Introduction to Inferential Statistics

Matthew Chen

September 8, 2025

Contents

In inferential statistics, we make the assumption that collected data are part of a larger population, and that by analyzing the collected data, we can infer properties of the parent distribution. The theory introduced in these notes are foundational for understanding the field of statistics.

1 Limit Theorems for Sample Mean

1.1 Sample Mean

A large part of the study of statistics is to understand the sum of iid random variables $X_1, ..., X_n$ with common mean $\mu = \mathbb{E}X_i$ and variance $\sigma^2 = Var(X_i) = \mathbb{E}X_i^2 - (\mathbb{E}X_i)^2$. We define the sample mean as

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

which has expectation

$$\mathbb{E}\bar{X} = \mu$$

and variance

$$Var(\bar{X}) = Var\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

1.2 Notions of Convergence

Deterministically, a sequence of numbers a_n converges to a if for any $\epsilon > 0$ there exists $N = N_\epsilon$ such that for every $n \geq N_\epsilon$ we have that $|a_n - a| \leq \epsilon$. For such as case, we denote $a_n \to a$ or $\lim_{n\to\infty} a_n = a$.

In contrast, sequences of random variables require different notions of convergence, which include convergence in probability, convergence in distribution, and convergence almost surely.

1.2.1 Convergence in Probability

A sequence of random variables X_n converges to a random variable X if for any $\epsilon > 0$ we have that

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

That is, for a sequence of converging random variables, the event that X_n is at least ϵ units away from X becomes very unlikely for large n. This notion of convergence is prevalent in statistics, for example, a statistical estimator is deemed *consistent* if it converges in probability to the estimated quantity. Convergence in probability is also established by the weak law of large numbers.

Often, we denote convergence in probability as the following

$$plim_{n\to\infty} X_n = X$$

$$X_n \xrightarrow{p} X$$

1.2.2 Convergence in Distribution

The idea behind convergence in distribution is that the distribution associated with a member of the sequence X_n becomes arbitrarily close to the specified distribution of X as n grows large. That is, X_n converges in distribution to X if the cumulative distribution function (cdf) of X_n converges to the cdf of X.

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

where $F_X(x) = \mathbb{P}(X \leq x)$ is the cumulative distribution function of X.

Convergence in distribution is the weakest notion of convergence, and is implied by convergence in probability. We can denote convergence in distribution as

$$X_n \overset{d}{\to} X$$

1.2.3 Almost Sure Convergence

The idea behind almost sure convergence is that events where the sequence X_n does not converge to X has probability zero, i.e.

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

Almost sure convergence is the strongest notion of convergence, and implies convergence in probability which in turn implies convergence in distribution. For example, almost sure convergence is provided by the strong law of large numbers. We denote almost sure convergence as

$$X_n \overset{a.s.}{\to} X$$

1.3 Weak Law of Large Numbers (LLN)

The weak law of large numbers provides that the sample mean converges to the true mean μ in probability. We first show Markov's Inequality to show Chebyshev's Inequality from which the law of large numbers follows.

1.3.1 Markov's Inequality

Markov's Inequality states that if X is a *non-negative* random variable and for a *non-negative* scalar $t > 0$ we have that

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}X}{t}$$

Proof. Let $A = \{X > t\}$ be the event that X is greater than t and let $A^c = \{X \leq t\}$ be the complement of A. Let I_A and I_{A^c} be indicator random variables for the events A and A^c, respectively. We want to show that

$$\mathbb{E}X \geq t\mathbb{P}(X > t)$$

Starting from the left hand side, we have

$$\begin{aligned}
\mathbb{E}X &= \mathbb{E}\left((I_A + I_{A^c})X\right) \\
&= \mathbb{E}(I_A X) + \mathbb{E}(I_{A^c} X) \\
&\geq \mathbb{E}(I_A X) \\
&\geq t\mathbb{E}I_A \ (if\ A\ occurs,\ X > t) \\
&= t\mathbb{P}(A) = t\mathbb{P}(X > t)
\end{aligned}$$

1.3.2 Chebyshev's Inequality

Chebyshev's Inequality states that for *any* random variable X and $t > 0$, then

$$\mathbb{P}\left(|X - \mathbb{E}X| > t\right) \leq \frac{Var(X)}{t^2}$$

Proof. Starting from the left hand side, we have

$$\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}X| > t\right) &= \mathbb{P}\left((X - \mathbb{E}X)^2 > t^2\right) \\
&\leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{t^2} \ (by\ Markov's\ Inq.) \\
&= \frac{Var(X)}{t^2}
\end{aligned}$$

1.3.3 (Weak) Law of Large Numbers

Let $X_1, ..., X_n$ be iid random variables with mean μ and variance σ^2. Then $\bar{X}$ converges to μ in probability. That is, for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X} - \mathbb{E}X| > \epsilon) = 0$$

Proof. We can show the weak LLN easily with Chebyshev's Inequality. We have

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X} - \mathbb{E}X| > \epsilon) \leq \lim_{n\to\infty} \frac{Var(\bar{X})}{\epsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

Noting that the left hand side is non-negative, then $\lim_{n\to\infty} \mathbb{P}(|\bar{X} - \mathbb{E}X| > \epsilon) = 0$.

1.4 Central Limit Theorem

The central limit theorem gives an approximate distribution for $\bar{X}$. Specifically, a properly scaled sample mean $\bar{X}$ converges in distribution to the standard normal distribution. That is,

$$\frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} = \sqrt{\frac{n}{\sigma^2}}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, 1)$$

or equivalently,

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

Proof. For simplicity, assume $\mu = 0$ and $\sigma^2 = 1$ (we can always standardize X_i first). Let

$$Z_n = \frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} = \frac{X_1 + ... + X_n}{\sqrt{n}}$$

It suffices to show that the moment generating function (MGF) of Z_n, $M_{Z_n}(t)$, converges to the MGF of the standard normal distribution $M_Z(t) = e^{t^2/2}$ in the limiting case $n \to \infty$. This strategy is easier than using the cdf's directly. Let $M(t)$ be the MGF of X_i.

Starting with the MGF of Z_n, we have

$$\begin{aligned}
M_{Z_n}(t) &= \mathbb{E}e^{tZ_n} \\
&= \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}(X_1+...+X_n)}\right) = \mathbb{E}\left(e^{\frac{tX_1}{\sqrt{n}}}...e^{\frac{tX_n}{\sqrt{n}}}\right) \\
&= \left(\mathbb{E}e^{\frac{tX_1}{\sqrt{n}}}\right)^n \ (since X_1, ..., X_n \ independent) \\
&= \left(M(t/\sqrt{n})\right)^n
\end{aligned}$$

Since $t/\sqrt{n} \to 0$ as $n \to \infty$, we apply the Taylor expansion $g(x) = g(0) + g'(0)x + \frac{g''(0)}{2}x^2 + R(x)$ where the remainder term satisfies $R(x)/x^2 \to 0$ as $x \to 0$. This gives us

$$\begin{aligned}
M_{Z_n}(t) &= \left(M(t/\sqrt{n})\right)^n \\
&= \left(M(0) + M'(0)\frac{t}{\sqrt{n}} + M''(0)\frac{(t/\sqrt{n})^2}{2} + R(t/\sqrt{n})\right)^n \\
&= \left(1 + t^2/(2n) + R(t/\sqrt{n})\right)^n
\end{aligned}$$

The above is true from our assumptions of mean and variance. Recall that from a moment generating function, we have

$$\begin{aligned}
M(0) &= \mathbb{E}e^0 = 1 \\
M'(0) &= \mathbb{E}X_1 = \mu = 0 \\
M''(0) &= \mathbb{E}X_1^2 = Var(X_1) = 1
\end{aligned}$$

Also, from the Taylor expansion we have that

$$\frac{R(t/\sqrt{n})}{(t/\sqrt{n})^2} \to 0 \; as \; t/\sqrt{n} \to 0$$

which is equivalent to

$$nR(t/\sqrt{n}) \to 0 \; as \; n \to \infty$$

At this point, we can apply the result from calculus that $\left(1 + \frac{a_n}{n}\right)^n \to e^a$ as $n \to \infty$ if $a_n \to a$. Rewriting $M_{Z_n}(t)$ we have

$$\begin{aligned} M_{Z_n}(t) &= \left(1 + t^2/(2n) + R(t/\sqrt{n})\right)^n \\ &= \left(1 + \frac{1}{n}\left(\frac{t^2}{2} + nR(t/\sqrt{n})\right)\right)^n \end{aligned}$$

Since the MGF of the standard normal is $M_Z(t) = e^{t^2/2}$, we want

$$\frac{t^2}{2} + nR(t/\sqrt{n}) \to \frac{t^2}{2} \; as \; n \to \infty$$

which easily follows since $nR(t/\sqrt{n}) \to 0 \; as \; n \to \infty$. This concludes the proof.

Example. A fair coin is tossed 70 times. Estimate the probability that the number of heads is at most 37 using the CLT (note that we can use Binomial(70, 1/2) to calculate the exact probability).

Let $S = X_1 + ... + X_{70}$ be the number of heads, where $X_i \overset{iid}{\sim} Bernoulli(1/2)$. We want to solve

$$\mathbb{P}(S \leq 37) = \mathbb{P}(\bar{x} \leq 37/70)$$

We have that $\mathbb{E}\bar{x} = \mu = 1/2$ and $Var(\bar{x}) = \sigma^2/n = 1/(2^2 * 70)$. Then we have

$$\mathbb{P}(\bar{x} \leq 37/70) = \mathbb{P}\left(\frac{\bar{x} - 1/2}{\sqrt{1/280}} \leq \frac{37/70 - 1/2}{\sqrt{1/280}}\right) \approx \mathbb{P}\left(Z \leq \frac{37/70 - 1/2}{\sqrt{1/280}}\right)$$

which we can evaluate using

$$\Phi\left(\frac{37/70 - 1/2}{\sqrt{1/280}}\right)$$

where Φ is the standard normal CDF.

1.5 Delta Method

Recall that the central limit theorem tells us $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. The Delta Method answers the question about the limiting distribution of an arbitrary differentiable function of $\bar{x}$, say $g(\bar{x})$.

Specifically, suppose $\hat{\theta}_n$ is an estimate of θ with limiting distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2)$$

Let g be any differentiable function with $g'(\theta) \neq 0$. Then by the Delta Method we have

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \to \mathcal{N}(0, \sigma^2\left(g'(\theta)\right)^2)$$

For intuition, consider the first order Taylor expansion

$$g(\hat{\theta}_n) \approx g(\theta) + g'(\theta)(\hat{\theta}_n - \theta)$$

Then

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \approx g'(\theta)\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, (g'(\theta))^2\sigma^2)$$

Example. Let $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(p)$. Estimate the odds $p/(1-p)$ and the limiting distribution of the estimate.

We can first estimate p using $\bar{x}$. Then by the central limit theorem, we have

$$\sqrt{n}(\bar{x} - p) \to \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, p(1-p))$$

Let the function $g(p) = p/(1-p)$. We can estimate the odds by evaluating the estimate into the odds function. That is, we can use

$$g(\hat{p}) = g(\bar{x}) = \bar{x}/(1-\bar{x})$$

To obtain a limiting distribution, we can apply the Delta Method. First evaluating the derivative of g at p we have

$$g'(p) = \frac{1}{(1-p)^2}$$

Then by the Delta Method, we have

$$\begin{aligned}\sqrt{n}\left(\frac{\bar{x}}{1-\bar{x}} - \frac{p}{1-p}\right) &\to \mathcal{N}(0, \sigma^2(g'(p))^2) \\ &= \mathcal{N}\left(0, p(1-p)\frac{1}{(1-p)^4}\right) \\ &= \mathcal{N}\left(0, \frac{p}{(1-p)^3}\right)\end{aligned}$$

2 Methods of Estimation

In this section, we discuss methods to estimate the unknown parameters θ of a theorized statistical model or distribution by using n iid data observations $X_1, ..., X_n$ from the desired distribution.

2.1 Method of Moments

The method of moments (MOM) estimates statistical parameters by first estimating first and higher order moments. The method is useful to provide quick estimates, but may not provide the most efficient estimators.

Recall that the *kth moment* of a random variable X is defined

$$\mu_k = \mathbb{E}X^k = M_x^{(k)}(0)$$

Given $X_1, ..., X_n$ iid random variables with mean μ and variance σ^2 we have from the law of large numbers that $\bar{x}$ converges in probability to μ and from the central limit theorem, we have converge in distribution to a standard normal random variable. The LLN/CLT show that $\hat{\mu}_1 := \bar{x}$ is a good estimate of the first moment $\mu_1 = \mu$.

Further, if $X_1, ..., X_n$ are iid random variables, then $X_1^k, ..., X_n^k$ are also iid random variables with mean $\mu_k = \mathbb{E}X^k$. Then by the law of large numbers,

$$\hat{\mu}_k = \frac{X_1^k + ... + X_n^k}{n} \xrightarrow{p} \mu_k$$

and by the central limit theorem

$$\frac{\hat{\mu}_k - \mu_k}{\sqrt{Var(\hat{\mu}_k)}} \xrightarrow{d} N(0,1)$$

in distribution. That is, $\hat{\mu}_k$ can be used to estimate the kth moment.

Use of the method of moments is as follows. First, we find the first few moments of X as functions of the unknown parameters. For example, we can find

$$\begin{aligned}\mu_1 &= g_1(\theta_1, \theta_2)\\ \mu_2 &= g_2(\theta_1, \theta_2)\end{aligned}$$

Next, we can estimate the moments by find the sample moments $\hat{\mu}_1, \hat{\mu}_2$

$$\begin{aligned}\hat{\mu}_1 &= \frac{X_1 + ... + X_n}{n}\\ \hat{\mu}_2 &= \frac{X_1^2 + ... + X_n^2}{n}\end{aligned}$$

Replacing the moments by the sample moments, we have the system

$$\begin{aligned}\hat{\mu}_1 &\approx g_1(\theta_1, \theta_2)\\ \hat{\mu}_2 &\approx g_2(\theta_1, \theta_2)\end{aligned}$$

where the solution are the method of moments estimators

$$\begin{aligned}\hat{\theta}_1 &= h_1(\hat{\mu}_1, \hat{\mu}_2)\\ \hat{\theta}_2 &= h_2(\hat{\mu}_1, \hat{\mu}_2)\end{aligned}$$

Example. Normal distribution. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with unknown μ, σ^2. Estimate the unknown parameters with the Method of Moments.
Solving for the moments we have

$$\begin{aligned}\mu_1 &= \mathbb{E}X = \mu\\ \mu_2 &= \mathbb{E}X^2 = Var(X) + (\mathbb{E}X)^2 = \sigma^2 + \mu_2\end{aligned}$$

Solving using the sample moments, we have the estimates

$$\begin{aligned}\hat{\mu} &= \hat{\mu}_1 = \bar{X}\\ \hat{\sigma}^2 &= \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{X_1^2 + ... + X_n^2}{n} - \bar{X}^2\end{aligned}$$

Example. Exponential distribution. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} \mathit{Exponential}(\lambda)$ with unknown parameter λ. Estimate the unknown parameters with the Method of Moments.

Recall that the Exponential pdf is

$$f(x) = \lambda e^{-\lambda x},\ x \geq 0$$

Solving for the first moment

$$\mu_1 = \mathbb{E}X = \int_0^\infty x\lambda e^{-\lambda x} dx = 1/\lambda$$

so we can estimate λ using the first sample moment (sample mean) as follows

$$\hat{\lambda} = \frac{1}{\hat{\mu}_1} = \frac{1}{\bar{X}}$$

We can also find a limiting distribution for the estimate using the Delta Method. From the CLT we have

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) \to \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 1/\lambda^2)$$

Recall that the variance of an exponential random variable is $1/\lambda^2$. Previously we showed that

$$\lambda = 1/\mu_1$$

so setting $\lambda = g(\mu_1) = 1/\mu_1$ and $\hat{\lambda} = g(\hat{\mu}_1) = 1/\bar{x}$ we have

$$g'(\mu_1) = \frac{-1}{\mu_1^2} = \lambda^2$$

Then applying the Delta Method, we have

$$\begin{aligned}
\sqrt{n}(g(\hat{\mu}_1) - g(\mu_1)) &\to \mathcal{N}\left(0, \sigma^2[g'(\mu_1)]^2\right) \\
\sqrt{n}(\hat{\lambda} - \lambda) &\to \mathcal{N}\left(0, \frac{1}{\lambda^2}(\lambda^2)^2\right) \\
\sqrt{n}(\hat{\lambda} - \lambda) &\to \mathcal{N}\left(0, \lambda^2\right)
\end{aligned}$$

2.2 Method of Maximum Likelihood

Suppose we toss a coin with unknown head probability p and observe 40 heads out of 100 tosses. If we must choose one of the following p, which should we choose?

1. $p = 0.1$. Then $\mathbb{P}(S = 40) = \binom{100}{40}0.1^{40}(1-0.1)^{60} \approx 10^{-15}$
2. $p = 0.8$. Then $\mathbb{P}(S = 40) = \binom{100}{40}0.8^{40}(1-0.8)^{60} \approx 10^{-18}$
3. $p = 0.5$. Then $\mathbb{P}(S = 40) = \binom{100}{40}0.5^{40}(1-0.5)^{60} \approx 0.01$

In such a case, it is reasonable to choose $p = 0.5$ since it corresponds with the highest probability of observing 40 heads. Similarly, the method of maximum likelihood seeks to estimate parameters by maximizing the likelihood of the observed data, i.e. we choose our statistical model so that it is the most probable to observe the observed data.

Let $X_1, ..., X_n$3w be random variables with joint density $f(x_1, ..., x_n|\theta)$ that depends on the parameter θ. We observe the data $(X_1 = x_1), (X_2 = x_2), ..., (X_n = x_n)$. Then holding the data fixed, the likelihood function is a function of θ

$$L(\theta) = f(x_1, ..., x_n|\theta)$$

The *maximum likelihood estimator* (MLE) of θ is defined as

$$\hat{\theta} = \underset{\theta}{argmax}\ L(\theta)$$

Example. Bernoulli distribution. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(p)$. Estimate p using the MLE.

Recall that the pmf for the $Bernoulli(p)$ distribution is

$$p(x) = \mathbb{P}(X = x) = p^x(1-p)^{1-x},\ x \in [0, 1]$$

Then by independence, the joint distribution over $X_1, ..., X_n$ is the product of the individual pmf's

$$\begin{aligned}
p(x_1, ..., x_n) &= p(x_1)p(x_2)...p(x_n) \\
&= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}...p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_i x_i}(1-p)^{\sum_i(1-x_i)}
\end{aligned}$$

Note that since the logarithm is monotonically increasing, we can apply it to the joint and arrive at the same optimal p. Then the log-likelihood is

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

We are solving for $\hat{p} = \underset{p}{argmax}\ \ell(p)$. Setting $\ell'(p) = 0$, we have

$$\frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-x_i) = 0$$

Using the property that $\frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d}$, we have

$$\frac{\sum_i x_i}{p} = \frac{\sum_i (1-x_i)}{1-p} = \frac{n}{1}$$

which we can rearrange to obtain

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note that it can be shown that $\ell(p)$ is concave, that is, $\ell''(p) < 0$ and the solution is the unique maximum.

Example. Poisson distribution. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the MLE of λ.

Recall that the pmf for the Poisson distribution is

$$p(x) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Then by independence the joint distribution is

$$p(x_1, ..., x_n) = \frac{e^{-\lambda}\lambda_1^x}{x_1!} ... \frac{e^{-\lambda}\lambda_n^x}{x_n!} = \frac{e^{-\lambda n}\lambda^{(\sum_i x_i)}}{x_1!x_2!...x_n!}$$

And the log-likelihood is

$$\ell(\lambda) = -\lambda n + log(\lambda)\sum_{i=1}^{n} - log(x_1!...x_n!)$$

Solving for $\ell'(\lambda) = 0$, we have

$$-n + \frac{1}{\lambda}\sum_{i=1}^{n} = 0$$

Which can be rearranged as

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

As before, it can be shown that $\ell''(\lambda) < 0$ everywhere so that the solution is the unique maximum.

Example. Normal distribution. Suppose we have $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Find the MLE's for μ and σ.

Recall that the normal pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\ x \in \mathbb{R}$$

With joint distribution

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

Then the log-likelihood is

$$\begin{aligned}\ell(\mu, \sigma^2) &= -nlog\left((2\pi\sigma^2)^{1/2}\right) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \\ &= \frac{-n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\end{aligned}$$

Solving for $\partial\ell/\partial\mu = 0$ we have

$$\begin{aligned}\frac{2}{2\sigma^2}\sum_{i=1}^n (x_i - \mu) &= 0 \\ \sum_{i=1}^n (x_i - \mu) &= 0 \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^n x_i\end{aligned}$$

Let $\tau = \sigma^2$. Solving for $\partial\ell/\partial\tau = 0$ we have

$$\begin{aligned}\frac{-n}{2\tau} + \frac{1}{2\tau^2}\sum_{i=1}^n (x_i - \mu)^2 &= 0 \\ \hat{\tau} = \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

As before, we can verify that the second derivatives are negative.

Example. Suppose $X_1, ..., X_n \overset{iid}{\sim} Uniform(0, \theta)$ where θ is an unknown parameter. Find the MLE for θ.

Recall that the pdf for the Uniform distribution is

$$f(x) = \frac{1}{\theta}1\{x \in [0, \theta]\}$$

Then by independence we have that the joint pdf is

$$\begin{aligned}f(x_1, ..., x_n) &= \frac{1}{\theta^n}1\{x_1 \in [0, \theta]\}1\{x_2 \in [0, \theta]\}...1\{x_n \in [0, \theta]\} \\ &= \frac{1}{\theta^n}1\{x_{max} \in [0, \theta]\}1\{x_{min} \in [0, \theta]\}\end{aligned}$$

Unlike the other example, the joint is not differentiable. However, we make the observation that for the likelihood to be maximized, $I = 1$ which requires $x_{max} \leq \theta$ and $x_{min} \geq 0$. Under these conditions, the likelihood becomes $1/\theta^n$, a decreasing function in θ. Thus, to maximize the likelihood, we choose the smallest possible θ under the stated constraints and arrive at

$$\hat{\theta} = x_{max}$$

This is a reasonable estimate!

2.2.1 Invariance Property of the MLE

It can be shown that if $\hat{\theta}$ is the MLE for θ, then for some arbitrary function $\tau(\theta)$, $\tau(\hat{\theta})$ is the MLE for $\tau(\theta)$. This leads to the "plug in" principle. For example, the MLE for variance $p(1-p)$ in the Bernoulli example is $\hat{p}(1-\hat{p})$. Similarly, in the Gaussian example, we evaluate for $\bar{x}$ instead of μ to arrive at the standard deviation $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

2.2.2 Exponential Families

As an aside, we define a family of pdfs or pmfs to be the *exponential family* if can be expressed in the form

$$f(x;\theta) = h(x)c(\theta)exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right) = exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x) + C(\theta) + H(x)\right)$$

For example, the Bernoulli distribution is in the exponential family and its pmf can be written as

$$p(x) = p^x(1-p)^{1-x} = exp\left(xlog(p) + (1-x)log(1-p)\right)$$

Other examples of exponential family distributions include Normal, Chi-squared, Poisson, Binomial, Negative Binomial, Geometric, Exponential, Gamma, and Beta. The form of the exponential family is chosen for mathematical convenience.

2.3 Bayesian Estimators

In the classical "Frequentist" approach, the parameter θ is thought be an unknown but fixed quantity that represents the "true" population distribution. Estimates or confidence intervals around θ can be formed as a function of a random sample is drawn from the population. For example, constructed confidence interval bounds are functions of random data, and represent the probability that the random bounds capture the true (fixed) θ. Note that the data is treated as random variables (and so are functions of the data), while the parameter is treated as fixed.

In the Bayesian approach, θ is treated as a random variable that follows a prior distribution, denoted $\pi(\theta)$, representing the experimenter's belief. The observed data is treated as evidence that updates our belief about θ, that is, we update the prior $\pi(\theta)$ to for the posterior $\pi(\theta|x)$ given the data x using Bayes' Theorem

$$\pi(\theta|x) = \frac{\pi(\theta,x)}{\pi(x)} = \frac{\pi(\theta)\pi(x|\theta)}{\int_{-\infty}^{+\infty}\pi(\theta,x)d\theta} = \frac{\pi(\theta)f(x|\theta)}{\int_{-\infty}^{+\infty}\pi(\theta)f(x|\theta)d\theta}$$

We then use the posterior to infer about the parameter θ. For example, we can use the mean of the posterior distribution as a point estimate, or we can construct credible intervals that represent the probability of the parameter being within a certain interval, given the data at hand and our prior belief. Unlike confidence intervals, here the credible interval bounds are treated as fixed and the parameter as random. The data is treated as fixed in the sense that we form the posterior conditioned on the data that is observed.

Example. Lifetime of components. Suppose $X_1,...,X_n \overset{iid}{\sim} Exponential(\theta)$. We assume a Gamma prior $\theta \sim \Gamma(\alpha,\beta)$. Find the posterior distribution.

Recall that the Exponential pdf is

$$f(x|\theta) = \theta e^{-\theta x},\ x \geq 0$$

and the Gamma pdf with parameters α and β is

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}1\{\theta \in (0,\infty)\}$$

where Γ is the gamma function.

We first solve for the joint, that is

$$\begin{aligned}\pi(\theta, x_1, ..., x_n) &= \pi(\theta)\pi(x_1, ..., x_n|\theta)\\ &= \pi(\theta)\pi(x_1|\theta)...\pi(x_n|\theta) \ (since\ x_1, ..., x_n\ indp)\\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\left(\theta^n e^{-\theta\sum_i x_i}\right)1\{\theta \in (0,\infty)\}\\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha+n-1}e^{-\theta\left(\beta+\sum_i x_i\right)}1\{\theta \in (0,\infty)\}\end{aligned}$$

The marginal distribution is

$$\begin{aligned}\pi(x_1, ..., x_n) &= \int_{-\infty}^{+\infty}\pi(\theta, x_1, ..., x_n)d\theta\\ &= \int_0^{+\infty}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha+n-1}e^{-\theta\left(\beta+\sum_i x_i\right)}d\theta \ (note\ \theta \geq 0)\end{aligned}$$

Finally, we have that the posterior is

$$\begin{aligned}\pi(\theta|x_1, ..., x_n) &= \frac{\pi(\theta, x_1, ..., x_n)}{\pi(x_1, ..., x_n)}\\ &= \frac{\left(\beta+\sum_i x_i\right)^{\alpha+n}}{\Gamma(\alpha+n)}\theta^{\alpha+n-1}e^{-\theta\left(\beta+\sum_i x_i\right)}1\{\theta \in (0,\infty)\}\\ &\sim \Gamma\left(\alpha+n, \beta+\sum_{i=1}^n x_i\right)\end{aligned}$$

The posterior follows a Gamma distribution with parameters $\alpha+n$ and $\beta+\sum_i x_i$. As $n \to \infty$, the posterior distribution will concentrate around the "true" θ.

Note that the marginal $\pi(x_1, ..., x_n)$ is a constant that does not depend on θ. It only serves as a factor so that the posterior integrates to one. We only need to show that the joint distribution is proportional to a known distribution to find the posterior.

2.3.1 Conjugate Family

Let $F = \{f(x|\theta), \theta \in \Omega\}$ be a class of pdf's or pmf's indexed by θ. A class Π of prior distributions π is a conjugate family for F if the posterior $\pi(\theta|x) \in \Pi$ for all $f \in F$, $\pi \in \Pi$ and all x. That is, the prior and posterior belong to the same family of distributions for some class of data distribution $f(x|\theta)$.

- *Gamma* is conjugate for *Exponential, Poisson*
- *Beta* is conjugate for *Bernoulli, Binomial, Negative Binomial, Geometric*
- *Normal* is conjugate for *Normal*

2.3.2 Loss Function

In the Bayesian approach, for each estimate a, the expected loss functions $L(\theta, a)$ is considered to produce a point estimate. We have by LOTUS

$$\mathbb{E}[L(\theta, a)|x] = \int_\Omega L(\theta, a)\pi(\theta|x)d\theta$$

A Bayesian estimator of θ is defined as

$$\delta(x) = \underset{a}{argmin} \mathbb{E}[L(\theta,a)|x]$$

For example, with the most common loss we have square loss, i.e. $L(\theta,a) = (\theta-a)^2$. Then we have that

$$\mathbb{E}[(\theta-a)^2|x] = \mathbb{E}[(\theta-\mathbb{E}(\theta|x)+\mathbb{E}(\theta|x)-a)^2|x] = \mathbb{E}[(\theta-\mathbb{E}(\theta|x))^2|x] + [\mathbb{E}(\theta|x)-a]^2$$

which is minimized when $a = \delta(x) = \mathbb{E}(\theta|x)$ as the first term does not depend on a. That is, using the square error loss produces an estimate representing the mean of the posterior.

Similarly, when taking the absolute loss $L(\theta,a) = |\theta-a|$, the estimate $\delta(x)$ is the median of the posterior.

Example. Clinical Trials. Suppose the result of a treatment is $X_1,...,X_n \overset{iid}{\sim} Bernoulli(p)$. Using a Beta prior $p \sim Beta(\alpha,\beta)$, estimate p using square loss.

Recall that the Beta density is

$$f(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}1\{p \in [0,1]\}$$

From the Bernoulli distribution, we have

$$\pi(x_1,...,x_n|p) = p^{\sum_i x_i}(1-p)^{\sum_i(1-x_i)}$$

The the joint distribution is

$$\begin{aligned}\pi(p,x_1,...,x_n) &= f(p|\alpha,\beta)\pi(x_1,...,x_n|p)\\ &= \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}p^{\sum_i x_i}(1-p)^{\sum_i(1-x_i)}1\{p \in [0,1]\}\\ &= \frac{1}{B(\alpha,\beta)} p^{\alpha+\sum_i x_i-1}(1-p)^{\beta+\sum_i(1-x_i)-1}1\{p \in [0,1]\}\end{aligned}$$

where $B(\alpha,\beta)$ is the Beta function. We see that the joint is proportional to a new Beta distribution. Then the posterior is

$$p|x \sim Beta\left(\alpha + \sum_i x_i, \beta + \sum_i (1-x_i)\right)$$

Recall that the Beta mean is $\alpha/(\alpha+\beta)$. Then our Bayes estimator is

$$\hat{p} = \frac{\alpha + \sum_i x_i}{\alpha+\beta+n}$$

Example. Customer Arrivals. Suppose $X_1,...,X_n \overset{iid}{\sim} Poisson(\theta)$ with Gamma prior, $\theta \sim Gamma(\alpha,\beta)$.

We have that the joint is proportional to a new Gamma distribution

$$\begin{aligned}\pi(\theta,x_1,...,x_n) &= \pi(\theta)\pi(x_1,...,x_n|\theta)\\ &\propto \theta^{\alpha-1}e^{-\beta\theta}[e^{-n\theta}\theta^{\sum_i x_i}]1\{\theta \in (0,\infty)\}\\ &= \theta^{\alpha+\sum_i x_i-1}e^{-\theta(\beta+n)}1\{\theta \in (0,\infty)\}\end{aligned}$$

Then the posterior

$$\theta|x \sim Gamma(\alpha + \sum_i x_i, \beta+n)$$

Recall that the Gamma mean is α/β, so using square loss, we estimate

$$\hat{\theta} = \frac{\alpha + \sum_i x_i}{\beta+n}$$

Note how for both the Poisson and Bernoulli example, the estimates are approximately $\bar{x}$ for large n.

3 Fisher Information and Cramer-Rao Lower Bound Efficiency

3.1 Best Unbiased Estimators

3.1.1 Mean Square Error

The mean square error (MSE) of an estimate $\hat{\theta}$ of θ is

$$MSE := \mathbb{E}\left(\hat{\theta}-\theta\right)^2$$

We can decompose MSE into bias and variance

$$\begin{aligned}MSE &= \mathbb{E}\left(\hat{\theta}-\theta\right)^2\\ &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta\right)^2\\ &= \mathbb{E}\left((\hat{\theta}-\mathbb{E}\hat{\theta})^2+(\mathbb{E}\hat{\theta}-\theta)^2+2(\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\right)\\ &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2+\left(\mathbb{E}\hat{\theta}-\theta\right)^2+2(\mathbb{E}\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\\ &= Var(\hat{\theta})+Bias(\hat{\theta})^2\end{aligned}$$

An estimator that has good MSE will have small combined variance and bias.

3.1.2 Uniform Minimum Variance Unbiase Estimator (UMVUE)

There is no one "best MSE" estimator for all values of a parameter; the class of all estimators is too large. Instead, we consider only unbiased estimators and choose the estimator with the smaller variance. A uniform minimum variance unbiased estimator $\hat{\theta}$ satisfies both $\mathbb{E}\hat{\theta}=\theta$ (unbiased) and $Var(\hat{\theta}) \leq Var(W)$ where W is any unbiased estimator of θ (minimum variance).

3.2 Cramer-Rao Lower Bound

Let $X_1, ..., X_n$ be iid random variables with density $f(x|\theta)$. Let $W = W(X_1, ..., X_n)$ be an *unbiased* estimator for $\tau(\theta)$, i.e. $\tau(\theta) = \mathbb{E}W$. Then the variance of W is lower bounded by

$$Var_\theta(W) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta W\right)^2}{n\mathbb{E}\left(\frac{\partial}{\partial\theta}logf(X_1|\theta)\right)^2} = \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta W\right)^2}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information

$$I(\theta) = \mathbb{E}\left(\frac{\partial}{\partial\theta}logf(X_1|\theta)\right)^2 = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}logf(X_1|\theta)\right)$$

Example. Let $X_1, ..., X_n \overset{iid}{\sim} Poisson(\theta)$. Show that the MLE achieves the Cramer-Rao lower bound.

The MLE for Poisson θ is $\hat{\theta} = \bar{x}$ which is unbiased, i.e. $\mathbb{E}\bar{X} = \theta$ with variance $Var(\bar{X}) = \frac{1}{n}Var(X_1) = \theta/n$.

Next, we find the Fisher information. Recall that the Poisson pmf is $p(x) = e^{-\theta}\theta^x/x!$ so taking the logarithm and its second derivative we have

$$\begin{aligned}log(p(x)) &= -\theta + xlog(\theta) - log(x!)\\ \frac{\partial}{\partial\theta}log(p(x)) &= -1+\frac{x}{\theta}\\ \frac{\partial^2}{\partial\theta^2}log(p(x)) &= \frac{-x}{\theta^2}\end{aligned}$$

So the Fisher information is

$$I(\theta) = -\mathbb{E}\left(\frac{-X_1}{\theta^2}\right) = \frac{1}{\theta^2}\mathbb{E}X_1 = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

In this case, the statistic $W(X) = \bar{X}$ is an unbiased estimator for $\tau(\theta) = \theta$. Then we have

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta W = \frac{\partial}{\partial\theta}\theta = 1$$

And the Cramer-Rao lower bound is

$$Var_\theta(W) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(W)\right)^2}{nI(\theta)} = \theta/n$$

The variance of the MLE is also θ/n so it achieves the Cramer-Rao lower bound.

3.2.1 Attainment

Let $X_1, ..., X_n$ be iid random variables with likelihood $L(\theta) = \prod_{i=1}^n f(x_i|\theta)$. If W is an estimator of $\tau(\theta)$, then it attains the Cramer-Rao lower bound if and only if

$$\frac{\partial}{\partial\theta}logL(\theta) = \alpha(\theta)[W - \tau(\theta)]$$

where α is some arbitrary function of θ.

Example. Suppose $X_1, ..., X_n \overset{iid}{\sim} Bernoulli(p)$. Let $\hat{p} = \bar{x}$ be an estimator for p. Show that $\hat{p}$ attains the CR-lower bound.

We have that the likelihood function is

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

Taking the derivative of the log-likelihood, we have

$$\begin{aligned}
\frac{\partial}{\partial p}logL(p) &= \frac{\partial}{\partial p}\left(log(p)\sum_i x_i + log(1-p)\sum_i(1-x_i)\right) \\
&= \frac{\sum_i x_i}{p} - \frac{\sum_i(1-x_i)}{1-p} \\
&= \sum_i\left(x_i\left(\frac{1}{p} + \frac{1}{1-p}\right)\right) - \frac{n}{1-p} \\
&= \frac{\sum_i x_i}{p(1-p)} - \frac{np}{p(1-p)} \\
&= \frac{n}{p(1-p)}[\bar{x} - p]
\end{aligned}$$

Setting $\alpha(p) = \frac{n}{p(1-p)}$, we see that the estimator attains the CR-lower bound.

Example. Suppose $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Show that $\hat{\lambda} = \bar{x}$ achieves the CR-lower bound.

The joint-likelihood is

$$L(\lambda) = \prod_{i=1}^n e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}$$

Taking the logarithm and differentiating

$$\begin{aligned}\frac{\partial}{\partial\lambda}logL(\lambda) &= \frac{\partial}{\partial\lambda}\left(-n\lambda + log(\lambda)\sum_i x_i\right)\\ &= -n + \frac{1}{\lambda}\sum_i x_i\\ &= \frac{n}{\lambda}[\bar{x} - \lambda]\end{aligned}$$

Setting $\alpha(\lambda) = n/\lambda$, we see that the estimator achieves the CR-lower bound.

3.3 Limiting Distribution of the MLE

Asymptotically, the MLE is unbiased and attains the Cramer-Rao lower bound. Specifically, let $X_1, ..., X_n$ be iid random variables with pdf/pmf $f(x|\theta)$. Under some smoothness conditions of f, as $n \to \infty$, the MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta))$$

where $v(\theta)$ is the Cramer-Rao lower bound

$$v(\theta) = \frac{1}{-\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}logf(X_1|\theta)\right)} = \frac{1}{I(\theta)}$$

Sketch of proof. Let $X_1, ..., X_n \overset{iid}{\sim} f(x|\theta)$ with log-likelihood $\ell(\theta) = \sum_{i=1}^n logf(x_i|\theta)$. By the definition of the MLE, we have that $\ell'(\hat{\theta}) = 0$. We can approximate $\ell'(\hat{\theta})$ with a Taylor expansion around the true θ so that

$$\begin{aligned}0 = \ell'(\hat{\theta}) &\approx \ell'(\theta) + \ell''(\theta)(\hat{\theta} - \theta)\\ \hat{\theta} - \theta &= \frac{-\ell'(\theta)}{\ell''(\theta)}\\ \sqrt{n}(\hat{\theta} - \theta) &= -\sqrt{n}\frac{\ell'(\theta)}{\ell''(\theta)}\end{aligned}$$

Evaluating, we have

$$\sqrt{n}(\hat{\theta} - \theta) = -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}logf(x_i|\theta)}{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}logf(x_i|\theta)}$$

In the numerator, the goal is to apply the central limit theorem. First, we find the expectation of $\frac{\partial}{\partial\theta}f(X|\theta)$.

$$\begin{aligned}\mathbb{E}\left(\frac{\partial}{\partial\theta}logf(X|\theta)\right) &= \mathbb{E}\left(\frac{1}{f(X|\theta)}\frac{\partial}{\partial\theta}f(X|\theta)\right)\\ &= \int_{\mathbb{R}}\frac{1}{f(x|\theta)}\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)f(x|\theta)dx\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f(x|\theta)dx\\ &= 0\ (since\ the\ pdf\ integrates\ to\ 1)\end{aligned}$$

Similarly, we find the variance of $\frac{\partial}{\partial\theta}f(X|\theta)$.

$$Var\left(\frac{\partial}{\partial\theta}f(X|\theta)\right) = \mathbb{E}\left(\left(\frac{\partial}{\partial\theta}f(X|\theta)\right)^2\right) = I(\theta)$$

Thus, we have n iid random variables with mean 0 and variance $I(\theta)$. Then by the central limit theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}f(X_i|\theta)\xrightarrow{d}\mathcal{N}(0,I(\theta))$$

In the denominator, we apply the law of large numbers, which gives us

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}logf(X_i|\theta)\xrightarrow{p}\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}logf(X|\theta)\right)=-I(\theta)$$

Slutsky's Theorem has that if X_n converges to an random element X in distribution, and Y_n converges in probability to a constant c then $X_nY_n\xrightarrow{d}cX$. Applying Slutsky's Theorem to the numerator and denominator, we have that

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}\frac{1}{I(\theta)}\mathcal{N}(0,I(\theta))=\mathcal{N}\left(0,\frac{1}{I(\theta)}\right)$$

as desired.

Note: if we have that $\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{I(\theta)}\right)$ then we have the pivotal quantity $\sqrt{nI(\theta)}(\hat{\theta}-\theta)\xrightarrow{d}\mathcal{N}(0,1)$ to construct a confidence interval, i.e.

$$\mathbb{P}\left(\hat{\theta}-\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{nI(\theta)}}<\theta<\hat{\theta}+\frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{nI(\theta)}}\right)=1-\alpha$$

where Φ^{-1} is the inverse-cdf (quantile function) of the standard normal. We can further make the approximation that $I(\theta)\approx I(\hat{\theta})$.

3.3.1 General Case of the MLE Limit

Let $X_1,...,X_n$ be iid random variables with pdf/pmf $f(x|\theta)$. Under some smoothness conditions of f, as $n\to\infty$, the MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\tau(\hat{\theta}_n)-\tau(\theta))\xrightarrow{d}\mathcal{N}(0,v(\theta))$$

where

$$v(\theta)=\frac{(\tau'(\theta))^2}{I(\theta)}$$

is the Cramer-Rao lower bound. The general case is true by the Delta method.

Example. Suppose $X_1,...,X_n\overset{iid}{\sim}Exponential(\theta)$. Estimate the mean $\mu=1/\theta$ using the MLE $\hat{\theta}=1/\bar{x}$ and find the limiting distribution.

We can estimate μ using $\tau(\hat{\theta})=1/\hat{\theta}=\bar{x}$ (recall the invariance property of the MLE).

The Fisher's information to is

$$I(\theta)=-\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}log\left(\theta e^{-\theta X}\right)\right)=-\mathbb{E}\left(\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}-X\right)\right)=\frac{1}{\theta^2}$$

We also have

$$\tau'(\theta)=-\frac{1}{\theta^2}$$

So the Cramer-Rao lower bound is

$$\frac{(\tau'(\theta))^2}{I(\theta)}=\frac{1}{\theta^2}$$

Finally, we have that

$$\sqrt{n}(\tau(\hat{\theta}_n)-\tau(\theta))=\sqrt{n}(\bar{x}-\mu)\xrightarrow{d}\mathcal{N}\left(0,\frac{1}{\theta^2}\right)$$

3.3.2 Asymptotic Relative Efficiency

Suppose we have estimators W and V for $\tau(\theta)$ satisfying

$$\sqrt{n}(W - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_w^2)$$
$$\sqrt{n}(V - \tau(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma_v^2)$$

The *asymptotic relative efficiency* of V with respect to W is defined as σ_w^2/σ_v^2.

3.4 Second Order Delta Method

Let $\hat{\theta}_n$ be an estimate of θ with limiting distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2)$$

Let g be any differentiable function with $g'(\theta) = 0$ and $g''(\theta) \neq 0$. Then we have that

$$n(g(\hat{\theta}_n) - g(\theta)) \to \left(\frac{\sigma^2 g''(\theta)}{2} \chi_1^2 \right)$$

where χ_1^2 is a chi-squared random variable with one degree of freedom.

4 Sufficient Statistics

Let $X_1, ..., X_n$ be and iid sample from a distribution indexed by parameter θ. Let $T = r(X_1, ..., X_n)$ be statistic. T is a sufficient statistic if the conditional distribution of the data given the sufficient statistic $T = t$ does not depend on the parameter θ. That is, the conditional joint distribution $X_1, ..., X_n | T = t, \theta$ is the same for all θ. Intuitively, a statistic calculated from a sample is sufficient if the sample itself can provide no additional information as to the value of θ.

4.1 Factorization Theorem

Let $X_1, ..., X_n$ form a random sample from pdf or pmf $f(x|\theta)$, where $\theta \in \Omega$ is unknown. A statistic $T = r(X_1, ..., X_n)$ is a sufficient statistic for θ if and only if the joint $f(x_1, ..., x_n|\theta)$ can be factored as follows for all x and all values of $\theta \in \Omega$

$$f(x_1, ..., x_n|\theta) = u(x)v(r(x), \theta)$$

where u and v are non-negative functions.

Proof. We will prove the factorization theorem in the discrete case. Denote the joint probability density of X and T by $f(x, t|\theta)$.

Suppose that T is a sufficient statistic. Since T is a function of X, we have that $f(x, t|\theta) = f(x|\theta)$ as long as $T = r(x) = t$ (otherwise we have a disjoint outcome and the density is zero). Then we have that

$$f(x|\theta) = f(x, t|\theta) = f(x|t, \theta)f(t|\theta)$$

By the definition of sufficient statistics, we have that $f(x|t) = f(x|t, \theta)$. Setting $u(x) = f(x|t)$ and $v(t, \theta) = f(t|\theta)$, we have that the expression can be factorized per the factorization theorem.

Conversely, suppose that we have $f(x|\theta) = u(x)v(t,\theta)$. We have that

$$\begin{aligned} f(x|t,\theta) &= \frac{f(x,t|\theta)}{f(t|\theta)} \\ &= \frac{f(x|\theta)}{f(t|\theta)} \\ &= \frac{u(x)v(t,\theta)}{f(t|\theta)} \end{aligned}$$

By the law of total probability, we have that

$$\begin{aligned} f(t|\theta) &= \sum_{x:r(x)=t} f(x,t|\theta) \\ &= \sum_{x:r(x)=t} f(x|\theta) \\ &= \sum_{x:r(x)=t} u(x)v(t,\theta) \\ &= v(t,\theta) \sum_{x:r(x)=t} u(x) \end{aligned}$$

Substituting this expression into the first expression, we have that

$$f(x|t,\theta) = \frac{u(x)v(t,\theta)}{v(t,\theta)\sum_{x:r(x)=t} u(x)} = \frac{u(x)}{\sum_{x:r(x)=t} u(x)}$$

which shows that $T = t$ is a sufficient statistic since $f(x|t,\theta)$ does not depend on θ.

4.1.1 Corollary to Factorization Theorem

A statistic $T = r(X)$ is sufficient if and only if, no matter what prior we use, the posterior distribution of θ depends only on the data through the value of T.

Example. Suppose $X_1, ..., X_n \overset{iid}{\sim} Poisson(\theta),\ \theta > 0$. Find a sufficient statistic.

Writing the joint distribution

$$f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) e^{-n\theta}\theta^{\sum_i x_i}$$

Let $u(x) = \prod_{i=1}^{n} \frac{1}{x_i!}$, $v(r(x),\theta) = e^{-n\theta}\theta^{\sum_i x_i}$. It follows that the sufficient statistic is $r(x) = \sum_i x_i$. Note that the sufficient statistic directly interacts with the parameter θ.

5 Multivariate Transformation

5.1 Differentiable Transformation

Let $X = (X_1, ..., X_n)$ have joint density $f_X(x)$ with support $S = \{x : f_X(x) \neq 0\}$. Consider a *one-to-one*, *differentiable* transformation from S to T, $r : \mathbb{R}^n \to \mathbb{R}^n$. Denote Y as the transformed variable

$$Y = r(X) = (r_1(X), ..., r_n(X))$$

and denote the inverse map

$$X = r^{-1}(Y) = (r_1^{-1}(Y), ..., r_n^{-1}(Y))$$

Then the joint density function of Y is given by

$$f_Y(y) = f_X(r^{-1}(y))|J|, \ h \in T$$

where J is the determinant of the Jacobian matrix

$$J = det\left(\frac{\partial r_i^{-1}(y)}{\partial y_j}\right)_{ij}$$

5.1.1 Univariate case

Consider the transformation $Y = r(X)$ where r is monotone increasing. Starting with the cdf of Y, we have that

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

Differentiating to find the pdf

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(r^{-1}(y)) \\ &= \frac{dF_X(r^{-1}(y))}{d(r^{-1}(y))}\frac{d(r^{-1}(y))}{dy} \\ &= f_X(r^{-1}(y))\frac{d}{dy}r^{-1}(y) \end{aligned}$$

The result is positive since r is increasing.

Example. Let (X_1, X_2) be a random vector with density $f_X(x_1, x_2) = 4x_1x_2, \ (x_1, x_2) \in (0, 1)^2$. Find the distribution of X_1/X_2.

Consider the transformation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = r\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1/x_2 \\ x_1 \end{pmatrix}$$

The inclusion of $y_2 = x_1$ is set so the transformation is one-to-one. Taking the inverse of the transformation, we have

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = r^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ y_2/y_1 \end{pmatrix}$$

Then the Jacobian matrix is

$$D_y r^{-1}(y) = \begin{pmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{pmatrix} = \begin{pmatrix} \partial y_2/\partial y_1 & \partial y_2/\partial y_2 \\ \partial(y_2/y_1)/\partial y_1 & \partial(y_2/y_1)/\partial y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -y_2/y_1^2 & 1/y_1 \end{pmatrix}$$

with determinant

$$J = detD_y r^{-1} = \frac{y_2}{y_1^2}$$

Solving for the joint distribution of Y, we have

$$f_Y(y) = f_X(r^{-1}(y))|J| = 4y_2\left(\frac{y_2}{y_1}\right)\left(\frac{y_2}{y_1^2}\right) = 4\left(\frac{y_2}{y_1}\right)^3$$

Finally, we can find the marginal distribution of Y_1 by integrating out the dependence on Y_2

$$f_{Y_1}(y_1) = \int_{-\infty}^{+\infty} f_Y(y_1, y_2)dy_2 = \int_{-\infty}^{+\infty} 4\left(\frac{y_2}{y_1}\right)^3 dy_2$$

5.2 Linear Transformation

Let $X = (X_1, ..., X_n)$ be a random vector with joint distribution $f_X(x)$. Consider the transformation $Y = AX$ where A is a square non-singular matrix, i.e. $detA \neq 0$ and A is invertible. Then we have that the joint distribution of Y is

$$f_Y(y) = \frac{1}{|detA|} f_X(A^{-1}y)$$

Proof. $f_Y(y) = f_X(r^{-1}(y))|J| = f_X(A^{-1}y)|detA^{-1}| = f_X(A^{-1}y)/|detA|$

5.2.1 Invariance of the Standard Normal

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ with $X = (X_1, ..., X_n)$. If A is an orthonormal matrix, then AX is also a vector of iid $\mathcal{N}(0,1)$ random variables.

Proof. If A is orthonormal, we have by definition that $AA^T = A^TA = I$, and $A = A^{-1}$. That is, $a_i^T a_j = 0, \ \forall i \neq j$ and $a_i^T a_i = 1$. Then we have that

$$||Ax||^2 = (Ax)^T(Ax) = \left(\sum_i x_i a_i\right)^T \left(\sum_j x_j a_j\right) = \sum_{ij} x_i x_j (a_i^T a_j) = \sum_i x_i^2 (a_i^T a_i) = ||x||^2$$

Also, we apply the properties that for square matrices, $det(AB) = det(A)det(B)$ and $det(A) = det(A^T)$ to show that $detA = 1$

$$det(I) = det(AA^T) = det(A)det(A^T) = (det(A))^2 = 1$$

The joint distribution of X (multivariate standard normal) is

$$f_X(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{||x||^2}{2}\right)$$

Considering the transformation $Y = AX$, the joint distribution of Y is

$$f_Y(y) = \frac{1}{|detA|} f_X(A^{-1}y) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{||A^{-1}y||^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left(-\frac{||y||^2}{2}\right)$$

which is a standard normal distribution.

6 Sampling Distributions

6.1 Chi-squared Distribution

A chi-squared distribution with m degrees of freedom is a $Gamma(m/2, 1/2)$ distribution by definition. Recall that a $Gamma(\alpha, \beta)$ has pdf

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

and moment generating function (MGF)

$$M(t) = \mathbb{E}e^{tX} = \left(\frac{\beta}{\beta - t}\right)^\alpha$$

The square of a standard normal random variable follows the chi-squared distribution with one degrees of freedom. That is, if $X \sim \mathcal{N}(0,1)$ then $X^2 \sim \chi_1^2$.

Proof.

$$F_{X^2}(x) = \mathbb{P}(X^2 \leq x) = \mathbb{P}(-\sqrt{x} \leq X \leq +\sqrt{x}) = \mathbb{P}(X \leq \sqrt{x}) - \mathbb{P}(X \leq -\sqrt{x}) = F_X(\sqrt{x}) - F_X(-\sqrt{x})$$

Differentiating to find the pdf

$$f_{X^2}(x) = f_X(\sqrt{x})\frac{1}{2\sqrt{x}} + f_X(-\sqrt{x})\frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{x}}f_X(\sqrt{x}) = \frac{1}{\sqrt{2\pi}}x^{1/2-1}e^{-x/2} \sim Gamma(1/2, 1/2)$$

Further, the sum of m independent squared standard normal random variables follows the chi-squared distribution with m degrees of freedom. That is, if $X_1, ..., X_m \overset{iid}{\sim} \mathcal{N}(0,1)$ then $X_1^2 + X_2^2 + ... + X_m^2 \sim \chi^2_m$.

Proof.
We will show that the MGF's for χ^2_m and $X_1^2 + ... + X_m^2$ are the same. Starting with χ^2_m, we have that

$$M_{\chi^2_m}(t) = \left(\frac{1/2}{1/2 - t}\right)^{m/2} = \left(\frac{1}{1-2t}\right)^{m/2}$$

Since $X_1, ..., X_m$ are independent, we have that the MGF of $X_1^2 + ... + X_m^2$ is the product of m MGF's of X_1^2

$$\left(M_{X_1^2}(t)\right)^m = \left(\left(\frac{1/2}{1/2 - t}\right)\right)^m = \left(\frac{1}{1-2t}\right)^{m/2}$$

6.1.1 Sample Mean and Variance

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then the MLE's follow

$$\hat{\mu} = \bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \sim (\sigma^2/n)\chi^2_{n-1}$$

We also have that $\hat{\mu}$ and $\hat{\sigma}^2$ are independent.

Sketch of proof. Assume that $X_1, ..., X_n \sim \mathcal{N}(0,1)$ with $X = (X_1, ..., X_n)$. We can construct an orthonormal transformation $Y = AX$ such that $Y_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \sim \mathcal{N}(0,1)$, i.e. all the elements in the first row of A is $1/\sqrt{n}$. Due to the invariance property of the standard normal, Y is also a vector of standard normal random variables.

Recall that $||Y|| = ||AX|| = ||X||$. Then we have that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2 = \sum_{i=1}^n X_i^2 - Y_1^2$$
$$= ||X||^2 - Y_1^2 = ||Y||^2 - Y_1^2 = \sum_{i=2}^n Y_i^2$$

Thus, the sample variance is a scaled sum of $n-1$ chi-squared random variables, so the sample variance is a scaled χ^2_{n-1} random variable. We also observe that it is independent of Y_1, which is a scaled sample mean. That is, the sample mean and sample variances are independent.

6.2 t-Distribution

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then $Z = \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim \mathcal{N}(0,1)$. If σ^2 is known, we can directly use the standard normal distribution to construct a confidence interval for μ. If σ^2 is unknown, we can estimate it by substituting the sample variance $\hat{\sigma}^2$. Replacing σ^2 with $\hat{\sigma}^2$ will add more variability Z, since $\hat{\sigma}^2$ itself is a random variable.

In such a case, we can use the t distribution. A t distribution with m degrees of freedom has density

$$f_{t_m}(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma(m/2)\sqrt{m\pi}}\left(1+\frac{x^2}{m}\right)^{-(m+1)/2}$$

Notably, X follows a t_m distribution is

$$X = \frac{Z}{\sqrt{Y/m}}$$

where $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_m$ and Z, Y are independent.

Proof. Let $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_m$ and Z, Y are independent. Then we have joint distribution

$$f_{ZY}(z,y) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}$$

Consider the transformation

$$\begin{pmatrix} x \\ w \end{pmatrix} = r\begin{pmatrix} z \\ y \end{pmatrix} = \begin{pmatrix} z/\sqrt{y/m} \\ y \end{pmatrix}$$

and its inverse

$$\begin{pmatrix} z \\ y \end{pmatrix} = r^{-1}\begin{pmatrix} x \\ w \end{pmatrix} = \begin{pmatrix} x\sqrt{w/m} \\ w \end{pmatrix}$$

Then the determinant of the Jacobian is

$$det\begin{pmatrix} \partial\left(x\sqrt{w/m}\right)/\partial x & \partial\left(x\sqrt{w/m}\right)/\partial w \\ \partial w/\partial x & \partial w/\partial w \end{pmatrix} = det\begin{pmatrix} \sqrt{w/m} & \partial\left(x\sqrt{w/m}\right)/\partial w \\ 0 & 1 \end{pmatrix} = \sqrt{w/m}$$

Solving for the joint distribution of X and W we have

$$\begin{aligned} f_{XW}(x,w) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2 w}{2m}}\frac{(1/2)^{m/2}}{\Gamma(m/2)}w^{m/2-1}e^{-w/2}\sqrt{w/m} \\ &= \frac{2^{-m/2}}{\Gamma(m/2)\sqrt{2\pi m}}w^{\frac{m+1}{2}-1}e^{-w\left(\frac{x^2}{2m}+\frac{1}{2}\right)} \end{aligned}$$

Marginalizing for X

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XW}(x,w)dw \\ &= \frac{2^{-m/2}}{\Gamma(m/2)\sqrt{2\pi m}}\frac{\Gamma\left(\frac{m+1}{2}\right)}{2^{-\frac{m+1}{2}}\left(\frac{x^2}{m}+1\right)\left(\frac{m+1}{2}\right)} \\ &= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma(m/2)\sqrt{m\pi}}\left(1+\frac{x^2}{m}\right)^{-(m+1)/2} \end{aligned}$$

So $X = Z/\sqrt{Y/m}$ follows a t_m distribution.

Practically, we can use

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\hat{\sigma}} \sim \frac{\sigma\mathcal{N}(0,1)}{\sqrt{\sigma^2\chi^2_{n-1}/(n-1)}} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} = t_{n-1}$$

to build confidence intervals for the sample mean under the normal distribution with unknown variance. Note that the sample variance $\hat{\sigma}^2$ here is the unbiased version so that the chi-squared random variable is normalized by $n-1$ (per the definition of the t-distribution).

6.2.1 Properties of the t-distribution

The t-distribution is symmetric, with heavier tails compared to the standard normal distribution. The mean exists and is equal to 0. The variance also exists and is equal to $m/(m-2)$. Finally, the kth moment exists so long as $k < m$; the moment generating function (MGF) does not exist.

6.3 F-Distribution

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y_1, ..., Y_m \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y^2)$. Suppose we are interested in comparing the variances of the two populations, σ_x^2/σ_y^2. This information is captured in the ratio of sample variances, $\hat{\sigma}_x^2/\hat{\sigma}_y^2$. The sample variances both follow χ^2 distributions. Assuming the two sample variances are independent, then their ratio can be characterized by the F distribution. Specifically, the distribution of the ratio of two independent χ^2 random variables with degrees of freedom p, q respectively follows an $F_{p,q}$ distribution with density

$$f(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{x^{p/2}-1}{(1+(p/q)x)^{(p/q)/2}},\ x > 0$$

For example, in regression analysis we are interested in finding if the regression sum of squares (SSR) is large compared to the error sum of squares (SSE). Both are χ^2 distributed and are independent, so the regression F test is derived normalizing them by their degrees of freedom (i.e. we compare the ratio of mean squares instead of the sum of squares).

7 Confidence Intervals

In frequentist statistics, the confidence interval is constructed so that it contains the the true parameter most of the time (the proportion that intervals contain the true parameter is called the confidence level, $\gamma = 1-\alpha$). Confidence intervals are random, that is, a function of the random data. The probability that the interval (as random variables) contain the true parameter is the confidence level; this is not to be confused with the probability that the parameter is in between a fixed interval, since how can a fixed parameter be random? Rather, out of all intervals constructed, γ of them will contain the true parameter.

7.1 Intervals for Normal Mean

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. Using $\bar{X}$ to estimate the mean μ, recall that $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) \sim \mathcal{N}(0,1)$. Then we can construct a $1-\alpha$ level confidence interval by solving for a, b in the following

$$\mathbb{P}\left(a < \frac{\sqrt{n}}{\sigma}(\bar{X}-\mu) < b\right) = 1-\alpha$$

based on the standard normal distribution.

If σ^2 is unknown, we know that $\frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}-\mu) \sim t_{n-1}$, so we can solve for the confidence interval

$$\mathbb{P}\left(a < \frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}-\mu) < b\right) = 1-\alpha$$

based on a t_{n-1} distribution.

Note that the distribution of the statistics $\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu)$ or $\frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}-\mu)$ do not depend on the unknown parameters. These statistics are called pivots.

7.2 Finding Pivots

Let $X = (X_1, ..., X_n) \sim f(x|\theta)$ be a random sample. A statistic $Q = Q(X, \theta)$ is a pivot if its distribution does not depend on θ.

7.2.1 Location/Scale Family

Let $f(x)$ be any pdf. The family of pdfs $f(x - \mu)$ indexed by location parameter μ is called the location family with standard pdf f. Similarly, the family of pdfs $\sigma^{-1} f(\sigma^{-1} x)$ indexed by scale parameter σ is called the scale family with standard pdf f.

Location: If $X_1, ..., X_n \overset{iid}{\sim} f(x-\mu)$, then $\bar{X} - \mu$ is a pivot. That is, then $Y_1 = X_1 - \mu, ..., Y_n = X_n - \mu \overset{iid}{\sim} f(x)$. We have that

$$\bar{X} - \mu = \frac{1}{n}\left(\sum_{i=1}^{n}(Y_i + \mu) - \mu\right) = \bar{Y}$$

Since Y_i follows the standard pdf $f(x)$, the distribution of $\bar{Y}$ does not depend on μ and is a pivot.

Scale: If $X_1, ..., X_n \overset{iid}{\sim} \sigma^{-1} f(\sigma^{-1} x)$ then $\sigma^{-1}\bar{X}$ is a pivot. That is, $Y_1 = \sigma^{-1} X_1, ..., Y_n = \sigma^{-1} X_n \overset{iid}{\sim} f(x)$. We have that

$$\sigma^{-1}\bar{X} = \frac{1}{n\sigma}\sum_{i=1}^{n}\sigma Y_i = \bar{Y}$$

whose distribution does not depend on σ.

Location-scale: If $X_1, ..., X_n \overset{iid}{\sim} \sigma^{-1} f(\sigma^{-1}(x - \mu)$ then $(\bar{X} - \mu)/\hat{\sigma}$ is a pivot, where $\hat{\sigma}$ is the sample standard deviation.

Example. Normal family.
Location: $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ so $X_1 - \mu, ..., X_n - \mu \overset{iid}{\sim} \mathcal{N}(0, 1)$
Scale: $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ so $\sigma^{-1} X_1, ..., \sigma^{-1} X_n \overset{iid}{\sim} \mathcal{N}(0, 1)$
Scale-location: $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ so $(X_1 - \mu)/\hat{\sigma}, ..., (X_n - \mu)/\hat{\sigma} \overset{iid}{\sim} t$

The strategy is to start with the sufficient statistic, modify the sufficient statistic so the distribution does not depend on the unknown parameters, and finally, find the exact distribution of the modified statistic.

Example. Exponential distribution.
Let $X_1, ..., X_n \overset{iid}{\sim} \mathit{Exponential}(\theta)$ with density $f(x|\theta) = \theta e^{-\theta x}, x \geq 0$. Note that the Exponential distribution is $\mathit{Gamma}(1, \theta)$. This is an example of the scale family, i.e. the distribution of θX does not depend on θ. Let $g(x) = \theta x$ with inverse $g^{-1}(y) = \theta^{-1} y$. Then the pdf of θX is

$$f_{\theta X}(x) = \theta e^{-\theta(\theta^{-1}x)}\frac{d}{dx}(\theta^{-1}x) = \frac{\theta}{\theta}e^{-x} = e^{-x},\ x \geq 0$$

which does not depend on the unknown parameter θ, i.e. $\theta X \sim \mathit{Gamma}(1, 1)$.

We start with the sufficient statistic, which for the Exponential distribution is

$$T = \sum_i X_i = X_1 + ... + X_n$$

If $X_1, ..., X_n \overset{iid}{\sim} \mathit{Gamma}(1, \theta)$, then $X_1 + ... + X_n \sim \mathit{Gamma}(n, \theta)$. We can show this using the moment generating function (MGF). Recall that the Gamma MGF is $M_{X_1}(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$ so we have

$$M_{X_1 + ... + X_n}(t) = (M_{X_1}(t))^n = \left(\frac{\beta}{\beta - t}\right)^{\alpha n}$$

Setting $\alpha = 1$ and $\beta = \theta$ we have our desired result. Using this result, we have that the statistic

$$\theta T = \theta(X_1 + ... + X_n) \sim Gamma(n, 1)$$

and is a pivot. That is, we can construct confidence intervals by solving

$$\mathbb{P}(a < \theta T < b) = 1 - \alpha$$
$$\mathbb{P}\left(\frac{a}{T} < \theta < \frac{b}{T}\right) = 1 - \alpha$$

Since the sampling distribution of θT does not depend on θ, we are able to set a, b according to the confidence level.

7.2.2 Finding Pivot by Density

Let $T \sim f(t|\theta)$ be a statistic (usually we start with the sufficeint statistic) with density of the form

$$f(t|\theta) = g(Q(t,\theta)) \left|\frac{\partial}{\partial t} Q(t,\theta)\right|$$

where $Q(t, \theta)$ is a monotone function of t for all θ. Then, $Q(T, \theta)$ is a pivot. The proof follows from the multivariate transformation formula.

7.2.3 Finding Pivot by CDF

If X is a continuous random variable with cdf $F_X(x)$ then the random variable $F_X(X) \sim Uniform(0, 1)$

Proof. Let $Y = F_X(X)$. Then

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

which is the $Uniform(0, 1)$ cdf.

Consequently, if $T = T(X_1, ..., X_n)$ is a (sufficient) statistic with cdf $F_T(t|\theta)$ then $F_T(T|\theta) \sim Uniform(0, 1)$ is a pivot. That is,

$$\mathbb{P}(\alpha_1 \leq F_T(T|\theta) \leq 1 - \alpha_2) = 1 - (\alpha_1 + \alpha_2)$$

yields an $\alpha = \alpha_1 + \alpha_2$ level confidence interval.

Example. Location-Exponential Interval. Let $X_1, ..., X_n \overset{iid}{\sim} f(x|\mu) = e^{-(x-\mu)} 1\{x \in [\mu, \infty)\}(x)$. Construct a confidence interval for μ.

We begin by finding a sufficient statistic. Writing out the joint distribution

$$\begin{aligned} f(x_1, ..., x_n) &= e^{-\sum_i (x_i - \mu)} \prod_{i=1}^{n} 1\{x \in [\mu, \infty)\}(x_i) \\ &= e^{-\sum_i x_i} e^{n\mu} 1\{x \in [\mu, \infty)\} \left(\underset{1 \leq i \leq n}{min}\ x_i\right) \end{aligned}$$

By the Factorization Theorem, we see that $T = \underset{1 \leq i \leq n}{min}\ X_i$ is a sufficient statistic.

Next, we find the cdf of $T = min_i X_i$. Applying order statistics, we have

$$\begin{aligned} 1 - F_T(x) = 1 - \mathbb{P}(minX_i \leq x) = \mathbb{P}(minX_i > x) &= \prod_{i=1}^{n} \mathbb{P}(X_i > x) \\ &= \left(\int_x^\infty e^{-(t-\mu)} dt \right)^n, \ x > \mu \\ &= \left(e^{-(x-\mu)} \right)^n, \ x > \mu \\ &= e^{-n(x-\mu)}, \ x > \mu \end{aligned}$$

Rearranging, we have the cdf of T

$$F_T(x) = \left(e^{-n(x-\mu)} \right) 1\{x \in [\mu, \infty)\}(x)$$

Then $F_T(T) \sim Uniform(0,1)$

$$\begin{aligned} F_T(T) &= \left(e^{-n(T-\mu)} \right) 1\{x \in [\mu, \infty)\}(T) \\ &= \left(e^{-n(T-\mu)} \right) 1\{x \in [0, T)\}(\mu) \end{aligned}$$

That is, we can solve

$$\mathbb{P}\left(\alpha_1 < \left(e^{-n(T-\mu)} \right) 1\{x \in [0,T)\}(\mu) < 1 - \alpha_2 \right)$$

to obtain a confidence interval for μ. First considering the upper bound, and requiring that $\mu \leq T$

$$\begin{aligned} \alpha_1 &\leq 1 - e^{-n(T-\mu)} \\ e^{-n(T-\mu)} &\leq 1 - \alpha_1 \\ -n(T-\mu) &\leq log(1-\alpha_1) \\ T - \mu &\geq -\frac{log(1-\alpha_1)}{n} \\ \mu &\leq T + \frac{log(1-\alpha_1)}{n} \end{aligned}$$

Since this interval is tighter than $\mu \leq T$ itself, we select it for the upper bound of the confidence interval. Using a similar calculation, we have for the lower bound

$$\left(e^{-n(T-\mu)} \right) 1\{x \in [0,T)\}(\mu) < 1 - \alpha_2$$

While the case where $\mu > T$ satisfies the inequality since $0 \leq 1 - \alpha_2$, it is obsolete from the upper-bound. Considering the case where $\mu \leq T$ we have

$$1 - e^{-n(T-\mu)} \leq 1 - \alpha_2$$

which yields

$$\mu \geq T + \frac{log(\alpha_2)}{n}$$

In all, we have the exact $\alpha = \alpha_1 + \alpha_2$ level confidence interval for μ as

$$\left(\frac{log(\alpha_2)}{n}, \frac{log(1-\alpha_1)}{n} \right)$$

7.3 Choosing α_1, α_2

Let $T = T(X, \theta) \sim f(x)$ be a pivot with known density f. Then

$$\mathbb{P}(a \leq T \leq b) = 1 - \alpha = 1 - (\alpha_1 + \alpha_2)$$

yields a confidence set for θ. Assuming the confidence set is an increasing function of $b - a$, that is, smaller $b - a$ correspond to smaller confidence intervals for θ, and assuming a unimodal pdf, that is, there exist x^* such that $f(x)$ is non-decreasing on $(-\infty, x^*)$ and non-increasing on (x^*, ∞). Then the optimal choice for a, b satisfies

$$a \leq x^* \leq b \ s.t. \ f(a) = f(b)$$

where

$$\int_a^b f(x) dx = 1 - \alpha$$

7.4 Nonparametric Bootstrap

So far, we have relied on making parametric assumptions to find pivots and their distribution, and to ultimately construct confidence intervals. With the bootstrap, we take a completely different approach and make no assumptions on the underlying distribution.

The bootstrap is a simulation-based method of understanding the properties of a statistical estimate, usually consisting of sampling with replacement. In the nonparametric bootstrap, the underlying idea is to approximate the true population data distribution by the empirical distribution that places mass $1/n$ for each observed data value $X_1, ..., X_n$. This results in the empirical cdf as follows

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$$

or the fraction of data points less than x. This approximates the true cdf $F(x)$.

In the above, let $Y_i = 1\{X_i \leq x\} \sim Bernoulli(\mathbb{P}(X_i \leq x))$. Then we have that

$$\begin{aligned} \mathbb{E}Y_i &= \mathbb{P}(X_i \leq x) = F(x) \\ Var(Y_i) &= \mathbb{P}(X_i \leq x)(1 - \mathbb{P}(X_i \leq x)) = F(x)(1 - F(x)) \end{aligned}$$

As a consequence

$$\begin{aligned} \mathbb{E}F_n(x) &= F(x) \\ Var(F_n(x)) &= \frac{Var(Y_i)}{n^2} = \frac{F(x)(1 - F(x))}{n} \end{aligned}$$

This tells us that the empirical distribution $F_n(x)$ is an *unbiased* estimator for the true $F(x)$, and the law of large numbers gives us $F_n(x) \xrightarrow{p} F(x)$. We can actually show almost-sure uniform convergence of F_n to F (see Glivenko-Cantelli theorem) such that

$$||F_n - F||_\infty = \underset{x \in \mathbb{R}}{sup} |F_n(x) - F(x)| \xrightarrow{a.s.} 0$$

In all, $F_n(x)$ is a very reasonable estimate of $F(x)$, especially for large n.

Next, we can draw iid samples denotes $X_1^*, ..., X_n^*$ from the empirical distribution; this is equivalent to sampling with replacement n values from the original data sample $X_1, ..., X_n$. Doing this many times (denote B simulations) helps us understand the sampling distribution of a statistic of interest. For example, if we were interested in the standard error of a statistic $T = T(X_1, ..., X_n)$ then we would run the following algorithm

```
for $i = 1, ..., B$ do
    • Simulate $X_1^*, ..., X_n^*$ from the empirical distribution, drawing $n$ samples with replacement from the original data $X_1, ..., X_n$
    • Compute the statistic $T^* = T(X_1^*, ..., X_n^*)$ on the bootstrap data sample
end for
return The empirical standard deviation of $T^*$ across the $B$ simulations
```

We have essentially simulated the sampling distribution of the statistic T using the empirical distribution using a large number (B) of bootstrap samples. At a high level, this works because F_n is a good approximation of F, and because calculating the statistic on resampled data captures its variability. Naturally, we can use the same idea to build confidence intervals. Recall that in Frequentist statistics, the idea of behind a $1-\alpha$ confidence interval is if we repeatedly take samples from the population distribution and calculate similar confidence intervals, $1-\alpha$ of the intervals will contain the truth. We can mimic such a procedure using bootstrapping by repeatedly taking samples from the empirical distribution.

In the nonparametric setting, we can define a parameter, or rather a population quantity of interest, as some function of the cdf F (formally called statistical functionals), denoted as $\theta = T(F)$. For example, the mean of $X \sim F$ can be thought of as $\mathbb{E}X = T_{mean}(F) = \int x dF(x) = \int xp(x)dx$ if X has a pdf (in which case, $dF(x) = p(x)dx$). If X has a pmf, then we can define $\int xdF(x) = \sum_x xp(x)$. Similarly, the median of X (or similarly any quantile) can be thought of as $T_{median}(F) = F^{-1}(0.5)$.

Bootstrap estimators estimate the population quantity of interest by using *plug-in* estimators. Specifically, we estimate $\theta = T(F)$ using an empirical distribution $\hat{\theta} = T(F_n)$ over $X_1, ..., X_n$. For example, we have that $T_{mean}(F_n) = \sum_{i=1}^n X_i p_n(x) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$ which is the sample mean. Similarly, we have that $T_{median}(F_n) = F_n^{-1}(0.5)$ which is the sample median.

It is helpful to see $\hat{\theta} = T(F_n)$ as the "truth" for the empirical distribution F_n. Using bootstrap samples, i.e. repeated samples from F_n, we find the sampling distribution of the plug-in estimators and create a confidence interval for $\hat{\theta}$. Roughly speaking, the bootstrap confidence interval is valid because if we create a valid confidence interval for the true value on F_n, $\hat{\theta} = T(F_n)$, we also create a valid interval for the true value on F, $\theta = T(F)$, asymptotically since $F_n \to F$. We are using the behavior of $\hat{\theta}^*$ around $\hat{\theta}$ (bootstrap world, distribution F_n) to learn about the behavior of $\hat{\theta}$ around θ (real world, distribution F).

7.4.1 Percentile Bootstrap Interval

Suppose we are interested in constructing a confidence interval for θ, using some estimate $\hat{\theta}$. Let $\hat{\theta}^*(\alpha/2)$ and $\hat{\theta}^*(1-\alpha/2)$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated values of $\hat{\theta}^*$. The most straightforward way to construct the confidence interval would be to simply take

$$\left(\hat{\theta}^*(\alpha/2), \hat{\theta}^*(1-\alpha/2)\right)$$

which represents a $1-\alpha$ coverage probability of the bootstrap sampling distribution. Again, the rationale behind the percentile bootstrap interval follows directly from the reasoning that an interval for the true parameter for F_n, $\hat{\theta}$, will converge to an interval for the true parameter for F, θ, since F_n converges to F asymptotically. However, this approach does not directly address differences between θ for F and $\hat{\theta}$ for F_n.

7.4.2 Basic Bootstrap Interval

Instead, we can argue that deviations between the truth θ and samples $\hat{\theta}$ in the real world (distributed using F) should be a good approximation for the deviations between the truth $\hat{\theta}$ and samples $\hat{\theta}^*$ in the bootstrap world (distributed using F_n). That is, the behavior of $\theta - \hat{\theta}$ is approximately the same as the behavior of $\hat{\theta} - \hat{\theta}^*$. Then we can construct a "pivot" as follows

$$\begin{aligned}
1 - \alpha &\approx \mathbb{P}(\hat{\theta}^*(\alpha/2) < \hat{\theta}^* < \hat{\theta}^*(1 - \alpha/2)) \\
&= \mathbb{P}(\hat{\theta} - \hat{\theta}^*(\alpha/2) > \hat{\theta} - \hat{\theta}^* > \hat{\theta} - \hat{\theta}^*(1 - \alpha/2)) \\
&\approx \mathbb{P}(\hat{\theta} - \hat{\theta}^*(\alpha/2) > \theta - \hat{\theta} > \hat{\theta} - \hat{\theta}^*(1 - \alpha/2)) \\
&= \mathbb{P}(2\hat{\theta} - \hat{\theta}^*(\alpha/2) > \theta > 2\hat{\theta} - \hat{\theta}^*(1 - \alpha/2))
\end{aligned}$$

So a $1 - \alpha$ confidence interval can be taken as

$$\left(2\hat{\theta} - \hat{\theta}^*(1 - \alpha/2), 2\hat{\theta} - \hat{\theta}^*(\alpha/2)\right)$$

In practice, basic and percentile intervals can be very similar, especially when the distribution of $\hat{\theta}^*$ around $\hat{\theta}$ is symmetric.

The main advantage of the bootstrap is that it enables statistical inference without making (sometimes unreasonable) assumptions on the distribution. Before the bootstrap, many estimators had to be seen in the form of sample averages so that the Central Limit Theorem can be applied to build approximate confidence intervals around them. It can be shown, for sufficiently large n, that the bootstrap is never worse than a CLT approximation.

8 Hypothesis Testing

The goal of hypothesis testing is to determine whether or not the data sufficiently supports a certain hypothesis. Usually, this is stated as the comparison between a *null hypothesis*, denoted as H_0, and an *alternative hypothesis*, denoted as H_1. This is done by computing a *test statistic* against the null hypothesis; that is, if the data is unlikely to appear under the null hypothesis, we reject the null and accept the alternative.

There is a risk of making the wrong decision. We denote *Type I error* as rejecting H_0 when it is true (false positive). This probability is called the significance level of the test, denoted by α. α is usually fixed to determine the *decision rule* of when to reject H_0, for example, $\alpha = 0.05$ is a common threshold. Similarly, *Type II error* is rejecting H_1 when it is true (false negative). This probability is denoted as β. The *power* of the test, $1 - \beta$, is the probability of correctly accepting H_1. The ideal test has small α and large power.

8.1 Likelihood Ratio Test

The likelihood ratio test is a class of hypothesis tests that compares the likelihood functions under the null hypothesis and under the alternative hypothesis as a ratio. If this ratio is small, then there is evidence that the alternative is more likely so we reject the null hypothesis.

Take for example a normal distribution that has unknown mean that is either μ_0 (null hypothesis) or μ_1 (alternative hypothesis), and known variance σ^2. Suppose we have that $\mu_0 \leq \mu_1$. Let $X_1, ..., X_n$ be iid random variables representing the observed data. We want to test using likelihood ratio

$$\begin{aligned}
H_0 &: \mu = \mu_0 \\
H_1 &: \mu = \mu_1
\end{aligned}$$

Under the null hypothesis we have likelihood

$$f_0(x_1, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right)$$

Similarly, under the alternative hypothesis we have likelihood

$$f_1(x_1, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right)$$

Then the ratio between null and alternative likelihoods is

$$\begin{aligned}\Lambda = \frac{f_0(x_1, ..., x_n)}{f_1(x_1, ..., x_n)} &= \prod_{i=1}^{n} exp\left(\frac{1}{2\sigma^2}\left((x_i - \mu_1)^2 - (x_i - \mu_0)^2\right)\right) \\ &= exp\left(\frac{1}{\sigma^2}(\mu_0 - \mu_1)\sum_{i=1}^{n}\left(x_i - \frac{\mu_0 + \mu_1}{2}\right)\right) \\ &= exp\left(\frac{n}{\sigma^2}(\mu_0 - \mu_1)\left(\bar{x} - \frac{\mu_0 + \mu_1}{2}\right)\right)\end{aligned}$$

Suppose that we compare Λ to a constant, that is, we reject H_0 if Λ is sufficiently small compared to a threshold $\Lambda < c$. Moving around constants, we have

$$\Lambda < c \iff \bar{x} - \frac{\mu_0 + \mu_1}{2} > c' \iff \bar{x} > c''$$

We have the result that under the likelihood ratio test, the decision rule is to reject the null hypothesis if the test statistic $\bar{X}$ is sufficiently *larger* than a constant. The only remaining question is to determine the size of such a constant. We determine this by controlling for the Type I error (the probability that $\bar{X}$ is greater than c'' (reject null) even if it is true $\bar{X}$ follows the null distribution). In other words, while there is α chance for Type I error for when the null is true, observed values this extreme are usually taken as evidence against the null hypothesis

$$\alpha = \mathbb{P}(reject\ H_0|H_0\ true) = \mathbb{P}(\bar{X} > c''|X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu_0, \sigma^2)) = \mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c'' - \mu_0}{\sigma/\sqrt{n}}|H_0\right)$$

Under the null hypothesis, the left hand side is a standard normal random variable, so that the constant on the right hand side should be equal to $z(1 - \alpha)$, the $1 - \alpha$ quantile of the standard normal distribution, if we are to control for α-level Type I error. Rearranging, we have that the final rejection region is if

$$\bar{x} > \frac{\sigma}{\sqrt{n}} z(1 - \alpha) + \mu_0$$

where the right hand side c'' is called the *critical value* and $\bar{x}$ denotes the observed data.

Another equivalent way to decide whether to reject H_0 is to see if the *p-value* is less than α. The *p-value* is the probability that the test statistic $\bar{X}$ as a random variable under the null distribution, is more extreme than the observed data $\bar{x}$. In this example, the p-value is defined as

$$p = \mathbb{P}(\bar{X} > \bar{x}|H_0)$$

If $p = \alpha$, then the observed data $\bar{x}$ coincides with the critical value c''. If $p < \alpha$, the observed data is more extreme than the critical value, in this case, $\bar{x} > c''$. The two ways of deciding to reject H_0 are equivalent.

Note that only the null distribution was used to perform the test; we only accounted for how extreme the observed data is if the null were true while controlling for Type I error.

8.2 Power Calculation

Recall that power is the probability of rejecting H_0 when H_1 is true

$$\mathbb{P}(reject\ H_0|H_1\ true) = 1 - \beta$$

For a given α, we want power to be as high as possible. While there is no simple relationship between α and $1-\beta$, as α increases, $1-\beta$ also increases. For example, if we accept all alternatives ($\alpha = 1$) then $1-\beta$ is also 1.

Take the previous example where $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and we test $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1$ where $\mu_0 \leq \mu_1$. Recall that we reject H_0 if $\bar{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} z(1-\alpha)$. Then the power is

$$\begin{aligned} 1 - \beta &= \mathbb{P}(reject\ H_0 | H_1\ true) \\ &= \mathbb{P}\left(\bar{X} > \mu_0 + \frac{\sigma}{\sqrt{n}} z(1-\alpha) | \bar{X} \sim \mathcal{N}(\mu_1, \sigma^2/n)\right) \end{aligned}$$

To solve for this probability, we can rearrange the left hand side so that it follows a standard normal distribution under the alternative hypothesis. Then we have

$$\begin{aligned} 1 - \beta &= \mathbb{P}\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 + \frac{\sigma}{\sqrt{n}} z(1-\alpha) - \mu_1}{\sigma/\sqrt{n}} | H_1\right) \\ &= \mathbb{P}\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu_1) + z(1-\alpha) | H_1\right) \end{aligned}$$

Note that when $n \to \infty$, then $LHS \to -\infty$ so the power is 1. When $\mu_0 = \mu_1$, the power is the same as α. When $\mu_0 << \mu_1$, $\mu_0 - \mu_1$ is more negative and the power increases (that is, power is greater when the two hypotheses are more distinct). These observations are consistent with what we would expect.

8.2.1 Neyman-Pearson Lemma

So far, we have discussed *simple hypotheses* where each parameter can only take one value, and the probability distribution is completely specified by the hypothesis. In contrast, a *composite hypothesis* is when at least one parameter can take more than one value.

The Neyman-Pearson Lemma says that if H_0 and H_1 are simple hypothesis, the likelihood ratio test is the most powerful. Specifically, compared to a test that rejects H_0 whenever the likelihood ratio is less than c_α with significance level α and power $1 - \beta$, any other test for which the significance level is at most α has power at most $1 - \beta$.

8.2.2 Duality Between Confidence Intervals and Hypothesis Testing

If a $1 - \alpha$ level confidence interval for θ is (a, b), then we reject the null if $\theta \notin (a, b)$ and accept the null otherwise at the significance level α. Similarly, if we accept the null that $\theta = \theta_0$ for all $\theta_0 \in (a, b)$ and reject the null for all $\theta_0 \notin (a, b)$ at level α, then (a, b) is a $1 - \alpha$ confidence interval for θ.

8.3 Generalized Likelihood Ratio Test

The likelihood ratio test is optimal for testing simple hypotheses. However, when the test is not simple, we turn to generalized likelihood ratio tests. These are not generally optimal, but perform reasonably well. They play the same role as MLE's do in estimation.

Suppose that $X = (X_1, ..., X_n)$ has joint density $f(x|\theta)$ where $x = (x_1, ..., x_n)$ and consider the test

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \\ H_1 &: \theta \in \Omega_1 \end{aligned}$$

for some sets Ω_0, Ω_1. It is natural to compare the maximum likelihood under the two hypothesis, that is, $\underset{\theta\in\Omega_0}{max} f(x|\theta)$ and $\underset{\theta\in\Omega_1}{max} f(x|\theta)$. Let $\Omega = \Omega_0 \cup \Omega_1$. We can study the ratio

$$\Lambda = \frac{\underset{\theta\in\Omega_0}{max}\ f(x|\theta)}{\underset{\theta\in\Omega}{max}\ f(x|\theta)}$$

This ratio is convenient because if $\Omega = \mathbb{R}$, then the denominator is maximized at the MLE. As before, we reject the null if $\Lambda < c$ for some constant c such that $\mathbb{P}(\Lambda < c|H_0) = \alpha$ for an α level test.

Example. Normal mean. Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with unknown μ and known variance σ^2. Consider the test

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

We have joint distribution

$$f(x_1, ..., x_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

As for the likelihood ratio, Ω_0 has only one element, μ_0. Meanwhile, the denominator is maximized at the MLE $\bar{x}$. Then we have

$$\begin{aligned} \Lambda &= \frac{f(x_1, ..., x_n|\mu = \mu_0)}{f(x_1, ..., x_n|\mu = \bar{x})} \\ &= exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\mu_0)^2 - (x_i-\bar{x})^2\right)\right) \\ &= exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\bar{x}-\mu_0)(2x_i-\mu_0-\bar{x})\right) \\ &= exp\left(-\frac{1}{2\sigma^2}(\bar{x}-\mu_0)(2n\bar{x}-n\mu_0-n\bar{x})\right) \\ &= exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu_0)^2\right) \end{aligned}$$

We reject H_0 if $\Lambda < c \iff (\bar{x}-\mu_0)^2 > c' \iff |\bar{x}-\mu| > c''$. As before, we determine c'' by controlling Type I error, that is, we set

$$\begin{aligned} \alpha &= \mathbb{P}(reject\ H_0|H_0\ true) \\ &= \mathbb{P}(|\bar{X}-\mu| > c''|H_0) \\ &= \mathbb{P}\left(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} > \frac{c''}{\sigma/\sqrt{n}}|H_0\right) \end{aligned}$$

Under the null hypothesis, the left hand side is a standard normal random variable, so we have that

$$c'' = \frac{\sigma}{\sqrt{n}} z(1-\alpha/2)$$

where $z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution. Our result is consistent with what is commonly known as the two-tailed z-test.

8.3.1 Asymptotic Test

Under the null hypothesis H_0 and smoothness conditions on the probability distribution, we have that

$$-2log\Lambda \xrightarrow{d} \chi^2_d$$

where χ^2_d is a chi-squared random variable with d degrees of freedom; d is the difference between the number of free parameters in Ω and the number of free parameters in Ω_0.

We reject the null hypothesis when Λ is small, or equivalently, if $-2log\Lambda$ is large. Asymptotically, we can compare to the χ^2_d distribution, rejecting the null hypothesis when $-2log\Lambda > c$ where c is the $1-\alpha$ quantile of χ^2_d for an α-level test.

In the previous normal mean example, the result is exact. We have $d = 1$ since Ω had one free parameter, and Ω_0 had no free parameters since it only had one point. Then we have

$$-2log\Lambda = \frac{n(\bar{x} - \mu_0)^2}{\sigma^2} \sim \chi^2_1$$

Example. Mendel's Experiment.
Recall the multinomial distribution, which generalizes the binomial distribution for r different outcomes and n independent trials. For each trail, the probabilities of the r outcomes are $p_1, ..., p_r$, respectively. Let $N_1, ..., N_r$ be the total number of outcomes of each type. Then we have

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, ..., N_r = n_r) = \binom{n}{n_1, ..., n_r} p_1^{n_1} ... p_r^{n_r}$$

Recall the multinomial coefficient, representing the number of ways to split n items into r groups is defined as

$$\binom{n}{n_1, ..., n_r} = \frac{n!}{n_1!n_2!...n_r!}$$

Mendel's Experiments studied the different traits of pea plants to learn about genetics - we can model the frequency of each trait using a multinomial distribution. Take for example the data below that compares 4 different traits of peas and their occurrence frequency

Type	Smooth yellow	Smooth green	Wrinkled yellow	Wrinkled green
Frequency (theory)	9/16	3/16	3/16	1/16
Observed counts (n=556)	315	108	012	31

We want to test the null hypothesis that the scientific theory is true, against the alternative that the theory is false. In the alternative hypothesis, we have that the parameters are free except for the constraint that they are non-negative and sum to one. That is, we have

$$\Omega_0 = \{p : p_1 = 9/16, p_2 = 3/16, p_3 = 3/16, p_4 = 1/16\}$$
$$\Omega = \Omega_0 \cup \Omega_1 = \{p : p_1 \geq 0, p_2 \geq 0, p_3 \geq 0, p_4 \geq 0, \sum_i p_i = 1\}$$

with likelihood ratio

$$\Lambda = \frac{\binom{n}{n_1, ..., n_r} p_1^{n_1} p_2^{n_2} ... p_r^{n_r}}{\binom{n}{n_1, ..., n_r} \hat{p}_1^{n_1} \hat{p}_2^{n_2} ... \hat{p}_r^{n_r}}$$

where $\hat{p}_1, ..., \hat{p}_r$ are the solutions to $\underset{p \in \Omega}{argmax} \binom{n}{n_1, ..., n_r} p_1^{n_1} p_2^{n_2} ... p_r^{n_r}$, that is, the MLE for p.

Solving for the MLE for p, we can set $p_r = 1 - \sum_{i=1}^{r-1} p_i$ and let the other $p_1, ..., p_{r-1}$ vary freely so that the constraint $\sum_{i=1}^{r} p_i = 1$ is enforced. Then the log-likelihood is

$$\ell(p) = log\binom{n}{n_1, ..., n_r} + \sum_{i=1}^{r} n_i log(p_i) = log\binom{n}{n_1, ..., n_r} + \sum_{i=1}^{r-1} n_i log(p_i) + n_r log\left(1 - \sum_{i=1}^{r-1} p_i\right)$$

Setting the partial derivatives $\partial\ell/\partial p_i = 0$ we have that for all $i = 1, ...r - 1$

$$\frac{n_i}{p_i} - \frac{n_r}{1 - \sum_{j=1}^{r-1} p_j} = 0$$

Recall the fact that $\frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d}$. Then we have

$$\frac{n_1}{p_1} = \frac{n_2}{p_2} = ... = \frac{n_{r-1}}{p_{r-1}} = \frac{n_r}{1 - \sum_{j=1}^{r-1} p_j} = \frac{n}{1}$$

Rearranging, we have

$$\hat{p}_1 = \frac{n_1}{n}, \hat{p}_2 = \frac{n_2}{n}, ..., \hat{p}_{r-1} = \frac{n_{r-1}}{n}$$

and

$$\hat{p}_r = 1 - \sum_{j=1}^{r-1} \hat{p}_j = \frac{n_r}{n}$$

Under the null hypothesis, we have the asymptotic test

$$\begin{aligned} -2log\Lambda &= -2\sum_{i=1}^{r} n_i log\left(\frac{p_i}{\hat{p}_i}\right) \\ &= -2\left(315log\left(\frac{9/16}{315/556}\right) + ... + 31log\left(\frac{1/16}{31/556}\right)\right) \\ &\sim \chi^2_{(3)} \end{aligned}$$

We have 3 degrees of freedom since there are $4 - 1$ free parameters in Ω, and no free parameters in Ω_0 since it consists of a fixed value. We reject the null hypothesis if Λ is small, or equivalently if $-2log\Lambda$ is large which we can compare directly to the $\chi^2_{(3)}$ distribution controlling for α, the Type I error rate.

8.3.2 Power Function

Let δ represent a test procedure. The function $\pi(\theta|\delta)$ is called the *power function* of the test δ. Let T be the test statistic and R be the rejection region, the power function is defined as

$$\pi(\theta|\delta) = \mathbb{P}(T \in R|\theta), \ \theta \in \Omega$$

The power function specifies the probability of rejecting the null hypothesis H_0 for every possible parameter θ. That is, the ideal power function has value zero for all $\theta \in \Omega_0$ (no Type I error) and value one for all $\theta \in \Omega_1$ (power of 1; no Type II error). Test procedures with ideal power functions in practice are very rare.

8.4 Uniformly Most Powerful Test

Recall that when the null and alternative hypotheses are simple, the likelihood ratio test (LRT) is power optimal, i.e. among all test with significance level at most that of the LRT, the LRT has the largest power. For composite hypotheses, a modified version of this test is still optimal in the special case when

- $H_0 : \theta \leq \theta_0 \ vs. \ H_1 : \theta > \theta_0$ or $H_0 : \theta \geq \theta_0 \ vs. \ H_1 : \theta < \theta_0$
- The family of distributions has a monotone likelihood ratio in some statistic T

8.4.1 Monotone Likelihood Ratio (MLR)

Suppose $X = (X_1, ..., X_n) \sim f(x|\theta)$ and $T = T(X)$. If for every $\theta_1 < \theta_2$, the likelihood ratio $f(x|\theta_2)/f(x|\theta_1)$ is an monotone increasing (or decreasing) function of T, then we say the distribution of X has an increasing (or decreasing) likelihood ratio.

For example, suppose we want to test $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta = \theta_2 > \theta_1$. Then the LRT rejects H_0 when Λ is sufficiently small, or, $\Lambda^{-1} = f(x|\theta_2)/f(x|\theta_1)$ is sufficiently large. If Λ^{-1} is monotone increasing with respect to T, then rejecting the null when Λ^{-1} is sufficiently large is equivalent to rejecting the null when T is sufficiently large. That is, rejecting the null if $T > c$ is an LRT.

Example. Show that $Bernoulli(p)$ has a MLR.
For $p_2 > p_1$ we have

$$\begin{aligned}\frac{f(x|p = p_2)}{f(x|p = p_1)} &= \frac{\prod_{i=1}^n p_2^{x_i}(1 - p_2)^{1-x_i}}{\prod_{i=1}^n p_1^{x_i}(1 - p_1)^{1-x_i}} \\ &= \left(\frac{p_2}{p_1}\right)^{\sum_i x_i}\left(\frac{1 - p_2}{1 - p_1}\right)^{\sum_i(1-x_i)} \\ &= \left(\frac{p_2/(1 - p_2)}{p_1/(1 - p_1)}\right)^{\sum_i x_i}\left(\frac{1 - p_2}{1 - p_1}\right)^n\end{aligned}$$

Set $T = \sum_i X_i$. Since the function $g(x) = x/(1 - x)$ is increasing, i.e. $g'(x) = \frac{1}{(1-x)^2} > 0$, $\frac{p_2/(1-p_2)}{p_1/(1-p_1)} > 1$ since $p_2 > p_1$. This shows that the likelihood ratio is increasing with T.

Example. Show that $Exponential(\theta)$ has a MLR.
Recall that $f(x) = \theta e^{-\theta x}$ is the pdf for the exponential distribution. Then for $\theta_2 > \theta_1$, we have

$$\begin{aligned}\frac{f(x|\theta = \theta_2)}{f(x|\theta = \theta_1)} &= \frac{\theta_2^n e^{-\theta_2 \sum_i x_i}}{\theta_1^n e^{-\theta_1 \sum_i x_i}} \\ &\propto e^{(\theta_1-\theta_2)\sum_i x_i}\end{aligned}$$

Since $\theta_1 < \theta_2$, we have that the likelihood ratio is decreasing with the statistic $T = \sum_i X_i$.

Example. Show that $\mathcal{N}(\mu, 1)$ has a MLR (that is, normal distribution with unknown mean and known variance).
For $\mu_2 > \mu_1$

$$\begin{aligned}\frac{f(x|\mu = \mu_2)}{f(x|\mu = \mu_1)} &= \frac{exp\left(-\frac{1}{2}\sum_i(x_i - \mu_2)^2\right)}{exp\left(-\frac{1}{2}\sum_i(x_i - \mu_1)^2\right)} \\ &= exp\left(\frac{1}{2}\sum_i((x_i - \mu_1)^2 - (x_i - \mu_2)^2)\right) \\ &\propto exp\left((\mu_2 - \mu_1)\sum_i x_i\right)\end{aligned}$$

Since $\mu_2 > \mu_1$, the likelihood ratio is increasing with the statistic $T = \sum_i X_i$.

8.4.2 One-sided Alternatives and Karlin-Rubin Theorem

Consider the one-sided alternative

$$\begin{aligned}H_0 &: \theta \le \theta_0 \\ H_1 &: \theta > \theta_0\end{aligned}$$

Suppose that the joint of X has an *increasing* MLR in the statistic T. Let c, α be constants such that

$$\mathbb{P}(T > c | \theta = \theta_0) = \alpha$$

The *Karlin-Rubin Theorem* states that the test that rejects H_0 if $T > c$ is a uniform most powerful test (UMP) at the significance level α. That is, among all tests with level at most α, this test has the highest power for all $\theta > \theta_0$.

Similarly, consider the other one-sided alternative

$$\begin{aligned} H_0 &: \theta \geq \theta_0 \\ H_1 &: \theta < \theta_0 \end{aligned}$$

Suppose that the joint of X has an *increasing* MLR in the statistic T. Let c, α be constants such that

$$\mathbb{P}(T < c | \theta = \theta_0) = \alpha$$

Then the test that rejects H_0 if $T < c$ is a uniform most powerful test at significance level α. That is, among all tests with level at most α, this test has the highest power for all $\theta < \theta_0$.

8.4.3 Two-sided Alternatives

Consider the two-sided alternative

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

and suppose that the joint of X has an *increasing* MLR in the statistic T.

In the case that $\theta > \theta_0$, the power-optimal test is $T > c$. Similarly, in the case that $\theta < \theta_0$, the power-optimal test is $T < c$. Since these are in conflict, there is no overall UMP. However, it is natural to consider the test that rejects the null if $T < c_1$ *or* $T > c_2$ so that

$$\begin{aligned} \mathbb{P}(T < c_1 | \theta = \theta_0) &= \alpha_1 \\ \mathbb{P}(T > c_2 | \theta = \theta_0) &= \alpha_2 \\ \alpha_1 + \alpha_2 &= \alpha \end{aligned}$$

8.5 Special Tests

8.5.1 t-Test for Normal Mean

Consider the one-sided t-test for the mean of $\mathcal{N}(\mu, \sigma^2)$ with unknown σ^2. Since there are two parameters, there is no UMP.

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

Recall that if $\mu = \mu_0$, then the statistic

$$U = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}} \sim t_{n-1}$$

We can reject the null $U > c$ where c is the $1 - \alpha$ quantile for t_{n-1}. This is a level α test, that is, the probability for Type I error *is at most* α. When $\mu = \mu_0$, the probability for Type I error is exactly α by design. For the case when $\mu \in \Omega_0$ but $\mu \neq \mu_0$ (null hypothesis is still true), that is, $\mu < \mu_0$, the probability for Type I error is less than α. Writing it out, we have

$$\begin{aligned}\mathbb{P}(U > c|\mu = \tilde{\mu} < \mu_0) &= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X} - \tilde{\mu})}{\hat{\sigma}} + \frac{\sqrt{n}(\tilde{\mu} - \mu_0)}{\hat{\sigma}} > c|\mu = \tilde{\mu} < \mu_0\right) \\ &= \mathbb{P}\left(t_{n-1} > c + \frac{\sqrt{n}(\mu_0 - \tilde{\mu})}{\hat{\sigma}}\right) < \alpha\end{aligned}$$

which is true since $\mu_0 - \tilde{\mu} > 0$.

Following the same logic, the probability of Type I error goes to zero as $\mu \to -\infty$, and the power of the test goes to one as $\mu \to \infty$. Also note this t-test turns out to be a generalized likelihood ratio test comparing $\Omega_0 = \{(\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\}$ to $\Omega = \Omega_0 \cup \Omega_1$ where $\Omega_1 = \{(\mu, \sigma^2) : \mu > \mu_0, \sigma^2 > 0\}$.

The same result holds for testing the other one-sided hypothesis

$$\begin{aligned}H_0 &: \mu \geq \mu_0 \\ H_1 &: \mu < \mu_0\end{aligned}$$

but instead, we reject the null hypothesis when $U < c$ and c is the α quantile for t_{n-1}. In the case of the two-sided hypothesis,

$$\begin{aligned}H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0\end{aligned}$$

we reject the null when $|U| > c$ and c is the $1 - \alpha/2$ quantile for t_{n-1}. Each of these tests are generalized likelihood ratio tests

8.5.2 Two-sample t-Test

Suppose we want to compare the means of *two* normal populations. Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma^2)$ represent a sample of size n from the first population and $Y_1, ..., Y_m \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma^2)$ represent a sample of size m from the second population. Consider the test

$$\begin{aligned}H_0 &: \mu_x \leq \mu_y \\ H_1 &: \mu_x > \mu_y\end{aligned}$$

Denote

$$\begin{aligned}S_X^2 &= \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ S_Y^2 &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2\end{aligned}$$

We use the test statistic

$$U = \frac{(n + m - 2)^{1/2}(\bar{X} - \bar{Y})}{\left(\frac{1}{n} + \frac{1}{m}\right)^{1/2} (S_X^2 + S_Y^2)^{1/2}}$$

We can reject the null if $U > c$ where c is the $1 - \alpha$ quantile for t_{n+m-2}. This is a level α generalized likelihood ratio test. Note that the probability of Type I error goes to zero as $\mu_x - \mu_y \to -\infty$ and the power goes to 1 as $\mu_x - \mu_y \to \infty$.

Similar results hold for the other one-sided alternative (i.e. reject when $U < c$) or two-sided alternative (reject when $|U| > c$).

8.5.3 F-test

While the two-sample t-test compares the means of two normal populations, it may also be of interest to compare their variances. Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$ represent a sample of size n from the first population and $Y_1, ..., Y_m \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y^2)$ represent a sample of size m from the second population. Consider the test

$$H_0 : \sigma_x^2 \leq \sigma_y^2$$
$$H_1 : \sigma_x^2 > \sigma_y^2$$

Consider the test statistic U, where S_X^2 and S_Y^2 are defined above

$$U = \frac{S_X^2/(n-1)}{S_Y^2/(m-1)}$$

We reject the null hypothesis when $U > c$, where c is the $1-\alpha$ quantile of the $F_{n-1,m-1}$ distribution. This is an α level generalized likelihood ratio test. Similar results hold for the other one-sided/two-sided hypothesis.

9 Some Applications: Statistical Models

9.1 Logistic (Binomial) Regression

Suppose a technology company would like to understand what factors influence whether or not a user clicks on an advertisement. The company has n observations of ad impressions, and for each observation i, has data on p different factors $x_{i1}, ..., x_{ip}$ such as the position of the ad, the size of the ad, the age of the visitor, the gender of the visitor, etc...We also know whether or not the visitor clicks on the ad, $Y_i = 1$ if the visitor clicks, and $Y_i = 0$ otherwise.

We can use the *logistic regression model*, a statistical model that assumes that each response Y_i is an independent random variable with distribution $Bernoulli(p_i)$. As with linear regression, the covariates x are treated as fixed and are conditioned on, i.e. we are modeling $Y_i|x_i$ with probability $p_i(x_i)$, but the notation is usually omitted. In logistic regression, the log-odds corresponding to p_i is modeled as a linear combination of the covariates and an added intercept, that is we model

$$log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

The intercept β_0 is the baseline log-odds if all other covariates are zero. The other coefficients $\beta_1, ..., \beta_p$ represent the increase or decrease in the log-odds given a unit increase in the respective covariate. Note that we can rewrite the above as

$$p_i(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}$$

9.1.1 Estimation of Logistic Regression Coefficients

As usual, we can estimate the regression coefficients using the MLE. Since $Y_1, ..., Y_n$ are independent Bernoulli random variables, the likelihood function is as follows

$$lik(\beta_0, ..., \beta_p) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \prod_{i=1}^{n}(1-p_i)\left(\frac{p_i}{1-p_i}\right)^{y_i}$$

For convenience, set $x_{i0} = 1$ for all i so that $\beta^T x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ where the vectors are defined $\beta = (\beta_0, ..., \beta_p)$ and $x_i = (1, x_{i1}, ..., x_{ip})$. Then the log-likelihood is

$$\begin{aligned}
\ell(\beta_0, ..., \beta_p) &= \sum_{i=1}^{n} \left(y_i log \left(\frac{p_i}{1-p_i} \right) + log(1-p_i) \right) \\
&= \sum_{i=1}^{n} \left(y_i \beta^T x_i + log(p_i) - \beta^T x_i \right) \\
&= \sum_{i=1}^{n} \left(y_i \beta^T x_i + log \left(\frac{exp(\beta^T x_i)}{1+exp(\beta^T x_i)} \right) - \beta^T x_i \right) \\
&= \sum_{i=1}^{n} \left(y_i \beta^T x_i - log \left(1 + exp(\beta^T x_i) \right) \right)
\end{aligned}$$

Note that from a machine learning perspective (i.e. focus on predictive power rather than statistical inference), maximizing the log-likelihood here is the same as minimizing the binary cross-entropy loss function.

We can compute the MLE numerically (e.g. by Newton-Raphson) by solving the system $\partial\ell/\partial\beta_j = 0$ for each $j = 0, ..., p$. For a function $f(x) = log(1 + e^x)$ we have first derivative $f'(x) = \frac{e^x}{1+e^x} = 1 - \frac{1}{1+e^x}$ and second derivative $f''(x) = \frac{e^x}{(1+e^x)^2}$. Then by chain rule we have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} \left(y_i - \frac{exp(\beta^T x_i)}{1+exp(\beta^T x_i)} \right) = 0$$

for which the solution gives the MLE estimates $\hat{\beta}_0, ..., \hat{\beta}_p$. To estimate the click through rate (conversion probability) for a new ad with covariate $\tilde{x}$, we can use the plug-in estimator

$$\hat{p}(\tilde{x}) = \frac{exp(\hat{\beta}^T \tilde{x})}{1+exp(\hat{\beta}^T \tilde{x})}$$

9.1.2 Confidence Intervals for Logistic Regression

To build confidence intervals around the estimated coefficients, we can apply the asymptotic theory for the MLE. To do so, we first compute the Fisher information matrix $I(\beta) = -\mathbb{E}_Y[\nabla^2 \ell(Y, \beta)]$ by finding the second partial derivatives of the likelihood function. We have

$$\frac{\partial^2 \ell}{\partial \beta_j \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} \frac{exp(\beta^T x_i)}{(1+\beta^T x_i)^2}$$

Note how the second derivatives do not depend Y, so that in calculating the Fisher information we do not need to take an expectation. If we define W as a nxn diagonal matrix as follows

$$W = W(\beta) := diag \left(\frac{exp(\beta^T x_1)}{(1+\beta^T x_1)^2}, ..., \frac{exp(\beta^T x_n)}{(1+\beta^T x_n)^2} \right)$$

with data matrix X with $x_1, ..., x_n$ in its rows, we can write the second partial derivatives as $\frac{\partial^2 \ell}{\partial \beta_j \beta_k} = -X_{:j}^T W X_{:k}$ where the vector $X_{:j}$ is the jth column of X. Finally, we have the Hessian of the likelihood is $\nabla^2 \ell = -X^T W X$ with Fisher Information matrix $I(\beta) = X^T W X$. Then the MLE estimates are asymptotically multivariate normally distributed with

$$\hat{\beta} \xrightarrow{d} \mathcal{N}(\beta, I(\beta)^{-1}) = \mathcal{N}(\beta, (X^T W X)^{-1})$$

Let $\hat{W} = W(\hat{\beta})$ be the plugin estimate of W so that we can estimate the standard error of $\hat{\beta}_j$ using $s(\hat{\beta}_j) = \sqrt{(X^T \hat{W} X)_{jj}^{-1}}$. Then we can construct a $1 - \alpha$ confidence interval for β_j using

$$\mathbb{P}(\hat{\beta}_j - z(1-\alpha/2)s(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + z(1-\alpha/2)s(\hat{\beta}_j)) = 1-\alpha$$

where $z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Note that the standard error estimates $s(\hat{\beta}_j)$ are only expected to be accurate when the model is *correctly specified*, that is, Y_i's are truly independent random variables following $Bernoulli(p_i)$, where the log-odds for each p_i are a linear combination of the covariates. The covariance matrix of $\hat{\beta}$ is given by the inverse Fisher information only in a correctly specified model. In some cases, e.g. in machine learning, logistic regression is used for binary classification problems even when the model does not accurately fit the data - as long as the classification accuracy is high. In such as a case, the above method for inference is invalid, though the standard error of $\hat{\beta}_j$ can be estimated using the nonparametric bootstrap.

9.1.3 Testing Logistic Regression Coefficients

Suppose we want to test if a particular coefficient is 0, for example

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0$$

To do so, we can fit a reduced model (under the null hypothesis) by solving for the MLE's $\hat{\beta}_0^{(0)}, ..., \hat{\beta}_{p-1}^{(0)}$ where the *pth* coefficient is removed. We can then compare the reduced model with the full model using the generalized likelihood ratio statistic

$$-2log\Lambda = -2log\frac{lik(\hat{\beta}_0^{(0)}, ..., \hat{\beta}_{p-1}^{(0)}, 0)}{lik(\hat{\beta}_0, ..., \hat{\beta}_p)}$$

which is asymptotically χ_1^2 distributed as $n \to \infty$ under the null hypothesis. An asymptotic level α test rejects the null hypothesis when $-2log\Lambda > c$, where c is the $1-\alpha$ quantile of χ_1^2.

9.2 Poisson Regression

Consider an example from neuroscience. Neurons communicate by firing electric signals, called action potentials or spikes, across a (biological) neural network. A series of a spikes over time for a single neuron is called a spike train. A common statistical model for a spike train is an in-homogeneous Poisson (point) process; that is, the number of spikes on a spike train is Poisson distributed, where the mean parameter depends on the location on the spike train. Specifically, suppose we have n times windows of length Δ, where Y_i is the number of spikes in the *ith* time window, and $Y_1, ..., Y_n$ are independent Poisson random variables with distribution $Y_i \sim Poisson(\lambda_i \Delta)$. For simplicity, suppose that $\Delta = 1$.

The spiking rate λ_i may be influenced by external sensory stimuli present in the *ith* time window, such as intensity of visible light, texture of object touched, etc...Let these external stimuli be encoded as p covariates $x_{i1}, ..., x_{ip}$. We can formulate a statistical model for λ_i based on these covariates. Specifically, the Poisson log-linear model (Poisson regression model) models independent $Poisson(\lambda_i)$ random variables using

$$log\lambda_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

or equivalently,

$$\lambda_i = exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})$$

Poisson regression is useful when modeling count data where there is no upper-limit to the possible number of counts observed. This is in contrast to binomial/logistic regression, where there is an upper limit (e.g. number of success in n trials).

9.2.1 Estimation of Poisson Regression Coefficients

Since we have $Y_1, ..., Y_n$ as independent Poisson random variables, we have likelihood

$$lik(\beta_0, ..., \beta_p) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

where λ_i is log-linear in the covariates. As before, set $x_{i0} = 1$ for all i so that $\beta^T x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ where the vectors are defined $\beta = (\beta_0, ..., \beta_p)$ and $x_i = (1, x_{i1}, ..., x_{ip})$ for convenience. We have log-likelihood

$$\begin{aligned}\ell(\beta_0, ..., \beta_p) &= \sum_{i=1}^{n} y_i log\lambda_i - \lambda_i - log(y_i!) \\ &= \sum_{i=1}^{n} y_i \beta^T x_i - exp(\beta^T x_i) - log(y_i!)\end{aligned}$$

The MLE's $\hat{\beta}_0, ..., \hat{\beta}_p$ are the solution to the system setting the partial derivatives $\partial\lambda/\partial\beta_j = 0$ for $j = 0, ..., p$.

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - exp(\beta^T x_i)) = 0$$

These can be solved numerically.

9.2.2 Inference for Poisson Regression

As with logistic regression, we can apply the asymptotic theory for the MLE to build confidence intervals around the estimated Poisson regression coefficients. To do so, we first compute the Fisher information matrix $I(\beta) = -\mathbb{E}_Y[\nabla^2 \ell(Y, \beta)]$ by finding the second partial derivatives of the likelihood function. We have

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} x_{ij} x_{ik} exp(\beta^T x_i) = -X_{:j}^T W X_{:k}$$

where the vectors of length n $X_{:j}$ and $X_{:k}$ are the jth and kth columns of the design matrix X, respectively, and W is a diagonal matrix defined as follows

$$W = W(\beta) := diag(exp(\beta^T x_1), ..., exp(\beta^T x_n))$$

Then the Hessian matrix of the likelihood is $\nabla^2 \ell(\beta) = -X^T W X$. Note how the second derivatives do not depend on Y, so that in calculating the Fisher Information we do not need to take an expectation. Then, $I(\beta) = X^T W X$. As n grows large, provided that the Poisson log-linear model is *correctly specified*, the MLE $\hat{\beta}$ is multivariate normally distributed with

$$\hat{\beta} \xrightarrow{d} \mathcal{N}(\beta, I(\beta)^{-1}) = \mathcal{N}(\beta, (X^T W X)^{-1})$$

Similar to logistic regression, we can estimate the standard error of $\hat{\beta}_j$ using $s(\hat{\beta}_j) = \sqrt{(X^T \hat{W} X)_{jj}^{-1}}$ where $\hat{W} = W(\hat{\beta})$ is the plugin estimate of W so that it does not depend on the unknown parameter β. Then an approximate $1 - \alpha$ confidence interval for β_j is

$$\mathbb{P}(\hat{\beta}_j - z(1 - \alpha/2)s(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + z(1 - \alpha/2)s(\hat{\beta}_j)) = 1 - \alpha$$

where $z(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

9.2.3 Overdispersion

The Poisson modeling assumption is often violated in practice, since it implies that the variance of Y_i must be equal to the mean. Frequently, the variance is larger than the mean, which results in what is known as *overdispersion.* However, as long as the mean assumption is correct, i.e,

$$log\mathbb{E}Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

then the MLE $\hat{\beta}$ in the Poisson model is unbiased for β even in the overdispersion case. The standard error estimate would be incorrect, though instead we can conduct the nonparametric bootstrap.

9.2.4 Generalized Linear Models (GLM)

Linear regression, logistic regression, and Poisson regression are all examples of *generalized linear models* (GLM). In a GLM, $Y_1, ..., Y_n$ are modeled as independent observations with distributions $Y_i \sim f(y|\theta_i)$ for some one-parameter family $f(y|\theta)$. We model a function of the parameter θ_i as a linear combination of the covariates

$$g(\theta_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a one-to-one function called the *link function.* For example, in linear regression, we have the parameter of interest $\theta \equiv \mu$ with link function $g(\mu) = \mu$ when modeling independent $\mathcal{N}(\mu_i, \sigma^2)$ random variables (treating σ^2 as a known fixed constant). In logistic regression, the parameter of interest was $\theta \equiv p$ with link function $g(p) = log\frac{p}{1-p}$ when modeling independent $Bernoulli(p_i)$ random variables. In Poisson regression, the parameter of interest was $\theta \equiv \lambda$ with link function $g(\lambda) = log\lambda$ when modeling independent $Poisson(\lambda_i)$ random variables.

The examples of link functions above are called *natural links.* Consider a change of variable for the parameter, $\theta = \eta(\theta)$ so that the pmf/pdf in terms of the new parameter η has form

$$f(y|\eta) = exp(\eta y - A(\eta))h(y)$$

for some functions A and h. This is called the *exponential family* form of the pdf/pmf, where η is called the natural parameter. The natural link sets the link function as the natural parameter; that is $g(\theta) = \eta(\theta)$. For example, we can write the $Bernoulli(p)$ pmf as

$$f(y) = p^y(1-p)^{1-y} = (1-p)\left(\frac{p}{1-p}\right)^y = exp\left(ylog\frac{p}{1-p} + log(1-p)\right)$$

so we can set $\eta(p) = log\frac{p}{1-p}$, $A(p) = log(1-p)$, and $h(y) = 1$. The natural link simply sets $g(p) = \eta(p) = log\frac{p}{1-p}$.

The natural link is mathematically convenient for likelihood-based inference. That is, the second partial derivatives of

$$logf(Y|\eta) = \eta Y - A(\eta) + logh(Y)$$

with respect to β does not depend on Y, since η is modeled as linear in β, i.e. $\eta_i = g(\theta_i) = \beta^T x_i$. The consequence of this is that when we derive the Fisher information matrix, we do not need to take the expectation over Y for $-\nabla^2 \ell(\beta)$.

However, from a pure modeling perspective, there is no intrinsic reason to believe that the natural link is the correct transformation of the parameter so that it is a linear combination of the covariates. Other choices of link functions are possible and commonly used, especially if they lead to a better fit of the data.

9.3 Survival Analysis: Proportional Hazards Model

9.3.1 Hazard Functions

Survival analysis is a field of statistics concerned with time-to-event/time-to-failure data. For example, consider a clinical trial where the efficacy of a drug for prolonging remission (the disappearance of cancer symptoms) induced by chemotherapy. Equivalently, we study the time until recurrence of the cancer. For the *ith* patient in the trial, denote T_i as the time until recurrence of the cancer.

We can model T_i as a continuous and positive random variable with cdf $F_i(t)$ and pdf $f_i(t) = F_i'(t)$. We can think about the distribution of T_i in terms of its *hazard function*, denoted $\lambda_i(t)$, which represents "instantaneous risk" at time t. The hazard function is defined as follows

$$\lambda_i(t) := \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}(T_i \leq t + \delta | T_i \geq t)$$

That is, $\delta\lambda_i(t)$ represents the probability that recurrence occurs sometime in the time window $[t, t+\delta]$ conditional on recurrence not having already occurred up until time t, for small δ. We have

$$\begin{aligned} \lambda_i(t) &= \lim_{\delta \to 0} \frac{\mathbb{P}(t \leq T_i \leq t + \delta)}{\delta \mathbb{P}(T_i \geq t)} \\ &= \lim_{\delta \to 0} \frac{F_i(t+\delta) - F_i(t)}{\delta(1 - F_i(t)} \\ &= \frac{f_i(t)}{1 - F_i(t)} \end{aligned}$$

The first statement follows from the definition of conditional probability, the second from the definition of the cdf, and the third from the definition of derivative.

For illustration, suppose that $T_i \sim$ *Exponential*(θ). Recall the Exponential pdf is $f_i(t) = \theta e^{-\theta t}$ with cdf $F_i(t) = 1 - e^{-\theta t}$. Evaluating into the hazard function, we have

$$\lambda_i(t) = \frac{\theta e^{-\theta t}}{1 - (1 - e^{-\theta t})} = \theta$$

In this case, the hazard function is constant in time, which is unsurprising due to the memoryless property of the exponential distribution. Here, the large θ is, the faster the exponential distribution decays and the more likely it is that cancer will occur at any next instant of time, which corresponds with a higher hazard.

9.3.2 Cox's Proportional Hazards Model

In the Cox proportional hazards model (Cox PH), we do not assume that T_i follows any particular distribution. Instead, we model the hazard function for T_i as follows

$$\lambda_i(t) = \lambda_0(t) exp(\beta_1 x_{i1} + ... + \beta_p x_{ip})$$

where $x_1, ..., x_p$ are p different covariates and $\beta_1, ..., \beta_p$ are its corresponding coefficients. For example, x_1 can be a 0-1 variable indicating if the patient has taken the drug or a placebo. The other covariates $x_2, ..., x_p$ can represent other factors, such as age, weight, race, etc., that may affect the cancer recurrence time. $\lambda_0(t)$ is called the *baseline hazard function*, which is an unknown function that controls the shape of the hazard function for *all* patients, and the scalar $exp(\beta_1 x_{i1} + ... + \beta_p x_{ip})$ controls the scale of the hazard function for each patient i. That is, we assume that the hazard function has the same shape across all patients, and the scale differs for each patient based on the covariates which act additively on $log\lambda_i(t)$. If we take the ratio of hazards for two patients i, j, $\lambda_i(t)/\lambda_j(t)$ is constant with time since the baseline hazard cancels out (hence the name "proportional hazards"). This ratio is determined by a linear combination of covariates from patient i and j, specifically, they control the log-hazards-ratio.

9.3.3 Censoring

In the clinical trail, the remission length for a patient may last longer than the duration for which the patient participates in the trial. In such as case, we do not observe the remission length T_i but only the fact that $T_i > \ell_i$ where ℓ_i is the length of time patient i participates in the trial. This patient is called *right-censored.* There is also *left-censoring*, which is less common but occurs when the event occurs *before* we begin observations. *Interval-censoring* is when we only know the observation occurs in some time window, and not the precise time. For example, interval censoring occurs if we only check on the patients at the beginning and end of the day. The method of estimation and inference described below for the Cox PH model directly accounts for right-censoring. We make the assumption that ℓ_i is constant and known, and that the remission length T_i does not depend on ℓ_i.

9.3.4 Estimating Cox Regression Coefficients

Usually, the coefficients $\beta_1, ..., \beta_p$ are of more interest than the baseline hazard function $\lambda_0(t)$ itself. To estimate these coefficients, we usually write out their likelihood function and find the MLE, however, these will depend on the unknown $\lambda_0(t)$ unless we make parametric assumptions for the baseline hazard. Instead, we circumvent this problem by deriving the *partial likelihood function.*

We begin by considering the order statistics $t_{(1)} < t_{(2)} < ... < t_{(m)}$ of *observed* remission lengths. We make the assumption that there are no ties (this is reasonable since we model continuous remission lengths). This means that each $t_{(k)}$ only corresponds to one patient, and m is the number of non-right-censored patients. Let $\mathcal{R}(t_{(k)})$ be the *risk set* at time $t_{(k)}$ representing the set of observations that are "at risk" to fail during time $t_{(k)}$, that is, the set of patients that have not yet left the study (been right-censored) and are still in remission immediately before time $t_{(k)}$. The probability that individual/patient k (denote as I_k) actually fails at time $t_{(k)}$, conditional on *some* patient from $\mathcal{R}(t_{(k)})$ failing at time $t_{(k)}$ is as follows

$$\begin{aligned}\mathbb{P}(I_k \ fails \ at \ t_{(k)} | \mathcal{R}(t_{(k)}) \ fails \ at \ t_{(k)}) &= \frac{\mathbb{P}(I_k \ fails \ at \ t_{(k)} \cap \mathcal{R}(t_{(k)}) \ fails \ at \ t_{(k)})}{\mathbb{P}(\mathcal{R}(t_{(k)}) \ fails \ at \ t_{(k)})} \\ &= \frac{\mathbb{P}(I_k \ fails \ at \ t_{(k)})}{\mathbb{P}(\mathcal{R}(t_{(k)}) \ fails \ at \ t_{(k)})} \ (since \ I_k \in \mathcal{R}(t_{(k)})) \\ &= \frac{\lambda_k(t_{(k)})}{\sum_{i \in \mathcal{R}(t_{(k)})} \lambda_i(t_{(k)})}\end{aligned}$$

Under the Cox proportional hazards model, this ratio is

$$\frac{\lambda_0(t_{(k)}) exp(\beta_1 x_{k1} + ... + \beta_p x_{kp})}{\sum_{i \in \mathcal{R}(t_{(k)})} \lambda_0(t_{(k)}) exp(\beta_1 x_{i1} + ... + \beta_p x_{ip})}$$

Importantly, the unknown baseline hazard $\lambda_0(t_{(k)})$ cancels out. Finally, we formulate the partial likelihood function as the product of these conditional probabilities over all observed failure times as follows

$$plik(\beta_1, ..., \beta_p) = \prod_{k=1}^{m} \frac{\exp(\beta_1 x_{k1} + ... + \beta_p x_{kp})}{\sum_{i \in \mathcal{R}(t_{(k)})} exp(\beta_1 x_{i1} + ... + \beta_p x_{ip})}$$

Intuitively, the partial likelihood captures for each observed failure time $t_{(k)}$, the chance that patient k fails at time $t_{(k)}$ as opposed to the other patients that could have failed at time $t_{(k)}$. In a similar spirit to the MLE, we want this value to be as high as possible so that the model aligns with observation. That is, we estimate the model coefficients by maximizing the partial likelihood. Consider the log-partial-likelihood

$$\ell(\beta_1, ..., \beta_p) = \sum_{k=1}^{m} \left(\beta_1 x_{k1} + ... + \beta_p x_{kp} - log \sum_{i \in \mathcal{R}(t_{(k)})} exp(\beta_1 x_{i1} + ... + \beta_p x_{ip}) \right)$$

We solve the system setting each $\partial\ell/\partial\beta_j = 0$ for $j = 1, ..., p$ to arrive at the maximum partial-likelihood estimates $\hat{\beta}_1, ..., \hat{\beta}_p$.

9.3.5 Inference for Cox Regression

The generalized likelihood ratio test can be applied to the coefficients estimated by maximum partial-likelihood (non-trivial). Suppose we want to test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

corresponding to x_1, the covariate determining if the patient took the drug or a placebo. Under H_0, we can construct a "reduced" model where the first covariate is omitted and the remaining coefficients $\hat{\beta}_2^{(0)}, ..., \hat{\beta}_p^{(0)}$ are fitting using the same procedure as before with maximum partial likelihood. We can use the test statistic

$$-2log\Lambda = -2log\frac{plik(0, \hat{\beta}_2^{(0)}, ..., \hat{\beta}_p^{(0)})}{plik(\hat{\beta}_1, ..., \hat{\beta}_p)}$$

which is asymptotically χ_1^2 distributed as $n \to \infty$ under the null hypothesis. An asymptotic level α test rejects the null hypothesis when $-2log\Lambda > c$, where c is the $1 - \alpha$ quantile of χ_1^2.