. In nonparametric problems, we shall see that the situation is quite different.

1.1 Notation

We will need the following notation. Let Φ be the cdf of a standard Normal random variable Z. For $0 < \alpha < 1$, let

$$z_\alpha = \Phi^{-1}(1 - \alpha).$$

Hence,

$$P(Z > z_\alpha) = \alpha \quad \text{and} \quad P(Z < -z_\alpha) = \alpha.$$

Sometimes we will write X^n to mean $(X_1, \ldots, X_n)$.

2 Hypothesis Testing

The classical statistical hypothesis testing framework (as with much of statistics) originated with Fisher.

Example 1: The story goes that a colleague of Fisher claimed to be able to distinguish if in an English tea, milk was added before water (or the other way around). Fisher proposed to give her 8 cups of tea, 4 of which had milk first, and 4 of which had tea first in a random order. The point was roughly, that if she was "labeling" at random then she would have a small chance $1/\binom{8}{4} = 1/70 = 0.014$ of getting every cup right. In his description, the null hypothesis was that she had no ability to distinguish. She actually got them all correct, which would have happened by chance with probability 0.014. He concluded that since this probability was less that 0.05 that it was "statistically significant". Notice the asymmetry in this description: only a null hypothesis is actually specified (i.e. there is no alternative hypothesis – it is in some sense implicit), i.e. the null hypothesis is often special. Furthermore, there is an arbitrary choice of a cut-off 0.05 below which we declare something is significant.

Hypothesis testing is really everywhere. It would probably alarm you to know how many policy decisions, nutrition decisions, scientific results live or die on the basis of hypothesis tests.

Exaample 2. In July, Castillo et al reported on a study about treating Covid with Vitamin D. 76 patients were randomized to treatment or no treatment. The outcome was whether they needed to be put in the ICU or not. The null hypothesis that there is no value in using Vitamin D had a p-value less than .001.

Example 3: A couple of typical examples to emphasize again why the null might really be special. A common example is in forensics. Things like fingerprint matches, DNA matches, deciding whether pieces of glass match in their chemical composition etc. are actually problems of a statistical nature. Here perhaps following the "innocent till proven guilty" adage, the null hypothesis is that the defendant is innocent. We then need to review evidence and choose to either reject or fail to reject (i.e. acquit) the defendant. It is perhaps clear that there in many cases is a heavier price for false convictions and so it makes sense to control this error. Indeed, deciding how to choose a significance level in this context is a huge debate.

3 The formal framework

Let $X_1, \ldots, X_n \sim p(x; \theta)$. Suppose we want to know if $\theta = \theta_0$ or not, where θ_0 is a specific value of θ. For example, if we are flipping a coin, we may want to know if the coin is fair; this corresponds to $\theta = 1/2$. If we are testing the effect of two drugs — whose means effects are θ_1 and θ_2 — we may be interested to know if there is no difference, which corresponds to $\theta_1 - \theta_2 = 0$.

We formalize this by stating a *null hypothesis* H_0 and an alternative hypothesis H_1. For

example:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad \theta \neq \theta_0.$$

More generally, consider a parameter space Θ. We consider

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$. If Θ_0 consists of a single point, we call this a *simple null hypothesis*. If Θ_0 consists of more than one point, we call this a *composite null hypothesis*.

Example 1 $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}. \quad \square$$

The question is not whether H_0 is true or false. The question is whether there is sufficient evidence to reject H_0, much like a court case. Our possible actions are: reject H_0 or retain (don't reject) H_0.

	Decision	
	Retain H_0	Reject H_0
H_0 true	$\surd$	Type I error (false positive)
H_1 true	Type II error (false negative)	$\surd$

4 Constructing Tests

Hypothesis testing involves the following steps:

1. Choose a *test statistic* $T_n = T_n(X_1, \ldots, X_n)$.
2. Choose a rejection region $R \subset \mathcal{X}^n$. Often this has the form

   $$R = \Big\{(x_1, \ldots, x_n) : \ T_n(x_1, \ldots, x_n) > t\Big\}$$

   for some t.

3. If $(X_1, \ldots, X_n) \in R$ we reject H_0 otherwise we retain H_0.

Although you can define the rejection region without an associated test statistic, often it will be the case that R will be defined in terms of the test statistic, i.e. we simply reject if the test statistic takes an "extreme value". We define the *test function* ϕ by:

$$\phi(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } (x_1, \ldots, x_n) \in R \\ 0 & \text{otherwise.} \end{cases}$$

Example 2 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. *Suppose we test*

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

Let $T_n = n^{-1} \sum_{i=1}^n X_i$ *and*

$$R = \left\{ (x_1, \ldots, x_n) : \ |T_n(x_1, \ldots, x_n) - 1/2| > \delta \right\}.$$

So we reject H_0 *if* $|T_n - 1/2| > \delta$.

We need to choose T and R so that the test has good statistical properties. We will consider the following tests:

1. The Neyman-Pearson Test
2. The Wald test
3. The Likelihood Ratio Test (LRT)
4. The permutation test.

Before we discuss these methods, we first need to talk about how we evaluate tests.

5 Error Rates and Power

Suppose we reject H_0 when $(X_1, \ldots, X_n) \in R$. Define the *power function* by

$$\beta(\theta) = P_\theta((X_1, \ldots, X_n) \in R).$$

We want $\beta(\theta)$ to be small when $\theta \in \Theta_0$ and we want $\beta(\theta)$ to be large when $\theta \in \Theta_1$. The general strategy is:

1. Fix $\alpha \in [0,1]$.
2. Now try to maximize $\beta(\theta)$ for $\theta \in \Theta_1$ subject to $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$.

Notice the asymmetry that we always favor the null hypothesis and only consider tests that control the Type-I error.

We need the following definitions. A test is *size* α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

A test is *level* α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

A size α test and a level α test are almost the same thing. The distinction is made bcause sometimes we want a size α test and we cannot construct a test with exact size α but we can construct one with a smaller error rate.

Example 3 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose we test*

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta > \theta_0.$$

This is called a **one-sided alternative**. *Suppose we reject H_0 if $T_n > c$ where*

$$T_n = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}.$$

Then

$$\begin{aligned} \beta(\theta) &= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) = P_\theta\left(\frac{\overline{X}_n - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right), \end{aligned}$$

where Φ is the cdf of a standard Normal and $Z \sim \Phi$. Now

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = 1 - \Phi(c).$$

To get a size α test, set $1 - \Phi(c) = \alpha$ so that

$$c = z_\alpha$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$. Our test is: reject H_0 when

$$T_n = \frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > z_\alpha.$$

Example 4 *$X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Suppose*

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

This is called a **two-sided** *alternative. We will reject H_0 if $|T_n| > c$ where T_n is defined as before. Now*

$$\begin{aligned}
\beta(\theta) &= P_\theta(T_n < -c) + P_\theta(T_n > c) \\
&= P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} < -c\right) + P_\theta\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} > c\right) \\
&= P\left(Z < -c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(-c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) + \Phi\left(-c - \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

since $\Phi(-x) = 1 - \Phi(x)$. The size is

$$\beta(\theta_0) = 2\Phi(-c).$$

To get a size α test we set $2\Phi(-c) = \alpha$ so that $c = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}$. The test is: reject H_0 when

$$|T| = \left|\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

When $\alpha = .05$, $z_{\alpha/2} = 1.96 \approx 2$. In this case we reject when $|T| > 2$.

6 The Neyman-Pearson Test

Let $\mathcal{C}_\alpha$ denote all level α tests. A test in $\mathcal{C}_\alpha$ with power function β is **uniformly most powerful (UMP)** if the following holds: if β' is the power function of any other test in $\mathcal{C}_\alpha$ then $\beta(\theta) \geq \beta'(\theta)$ for all $\theta \in \Theta_1$.

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. (Simple null and simple alternative.)

Theorem 5 *Let $L(\theta) = p(X_1, \ldots, X_n; \theta)$ and*

$$T_n = \frac{L(\theta_1)}{L(\theta_0)}.$$

Suppose we reject H_0 if $T_n > k$ where k is chosen so that

$$P_{\theta_0}(X^n \in R) = \alpha.$$

This test is a UMP level α test.

One nice thing about this is that it is a "general recipe" for doing a hypothesis test. The drawback of course is that it only applies to the restricted class of simple versus simple tests. The Neyman-Pearson test, despite its restricted applicability is a very important conceptual contribution. When it is applicable it is an optimal test. This is often called the Neyman-Pearson Lemma.

Proof of the Neyman-Pearson Lemma. Let us denote the test function of the NP test as ϕ_{NP} and the test function of any other test we want to compare against as ϕ_A. The test function simply takes the value 1 if the test rejects and 0 otherwise. To ease notation we will assume that $n = 1$. Let $f_0(x) = L(\theta_0; x)$ and $f_1(x) = L(\theta_1; x)$. So with this notation, we reject if:

$$\frac{f_1(x)}{f_0(x)} \geq k.$$

To prove the NP Lemma, we will first argue that the following is true:

$$\int_x \underbrace{(\phi_{NP}(x) - \phi_A(x))}_{U_1} \underbrace{(f_1(x) - kf_0(x))}_{U_2} dx \geq 0.$$

To see this we can just consider some cases:

1. If both tests reject or if both tests accept then the inequality is clearly true since the LHS is 0.
2. If NP rejects, and the test A accepts then $\phi_{NP}(x) = 1$, and $\phi_A(x) = 0$, so $U_1 \geq 0$. Since the NP test rejected the null we know that:
$$\frac{f_1(x)}{f_0(x)} \geq k,$$
so that $U_2 \geq 0$. So the inequality is true in this case.
3. If NP accepts and the test A rejects then both U_1 and U_2 are negative so the inequality is also true in this case.

So we can see that for every x, $U_1 \times U_2 \geq 0$ so it is true when we integrate over x. Now, we can rearrange this inequality to see that:

$$\begin{aligned}\int_x (\phi_{NP}(x) - \phi_A(x)) f_1(x) dx &\geq k \int_x (\phi_{NP}(x) - \phi_A(x)) f_0(x) dx \\ &= k \left(\underbrace{\int_x \phi_{NP}(x) f_0(x) dx}_{=\alpha} - \underbrace{\int_x \phi_A(x) f_0(x) dx}_{\leq \alpha} \right) \\ &\geq 0.\end{aligned}$$

This proves the NP lemma, i.e. that the power of the NP test is larger than the power of any other test. □

Now we develop some tests that are useful in other more complex settings.

7 The Wald Test

When we are testing a simple null hypothesis against a possibly composite alternative, the NP test is no longer applicable. A general method is the Wald test. We are interested in testing the hypotheses in a parametric model:

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta \neq \theta_0. \end{aligned}$$

The Wald test most generally is based on an asymptotically normal estimator, i.e. we suppose that we have access to an estimator $\widehat{\theta}$ which, under the null, satisfies the property that:

$$\frac{\widehat{\theta} - \theta_0}{se_0} \xrightarrow{d} N(0,1)$$

where $se_0 = \sqrt{\text{Var}}(\widehat{\theta})$ is the standard deviation of $\widehat{\theta}$ under the null. In this case, we could consider the statistic:

$$T_n = \frac{\widehat{\theta} - \theta_0}{se_0}$$

or, if se_0 is not known, we use

$$T_n = \frac{\widehat{\theta} - \theta_0}{\widehat{se_0}}.$$

Under the null $T_n \xrightarrow{d} N(0,1)$, so we simply reject the null if: $|T_n| \geq z_{\alpha/2}$. This controls the Type-I error only asymptotically (i.e. only if $n \to \infty$) but this is relatively standard in applications. That is

$$P_{\theta_0}(|T_n| \geq z_{\alpha/2}) \to \alpha.$$

It is also valid to use the statistic

$$T_n = \frac{\widehat{\theta} - \theta_0}{\widehat{se}}$$

where $\widehat{se}$ is any consistent estimate of the standard error; it's not necessary to assume H_0 is true when estimating the standard error. (This follows from Slutsky's theorem and the continuous mapping theorem.)

Example: Suppose that $X_1, \ldots, X_n \sim \text{Ber}(p)$, and the null is that $p = p_0$. Defining $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. Let

$$T_n = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

which has an asymptotic $N(0,1)$ distribution. As mentioned above, we can also use

$$T_n = \frac{\widehat{p} - p_0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}.$$

Observe that this alternative test statistic also has an asymptotically standard normal distribution under the null. Its behaviour under the alternate is a bit more pleasant as we will see.

7.1 Power of the Wald Test

To get some idea of what happens under the alternate, suppose we are in some situation where the MLE has "standard asymptotics", i.e. $\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow N(0, 1/(I_1(\theta)))$. Suppose that we use the statistic:

$$T_n = \sqrt{nI_1(\widehat{\theta})}(\widehat{\theta} - \theta_0),$$

and that the true value of the parameter is $\theta_1 \neq \theta_0$. Let us define:

$$\Delta = \sqrt{nI_1(\theta_1)}(\theta_0 - \theta_1),$$

then the probability that the Wald test rejects the null hypothesis is asymptotically:

$$1 - \Phi\left(\Delta + z_{\alpha/2}\right) + \Phi\left(\Delta - z_{\alpha/2}\right).$$

You will prove this on your HW (it is some simple re-arrangement, similar to what we have done previously when computing the power function in a Gaussian model). There are some aspects to notice:

1. If the difference between θ_0 and θ_1 is very small the power will tend to α, i.e. if $\Delta \approx 0$ then the test will have trivial power.

2. As $n \to \infty$ the two Φ terms will approach either 0 or 1, and so the power will approach 1.

3. As a rule of thumb the Wald test will have non-trivial power if $|\theta_0 - \theta_1| \gg \frac{1}{\sqrt{nI_1(\theta_1)}}$.

8 Likelihood Ratio Test (LRT)

To test composite versus composite hypotheses the general method is to use something called the (generalized) likelihood ratio test. We want to test:

$$\begin{aligned} H_0 : &\quad \theta \in \Theta_0 \\ H_1 : &\quad \theta \notin \Theta_0. \end{aligned}$$

This test is simple: reject H_0 if $\lambda(X_1, \ldots, X_n) \leq c$ where

$$\lambda(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}$$

where $\widehat{\theta}_0$ maximizes $L(\theta)$ subject to $\theta \in \Theta_0$.

We can simplify the LRT by using an asymptotic approximation. This fact that the LRT generally has a simple asymptotic approximation is known as *Wilks' phenomenon.* First, some notation:

Notation: Let $W \sim \chi^2_p$. Define $\chi^2_{p,\alpha}$ by

$$P(W > \chi^2_{p,\alpha}) = \alpha.$$

We let $\ell(\theta)$ denote the log-likelihood in what follows.

Theorem 6 *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ where $\theta \in \mathbb{R}$. Under H_0,*

$$-2 \log \lambda(X_1, \ldots, X_n) \rightsquigarrow \chi^2_1.$$

Hence, if we let $T_n = -2 \log \lambda(X^n)$ then

$$P_{\theta_0}(T_n > \chi^2_{1,\alpha}) \to \alpha$$

as $n \to \infty$.

Proof: Using a Taylor expansion:

$$\ell(\theta) \approx \ell(\widehat{\theta}) + \ell'(\widehat{\theta})(\theta - \widehat{\theta}) + \ell''(\widehat{\theta}) \frac{(\theta - \widehat{\theta})^2}{2} = \ell(\widehat{\theta}) + \ell''(\widehat{\theta}) \frac{(\theta - \widehat{\theta})^2}{2}$$

and so

$$\begin{aligned} -2\log\lambda(x_1,\ldots,x_n) &= 2\ell(\widehat{\theta}) - 2\ell(\theta_0) \\ &\approx 2\ell(\widehat{\theta}) - 2\ell(\widehat{\theta}) - \ell''(\widehat{\theta})(\theta_0-\widehat{\theta})^2 = -\ell''(\widehat{\theta})(\theta_0-\widehat{\theta})^2 \\ &= \frac{-\frac{1}{n}\ell''(\widehat{\theta})}{I_1(\theta_0)}(\sqrt{nI_1(\theta_0)}(\widehat{\theta}-\theta_0))^2 = A_n \times B_n. \end{aligned}$$

Now $A_n \xrightarrow{p} 1$ by the WLLN and $\sqrt{B_n} \rightsquigarrow N(0,1)$. The result follows by Slutsky's theorem. ■

Example 7 *$X_1,\ldots,X_n \sim \text{Poisson}(\lambda)$. We want to test $H_0: \lambda = \lambda_0$ versus $H_1: \lambda \neq \lambda_0$. Then*

$$-2\log\lambda(x^n) = 2n[(\lambda_0 - \widehat{\lambda}) - \widehat{\lambda}\log(\lambda_0/\widehat{\lambda})].$$

We reject H_0 when $-2\log\lambda(x^n) > \chi^2_{1,\alpha}$.

Now suppose that $\theta = (\theta_1,\ldots,\theta_k)$. Suppose that $H_0: \theta \in \Theta_0$ fixes some of the parameters. Then, under conditions,

$$T_n = -2\log\lambda(X_1,\ldots,X_n) \rightsquigarrow \chi^2_\nu$$

where

$$\nu = \dim(\Theta) - \dim(\Theta_0).$$

Therefore, an asymptotic level α test is: reject H_0 when $T_n > \chi^2_{\nu,\alpha}$.

Example 8 *Consider a multinomial with $\theta = (p_1,\ldots,p_5)$. So*

$$L(\theta) = p_1^{y_1}\cdots p_5^{y_5}.$$

Suppose we want to test

$$H_0: p_1 = p_2 = p_3 \quad \text{and} \quad p_4 = p_5$$

versus the alternative that H_0 is false. In this case

$$\nu = 4 - 1 = 3.$$

The LRT test statistic is

$$\lambda(x_1,\ldots,x_n) = \frac{\prod_{j=1}^5 \widehat{p}_{0j}^{Y_j}}{\prod_{j=1}^5 \widehat{p}_j^{Y_j}}$$

where $\widehat{p}_j = Y_j/n$, $\widehat{p}_{01} = \widehat{p}_{02} = \widehat{p}_{03} = (Y_1+Y_2+Y_3)/n$, $\widehat{p}_{04} = \widehat{p}_{05} = (1-3\widehat{p}_{01})/2$. Now we reject H_0 if $-2\lambda(X_1,\ldots,X_n) > \chi^2_{3,\alpha}$. □

9 p-values

When we test at a given level α we will reject or not reject. It is useful to summarize what levels we would reject at and what levels we would not reject at.

The p-value is the smallest α at which we would reject H_0.

In other words, we reject at all $\alpha \geq p$. So, if the pvalue is 0.03, then we would reject at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Hence, to test at level α, we reject when $p < \alpha$.

Theorem 9 *Suppose we have a test of the form: reject when $T(X_1, \ldots, X_n) > c$. Then the p-value is*

$$p = \sup_{\theta \in \Theta_0} P_\theta(T_n(X_1^*, \ldots, X_n^*) \geq T_n(x_1, \ldots, x_n))$$

where $x_1, \ldots, x_n$ are the observed data and $X_1^*, \ldots, X_n^* \sim p_{\theta_0}$.

Example 10 *$X_1, \ldots, X_n \sim N(\theta, 1)$. Test that $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. We reject when $|T_n|$ is large, where $T_n = \sqrt{n}(\overline{X}_n - \theta_0)$. Let t_n be the observed value of T_n. Let $Z \sim N(0, 1)$. Then,*

$$p = P_{\theta_0}\left(|\sqrt{n}(\overline{X}_n - \theta_0)| > t_n\right) = P(|Z| > t_n) = 2\Phi(-|t_n|).$$

The p-value is a random variable. Under some assumptions that you will see in your HW the p-value will be uniformly distributed on $[0, 1]$ under the null.

Important. Note that p is NOT equal to $\mathbb{P}(H_0|X_1, \ldots, X_n)$. The latter is a Bayesian quantity which we will discuss later.

Lecture Notes 17
36-705

Now we will take a look at some specific hypothesis testing problems. And we shall depart

1 Goodness-of-fit testing

Let $X_1, \ldots, X_n \sim P$. We want to test:

$$\begin{aligned} H_0 : &\quad P = P_0 \\ H_1 : &\quad P \neq P_0, \end{aligned}$$

for some fixed, known distribution P_0. As an example, suppose that P is multinomials on k categories. The null distribution just a vector of probabilities $(p_{01}, \ldots, p_{0k})$, with $p_{0i} \geq 0$, $\sum_i p_{0i} = 1$. We could use the LRT but here we introduce another popular test.

Given a sample $X_1, \ldots, X_n$ you can reduce it to a vector of counts $(Z_1, \ldots, Z_k)$ where Z_i is the number of times we observed the i-th category. Let

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{k} \frac{(Z_i - np_{0i})^2 - np_{0i}}{np_{0i}}.$$

On your HW you will show that asymptotically this test statistic, under the null, has a χ^2_{k-1} distribution. This is called Pearson's χ^2 test. More generally, we can perform any goodness-of-fit test by reducing to a multinomial test by binning, i.e. you define a sufficiently find partition of the domain, this induces a multinomial p_0 under the null which you then test using Pearson's test.

2 Two-sample Testing

Another popular hypothesis testing problem is the following: you observe $X_1, \ldots, X_{n_1} \sim P$ and $Y_1, \ldots, Y_{n_2} \sim Q$, and want to test if:

$$\begin{aligned} H_0 : &\quad P = Q \\ H_1 : &\quad P \neq Q. \end{aligned}$$

Assume first that P and Q are in the same parametric family $(P_\theta : \ \theta \in \Theta)$ So $p = p(x; \theta_1)$ and $q = p(x; \theta_2)$ for some θ_1, θ_2. We want to test $H_0 : \theta_1 = \theta_2$. If the parameter is scalar, the Wald test statistic is

$$T = \frac{\widehat{\theta}_1 - \widehat{\theta}_2}{se}$$

where

$$se^2 = \frac{se_1^2}{n_1} + \frac{se_2^2}{n_2}$$

and se_1 and se_2 are estimates of the standard errors of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. If θ is a vector, we can use the $LRT = -2\log\lambda$ where

$$\lambda = \frac{\prod_i p(X_i;\widehat{\theta})\prod_i p(Y_i;\widehat{\theta})}{\prod_i p(X_i;\widehat{\theta}_1)\prod_i p(Y_i;\widehat{\theta}_2)}$$

where $\widehat{\theta}$ is the mle under H_0 obtained by combining both samples. If θ is a vector of length k, then under the null there are k parameters and under the alternative there are $2k$ parameters. The difference is k. So the LRT converges to χ^2_k under H_0.

Consider again the multinomial setting where P and Q are multinomials on k categories. Then there is a version of the χ^2 test that is commonly used. Let us define $(Z_1,\ldots,Z_k)$ and $(Z'_1,\ldots,Z'_k)$ to be the counts in the X and Y sample respectively. We can define for $i \in \{1,\ldots,k\}$,

$$\widehat{c}_i = \frac{Z_i + Z'_i}{n_1+n_2}.$$

The two-sample χ^2 test is then:

$$T_n = \sum_{i=1}^{k}\left[\frac{(Z_i - n_1\widehat{c}_i)^2}{n_1\widehat{c}_i} + \frac{(Z'_i - n_2\widehat{c}_i)^2}{n_2\widehat{c}_i}\right].$$

This is a bit harder to see but under the null this statistic also has a χ^2_{k-1} distribution.

3 The Permutation Test

For two-sample testing we can determine the cutoff in a different way without resorting to asymptotics and without assuming a parametric model.

A typical example is in a drug trial where one set of people are given a drug and the other set are given a placebo. We then would like to know if there is some difference in the outcomes of the two populations or if they are identically distributed.

Let $T(X_1,\ldots,X_m,Y_1,\ldots,Y_n)$ be any test statistic. For example,

$$T(X_1,\ldots,X_m,Y_1,\ldots,Y_n) = \left|\frac{1}{m}\sum_{i=1}^{m}X_i - \frac{1}{n}\sum_{i=1}^{n}Y_i\right|.$$

Let us denote the value of the test statistic computed on the observed data by T_{obs}.

The idea of the permutation test is simple. Define $N = m + n$ and consider all $N!$ permutations of the data $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. For each permutation we could compute our test statistic T. Denote these as $T_1, \ldots, T_{N!}$. The key observation is: **under the null hypothesis each value $T_1, \ldots, T_{N!}$ has the *same* distribution (even if we do not know what it is).**

Suppose we reject for large values of T. Then we could simply define the p-value as:

$$\text{p-value} = \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{I}(T_i > T_{\text{obs}}).$$

It is important to note that this is an exact p-value, i.e. no asymptotic approximations are needed to show that rejecting the null when this p-value is less than α controls the Type I error at α. Here is a toy-example:

Example 2: Suppose we observe $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1)$ be the difference in means, i.e. $T(X_1, X_2, Y_1) = 2$. The permutations are:

permutation	value of T
(1,9,3)	2
(9,1,3)	2
(1,3,9)	7
(3,1,9)	7
(3,9,1)	5
(9,3,1)	5

We could use this to calculate the p-value by counting how often we got a larger value than 2:

$$\text{p-value} = \frac{4}{6} = 0.66,$$

so most likely we would not reject the null hypothesis in this case. Typically, we do not calculate the exact p-value (although in principle we could) since evaluating $N!$ test statistics would take too long for large N. Instead we approximate the p-value by drawing a few random permutations and using them. This leads to the following algorithm for computing the p-value using a permutation test:

Algorithm for Permutation Test

1. Compute the observed value of the test statistic $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.
2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

We first show that the permutation test that we covered last time actually controls the Type I error, and then move on to the problem of multiple testing which will occupy us for a couple of lectures.

4 Analyzing the permutation test for two-sample testing

We observe $X_1, \ldots, X_{n_1} \sim P$ and $Y_1, \ldots, Y_{n_2} \sim Q$, and want to test if:

$$\begin{aligned} H_0 &: \quad P = Q \\ H_1 &: \quad P \neq Q. \end{aligned}$$

Let us introduce some notation: we suppose we are given a test statistic T which is a function of the observed data, for instance:

$$T(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \left| \frac{1}{n_1}\sum_{i=1}^{n_1} X_i - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \right| := t_{\text{obs}}.$$

We let $N = n_1 + n_2$, and denote the permutations of the data by $\{Z_1, \ldots, Z_{N!}\}$. We let:

$$\phi_{\text{perm}}(Z_{\text{obs}}) = \mathbb{I}\left[\left(\frac{1}{N!}\sum_{i=1}^{N!} \mathbb{I}(T(Z_i) > t_{\text{obs}}) \right) < \alpha \right].$$

We claim that:

$$\mathbb{P}_{H_0}(\phi_{\text{perm}}(Z_{\text{obs}}) = 1) \leq \alpha.$$

Proof: Note that the permutation test would reject the null only for test statistics that are in the upper α-quantile of the distribution of test statistics, i.e.:

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \phi_{\text{perm}}(Z_i).$$

Taking the expectation over Z_i under the null we obtain that,

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{E}_{H_0}[\phi_{\text{perm}}(Z_i)].$$

Under the null hypothesis each dataset Z_i has the same distribution as Z_{obs} so we obtain that:

$$\alpha \geq \frac{1}{N!} \sum_{i=1}^{N!} \mathbb{E}_{H_0}[\phi_{\text{perm}}(Z_{\text{obs}})],$$

i.e. that,

$$\mathbb{P}_{H_0}(\phi_{\text{perm}}(Z_{\text{obs}}) = 1) \leq \alpha,$$

as desired. Also note that up to some small quantization error (since the p-values that the permutation test produces are discrete), all of the above inequalities are actually equalities, i.e. the permutation test has Type I error that is very close to α.

5 Multiple Testing

Testing many hypotheses at once, is called multiple testing. The problem of multiple testing is one that is fundamental to a lot of science. Typical modern scientific discovery does not proceed in a simple fashion where we have a single hypothesis that we would like to test.

A good example is in the analysis of gene expression data. We measure the expression of tens of thousands of genes and we would like to know if any of them are associated with some phenotype (for example whether a person has a disease or not). Typically, the way this is done is that the scientist does tens of thousands of hypothesis tests, and then reports the associations that are significant, i.e. reports the tests where the null hypothesis was rejected.

This is very problematic:

Suppose we did 1000 hypothesis tests, and for each of them rejected the null when the p-value was less than $\alpha = 0.05$. How many times would you expect to falsely reject the null hypothesis? The answer is we would expect to reject the null hypothesis 50 times. So we really cannot report all the discovered associations (rejections) as significant because we expect many false rejections.

Another example is in vaccine trials. If we keep testing whether a vaccine is effective as time goes on, we will end up doing many tests.

The multiple testing problem is behind a lot of the "reproducibility crisis" of modern science. Many results that have been reported significant cannot be reproduced simply because they are false rejections. Too many false rejections come from doing multiple testing but not properly adjusting your tests to reflect the fact that many hypothesis tests are being done. The basic question is how to we adjust our p-value cutoffs to account for the fact that multiple tests are being done.

5.1 The Family-Wise Error Rate

We first need to define what the error control we desire is. Recall, the Type I error controls the probability of falsely rejecting the null hypothesis. We have seen that in order to control the Type I error we can simply threshold the p-value, i.e rejecting the null if the p-value $\leq \alpha$ controls the Type I error at α.

One possibility is to control the probability that we falsely reject *any* null hypothesis. This is called the Family-Wise Error Rate (FWER). The FWER is the probability of falsely rejecting the null hypothesis even once amongst the multiple tests. The basic question is then: how do we control the FWER?

5.2 Sidak correction

Suppose we do d hypothesis tests, and want to control the FWER at α. The Sidak correction says to reject any test if the p-value is smaller than:

$$\text{p-value} \leq 1 - (1 - \alpha)^{1/d} = \alpha_t,$$

so we reject any test if its p-value is less than α_t.

The main result is that: if the p-values are all *independent* then the FWER $\leq \alpha$.

Proof: Suppose that all the null hypotheses are true (this is called the *global null*). You can easily see that if this is not the case you can simply ignore all the tests for which the null is false. The probability of falsely rejecting a fixed test is α_t, so we correctly fail to reject it with probability $1 - \alpha_t$.

Since the p-values are all independent the probability of falsely rejecting any null hypothesis is:

$$\text{FWER} = 1 - (1 - \alpha_t)^d = \alpha.$$

5.3 Bonferroni correction

The main problem with the Sidak correction is that it requires the independence of p-values. This is unrealistic especially if you compute the test statistics for the different tests on the same set of data. The Bonferroni correction instead uses the union bound to avoid this assumption.

The Bonferroni correction says we reject any test if the p-value is smaller than:

$$\text{p-value} \leq \frac{\alpha}{d}.$$

The main result is that: The FWER $\leq \alpha$.

Proof: Suppose again that the global null is true. In this case,

$$\text{FWER} = \mathbb{P}\left(\bigcup_{i=1}^{d} \text{reject } H_{0i}\right) \leq \sum_{i=1}^{d} \mathbb{P}\left(\text{reject } H_{0i}\right) \leq \sum_{i=1}^{d} \frac{\alpha}{d} = \alpha,$$

where the first inequality follows from the union bound.

5.4 Holm's procedure

There are many possible improvements to the Bonferroni procedure. For instance, suppose that I told you that exactly (or at most) d_0 of the null hypotheses are truly nulls. Then you can see that we could have used the cut-off of $\frac{\alpha}{d_0}$ and still maintained control over the FWER.

As a thought experiment consider the following setting. You conduct $d = 5$ experiments and you observe p-values of $(0.7, 0.02, 0, 0, 0)$.

Intuitively, it seems like since we are absolutely sure that the last three experiments are non-nulls we should be able to use the cut-off of $\alpha/2$ for the remaining two tests, and still control the FWER.

At a high-level it seems intuitively clear to us that other p-values for $\{p_j\}_{j \neq i}$ contain information at least about the number of null hypotheses and we can use this to relax the correction for p_i. Holm's procedure translates this intuition into a rigorous procedure.

1. Order the p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.
2. If $p_{(1)} < \frac{\alpha}{d}$ then reject $H_{(1)}$ and move on, else stop and accept all H_i.
3. If $p_{(2)} < \frac{\alpha}{d-1}$ then reject $H_{(2)}$ and move on, else stop and accept $H_{(2)}, \ldots, H_{(d)}$.
   $\vdots$
4. If $p_{(d)} < \alpha$, then reject $H_{(d)}$, else accept $H_{(d)}$.

More succinctly, let

$$i^* = \min \left\{ i : p_{(i)} > \frac{\alpha}{d - i + 1} \right\},$$

and reject all $H_{(i)}$ for $i < i^*$.

Holm's procedure controls the FWER at level α. Importantly, Holm's procedure does not require independence of the p-values, and it strictly dominates the Bonferroni procedure.

Proof: Let I_0 denote the indices of the true nulls. First let us make an observation: if

$$\min_{i \in I_0} p_i > \frac{\alpha}{d_0},$$

then we reject none of the true nulls. This is because the first time we encounter a true null we would compare it to a threshold that is at most α/d_0, and if we fail to reject it we would not reject any of the other true nulls.

So the FWER is:

$$\text{FWER} \leq \mathbb{P}\left(\min_{i \in I_0} p_i \leq \frac{\alpha}{d_0} \right) \leq \alpha,$$

by the union bound.

5.5 Something to think about

In the above discussion we assumed that there was a single scientist doing a bunch of tests so he could appropriately correct his procedure for the multiple testing problem. One thing to ponder is really what error rate should we be controlling, i.e. maybe I am the editor of a journal, and I want to ensure that across all articles in my journal the FWER is $\leq \alpha$. Maybe I want this to be true across the entire field? Should I be adjusting my p-values for people in other disciplines? Sounds absurd but it actually makes sense if you think about each of these procedures and their implications for reproducibility.

6 False Discovery Rate

Suppose that we tested $d = 1000$ genes for association with some disease, we got a 1000 p-values, and 100 of them were less than 0.01. We'd expect that roughly $0.01d_0 \leq 0.01d = 10$ of these to be falsely rejected nulls, and perhaps this is not a bad tradeoff, i.e. if we rejected 100 nulls, we would spend only 10% of our time on falsely rejected nulls, i.e. we would make 90 real discoveries.

This suggests using a different error criterion. The FDR (false discovery rate) is the expected number of false rejections divided by the number of rejections.

Denote the number of false rejections as V, and the total number of rejections as R. Then the false discovery *proportion* is:

$$\text{FDP} = \begin{cases} \frac{V}{R} & \text{if } \ R > 0 \\ 0 & \text{if } \ R = 0. \end{cases}$$

The FDR is then defined as:

$$\text{FDR} = \mathbb{E}[\text{FDP}].$$

In this notation we can see that the FWER is:

$$\text{FWER} = \mathbb{P}(V \geq 1).$$

We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

Lecture 18
36-705

Recall that the FDR (false discovery rate) is

$$FDR = \mathbb{E}[FDP]$$

where the FDP (false discovery proportion) is

$$FDP = \frac{\text{number of false rejections}}{\text{number of rejections}} = \frac{V}{R}$$

where the ratio is defined to be 0 if there are no rejections. We will next consider how one can control the FDR. We will describe a procedure known as the Benjamini-Hochberg (BH) procedure.

0.1 The BH procedure

The BH procedure is one that controls the FDR under independence (i.e. the p-values are independent). There is a different version this procedure that works under dependence.

The procedure is:

1. Suppose we do d tests. Let us take the p-values $p_1, \ldots, p_d$, and sort them, i.e. we create the list: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(d)}$.

2. Define the thresholds:

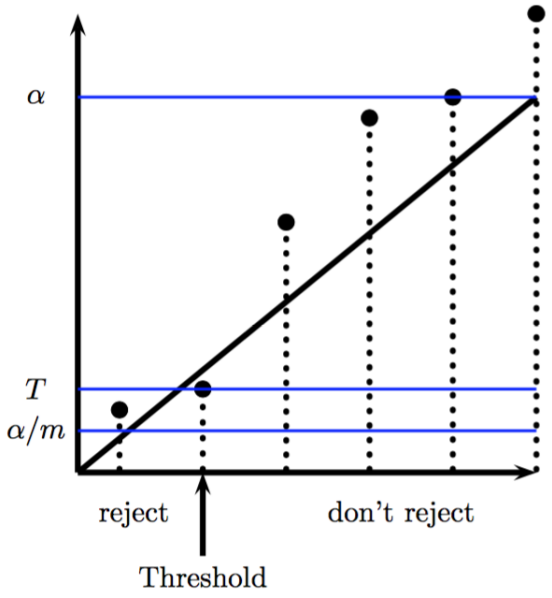
$$t_i = \frac{i\alpha}{d}.$$

3. Find the largest $i_{\max}$ such that

$$i_{\max} = \arg\max_i \{i : p_{(i)} < t_i\}.$$

4. Reject all nulls up to and including $i_{\max}$.

This might seem a bit confusing but here is a simple picture:



0.2 Properties of FDR

We have now seen a procedure that controls the FDR under some assumptions. One question of interest is how does FDR control compare to FWER control? Another is just: how do we interpret FDR control?

Interpreting FDR control:

The way to think about FDR control is: if we repeat our experiment many times, on average we control the FDP. This is not a statement about the individual experiment we did conduct, and really it does not say much about how likely it is that on a given experiment we have an FDP that is larger than a threshold (think about using Markov's inequality).

Connection to FWER:

1. The first connection is that under the global null (when all the null hypotheses are true) FDR control is equivalent to FWER control.

   Proof: Under the global null, any rejection is a false rejection. There are two possibilities: either we do not reject anything: in this case the FDP $= 0$. If we do reject any null hypothesis then our FDP is 1 (since $V = R$). So we have that:

   $$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{P}(V > 0) \times 1 + \mathbb{P}(V = 0) \times 0 = \mathbb{P}(V > 0) = \text{FWER}.$$

2. The second connection is that the FWER $\geq$ FDR always. This implies that controlling the FWER implies FDR control.

Proof: We can see that the following is a simple upper bound on the FDP:

$$\text{FDP} \leq \mathbb{I}(V \geq 1),$$

since if $V = 0$, FDP $= 0$, and if $V > 0$ then $V/R \leq 1$. Taking expectations of this expression gives:

$$\text{FDR} \leq \mathbb{P}(V \geq 1) = \text{FWER}.$$

The flip-side of this is that FDR control is less stringent so if this is the correct measure for you then you will have *more* power by controlling FDR (rather than controlling FWER).

1 Proving that BH controls FDR

The main result is the following:

Theorem: Suppose that the p-values are independent, the BH procedure controls the FDR at level α. In fact,

$$\text{FDR} \leq \frac{d_0 \alpha}{d} \leq \alpha.$$

Proof Intuition: Suppose that the BH procedure returned a value $i_{\max}$ then we know that,

$$p_{(i_{\max})} < \frac{i_{\max} \alpha}{d}.$$

We have rejected $i_{\max}$ hypotheses. At the cut-off $\frac{i_{\max}\alpha}{d}$ we expect that $\frac{d_0 i_{\max}\alpha}{d}$ nulls to be rejected. This gives us that the FDR should be roughly:

$$\text{FDR} \approx \frac{d_0 i_{\max} \alpha}{d i_{\max}} = \frac{d_0 \alpha}{d} \leq \alpha.$$

Formalizing this argument is a bit intricate: notice that $i_{\max}$ is a random variable and furthermore the numerator and denominator in the FDP are not independent random variables so we need to be careful while taking the expectation of the ratio.

Here is a proof from Emmanuel Candes' Stat 300c notes at Stanford. These notes are in general a great resource that delve much deeper into theoretical aspects of multiple testing.

Proof: When $d_0 = 0$ there are no false discoveries so there is nothing to prove. We will suppose that $d_0 \geq 1$, and denote the set of nulls as I_0. Let us define:

$$V_i = \mathbb{I}(H_i \text{ is rejected}),$$

then we can write the FDP as:

$$\text{FDP} = \sum_{i \in I_0} \frac{V_i}{\max\{R, 1\}},$$

notice that taking the max in the denominator just avoids the 0/0 problem, and is a shorthand way of writing the FDP. Suppose we could prove that:

$$\mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{\alpha}{d},$$

then we are done since,

$$\text{FDR} = \sum_{i \in I_0} \mathbb{E}\left[\frac{V_i}{\max\{R, 1\}}\right] = \frac{d_0 \alpha}{d}.$$

To prove the claim we first re-write:

$$\frac{V_i}{\max\{R, 1\}} = \sum_{k=1}^{d} \frac{V_i \mathbb{I}(R = k)}{k},$$

noting that if $R = 0$ both the LHS and RHS are 0. We now need to make some further observations:

1. Suppose that there are k rejections, then we can rewrite:

$$V_i = \mathbb{I}(H_i \text{ is rejected}) = \mathbb{I}(p_i \leq k\alpha/d).$$

2. Suppose that $p_i \leq \alpha k/n$, then we take p_i and set it to 0, and denote the number of rejections as $R(p_i \to 0)$ and note that $R(p_i \to 0)$ is exactly the same as R. On the other hand if $p_i > \alpha k/n$ then $V_i = 0$. So we can write:

$$V_i \mathbb{I}(R = k) = V_i \mathbb{I}(R(p_i \to 0) = k).$$

Now, returning to the main thread suppose we considered the conditional expectation:

$$\begin{aligned}\mathbb{E}\left[\frac{V_i \mathbb{I}(R = k)}{k} \Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right] &= \frac{\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)\mathbb{I}(R(p_i \to 0) = k) | p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d]}{k} \\ &= \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d},\end{aligned}$$

where we use the fact that conditional on the other p-values $\mathbb{I}(R(p_i \to 0) = k)$ is deterministic and that the p-values have uniform distribution under the null, and that the nulls are independent so that:

$$\mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d) | p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d] = \mathbb{E}[\mathbb{I}(p_i \leq k\alpha/d)] = k\alpha/d.$$

Now, by iterated expectations:

$$\mathbb{E}\left[\frac{V_i}{\max\{R,1\}}\right] = \sum_{k=1}^{d} \mathbb{E}\left[\mathbb{E}\left[\frac{V_i \mathbb{I}(R=k)}{k} \Big| p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_d\right]\right]$$
$$= \sum_{k=1}^{d} \frac{\mathbb{I}(R(p_i \to 0) = k)\alpha}{d} = \frac{\alpha}{d},$$

which was the claim we needed to prove. □

Today we will discuss confidence sets and ways to construct them. We have discussed point estimation so far where the goal is to construct an estimate $\widehat{\theta}(X_1, \ldots, X_n)$ of some parameter θ after observing $\{X_1, \ldots, X_n\}$.

The setting here is that we have a statistical model (i.e. a collection of distributions) $\mathcal{P}$. Let $C_n(X_1, \ldots, X_n)$ be a set constructed using the observed data $X_1, \ldots, X_n$. This is a random set. C_n is a $1-\alpha$ confidence set for a parameter θ if:

$$P(\theta \in C_n(X_1, \ldots, X_n)) \geq 1 - \alpha, \quad \text{for all} \;\; P \in \mathcal{P}.$$

This means that no matter which distribution in $\mathcal{P}$ generated the data, the interval guarantees the coverage property described above. Some people would refer to such intervals as *honest* confidence intervals to make explicit the fact that the coverage is uniform over the model.

At a high-level, the confidence interval gives us some idea of how precise our estimate of the unknown parameter is, i.e. a wide interval indicates that our (point) estimate is imprecise.

Example: Suppose that we considered, $X_1, \ldots, X_n \sim U[0, \theta]$. Then we could construct the usual point estimate $\widehat{\theta} = X_{(n)}$. We could perhaps consider two types of confidence intervals:

$$C_1 = [a_1\widehat{\theta}, b_1\widehat{\theta}], \quad 1 \leq a_1 \leq b_1$$
$$C_2 = [\widehat{\theta} + a_2, \widehat{\theta} + b_2], \quad a_2, b_2 \geq 0.$$

Let us try to calculate the coverage probabilities of these two types of intervals. As a preliminary observe that:

$$\mathbb{P}(\widehat{\theta} \leq t) = \left(\frac{t}{\theta}\right)^n, \quad \text{for} \;\; 0 \leq t \leq \theta.$$

1. C_1: We can compute that,

$$\mathbb{P}(\theta \in C_1) = \mathbb{P}(\widehat{\theta} \leq \theta/a_1, \widehat{\theta} \geq \theta/b_1)$$
$$= \mathbb{P}(\widehat{\theta} \leq \theta/a_1) - \mathbb{P}(\widehat{\theta} \leq \theta/b_1)$$
$$= \left(\frac{1}{a_1}\right)^n - \left(\frac{1}{b_1}\right)^n.$$

So we have that for instance choosing $a_1 = 1$, $b_1 = \left(\frac{1}{\alpha}\right)^{1/n}$ guarantees us that this confidence interval has coverage probability exactly $1 - \alpha$.

2. C_2: Similarly we have that,

$$\begin{aligned}\mathbb{P}(\theta \in C_2) &= \mathbb{P}(\widehat{\theta} \leq \theta - a_2, \widehat{\theta} \geq \theta - b_2) \\ &= \mathbb{P}(\widehat{\theta} \leq \theta - a_2) - \mathbb{P}(\widehat{\theta} \leq \theta - b_2) \\ &= \left(\frac{\theta - a_2}{\theta}\right)^n - \left(\frac{\theta - b_2}{\theta}\right)^n.\end{aligned}$$

Notice that now the coverage probability depends on the unknown parameter θ (which is undesirable). Furthermore, if we choose any constants (a_2, b_2) (say depending only on the desired coverage probability α), then as $\theta \to \infty$ we have that the interval has coverage probability that tends to 0, i.e. the interval is not honest for any constants (a_2, b_2).

We will now discuss a few different ways of constructing confidence intervals. Although superficially different most of these methods are roughly the same.

2 Inverting a test

We discussed this method in the last lecture. We suppose that we have a (family of) test(s) for the hypotheses:

$$\begin{aligned}H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0.\end{aligned}$$

These tests have a rejection region and a corresponding acceptance region (where we fail to reject the null). Denote the acceptance region for the test of $\theta = \theta_0$ as $A(\theta_0)$. This is a subset of the sample space.

Given observed data $\{X_1, \ldots, X_n\}$ we consider the random set:

$$C(X_1, \ldots, X_n) = \{\theta_0 : \{X_1, \ldots, X_n\} \in A(\theta_0)\}.$$

Our confidence set is simply the set of parameters θ_0 that we would fail to reject using our family of tests. If our family of tests has level α then the set $C(X_1, \ldots, X_n)$ is a $1 - \alpha$ confidence set.

To see this observe that since our test controls the Type I error we have that for any parameter θ_0,

$$\mathbb{P}_{\theta_0}(\{X_1, \ldots, X_n\} \notin A(\theta_0)) \leq \alpha,$$

so with probability at least $1-\alpha$ we have that, $\{X_1, \ldots, X_n\} \in A(\theta_0)$ and hence that $\theta_0 \in C(X_1, \ldots, X_n)$.

We can also construct tests using confidence intervals, i.e. consider the test that rejects the null hypothesis $\theta = \theta_0$ if $\theta_0 \notin C(X_1, \ldots, X_n)$, then if $C(X_1, \ldots, X_n)$ is a $1-\alpha$ confidence interval this test has level α, i.e.

$$\mathbb{P}_{\theta_0}(\text{reject null } \theta = \theta_0) = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X_1, \ldots, X_n)) \leq \alpha.$$

Let us quickly re-visit the uniform example. Suppose we observe $X_1, \ldots, X_n \sim U[0, \theta]$ and would like to construct a confidence interval, then one method would be to invert the LRT, i.e. we compute the likelihood-ratio for testing $H_0 : \theta = \theta_0$ as if $\theta \geq \max_i X_i$ then:

$$\text{LR} = \frac{\frac{1}{\theta_0^n}}{\frac{1}{(\max_i X_i)^n}} = \frac{(\max_i X_i)^n}{\theta_0^n},$$

and we note that we reject the null for small values of this quantity, i.e. we reject the null if

$$\frac{(\max_i X_i)^n}{\theta_0^n} \leq k_\alpha,$$

for an appropriate choice of k_α. So if we consider the confidence interval obtained by inverting this test, we see that it has the form:

$$C(X_1, \ldots, X_n) = \left\{ \theta : \max_i X_i \leq \theta \leq \frac{\max_i X_i}{k_\alpha^{1/n}} \right\},$$

which is precisely the type of multiplicative interval that we studied in the last lecture (we also calculated a value for k_α that ensures that $C(X_1, \ldots, X_n)$ has coverage $1-\alpha$ in that lecture). This just highlights that in this case, we could have obtained the right kind of interval in a less ad hoc manner (by inverting the LRT).

Example: Suppose we observe $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$, and want to construct a confidence interval for λ.

As our test, suppose we use the LRT, i.e. we define the likelihood ratio:

$$\Lambda = \frac{\lambda_0^n \exp(-\lambda_0 \sum_i X_i)}{(\frac{1}{\overline{X}})^n \exp(-n)}.$$

The acceptance region has the form:

$$A(\lambda_0) = \left\{ \{X_1, \ldots, X_n\} : \left(\lambda_0 \sum X_i\right)^n \exp(-\lambda_0 \sum X_i) \geq k_\alpha(\lambda_0) \right\},$$

where $k_\alpha(\lambda_0)$ needs to be chosen appropriately to control the Type I error. Observe that since $X_i \sim \text{Exp}(\lambda_0)$, $\lambda_0 X_i \sim \text{Exp}(1)$, so $k_\alpha(\lambda_0)$ does not depend on λ_0. Once we determine the cut-off we would obtain the confidence interval by collecting:

$$C(X_1, \ldots, X_n) = \{\lambda : \left(\lambda \sum X_i\right)^n \exp(-\lambda \sum X_i) \geq k_\alpha\},$$

which is an expression that can be solved numerically. Determining k_α and then finding the confidence set can be quite tedious to do exactly (see the Casella and Berger book) and an alternative would be to use large-sample (asymptotic approximations).

3 Inverting Probability Inequalities

In some simple cases, we can use tail bounds to derive confidence intervals. These typically have the advantage of being exact, finite-sample intervals. However, they are rarely used in practice for many reasons including: (1) we do not always have tail bounds for estimators of interest (2) there are usually imprecisely known constants in tails bounds (3) related to (2) they are often very conservative (i.e. the intervals are often too wide to be useful).

Here are a couple of examples:

Example 1 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. By Hoeffding's inequality:*

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(|\widehat{p} - p| > \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$ where $C = (\widehat{p} - \epsilon_n, \widehat{p} + \epsilon_n)$.

Example 2 *Let $X_1, \ldots, X_n \sim F$. Suppose we want a* **confidence band** *for F. We can use VC theory. Remember that*

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Let

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Then

$$\mathbb{P}\left(\sup_x |F_n(x) - F(x)| > \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}\right) \leq \alpha.$$

Hence,

$$P_F(L(t) \leq F(t) \leq U(t) \text{ for all } t) \geq 1 - \alpha$$

for all F, where

$$L(t) = \widehat{F}_n(t) - \epsilon_n, \quad U(t) = \widehat{F}_n(t) + \epsilon_n.$$

We can improve this by taking

$$L(t) = \max\left\{\widehat{F}_n(t) - \epsilon_n,\ 0\right\}, \quad U(t) = \min\left\{\widehat{F}_n(t) + \epsilon_n,\ 1\right\}.$$

3.1 Pivots

Another useful way of attempting to construct confidence intervals is to base the intervals on *pivots*. A pivot is a function of the data and the unknown parameter θ – $Q(X_1, \ldots, X_n, \theta)$ – whose distribution does not depend on θ.

Let us consider two examples:

1. Suppose that $X_1, \ldots, X_n \sim N(\theta, 1)$ then we can see that $Q(X_1, \ldots, X_n) = \overline{X_n} - \theta \sim N(0, 1/n)$ and so the distribution of Q does not depend on θ.
2. Suppose we consider $X_1, \ldots, X_n \sim U[0, \theta]$ and we consider the function:

   $$Q(X_1, \ldots, X_n, \theta) = \frac{\max_i X_i}{\theta},$$

   has distribution:

   $$P(Q(X_1, \ldots, X_n, \theta) \leq t) = \begin{cases} t^n & 0 \leq t \leq 1 \\ 1 & t \geq 1. \end{cases}$$

   Once again the distribution does not depend on θ.

Given a pivot we can construct confidence intervals in a simple way. Since the distribution of Q does not depend on θ, we can find a, b which do not depend on θ such that:

$$\mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

Now, we construct our confidence interval as:

$$C(X_1, \ldots, X_n) = \{\theta : a \leq Q(X_1, \ldots, X_n, \theta) \leq b\}.$$

By our construction:

$$\mathbb{P}_\theta(\theta \in C(X_1, \ldots, X_n)) = \mathbb{P}_\theta(a \leq Q(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha.$$

Going back to our two examples we find that we will once again obtain the now standard intervals for the two problems (the additive interval for the Gaussian mean, and the multiplicative scale interval for the uniform parameter).

4 Tests Versus Confidence Intervals

Confidence intervals are more informative than tests. Intuitively, p-values are more informative than an accept/reject decision because it summarizes all the significance levels for which we would reject the null hypothesis. Similarly, a confidence interval is more informative that a test because it summarizes all the parameters for which we would (fail to) reject the null hypothesis. More practically, a confidence interval tells us something about the "effect size" as well as something about the uncertainty in our estimate of the "effect size".

Look at Figure 1. Suppose we are testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. We see 5 different confidence intervals. The first two cases (top two) correspond to not rejecting H_0. The other three correspond to rejecting H_0. Reporting the confidence intervals is much more informative than simply reporting "reject" or "don't reject."

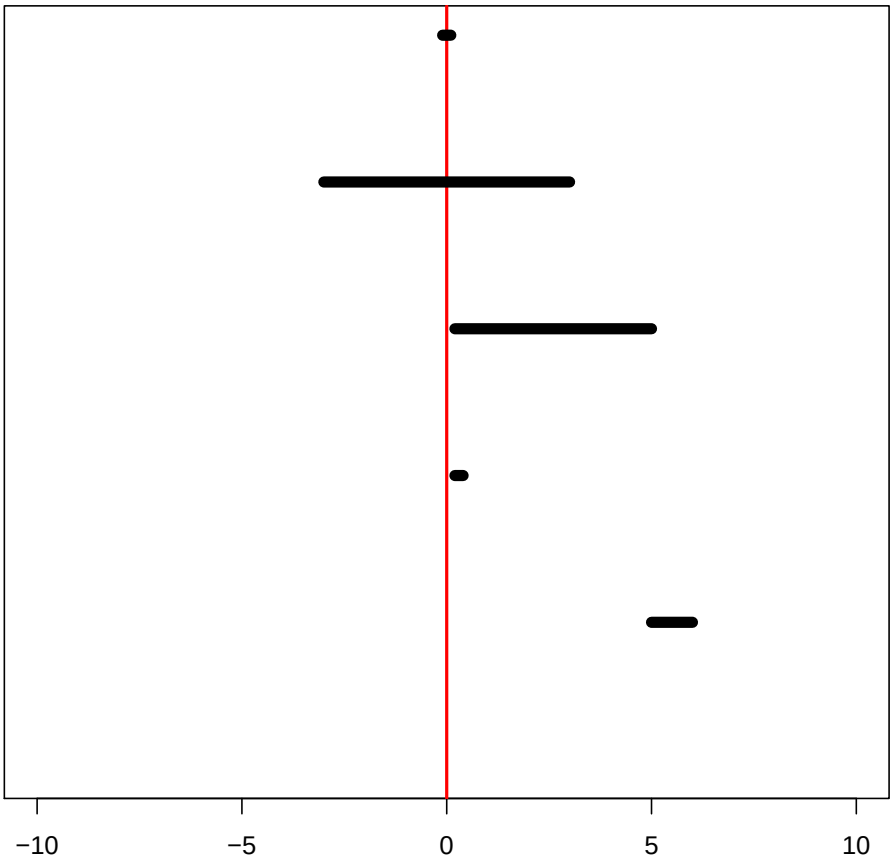


Figure 1: Five examples: 1. Not significant, precise. 2. Not significant, imprecise. 3. Barely significant, imprecise. 4. Barely significant, precise. 5. Significant and precise.

Lecture Notes 19
36-705

So far we have focused on parametric models. Now we will turn our attention to nonparametric inference. In particular, we will discuss estimating quantities known as *statistical functionals.*

1 Statistical Functional

A statistical functional is a map ψ that maps a distribution P to a real number (or vector). Examples include:

the mean: $\psi(P) = \int xp(x)dx$

the variance $\psi(P) = \int x^2 p(x)dx - \left(\int xp(x)dx\right)^2$

the median $\psi(P) = F^{-1}(1/2)$ where F is the cdf

Sometimes people refer to an unknown statistical functional as a parameter. This should not be confused with the idea of a parameter in a parametric model.

At this point, let me remind you of some notation. If g is any function then we write

$$\int g(x)dP(x) = \begin{cases} \int g(x)p(x)dx & \text{if } X \text{ is continuous} \\ \sum_j g(x_j)p(x_j) & \text{if } X \text{ is discrete.} \end{cases}$$

We also write this as $\int g(x)dF(x)$ where F is the cdf.

2 Plug-In Estimators

Let $X_1, \ldots, X_n \sim P$. Recall that the empirical distribution P_n is the distribution that puts mass $1/n$ at each data point. Thus

$$P_n(A) = \frac{1}{n}\sum_i I(X_i \in A).$$

The corresponding cdf — the empirical cdf — is

$$F_n(t) = \frac{1}{n}\sum_i I(X_i \leq t).$$

If g is any function then

$$\int g(x)dP_n(x) = \int g(x)dF_n(x) = \frac{1}{n}\sum_i g(X_i).$$

If $\psi(P)$ is a statistical functional, the *plug-in estimator* is

$$\widehat{\psi}_n = \psi(P_n).$$

For example, if $\psi(P) = \int x dP(x)$ is the mean then the plug-in estimator is

$$\widehat{\psi}_n = \psi(P_n) = \int x dP_n(x) = \frac{1}{n}\sum_i X_i.$$

If

$$\psi(P) = \int (x-\mu)^2 dP(x) = \int x^2 dP(x) - \left(\int x dP(x)\right)^2$$

is the variance then plug-in estimator is

$$\begin{aligned}\widehat{\psi}_n = \psi(P_n) &= \int x^2 dP_n(x) - \left(\int x dP_n(x)\right)^2 = \frac{1}{n}\sum_i X_i^2 - \left(\frac{1}{n}\sum_i X_i\right)^2 \\ &= \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2.\end{aligned}$$

Let's consider a bivariate example. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$. The covariance is

$$\psi(P) = \mathbb{E}[XY] - \mathbb{E}[X]\ \mathbb{E}[Y] = \int xy dP(x,y) - \int x dP(x) \int y dP(y)$$

and the plug-in estimator is

$$\widehat{\psi}_n = \frac{1}{n}\sum_i X_i Y_i - \overline{X}_n \overline{Y}_n = \sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n).$$

Is plug-in estimation a good idea? It depends. If the functional ψ satisfies some weak regularity conditions and P is well-behaved (for example, has some moments) that $\widehat{\psi}_n$ can be a good estimator. We won't go into details on this here.

The next question is: how do we do inference for a statistical functional? We'll dicuss two approaches: influence functions and the bootstrap.

3 Influence Functions

Let δ_x denote a point mass at x. The influence function for a statistical functional ψ is defined by

$$\varphi(x) = \lim_{\epsilon \to 0} \frac{\psi((1-\epsilon)P + \epsilon\delta_x) - \psi(P)}{\epsilon}.$$

For example, if $\psi(P)$ is the mean of P then

$$\frac{\psi((1-\epsilon)P + \epsilon\delta_x) - \psi(P)}{\epsilon} = \frac{(1-\epsilon)\psi(P) + \epsilon x) - \psi(P)}{\epsilon} = x - \psi(P).$$

Hence $\varphi(x) = x - \psi(P)$.

Let's consider another example. Suppose that $\psi(P)$ is the variance σ^2, that is, $\psi(P) = \int x^2 dP(x) - (\int x dP(x))^2$. Let μ denote the mean. Then

$$\psi((1-\epsilon)P + \epsilon\delta_x) - \psi(P) = (1-\epsilon)\int x^2 dP(x) + \epsilon x^2 - [(1-\epsilon)\mu + \epsilon x]^2 - (\int x^2 dP(X) - \mu^2)$$

and so

$$\varphi(x) = \lim_{\epsilon \to 0} \frac{\psi((1-\epsilon)P + \epsilon\delta_x) - \psi(P)}{\epsilon} = x^2 - \int x^2 dP(x) - 2\mu x.$$

Notice that the influence function φ is itself a statistical functional: it depends on P. For example, if ψ is the mean then $\varphi(x) = x - \psi = x - \psi(P)$. So we can write

$$\varphi(x) = \varphi(x, P).$$

The empirical influence function is an estimate of the influence function obtained by replacing P by P_n, that is,

$$\widehat{\varphi}(x) = \varphi(x, P_n).$$

So, for example, when $\varphi(x) = x - \psi = x - \psi(P)$ for the mean, the empirical influence function is

$$\widehat{\varphi}(x) = x - \psi(P_n) = x - \overline{X}_n.$$

Now we can use the following result.

Theorem 1 *If ψ satisfies some regularity conditions then*

$$\sqrt{n}(\psi(P_n) - \psi(x)) \rightsquigarrow N(0, \tau^2)$$

where

$$\tau^2 = \int \varphi^2(x) dP(x).$$

A consistent estimate of τ^2 *is*

$$\widehat{\tau}^2 = \frac{1}{n}\sum_i \widehat{\varphi}(X_i).$$

Hence, an asymptotic $1-\alpha$ *confidence interval for* $\psi(P)$ *is*

$$\widehat{\psi}_n \pm \frac{z_{\alpha/2}\widehat{\tau}}{\sqrt{n}}.$$

In the case where $\psi(P)$ is the mean we see that

$$\widehat{\tau}^2 = \frac{1}{n}\sum_i \widehat{\varphi}^2(X_i) = \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2 = S_n^2$$

and the confidence interval is

$$\overline{X}_n \pm \frac{z_{\alpha/2}S_n}{\sqrt{n}}.$$

Of course, we did not need all this machinery to arrive at this confidence interval, but in more complicated cases these methods can be very useful. Let's consider another example.

Let $\psi(P)$ be the r^{th} quantile for $0 < r < 1$. Assume that the cdf is strictly increasing so that

$$\psi(P) = F^{-1}(r).$$

The plug-in estimator $\widehat{\psi}_n$ is the r^{th} sample quantile

$$\widehat{\psi}_n = \inf\{x : F_n(x) \geq r\}.$$

The influence function is

$$\varphi(x) = \begin{cases} \frac{r-1}{p(\psi)} & x \leq \psi \\ \frac{r}{p(\psi)} & x > \psi \end{cases}$$

where p is the density function. Hence,

$$\tau^2 = \int \varphi^2(x)dP(x) = \frac{r(1-r)}{p^2(\psi)}.$$

To estimate τ^2 we would need to estimate the density p. We'll discuss how to do that later in the course. However, there are simpler ways to get confidence intervals for quantiles.

There are many subtle technicalities associated with influence functions. These are beyond the scope of the course but if you are interested, search for *semiparametric inference*.

Lecture Notes 20
36-705

We have discussed plug-in estimators and influence functions. Today we consider a different nonparametric approach for getting confidence intervals for plug-in estimators: the bootstrap.

1 Monte Carlo

Before we get to the bootstrap, we should briefly discuss the Monte Carlo method.

Let g be a function and let P be a distribution. Suppose we want to know the mean of g, that is $\mathbb{E}[g(X)] = \int g(x)p(x)dx$. One way to do this is to do the integral $\int g(x)p(x)dx$. Another approach is simulation, also known as Monte Carlo. We draw a large sample $X_1, \ldots, X_B \sim P$. Then, by the law of large numbers

$$\frac{1}{B}\sum_{j=1}^{B} g(X_j) \xrightarrow{P} \mathbb{E}[g(X)].$$

Since we can simulate as many observations as we want, we can make the estimate very close to $\mathbb{E}[g(X)]$.

The same is true for the variance. We can get the variance of $g(X)$ by integration:

$$\text{Var}[g(X)] = \int g^2(x)p(x) - \left(\int g(x)p(x)\right)^2.$$

But we can also compute the sample variance from the simulated values and, again, by the law of large numbers

$$\frac{1}{n}\sum_j (g(X_j) - \overline{g})^2 \xrightarrow{P} \text{Var}[g(X)]$$

where $\overline{g} = \frac{1}{B}\sum_j g(X_j)$.

Now suppose that $T = g(X_1, \ldots, X_n)$ is a function of n iid variables. The mean is

$$\int \cdots \int g(x_1, \ldots, x_n)p(x_1)\cdots p(x_n)dx_1 \ \ldots \ dx_n$$

which is an n-dimensional integral. We can still use Monte-Carlo if we draw samples of size n each time. When we draw $X_1, \ldots, X_n \sim p$, we can think of this as one draw from the joint

density $p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n)$. In other words we do the following:

$$\begin{aligned} \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_1 = g(X_1, \ldots, X_n) \\ \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_2 = g(X_1, \ldots, X_n) \\ & \vdots \\ \text{draw } X_1, \ldots, X_n \sim P \quad & \text{compute } T_B = g(X_1, \ldots, X_n). \end{aligned}$$

Then $T_1, T_2, \ldots$ are draws from the distribution of $T = g(X_1, \ldots, X_n)$. Again, by the law of large numbers, as $B \to \infty$,

$$\frac{1}{B} \sum_{j=1}^{B} T_j \xrightarrow{P} \mathbb{E}[T] = \mathbb{E}[g(X_1, \ldots, X_n)]$$

and

$$\frac{1}{n} \sum_j (T_j - \overline{T})^2 \xrightarrow{P} \text{Var}[T] = \text{Var}[g(X_1, \ldots, X_n)]$$

where $\overline{T} = \frac{1}{B} \sum_j T_j$.

2 Bootstrap Variance Estimation

Let $X_1, \ldots, X_n \sim P$ and let $T = g(X_1, \ldots, X_n)$ be some statistic. Of course, the case we have in mind is that $T = g(X_1, \ldots, X_n)$ is an estimator of some parameter. Our goal is to estimate the standard error, that is the standard deviation of T. As a concrete example, think of $T = g(X_1, \ldots, X_n)$ as the median of the data.

If we knew P, we could use Monte Carlo to estimate $\tau^2 = \text{Var}[T]$. The idea of the bootstrap is to estimate P with the empirical distribution P_n. In other words, τ^2 is a statistical functional so we can write it as $\tau^2(P)$. We will estimate $\tau^2(P)$ with $\tau^2(P_n)$. Computing $\tau^2(P_n)$ is not easy to do analytically but now we can use Monte Carlo. We just need to simulate many times from P_n. When we draw a sample from P_n we usually denote the draws by X_i^*. So

$$X_1^*, \ldots, X_n^* \sim P_n$$

denotes a sample from P_n. We call $X_1^*, \ldots, X_n^*$ a bootstrap sample.

Specifically:

$$\begin{aligned} \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_1 = g(X_1^*, \ldots, X_n^*) \\ \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_2 = g(X_1^*, \ldots, X_n^*) \\ & \vdots \\ \text{draw } X_1^*, \ldots, X_n^* \sim P_n \quad & \text{compute } T_B = g(X_1^*, \ldots, X_n^*). \end{aligned}$$

Again, by the law of large numbers, as $B \to \infty$,

$$\widehat{\tau}^2 = \frac{1}{n}\sum_j (T_j - \overline{T})^2 \xrightarrow{P} \tau^2(P_n)$$

where $\overline{T} = \frac{1}{B}\sum_j T_j$. Note that there are two things going on:

1. We estimate $\tau^2(P)$ with $\tau^2(P_n)$
2. We approximate $\tau^2(P_n)$ with the Monte Carlo approximation $\widehat{\tau}^2$.

These are two distinct ideas. The first is plug-in estimation an the second is Monte Carlo.

How do we draw a sample from P_n? Remember that P_n puts mass $1/n$ at eachy data point. The distribution looks like this:

value	X_1	X_2	$\cdots$	X_n
mass	$1/n$	$1/n$	$\cdots$	X_n

To draw X_1^* we just draw one datapoint at random. To draw X_2^* we again draw one datapoint at random. We repeat this n times to get one bootstrap sample. Note that this is equivalent to drawing n times from the data *with replacement*. Draw a point; put it back; drsw a point; put it back; etc. For this reason, people often describe drawing a bootstrap sample as *resampling the data*. But is is best regarded as drawing n times from P_n.

Now we can use the bootstrap for statistical inference. Suppose that $\widehat{\psi}_n = g(X_1, \ldots, X_n)$ is an estimator. For example, it could be a plug-in estimator. Now we apply the bootstrap method. We sample n observations from P_n and re-compute the estimator. We repeat B times to get $\widehat{\tau}$ which is the estimated standard error of $\widehat{\psi}_n$.

3 Bootstrap Confidence Intervals

We can also use the bootstrap to get a confidence interval for ψ. In fact, I will describe three methods.

Method 1. If $\widehat{\psi}_n$ is asymptotically Normal then a $1-\alpha$ confidence interval is

$$\widehat{\psi}_n \pm z_{\alpha/2}\widehat{\tau}$$

where $\widehat{\tau}$ is the bootstrap estimate of the standard error.

Method 2: The Percentile Interval. Let $\widehat{\psi}_1^*, \ldots, \widehat{\psi}_B^*$ denote the bootstrap values of the estimator. The percentile confidence interval is

$$C_n = [\widehat{\psi}^*_{(\alpha/2)}, \widehat{\psi}^*_{(1-\alpha/2)}]$$

where $\widehat{\psi}^*_{(\alpha/2)}$ is the $\alpha/2$ quantile of $\widehat{\psi}^*_1, \ldots, \widehat{\psi}^*_B$ and $\widehat{\psi}^*_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of $\widehat{\psi}^*_1, \ldots, \widehat{\psi}^*_B$.

Method 3: The Basic Bootstrap (Reverse Perentile). Suppose for a moment that we knew the distribution

$$G_n(t) = P(\sqrt{n}(\widehat{\psi}_n - \psi) \leq t).$$

Let $g_{\alpha/2} = G_n^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = G_n^{-1}(1-\alpha/2)$. Let

$$C_n = \left[\widehat{\psi}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}}, \widehat{\psi}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right].$$

Now

$$\begin{aligned}\mathbb{P}(\psi \in C_n) &= \mathbb{P}\left(g_{\alpha/2} \leq \sqrt{n}(\widehat{\psi}_n - \psi) \leq g_{1-\alpha/2}\right) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.\end{aligned}$$

This interval looks strange because you are used to Normal-based intervals. In fact, if G_n is Normal, this interval can be re-written to look like the usual interval due to the symmetry of the Normal.

We do not know G_n so we can't use this interval. But we can estimate G_n with the bootstrap. We define

$$\widehat{G}_n(t) = \frac{1}{n}\sum_{j=1}^{B} I(\sqrt{n}(\widehat{\psi}^*_j - \widehat{\psi}) \leq t).$$

We then estimate $g_{\alpha/2} = G_n^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = G_n^{-1}(1-\alpha/2)$ with $\widehat{g}_{\alpha/2} = \widehat{G}_n^{-1}(\alpha/2)$ and $\widehat{g}_{1-\alpha/2} = \widehat{G}_n^{-1}(1-\alpha/2)$. The confidence interval is

$$C_n = \left[\widehat{\psi}_n - \frac{\widehat{g}_{1-\alpha/2}}{\sqrt{n}}, \widehat{\psi}_n - \frac{\widehat{g}_{\alpha/2}}{\sqrt{n}}\right].$$

Note that

$$\widehat{g}_{(\alpha/2)} = \sqrt{n}(\widehat{\psi}^*_{(\alpha/2)} - \widehat{\psi})$$

and

$$\widehat{g}_{1-(\alpha/2)} = \sqrt{n}(\widehat{\psi}^*_{1-(\alpha/2)} - \widehat{\psi})$$

so that

$$\widehat{\psi} - \frac{\widehat{g}_{1-(\alpha/2)}}{\sqrt{n}} = 2\widehat{\psi} - \widehat{\psi}^*_{1-(\alpha/2)}$$

and

$$\widehat{\psi} - \frac{\widehat{g}_{(\alpha/2)}}{\sqrt{n}} = 2\widehat{\psi} - \widehat{\psi}^*_{(\alpha/2)}.$$

Therefore, we can write

$$C_n = \left[2\widehat{\psi} - \widehat{\psi}^*_{1-(\alpha/2)}, 2\widehat{\psi} - \widehat{\psi}^*_{(\alpha/2)}\right].$$

Again, it looks weird but it follows from the calculations.

4 The Parametric Bootstrap

The bootstrap can also be used for parametric models. Instead of drawing $X_1^*, \ldots, X_n^* \sim P_n$ we instead draw $X_1^*, \ldots, X_n^* \sim p(x; \widehat{\theta})$. The res is the same.

5 Variants

There are many many many papers that have been written about the bootstrap. There are many different versions: the block bootstrap for time-series, the residual bootstrap or the wild bootstrap for regression, the smooth bootstrap, the bias-corrected bootstrap, and many others.

6 Why Does the Bootstrap Work?

We want that the quantiles of the bootstrap distribution of our statistic should be close to the quantiles its actual distribution. Let

$$\widehat{F}_n(t) = \mathbb{P}_n(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n),$$

be the CDF of the bootstrap distribution, and

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t),$$

be the CDF of the true sampling distribution of our statistic. We want to show that

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \to 0.$$

This turns out to be true in quite a bit of generality, only requiring mild conditions (Hadamard differentiability) but we will prove it in the simplest case: when $\widehat{\theta}_n$ is a sample mean. In this case there are much simpler ways to construct confidence intervals (using Normal approximations) but that is not really the point.

Suppose that $X_1, \ldots, X_n \sim P$ where X_i has mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.

Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t). \tag{1}$$

We want to show that

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \leq t \;\Big|\; X_1, \ldots, X_n\Big)$$

is close to F_n.

Theorem 1 (Bootstrap Theorem) *Suppose that* $\mu_3 = \mathbb{E}|X_i|^3 < \infty$. *Then,*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

To prove this result, let us recall that Berry-Esseen Theorem.

Theorem 2 (Berry-Esseen Theorem) *Let* $X_1, \ldots, X_n$ *be i.i.d. with mean* μ *and variance* σ^2. *Let* $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. *Let* $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ *be the sample mean and let* Φ *be the cdf of a* $N(0,1)$ *random variable. Let* $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. *Then*

$$\sup_z \Big|\mathbb{P}(Z_n \leq z) - \Phi(z)\Big| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \tag{2}$$

Proof of the Bootstrap Theorem. Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \mathrm{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n)|X_1, \ldots, X_n)$. Now, by the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\leq \sup_t |F_n(t) - \Phi_\sigma(t)| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0, \sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned}\mathrm{I} &= \sup_t |F_n(t) - \Phi_\sigma(t)| = \sup_t \Big|\mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t\right) - \mathbb{P}\left(\sigma Z \leq t\right)\Big| \\ &= \sup_t \left|\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \leq \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \leq \frac{t}{\sigma}\right)\right| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}.\end{aligned}$$

Using the same argument on the third term, we have that

$$\mathrm{III} = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \leq \frac{33}{4}\frac{\widehat{\mu}_3}{\widehat{\sigma}^3\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1} |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. By the strong law of large numbers, $\widehat{\mu}_3$ converges almost surely to μ_3 and $\widehat{\sigma}$ converges almost surely to σ. So, almost surely, for all large n, $\widehat{\mu}_3 \leq 2\mu_3$ and $\widehat{\sigma} \geq (1/2)\sigma$ and $\mathrm{III} \leq \frac{33}{4}\frac{4\mu_3}{\sqrt{n}}$. From the fact that $\widehat{\sigma} - \sigma = O_P(\sqrt{1/n})$ it may be shown that $\mathrm{II} = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_P(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.) This completes the proof. □

So far we have focused on the mean. Similar theorems may be proved for more general parameters. The details are complex so we will not discuss them here.

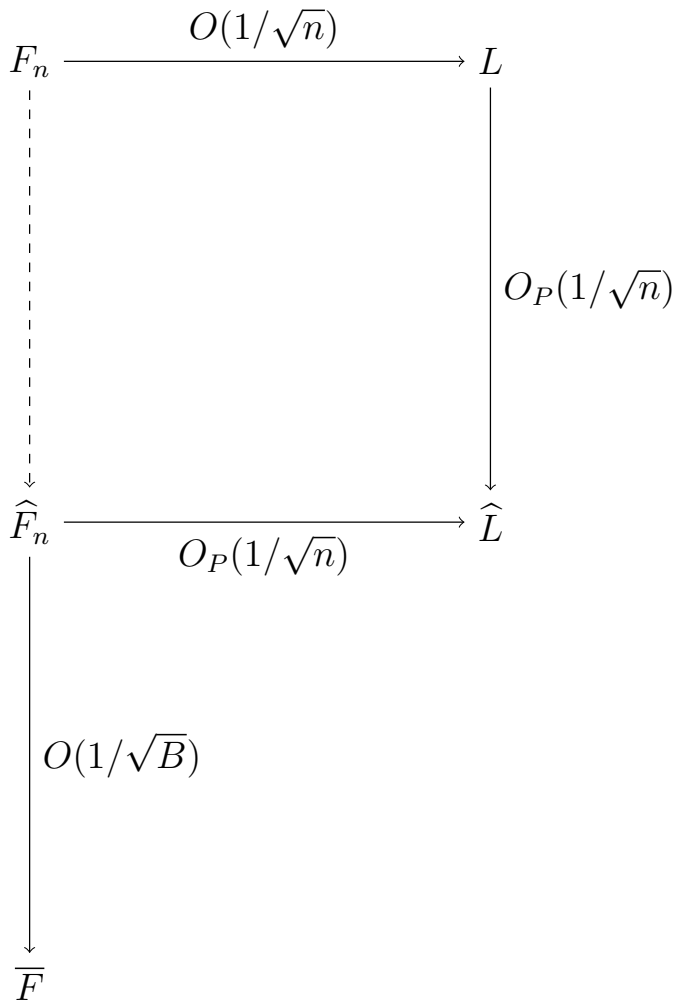


Figure 1: *The distribution* $F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$ *is close to some limit distribution* L*. Similarly, the bootstrap distribution* $\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n)$ *is close to some limit distribution* $\widehat{L}$*. Since* $\widehat{L}$ *and* L *are close, it follows that* F_n *and* $\widehat{F}_n$ *are close. In practice, we approximate* $\widehat{F}_n$ *with its Monte Carlo version* $\overline{F}$ *which we can make as close to* $\widehat{F}_n$ *as we like by taking* B *large.*

7 Failure of the Bootstrap

As usual when we need a counterexample we try the uniform distribution. Suppose that $X_1, \ldots, X_n \sim U[0, \theta]$ and we try to bootstrap the MLE to construct a confidence interval for θ. The mle is $X_{(n)}$. This point is contained in the bootstrap sample with probability

$$1 - (1 - 1/n)^n \approx .63.$$

So the bootstrap distribution puts mass .63 at the single point $X_{(n)}$. But we know that $n(X_{(n)} - \theta)$ has an exponential distribution. So the bootstrap distribution does not resemble the true distribution.

Lecture Notes 21
36-705

1 Causal Inference

Much of statistics and machine learning focuses on questions of association. Are X and Y correlated? Is X predictive of Y, and so on.

In many applications however, our questions are inherently causal: is a medication effective against a disease? Do masks prevent the spread of Covid? Was someone fired because of their age? Does making an ad larger on a website make people buy more?

These are not questions of association. Aspirin is strongly associated with headaches but we don't think that aspirin causes headaches. We often experience turbulence aafter the seat belt sign comes on in a plane. The association is strong. But turning on the seaat belt sign does not cause turbulence. This is what we mean by the phrase: "correlation does not imply causation."

2 The Potential Outcomes Framework

There are two essentially equivalent languages for causation: the first is called potential outcomes or counterfactuals. The second is structural equation models or directed acyclic graphs. We'll start with the first one.

Suppose we have two random variables (A, Y) where A is an exposure or treatment and Y is an outcome. For now, assume that A is binary such as "take aspirin $(A = 1)$" and "don't take aspirin $(A = 0)$." A typical dataset looks like this:

A	1	1	1	1	0	0	0	0
Y	97	76	83	93	100	89	13	67

Now introduce more random variables called *potential outcomes* (or *counterfactuals*). Let $Y(0)$ be the outcome that would have been observed if $A = 0$ and let $Y(1)$ be the outcome that would have been observed if $A = 1$. Causal questions involve comparisons of these two potential outcomes. Note that

$$Y = \begin{cases} Y(0) & \text{if } A = 0 \\ Y(1) & \text{if } A = 1. \end{cases}$$

We can write this as

$$Y = Y(A)$$

or

$$Y = (1 - A)Y(0) + AY(1).$$

So now we have four random variaables $(Y, A, Y(0), Y(1))$ where Y is related to $Y(0)$ and $Y(1)$ by the above consistency relations. Our data set now looks like this:

A	1	1	1	1	0	0	0	0
Y	97	76	83	93	100	89	13	67
Y(0)	?	?	?	?	100	89	13	67
Y(1)	97	76	83	93	?	?	?	?

Much of the data are missing because we don't observe $Y(0)$ when $A = 1$ and we don't observe $Y(1)$ when $A = 0$.

More generally, if $A \in \mathbb{R}$ then the set of counterfactuals is $(Y(a) : \ a \in \mathbb{R})$. In this case there are infinitely many counterfactuals. The observed Y is

$$Y = Y(A).$$

You can think of $Y(a)$ as a curve and we get to observe $Y(a)$ evaluated at A.

While all of this might seem rather obvious, thinking formally about treatment and control, and the potential outcomes is extremely important to causal inference. A point of particular emphasis is that if you are asking a causal question, ideally you need to be able to meaningfully say what the "treatment" is and what the potential outcomes are.

Here are a few examples of statements:

1. "Aspirin cures headaches." In order to cast this is the potential outcomes framework we could imagine that for a person with a headache (a unit) we could either give the person aspirin (treatment) or a placebo (control), and observe the corresponding potential outcome.

2. "She has long hair because she is a girl." This sounds like a causal statement so we should be able to describe the experiment. Is a unit a girl/boy? What exactly is a treatment? Can we meaningfully say what the potential outcomes are?

For some causal questions we can naturally define an associated "experiment". Murky causal questions are ubiquitous, and are in some sense interesting and challenging.

3 Causal Estimands

There are many possible parameters of interest. For example, $\mathbb{E}[Y(a)]$ which is the outcome if everyone had $A = a$. Here is some other notation that is sometimes used:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y|\text{set } A = a] = \mathbb{E}[Y|\text{do } A = a].$$

In general, $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y|A=a]$!

When A is binary, it is often of interest to estimate the *average treatment effect (ATE)*

$$\psi = \mathbb{E}[Y(1)] - \mathbb{E}[Y(1)].$$

Think of this as the mean of Y if everyone took treatment minus the mean of Y if nobody took treatment. In prediction and machine learning one instead focuses on quantities like

$$\alpha = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$$

which is not, in general, the same as ψ. The latter is some measure of association.

How are we going to estimate ψ?

4 Randomized Experiments

Suppose that A was randomly assigned. (Think of the vaccine trials for covid.) In that case, A is independent of $(Y(0), Y(1)$ which we write as

$$A \perp\!\!\!\perp (Y(0), Y(1))$$

then we have

$$\alpha = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] = \mathbb{E}[Y(1)|A=1] - \mathbb{E}[Y(0)|A=0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \psi.$$

Randomization ensures that association IS causaation. And we can estimate α easily. Suppose, for example, that we assigned treatment by flipping a coin. Let

$$\widehat{\alpha} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} \equiv \overline{Y}_1 - \overline{Y}_0$$

where $n_1 = \sum_i I(A_i = 1)$ and $n_0 = \sum_i I(A_i = 0)$. It is easy to see that $\sqrt{n}(\overline{Y}_1 - \overline{Y}_0) \rightsquigarrow N(\theta, \tau^2)$ where $\tau^2 = 2\sigma_1^2 + 2\sigma_2^2$ and $\sigma_j^2 = \text{Var}[Y|A=j]$. Inference is easy. This is why those companies are spending millions of dollars doing randomized trials.

5 Hypothesis testing: Fisher's Exact p-values

Fisher was one of the first to understand the power of a randomized trial. In agricultural experiments, he advocated randomized experiments in order to draw rigorous causal conclusions. A natural subsequent problem is: given an estimate of the causal effect, assess its significance (or construct confidence intervals for it).

Fisher gave a way to construct valid p-values under what is called the *sharp null*, i.e. the null hypothesis that for every unit i the potential outcomes are the same under the treatment and control, i.e. the treatment has no effect. The method is reminiscent of the permutation method we used for two-sample testing.

Suppose we test $H_0 : \theta = 0$ by rejecting when $|\widehat{\alpha}|$ is large. Under the null hypothesis, we can determine both potential outcomes $Y_i(0)$ and $Y_i(1)$ for all the units.

We can now use the permutation method. Say there are n subjects and m were treated. Permute the values of A_i and let T' denote the m units who receive treatment: then our estimate would be:

$$\widehat{\psi}_{T'} = \frac{1}{m}\sum_{i\in T'} Y_i(1) - \frac{1}{n-m}\sum_{i\notin T'} Y_i(0),$$

where we can use the sharp null hypothesis to "fill in" the potential outcomes we do not observe. We can repeat this many times (say B) and compute the p-value:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\widehat{\psi}_{T_b}| \geq |\widehat{\psi}|).$$

It is easy to verify that this is a valid p-value.

6 Confounding

For many policy questions, we cannot actually do a randomized trial. For instance, if I wanted to know if smoking caused lung cancer, there are ethical issues with trying to run a randomized trial. In this case, we have to use *observational* i.e. we have information about many people who are smokers and not, and whether they have lung cancer or not. It is clear that we can measure the correlation between smoking and lung cancer: the main question is when, if ever, can we claim a causal relationship?

Here is a motivating example: Suppose that our population has two kinds of people, those who are always healthy ($Y_i(1) = Y_i(0) = 1$) irrespective of whether they take the treatment or not, and those who are always unhealthy ($Y_i(1) = Y_i(0) = 0$) irrespective of whether they take the treatment or not. Then $Y_i(1) - Y_i(0) = 0$ for all i so there is no causal effect. Suppose further that mostly healthy people take the treatment, while the unhealthy ones do not take the treatment. The causal effect is $\psi = 0$, but the estimator above would yield, $\widehat{\psi} \approx 1$, and we might incorrectly conclude that the treatment is beneficial. The data would look like this:

A	1	1	1	1	0	0	0	0
Y	1	1	1	1	0	0	0	0
Y(0)	1	1	1	1	0	0	0	0
Y(1)	1	1	1	1	0	0	0	0

Suppose however, that we knew who the healthy people were and who the unhealthy people were (we could gather such information by asking people questions about their lifestyle and other things). Then we could try to compare healthy people who took the treatment with healthy people who did not and similarly compare unhealthy people who took the treatment with unhealthy people who did not (and then try to combine these two estimates in some way). In this case, when we compared two healthy people who took the treatment and who did not we would see the treatment had no effect, and similarly for the unhealthy ones. We would correctly conclude that the treatment has no effect.

The key assumption that makes causal inference from observational data possible is the assumption of *no unmeasured confounding* or *selection on observables* or *ignorability*. Formally, we suppose that we have access to covariates X (think demographic information) such that,

$$A \perp\!\!\!\perp (Y(1), Y(0))|X.$$

This is an assumption. Roughly the assumption is plausible in settings where we believe we can measure all of the covariates that explain the decision to take the treatment. We also need the assumption that $\mathbb{P}(A = 1|X = x)$ is bounded away from 0 and 1, so that every individual has some non-zero chance of being either treated or in the control group.

One way to think about this assumption, is that conditional on X we have a randomized trial: the treatment is independent of the potential outcomes. So if we condition on the confounders X we no longer have any selection bias.

In what follows we will assume we have random variables $(X, A, Y, Y(0), Y(1))$ where

$$Y = AY(1) + (1 - A)Y(0) = Y(A).$$

7 Identification under no unmeasured confounding

We want to estimate:

$$\psi = \mathbb{E}[Y(1) - Y(0)]$$

assuming that

$$A \perp\!\!\!\perp (Y(1), Y(0))|X.$$

Now

$$\begin{aligned}\mathbb{E}[Y(1)] &= \int \mathbb{E}[Y(1)|X = x]p(x)dx = \int \mathbb{E}[Y(1)|X = x, A = 1]p(x)dx \\ &= \int \mathbb{E}[Y|X = x, A = 1]p(x)dx = \int \mu_1(x)p(x)dx\end{aligned}$$

where $\mu_a(x) = \mathbb{E}[Y|X = a, A = a]$. Note that thus is NOT equal to $\mathbb{E}[Y|A = 1] = \int \mu_a(x)p(x|1)dx$. Similarly, $\mathbb{E}[Y(0)] = \int \mu_0(x)p(x)dx$. So

$$\psi = \mathbb{E}[Y(1) - Y(0)] = \int [\mu_1(x) - \mu_0(x)]p(x)dx.$$

This is a function of the observed data (X, A, Y) so we can estimate it.

In the case that A is continuous, the same argument shows that

$$\mathbb{E}[Y(a)] = \int \mu_a(x)p(x).$$

8 Estimation under no unmeasured confounding

The most direct way to estimate ψ is to estimate:

$$\mu_0(x) = \mathbb{E}[Y|X = x, W = 0]$$
$$\mu_1(x) = \mathbb{E}[Y|X = x, W = 1].$$

These are two functions of the covariates X, one of them is the average outcome of the treatment group as a function of the covariates, and the other is the average outcome of the control group as a function of the covariates.

Estimating a conditional expectation is a problem is probably the most common problem in statistics – it is known as *regression*. We will delve into this formally in the next few lectures but for now let us suppose that someone hands us estimators $\widehat{\mu}_0$ and $\widehat{\mu}_1$ of these two functions.

Then we can compute the plug-in estimator:

$$\widehat{\psi} = \widehat{\mathbb{E}}_X \left[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right] = \frac{1}{n}\sum_{i=1}^{n} \left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right]$$

which is just the average of the difference between two regression functions. One approximately correct way to think about this is that we are using regression to impute the missing potential outcomes for each individual.

There are other ways to try to estimate ψ. The other popular estimator is called the inverse propensity score estimator. The *propensity score* is

$$\pi(x) = \mathbb{P}(A = 1|X = x),$$

which represents the probability that a unit with covariates x receives treatment. Note that,

$$\mathbb{E}[A|X=x] = \pi(x)$$
$$\mathbb{E}[1-A|X=x] = 1-\pi(x).$$

Let $p(y|x,a)$ denote the density of Y given x and a and recall that $\pi(x) = \mathbb{P}(A=1|X=x)$. So, when $a=1$,

$$p(x,a,y) = p(x)p(a|x)p(y|x,a) = p(x)\pi(x)p(y|x,1)$$

and when $a=0$,

$$p(x,a,y) = p(x)p(a|x)p(y|x,0) = p(x)(1-\pi(x))p(y|x,0).$$

So, for $a=1$,

$$\begin{aligned}
\mathbb{E}[Y(1)] &= \int \mathbb{E}[Y|X=x, A=1]p(x)dx = \int\int yp(y|x,1)p(x)dxdy \\
&= \int\int \frac{y}{\pi(x)} p(y|x,1)\pi(x)p(x)dxdy \\
&= \int\int \frac{y}{\pi(x)} p(x,a=1,y)dxdy \\
&= \int\int \frac{ay}{\pi(x)} p(x,a=1,y)dxdy \\
&= \sum_{a=0}^{1} \int\int \frac{ay}{\pi(x)} p(x,a,y)dxdy \\
&= \mathbb{E}\left[\frac{AY}{\pi(X)}\right].
\end{aligned}$$

Similarly,

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{(1-A)Y}{1-\pi(X)}\right].$$

Therefore,

$$\psi = \mathbb{E}\left[\frac{YA}{\pi(X)}\right] - \mathbb{E}\left[\frac{Y(1-A)}{(1-\pi(X))}\right].$$

This suggests the estimator

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iA_i}{\pi(X_i)} - \frac{Y_i(1-A_i)}{1-\pi(X_i)}\right].$$

This is called the Horvitz-Thompson estimator or the inverse probability weighted (IPW) estimator. This requires that $\pi(x)$ be known as it would be in a randomized experiment. Otherwise we have to insert an estimate of $\pi(x)$. This is again a problem of regression except the outcome is binary.

9 Advanced topics

This is just the tip of the iceberg. If you take a course in Causal Inference you will see many other interesting things such as:

1. No unmeasured confounding is just one assumption that leads to identification of a causal effect. More broadly, in economics, political science and other fields people look for what are called natural experiments, i.e. roughly some subset of the population for which the assignment to treatment/control is nearly random.

2. Even in a randomized trial you might have something called non-compliance, i.e. some people don't do what they are told. In this case, you need to adjust your estimates. This is a canonical example of something called an instrumental variable problem.

3. There are many things beyond the average treatment effect that you might want to estimate. They all have different assumptions under which they are identified (i.e. can be written in terms of observable quantities) and there are different strategies to estimate them.

4. There is a very nice/simple way to combine the regression-based and propensity-score based estimators from above to construct what are called *doubly robust* estimators. These have the property that they are consistent if you can estimate either the regression function or the propensity score well (i.e. you do not need to estimate both well).

5. The plug-in estimator $\widehat{\psi} = n^{-1} \sum_i \int [\widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)]$ is not optimal. Finding optimal estimators of functionals is part of semiparametric theory.

6. There are many different languages for talking about causality and causal inference. We used potential outcomes. Many people use structural equation models and directed graphs. These lead to the same formulas for causal effects. We might revisit this later.

Lecture Notes 22
36-705

Our goal here is to discuss some basic results in *high-dimensional statistics*. The starting point is the Gaussian sequnce model.

1 The Gaussian Sequence Model

Let

$$Y_i = \theta_i + \epsilon_i, \quad i = 1, \ldots, d$$

where $\epsilon_i \sim N(0, \sigma^2/n)$. To understand why we divided the variance by n in the model, you should observe that this corresponds to taking n i.i.d. observations and averaging them. For example, suppose that Y_i is the average of $X_{i1}, \ldots, X_{in} \sim N(\theta_i, \sigma^2)$.

To think about this as a high-dimensional problem, we just assume that $d \to \infty$ as $n \to \infty$, i.e. d is not assumed to be a constant so the number of parameters we want to estimate grows with n.

Recall that the minimax estimator is

$$\widehat{\theta} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_d \end{bmatrix},$$

and its ℓ_2 risk is:

$$R(\widehat{\theta}, \theta) = \mathbb{E}\left[\sum_{i=1}^{d} \epsilon_i^2\right] = \frac{\sigma^2 d}{n},$$

so if $d \gg n$ then we cannot consistently estimate θ. This is a form of the curse of dimensionality but it is much milder than in non-parametric problems. You can see the rate is the "usual parametric rate".

The Gaussian sequence model is an extremely rich model. Due to something known as Le Cam's equivalence, one can reduce many parametric and non-parametric problems (including things like density estimation and non-parametric regression) to sequence model problems, with constraints on the vector θ. Many things we understand about rates of convergence are seen most clearly in this model.

In order to estimate θ in the high-dimensional setting where $d > n$ we need some assumptions on θ. A natural assumption from a practical standpoint is sparsity: we assume that the true underlying θ has many entries which are 0 or nearly zero.

What are natural estimators in this case? A couple of popular ones are based on thresholding:

1. **Hard Thresholding:** Here we use the estimator:

$$\widehat{\theta}_i = Y_i \mathbb{I}(|Y_i| \geq t), \quad \forall\ i \in \{1, \ldots, d\},$$

   where $t > 0$ is some threshold that we need to select.

2. **Soft Thresholding:** One that is closer in spirit to the LASSO (its regression counterpart) is based on soft thresholding, i.e.

$$\widehat{\theta}_i = \text{sign}(Y_i) \max\{|Y_i| - t, 0\}, \quad \forall\ i \in \{1, \ldots, d\},$$

   where $t > 0$ is some threshold that we need to select. Soft thresholding sets any entry to zero if its absolute value is smaller that t (same as hard thresholding) but shrinks other values by t.

There is a different way to motivate these estimators as solutions to (regularized) least-squares problems.

1. **Classical Estimator:** The classical estimator is the solution to the least-squares problem:

$$\widehat{\theta} = \arg\min_a \frac{1}{2}\|Y - a\|_2^2.$$

2. **Hard Thresholding Estimator:** The hard-thresholding estimator is the solution to the problem:

$$\widehat{\theta} = \arg\min_a \frac{1}{2}\|Y - a\|_2^2 + \frac{t^2}{2}\sum_{i=1}^{d} \mathbb{I}(a_i \neq 0).$$

   The penalty here is known as the ℓ_0 penalty, it penalizes solutions that are non-sparse. You should convince yourself that the solution is hard thresholding.

3. **Soft Thresholding Estimator:** The soft-thresholding estimator is the solution to the problem:

$$\widehat{\theta} = \arg\min_a \frac{1}{2}\|Y - a\|_2^2 + t\sum_{i=1}^{d} |a_i|.$$

   Showing that this is equivalent to the soft-thresholding estimator is a little bit more work (and requires some basic sub-gradient calculus) so we'll skip it.

A basic question is then: what is the risk of the hard/soft thresholding estimators? They will turn out to be nearly identical for appropriate choices of the penalty so we will analyze the hard-thresholding estimator here.

Maximum of Gaussians: Before we continue we take another detour to study the maximum of Gaussian RVs. Here is a lemma:

Lemma 1 *Suppose that,* $\epsilon_1, \ldots, \epsilon_d \sim N(0, \sigma^2)$. *Then with probability at least* $1 - \delta$,

$$\max_{i=1} |\epsilon_i| \leq \sigma \sqrt{2 \log(2d/\delta)}.$$

Proof: Recall, our Gaussian tail bound: if $\epsilon \sim N(0, \sigma^2)$:

$$\mathbb{P}(|\epsilon| \geq t) \leq 2 \exp(-t^2/(2\sigma^2)),$$

so by the union bound we obtain that,

$$\mathbb{P}(\max_i |\epsilon_i| \geq t) \leq 2d \exp(-t^2/(2\sigma^2)),$$

which implies the desired lemma. ■

With this lemma we can analyze the hard-thresholding estimator, and obtain the following theorem. (The constants can be improved by a more refined analysis.)

Theorem 2 *Let* $\widehat{\theta}$ *be the hard thresholding estimator with threshold*

$$t = 2\sigma \sqrt{\frac{2 \log(2d/\delta)}{n}}.$$

Then with probability at least $1 - \delta$,

$$\|\widehat{\theta} - \theta\|_2^2 \leq 9 \sum_{i=1}^{d} \min \left\{ \theta_i^2, \frac{t^2}{4} \right\}.$$

Proof: We condition on the event from the previous lemma, i.e. that

$$\max_{i=1} |\epsilon_i| \leq \sigma \sqrt{2 \log(2d/\delta)/n} \leq \frac{t}{2}.$$

Now, observe that,

$$\|\widehat{\theta} - \theta\|_2^2 = \sum_{i=1}^{d} (\widehat{\theta}_i - \theta_i)^2,$$

so we can consider each co-ordinate separately. Let us consider some cases:

1. If for any co-ordinate $|\theta_i| \leq \frac{t}{2}$ our estimate is 0, so our risk for that coordinate is simply θ_i^2.

2. If $|\theta_i| \geq \frac{3t}{2}$ our estimate is $\widehat{\theta}_i = Y_i$ so our risk is simply $\epsilon_i^2 \leq \frac{t^2}{4}$.

3. If $\frac{t}{2} \leq |\theta_i| \leq \frac{3t}{2}$, then our risk,

$$(\widehat{\theta}_i - \theta_i)^2 = (Y_i \mathbb{I}(|y_i| \geq t) - \theta_i)^2 = \theta_i^2 \mathbb{I}(|Y_i| < t) + \epsilon_i^2 \mathbb{I}(|Y_i| \geq t) \leq \max\{\epsilon_i^2, \theta_i^2\} \leq \frac{9t^2}{4}.$$

Putting these together we see that,

$$\|\widehat{\theta} - \theta\|_2^2 \leq 9 \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{t^2}{4}\right\}.$$

■

Note that we have provided a high probability bound. With some more work we can get a similar bound for $\mathbb{E}||\widehat{\theta} - \theta||^2$.

2 Interpreting the bound

We have seen that the risk of the hard-thresholding estimator is upper bounded by,

$$R(\widehat{\theta}, \theta) \lesssim \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{\sigma^2 \log(d)}{n}\right\}.$$

In the worst case, all of the θ_is are non-zero or large, and we obtain that the risk is upper bounded by $\sigma^2 d \log d/n$, which is almost the same as that of the classical estimator (except for the log-factor which you can eliminate by a more careful analysis).

On the other hand if θ is s-sparse, i.e. only s of its entries are non-zero then you observe that the risk looks like:

$$R(\widehat{\theta}, \theta) \lesssim \frac{\sigma^2 s \log(d)}{n},$$

which means that the hard-thresholding estimator is consistent even if $d \gg n$, so long as $s \log(d)/n \to 0$. In fact you can obtain non-trivial estimates even when d is exponentially larger than n. This is quite miraculous: we can avoid the curse of dimensionality in a parametric problem if the target parameter θ is sufficiently structured.

Perhaps one might not expect the vector θ to be exactly sparse but only approximately so, i.e. in some meaningful sense most of its entries are small. There are various ways to

measure sparsity and these will all lead to different, interesting bounds on the risk. Just to get a flavor of this idea, suppose we considered ℓ_1 sparsity, i.e.

$$\sum_{i=1}^{d} |\theta_i| \leq R,$$

for some radius R. Then we can see that, the number of entries of θ larger than R/k is at most k, for any k. Put another way, the numer of i such that $|\theta_i| > C$ is at most R/C.

So for any k, we can use the previous risk bound to obtain:

$$\begin{aligned} R(\widehat{\theta}, \theta) &\lesssim \sum_{i=1}^{d} \min\left\{\theta_i^2, \frac{\sigma^2 \log(d)}{n}\right\} \\ &\lesssim \sum_{i:\theta_i^2 \geq \sigma^2 \log(d)/n} \frac{\sigma^2 \log(d)}{n} + \sum_{i:\theta_i^2 \leq \sigma^2 \log(d)/n} \theta_i^2. \end{aligned}$$

The number of entries of the vector θ that can exceed $\sigma\sqrt{\log(d)/n}$ is at most $\sqrt{n}R/\sigma\sqrt{\log(d)}$, we obtain that bound that,

$$\begin{aligned} R(\widehat{\theta}, \theta) &\lesssim R\sigma\sqrt{\frac{\log(d)}{n}} + \sum_{i:\theta_i^2 \leq \sigma^2 \log(d)/n} \theta_i^2 \\ &\lesssim R\sigma\sqrt{\frac{\log(d)}{n}} + \sigma\sqrt{\frac{\log(d)}{n}} \sum_{i:\theta_i^2 \leq \sigma^2 \log(d)/n} |\theta_i| \\ &\lesssim 2R\sigma\sqrt{\frac{\log(d)}{n}}. \end{aligned}$$

Notice that the rate of convergence is different from the s-sparse case, roughly behaving as $1/\sqrt{n}$ instead of $1/n$. Ignoring this distinction however, the result should again surprise you – we are not even assuming that the unknown vector θ is sparse, just that is has ℓ_1-norm that is controlled, and once again we can obtain consistent estimators when $d \gg n$. More generally, there are many ways in which we can measure sparsity or impose structure on the unknown parameter, and depending on the structural assumption we might obtain improved rates of convergence.

We will see that similar things happen in high-dimensional regression (under appropriate assumptions), and are well-understood now to happen in many other interesting models. This is the area of high-dimensional statistics: the main features are we do not assume the dimension of the model is fixed as $n \to \infty$ but need structural assumptions of various kinds (typically variants of sparsity) to obtain fast rates of convergence.

Lecture Notes 23
36-705

We begin with a quick review of low dimensional linear regression, before turning our attention to high-dimensional regression with the LASSO.

1 Low Dimensional Linear Regression – Review

Linear regression is a tool to approximate the conditional expectation

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

with a linear function of X. We observe iid pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and we use the model

$$Y_i = \beta^T X_i + \epsilon_i,$$

where $Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d$ and $\epsilon_i \sim N(0, \sigma^2)$. Usually the first coordinate of X_i is 1 for all i which means that β_1 is the intercept. The design matrix is the $n \times d$ matrix $\mathbb{X}$ where $\mathbb{X}_{ij}$ is the j^{th} coordinate of X_i. We let

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T = \frac{1}{n}\mathbb{X}^T\mathbb{X}.$$

Least Squares: The least squares estimator is

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2$$

and is given by

$$\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\alpha} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

where $\widehat{\alpha} = \frac{1}{n}\sum_i X_i Y_i$ and $\mathbb{Y} = (Y_1, \ldots, Y_n)$.

Exercise. Show that the least squares estimator is also the (conditional) maximum likelihood estimator, that is, $\widehat{\beta}$ maximizes $\prod_i p(Y_i|X_i;\beta)$.

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Many of the calculations are done conditional on $\mathcal{X}$. This simplifies the calculations but it leads to valid inferences. For example, suppose that C is a confidence set for β conditional on $\mathcal{X}$. So $P(\beta \in C|\mathcal{X}) = 1 - \alpha$. Then

$$P(\beta \in C) = \mathbb{E}[P(\beta \in C|\mathcal{X})] = 1 - \alpha.$$

We can write

$$\mathbb{Y} = \mathbb{X}\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. So, conditional on $\mathcal{X}$,

$$\begin{aligned}\widehat{\beta} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\beta + \epsilon) \\ &= \beta + (\mathbb{X}^T\mathbb{X})^{-1}\epsilon \stackrel{d}{=} N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).\end{aligned}$$

This implies that, conditional on $\mathcal{X}$, $\widehat{\beta}_j \sim N(\beta_j, \tau_j^2)$ where τ_j^2 is the j^{th} diagonal element of $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$. A consistent estimaator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-d}\sum_i \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon}_i = Y_i - \widehat{\beta}^T X_i$. Hence, an asymptotic $1-\alpha$ confidence interval is $\widehat{\beta}_j \pm z_{\alpha/2}\widehat{\tau}$. The fitted values are

$$\widehat{\mathbb{Y}} = \mathbb{X}\widehat{\beta} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$$

where $\widehat{Y}_i = \widehat{\beta}^T X_i$. Note that

$$\widehat{\mathbb{Y}} = HY$$

where

$$H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$$

is called the hat matrix. This means that $\widehat{\mathbb{Y}}$ is the projection of $\mathbb{Y}$ onto the column space of $\mathbb{X}$.

There are several other quantities of interest in linear regression:

1. The in-sample prediction error:

$$\frac{1}{n}\sum_i (\widehat{Y}_i - Y_i)^2 = \frac{1}{n}||\widehat{\mathbb{Y}} - \mathbb{Y}||^2.$$

2. The out-of-sample prediction error:

$$\mathbb{E}[(\widehat{Y} - Y)^2]$$

   where (X, Y) is a new point and the expectation is over both the randomness in $\widehat{\beta}$ and in the new sample.

3. The ℓ_2 error $\mathbb{E}[\|\widehat{\beta} - \beta\|_2^2]$.

4. The support recovery error (makes most sense when β^* is sparse):

$$\mathbb{P}(\text{supp}(\widehat{\beta}) \neq \text{supp}(\beta^*)).$$

Let us review these quantities for low-dimensional regression.

In-sample prediction error. Note that

$$\frac{1}{n}||\widehat{\mathbb{Y}} - \mathbb{Y}||^2 = \frac{1}{n}||\mathbb{X}\widehat{\beta} - \mathbb{X}\beta - \epsilon||^2 = \frac{1}{n}||\mathbb{X}\widehat{\beta} - \mathbb{X}\beta||^2 + \frac{1}{n}||\epsilon||^2 - 2\frac{1}{n}\langle \epsilon, \mathbb{X}(\widehat{\beta} - \beta)\rangle.$$

The second term is the unavoidable error: our estimate $\widehat{\beta}$ has no effect on it. The last term has mean 0 and concentrates to 0 quickly. So people focus on the first term. Recall that, conditional on $\mathcal{X}$,

$$\widehat{\beta}\ N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$$

so that

$$\mathbb{X}\widehat{\beta}\ N(\mathbb{X}\beta, \sigma^2 H).$$

that is

$$\Delta \equiv \mathbb{X}\widehat{\beta} - \mathbb{X}\beta \sim N(0, \sigma^2 H).$$

Now we use the following fact: if $W \sim N(\mu, \Sigma)$ then $\mathbb{E}[W^T A W] = \text{trace}(A\Sigma) + \mu^T A \mu$ where tr denotes the trace (sum of the diagonal elements). Hence,

$$\mathbb{E}[\Delta^T \Delta] = \sigma^2 \text{tr}(H) = \sigma^2 \text{tr}(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X})$$
$$= \sigma^2 \text{tr}(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)$$
$$\sigma^2 \text{tr}(\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}) = \sigma^2 \text{tr}(I) = \sigma^2 d.$$

Finally,

$$\mathbb{E}\left[\frac{\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2}{n} \,\middle|\, \mathcal{X}\right] = \frac{\sigma^2 d}{n}.$$

It follows that

$$\mathbb{E}\left[\frac{\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2}{n}\right] = \frac{\sigma^2 d}{n}.$$

ℓ_2 **error**. Again, under our assumptions we know that, conditional on $\mathcal{X}$,

$$\widehat{\beta} \sim N(\beta^*, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}).$$

Thus

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2 \ \Big|\ \mathcal{X}\Big] = \sigma^2 \text{tr}(\mathbb{X}^T\mathbb{X})^{-1})$$

and so

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2\ \Big] = \sigma^2 \mathbb{E}[\text{tr}(\mathbb{X}^T\mathbb{X})^{-1}] = \frac{\sigma^2}{n}\mathbb{E}[(\widehat{\Sigma})^{-1}]$$

where we recall that $\widehat{\Sigma} = n^{-1}\mathbb{X}^T\mathbb{X}$. Suppose that $\widehat{\Sigma}$ has eigenvalues bounded from below by $c > 0$. Then $\widehat{\Sigma}$ has eigenvalues bounded from above by $1/c$. Recall that the trace is equal to the sum of the eigenvalues. Therefore,

$$\mathbb{E}\Big[||\widehat{\beta} - \beta^*||^2\Big] \leq \frac{\sigma^2 d}{cn}.$$

2 High-dimensional Regression

In high-dimensional regression, we are interested in the setting where the covariate distribution has dimension $d \gg n$. The first thing to observe is that even if our old analysis worked (it does not) the prediction error and ℓ_2 error both scale as $\sigma^2 d/n$ which does not go to 0 as we increase the sample-size, which would mean that our methods are inconsistent. From a minimax perspective, it turns out that this is unavoidable, i.e. it is impossible to consistently estimate the regression vector β, when $d \gg n$, and we need to make structural assumptions to make progress.

Also, the least-squares estimator is no longer well-defined. To see this, observe that the assumption that $\widehat{\Sigma}$ is invertible (which is completely benign in low-dimensions) can never hold in high-dimensions. In particular the matrix,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T,$$

has rank at most n (it is a sum of rank 1 matrices) and is a $(d \times d)$ matrix, so is clearly not invertible if $d > n$. The way to picture this is that in high-dimensions there will be many vectors β such that, $Y = \mathbb{X}\beta$ which have least squares error of 0 (i.e. exactly pass through all the samples).

This is a form of over-fitting, and one way to avoid this is to use regularization. This is roughly equivalent to imposing some type of structure on the unknown β and then attempting to recover β by leveraging this structure. We will focus on versions of sparsity, i.e. settings where β is either exactly sparse (i.e. has s non-zero entries) or is approximately sparse (i.e. has bounded ℓ_1 norm).

Analogous to the Gaussian sequence model there are two estimators that one might consider:

1. **Hard-Thresholding type estimator:** The analog of hard thresholding is:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \frac{t^2}{2}\sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0).$$

   This is usually called best-subset regression. The best way to think about the nomenclature is to consider a closely related estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} \mathbb{I}(\beta_i \neq 0) \leq k,$$

   where now we have a different tuning parameter $k > 0$ (instead of t). You should be able to (with some effort) convince yourself of the fact that these two programs are

exactly equivalent, i.e. if you fix any $t > 0$ and solve the first program, then there is some k for which you obtain exactly the same solution. The first form is sometimes called the penalized-form and the second is called the constrained-form.

The natural way to implement the second estimator would be to enumerate all subsets of size k, fit a regression on this subset and then pick the subset, and estimate β that has lowest mean -squared error. Hence the name, "best-subset regression". But this is computationally infeasible.

2. **Soft-Thresholding type estimator:** The analog of soft thresholding is known as the LASSO, i.e. the Least Absolute Selection and Shrinkage Operator,

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + t\sum_{i=1}^{d} |\beta_i|.$$

Analogous to the above, one can consider a closely related estimator:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2,$$
$$\text{subject to } \sum_{i=1}^{d} |\beta_i| \leq k,$$

again there is an equivalence, i.e. every value of t corresponds to some value of k. This program is a convex program, and simple methods (roughly, gradient descent with tweaks) can be used to solve it quite fast. There is typically no closed-form solution but that is not a huge problem.

This brings us to an important distinction between the Gaussian sequence model and regression. In the Gaussian sequence model (no X) both of these programs had simple closed-form solutions, whereas now this is no longer the case. More importantly, best-subset is computationally intractable but the LASSO is not.

With this motivation in place, let us study the prediction error of the LASSO. We begin with some assumptions, for simplicity we will study the constrained form of the LASSO, and further we will just assume that the tuning parameter k is chosen to be exactly $\|\beta^*\|_1$ where β^* is the true value of β. In practice, one might choose this tuning parameter by cross-validation or some other method.

To simplify our calculations we will also assume the design matrix X is column-normalized, i.e. for each column j of the matrix:

$$\sum_{i=1}^{n} \mathbb{X}_{ij}^2 = n.$$

You can ensure this by re-normalizing every column of X. This does change β^* (and its ℓ_1 norm).

Theorem 1 *Suppose we consider the constrained-LASSO with $k = \|\beta^*\|_1$ where β^* denotes the true value. Then, with probability at least $1 - \delta$:*

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq 4\sigma\|\beta^*\|_1\sqrt{\frac{2\log(2d/\delta)}{n}}.$$

This bound is exactly analogous to the bound on the error of the hard/soft-thresholding estimator in the Gaussian sequence model when we assumed that the ℓ_1 norm of the mean vector θ^* was bounded. Notice again, that the prediction error goes to 0 with n, even in settings where $d \gg n$.

This result is due to Greenshtein and Ritov and really kicked off the wave of high-dimensional statistics. It showed that high-dimensional prediction was possible (at least in the linear model). Several later works showed that under stronger assumptions, one could achieve small ℓ_2 error and even exactly identify the non-zero components of β^* (i.e. do feature selection) in the high-dimensional setting. Furthermore, most of these phenomena generalize to general parametric models (for instance, high-dimensional logistic regression, high-dimensional graphical model estimation and so on).

Proof: To prove this we note that, since we selected the tuning parameter to be equal to $\|\beta^*\|_1$, the vector β^* is feasible for the program and $\widehat{\beta}$ is optimal, so we have the so-called basic inequality:

$$\frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\widehat{\beta}\|_2^2 \leq \frac{1}{2n}\|\mathbb{Y} - \mathbb{X}\beta^*\|_2^2,$$

where we divided both sides by n for convenience. Re-arranging this inequality we obtain that,

$$\frac{1}{2n}\|\mathbb{E}X\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq \frac{1}{n}\langle\epsilon,\, \mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\rangle = \langle\frac{\mathbb{X}^T\epsilon}{n},\, \widehat{\beta} - \beta^*\rangle,$$

where ϵ is the noise in the linear model. Holder's inequality tells us that for any two vectors $a, b \in \mathbb{R}^d$,

$$\langle a,\, b\rangle \leq \left(\max_{i=1} a_i\right)\left(\sum_{i=1}^{d}|b_i|\right).$$

Applying this inequality we obtain,

$$\frac{1}{n}\|\mathbb{X}\widehat{\beta} - \mathbb{X}\beta^*\|_2^2 \leq 2\|\widehat{\beta} - \beta^*\|_1 \max_{i=1}\frac{X_i^T\epsilon}{n}$$

where X_i denotes the i-th column of the design matrix. Now, by the triangle inequality, $\|\widehat{\beta} - \beta^*\|_1 \leq 2\|\beta^*\|_1$ (recall that we constrained our optimal solution to have ℓ_1 norm at most $\|\beta^*\|_1$), so it only remains to bound $\max_{i=1}^d \frac{X_i^T\epsilon}{n}$.

Each entry here, conditional on $\mathcal{X}$, has a Gaussian distribution with mean 0 and variance $\sigma^2 \|X_i\|_2^2 / n^2 \leq \sigma^2 / n$, using our column normalization assumption. So with probability at least $1 - \delta$, we have that,

$$\max_{i=1} \frac{X_i^T \epsilon}{n} \leq \sigma \sqrt{\frac{2 \log(2d/\delta)}{n}},$$

and combining these facts we obtain the desired bound.

Lecture Notes 24
36-705

Today we will discuss Bayesian Inference.

1 Bayesian Inference

We have already talked about the mechanics of Bayesian inference when we discussed constructing point estimates by treating the parameters as random with a prior, and the computing and summarizing the posterior. What we really did not talk about yet, and will spend a small amount of time talking about now is the philosophy of Bayesian inference.

The philosophical distinction between Bayes and frequentists is deep. We have so far followed the frequentist framework, where, to us a probability is representing some type of long run frequency, i.e. when we say the probability that our estimator is close to some unknown "true" parameter with probability at least $1-\delta$ we are really imagining repeating this (or some other) experiment many many times and then our guarantees will be correct for at least $1-\delta$ of these experiments. Similarly, with confidence intervals, we imagine many people across the world construct confidence intervals and our guarantee is that 95% of those intervals would trap the true parameter, i.e. **the goal of frequentist inference is to create procedures with long run guarantees.**

Moreover, the guarantees should be uniform over θ if possible. For example, a confidence interval traps the true value of θ with probability $1-\alpha$, no matter what the true value of θ is. **In frequentist inference, procedures are random while parameters are fixed, unknown quantities.**

In the *Bayesian approach*, probability is regarded as a measure of **subjective degree of belief**. One can view the Bayesian approach as a way to manipulate beliefs. Beliefs are then assumed to follow the rules of normal probabilities by a notion called *coherence*. In this framework, everything, including parameters, is regarded as random. These procedures do not have to satisfy frequency guarantees.

A summary of the main ideas is in Table 1.

2 The mechanics of Bayesian Inference

Roughly, the setup in Bayesian Inference is exactly the same as in frequentist inference: we begin by specifying a statistical model, i.e. a collection of distributions $\{P_\theta : \theta \in \Theta\}$.

The main distinction is that we now treat the parameter θ as random, and encode our "prior beliefs" about the value of the parameter in a distribution π.

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
θ	random variable	fixed
X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

Table 1: Bayesian versus Frequentist Inference

We assume that the observed data, is from the *conditional distribution*, conditional on some realization of the random parameter, i.e. the setup is:

$$\begin{aligned} \theta &\sim \pi \\ \{X_1, \ldots, X_n\}|\theta &\sim P_\theta. \end{aligned}$$

We do not observe θ but can compute our "posterior belief" using Bayes' rule, i.e.:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)}{\int_\theta \mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)} \propto \mathcal{L}(\theta)\pi(\theta)$$

i.e. while the frequentist treats the likelihood as just a function of θ, the Bayesian (weights and) normalizes the likelihood and interprets it as a distribution over Θ.

We have seen examples of this whole thing before, except rather than treat the posterior as an object of interest, we used it to obtain point estimates and we focused on the frequentist properties of the resulting point estimate.

3 The goals of Bayesian inference

In frequentist inference the goal was: create procedures that have good frequency properties.

In Bayesian inference the goal is to write down a prior that captures your prior belief and compute the posterior; then you are essentially done.

4 Bayesian confidence sets and Frequentist guarantees

Once we have a posterior distribution, we can construct what are called credible sets: they are the Bayesian analogue of confidence sets but are quite different.

A $1-\alpha$ credible set/interval is simply any set C_α to which the posterior assigns $1-\alpha$ mass, i.e.

$$\int_{C_\alpha} \pi(\theta|X_1,\ldots,X_n)d\theta = 1-\alpha.$$

Once again notice that the thing that is random is θ, the data is conditioned on (i.e. fixed). The set C_α is fixed (i.e. not random) here, unlike in a frequentist confidence interval. These intervals do not typically have frequency guarantees, and we will see examples of this.

If one is interested in the frequency properties of Bayesian inference then one might also be interested in some notion of frequentist consistency and rates of convergence. The typical way to formulate frequentist consistency is via something called *posterior contraction*, i.e. in the frequentist setup (where θ^* is fixed, unknown) we want that our posterior concentrates around the true value of the parameter (consistency) and does so quickly (rates of convergence).

Formally, let $B_\epsilon = \{\theta : \ ||\theta-\theta^*|| \leq \epsilon\}$ be a ball around the true value θ^* and let

$$Q = \int_{B_\epsilon} p(\theta|X_1,\ldots,X_n)d\theta$$

be the posterior probability of B_ϵ. Now we think of $Q \equiv Q(X_1,\ldots,X_n)$ as a statistic in the frequentist sense: the randomness is in $X_1,\ldots,X_n$. The posterior is consistent if $Q \xrightarrow{P} 1$ for every $\epsilon > 0$ and every θ^*. The convergence is with respect to $p(x_1,\ldots,x_n|\theta^*)$. We say that the rate of convergence is ϵ_n if the above is true when we let $\epsilon = \epsilon_n$ for some $\epsilon_n \to 0$.

In finite dimensional parametric models, if the prior puts positive mass around each parameter value, the posterior will typically be consistent and the rate of convergence will be $O(n^{-1/2})$.

5 Bernstein-von Mises theorem

The Bernstein-von Mises theorem guarantees us that in fixed-dimensional problems, under the assumption that the prior is continuous, and (strictly) positive in a neighborhood around θ^*, the posterior is close to a Gaussian, i.e.

$$\left\| \pi(\theta|X_1,\ldots,X_n) - N\left(\widehat{\theta}_n, \frac{I(\widehat{\theta}_n)^{-1}}{n}\right) \right\|_{\text{TV}} \to 0,$$

where $\widehat{\theta}_n$ is the MLE and the distance between the two distributions is the total-variation distance, i.e. for two distributions with densities p, q:

$$\|p-q\|_{\text{TV}} = \frac{1}{2}\int |p(x)-q(x)|dx.$$

We might discuss this further at some point but for now you should take away that the posterior is very close to a Gaussian centered at the MLE with rapidly shrinking variance.

One immediate consequence is that credible intervals will be roughly identical to the usual Wald interval (based on the MLE) as $n \to \infty$. The key take away: in fixed dimension, large sample-size problems, under some conditions Bayesian procedures will behave like standard frequentist inference.

The condition "(strictly) positive in a neighborhood around θ^*" is extremely strong in high-dimensions or in a non-parametric problem. In general, when the parameter space is large you should be suspicious of this assumption.

6 Where do priors come from?

The purist Bayesian view is that the prior should truly be an encoding of your prior beliefs. Often we choose priors by convenience (recall our examples from the minimax lecture). Some might argue that in many cases the priors do not matter but this is only rigorously true in low-dimensional, parametric problems.

Some might also choose priors based on the data; this is known as *empirical Bayes*. This is often a good idea but some would consider it to not strictly adhere to the Bayesian philosophy.

Some have argued for what are called non-informative priors, i.e. priors that somehow capture complete ignorance about the parameter. The natural first attempt would be to say that we take $\pi(\theta) \propto 1$, however this has some drawbacks. If we have no information about the parameter θ then presumably we should also have no information about some transformation about the parameter, i.e. say θ^2. However, if you transform the flat prior to a prior on θ^2 it will not be flat. A prior that is in fact invariant under transformations is called Jeffreys prior where we choose $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information for the model under consideration. You will verify this in your HW.

7 Priors = Regularizers?

A slightly different viewpoint that is often articulated is that one can view priors as regularizers.

Most regularized frequentist estimators, for instance estimating Bernoulli probabilities with "Laplace smoothing" (i.e. adding psuedo-counts) is just the posterior mean with a Beta prior. The LASSO regression estimator or Ridge regression estimator are just the posterior mode with either a Laplace prior or a Gaussian prior (respectively). Relatedly, one can derive model complexity regularizers with appropriate model complexity dependent priors.

The argument is that "many sensible frequentists procedures (ones with strong guarantees) are just posterior summaries with particular priors."

This argument should be taken with a grain of salt: lets focus on the LASSO, which is the posterior mode with a Laplace prior. As we have discussed previously the LASSO has some very desirable properties (high-dimensional prediction consistency) and so one might wonder does the (full) posterior have nice properties in a high-dimensional setting?

The answer turns out to be no: *once you leave the realm of the Bernstein-von Mises theorem (fixed d, growing n) things can break down.* In particular, in the high-dimensional regression problem the posterior itself does not meaningfully concentrate and sampling from the posterior will lead to completely meaningless inference (from a frequentist point of view). Essentially only the posterior mode has nice properties, the rest of the posterior is useless.

8 Failure of credible intervals

We have said many times now that things can break down in high-dimensions. This is particularly alarming if we treat credible intervals as confidence intervals. They are not valid confidence intervals. Let us verify this in a simple example.

Suppose we are in the Gaussian sequence model, i.e. we observe,

$$Y_i = \theta_i + \epsilon_i, \quad i \in \{1, \dots, d\},$$

and $\epsilon_i \sim N(0, \sigma^2/n)$.

We choose a flat prior (although this is not really crucial it makes the calculations simpler), i.e. $\pi(\theta_1, \dots, \theta_d) \propto 1$. This is an example of something called an *improper prior*, i.e. it is not really a valid distribution. We can still use the usual mechanics to obtain a valid posterior.

Our goal is to construct a confidence interval for the parameter $\mu = \sum_{i=1}^d \theta_i^2$. Since the prior is flat the posterior is easy to compute and in particular, the posterior factorizes over the parameters (since the prior is flat and the likelihood factorizes) and we have:

$$\pi(\theta_i | Y_1, \dots, Y_d) \stackrel{d}{=} N(Y_i, \sigma^2/n).$$

The posterior for $\mu | Y_1, \dots, Y_d$ is σ^2/n times a non-central χ^2 distribution, with d degrees of freedom and non-centrality parameter $\lambda = (n/\sigma^2) \sum_{i=1}^d Y_i^2$. The mean of the posterior for μ is $\sum_{i=1}^d (Y_i^2 + \sigma^2/n)$, and the variance of the posterior for μ is $4\sigma^2(\sum_{i=1}^d Y_i^2)/n + 2\sigma^4 d/n^2$.

If we were to examine frequentist properties, we would fix a θ, and then mean of the posterior mean is

$$\mathbb{E}\left[\sum_{i=1}^d (Y_i^2 + \sigma^2/n)\right] = \mu + 2\sigma^2 d/n$$

and the standard deviation of the posterior is on the order of $\sigma\sqrt{\mu/n} + \sigma^2\sqrt{d}/n$. So the posterior is centered at the wrong point, and its spread is quite small. Using these two facts along with Chebyshev's inequality, you can see that a posterior credible interval will have coverage that $\to 0$ as $d \to \infty$.

Lecture Notes 25
36-705

Today we will discuss the problem of *model selection*. Let's start with some examples.

Example 1: A convenient, flexible parametric family is the mixture of Gaussians:

$$p_\theta(x) = \sum_{i=1}^{k} \pi_i N(\mu_i, \Sigma_i)$$

where $\sum_i \pi_i = 1$. The parameters are $\theta = (\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k)$. We also need to choose the number of mixture components k and this is a model selection problem where we have a sequence of models $\mathcal{M}_1, \ldots, \mathcal{M}_k$ indexed by the number of components.

Example 2: Polynomial order in regression. Suppose you use a polynomial to model the regression function:

$$m(x) = \mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p.$$

You will need to choose the order of polynomial p. We can think of this as a sequence of models $\mathcal{M}_1, \ldots, \mathcal{M}_p, \ldots$ indexed by p.

Example 3: AR model. We have always assumes that the data are iid. An example of non iid data is a time series where we expect the data to be correlated over time. Consider a time series $Y_1, Y_2, \ldots$. A common model is the AR (autoregressive model):

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_k Y_{t-k} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2)$. The number k is called the order of the model. We need to choose k.

Example 4: In survival analysis we want to model lifetime Y which is a non-negative random variable. Two common models are the Weibull distribution $p_\theta(x) = (k/\lambda)(x/\lambda)^{k-1} e^{-(x/\lambda)^k}$ and the lognormal, where $\log X \sim N(\mu, \sigma^2)$. This defines two different models $\mathcal{M}_1$ and $\mathcal{M}_2$.

Notice that models are often nested with increasing complexity. Choosing a large k can lead to overfitting. Just picking the model with the highest likelihood will lead to always picking the biggest model. There are two different goals in model selection:

1. Find the model that gives the best prediction (without assuming that any of the models are correct). This is equivalent to: find the model whose estimated distribution is closest to the true distribution.

2. Assume one of the models is the true model (the smallest model containing the true density) and find the true model.

We'll talk about the following methods:

Cross-validation
AIC
BIC
Bayes.

Perhaps the only slightly counter-intuitive fact you need to remember is that when there is a true model methods like cross-validation can fail to find it.

The basic take-aways are:

1. If your goal is prediction, you have a reasonable sample-size and you have a reasonable computation budget use cross-validation.

2. If your goal is prediction, but you either have too small a sample or you have a very low computational budget, you should consider using AIC.

3. If your goal is selecting the true model you should use BIC.

1 The Methods

CV: Cross-validation has different versions but the procedure is probably roughly familiar to you. We train our models on a subset of the data and then evaluate and choose between the models on the rest of the data. We then potentially re-shuffle the data, repeat, and combine the results in some way. We will simplify this and just suppose throughout this lecture that we do a train-test split.

AIC: AIC (Akaike Information Criterion) is a model selection rule that does not use any sample-splitting. Informally, it is best understood as an asymptotic approximation to CV. Formally, Stone showed in a classic paper that under assumptions AIC and CV are asymptotically equivalent when using the MLE for each model.

BIC. BIC (Bayesian Information Criterion) can be thought of as an asymptotic approximation of a Bayesian approach.

2 Cross Validation

Denote the sample size by $2n$. Split the data into two groups $\mathcal{D}_1$ and $\mathcal{D}_2$. (In practice we often use many groups.) Using $\mathcal{D}_1$ we find the mle $\widehat{\theta}_j$ for model $\mathcal{M}_j$. Let $p_j(x) = p(x; \widehat{\theta}_j)$ be te estimated density from model $\mathcal{M}_j$. The idea is to choose j which minimizes $K(p, p_j)$

where p is the true density and

$$K(p, p_j) = \int p(x) \log \left(\frac{p(x)}{p_j(x)} \right) dx$$

is the Kullback-Leibler distance. Notice that we do not necessarily assume that the true density p is in any of the models.

Notice that minimizing $K(p, p_j)$ is the same as maximizing

$$R_j = \int p(x) \log p_j(x) dx.$$

We can estimate R_j from the second sample $\mathcal{D}_2$ by

$$\widehat{R}_j = \frac{1}{n} \sum_{i \in \mathcal{D}_2} \log p(X_i; \widehat{\theta}_j).$$

The score $\widehat{R}_j$ can also be thought of as a measure of how well we can predict future observations. If we did not split the data, $\widehat{R}_j$ would be biased: larger models will always lead to a larger score.

We can now use the LLN to argue that if the test-set size goes to ∞ then our risk estimates converge to their expectations, and then we will find the model/estimate with the lowest KL to the true model. Let's make this more precise. Assume that $|\log p_\theta(X)| \leq B$ for every θ and X that we care about (this can be relaxed using more complex techniques). From Hoeffding's inequality and the union bound:

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2M \exp(-2n\epsilon^2/(4B^2)).$$

Let

$$\epsilon_n = \sqrt{\frac{4B^2 \log(2M/\alpha)}{n}}.$$

Then

$$\mathbb{P}(\max_i |R_i - \mathbb{E}(R_i)| \geq \epsilon_n) \leq \alpha.$$

Let $\widehat{i} = \arg\min_i R_i$ be the selected model and let $i^* = \arg\min_i \mathbb{E}(R_i)$ be the best model. Then, with probability at least $1 - \alpha$:

$$\mathbb{E}(R_{\widehat{i}}) \leq R_{\widehat{i}} + \epsilon_n \leq R_{i^*} + \epsilon_n \leq \mathbb{E}(R_{i^*}) + 2\epsilon_n.$$

So the model we select will be sub-optimal by at most $2\epsilon_n$. In regression, we would use exactly the same reasoning, but just replace the risk with the squared loss. Reasoning about

K-fold cross-validation turns out to be much more challenging, because the data re-use breaks independence assumptions.

The analysis above should remind you of the analysis we did before of Empirical Risk Minimization. The goals are slightly different, as is the final guarantee. It is worth thinking about what exactly the data splitting buys you. In particular, we do not require uniform convergence of the empirical to the true risk over all the model classes $\mathcal{M}_1, \ldots, \mathcal{M}_M$, rather we only require a good estimate of the risk for the *fixed* models indexed by $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$.

3 AIC

Suppose we don't want to use data-splitting because of lack of data or lack of computational resources. We could try to estimate R_j by

$$\widehat{R}_j = \frac{1}{n}\sum_i \log p(X_i; \widehat{\theta}_j) = (1/n)\ell_j(\widehat{\theta}_j)$$

but, as we discussed above, this is very biased. And the more parameters there are, ther more biased this will be. Akaike proved that the bias is approximately equal to d_j/n where d_j is the dimension of the model $\mathcal{M}_j$. So a bias-adjusted estimator is

$$\widehat{R}_j = (1/n)[\ell_j(\widehat{\theta}_j) - d_j].$$

For a variety of historical reasons, we will multiple by $2n$ which does not affect which model is the maximizer. This leads to the criterion

$$\mathrm{AIC}_j = 2\ell_j(\widehat{\theta}_j) - 2d_j.$$

We choose j to maximize AIC_j. It is best to think of this as an approximation to CV.

4 BIC

. The BIC crierion is

$$\mathrm{BIC}_j = \ell_j(\widehat{\theta}_j) - \frac{d_j}{n}\log n.$$

This is basically AIC but with a harsher penalty. This approach was proposed by Gideon Schwartz based on the following Bayesian argument. We place prior probabilities $\nu_1, \ldots, \nu_k$ for each model. Then we put priors $p_j(\theta_j)$ for the parameters in model $\mathcal{M}_j$. By Bayes' theorem

$$p(\mathcal{M}_j|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\mathcal{M}_j)\nu_j}{\sum_s p(X_1, \ldots, X_n|\mathcal{M}_s)\nu_s} = \frac{\nu_j \int \mathcal{L}_j(\theta_j)d\theta_j}{\sum_s \nu_s \int \mathcal{L}_s(\theta_s)d\theta_s}.$$

Schwartz showed that

$$\log p(\mathcal{M}_j|X_1,\ldots,X_n) \approx \text{BIC}_j.$$

Suppose that the true density is in one of the models. Let $\mathcal{M}_j$ be the smallest model that contains p. Then it can be shown that $\widehat{j} \xrightarrow{P} j$ where $\widehat{j}$ is the model chosen by BIC. We'll return to this point later.

5 Hypothesis Testing

In some cases we can frame model selection as a hypothesis testing problem. For example, suppose that $X_1,\ldots,X_n \sim N(\theta,1)$ and that are two models are

$$\mathcal{M}_0 = \{\theta : \theta = 0\},$$
$$\mathcal{M}_1 = \{\theta : \theta \in \mathbb{R}\}.$$

We can select a model by testing $H_0 : \theta = 0$ versus $H_0 : \theta = 1$. Notice that, if $\theta = 0$, we will choose the wrong model with probability α. But that's ok since CV and AIC also can choose the wrong model. But notice that hypothesis testing doesn't provide any guarantee about the selected model in terms of KL distance, for example.

6 Choosing the True Model?

Suppose that the true density is in one of the models. Let $\mathcal{M}_j$ be the smallest model that contains p. We have seen that, asymptotically, BIC chooses this model. Noe we will show that CV and AIC don't do this. But thus is not a problem since that's not their goal. We'll focus on the example above where the data re Normal and we are choosing between $\mathcal{M}_0$ and $\mathcal{M}_1$.

Let us denote the mean of the training samples as $\widehat{\mu}_{\text{tr}}$ and the mean of the testing samples as $\widehat{\mu}_{\text{te}}$. Then

$$\widehat{\theta}_0 = 0, \qquad \widehat{\theta}_1 = \widehat{\mu}_{\text{tr}}.$$

Focus on the case where $\theta = 0$. The difference between the cross-validation loss for the two models is (proportional to):

$$\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_{\text{tr}})^2 = -\widehat{\mu}_{\text{tr}}^2 + 2\widehat{\mu}_{\text{tr}}\widehat{\mu}_{\text{te}},$$

and we select the wrong model if this quantity is greater than 0. We can re-write this as: we select the wrong model if

$$(\sqrt{n}\widehat{\mu}_{\text{tr}})^2 - 2(\sqrt{n}\widehat{\mu}_{\text{tr}})(\sqrt{n})\widehat{\mu}_{\text{te}}) < 0.$$

Now we observe that, $\sqrt{n_{\text{tr}}}\widehat{\mu}_{\text{tr}}$ and $\sqrt{n_{\text{te}}}\widehat{\mu}_{\text{te}}$ are each independent $N(0,1)$ variables, so the probability of choosing the wrong model is

$$P(Z_1^2 - 2Z_1 Z_2 < 0)$$

where $Z_1, Z_2 \sim N(0,1)$. This is about 0.35 (no matter how large n is).

Again, this is not really a problem because choosing the true model is not the goal. Also, even if we choose the wrong model, $\widehat{\theta}_1$ will be close to 0 so $p(x;\widehat{\theta}_0)$ will be close to $N(0,1)$.

Now consider AIC. We select $\mathcal{M}_1$ if

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \geq \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^2 + \frac{1}{n}.$$

Re-arranging this we see that we would select the wrong model (i.e. Model 1) if,

$$\widehat{\mu}^2 \geq \frac{2}{n}.$$

This intuitively makes sense: if the mean is small in absolute value we select Model 0 and otherwise we select Model 1. Now $n\widehat{\mu}^2 \sim \chi_1^2$ so the probability of choosing $\mathcal{M}_1$ is $P(\chi_1^2 > 2) = 0.16$.

Now consder BIC. You can go through exactly the same calculation as above and see that BIC would select the wrong model if,

$$\widehat{\mu}^2 \geq \frac{\log n}{n},$$

which has probability $P(\chi_1^2 > \log n) \to 0$. If $\mu \neq 0$, you can check that the probability of choosing Model 1 goes to 1. Thus BIC is model seletion consistent. Thus is true of BIC more generally.

Finally, let us return to hypothesis testing. We would reject the null if

$$n\widehat{\mu}^2 \geq \chi_{1,\alpha}^2,$$

and this controls the Type I error (i.e. the error of incorrectly selecting the more complex model) at α. More generally, we could imagine testing between pairs of models using the LRT (using Wilks result for the asymptotic distributions). This procedure is very similar to AIC but inflates the penalty just enough to ensure that we have some specified error control.

7 What to do?

The derivations of AIC and BIC depend on many technical assumptions about the model. CV does not depend on any model assumptions. For this reason, CV is the better choice if it is feasible.

A difficult problem that we have not considered is how to account for model selection when doing inference. This is a complicated topic. The simplest think, if we have lots of data, is to keep some hold out data. After model selection, we use the hold out data which is not affected by the model selection process.

Another issue that we have not considered is interpretability. Getting good predictions is not the only goal. We might be willing to sacrifice a bit of prediction accuracy to have a more interpretable model. This is an area of active research.

Lecture Notes 26
36-705

Today we will discuss nonparametric density estimation and nonparametric regression. First, we need to define kernels.

1 Kernels

A kernel function $K(x)$ for $x \in \mathbb{R}$ is a function K such that $\int K(x)dx = 1$ and K is symmetric, i.e. $\int xK(x)dx = 0$. We will also assume that $K(x) \geq 0$ and $\int x^2K(x)dx$. Examples are: the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

the boxcar kernel

$$K(x) = I(|x| < 1/2)$$

and the Epanechnikov

$$K(x) = \frac{3}{4}(1-x^2)I(|x| < 1).$$

Given a kernel K and a number $h > 0$ called the bandwidth, we define

$$K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right).$$

Similarly, for $x \in \ mathbbR^d$ we define $K : \mathbb{R}^d \to \mathbb{R}$ where K is symmetric and integrates to 1. Then, given a symmetric positive definite bandwidth matrix H we define

$$K_H(x) = \frac{1}{|H|}K(H^{-1}x).$$

A common choice is to take $H = hI$ and

$$K_H(x) = \frac{1}{h^d}\prod_j K\left(\frac{x_j}{h}\right)$$

where K is a one-dimensional kernel.

2 Non-parametric Density Estimation

Let $Y_1, \ldots, Y_n \sim p$. We'll focus on the case $Y_i \in \mathbb{R}$. We want to estimate p nonparametrically. A common estimator is the kernel density estimator defined by

$$\widehat{p}(y) = \frac{1}{n}\sum_i K_h(Y_i - y)$$

where K_h is a kernel with bandwidth h. You should think of $h = h_n$ as a number that decreases with sample size. We will assume that $p''(y) < \infty$.

Let's analyze this estimator. First

$$\begin{aligned}
\mathbb{E}[\widehat{p}(y)] &= \int K_h(u-y)p(u)du = \int K(t)p(y+th)dt \\
&\approx \int K(t)\left[p(y) + thp'(y) + \frac{t^2h^2}{2}p''(y)\right]dt \\
&= p(y) + hp'(y)\int tK(t)dt + \frac{h^2p''(y)}{2}\int t^2K(t)dt \\
&= p(y) + c_1(y)h^2
\end{aligned}$$

where $c_1(y) = p''(y)\int t^2K(t)dt/2$. So the bias is $c_1(y)h^2$.

Now we find the variance. We have

$$\text{Var}[\widehat{p}(y)] = \frac{1}{n}\text{Var}[K_h(Y-y)] = \frac{1}{n}\mathbb{E}[K_h^2(Y-y)] - \frac{1}{n}(\mathbb{E}[K_h(Y-y)])^2.$$

Now

$$\begin{aligned}
\mathbb{E}[K_h^2(Y-y)] &= \int \frac{1}{h^2}K^2((u-y)/h)p(u)du = \frac{1}{h}\int K^2(t)p(y+th)dt \\
&= \frac{1}{h}\int K^2(t)[p(y) + thp'(y) + \cdots]dt \approx \frac{c_2p(y)}{h}
\end{aligned}$$

where $c_2 = \int t^2K(t)dt$. Next

$$(\mathbb{E}[K_h(Y-y)])^2 \approx (p(y) + c_1(y)h^2)^2 \approx p^2(y).$$

So

$$\text{Var}[\widehat{p}(y)] \approx \frac{c_2np(y)}{nh} + \frac{p(y)}{n} \approx \frac{p(y)}{nh}.$$

Note that, if $h \to 0$ and $nh \to \infty$ as $n \to \infty$ then the bias and variance go to 0 and hence $\widehat{p}(y) \xrightarrow{P} p(y)$.

Next, consider the *integrated mean squared error* IMSE:

$$\text{IMSE} = \mathbb{E}[\int (\widehat{p}(y) - p(y))^2 dy] = \int \mathbb{E}[(\widehat{p}(y) - p(y))^2] dy$$
$$= \int \left(c_1^2(y) h^4 + \frac{c_2 p(y)}{nh} \right) dy = c_1 h^4 + \frac{c_2}{nh}$$

where $c_1 = \int c^2(y) dy$.

Here we say the bias-variance tradeoff. As h increases, the bias increases and the variance decreases and vice versa. The IMSE is minimized by choosing

$$h_n = \left(\frac{c_2}{4 c_1 n} \right)^{1/5} \approx \left(\frac{1}{n} \right)^{1/5}.$$

With this choice, we see that

$$\text{IMSE} = O\left(\frac{1}{n} \right)^{4/5}.$$

In practice, h is usually chosen by a version of cross-validation. In d dimensions it turns out that the IMSE is $O(n^{-4/(4+d)})$. The effect of dimension is brutal and is called the curse of dimensionality.

3 Non-parametric Regression

We observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ and our goal is to estimate the regression function

$$r(x) = \mathbb{E}[Y | X = x].$$

We *integrated* squared loss

$$L(\widehat{r}, r) = \int (\widehat{r}(x) - r(x))^2 dx.$$

The risk is then

$$R(\widehat{r}, r) = \mathbb{E}\left(\int (\widehat{r}(x) - r(x))^2 dx \right).$$

We will assume that $r''(y) < \infty$.

As in the case of point estimation we have a bias variance decomposition. First we define the point-wise bias:

$$b(x) = \mathbb{E}(\widehat{r}(x)) - r(x),$$

and the point-wise variance:

$$v(x) = \mathbb{E}\Big(\widehat{r}(x) - \mathbb{E}(\widehat{r}(x))\Big)^2.$$

Now, as before we can verify that:

$$R(\widehat{r}, r) = \int b^2(x)dx + \int v(x)dx.$$

A natural strategy in non-parametric regression is to locally average the data, i.e. our estimate of the regression function at any point will be the average of the Y values in a small neighborhood of the point.

The width of this neighborhood will determine the bias and variance. Too large a neighborhood will result in high bias and low variance (this is called oversmoothing) and too small a neighborhood will result in low bias but large variance (this is known as undersmoothing).

4 Optimal Regression Function

Suppose we knew the joint distribution over (X, Y). One could alternatively begin by defining the risk of an estimate $\widehat{r}$ as

$$R(\widehat{r}) = \mathbb{E}(Y - \widehat{r}(X))^2.$$

This risk simply measures the prediction error, i.e. the expected error we make in predicting Y when we use the function $\widehat{r}(X)$. This risk is minimized by the conditional expectation, i.e. we have the following theorem.

Theorem 1 *The risk R is minimized by*

$$r(x) = \mathbb{E}(Y|X = x).$$

Proof: Let $g(x)$ be any function of x. Then

$$\begin{aligned}
R(g) &= \mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - r(X) + r(X) - g(X))^2 \\
&= \mathbb{E}(Y - r(X))^2 + \mathbb{E}(r(X) - g(X))^2 + 2\mathbb{E}((Y - r(X))(r(X) - g(X))) \\
&\geq \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}((Y - r(X))(r(X) - g(X))) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\mathbb{E}\left((Y - r(X))(r(X) - g(X)) \;\middle|\; X\right) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\Big((\mathbb{E}(Y|X) - r(X))(r(X) - g(X))\Big) \\
&= \mathbb{E}(Y - r(X))^2 + 2\mathbb{E}\Big((r(X) - r(X))(r(X) - g(X))\Big) \\
&= \mathbb{E}(Y - r(X))^2 = R(r).
\end{aligned}$$

■

5 Kernel Regression

One of the most basic ways of doing non-parametric regression is called kernel regression. We will analyze kernel regression when we only have one covariate. The general case is not very different. The estimator is defined as:

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where the weights assign more importance to points near x. This is called a kernel regressor when the weights are chosen according to a kernel, i.e. we have weights:

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} = \frac{K_h(X_i - x)}{\sum_j K_h(X_i - x)}$$

where, as before, the bandwidth h controls the amount of smoothing.

To analyze this estimator, note that we can write

$$Y_i = r(X_i) + \epsilon_i$$

where ϵ_i has mean 0. Now

$$\begin{aligned}
\widehat{r}(x) &= \frac{\sum_i Y_i K_h(X_i - x)}{\sum_i K_h(X_i - x)} = \frac{\frac{1}{n}\sum_i Y_i K_h(X_i - x)}{\frac{1}{n}\sum_i K_h(X_i - x)} \\
&= \frac{\frac{1}{n}\sum_i Y_i K_h(X_i - x)}{\widehat{p}(x)} = \frac{\frac{1}{n}\sum_i Y_i K_h(X_i - x)}{p(x) + o_P(1)} \\
&\approx \frac{\frac{1}{n}\sum_i Y_i K_h(X_i - x)}{p(x)}.
\end{aligned}$$

Let's find the mean and variance of the numerator. We have

$$\begin{aligned}
\mathbb{E}[Y K_h(X - x)] &= \int\int y K_h(u - x) p(x, y) du dy = \int K_h(u - x) \int y p(y|u) dy p(u) du \\
&= \int K_h(u - x) r(u) p(u) du = \int K(t) r(x + th) p(x + th) dt \\
&\approx \int K(t) \left[r(x) + thr'(x) + \frac{t^2 h^2}{2} r''(x) \right] \left[p(x) + thp'(x) + \frac{t^2 h^2}{2} p''(x) \right] dt \\
&= r(x)p(x) + \frac{ch^2}{2}[r(x)p''(x) + 2r'(x)p'(x) + r''(x)p(x)]
\end{aligned}$$

where $c = \int t^2 K(t)$. Hence,

$$\mathbb{E}[\widehat{r}(x)] = r(x) + Ch^2.$$

By a similar calculation

$$\text{Var}[\widehat{r}(x)] = \frac{C}{nh}.$$

We conclude that

$$\text{IMSE} = ch^4 + \frac{c}{nh}$$

where now we use c generically to define constants. As in density estimation, the best bandwidth is $h_n \asymp n^{-1/5}$ and the risk is $n^{-4/5}$.

The analysis reveals that the bias depends on $p'(x)$ and $p(x)$. These terms can be removed by using better estimators.

6 The general case

So far we assumes that $r''(y) < \infty$. More generally, suppose that the β^{th} derivative of $r(x)$ is bounded, and we are in d-dimensions. In this case the bias will be roughly:

$$b^2(x) \approx h^{2\beta},$$

and the variance:

$$v(x) \approx \frac{1}{nh^d},$$

and balancing these will lead to the rate of convergence:

$$R(\widehat{r}, r) \approx n^{-2\beta/(2\beta+d)}.$$

This reveals another crucial feature of non-parametrics. In linear regression, the rate of convergence is typically something like:

$$R(\widehat{\beta}, \beta) \approx \frac{d}{n}.$$

In both cases, the situation gets worse as d increases, however in non-parametrics the situation gets *exponentially* worse. This is often colloquially referred to as the *curse of dimensionality.*

7 RKHS regression

There is another method also referred to as kernel regression. More precisely, it is Reproducing Kernel Hilbert Space (RKHS) regression. We will not cover this in much detail but here is te general idea.

A symmetric bivariate function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{x_i\}_{i=1}^n$ where each $x_i \in \mathcal{X}$, the $n \times n$ matrix K with elements $K_{ij} := K(x_i, x_j)$ is positive semidefinite.

Here are a few standard examples:

1. Linear kernel: When $\mathcal{X} = \mathbb{R}^d$ then $K(x_i, x_j) = \langle x_i, x_j \rangle = \sum_{u=1}^d x_{iu} x_{ju}$, is the linear kernel and is PSD.
2. Polynomial kernel: Again when $\mathcal{X} = \mathbb{R}^d$ then $K(x_i, x_j) = (\langle x_i, x_j \rangle)^m$, is the homogenous polynomial kernel of degree $m \geq 2$. This kernel is also PSD. The inhomogenous polynomial kernel $K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^m$ is also PSD.
3. Gaussian kernel: Perhaps the most popular kernel in machine learning is the Gaussian kernel. Here we take $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/(2\sigma^2))$.

Given dataa $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ we are going to estimate the function r by a function r_α which we will assume has the form:

$$r_\alpha(x) = \sum_{i=1}^n \alpha_i K(x_i, x),$$

where we need to estimate the α_i's. To do this we will minimize a least-squares type objective:

$$\widehat{\alpha} = \arg\min_\alpha \frac{1}{2} \sum_{i=1}^n (Y_i - r_\alpha(X_i))^2 + \lambda \text{Pen}(r_\alpha).$$

The penalty we will use is something called an RKHS norm penalization and it takes the form:

$$\text{Pen}(r_\alpha) = \alpha^T K \alpha,$$

where K is the gram matrix, i.e. $K_{ij} = K(x_i, x_j)$. This penalty encourages the function to be smooth but this is not easy to see without going into more detail. Observe that we can write:

$$\begin{bmatrix} r_\alpha(X_1) \\ r_\alpha(X_2) \\ \vdots \\ r_\alpha(X_n) \end{bmatrix} = K\alpha,$$

so the RKHS regression objective simplifies to:

$$\widehat{\alpha} = \arg\min_{\alpha} \frac{1}{2}\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha,$$

which we can solve in closed form (just by taking derivatives and setting to zero) as:

$$\widehat{\alpha} = (K + \lambda I)^{-1} Y.$$

Our estimated regression function then takes the form:

$$r_{\widehat{\alpha}}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(X_i, x).$$

Superficially there are similarities between RKHS regression and kernel regression. They both produce a function whose value at a given point is a weighted combination of the Y_i values at other points. In kernel regression the weights are easy to interpret, while in RKHS regression the weights are the solution to a least squares problem and are not directly interpretable.

From a practical standpoint, RKHS regression typically has two tuning parameters: the penalty parameter λ and usually some RKHS parameter (for instance the RKHS kernel bandwidth for a Gaussian kernel).

There are two types of problems one could ponder: (1) we want to fit a function that is Lipschitz or Holder smooth (as we analyzed in the first half): in this case, it is perhaps natural to use kernel regression and somewhat more artificial to use RKHS regression (2) we want to fit a function in a particular RKHS, in this case it is perhaps more natural to use RKHS regression.

From a theoretical standpoint, RKHS regression is usually analyzed using variants of the Rademacher complexity results we saw earlier in the course, i.e. they are not directly analyzed in terms of the bias and variance (this is because the RKHS regression procedure is naturally viewed as ERM over an RKHS). This means that the rates of convergence are typically specified in terms of properties of the RKHS and the data-generating distribution, i.e. a typical measure of complexity of an RKHS is the decay-rate of eigenvalues of the kernel gram matrix. This is quite unlike kernel regression where the function class is something simple (Lipschitz functions), and the measure of complexity is just the smoothness of the function.

RKHS regression is not the only alternative to kernel regression. Often you will see methods like k-NN regression (where you predict at a point by averaging the y values of the k-closest points), local polynomial regression (where you chop up the domain and fit (low-degree) polynomials in each piece of the domain) and orthogonal series estimators or projection estimators (where you expand the regression function in a orthogonal basis – say of sine/cosine type functions – and then estimate the coefficients in this basis).

Lecture Notes 27
36-705

Today we will discuss distances and metrics between distributions that are useful in statistics. We will discuss them in two contexts:

1. There are metrics that are analytically useful in a variety of statistical problems, i.e. they have intimate connections with estimation and testing.

2. There are metrics that are useful in data analysis, i.e. given data we want to measure some notion of distance between (the distribution of) subsets of the data and use this in some way.

There is of course overlap between these contexts and so there are distances that are useful in both, but they are motivated by slightly different considerations.

1 The Fundamental Statistical Distances

There are four notions of distance that have an elevated status in statistical theory. Let P, Q be two probability measures with densities p and q.

1. Total Variation: The TV distance between two distributions is:

$$\text{TV}(P,Q) = \sup_A |P(A) - Q(A)| = \sup_A \left| \int (p(x) - q(x))dx \right|,$$

where A is just any measurable subset of the sample space, i.e. the TV distance is measuring the maximal difference between the probability of an event under P versus under Q.

The TV distance is equivalent to the ℓ_1 distance between the densities, i.e. one can show that:

$$\text{TV}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx.$$

One can also write the TV distance as:

$$\text{TV}(P,Q) = \sup_{\|f\|_\infty \leq 1} \left| \mathbb{E}_P[f] - \mathbb{E}_Q[f] \right|.$$

2. The χ^2 divergence: The χ^2 divergence is defined for distributions P and Q such that Q dominates P, i.e. if $Q(A) = 0$ for some set A then it has to be the case that $P(A)$ is also 0. For such distributions:

$$\chi^2(P, Q) = \int_{\{x:q(x)>0\}} \frac{p^2(x)}{q(x)} dx - 1.$$

Alternatively, one can write this as:

$$\chi^2(P, Q) = \int_{\{x:q(x)>0\}} \frac{(p(x) - q(x))^2}{q(x)} dx.$$

3. Kullback-Leibler divergence: Again we suppose that Q dominates P. The KL divergence between two distributions:

$$\mathrm{KL}(P, Q) = \int \left(\log \frac{p(x)}{q(x)} \right) p(x) dx.$$

4. Hellinger distance: The Hellinger distance between two distributions is,

$$H(P, Q) = \left[\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \right]^{1/2},$$

i.e. the Hellinger distance is the ℓ_2 norm between $\sqrt{p}$ and $\sqrt{q}$. It might seem at first a bit weird to consider the ℓ_2 norm between $\sqrt{p}$ and $\sqrt{q}$ rather than p and q, but this turns out to be the right thing to do for statistical applications. The use of Hellinger is various statistical contexts was popularized by Lucien Le Cam, who advocated strongly (and convincingly) for thinking of square-root of the density as the central object of interest.

The Hellinger distance is also closely related to what is called the *affinity* (or Bhattacharya coefficient):

$$\rho(P, Q) = \int \sqrt{p(x)q(x)} dx.$$

In particular, note the equality:

$$H^2(P, Q) = 2 \left(1 - \rho(P, Q)\right).$$

All of these fundamental statistical distances are special cases of what are known as f-divergences. The field of information theory has devoted considerable effort to studying families of distances (α, β, ϕ, f-divergences) and so on, and this has led to a fruitful interface between statistics and information theory. An f-divergence is defined for a *convex* function f with $f(1) = 0$:

$$D_f(P, Q) = \int q(x) f \left(\frac{p(x)}{q(x)} \right) dx.$$

You can look up (for instance on Wikipedia) which functions lead to each of the divergences we defined above.

2 Hypothesis Testing Lower Bounds

A basic hypothesis testing problem is the following: suppose that we have two distributions P_0 and P_1, and we consider the following experiment: I toss a fair coin and if it comes up heads I give you a sample from P_0 and if it comes up tails I give you a sample from P_1. Let $T = 0$ if the coin comes up heads and $T = 1$ otherwise. You only observe the sample X, and need to tell me which distribution it came from.

This is exactly like our usual simple versus simple hypothesis testing problem, except I pick each hypothesis with probability $1/2$. Now, suppose you have a test $\Psi : X \mapsto \{0, 1\}$. Define its error rate as:

$$\mathbb{P}(\Psi(X) \neq T) = \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right].$$

Roughly, we might believe that if P_0 and P_1 are close in an appropriate distance measure then the error rate of the best possible test should be high and otherwise the error rate should be low. This is known an Le Cam's Lemma.

Lemma 1 *For any two distributions P_0, P_1, we have that,*

$$\inf_{\Psi} \mathbb{P}(\Psi(X) \neq T) = \frac{1}{2}\left(1 - TV(P_0, P_1)\right).$$

Before we prove this result we should take some time to appreciate it. What Le Cam's Lemma tells us is that if two distributions are close in TV then *no test* can distinguish them. In some sense TV is the right notion of distance for statistical applications. In fact, in theoretical CS, the TV distance is sometimes referred to as the *statistical distance*. It is also the case that the likelihood ratio test achieves this bound exactly (should not surprise you since it is a simple-vs-simple hypothesis test).

Proof: For any test Ψ we can denote its acceptance region A, i.e. if $X \in A$ then $\Psi(X) = 0$. Then,

$$\begin{aligned}\frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] &= \frac{1}{2}\left[\mathbb{P}_0(X \notin A) + \mathbb{P}_1(X \in A)\right] \\ &= \frac{1}{2}\left[1 - (\mathbb{P}_0(X \in A) - \mathbb{P}_1(X \in A))\right].\end{aligned}$$

So to find the best test we simply minimize the RHS or equivalently:

$$\begin{aligned}\inf_{\Psi} \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] &= \frac{1}{2}\left[1 - \sup_{A}(\mathbb{P}_0(X \in A) - \mathbb{P}_1(X \in A))\right] \\ &= \frac{1}{2}\left(1 - \mathrm{TV}(P_0, P_1)\right).\end{aligned}$$

Close analogues of Le Cam's Lemma hold for all of the other divergences above, i.e. roughly, if any of the χ^2, Hellinger or KL divergences are small then we cannot reliably distinguish between the two distributions. If you want a formal statement see Theorem 2.2 in Tsybakov's book.

3 Tensorization

Given the above fact, that all of the distances we defined so far are in some sense "fundamental" in hypothesis testing, a natural question is why do we need all these different distances?

The answer is a bit technical, but roughly, when we want to compute a lower bound (i.e. understand the fundamental statistical difficulty of our problem) some divergences might be easier to compute than others. For instance, it is often the case that for mixture distributions the χ^2 is easy to compute, while for many parametric models the KL divergence is natural (in part because it is closely related to the Fisher information. Knowing which divergence to use when is a bit of an art but having many tools in your toolbox is always useful.

One natural thing that will arise in statistical applications is that unlike the above setting of Le Cam's Lemma we will observe n i.i.d. samples $X_1, \ldots, X_n \sim P$, rather than just one sample. Everything we have said so far works in exactly the same way, except we need to calculate the distance between the product measures, i.e. $d(P^n, Q^n)$, which is just the distance between the distributions:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i).$$

For the TV distance this turns out to be quite difficult to do directly. However, one of the most useful properties of the Hellinger and KL distance is that they *tensorize*, i.e. they behave nicely with respect to product distributions. In particular, we have the following useful relationships:

$$\begin{aligned}
&\mathrm{KL}(P^n, Q^n) = n\mathrm{KL}(P, Q) \\
&H^2(P^n, Q^n) = 2\left(1 - \left(1 - \frac{H^2(P, Q)}{2}\right)^n\right) = 2\left(1 - \rho(P, Q)^n\right),
\end{aligned}$$

where $\rho(p, q) = \int \sqrt{pq}$ is the affinity defined earlier. The key point is that when we see n i.i.d samples it is easy to compute the KL and Hellinger.

4 Hypothesis Testing Upper Bounds

One can ask if there are analogous upper bounds, i.e. for instance if the distance between P_0 and P_1 gets larger, are there quantitatively better tests for distinguishing them?

For Hellinger and TV the answer turns out to be yes (and for χ^2 and KL the answer is yes under some assumptions). Formally, given n samples from either P_0 or P_1 you can construct tests that distinguish between P_0 and P_1 such that for some constant $c_1, c_2 > 0$:

$$\inf_{\Psi} \frac{1}{2} \left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq c_1 \exp(-c_2 n \mathrm{TV}(P_0, P_1)),$$

and similarly

$$\inf_{\Psi} \frac{1}{2} \left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq c_1 \exp(-c_2 n H^2(P_0, P_1)).$$

These are sometimes called large-deviation inequalities (and in most cases precise constants are known).

It turns out that even for distinguishing two hypotheses that are separated in the Hellinger distance, the likelihood ratio test is optimal (and achieves the above result). The proof is short and elegant.

Proof: We recall the elementary bound:

$$\log x \leq x - 1, \quad \text{for all } x \geq 0.$$

Which in turn tells us that:

$$\log \rho(P_0, P_1) \leq \log \rho(P_0, P_1) - 1 = -\frac{H^2(P_0, P_1)}{2}.$$

So now let us analyze the LRT which rejects the null if:

$$\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1.$$

Let us study its Type I error (its Type II error bound follows essentially the same logic). We note that:

$$\begin{aligned} P_0\left(\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1\right) &= P_0\left(\prod_{i=1}^{n} \frac{P_1(X_i)}{P_0(X_i)} \geq 1\right) \\ &= P_0\left(\prod_{i=1}^{n} \sqrt{\frac{P_1(X_i)}{P_0(X_i)}} \geq 1\right) \\ &\leq \mathbb{E}_{P_0} \prod_{i=1}^{n} \sqrt{\frac{P_1(X_i)}{P_0(X_i)}}, \end{aligned}$$

using Markov's inequality. Now, using independence we see that:

$$\begin{aligned} P_0\left(\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1\right) &\leq \left[\mathbb{E}_{P_0}\sqrt{\frac{P_1(X)}{P_0(X)}}\right]^n \\ &= \exp(n \log \rho(P_0, P_1)) \\ &\leq \exp(-nH^2(P_0, P_1)/2). \end{aligned}$$

Putting this together with an identical bound under the alternate we obtain,

$$\inf_{\Psi} \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq \exp(-nH^2(P_0, P_1)/2).$$

The result is quite nice – it says that the LRT can distinguish two distributions reliably provided their Hellinger distance is large compared to $1/\sqrt{n}$. Furthermore, the bound has an exponential form so you might imagine that it will interact nicely with a union bound (in a multiple testing setup where we want to distinguish between several distributions).

5 Inequalities

Given the fact that these four distances are fundamental and that they have potentially different settings where they are easy to use it is also useful to have inequalities that relate these distances.

The following inequalities reveal a sort of hierarchy between the distances:

$$\mathrm{TV}(P, Q) \leq H(P, Q) \leq \sqrt{\mathrm{KL}(P, Q)} \leq \sqrt{\chi^2(P, Q)}.$$

This chain of inequalities should explain why it is the case that if any of these distances are too small we cannot distinguish the distributions. In particular, if any distance is too small then the TV must be small and we then use Le Cam's Lemma.

There are also reverse inequalities in some cases (but not all). For instance:

$$\frac{1}{2}H^2(P, Q) \leq \mathrm{TV}(P, Q) \leq H(P, Q),$$

so up to the square factor Hellinger and TV are closely related.

6 Distances from parametric families

We have encountered these before: they are usually only defined for parametric families:

1. Fisher information distance: for two distributions $P_{\theta_1}, P_{\theta_2}$ we have that,

$$d(P_{\theta_1}, P_{\theta_2}) = (\theta_1 - \theta_2)^T I(\theta_1)^{-1}(\theta_1 - \theta_2).$$

2. Mahalanobis distance: for two distributions $P_{\theta_1}, P_{\theta_2}$, with means μ_1, μ_2 and covariances Σ_1, Σ_2, the Mahalanobis distance would be:

$$d(P_{\theta_1}, P_{\theta_2}) = (\mu_1 - \mu_2)^T \Sigma_1^{-1}(\mu_1 - \mu_2).$$

   This is just the Fisher distance for the Gaussian family with known covariance.

7 Robustness to Model Misspecification

Another strong motivation for studying estimation or testing in various metrics stems from robustness considerations. We have already seen that Maximum Likelihood inherits a certain type of KL-robustness, i.e. if we observe samples $X_1, \ldots, X_n \sim P$ where possibly $P \notin \mathcal{P}_\theta$, then MLE still makes sense and asymptotically (under some regularity conditions) will find us a distribution $P_{\widehat{\theta}}$ such that,

$$\mathrm{KL}(P, P_{\widehat{\theta}}) \leq \mathrm{KL}(P, P_\theta) \quad \forall \ \theta \in \Theta.$$

One way to interpret this statement is that the MLE is robust to model-misspecification in the KL distance, i.e. if P was close to $\mathcal{P}_\theta$ in the sense that for some $P_\theta \in \mathcal{P}_\theta$, $\mathrm{KL}(P, P_\theta) \leq \epsilon$ then the MLE would automatically find us a distribution which was asymptotically at most ϵ-far in KL from P.

Of course, this is not the only notion of model mis-specification that we might care about, and a general idea is to tailor the estimation procedure to the notion of model mis-specification that we expect.

These lead to so-called minimum distance estimators. These are estimators that attempt to find a distribution in $\mathcal{P}_\theta$ that is close to the samples or the true distribution P in the some distance – say the KL/TV/Hellinger etc. Minimum distance estimators are typically robust to mis-specification in their native distance, i.e. the TV minimum distance estimator will be robust to model-misspecification in TV.

Classically, outlier robustness was studied in something called the Huber ϵ-contamination model, where we observe samples:

$$X_1, \ldots, X_n \sim (1 - \epsilon)P_\theta + \epsilon Q,$$

where Q is an arbitrary distribution, i.e. ϵ-fraction of the samples are arbitrarily corrupted (outliers). It is easy to see that,

$$\mathrm{TV}((1 - \epsilon)P_\theta + \epsilon Q, P_\theta) \leq \epsilon,$$

so that Huber's model is very closely related to model mis-specification in TV. As a result, the TV minimum distance estimator is very robust to outliers. The main drawback is however a computational one: the TV minimum distance estimator is often difficult to compute.

8 Distances for data analysis

Although the distances we have discussed so far are fundamental for many analytic purposes, we do not often use them in data analysis. This is because they can be difficult to estimate in practice. Ideally, we want to roughly be able to trade-off how "expressive" the distance is versus how easy it is to estimate (and the distances so far are on one end of the spectrum).

To be a bit more concrete lets think about a typical setting (closely related to two-sample testing that we have discussed earlier): we observe samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$, and we want to know how different the two distributions are, i.e. we want to *estimate* some divergence between P and Q given samples.

This idea has caught on again recently because of GANs, where roughly we want to generate samples that come from a distribution that is close to the distribution that generated the training samples, and often we do this by estimating a distance between the training and generated distributions and then trying to make it small.

Another popular task is *independence testing* or *measuring dependence* where we are given samples $(X_1, Y_1), \dots, (X_n, Y_n) \sim P_{XY}$ and we want to estimate/test how far apart P_{XY} and $P_X P_Y$ are (i.e. we want to know how far apart the joint and product of marginals are).

8.1 Integral Probability Metrics

If two distributions P, Q are identical then it should be clear that for any (measurable) function f, it must be the case that:

$$\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}[f(Y)].$$

One might wonder if the reverse implication is true, i.e. is it that if $P \neq Q$ then there must be some *witness* function f such that:

$$\mathbb{E}_{X\sim P}[f(X)] \neq \mathbb{E}_{Y\sim Q}[f(Y)].$$

It turns out that this statement is indeed true. In particular, we have the following lemma.

Lemma 2 *Two distributions P, Q are identical if and only if for every continuous function $f \in C(\mathcal{X})$*

$$\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}[f(Y)].$$

This result suggests that to measure the distance between two distributions we could use a so-called integral probability metric (IPM):

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]|,$$

where $\mathcal{F}$ is a class of functions. We note that the TV distance is thus just an IPM with $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$. This class of functions (as well as the class of all continuous functions) is too large to be useful statistically, i.e. it is the case that these IPMs are not easy to estimate from data, so we instead use function classes that have *smooth* functions.

8.2 Wasserstein distance

A natural class of smooth functions is the class of 1-Lipschitz functions, i.e.

$$\mathcal{F}_L = \{f : f \text{continous }, |f(x) - f(y)| \leq \|x - y\|\}.$$

The corresponding IPM is known as the Wasserstein 1 distance:

$$W_1(P, Q) = \sup_{f \in \mathcal{F}_L} |\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]|.$$

As with the TV there are many alternative ways of defining the Wasserstein distance. In particular, there are very nice interpretations of Wasserstein as a distance between "couplings" and as a so-called transportation distance.

The Wasserstein distance has the somewhat nice property of being well-defined between a discrete and continuous distribution, i.e. the two distributions you are comparing do not need to have the same support. This is one of the big reasons why they are popular in ML.

In particular, a completely reasonable estimate of the Wasserstein distance between two distributions, given samples from each of them is the Wasserstein distance between the corresponding empirical measures, i.e. we estimate:

$$\widehat{W_1}(P, Q) = W_1(\mathbb{P}_n, \mathbb{Q}_n),$$

where $\mathbb{P}_n$ (for instance) is the distribution that puts mass $1/n$ on each sample point.

There are many other nice ways to interpret the Wasserstein distance, and it has many other elegant properties. It is central in the field of optimal transport. A different expression for the Wasserstein distance involves coupling: i.e. a joint distribution J over X and Y, such that the marginal over X is P and over Y is Q. Then the W_1 distance (or more generally W_p distance) is:

$$W_p^p(P, Q) = \inf_J \mathbb{E}\|X - Y\|_2^p.$$

One can also replace the Euclidean distance by any metric on the space on which P and Q are defined. More generally, there is a way to view Wasserstein distances as measuring the cost of optimally moving mass of the distribution P to make it look like the distribution Q (hence, the term optimal transport).

The Wasserstein distance also arises frequently in image processing. In part, this is because the Wasserstein barycenter (a generalization of a mean) of a collection of distributions preserves the shape of the distribution. This is quite unlike the "usual" average of the distributions.

8.3 Maximum Mean Discrepancy

Another natural class of IPMs is where we restrict $\mathcal{F}$ to be a unit ball in a Reproducing Kernel Hilbert Space (RKHS). An RKHS is a function space associated with a kernel function k that satisfies some regularity conditions. We will not be too formal about this but you should think of an RKHS as another class of smooth functions (just like the 1-Lipschitz functions) that has some very convenient properties. A kernel is some measure of similarity, a commonly used one if the RBF kernel:

$$k(X, Y) = \exp\left[-\frac{\|X - Y\|_2^2}{2\sigma^2}\right]$$

The MMD is a typical IPM:

$$\text{MMD}(P, Q) = \sup_{f \in \mathcal{F}_k} |\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]|,$$

but because the space is an RKHS with a kernel k, it turns out we can write this distance as:

$$\text{MMD}^2(P, Q) = \mathbb{E}_{X,X' \sim P} k(X, X') + \mathbb{E}_{Y,Y' \sim Q} k(Y, Y') - 2\mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).$$

Intuitively, we are contrasting how similar the samples from P look to each other and Q look to each other to how similar the samples of P are to Q. If P and Q are the same then all of these expectations should be the same and the MMD will be 0.

The key point that makes the MMD so popular is that it is completely trivial to estimate the MMD since it is a bunch of expected values for which we can use the empirical expectations. We have discussed U-statistics before, the MMD can be estimated by a simple U-statistic:

$$\widehat{MMD}^2(P, Q) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(X_i, X_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(Y_i, Y_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(X_i, Y_j).$$

In particular, if the kernel is bounded then we can use a Hoeffding-style bound to conclude that we can estimate the MMD at $1/\sqrt{n}$ rates.

However, the usefulness of the MMD hinges not on how well we can estimate it but on how strong a notion of distance it is, i.e. for distributions that are quite different (say in the TV/Hellinger/...distances), is it the case that the MMD is large? This turns out to be quite a difficult question to answer.

9 Fano's Inequality

We focused primarily on the role of f-divergences in testing but they are equally fundamental in providing lower bounds for estimation. In estimation we obtain samples $X_1, \ldots, X_n \sim P_\theta$ where $P_\theta \in \mathcal{P}_\Theta$, and our goal is to estimate θ (say with small ℓ_2 error).

In an intuitive sense, estimation is a lot like like a multiple hypothesis testing problem of the following form – I give you samples from one of M distributions $\{P_{\theta_1}, \ldots, P_{\theta_M}\}$ and I ask you to figure out which one it was. Our estimator Ψ in this context simply takes in n samples and returns an index $\{1, \ldots, M\}$.

We could imagine the setting where we sample an index u uniformly from $\{1, \ldots, M\}$ and generate n samples from P_{θ_u}. We define the following notion of error:

$$\text{err} = P(\Psi(X_1, \ldots, X_n) \neq u).$$

Fano's inequality (and others like it) relate how hard this testing problem is (i.e. they lower bound err) to a function of some distance between the distributions $\{P_{\theta_1}, \ldots, P_{\theta_M}\}$.

Suppose that $M \geq 3$, and for some small constant $c_1 > 0$:

$$\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathrm{KL}(P_{\theta_i} \| P_{\theta_j}) \leq c_1 \log M,$$

then for some other $c_2 > 0$, err $> c_2$. In words, Fano's inequality says if the average pairwise KL divergence is small then multiple testing problem is difficult.

So how does this relate to estimation? Suppose we additionally ensure that for all pairs (i, j) we have that $\|\theta_i - \theta_j\|_2^2 \geq (2\delta)^2$, then it is easy to verify that the minimax estimation error:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \geq \inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_1, \ldots, \theta_M\}} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \geq \inf_{\widehat{\theta}} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{X_1, \ldots, X_n \sim P_{\theta_i}} \|\widehat{\theta} - \theta_i\|_2^2 \geq c_2 \delta^2,$$

using Markov's inequality.

So now we can try to understand the Fano inequality game. To produce a tight lower bound, we need to find many hypothesis (large M) such that their average KL divergence is sufficiently small, but their corresponding parameters are well-separated. Intuitively, this

should make sense, if we can find parameters that are well-separated but the underlying distributions are very close, then the estimation problem should be difficult.

An Application: So why do we need Fano's inequality? Suppose we consider establishing a lower bound for estimating the mean of a Normal distribution. We have already seen that the minimax error is at least $\sigma^2 d/n$ (but this required a complicated Bayes argument). On the other using a simple versus simple testing problem (with two separated Normals) we can easily show via Le Cam's lemma that the error is at least σ^2/n. To get the right dimension dependence we will need to use the full power of Fano's inequality.

Let us see how this works. We need to know a few facts: one is that there is a packing of the radius 4δ sphere of size $c2^d$ (for some $c > 0$), such that:

$$\|\theta_i\|_2 \leq 4\delta, \text{ and } \|\theta_i - \theta_j\| \geq 2\delta$$

i.e. there are roughly 2^d vectors in the 4δ-sphere that are well-separated (i.e. are separated by at least 2δ).

Additionally, we can calculate the KL divergence between n-samples from $N(\theta_i, I_d)$ and $N(\theta_j, I_d)$,

$$\mathrm{KL}(P_{\theta_i}, P_{\theta_j}) = \frac{n}{2\sigma^2}\|\theta_i - \theta_j\|_2^2 \leq \frac{cn\delta^2}{\sigma^2}.$$

So (ignoring constants) if $n\delta^2 \ll \sigma^2 d$, then our minimax estimation error is at least δ^2. So this gives us that the minimax estimation error is at least $c\sigma^2 d/n$.