S&DS 242: Theory of Statistics

1/26

$\hat{p}$ is a value calculated from observed data

↳ A statistic

↳ probability distribution is called sampling distribution

If the mean $\hat{p} = p$ it is unbiased

Sampling without Replacement

$E(\hat{p}) = p$ still

$$Var[\hat{p}] = \langle \hat{p}^2 \rangle - \langle \hat{p} \rangle^2$$

$$\hat{p} = \left(\frac{x_1 + x_2 + \dots x_n}{n} \right)$$

$$\langle \hat{p}^2 \rangle = \frac{1}{n^2} \langle x_1^2 + \dots x_n^2 + 2x_1x_2 + \dots 2x_{n-1}x_n \rangle$$

$$= \frac{1}{n^2} \langle n x_1^2 + \binom{n}{2} \langle 2x_1x_2 \rangle \rangle$$

↑ Expected Values are same for each X_i

$$= \frac{1}{n^2} \cdot \langle x_1^2 \rangle + \frac{n-1}{n} \langle 2x_1x_2 \rangle$$

$\langle x_1^2 \rangle = \langle x_1 \rangle$ since $x_1 = 0$ or 1

$\langle 2x_1x_2 \rangle = P[X_1 = 1 \text{ and } X_2 = 1]$

$= P[X_1 = 1] \cdot P[X_2 = 1 \mid X_1 = 1]$

$= p \cdot p\frac{N-1}{N-1}$

$$Var(\hat{p}) = \langle \hat{p}^2 \rangle - \langle \hat{p} \rangle^2$$

$$= \frac{1}{n} \cdot p + \frac{n-1}{n} \cdot p \cdot \frac{pN-1}{N-1}$$

$$= \frac{p(1-p)}{n} \underbrace{\left(1 - \frac{n-1}{N-1} \right)}_{\text{Correction factor}}$$

Some stuff on Central Limit Theorem

Types of Inference Questions

- Hypothesis Testing : Asking a yes or no question about the distribution
- Estimation : Determining the distribution or some characteristic
- Confidence Interval

1/28

Random Variables

Discrete: finite our countably infinite number of possible values

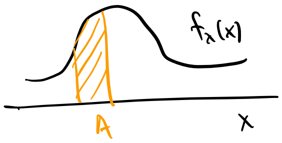
Probability Mass Function: $f_X(x) = \mathbb{P}[X=x]$

For a set of values A $\mathbb{P}[X \in A] = \sum_{x \in A} f_X(x)$ $=1$ when A is all possible

Continuous: takes any real value

Probability Density Function:

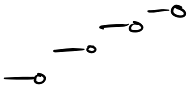
$$\mathbb{P}[X \in A] = \int_A f_X(x)\,dx$$



$$\int_R f_X(x)\,dx = \mathbb{P}[X \in R] = 1$$

Cumulative Distribution Function (CDF): $F_X(x) = \mathbb{P}[X \le x]$

Discrete CDF



$$F_X(x) = \sum_{y: y \le x} f_X(y)$$

Continous case: $F_X(x) = \int_{-\infty}^{\infty} f_X(y)\,dy$

Properties of CDF $F_X(x)$

Increasing : For any $x \le y$ always $F_X(x) \le F_X(y)$

As $x \to -\infty$ $F_X(x) \to 0$ as $x \to \infty$ $f_X(x) \to 1$

If $F_X(x)$ is strictly increasing and continuous, then $F_X(x)$ has an inverse function F_X^{-1} $: (0,1) \to \mathbb{R}$ called quantile function of X

$F_X^{-1}(t)$ satisfies $\mathbb{P}[X \le x] = t$ i.e. $F_X^{-1}(t)$ is the t^{th} quantile of X

Expected Value and Variance

$$\mathbb{E}[X] = \sum_{x \in X} x \cdot f_X(x) \qquad \mathbb{E}[X] = \int_R x f(x)\,dx$$

$g: \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[g(x)] = \sum_{x \in X} g(x) f_x(x)$$

$$\mathbb{E}[g(x)] = \int_{\mathbb{R}} g(x) f_x(x) dx$$

Expectation is linear!!!

$$Var[X] = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - \mathbb{E}(X)^2$$

Translation invariant: $Var[X+c] = Var[X]$

Multiply by constant: $Var[cX] = c^2 Var[X]$

Independent RV: $Var[X_1 + X_2] = Var[X_1] + Var[X_2]$

Dependent RV: $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$

The standard deviation of X is $SD(X) = \sqrt{Var(X)}$

A Bernoulli r.v. $X \sim Bernoulli(p)$ takes all possible values 0 and 1

The PMF is $f_x(1) = \mathbb{P}[X=1] = p$

$f_X(0) = \mathbb{P}[X=0] = 1-p$

Mean: $\mathbb{E}(X) = p \cdot 1 + (1-p) \cdot 0 = p$

Variane: $Var[X] = \mathbb{E}(X^2] - \mathbb{E}(X)^2$

$= E(X) - E(X^2)$ since $X^2 = X$ when $X=0$ or 1

$= p - p^2 = p(1-p)$

A Binomial distribution is repeated bernoulli trials

$X = X_1 + X_2 + \dots X_n \sim Binomial(n,p)$

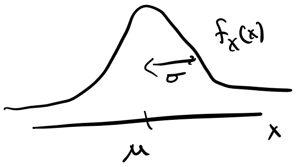
PMF is $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, 1, \dots n\}$

Mean: $\mathbb{E}(X) = \mathbb{E}(X_1 + \dots + X_n] = E(X_1] + \dots + \mathbb{E}[X_2] = n \cdot p$

Variance: $Var[X] = Var[X_1 + \dots X_n] = Var[X_1] + \dots Var[X_n] = np(1-p)$

A normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ is a continous r.v. on $\mathbb{R}$,

PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$



$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \mu$$

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \sigma^2$$

A Gamma r.v. $X \sim \text{Gamma}(\alpha, \beta)$ (for $\alpha, \beta > 0$) is continous

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0$$

$$\Gamma(\alpha) = \int_{-\infty}^{x} x^{\alpha-1} e^{-x} \, dx$$

For integer $n \geq 1$, $\Gamma(n) = (n-1)!$

Joint Distributions

Discrete case: Joint PMF $f_{X_1, \dots X_k}(x_1 \dots x_k) = \mathbb{P}[X_1 = x_1 \dots X_k = x_k]$

Continous Case: Joint PDF $f_{X_1 \dots X_k}(x_1 \dots x_k)$

$$\mathbb{P}[(X_1 .. X_k) \in A] = \iint_A f_{X_1 .. X_k}(x_1 \dots x_k) \, dx_1 \dots dx_k$$

Something about multi-nomial
↳ generalized binomia

$$f_{X_1 .. X_k}(x_1 \dots x_k) = \binom{n}{x_1 \dots x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

$$\frac{n!}{x_1! \, x_2! \dots x_k!}$$

Random Variable $X_1 .. X_k$ are independent

$$f_{X_1 \dots X_k}(x_1 \dots x_k) = f_{X_1}(x_1) \times f_{X_2}(x_2) \dots \times f_{X_k}(x_k)$$

Properties of independent r.v

i) $\mathbb{P}[X_1 \in A_1 \text{ and } X_2 \in A_2 \text{ and } \dots X_k \in A_k]$

$$= \mathbb{P}[X_1 \in A_1] \times \mathbb{P}[X_2 \in A_2] \times \dots \times \mathbb{P}[X_k \in A_k]$$

ii) For any function $g_1 \dots g_k : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[g_1(X_1) \cdot g_2(X_2) \dots g_k(X_k)] = \mathbb{E}[g_1(X_1)] \dots \mathbb{E}[g_k(X_k)]$$

E.g. If X, Y are independent, then $\mathbb{E}[XY] = \mathbb{E}(X)\,\mathbb{E}(Y)$

Covariance between r.v.'s X, Y

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

$$\text{Cov}[X,X] = \text{Var}[X]$$

Translationally Invariant: $\text{Cov}[X+a, Y+b] = \text{cov}[X,Y]$

Bilinear: $\text{cov}[aX, bY] = ab\, \text{cov}[X,Y]$

If X_i and X_j are independent

$$\text{Cov}[X_i, X_j] = \mathbb{E}(X_i X_j] - \mathbb{E}(X_i)\,\mathbb{E}(X_j]$$
$$= E(X_i]\, E(X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0$$

Independent → Cov $= 0$

Correlation

$$\text{corr}(X,Y) = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}[Y]}}$$

1/31

Def: The moment generating function (MGF) of a random variable X is given by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

ex. Normal MGF

Let $X \sim \mathcal{N}(0,1)$. What is MGF of $M_X(t)$

$$M_X(t) = \mathbb{E}[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$$

$$= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\,dx$$

$$= e^{-t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} = e^{-t^2/2}$$

PDF of $N(t,1)$

If $M_X(t)$ is finite in a small interval around 0, then it uniquely determines the distribution of X

Theorem: Let X and Y be different random variables where MGF$_s$ are finite in an interval $(-h,h)$ around 0, and

$M_X(t) = M_Y(t)$ for all $t \in (-h,h)$. Then X and Y have the same distribution

If $X_1 \dots X_n$ are independent r.v.'s then

$$M_{X_1+\dots+X_n}(t) = \mathbb{E}\left[e^{t(X_1+\dots X_n)}\right]$$

$$= \mathbb{E}\left[e^{tX_1}\right] \times \dots \times \mathbb{E}\left[e^{tX_n}\right] = M_{X_1}(t) \times \dots M_{X_n}(t)$$

Multivariate Normal Distribution

In k dimensions it is a continuous distribution for $(X_1 \dots X_k) \in \mathbb{R}^k$

It is specified by

mean vector : $\mu \in \mathbb{R}^k$

covariance matrix : $\Sigma \in \mathbb{R}^{k \times k}$

k-dimensional generalization of the normal $\mathcal{N}(\mu, \sigma^2)$

Def: $(X_1 \dots X_k)$ are multivariate normal if for my constants $a_1 \dots a_k \in \mathbb{R}$, the linear combination $a_1X_1 + a_2X_2 + \dots a_kX_k$ has a normal distribution. More specifically, $(X_1 .. X_k) \sim \mathcal{N}(\mu, \Sigma)$ if in addition to this property

$\mathbb{E}[X_i] = \mu_i$ and $\mathrm{Var}[X_i] = \Sigma_{ii}$ for each $i = 1 \dots k$

$\mathrm{Cov}[X_i, X_j] = \Sigma_{ij}$ for $i \neq j$

Note: This implies each individual r.v. is normal

You can prove this via expanding moment generating functions

$\mathrm{Cov}[X,Y] = 0$ implies independence for bivariate normal

Theorem: If $\mathbb{X} = (X_1 .. X_k)$ is multivariate normal and $\mathbb{X}_1$ and $\mathbb{X}_2$ are two subvectors that are in pairwise uncorrelated, then $\mathbb{X}_1$ and $\mathbb{X}_2$ are independent.

Sampling Distribution of Statistics

For data $X_1 \dots X_n$ a statistic $T(X_1 \dots X_n)$ is any value computed from this data

- Sample mean: $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$
- Sample variance: $S^2 = \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\right)$
- Sample Range: $R = \max(X_1 \dots X_n) - \min(X_1 \dots X_n)$

2/2

A statistic is any function of data

Randomness of data induces a sampling distribution

Chi-Squared Distribution

$X_1, \dots X_n \overset{IID}{\sim} N(0,1)$

Sampling distribution of $X_1^2 + \dots + X_n^2 \sim \chi^2_n$ with n degrees of freedom

$$M_{X_1^2+\dots+X_n^2} = M_{X_1^2}(t) \times \dots \times M_{X_n^2}(t)$$

$$M_{X_i^2}(t) = \mathbb{E}\left[e^{tX_i^2}\right] = \int_{-\infty}^{\infty} e^{tx^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(t-1/2)x^2} \, dx$$

$t \geq 1/2$ blows up

$$M_{X_i^2}(t) = \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \sqrt{\frac{1-2t}{2\pi}} \, e^{-\frac{1}{2}(1-2t)x^2} \, dx$$

just MGF of Gamma $\left(\frac{1}{2}, \frac{1}{2}\right)$ So $X_i^2 \sim$ Gamma $(1/2, 1/2)$

$$M_{X_1^2 + \dots X_n^2}(t) = \begin{cases} \infty & t \geq 1/2 \\ (1-2t)^{-n/2} & t < 1/2 \end{cases} \leftarrow \text{Gamma}\left(\tfrac{n}{2}, 1/2\right)$$

Distributions that are difficult to study will be studied by approximation

Simulate via R

Asymptotic approximations

1) Faster

2) Theoretical Understanding

Weak Law of Large Numbers

Suppose $X_1 .. X_n$ are IID, with $\mathbb{E}[X_i] = \mu$ and $\text{Var}[X_i] < \infty$. Let

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$$

$\bar{X} \to \mu$ as $n \to \infty$

Central Limit Theorem

Suppose $X_1, \dots X_n$ are iid with $\mathbb{E}[X_i]=\mu$ and var $[X_i]=\sigma^2$

Let $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ Then

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) \to N(0,1) \quad \text{is distribution as } n \to \infty$$

Normal Distribution Approximation accuracy depends on

i) Sample size

ii) Skewness

iii) Heaviness of tails

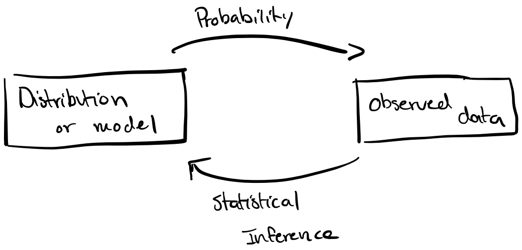
Continuous Mapping Theorem

Let $g(x)$ be a continuous function of x. As $n \to \infty$

i) If $S_n \to Z$ in distribution, then $g(S_n) \to g(Z)$ in dist

ii) Analogous for probability

2/7



Einstein's Theory of Brownian Motion

Suppose the particle is at position $P_t \in \mathbb{R}^2$ at time t. Then at time $t+\Delta t$, the position $P_{t+\Delta t}$ is random and has a bivariate normal distribution around P_t

$$P_{t+\Delta t} - P_t \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}\right)$$

$$\sigma^2 = \frac{RT}{3\pi\eta r N_A}(\Delta t) \leftarrow$$

how tf he derive this

Hypothesis test is a binary question about the distribution of the data

Accept/reject null hypothesis H_0 in favor of alternative hypothesis H_1

For the magicians die. Null hypothesis: die is fair

$H_0: (X_1 \dots X_6) \sim \text{Multinomial}(n, (1/6 \dots 1/6))$

$H_1: (X_1 \dots X_6) \sim \text{Multinomial}(n, (P_1 \dots P_6))$

$P_1 \dots P_6 \neq (1/6, \dots 1/6)$

Setting up notation for brownian motion experiment

$(X_1, Y_1) = P_1 - P_0$

$(X_2, Y_2) = P_2 - P_1$

$\vdots$

$(X_n, Y_n) = P_n - P_{n-1}$

displacements are measured every 30 sec

$H_0 : (X_1, Y_1) \ldots (X_n, Y_n) \overset{iid}{\sim} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2.23e-7 & 0\\0 & 2.23e-7\end{pmatrix}\right)$ ← Einstein's Theory

$H_1 : (X_1, Y_1) \ldots (X_n, Y_n) \overset{iid}{\sim} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\0 & \sigma^2\end{pmatrix}\right)$ ← Variance is wrong

Many other possibilities

Neyman Pearson Paradigm

Hypothesis testing is a binary test

Does that provide sufficiently strong evidence to reject H_0, in favor of H_1?

Default assumption is that H_0 is true

A test statistic T is a statistic computed from data that provides evidence against H_0 when extreme valued

1) How can we design a test statistic

2) How can you use it?

1) How can we design a test statistic?

For the brownian motion example:

$\bar{R} = \frac{1}{n}(R_1 + R_2 + \ldots R_n)$ $\quad R_i = X_i^2 + Y_i^2$ average distance in 30 sec interval

$\mathbb{E}(\bar{R}) = \mathbb{E}(R_i) = E(X_i^2) + \mathbb{E}(Y_i^2) = 4.46e-7$

Extreme values of $\bar{R}$ reject null hypothesis

Can also compare against $(2.23e-7)\,\chi_2^2$

Plot expected distribution against histogram

Hanging histogram plots $O_i - E_i$ for each histogram bin where O_i is the observed count in bin i and E_i is the theoretical expected count

Test statistic $T = \sum_i (O_i - E_i)^2$

Large T indicates rejection of null hypothesis

You can stabilize the variance by plotting $\frac{O_i - E_i}{\sqrt{E_i}} = \frac{O_i - E_i}{\sqrt{np_i}}$ so $\mathbb{E}\left[\left(\frac{O_i - E_i}{\sqrt{np_i}}\right)^2\right] = 1$

Called hanging chi-gram

$T = \sum_i \frac{(O_i - E_i)^2}{E_i}$ is the pearson chi-squared statistic for goodness of fit

Another alternative is to consider $\sqrt{O_i} - \sqrt{E_i}$

$\sqrt{O_i} - \sqrt{E_i} \simeq \frac{O_i - E_i}{2\sqrt{E_i}}$ via taylor expansion

Tukey's hanging histogram

QQ plot

plots sorted values $R_1 \dots R_n$ against $\frac{1}{n}, \frac{2}{n} \dots \frac{n}{n}$ quantiles of their hypothesized distribution

Values deviating from $y=x$ provide evidence against hypothesized distribution

Let $R_{(1)} < \dots < R_{(n)}$ be the sorted values of $R_1 \dots R_n$. The maximum vertical deviation is

$$T = \max_{i=1}^{n} \left| R_{(i)} - F^{-1}\left(\frac{i}{n}\right) \right|$$

F^{-1} is the quantile function ← inverse cdf

more deviation is higher quantiles (not as scrunched together)

Stabilize Quantile Spacing

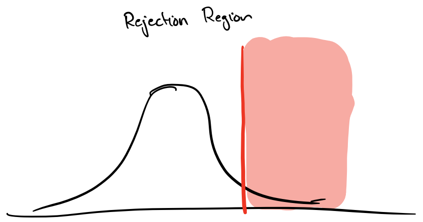
$$T = \max_{i=1}^{n} \left| F(R_{(i)}) - \frac{i}{n} \right|$$

Kolmogorov-Smirnov Statistic

2) What values of T would allow us to reject H_0?

Null distribution of T is the distribution of T under the assumption the null hypothesis is true

We divide space of possible T values into acceptance and rejection regions



Type I Error: probability that we reject H_0 when H_0 is true

In Neyman Pearson paradigm, we choose rejection error s.t.

Type I error $\leq \alpha$

for a specified significance level $\alpha \in (0,1)$

Generally don't want to specify a significance level

p-value is the smallest significance level that we would reject H_0

probability that the null distribution assigns values $\geq t_{obs}$ ← specifically for large T

2/9

Recap:

H_0 : Null hypothesis

H_1 : Alternative hypothesis

① Designing a good test statistic

② Choosing rejection region

Type I error: $\mathbb{P}_{H_0}[\text{reject } H_0] \le \alpha$ ← significance level

What is the best choice of test statistic?

Simple hypothesis

Def: A hypothesis is simple if it completely specifies its distribution

no unknown parameters

Neyman Pearson Lemma

Simple H_0 vs. Simple H_1

Def: The type II error for a simple alternative hypothesis H_1 is

$$B = P_{H_1}[\text{accept } H_0]$$

Power of the test is $1-B$

$$1-B = P_{H_1}[\text{reject } H_0]$$

Goal of hypothesis testing is to maximize power against H_1 while constraining type I error $\le \alpha$

Example: Consider $X \in \{1,2,3,4,5\}$ Two hypotheses

x	1	2	3	4	5
$f_0(x)$	0.2	0.2	0.2	0.2	0.2
$f_1(x)$	0	0.1	0.2	0.3	0.4

Let's say we want a test w/ significance level $\alpha = 0.4$

A hypothesis is defined by a rejection region $R \subseteq \{1,2,3,4,5\}$

If $x \in R$: Reject H_0

If $X \notin R$: Accept H_0

Type I Error: $\mathbb{P}_{H_0}[X \in R]$

Power $= \mathbb{P}_{H_1}[X \in R]$. To minimize power, set $R = \{4,5\}$. Then power is 0.7

More generally: Suppose the data is

$$\mathbf{X} = (X_1, \dots X_n)$$

Takes values $x = (x_1, .. x_n) \in \mathcal{X}$ is finite

H_0: $\mathbf{X}$ has joint pmf $f_0(x)$

H_1: $\mathbf{X}$ has joint pmf $f_1(x)$

To define a test, we need to define a region $R \subseteq \mathcal{X}$

If $X \in R$: Reject H_0

If $X \notin R$: Accept H_0

Want to find points of high f_1 and low f_0

Intuition: R should contain points $\mathbf{X}$ that have the smallest values of

$$L(\mathbf{X}) = \frac{f_0(x)}{f_1(x)}$$ ← increase in Type I error per unit increase in power

$L(\mathbf{X})$ is the likelihood ratio, Test rejects H_0 for small $L(\mathbf{X})$ is the likelihood ratio test

Analagous for continuous distributions

Neyman-Pearson Lemma: Let H_0 and H_1 be simple hypotheses. The significance level $\alpha \in (0,1)$. Suppose there is $c > 0$ s.t. the likelihood ratio test

rejects H_0 when $L(\mathbf{X}) < c$

accepts H_1 when $L(\mathbf{X}) \ge c$

has type I error of exactly α. Then for any other test of type I error $\le \alpha$, its power is at most the power of LRT.

Proof: Consider the discrete case. Let $R = \{\mathbf{X} : L(\mathbf{X}) < c\}$ be the rejection region of LRT.

Then R maximizes

$$\sum_{x \in R} (c f_1(x) - f_0(x))$$ among all subsets of $\mathcal{X}$

b/c if $x \in R$ then $\frac{f_0(x)}{f_1(x)} < c$ so $c f_1(x) - f_0(x) > 0$

if $X \notin R$ then $\frac{f_0(x)}{f_1(x)} \le c$ so $c f_1 - f_0 \le 0$

Consider another test w/ rejection region R'.

$$\sum_{X\in R} cf_1(x) - f_0(x) \geq \sum_{X\in R'} cf_1(x) - f_0(x)$$

$$\Rightarrow \quad c\Big(\underbrace{\sum_{X\in R} f_1(x)}_{\text{power of LRT}} - \underbrace{\sum_{X\in R'} f_1(x)}_{\text{power of other test}}\Big) \geq \underbrace{\sum_{X\in R} f_0(x)}_{\text{type I error of LRT}} - \underbrace{\sum_{X\in R'} f_0(x)}_{\text{type I error of other test}}$$

$$\geq 0$$

power of LRT $\geq$ power of other test

Examples:

$X_1 \dots X_n$ are normal

$H_0: X_1 \dots X_n \overset{iid}{\sim} N(0,1)$

$H_1: X_1 \dots X_n \overset{iid}{\sim} N(\mu,1)$

for some fixed $\mu > 0$.

Let's apply N-P lemma

$$f_0(\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-X_i^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\left(\frac{X_1^2}{2} + \dots \frac{X_n^2}{2}\right)}$$

$$f_1(\mathbf{X}) = \prod_{r=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(X_i-\mu)^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\left(\frac{(X_1-\mu)^2}{2} + \dots \frac{(X_n-\mu)^2}{2}\right)}$$

$$L(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} = \frac{e^{-\left(\frac{X_1^2}{2} + \dots \frac{X_n^2}{2}\right)}}{e^{-\left(\frac{(X_1-\mu)^2}{2} + \dots + \frac{(X_n-\mu)^3}{2}\right)}} = e^{-\left(\frac{X_1^2}{2} + \dots \frac{X_n^2}{2}\right) + \left(\frac{(X_1-\mu)^2}{2} + \dots \frac{(X_n-\mu)^2}{2}\right)}$$

$$= e^{\frac{n\mu^2}{2} - \mu(X_1 + \dots X_n)}$$

By the N-P lemma: Want to pick $c > 0$ s.t.

$$\underbrace{\mathbb{P}_{H_0}[L(\mathbf{X}) < c]}_{\text{type I error}} = \alpha$$

Then LRT which

rejects the null hypothesis when $L(\mathbf{X}) < c$

accepts the null hypothesis when $L(\mathbf{X}) \geq c$

is the most powerful test

Two observations:

$$L(\mathbf{X}) = e^{\frac{n\mu^2}{2} - \mu(X_1 + \dots + X_n)} =: f(\bar{X})$$

where $\bar{X} = \frac{(X_1 + \dots X_n)}{n}$ and $f(y) = e^{\frac{n\mu^2}{2} - n\mu y}$

This function is decreasing in y

so $L(\mathbf{X}) = f(\bar{X}) < c \iff \bar{X} > f^{-1}(c)$

we want to pick $f^{-1}(c)$ s.t.

$$\mathbb{P}_{H_0}[\bar{X} > f^{-1}(a)] = \alpha$$

but we know under the null the $X_1 \dots X_n \overset{iid}{\sim} N(0,1)$, $\bar{X} \sim N(0, 1/n)$. So $f^{-1}(c)$ should be the $(1-\alpha)^{th}$ quantile of $N(0, 1/n)$. i.e. $\frac{1}{\sqrt{n}} z(\alpha)$ where $z(\alpha)$ is the upper portion of $N(0,1)$

So the most powerful test is:

Reject H_0 if $\bar{X} > \frac{1}{\sqrt{n}} z(\alpha)$

Accept H_0 if $\bar{X} \leq \frac{1}{\sqrt{n}} z(\alpha)$

② No dependence of μ. The most powerful test is this same test for every $\mu > 0$

Lecture 7 (2/14/22)

Recap: Test $H_0 \sim f_0$
$H_1 \sim f_1$

Most powerful test is to:

Compute likelihood: $L(x) = \frac{f_0(x)}{f_1(x)}$

Pick a number $c > 0$ s.t. $P_{H_0}[L(x) < c] = \alpha$

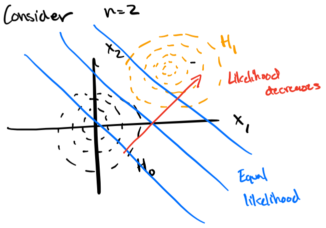
Reject H_0 when $L(x) < c$, accept H_0 when $L(x) \geq c$

Example:

$H_0 : X_1 \dots X_n \overset{iid}{\sim} N(0,1)$

$H_1 : X_1 \dots X_n \overset{iid}{\sim} N(\mu,1)$ $\mu > 0$ fixed and known

Consider $n=2$



Most powerful test is equivalent to

Compute $\bar{X} = \frac{X_1 + \dots + X_n}{n}$

Pick $c \in \mathbb{R}$ s.t. $\mathbb{P}_{H_0}[\bar{X} > c] = \alpha$

$c = \frac{1}{\sqrt{n}} z(\alpha)$

Reject H_0 if $\bar{X} > c$, accept H_0 if $\bar{X} \leq c$

Ex: $H_0 : X_1 \dots X_n \overset{iid}{\sim} \text{Bernoulli}(1/2)$

$H_1 : X_1 \dots X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$$L(\mathbb{X}) = \frac{f_0(x)}{f_1(x)} = \left(\frac{1}{2(1-p)}\right)^n \cdot \left(\frac{1-p}{p}\right)^{X_1 + \dots X_n}$$

Observation: $L(\mathbb{X})$ only depends on $X_1 \dots X_n$ via their sum

For $p > 1/2$ $\frac{1-p}{p} < 1$. So $L(\mathbb{X})$ is decreasing in $X_1 + \dots X_n$

Likelihood procedure is equivalently:

- Compute $S = X_1 + \dots X_n$
- Under H_0 $S \sim \text{Binomial}(n, 1/2)$

pick c as the upper α point $b_n(\alpha)$ of binomial $(n, 1/2)$

- Reject H_0 when $S > c$, acept when $S \leq c$

Composite Hypothesis and pivotal statistic

A hypothesis (either H_0 or H_1) that is not simple is composite

Ex. Let $n = 250$ be the number of students in STAT 242

Does 242 improve student knowledge of statistics?

Suppose each student takes

- diagnostic exam at start
- final exam at end

Let X_i be the difference (final - diagnostic) for student i.

Two formulations:

$H_0: X_1 \dots X_n \overset{iid}{\sim} N(0, \sigma^2)$ for some

$H_1: X_1 \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ $\mu, \sigma^2 > 0$

Both H_0 and H_1 are composite

$H_0: X_1 \dots X_n$ are iid from some pdf f w/ median 0

$H_1: X_1 \dots X_n$ are iid from some pdf f w/ median >0

Both H_0 and H_1 are composite

When testing composite null H_0:

- We want to ensure probability of Type I error is $\le \alpha$ for <u>every</u> possible data distribution described H_0
- Simplify design of the test by picking test statistic T whose distribution is the same under very distribution in H_0

↑ pivotal or distribution free test statistic

When testing composite Alternative H_1:

- Oftentimes no single test that maximizes power against all possible Alternatives
- We often design a test to balance the power against different alternatives

One-sample t-test

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

$H_0: \mu = 0$ $\qquad H_1: \mu > 0$

Q: what if σ^2 is known

In this case, we can first standardize our data: $Z_i = \frac{X_i}{\sigma}$

Then $Z_1 \sim N(\frac{\mu}{\sigma}, 1)$. The above is equivalent to testing $H_0: \frac{\mu}{\sigma} = 0$ vs. $H_1: \frac{\mu}{\sigma} > 0$ based on $Z_1 \dots Z_n$

The Neyman Pearson Lemma implies Most powerful level-α test is to reject if $\bar{Z} = \frac{\bar{X}}{\sigma} > \frac{1}{\sqrt{n}} z(\alpha)$

Idea: If sigma is unknown, let's estimate it from the data

Let $S^2 = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\right]$

Use instead

$$T = \frac{\sqrt{n}\,\bar{X}}{S}$$ ← One sample t-statistic

Pivotal Statistic!

Thm: Let $X_1 \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Let $\bar{X}$ and S^2 be the sample mean and variance. Then

1) S^2 is independent of $\bar{X}$

2) $S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$

Proof: WLOG $\mu = 0$

1) It sufffices to show $\bar{X}$ is independent of $(X_1 - \bar{X}, \dots X_n - \bar{X})$

↑ S_n is a simple function of $(X_1 - \bar{X}, \dots, X_n - \bar{X})$

$(\bar{X}, X_1-\bar{X}, \dots X_n-\bar{X})$ are all linear combinations of $X_1 \dots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$\Rightarrow (\bar{X}, X_1-\bar{X}, \dots X_n-\bar{X})$ is multivariate

So to show $\bar{X}$ is independent of $(X_1-\bar{X}, \dots X_n-\bar{X})$ its enough $\mathrm{Cov}[\bar{X}, X_i-\bar{X}] = 0$

We have:

$$\mathrm{Cov}[\bar{X}, X_i] = \mathrm{Cov}\left[\frac{1}{n}\sum_{j=1}^{n} X_j, X_i\right] = \frac{1}{n}\sum_{j=1}^{n} \underbrace{\mathrm{Cov}[X_j, X_i]}_{=0 \text{ for } i \neq j} = \frac{1}{n}\mathrm{Cov}[X_j, X_i] = \frac{1}{n}\mathrm{Var}[X_i] = \frac{\sigma^2}{n}$$

$$\mathrm{Cov}[\bar{X}, X_i] = \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

$\Rightarrow \mathrm{Cov}[\bar{X}, X_i-\bar{X}] = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$ so $\bar{X}$ and S^2 are independent

2) For the distribution of S^2, write $Z_i = \frac{X_i}{\sigma} \sim N(0,1)$

$$S^2 = \frac{1}{n-1}\left[(X_i-\bar{X})^2 + \dots + (X_n-\bar{X})^2\right]$$

$$\Rightarrow \underbrace{\frac{(n-1)}{\sigma^2} S^2}_{U} = (Z_1-\bar{Z})^2 + \dots + (Z_n-\bar{Z})^2$$

$$= (Z_1^2 - 2Z_1\bar{Z} + \bar{Z}^2) + \dots + (Z_n^2 - 2Z_n\bar{Z} + \bar{Z}^2)$$

$$= (Z_1^2 + \dots Z_n^2) - 2\bar{Z}\underbrace{(Z_1 + \dots + Z_n)}_{n\bar{Z}} + n\bar{Z}^2$$

$$= \underbrace{(Z_1^2 + \dots Z_n^2)}_{W} - \underbrace{(\sqrt{n}\,\bar{Z})^2}_{V}$$

This shows that $W = U + V$

we know $\bar{X}$ is independent of S^2

So V is independent of U

Here: $W \sim \chi^2_n$. $V \sim \chi^2_1$ b/c $\sqrt{n}\bar{Z} \sim N(0,1)$

$\Rightarrow U \sim \chi^2_{n-1}$.

Thus, $\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1} \Rightarrow S^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$

Returning to $T = \frac{\sqrt{n}\bar{X}}{S} = \frac{\sqrt{n}\ \bar{X}/\sigma}{S/\sigma}$

- Under H_0, $\frac{\sqrt{n}\bar{X}}{\sigma} \sim N(0,1)$ b/c $X_i \sim N(0, \sigma^2)$

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

- And $\frac{\sqrt{n}\bar{X}}{\sigma}$ is independent of S^2/σ^2

Def: If $Z \sim N(0,1)$, $U \sim \chi_n^2$ and Z and U are independent, then the distribution of $\frac{Z}{\sqrt{\frac{1}{n}U}}$ is called the t-distribution w/ n degrees of freedom

Remark: The preceeding thm explains why we use $\frac{1}{n-1}$ to define S^2:

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{\sigma^2}{n-1}\chi^2_{n-1}\right], \text{ hence } \mathbb{E}[\chi^2_{n-1}] = n-1 \quad \text{so} \quad \mathbb{E}[S^2] = \sigma^2 \quad \text{so } S^2 \text{ is unbiased for } \sigma^2$$

Lecture 8 (2/16/22)

Recap: Data $X_1 \dots X_n$ iid samples. Test whether distribution of X_i's is "centered around 0"

Formulation in Normal Setting:

$H_0: X_1 \dots X_n \overset{iid}{\sim} N(0,\sigma^2)$

$H_1: X_1 \dots X_n \overset{iid}{\sim} N(\mu,\sigma^2)$, $\mu > 0$

One-sample t-statistic

$T = \frac{\sqrt{n}\bar{X}}{S}$ where $\bar{X}$ is the average and $S^2 = \frac{1}{n-1}\left((X_1-\bar{X})^2 + \dots + (X_n-\bar{X})^2\right)$

What is distribution of t under H_0?

$$\left.\begin{array}{l} \cdot\ \frac{\sqrt{n}\bar{X}}{\sigma} \sim N(0,1) \\ \\ \cdot\ \frac{S^2}{\sigma^2} \sim \frac{1}{n-1}\chi^2_{n-1} \\ \\ \cdot\ \text{These two statistics are independent} \end{array}\right\} \text{So } T \sim t_{n-1}\text{, the } t\text{-distribution with } n\text{-1 degrees of freedom}$$

T-test: Reject H_0 when $T > t_{n-1}$

Formulation in a non-parametric: $X_1 \dots X_n \overset{iid}{\sim} f$

H_0: Median of distribution is 0

H_1: Median is greater than 0

Consider the sign statistic $S = \sum_{i=1}^{n} \overset{\text{Indicator}}{\mathbb{1}}\{X_i > 0\}$

$$\mathbb{1}\{X_i > 0\} = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i \le 0 \end{cases}$$

Under the distribution f described by H_0, $S \sim \text{Binomial}(n, 1/2)$

Let $b_n(\alpha)$ be upper-α point of binomial$(n, 1/2)$. Then the test that rejects for $S > b_n(\alpha)$ is the sign test

Remark: For large n we can approximate this test via a normal approximation

- By the CLT: $$\frac{\frac{S}{n} - \frac{1}{2}}{\sqrt{1/4}} \cdot \sqrt{n} \longrightarrow N(0,1)$$

$$\frac{S}{n} - \frac{1}{2} \simeq N(0, 1/4n) \Rightarrow S \simeq N\left(\frac{n}{2}, n/4\right)$$

- The $b_n(\alpha) \approx$ upper alpha point of $N(n/2, n/4)$

$$= \frac{n}{2} + \sqrt{\frac{n}{4}} \cdot \underset{\text{upper alpha point}}{z(\alpha)}$$

Approximate sign test by rejecting H_0

$$S > \frac{n}{2} + \sqrt{n/4}\, z(\alpha) \iff \sqrt{4n}\left(\frac{S}{n} - \frac{1}{2}\right) > z(\alpha)$$

The type I error is not exactly α but under H_0

$$\mathbb{P}[\text{Reject } H_0] \xrightarrow{n\to\infty} \mathbb{P}[\mathcal{N}(0,1) > z(\alpha)] = \alpha$$

type I error approaches α

Two-sample Tests

Ex. $X_1 \dots X_n$ are differences in exam scores for $n = 250$ students

Q: What if the exams are not equally difficult?

A: Take a separate group of $m=100$ students not in S&DS 242

Let $Y_1 \dots Y_m$ be the differences in scores

Informally, test whether the distribution of X_i's is larger than the Y_i's

Normal Model Function:

Suppose $X_1, \dots X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma^2)$

$Y_1 \dots Y_n \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma^2)$

$H_0: \mu_x = \mu_y$

$H_1: \mu_x > \mu_y$

More non-parametric formulation

Suppose $X_1 \dots X_n \overset{iid}{\sim} f$, $Y_1 \dots Y_n \overset{iid}{\sim} g$

$H_0: f = g$

$H_1: f$ stochastically dominates g

↑ If $X \sim f$ and $Y \sim g$ then for any $x \in \mathbb{R}$ $\mathbb{P}[X \geq x] \geq \mathbb{P}[Y \geq x]$

Two-sample t-test

Intuition: Look at difference of sample means $\bar{X} - \bar{Y}$

Reject $H_0: \mu_x = \mu$ in favor $H_1: \mu_x > \mu_y$

if $\bar{X} - \bar{Y}$ is "large enough"

What is the null distribution of $\bar{X} - \bar{Y}$

$$\bar{X} = \frac{X_1 + \dots X_n}{n} \sim \mathcal{N}\left(\mu_x, \frac{\sigma^2}{n}\right)$$

$$\bar{Y} = \frac{Y_1 + \dots Y_m}{m} \sim \mathcal{N}\left(\mu_y, \frac{\sigma^2}{m}\right)$$

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

Under $H_0: \mu_x = \mu_y$

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

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

If σ^2 were known, then reject H_0 if $\dfrac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} > z(\alpha)$

When σ^2 is unknown, estimate σ^2 from the data

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

Test Statistic

$$T = \frac{\bar{X}-\bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n}+\frac{1}{m}}} \leftarrow \text{pooled two-sample t-statistic}$$

Statistic is pivotal under H_0: Let $\mu = \mu_x = \mu_y$

Write $X_i = \mu + \sigma Z_i$ where Z_i's and W_j's are now $N(0,1)$

$Y_j = \mu + \sigma W_j$

Substituting into T: μ cancels from $\bar{X}-\bar{Y}$

σ cancels from $\frac{(\bar{X}-\bar{Y})}{S_{\text{pooled}}}$

What is the distribution of T?

$$\left.\begin{array}{l} \cdot\ \frac{\bar{X}}{\sigma} \sim N\left(\frac{\mu}{\sigma}, \frac{1}{n}\right) \\ \cdot\ \frac{\bar{Y}}{\sigma} \sim N\left(\frac{\mu}{\sigma}, \frac{1}{m}\right) \end{array}\right\} \quad \frac{\bar{X}-\bar{Y}}{\sigma} \sim N\left(0, \frac{1}{n}+\frac{1}{m}\right)$$

- $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1}$

- $\frac{1}{\sigma^2}\sum_{j=1}^{m}(Y_j-\bar{Y})^2 \sim \chi^2_{m-1}$

$$\Rightarrow \frac{S_{\text{pooled}}^2}{\sigma^2} = \frac{1}{m+n-2}\left(\overbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2}^{\chi^2_{n-1}} + \overbrace{\frac{1}{\sigma^2}\sum_{j=1}^{m}(Y_j-\bar{Y})^2}^{\chi^2_{m-1}}\right)$$

$$\sim \frac{1}{m+n-2}\cdot \chi^2_{m+n-2}$$

This implies $\mathbb{E}[S_{\text{pooled}}^2] = \sigma^2$, so this is unbiased

$$\text{So } T = \frac{\bar{X}-\bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\frac{1}{\sigma\sqrt{\frac{1}{n}+1/m}}(\bar{X}-\bar{Y})}{S_{\text{pooled}}/\sigma} = \frac{N(0,1)}{\sqrt{\frac{1}{m+n-2}\chi^2_{m+n-2}}}$$

$\sim t_{m+n-2}$, the t-distribution w/ m+n-2 degrees of freedom

The two sampled t-test rejects $H_0 = \mu_x = \mu_y$ in favor of $H_1: \mu_x > \mu_y$ when $T > t_{m+n-2}(\alpha)$

Remark: Assumed X_i and Y_i have the same variance

It's common to see applications where variances are not identical

Suppose $X_1 \dots X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$

$Y_1 \dots Y_n \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$

Test $H_0: \mu_x = \mu_y$ vs. $H_1: \mu_x > \mu_y$

Idea: Look at $\bar{X}-\bar{Y}$. Under H_0:

$$\bar{X}-\bar{Y} \sim N\left(0, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}\right)$$

Estimate σ_x^2 by sample var. $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$

σ_y^2 . . . S_y^2

Define $T_{\text{welch}} : \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}\cdot S_x^2 + \frac{1}{m}S_y^2}}$ ← Not exactly pivotal under H_0 and not exactly t-distributed

Mann-Whitney-Wilcoxon Rank-Sum Test

$X_1 \dots X_n \overset{iid}{\sim} f$ and $Y_1 \dots Y_m \overset{iid}{\sim} g$

Test $H_0: f=g$ vs. H_1: f stochastically dominates g

The M-W-W rank-sum statistic T is defined as:

① Pool all X_i's and Y_j's and rank them

Let smallest have rank 1

Largest have rank $m+n$

② Define T as the sum of ranks of only the Y_j's

Lecture 9 (2/21/22) - Permutation Tests

$X_1 \dots X_n \overset{iid}{\sim} f \qquad Y_1 \dots Y_m \overset{iid}{\sim} g$

$H_0: f=g \qquad H_1$: f stochastically dominates g

one-sided alternative

two-sided : $f \neq g$

Mann-Whitney-Wilcoxon Rank Sum

① Pool together $X_1 \dots X_n, Y_1 \dots Y_m$ and rank them smallest to largest

② T = sum of ranks $Y_1 \dots Y_m$

Under H_0, $X_1 \dots X_n, Y_1 \dots Y_m$ are all iid

By symmetry, equally likely for $Y_1 \dots Y_m$ to be any set of m ranks among $\{1,2 \dots m+n\}$

Thm: Under $H_0: f=g$

a) $\mathbb{E}[T] = \dfrac{m(m+n+1)}{2}$ and $\text{Var}[T] = \dfrac{mn(m+n+1)}{12}$

b) $\dfrac{T-\mathbb{E}[T]}{\sqrt{\text{Var}[T]}} \longrightarrow \mathcal{N}(0,1)$ in distribution as $n,m \to \infty$

Proof of part a

Let $N = m+n$. Define

$$I_k = \begin{cases} 1 & \text{rank } k \text{ observation is a } Y \\ 0 & \text{rank } k \text{ observation is a } X \end{cases}$$

$$T = \sum_{k=1}^{N} k I_k$$

Have the values k when $I_k=1$ is a simple random sample from $\{1-N\}$ under H_0

$$\mathbb{E}[I_k] = \mathbb{P}[I_k=1] = \frac{m}{N}$$

$$\mathbb{E}[T] = \sum_{k=1}^{N} k \cdot \mathbb{E}[I_k] = \sum_{k=1}^{N} k \cdot \frac{m}{n} = \frac{N(N+1)}{2} \cdot \frac{m}{N} = \frac{m(m+n+1)}{2}$$

$$\text{Var}[T] = \mathbb{E}[T^2] - (\mathbb{E}[T])^2$$

$$\mathbb{E}[T^2] = \mathbb{E}\left[\left(\sum_{k=1}^{N} kI_k\right)\left(\sum_{k=1}^{N} jI_j\right)\right]$$

$$= \sum_{k=1}^{N} jk\, \mathbb{E}[I_k]\mathbb{E}[I_j]$$

$$= \sum_{k=1}^{N} k^2 \underbrace{\mathbb{E}[I_k^2]}_{\frac{m}{N}} + 2 \sum_{j \neq k} jk \underbrace{\mathbb{E}[I_k I_j]}_{\frac{m}{n} \cdot \frac{m-1}{N-1}}$$

Apply: $\sum_{k=1}^{N} k^2 = \frac{N(N+1)(2N+1)}{6}$

$$2\sum_{j\neq k} jk = \left(\sum_{k=1}^{N} k\right)^2 - \sum_{k=1}^{N} k^2$$

$$= \left(\frac{N(N+1)}{2}\right)^2 - \frac{N(N+1)(2N+1)}{6}$$

$$\Rightarrow \mathbb{E}[T^2] = \frac{m}{N}\cdot\frac{N(N+1)(2N+1)}{6} + \frac{m}{N}\cdot\frac{m-1}{N-1}\left[\left(\frac{N(N+1)}{2}\right)^2 - \frac{N(N+1)(2N+1)}{6}\right]$$

$$\mathrm{Var}[T] = \mathbb{E}[T^2] - \mathbb{E}(T)^2$$

$$= \frac{mn(m+n+1)}{12}$$

To preform an asymptotic test

For large n,m distribution of T is $\approx \mathcal{N}\left(\frac{m(m+n+1)}{2}, \frac{mn(m+n+1)}{12}\right)$

To test against H_1: f stochastically dominates g

The T takes smaller value under H_1

Reject H_0 when $T < \frac{m(m+n+1)}{2} - \sqrt{\frac{mn(m+n+1)}{12}}\cdot z(\alpha)$

To test against $H_1: f\neq g$

Reject when $\left|T - \frac{m(m+n+1)}{2}\right| > \sqrt{\frac{mn(m+n+1)}{12}}\cdot z(\frac{\alpha}{2})$

Permutation Tests

$X_1 .. X_n \overset{iid}{\sim} f,\ Y_1 .. Y_m \sim g$

Test $H_0: f=g$ vs. $H_1: f\neq g$

Symmetry Underlying rank-sum test

Pool all observations as $\{Z_1 \dots Z_{n+m}\}$

ignoring their ordering

Given these pooled values $\{Z_1, \dots Z_{m+n}\}$ under H_0, its equaly likely for $X_1 \dots X_n\, Y_1 \dots Y_n$ to be any permutation of these values

Idea: For any test statistic $T(X_1, \dots X_n, Y_1 \dots Y_m)$

The permutation null distribution of T is its distribution upon:

1) Take a random permutation $X_1^* \dots X_n^*, Y_1^* \dots Y_m^*$

2) Compute $T(X_1^*, \dots X_n^*, Y_1^* \dots Y_m^*)$

↑ uniform distribution on all $(m+n)!$ possible possibilities

Can approximate this by simulation:

- Sample B uniformly random permutations
- Look at the distribution of the B values

A test of H_0 using T and this permutation null distribution is a permutation test

Permutation test is a conditional test

pivotal on H_0 conditional on $\{Z_1 \dots Z_{m+n}\}$

The test guarantees

$$\mathbb{P}_{H_0}[\text{reject } H_0 \mid \{Z_1 \dots Z_{m+n}\}] \leq \alpha \quad \text{for any set of } \{Z_1 \dots Z_{n+m}\}$$

This also holds unconditionally

$$\mathbb{P}_{H_0}[\text{reject } H_0] \leq \alpha$$

Example: Suppose $X_1 \dots X_n$ are objects and $Y_1 \dots Y_n$ are a second sample

Suppose we have definite distance $d(x, x')$ on $\mathcal{X}$

Here are 3 different statistics

- Average Distance Statistic

$$T_1 = 2\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} d(x_i, y_j) - \frac{1}{\binom{n}{2}}\sum_{\substack{i=1\\j=1}}^{n} d(x_i, x_j) - \frac{1}{\binom{m}{2}}\sum_{\substack{j=1\\i=1}}^{m} d(y_i, y_j)$$

- K-nearest neighbors

How many of its k-nearst numbers from its sample

T_2 = avg across all samples

- Freedmann - Rafety minimum spanning tree statistics
  - Constructing a minimum spanning tree from the pooled observations
  - Remove edges connected X with Y

T_3 = number of connected components

T_1, T_2, T_3 are not necessarly pivotal under H_0

Their full null distributions are complicated

Fisher's exact test

Randomly shuffle outcomes in table

Lecture 10 (2/23/22) - Effect size, power, and experimental design

Steps of a Scientific Study

1. Identify and formulate question of interest
2. Design experiment to answer this question
3. Visualize and explore collected data
4. Apply statistical procedure

Questions

- Predict in advance whether the study will succeed
- Predict size of study
- Why does experimental design influence our ability to ID this effect

<u>Case Study</u>: Stanford Peer Grading Experiment

Divide course into peer grading and control groups

Predict the power

Rejecting H_0 at desired level of significance

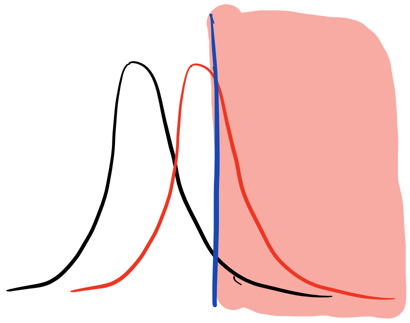
Power in the one-sample z-test

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

$H_0: \mu = 0$ vs. $H_1: \mu > 0$ assume σ^2 is known

Neyman - Pearson lemma tells us that the most powerful test rejects H_0 for large values of $\bar{X}$ ← z test

Null: $\bar{X} \sim N(0, \frac{\sigma^2}{n})$ Alternative: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$



Analytically: z-test rejects H_0 when $\frac{\sqrt{n}}{\sigma} \bar{X} > z(\alpha)$

this ensures type I error $= \alpha$

Under H_1: $\bar{X} \sim N(\mu, \sigma^2/n)$

$$\bar{X} = \mu + \frac{\sigma}{\sqrt{n}} Z \text{ where } Z \sim N(0,1)$$

$$\text{Power: } \mathbb{P}_{H_1}\left[\frac{\sqrt{n}}{\sigma} \bar{X} > z(\alpha)\right]$$

$$= \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}\left(\mu + \frac{\sigma}{\sqrt{n}} Z\right) > z(\alpha)\right]$$

$$= \mathbb{P}\left[Z > z(\alpha) - \sqrt{n}\frac{\mu}{\sigma}\right]$$

$$= \Phi\left(\sqrt{n}\frac{\mu}{\sigma} - z(\alpha)\right)$$

Power in comparing two Samples

Consider $X_1 \dots, X_n \sim N(\mu_x, \sigma^2)$ $Y_1 \dots Y_m \sim N(\mu_y, \sigma^2)$

$H_0: \mu_x = \mu_y$ $H_1: \mu_x > \mu_2$

Pooled two-sample

$$T = \frac{\bar{X}-\bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

Power = $\mathbb{P}_{H_1}\left[T > t_{m+n-2}(\alpha)\right]$

For large n, m

$$\approx \mathbb{P}_{H_1}\left[\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} > z(\alpha)\right]$$

Under H_1: $\bar{X}-\bar{Y} \sim N\left(\mu_X-\mu_Y, \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right)$

$$\bar{X}-\bar{Y} = \mu_X-\mu_Y+\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\cdot Z \quad \text{where } Z \sim N(0,1)$$

$$\text{Power} = \mathbb{P}\left[Z + \underbrace{\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\cdot\frac{\mu_X-\mu_Y}{\sigma}}_{d} > z(\alpha)\right] = \phi\left(d - z(\alpha)\right)$$

Power is increasing in

$$d = \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\cdot\frac{\mu_X-\mu_Y}{\sigma}$$

=> decreasing in $\frac{1}{n}+\frac{1}{m}$ so $\frac{1}{n}+\frac{1}{m}$ is minimized by

$$n = m = \frac{N}{2}$$

Predicting typical p-value

$$\text{p-value} = \mathbb{P}_{H_0}\left[T > t_{obs}\right] \approx 1-\phi(t_{obs})$$

Under H_1

$$t_{obs} \approx \frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} = Z + \frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}}\cdot\frac{\mu_X-\mu_Y}{\sigma}$$

median value of t_{obs} under H_1 is roughly d .

For $d=0.95$, this p-value is 0.17

Paired Design

split course into 2 units and swap groups between units

Consider paired differences

$$D_i = X_i - Y_i$$

If X_i, Y_i is bivariate normal then D_i has a normal distribution

$$\mathbb{E}(D_i) = \mu_X - \mu_Y$$

$$\text{Var}(D_i) = \text{Cov}(X_i - Y_i, X_i - Y_i) = 2\sigma^2(1-\rho)$$

Reduces problem to one-sample testing problem

Level α t-test

$$\frac{\sqrt{n}}{s}\bar{D} > t_{n-1}(\alpha)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i-\bar{D})^2$$ is sample variance

Difference between paired and unpaired is a

$\frac{1}{1-p}$ term

$1-p$ is the relative efficiency of the unpaired design to the paired design

paired test with n pairs has the same power as an unpaired design with sample size $\frac{n}{1-p}$ per group

Confounding variables

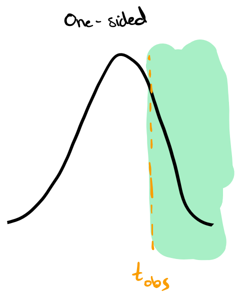
- Systematic biases
- Inflate variance

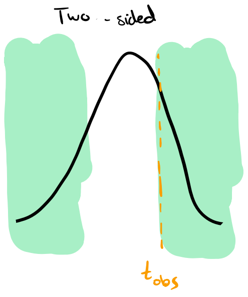
Lecture 11: Testing Multiple Hypotheses

If I test n null hypotheses at level α, all of which are true, then on average I'll falsely reject αn of them

Most statistical multiple-testing procedures take p-values as inputs

p-value is the smallest significance value at which the test rejects the null hypothesis





p-val is 0.036 (smallest value that rejects null)

p-value is 0.072

For a one-sided test with continuous test statistic T, we reject H_0 when T exceeds upper-α point of its null distribution

$$P = \mathbb{P}_{H_0}[T > t_{obs}] = 1 - F(t_{obs})$$

For a two-sided test, we reject H_0 when T is larger than upper-$\frac{\alpha}{2}$ or less than lower-$\frac{\alpha}{2}$

$$P = 2\cdot\min\left(F(t_{obs}), 1 - F(t_{obs})\right)$$

p-value is the probability under H_0 of observing a value of T that is more extreme than t_{obs}.

p-value can be considered as a test statistic itself

reject if $P \le \alpha$

P is uniform $(0,1)$ under H_0

Bonferroni Method:

For n different null hypothses, reject at α/n instead of α

Justification:

$$\mathbb{P}[\text{reject any null hypothesis}] = \mathbb{P}\left[\{\text{reject } H_0^{(1)}\} \cup \dots \cup \{\text{reject } H_0^{(n)}\}\right]$$

$$\le \mathbb{P}[\text{reject } H_0^{(1)}] + \dots + \mathbb{P}[\text{reject } H_0^{(n)}] \le \frac{\alpha}{n} + \dots + \frac{\alpha}{n} = \alpha$$

Family Wise Error Rate

n null hypotheses, n_0 are true nulls

$$\text{FWER} = \mathbb{P}[\text{reject any true null hypothesis}]$$

Controls FWER at level α guarantees FWER $\le \alpha$

Bonferroni Method controls FWER at α

False Discovery Rate

$$\text{False Discovery Proportion} = \frac{\text{number of true null hypotheses rejected}}{\text{number of total null hypotheses rejected}} = \frac{V}{R}$$ ← 0 when $V=0$ and $R=0$

Controls FDR at level α if FDR $\le \alpha$

Controlling FWER is more appropriate if the consequences of a single Type I error is high
result will be interpretted as truth

Controlling FDR is more appropriate if the test IDs candidate discoveries for further study
if false discoveries are acceptable as long most of the discoveries are correct

Estimating $\text{FDP} = \frac{V}{R}$ but we don't know V

We can estimate V since p-values are uniformly distributed $(0,1)$.

For a rejection value of t, we can expect tn_0 of the true nulls to have $p \le t$

$$V \approx tn_0$$

We don't know n_0 so

$$\widehat{\text{FDP}} = \frac{tn}{R}$$ ← estimate

Control $\widehat{\text{FDP}} = \frac{tn}{R(t)} \le \alpha$

number of rejections

goal is to find maximum t that satisfies this relation

For r rejections, $t = P_{(r)}$, the r^{th} smallest p-value and find largest r s.t.

$$\frac{P_{(r)}}{r} \cdot n \le \alpha \iff P_{(r)} \le \frac{\alpha r}{n}$$ ← Benjamini - Hochberg

1. Sort n p-values from smallest to largest
2. Find largest r s.t. $P_{(r)} < \frac{\alpha r}{n}$
3. Reject null hypothesis corresponding to $P_{(1)} \dots P_{(r)}$

Notice: $P_{(1)}$ is compared to $\frac{\alpha}{n}$, $P_{(2)} = \frac{2\alpha}{n}$, etc.

3/2 : Parametric models and the method-of-moments

Def: A parametric model is a family of distributions indexed by a small number of unknown parameters

e.g. $N(\mu,\sigma^2)$

Notation: Vector of parameters $\theta \in \mathbb{R}^k$

PDF / PMF as $f(x|\theta)$

^ not conditional

The set of allowable parameters is the parameter space

How to choose model to fit?

- What the data represents
- How the data arose?
- Visual examination of the data
- Considerations of complexity

Suppose we observe $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$

- How can we estimate θ
- How can we quantify our uncertainty

The Method of Moments

Suppose $\theta \in \mathbb{R}$ is a single unknown parameters

Pick θ so that the mean of $f(x|\theta)$ matches sample mean

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

Ex. The poisson distribution for parameter $\lambda > 0$ is a discrete distribution over non-negative integer counts

$$\text{PMF: } f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{for } x \in \{0,1,2,\dots\}$$

This distribution has mean λ. The M-O-M estimate is $\hat{\lambda} = \bar{X}$

Ex. The exponential distribution $\lambda > 0$

$f(x|\lambda) = \lambda e^{-\lambda x}$ mean: $\frac{1}{\lambda}$

equate $\frac{1}{\hat{\lambda}} = \bar{X}$

More generally, suppose the parameters are $\Theta \in \mathbb{R}^k$. Equating the theoretical mean w/ sample avg. gives 1 equation

To get k equations, consider the first k moments of $X \sim f(x|\theta)$

$$\left.\begin{aligned} \mu &= \mathbb{E}[X] \\ \mu_2 &= \mathbb{E}[X^2] \\ &\vdots \\ \mu_k &= \mathbb{E}[X^k] \end{aligned}\right\} \text{depend on } \Theta$$

M-O-M estimate

① Compute $\mu_1 \dots \mu_k$ in terms of Θ

② Equate theoretical moments w/ sample moments

③ Solve for Θ

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

$$\mu_1 = \mathbb{E}[X] = \mu$$

$$\mu_2 = \mathbb{E}[X^2] = \mathrm{Var}[X] + (\mathbb{E}X)^2 = \sigma^2 + \mu^2$$

Set $\hat{\mu} = \frac{1}{n}(X_1 + \dots X_n)$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\left(X_1^2 + \dots X_n^2\right)$$

Ex. $X_1 \dots X_n \overset{iid}{\sim} \text{Gamma}(\alpha, B)$. Recall Mean in $\frac{\alpha}{B}$

variance is $\frac{\alpha}{B^2}$

$$\mu_1 = \frac{\alpha}{B} \qquad \mu_2 = \frac{\hat{\alpha} + \hat{\alpha}^2}{\hat{B}}$$

Bias, Variance, and mean-squared error

Any estimate $\hat{\theta}$ for $\theta \in \mathbb{R}^k$ is a statistic.

The bias of $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta}] - \theta$, where $\mathbb{E}_\theta$ means expectation when the parameter is θ.

"whats the difference between average value of $\hat{\theta}$ and θ"

Standard error is just standard deviation of $\hat{\theta}$

$\sqrt{\text{Var}_\theta(\hat{\theta})}$

"How far is $\hat{\theta}$ typically from its average value"

Mean-squared-error of $\hat{\theta}$ is $\mathbb{E}[(\hat{\theta} - \theta)^2]$

MSE combines bias and standard error

MSE = Variance + Bias^2

Usually, MSE, bias, and variance all depend on θ

An estimator is said to be unbiased if bias = $\mathbb{E}_\theta[\hat{\theta}] - \theta = 0$ for every possible θ in the parameter space

Ex. Consider $X_1 \dots X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$

Recall M-O-M estimate for $\hat{\lambda} = \bar{X}$

① Bias : $\mathbb{E}_\lambda(\hat{\lambda}) = \mathbb{E}_\lambda\left[\frac{1}{n}(X_1 + \dots + X_n)\right] = \frac{1}{n}\left(\mathbb{E}(X_1) + \dots \mathbb{E}(X_n)\right) = \lambda$

so bias $\mathbb{E}_\lambda(\hat{\lambda}) - \lambda = 0$ $\hat{\lambda}$ is unbiased for λ

② Standard error : $\text{Var}_\lambda[\hat{\lambda}] = \text{Var}_\lambda\left[\frac{1}{n}(X_1 + \dots X_n)\right] = \frac{\lambda}{n}$

$$s.e. = \sqrt{\lambda/n}$$

③ $MSE = (s.e.)^2 + (\text{bias})^2 = \dfrac{\lambda}{n}$

3/7: Maximum Likelihood Estimator

Recap: Parametric model $f(X|\theta)$ parametrized by $\theta \in \mathbb{R}^k$

Estimate θ via method-of-moments

Compute $\mu_j = \mathbb{E}[X^j]$ in terms of θ for $j = 1 \dots k$

Equate values to sample values and solv

Def: The joint PMF or PDF of data $X_1 \dots X_n$ viewed as a function of the parameter $\theta \in \mathbb{R}^k$ is the likelihood function $\text{lik}(\theta)$

E.g. If $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$ then

$$\text{lik}(\theta) = f(X_1|\theta) \times \dots \times f(X_n|\theta)$$

Recall from NP

$H_0: \mathbf{X} \sim f_0$ vs. $H_1: \mathbf{X} \sim f_1$

Likelihood ratio statistic $L(\mathbf{X}) = \dfrac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$

In the context of parametric models, if

$$f_0(X_1 \dots X_n) = f(X_1|\theta_0) \times \dots \times f(X_n|\theta_0)$$

$$f_1(X_1 \dots X_n) = f(X_1|\theta_1) \times \dots \times f(X_n|\theta_1)$$

then $L(\mathbf{X}) = \dfrac{\text{lik}(\theta_0)}{\text{lik}(\theta_1)}$

Maximum Likelihood Estimator of θ is the value of θ in the parameter space that maximizes $\text{lik}(\theta)$

Examples

How to compute MLE $\hat{\theta}$?

Note: It's equivalent to maximize log likelihood

$$\ell(\theta) = \log \text{lik}(\theta)$$
$$= \sum_{i=1}^{n} \log f(X_i|\theta) \quad \text{if } X_1 \dots X_n \text{ are iid}$$

Let $X_1 \dots X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, Parametric space: $\lambda \in (0, \infty)$

The PMF is $f(x|\lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

Log Likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\lambda) = \sum_{i=1}^{n} -\lambda + X_i \log(\lambda) - \log(X_i!)$$

$$= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

To maximize λ, we want $0 = \ell'(\lambda)$

$$\ell'(\theta) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

First check: $\hat{\lambda}$ is in the parameter space

$\hat{\lambda} = \bar{X}$ is in $(0,\infty)$ not all $X_i = 0$

Second check: $\hat{\lambda}$ is a maximizer of the functio

$0 > \ell'(\lambda) \quad \lambda > \bar{X}$ and $0 < \ell'(\ell)$

$\Updownarrow \qquad\qquad \Updownarrow$

$\hat{\lambda} > \hat{\lambda} = X \qquad \lambda < \hat{\lambda} = \bar{X}$

$\ell(\lambda)$ is decreasing left of $\bar{x}$ and increasing right of $\bar{X}$

Alternatively check if $\ell''(\lambda) < 0$

$$\ell''(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i$$

This is negative for all $\lambda \in (0,\infty)$ so $\ell(\lambda)$ is concave

If all X_i's are 0, our estimate is $\hat{\lambda} =$

Example: Let $X_1 \dots X_n \overset{iid}{\sim} N(\mu,\sigma^2)$. The log likelihood function is

$$\ell(\mu,\sigma^2) = \sum_{i=1}^{n} \log f(X_i|\mu,\sigma^2)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}\right)$$

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

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

Maximize: $\ell(\mu,\sigma^2)$

First order condition

$\ell'(\theta) = 0$

$$\frac{\partial\ell}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) \iff 0 = \sum_{i=1}^{n} X_i - n\mu \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\frac{\partial\ell}{\partial\sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 \iff 0 = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i-\bar{X})^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Verifying Maximization

$0 < \frac{\partial\ell}{\partial\mu} \iff \mu < \bar{X}$ and $0 > \frac{\partial\ell}{\partial\mu} \iff \mu > \bar{X}$

Similarly for σ^2 by subbing $\mu = \bar{X}$

Example: Let $X_1 \dots X_n \overset{iid}{\sim} \text{Gamma}(\alpha,\beta)$ $\alpha,\beta>0$

Log Likelihood

$$\ell(\alpha,\beta) = \sum_{i=1}^{n} \log f(x_i|\alpha,\beta) = \sum_{i=1}^{n} \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}\right) = \sum_{i=1}^{n} \alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log(x_i) - \beta x_i$$

$$= n\alpha\log\beta - n\log(\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^{n}\log(x_i) - \beta\sum_{i=1}^{n} x_i$$

Maximize

$$\frac{\partial\ell}{\partial\alpha} = n\log\beta - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\log x_i$$

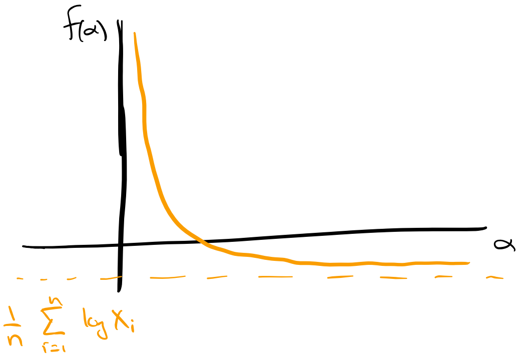
$$\frac{\partial\ell}{\partial\beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i$$

Solve 2nd equation in β so $\hat\beta = \hat\alpha/\bar X$

Substituting into first equation

$$0 = n\log\left(\frac{\hat\alpha}{\bar X}\right) - \frac{n\Gamma(\hat\alpha)}{\Gamma(\hat\alpha)} + \sum_{i=1}^{n}\log(x_i$$

$$0 = \underbrace{\log(\hat\alpha) - \frac{\Gamma(\hat\alpha)}{\Gamma(\hat\alpha)} - \log\bar X + \frac{1}{n}\sum_{i=1}^{n}\log x_i}_{f(\hat\alpha}$$



Observe $f(\alpha)$ is decreasing

$$-\log(\bar X) + \frac{1}{n}\sum_{i=1}^{n}\log(x_i) < 0$$

by Jensen's Inequality

So there must be a unique root to $0 = f(\hat\alpha)$

Then this $\hat\alpha$ and $\hat\beta = \frac{\hat\alpha}{\bar x}$ are the MLE's

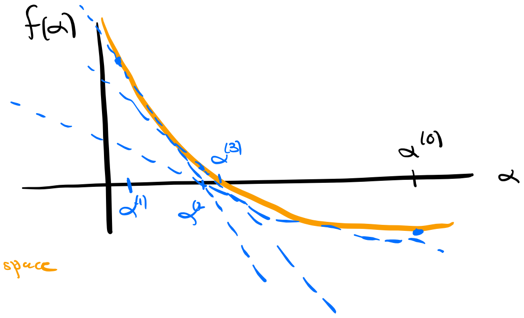
You can check that these are the maximizers

Usually no explicit form for MLE, Instead compute via optimization algorithm

Newton-Raphson method is a common approach

1. Initialize with a guess

Often use MOM estimate



2. linearize $\alpha^{(t)}$ using a Taylor expansion

$$0 \approx f(\alpha^t) + (\hat{\alpha} - \alpha^t) f'(\alpha^t)$$

$$\hat{\alpha} \approx \alpha^t - \frac{f(\alpha^t)}{f'(\alpha^t)}$$ ← reset α^t if α^t falls outside of the parameter space

3. Use linearized estimate for next iteration

This is different from M-O-M estimator

Example: Let $X_1 \dots X_K \sim \text{Multinomial}(n, (p_1 \dots p_K))$

$X_1 \dots X_K$ are not IID, they sum to n

Log likelihood function is the log PMF for $(X_1 \dots X_K)$

$$\ell(p_1 \dots p_K) = \log\left[\binom{n}{X_1, X_2, \dots X_K} p_1^{X_1} \cdot p_2^{X_2} \cdot \ \dots \ p_K^{X_K}\right]$$

$$= \log\binom{n}{X_1 \dots X_K} + \sum_{n=1}^{K} X_i \log p_i$$

The parameter space is the set $(p_1 \dots p_K)$ where $0 \le P_i \le 1$ and $P_1 + \dots + P_K = 1$

MLE maximizes $\ell(p_1 \dots p_K)$ subject to these constraints

Easiest way to do this is via Lagrange Multiplier Method

$$L(p_1 \dots p_K, \lambda) = \ell(p_1, \dots, p_K) + \overset{\text{Lagrange Multiplier}}{\lambda} (p_1 + \dots + P_K - 1)$$

$$= \log\binom{n}{X_1 \dots X_K} + \sum_{i=1}^{K} X_i \log(p_i) + \lambda(p_1 + \dots P_K - 1)$$

Set all partials of L to 0

$$0 = \frac{\partial L}{\partial p_i} = \frac{X_i}{P_i} + \lambda$$

$$0 = \frac{\partial L}{\partial \lambda} = P_1 + \dots + P_K - 1$$

Solve 1st eq to find $p_i = \frac{-X_i}{\lambda}$

Substitute into 2nd eqn:

$$0 = \frac{-X_1}{\lambda} - \frac{X_2}{\lambda} - \dots \frac{X_K}{\lambda} - 1$$

$$\lambda = -(X_1 + \dots X_K) = -n$$

$$\hat{p}_i = \frac{X_i}{n}$$ (fraction of observations in class i)

Rationale for Lagrange Multiplier

1) Fix any λ. Maximizing $L(p_1 \dots p_K, \lambda)$ subject to $P_1 + \dots P_K = 1$ is equivalent to maximizing $\ell(p_1 \dots p_K)$ under the same constraint

λ term cancels to 0

2) The unconstrained maximizer of $L(p_1 \dots p_k, \lambda)$ over $(p_1 \dots p_k)$ is $p_i = -\frac{X_i}{\lambda}$ as calalated above

3) Choosing $\lambda = -(X_1 + \dots X_k) = -n$, the unconstrained max satisfies the constraint

Implies that the solved p_i's are also constrained maximizers

from 1) we have that this solution maximizes $\ell(p_1 \dots p_n)$

Example: The genotypes AA, Aa, aa at a locus satisfy Hardy-Weinberg Equilibrium

Occur with probabilities

$(1-\theta)^2$, $2\theta(1-\theta)$ and θ^2

$\theta = [0,1]$ is the frequency of a

Count occurances of AA, Aa, aa in n samples can be modeled as multinomial

$$(X_1, X_2, X_3) \sim \text{Multinomial}\left(n, \left[(1-\theta)^2, 2\theta(1-\theta), \theta^2\right]\right)$$

Similar to previous example with single parameter $\theta = [0,1]$

Compute MLE for θ

$$\ell(\theta) = \log\left[\binom{n}{X_1, X_2, X_3} \left((1-\theta)^2\right)^{X_1} \left(2\theta(1-\theta)\right)^{X_2} \left(\theta^2\right)^{X_3}\right]$$

$$= \log \binom{n}{X_1 X_2 X_3} + (2X_1 + X_2)\log(1-\theta) + (X_2 + 2X_3)\log\theta$$

Maximize over θ (constraint already accounted for)

$$0 = \ell'(\theta) = -\frac{2X_1 + X_2}{1-\theta} + \frac{X_2 + 2X_3}{\theta}$$

$$\hat{\theta} = \frac{2X_3 + X_2}{2n}$$

3/9: Normal Approximation, confidence Interval

Example: Poisson Model

$X_1 \dots X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Let λ_0 be the true parameter

Both M-O-M and MLE find $\hat{\lambda} = \bar{X}$

From lecture 12: $\mathbb{E}_{\lambda_0}[\hat{\lambda}] = \lambda_0$ unbiased estimator

$\text{Var}_{\lambda_0}[\hat{\lambda}] = \frac{\lambda_0}{n}$ Standard error of $\sqrt{\frac{\lambda_0}{n}}$

By LLN: $\hat{\lambda} = \bar{X} \to \lambda_0$ in probability as $n \to \infty$

We say $\hat{\lambda}$ is consistent for λ

By CLT: $\sqrt{n}(\hat{\lambda} - \lambda_0) \to N(0, \lambda_0)$ in distribution as $n \to \infty$

Informally, $\hat{\lambda}$ has approximate distribution $N(\lambda_0, \frac{\lambda_0}{n})$ for large n

This allows us to make a confidence interval for λ_0

random interval containing λ_0 w/ a pre-specifies probability $1-\alpha \in (0,1)$

Let $z(\alpha/2)$ be the upper-α point of $N(0,1)$

CLT implies $\mathbb{P}\left[-\sqrt{\frac{\lambda_0}{n}}\, z(\alpha/2) \le \hat{\lambda}-\lambda_0 \le \sqrt{\frac{\lambda_0}{n}}\, z(\alpha/2)\right] \approx 1-\alpha$

Substituting $\sqrt{\hat{\lambda}/n}$ for $\sqrt{\lambda_0/n}$ we find

$$\mathbb{P}\left[-\sqrt{\frac{\hat{\lambda}}{n}}\, z(\alpha/2) \le \hat{\lambda}-\lambda_0 \le \sqrt{\frac{\hat{\lambda}}{n}}\, z(\alpha/2)\right] \approx 1-\alpha$$

We have lower and upper bounds of λ_0

$\Rightarrow \lambda_0$ belongs to the interval $\hat{\lambda} + \sqrt{\frac{\hat{\lambda}}{n}} \cdot z(\alpha/2)$ w/ probability $1-\alpha$

More formally: The coverage guarantee is

$$\mathbb{P}_{\lambda_0}\left[\lambda_0 \in \left[\hat{\lambda} - \sqrt{\hat{\lambda}/n}\cdot z(\alpha/2,\ \hat{\lambda} + \sqrt{\hat{\lambda}/n}\cdot z(\alpha/2\right]\right] \longrightarrow 1-\alpha \text{ as } n\to\infty$$

Because

- $\sqrt{n}(\hat{\lambda}-\lambda_0) \longrightarrow \mathcal{N}(0,\lambda_0)$ by CLT
- $\frac{1}{\sqrt{\hat{\lambda}}} \longrightarrow \frac{1}{\sqrt{\lambda_0}}$ by LLN
- $\frac{\sqrt{n}(\hat{\lambda}-\lambda_0)}{\sqrt{\hat{\lambda}}} \longrightarrow \frac{1}{\sqrt{\lambda_0}}\cdot\mathcal{N}(0,\lambda_0) = \mathcal{N}(0,1)$ by slutsky's Lemma

Asymptotic Normality of the MLE

Thm: Let $f(x|\theta)$ be a parametric model, where $\theta\in\mathbb{R}$. Let θ_0 be the true parameter, let $X_1,\dots X_n \overset{iid}{\sim} f(x|\theta_0)$.

Let $\hat{\theta}$ be the MLE.

Under some smoothness conditions for $f(x|\theta)$, as $n\to\infty$

a) $\hat{\theta}$ is a consistent estimator

$\hat{\theta} \longrightarrow \theta_0$ in probability as $n\to\infty$

b) $\hat{\theta}$ is asymptotic normal, and $\sqrt{n}(\hat{\theta}-\theta_0) \longrightarrow \mathcal{N}(0, 1/I(\theta_0))$ in distribution

$I(\theta)$ has two definitions

Single sample

$$I(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$$

- $I(\theta)$ is the Fisher Information
- $\frac{\partial}{\partial\theta}\log f(x|\theta)$ is the score

We will show $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] = 0$

so $I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right]$

- Implications
  - Asymptotically $\hat{\theta}$ is unbiased. The bias is smaller scale than $1/\sqrt{n}$

$\mathbb{E}[\sqrt{n}(\hat{\theta}-\theta_0)) \longrightarrow 0$, if bias was larger the constant term of the bias would be $\mathbb{E}$ of expression

- For large n, the standard error of the MLE is

$$S.E.\ \hat{\theta} = \sqrt{1/nI(\theta)}$$. Is on the order of $1/\sqrt{n}$

MSE = $(\text{bias})^2 + (\text{std. error})^2$ is dominated by std. error for large n

- Distribution of $\hat{\theta}$ is $\approx N(\theta_0, \frac{1}{nI(\theta_0)})$
- Confidence interval for θ_0 with coverage $1-\alpha$

$$\hat{\theta} \pm \sqrt{\frac{1}{nI(\hat{\theta})}} \cdot z(\alpha/2) \quad , \quad \sqrt{\frac{1}{I(\hat{\theta})}} \text{ estimates } \sqrt{\frac{1}{I(\theta_0)}}$$

Example: Consider again $X_1 \dots X_n \overset{iid}{\sim}$ Poisson (λ_0)

MLE is $\hat{\lambda} = \bar{X}$

$$\log f(X|\lambda) = \log \frac{e^{-\lambda}\lambda^{x}}{x!} = -\lambda + \log \lambda - \log(x!)$$

$$\frac{\partial}{\partial \lambda} \log f(X|\lambda) = -1 + \frac{X}{\lambda} \quad \leftarrow \text{Score}$$

$$\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda) = -\frac{X}{\lambda^2}$$

$$E_{\lambda_0}\left[\frac{\partial}{\partial \lambda} \log f(X|\lambda)\right] = E_{\lambda_0}\left[-1 + \frac{X}{\lambda}\right] = 0$$

$$I(\lambda_0) = \text{Var}_{\lambda_0}\left[-1 + \frac{X}{\lambda_0}\right] = \text{Var}\left[\frac{X}{\lambda_0}\right] = \frac{1}{\lambda_0^2}\text{Var}[X] = \frac{1}{\lambda_0}$$

Alternatively,

$$I(\lambda_0) = -E_{\lambda_0}\left[-X/\lambda_0^2\right] = -\frac{1}{\lambda_0^2}E[-X] = \frac{1}{\lambda_0}$$

Theorem shows $\sqrt{n}(\hat{\lambda} - \lambda_0) \to N(0, \frac{1}{I(\lambda_0)}) = N(0, \lambda_0)$

Proof Sketch

Consistency: The MLE $\hat{\theta}$ maximizes

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X|\theta)$$

Suppose θ_0 is the true parameter. For any fixed θ, this is the average of n IID random variables.

As $n \to \infty$, by LLN

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X|\theta) \to E_{\theta_0}\left[\log f(X_i|\theta)\right]$$

Implies under some conditions that the maximizer of LHS maximizes RHS

Maximizer of RHS is θ_0

$$E_{\theta_0}[\log f(X_i|\theta)] - E_{\theta_0}[\log f(X_i|\theta_0)] = E_{\theta_0}\left[\log \frac{f(X_i|\theta)}{f(X_i|\theta_0}\right] \le \log E_{\theta_0}\left[\frac{f(X_i|\theta)}{f(X_i|\theta_0)}\right]$$

Jensen's Inequality since log is concave

$$= \log \int \frac{f(x|\theta)}{f(x|\theta_0)} \cdot f(x|\theta_0)\,dx = \log \overbrace{\int f(x|\theta)\,dx}^{\text{PMF}} = \log 1 = 0$$

$$\Rightarrow \quad \mathbb{E}_{\theta_0}[\log f(x_i|\theta)] - \mathbb{E}_{\theta_0}[\log f(x_i|\theta_0)] \le 0$$

So this is maximized over θ at $\theta = \theta_0$

Fisher Information

$$\int f(x|\theta)\,dx = 1$$

$$\Rightarrow \quad 0 = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x|\theta)\,dx$$

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} \iff \frac{\partial}{\partial\theta} f(x|\theta) = \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)$$

$$\Rightarrow \quad 0 = \int\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\cdot f(x|\theta)\right)dx = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]$$

So the score has mean 0.

Differentiate a second time

$$0 = \frac{\partial}{\partial\theta}\int\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\,dx$$

product rule

$$= \int\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right)\cdot f(x|\theta)\,dx + \int\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)\underbrace{\left(\frac{\partial}{\partial\theta} f(x|\theta)\right)}_{\frac{\partial}{\partial\theta}\log f(x|\theta)\cdot f(x|\theta)}dx$$

$$= \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] + \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right]$$

$$\Rightarrow \quad -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right] = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right.$$

Asymptotical Normality

$$\sqrt{n}(\hat\theta - \theta_0) \to N(0, 1/I(\theta_0))$$

The MLE $\hat\theta$ maximizes $\ell(\theta) = \sum_{i=1}^{n}\log f(x|\theta)$

so $0 = \ell'(\hat\theta) \approx \ell'(\theta_0) + (\hat\theta - \theta)\cdot\ell''(\theta_0)$ ← Taylor Expansion

$$\hat\theta - \theta_0 \approx -\frac{\ell'(\theta_0)}{\ell''(\theta_0)}$$

$$\sqrt{n}(\hat\theta - \theta_0) \approx -\sqrt{n}\,\frac{\ell'(\theta_0)}{\ell''(\theta_0)} = \frac{-\frac{1}{\sqrt{n}}\,\ell'(\theta_0)}{-\frac{1}{n}\,\ell''(\theta_0)}$$

$$-\frac{1}{n}\,\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta_0)\right]$$

iid w/ mean $-\mathbb{E}_0\left[\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta)\right] = I(\theta_0)$ in pr

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \frac{1}{\sqrt{n}}\underbrace{\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i|\theta_0)}_{}$$

iid w/mean $=0$

$$\text{Var}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(x_i-\theta_0)\right] = I(\theta_0)$$

$\to N(0, I(\theta_0))$ by CLT

By Slutsky Lemma

$$\sqrt{n}(\hat{\theta}-\theta) \to \frac{1}{I(\theta_0)}\cdot N(0, I(\theta_0)) = N(0, 1/I(\theta))$$

3/14: Plug-in Estimates, delta Method

Recap: $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$, $\theta\in\mathbb{R}^k$

① Method of Moments

$$\text{Write } \mu = h(\theta) \begin{cases} \mu_1 = \mathbb{E}_\theta[X] = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \vdots \\ \mu_k = \mathbb{E}_\theta[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{cases}$$

Solve equations for θ

② Maximum Likelihood Estimation

$$\ell(\theta) = \sum_{i=1}^{n}\log f(x|\theta)$$

Maximize $\ell(\theta)$ over the parameter space to get $\hat{\theta}$

Thm: The MLE $\hat{\theta}$ is consistent for θ, as $n\to\infty$

$$\sqrt{n}(\hat{\theta}-\theta) \to N\left(0, \frac{1}{I(\theta)}\right) \quad \text{for } \theta\in\mathbb{R}$$

$$I(\theta) = \text{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] = -E_0\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$$

Asymptotic Confidence Interval level $1-\alpha$ for θ

$$\hat{\theta} \pm \sqrt{\frac{1}{nI(\theta)}}\cdot z(\alpha/2)$$

Estimating a function of θ

First estimate θ by $\hat{\theta}$, then use $g(\hat{\theta})$ as an estimate for $g(\theta)$

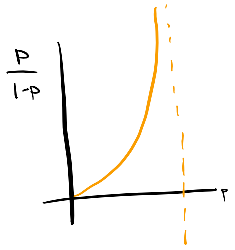
Example: Binomial coin, heads prob is p

If heads lose \$1
If tails win \$X

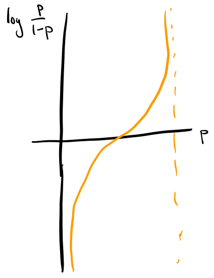
What value of X makes this game fair

Expected winnings : $p\cdot(-1) + (1-p)X = 0$

$X = \frac{p}{1-p}$ is called the odds of this game



Often consider log-odds or logit



Estimate $\log\frac{p}{1-p}$ from $X_1,\dots X_n \overset{iid}{\sim}$ Bernoulli (p)

① First estimate p, by $\bar{X}$

② Plug-in to get $\log\frac{\bar{X}}{1-\bar{X}}$ as an estimate of $\log\frac{p}{1-p}$

Example: Pareto Distribution

PDF: $f(X|x_0,\theta) = \theta x_0^{\theta} X^{-\theta-1}$ for $X \geq x_0$

Suppose we know $x_0=1$, we don't know θ, so

$$f(X|\theta) = \theta X^{-\theta-1} \quad \text{for } X \geq 1$$

Mean: $\frac{\theta}{\theta-1}$ when $\theta > 1$

Variance : $\frac{\theta}{(\theta-1)^2(\theta-2)}$ when $\theta > 2$

How to estimate mean from $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$

① Estimate θ

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x|\theta) = \sum_{i=1}^{n} \log\theta - (\theta+1)\log x_i$$

$$= n\cdot\log\theta - (\theta+1)\sum_{i=1}^{n}\log x_i$$

To maximize this over θ

$$0 = \ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}\log x_i$$

$$\Rightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n}\log X_i}$$ is the MLE for θ

② Estimate the mean $\frac{\theta}{\theta-1}$ by $\frac{\hat{\theta}}{\hat{\theta}-1}$

Delta Method

Goal: Quantify the uncertainity (asymptotically) for $g(\hat{\theta})$ based on uncertainity of $\hat{\theta}$ itself

Thm: If $g:\mathbb{R}\to\mathbb{R}$ is continuously differentiable at θ, and if $\sqrt{n}(\hat{\theta}-\theta) \to N(0,v(\theta))$ in distribution as $n\to\infty$, then

$$\sqrt{n}\left(g(\hat{\theta})-g(\theta)\right) \to N\left(0, g'(\theta)^2\cdot v(\theta)\right)$$ in distribution as $n\to\infty$

Proof sketch: Apply Taylor expansion of $g(\hat{\theta})$ around $\hat{\theta} = \theta$

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

Then, $\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \approx \sqrt{n}\left(\hat{\theta} - \theta\right) g'(\theta)$

$$\rightarrow g'(\theta) N(0, v(\theta) = N(0, g'(\theta)^2 \cdot v(\theta))$$

Example: Let $X_1 \dots X_n \overset{iid}{\sim}$ Bernouilli (p)

We estimate $\log \frac{p}{1-p}$ by $\log \frac{\bar{X}}{1-\bar{X}}$

Apply Delta method:

① By CLT: $\sqrt{n}(\bar{X} - p) \rightarrow N(0, p(1-p))$

② Let $g(p) = \log \frac{p}{1-p} = \log p - \log 1-p$

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

So, $\sqrt{n}\left(\log \frac{\bar{X}}{1-\bar{X}} - \log \frac{p}{1-p}\right) \rightarrow N\left(0, \frac{1}{p(1-p)}\right)$

Suppose we toss $n=100$ coins with 60 heads

$\bar{X} = 0.6$

log odds $\approx \frac{0.6}{1-0.6} = 0.41$

Standard error $= \sqrt{\frac{1}{n\bar{X}(1-\bar{X})}} \approx 0.2$

Asymptotic Level $1-\alpha$, confidence interval is $0.41 \pm 0.2 \cdot z(\frac{\alpha}{2})$

Back to Pareto Example

Let $X_1 \dots X_n \overset{iid}{\sim}$ Pareto $(1, \theta)$

Recall: MLE is $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i}$

Plug in estimate for mean $\frac{\theta}{\theta - 1}$ was $\frac{\hat{\theta}}{\hat{\theta} - 1}$

Apply delta method:

① Understand $\hat{\theta}$

$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, 1/I(\theta))$

Computing $I(\theta)$:

$$\log f(X|\theta) = \log\theta - (\theta+1)\log X$$

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = \frac{1}{\theta} - \log X$$

$$\frac{\partial^2}{\partial\theta^2}\log f(X|\theta) = -\frac{1}{\theta^2}$$

$$I(\theta) = -\mathbb{E}_\theta\left[-1/\theta^2\right] = \frac{1}{\theta^2}$$

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \to N(0,\theta^2)$$

② Set $g(\theta) = \dfrac{\theta}{\theta-1}$

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

so, $$\sqrt{n}\left(\frac{\hat{\theta}}{\hat{\theta}-1} - \frac{\theta}{\theta-1}\right) \to N\left(0, \left[-\frac{1}{(\theta-1)^2}\right]^2\theta^2\right) = N\left(0, \frac{\theta^2}{(\theta-1)^4}\right)$$

An alternative estimate for the mean is $\bar{X}$

For $\bar{x}$ estimate, we can apply clt

$$\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta-1}\right) \to N\left(0, \frac{\theta}{(\theta-1)^2(\theta-2)}\right)$$

Notice variance from MLE method is less than $\bar{x}$ estimate

Reduces impact of extreme observations

Standard Error for Method of Moments

Consider $\theta\in\mathbb{R}$. Estimate $\mu = \mathbb{E}_\theta[X]$ by $\bar{X}$

Suppose $\mu = h(\theta)$ for some function $h()$

Let g be the inverse function of h so $\theta = g(\mu)$

MOM estimate for $\theta = g(\mu)$ is $\hat{\theta} = g(\bar{x})$

For Standard Error

① By CLT, $\sqrt{n}(\bar{X}-\mu) \to N(0, v(\theta))$

$$v(\theta) = \mathrm{var}_\theta[X_i]$$

② By delta method

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\left(g(\bar{X}) - g(h(\theta))\right) \to N\left(0, g'(h(\theta))^2 \cdot v(\theta)\right)$$

Standard error of $\hat{\theta}$ is

$$\sqrt{\frac{g'(h(\theta))^2 \cdot v(\hat{\theta})}{n}}$$

Example: Consider $X_1, \dots X_n \overset{iid}{\sim}$ Exponential (λ)

PDF: $f(x|\lambda) = \lambda e^{-\lambda x}$ for $x > 0$

mean: $1/\lambda$ var: $1/\lambda^2$

Inverse of $h()$ is $\lambda = g(\mu) = 1/\mu$

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

Estimate std. error

① By CLT, $\sqrt{n}(\bar{x} - 1/\lambda) \to N(0, 1/\lambda^2)$

② $g'(\mu) = -1/\mu^2$ so $g'(1/\lambda) = -\lambda^2$

$$\sqrt{n}\left(1/\bar{x} - \lambda\right) = \sqrt{n}\left(g(\bar{x}) - g(1/\lambda)\right) \to N\left(0, g'(1/\lambda)^2 \cdot 1/\lambda^2\right) = N(0, \lambda^2)$$

Therefore std. error is $\approx \sqrt{\frac{\lambda^2}{n}}$ for large n, can estimate via $\sqrt{\frac{1}{n\bar{x}^2}}$

3/16: Cramer - Rao Bound, Asymptotic Efficiency

Recap: Pareto Model

PDF: $f(x|\theta) = \theta x^{-\theta-1}$ for $x \geq 1$

$X_1, \cdots X_n \overset{iid}{\sim} f(x|\theta)$

MLE $\hat{\theta}$: $\frac{n}{\sum_{i=1}^{n} \log x_i}$, $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \theta^2)$

Mean in this model: $\mu = \frac{\theta}{\theta - 1}$, $\sqrt{n}\left(\frac{\hat{\theta}}{\hat{\theta}-1} - \frac{\theta}{\theta-1}\right) \to N\left(0, \frac{\theta^2}{(\theta-1)^4}\right)$

Method-of-Moments Estimate for θ

Solve $\mu = \frac{\theta}{\theta - 1}$ for θ

$$\theta = g(\mu) = \frac{\mu}{\mu - 1}$$

M-O-M Estimator $\hat{\theta} = \frac{\bar{X}}{\bar{X}-1}$

Delta Method

① By CLT, $\sqrt{n}\left(\bar{X} - \frac{\theta}{\theta-1}\right) \longrightarrow N\left(0, \frac{\theta}{(\theta-1)^2(\theta-2)}\right)$

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

$\Rightarrow g'\left(\frac{\theta}{\theta-1}\right) = -(\theta-1)^2$

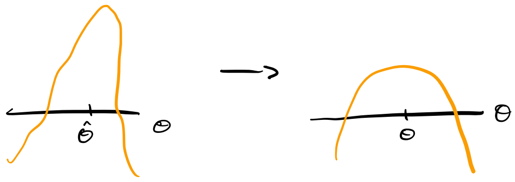
so $\sqrt{n}(\hat{\theta}-\theta) = \sqrt{n}\left(g(\bar{X}) - g\left(\frac{\theta}{\theta-1}\right)\right) \longrightarrow N\left(0, \frac{\theta}{(\theta-1)^2(\theta-2)} \cdot \left[-(\theta-1)^2\right]^2\right)$

$= N\left(0, \frac{\theta(\theta-1)^2}{\theta-2}\right)$

Geometry of the Fisher Information

Recall that as $n \to \infty$, fixing the parameter θ_0,

$$\frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta) \to E_{\theta_0}[\log f(x|\theta)] \text{ for every } \theta$$



Sample - log likelihood

Expected log-likelihood

Fisher information is the curvature of $\bar{\ell}(\theta)$ at $\theta = \theta_0$

$$\bar{\ell}(\theta) = \mathbb{E}_{\theta_0}[\log f(x|\theta)]$$

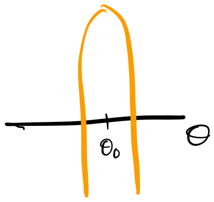
Reason: $I(\theta) = -\mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right]$

For a fixed true parameter θ_0

$$I(\theta_0) = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\Big|_{\theta=\theta_0}\right]$$

$$= -\frac{\partial^2}{\partial\theta^2} \underbrace{\mathbb{E}_{\theta_0}[\log f(x|\theta)]}_{\bar{\ell}(\theta)}\Bigg|_{\theta=\theta_0} = -\bar{\ell}''(\theta_0)$$

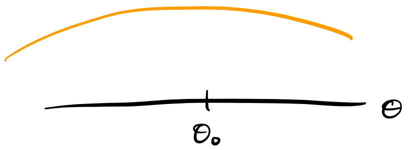
If $I(\theta_0)$ is large, then the curvature of $\bar{\ell}$ around θ_0 is big



If we perturb θ a bit from θ_0, then $\bar{\ell}(\theta)$ decreases a lot from $\bar{\ell}(\theta_0)$

Data contains a lot of information about θ_0

If $I(\theta_0)$ is small, then $\bar{\ell}$ would have small curvature around θ_0



perturbations from θ_0 do not result in large changes

Data contains less information about θ_0

Cramer-Rao Low Bound

Thm: Consider a parametric model $f(x|\theta)$ where $\theta \in \mathbb{R}$ is a single parameter. Let T be any unbiased estimator of θ based on n observations $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$. Then (under smoothness conditions)

$$Var_\theta[T] \geq \frac{1}{nI(\theta)}$$

Interpretation: If $T = \hat{\theta}$ is the MLE, then for large n, $Var[T] \approx \frac{1}{nI(\theta)}$

Any unbiased estimator can't achieve a smaller variance

Proof: Let $Z = \frac{\partial}{\partial\theta} \log f(x_1 \dots x_n|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(x_i|\theta)$

Recall: $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] = 0$

$Var_\theta\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right] = I(\theta)$

So, $\mathbb{E}_\theta[Z] = 0$ and $Var_\theta[Z] = nI(\theta)$

The correlation between Z and T belongs to $[-1,1]$

$$Cov_\theta[Z,T]^2 \leq Var_\theta[Z]\, Var[T]$$

$$Var_\theta[Z] = nI(\theta)$$

$$Cov_\theta[Z,T] = \mathbb{E}_\theta[(Z - \mathbb{E}_\theta Z)(T - \mathbb{E}_\theta T)]$$

$$= \mathbb{E}_\theta[(Z - \mathbb{E}_\theta Z)T]$$

$$= \mathbb{E}_\theta[ZT] \quad \text{bc} \quad \mathbb{E}_\theta Z = 0$$

Since T is unbiased for θ

$$\theta = \mathbb{E}_\theta[T] = \int T(x_1 \dots x_n) f(x_1 \dots x_n|\theta)\, dx_1 \dots dx_n$$

Derivative w/ respect to θ

$$1 = \int T(x_1, \dots x_n) \cdot \frac{\partial}{\partial \theta} f(x_1, \dots x_n | \theta)\, dx_1 \dots dx_n$$

Recall: $$\left[\frac{\partial}{\partial \theta} \log f(x_1, \dots x_n | \theta)\right] \cdot f(x_1, \dots x_n | \theta) = \frac{\partial}{\partial \theta} f(x_1, \dots x_n | \theta)$$

$$= \int \underbrace{T(x_1, \dots x_n)}_{T} \cdot \underbrace{\left[\frac{\partial}{\partial \theta} \log f(x_1, \dots x_n | \theta)\right]}_{Z} \cdot f(x_1, \dots x_n | \theta)\, dx_1 \dots dx_n$$

$$= \mathbb{E}_\theta [TZ]$$

Therefore, $\text{Cov}[T, Z] = 1$

$$\text{Var}_\theta [T] \geq \frac{1}{nI(\theta)}$$ ← Plugging everything

$\hat{\theta}$ is an asymptotically efficient estimator for θ

if $\sqrt{n}(\hat{\theta} - \theta) \to N(0, \frac{1}{I(\theta)})$ in distribution

MLE is asymptotically efficient

If two estimators $\hat{\theta}, \tilde{\theta}$ satisfy

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, u(\theta))$$

$$\sqrt{n}(\tilde{\theta} - \theta) \to N(0, v(\theta))$$

Then, $\frac{v(\theta)}{u(\theta)}$ is the asymptotic relative efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$

Interpretted as ratio of sample sizes required

$$\text{Var}[\hat{\theta}] \approx \frac{u(\theta)}{n} \qquad \text{Var}[\tilde{\theta}] = \frac{v(\theta)}{n}$$

For plug-in estimates

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to N(0, g'(\theta)^2 / I(\theta))$$

$$\text{Var}[g(\hat{\theta})] \approx \frac{g'(\theta)^2}{nI(\theta)}$$

Cramer-Rao Bound for plug-in estimates

In a parametric model $f(x|\theta)$ for $\theta \in \mathbb{R}$, if T is *any* unbiased estimator for $g(\theta)$ based on $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$

then $$\mathrm{Var}_\theta[T] \geq \frac{g'(\theta)^2}{n \cdot I(\theta)}$$

Analagous proof as for traditional Cramer-Rao lower bound

Fisher Information for multiple parameters

$\theta \in \mathbb{R}^k$, the fisher information matrix $I(\theta) \in R^{k \times k}$ is the matrix w/

$$(i,j) \text{ entry}: \ I(\theta)_{ij} = \mathrm{Cov}_\theta\left[\frac{\partial}{\partial\theta_i} \log f(x|\theta), \frac{\partial}{\partial\theta_j} \log f(x|\theta)\right]$$

$$= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(x|\theta)\right]$$

Thm: Let $f(x|\theta)$ be a parametric model where $\theta \in \mathbb{R}^k$

Let $\hat{\theta}$ be the MLE for θ based on n observations $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$

then, under smoothnesss conditions and assuming $I(\theta)$ is invertible

$$\underbrace{\sqrt{n}(\hat{\theta} - \theta)}_{\in \mathbb{R}^k} \to \underbrace{\mathcal{N}(0, \overset{\text{Inverse matrix}}{I(\theta)^{-1}})}_{\text{k-dimension multivariate normal}}$$

Example: Consider $X_1 \dots X_n \overset{iid}{\sim} \mathrm{Gamma}(\alpha, \beta) \quad \alpha, \beta > 0$

We can compute α, β via MLE

Use $I(\alpha, \beta)$ to quantify their uncertainity

$$\log f(x|\alpha,\beta) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1)\log x - \beta x$$

$$\frac{\partial}{\partial \alpha} \log f(X|\alpha,\beta) = \log \beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log X$$

$$\frac{\partial}{\partial \beta} \log f(X|\alpha,\beta) = \frac{\alpha}{\beta} - X$$

$$\frac{\partial^2}{\partial \alpha^2} \log f(X|\alpha,\beta) = \frac{\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} - \frac{\Gamma''(\alpha)}{\Gamma(\alpha)} = -\psi(\alpha)$$

← tri-gamma function

$$\frac{\partial^2}{\partial \alpha, \partial \beta} \log f(X|\alpha,\beta) = 1/\beta$$

$$\frac{\partial^2}{\partial \beta^2} \log f(X|\alpha,\beta) = \frac{-\alpha}{\beta^2}$$

So $$I(\alpha,\beta) = -\begin{pmatrix} -\psi(\alpha) & 1/\beta \\ 1/\beta & -\alpha/\beta^2 \end{pmatrix} = \begin{pmatrix} \psi(\alpha) & -1/\beta \\ -1/\beta & \alpha/\beta^2 \end{pmatrix}$$

$$I(\alpha,\beta)^{-1} = \frac{1}{\psi(\alpha)\frac{\alpha}{\beta} - 1/\beta^2} \begin{pmatrix} \alpha/\beta^2 & 1/\beta \\ 1/\beta & \psi(\alpha) \end{pmatrix}$$

Errors are positively correlated

3/28: Bayesian Inference

Parametric model: Data $\mathbb{X} = (X_1 \ldots X_n)$

Modeled by some distribution $f(\mathbb{X}, \theta)$ w/ parameter θ

Frequentist Perspective: Fixed true value we try to estimate

θ is non-random

Bayesian Perspective: Treat θ as a random variable with a distribution

Prior and Posterior Distributions

Review joint, marginal, and conditional distributions

Consider two r.v.'s X + Y

Joint PDF or PMF $f_{X,Y}(x,y)$

Marginal Distribution of X is given by

$$f_X(x) = \int f_{X,Y}(x,y)\,dy \quad \text{or} \quad f_X(x) = \sum_{y} f_{X,Y}(x,y)$$

continuous discrete

Conditional distribution of Y given X = x is then

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

This is a PDF or PMF over values y, fixing the value x

Similarly define

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

The Joint PDF/PMF factors in two ways

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) \cdot f_X(x)$$

$$= f_{X|Y}(x|y) \cdot f_Y(\,)$$

Bayesian Inference: Observed data $\mathbf{X} = (x_1 \ldots x_n)$

Parametric model for $\mathbf{X}$ with unknow parameter Θ

- Think about Θ as random

- Interpret the model fo $\mathbf{X}$ as the conditional distribution given $\Theta = \theta$

$$f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta$$

Θ is random w/ distribution $f_\Theta(\theta)$ — the prior distribution

=> This defines a joint distribution over both Θ and $\mathbf{X}$

$$f_{\mathbf{X},\Theta}(\mathbf{X},\theta) = f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta) \times f_\Theta(\theta)$$

We can factor this joint distribution

$$f_{\mathbf{X},\Theta}(\mathbf{X},\theta) = f_{\Theta|\mathbf{X}}(\theta|\mathbf{X}) \times f_{\mathbf{X}}(\mathbf{X})$$

$f_{\Theta}(\theta)$ is the prior distribution for Θ, ie. marginal distribution of Θ

$f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$ is our parametric model for $\mathbf{X}$ i.e. the likelihood function

$f_{\mathbf{X}}(\mathbf{x}) = \int f_{X,\Theta}(\mathbf{x},\theta)\,d\theta$ is the marginal distribution of $\mathbf{X}$.

Distribution of data averaging over Θ

$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ is the posterior distribution of Θ

This is our belief about the value of Θ, after seeing the data

Goal: Understand $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X},\Theta}(\mathbf{x},\theta)}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \times f_{\Theta}(\theta)}{f_{\mathbf{X}}(\mathbf{x})}$$

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) \times f_{\Theta}(\theta)$$

"posterior $\propto$ likelihood x prior"

Example: Coin, heads probability P

Consider a prior distribution for $P \sim$ Uniform $(0,1)$

Observe $X_1 \dots X_n \overset{iid}{\sim}$ Bernoulli(p) given $P=p$

What is the posterior of P?

Joint distribution of $X_1 \dots X_n, P$ is

$$f_{\boldsymbol{X},P}(\boldsymbol{X},P) = f_{\boldsymbol{X}|P}(\boldsymbol{X}|P) \times f_P(p)$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \times 1 \leftarrow \text{PDF uniform}$$

$$= p^s(1-p)^{n-s} \quad \text{where} \quad s = X_1 + \dots + X_n$$

Marginal PMF of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{X}) = \int_0^1 f_{\boldsymbol{X},p}(\boldsymbol{X},p)\,dp$$

$$= \int_0^1 p^s(1-p)^{n-s}\,dp$$

This is the beta integral

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Rightarrow f_{\boldsymbol{X}}(\boldsymbol{X}) = B(s+1, n-s+1)$$

Posterior Distribution of P given $\boldsymbol{X}=x$ is

$$f_{P|\boldsymbol{X}}(P|\boldsymbol{X}) = \frac{f_{\boldsymbol{X}P}(\boldsymbol{X},P)}{f_{\boldsymbol{X}}(\boldsymbol{X})}$$

$$= \frac{1}{B(s+1,n-s+1)} \cdot p^s(1-p)^{n-s} \leftarrow \text{PDF of Beta } (s+1, n-s+1)$$

Example: Can extend to a more general prior for P

Consider Beta (α, β) prior for P:

$$f_p(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1} (1-p)^{\beta-1} \quad \text{for } p \in (0,1)$$

$\alpha = \beta = 1$ gives Uniform prior

Posterior for P satisfies

$$f_{P|\mathbf{x}}(p|\mathbf{x}) \propto f_{\mathbf{x}|p}(\mathbf{x}|p) \times f_p(p)$$

$$\propto p^s (1-p)^{n-s} \cdot p^{\alpha-1} (1-p)^{\beta-1}$$

$$= p^{s+\alpha-1} (1-p)^{n-s+\beta-1}$$

This is proportional to Beta $(s+\alpha, n-s+\beta)$

Beta distribution is the posterior for P

Example: Observe counts $X_1 \dots X_n$ model as $X_1 \dots X_n \overset{iid}{\sim}$ Poisson (λ)

Treat parameter Λ as random

Take the prior $\Lambda \sim$ Gamma (α, β). This has PDF $f_\Lambda(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$ for $\lambda > 0$

What is the posterior distribution of Λ?

$$f_{\Lambda|\mathbf{x}}(\lambda|\mathbf{x}) \propto f_{(\mathbf{x}|\Lambda)}(\mathbf{x}|\lambda) \times f_\Lambda(\lambda)$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$\propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \times \lambda^{\alpha-1} e^{-\beta\lambda}$$

$$= \lambda^{\alpha + \sum_{i=1}^{n} x_i - 1} \times e^{-(\beta+\alpha)\lambda}$$

This is proportional to the Gamma $(\alpha + \sum_{i=1}^{n} x_i, \beta + \alpha)$

So this Gamma distribution is the posterior for Λ

Example : Observe $X_1 \dots X_n \overset{iid}{\sim} N(\theta, 1/\xi)$ where $\xi = 1/\sigma^2$ is the precision

① Assume that ξ is known, treat Θ as random

Consider a normal prior for $\Theta \sim N(\mu_{prior}, 1/\xi_{prior})$

$$f_{\Theta|X}(\theta|X) \propto f_{X|\Theta}(X|\theta) \times f_{\Theta}(\theta)$$

$$= \prod_{i=1}^{n} e^{-\frac{\xi}{2}(x_i-\theta)^2} \times e^{\frac{-\xi_{prior}}{2}(\theta-\mu_{prior})^2}$$

$$\propto \exp\left(\frac{-n\xi}{2}\theta^2 + \xi\left(\sum_{i=1}^{n} x_i\right)\theta - \frac{\xi_{prior}}{2}\theta^2 + \xi_{prior}\mu_{prior}\theta\right)$$

Suppose $\xi_{post} = n\xi + \xi_{prior}$ and $\mu_{post} = \dfrac{\xi \sum_{i=1}^{n} x_i + \xi_{prior}\mu_{prior}}{n\xi + \xi_{prior}}$

$$f_{\theta|x}(\theta|x)$$

some more algebra i dont fucking. know

$$N\left(\mu_{post}, \tfrac{1}{\xi_{post}}\right)$$

② Assume mean θ is known, precision ξ is unknown

Consider Gamma prior $\Xi \sim \text{Gamma}(\alpha, \beta)$

Then

$$f_{\Xi|X}(\xi|X) \propto \prod_{i=1}^{n} \sqrt{\xi}\, e^{-\frac{\xi}{2}(x_i-\theta)^2} \times \xi^{\alpha-1} e^{-\beta\xi}$$

$$\propto \xi^{\frac{\alpha+n}{2}-1} e^{-(\beta+\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}\xi}$$

GAMMA 1

Bayesian Point Estimates and credible Intervals

To get a numeric estimate for Θ:

- Mean of the posterior distribution
- Mode of the posterior ("MAP estimate")

To get an interval estimate for Θ:

- Define a Bayesian credible interval

I w/ coverage level $1-\alpha$

An interval containing $1-\alpha$ portion of the posterior distribution

$$\mathbb{P}[\Theta \in I \mid \mathbb{X}=x] = \int_I f_{\Theta|\mathbb{X}}(\Theta|\mathbb{X})\, d\theta = 1-\alpha$$

Common to use: lower $\frac{\alpha}{2}$-point to upper $\frac{\alpha}{2}$-point of the posterior

Example: $X_1 \ldots X_n \overset{iid}{\sim}$ Bernoulli(p)

Prior $P \sim Beta(\alpha, \beta)$

Recall: Posterior $P|\mathbb{X}=x \sim Beta(s+\alpha, n-s+\beta)$

where $s = X_1 + \cdots + X_n$

Estimate $\hat{p}$ by posterior mean

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

Distinct from $\frac{s}{n}$ as calculated by M-O-M and MLE

$\hat{p}$ is a weighted average of sample and prior means

Credible Interval

lower - 0.05 point to upper - 0.05 point of $Beta(s+\alpha, n-s+\beta)$

3/30: Bayesian Inference (cont'd)

Likelihood model: $\mathbb{X} = (X_1 \ldots X_n) \sim f_{X|\Theta}(X|\Theta)$

Prior: $f_\Theta(\Theta)$

Posterior: $f_{\Theta|X}(\Theta|X) \propto f_{X|\Theta}(X|\Theta) \times f_\Theta(\Theta)$

Examples:

① $X_1 \ldots X_n \overset{iid}{\sim}$ Bernoulli(p), prior $P \sim$ Uniform$(0,1)$

$\Rightarrow (P|\mathbb{X}) \sim Beta(X_1 + \cdots + X_n + 1, n-(X_1 + \cdots + X_n)+1)$

② $X_1 \ldots X_n \overset{iid}{\sim}$ Bernoulli(p), prior $P \sim Beta(\alpha, \beta)$

$\Rightarrow (P|\mathbb{X}) \sim Beta(X_1 + \cdots + X_n + \alpha, n-(X_1 + \cdots + X_n)+\beta)$

③ $X_1 \ldots X_n \overset{iid}{\sim}$ Poisson(λ), prior $\Lambda \sim$ Gamma(α, β)

$\Rightarrow (\Lambda|\mathbb{X}) \sim$ Gamma$(X_1 + \cdots X_n + \alpha, n+\beta)$

④ $X_1 \ldots X_n \overset{iid}{\sim} N(\Theta, 1/\xi)$

Known ξ, prior $\Theta \sim N(\mu_{prior}, 1/\xi_{prior})$

$$\Rightarrow (\Theta \mid \mathbf{X}) \sim N(\mu_{post}, 1/\xi_{post})$$

$$\mu_{post} = \frac{\sum_{i=1}^{n} X_n + (\xi_{prior}/\xi)\mu_{prior}}{n + (\xi_{prior}/\xi}, \quad \xi_{post} = n\xi + \xi_{prior}$$

Posterior means:

③ $\hat{\lambda} = \dfrac{X_1 + \dots + X_n + \alpha}{n + \beta}$

Different from MLE/MoM which is $\bar{X}$

- As if we had B extra samples summing to α

- $\hat{\lambda} = \dfrac{n}{n+\beta} \cdot \underset{\overset{\uparrow}{\text{Sample mean}}}{\bar{X}} + \dfrac{\beta}{n+\beta} \cdot \underset{\overset{\nwarrow}{\text{prior mean}}}{\dfrac{\alpha}{\beta}}$

④ Posterior Mean: $\hat{\Theta} = \mu_{post}$

$$\hat{\Theta} = \mu_{post} = \frac{n}{n+(\xi_{prior}/\xi)} \cdot \underset{\overset{\uparrow}{\text{Sample mean}}}{\bar{X}} + \frac{\xi_{prior}/\xi}{n+(\xi_{prior}/\xi)} \cdot \underset{\overset{\uparrow}{\text{prior mean}}}{\mu_{prior}}$$

A 70% Bayesian credible interval for Θ would be

$$\mu_{post} \pm \sqrt{\frac{1}{\xi_{prior}}} \cdot z(0.05)$$

Conjugate Priors and Improper Priors

A conjugate prior is a prior distribution when the resulting posterior has the same parametric form

- Beta prior → Bernoulli prob
- Gamma prior → Poisson rate
- Normal prior → Normal mean
- Gamma prior → Normal precision parameter

Unfortunately tend to be light-tailed (bias inference towards prior mean)

- Use heavier-tailed, non-conjugate priors for more robust inference

Goal is to use an uninformative prior

E.g. Poisson example $X_1 \dots X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$

Gamma(α, B) prior $\Rightarrow$ Posterior mean $\dfrac{X_1 + \dots + X_n + \alpha}{n + B}$

Uninformative prior would mean smaller values of α, B

At its limit : $\underset{\overset{\uparrow}{\text{not a proper probability distribution}}}{\text{Gamma}(0,0)} \propto \lambda^{-1}$

Since gamma(0,0) is not an actual distribution, we call it an improper prior

Improper priors can still yield proper posterior distributions

Bayesian vs. Frequentist Coverage Guarantees

Bayesian level-$(1-\alpha)$ credible interval guarantees

$$\mathbb{P}[\Theta \in I \mid \mathbf{X} = x] = 1 - \alpha$$

Frequentist Confidence Interval

$$\mathbb{P}_\theta[\theta \in I] = 1 - \alpha \text{ where } \mathbb{P}_\theta \text{ is over } X_1 \dots X_n \overset{iid}{\sim} f(x \mid \theta)$$

Example: Let $X_1 \dots X_n \overset{iid}{\sim} N(\theta, 1)$

MLE/MoM is that $\hat{\theta} = \bar{X}$

Under parameter θ, $\bar{X} \sim N(\theta, 1/n)$ so a frequentist level-α confidence interval is

$$\bar{X} \pm \frac{1}{\sqrt{n}} \cdot z(\alpha/2)$$

This guarantees $\mathbb{P}_\Theta\left[\Theta \in \bar{X} + \frac{1}{\sqrt{n}} \cdot z(\alpha/2)\right] = 1-\alpha$

For a bayesian analysis $\Theta \sim N(0, 1/\xi_{prior})$

$(\Theta | X) \sim N(\mu_{post}, 1/\xi_{prior})$

$$\mu_{post} = \frac{\sum_{i=1}^{n} x_i}{n+\xi_{prior}} = \frac{n}{n+\xi_{prior}} \bar{X}$$

$$\xi_{post} = n + \xi_{prior}$$

Level $(1-\alpha)$ Bayesian credible interval is

$$\frac{n}{n+\xi_{prior}} \cdot \bar{X} \pm \sqrt{\frac{1}{n+\xi_{prior}}} \cdot z(\alpha/2)$$

This guarantees : $\mathbb{P}\left[\Theta \in \frac{n}{n+\xi_{prior}} \bar{X} \pm \sqrt{\frac{1}{n+\xi_{prior}}} \cdot z(\alpha/2) \,\middle|\, X = x\right] = 1-\alpha$

Bayesian credible interval does guarantee that the frequentist coverage probability averaged according to the prior for Θ is $1-\alpha$

the average coverage probability with prior $f_\Theta(\theta)$ is $1-\alpha$

Normal Approximation

frequentist and bayesian approach converge for large n

For a fixed prior, as $n \to \infty$, the influence of the prior vanishes

mean and shape of posterior distribution are determined by data

Posterior will approach $N(\hat{\theta}, \frac{1}{nI(\theta)})$

$\hat{\theta}$ is MLE and $I(\theta)$ is Fisher Information

Bayesian Credible Interval will be $\hat{\theta} \pm \sqrt{\frac{1}{nI(\theta)}} \cdot z(\alpha/2)$

Heuristic Explanation

Let $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$ prior $f_\Theta(\theta)$

Define $\ell(\theta) = \sum_{i=1}^{n} \log(f(x_i|\theta))$ the total log likelihood

Posterior :

$$f_{\Theta|X}(\theta|X) \propto \text{likelihood} * \text{prior}$$

$$e^{\ell(\theta)} \times f_\Theta(\theta)$$

Taylor Expand for θ close to MLE

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2 \ell''(\hat{\theta})$$

$\ell'(\hat{\theta}) = 0$ be max

$$\frac{1}{n}\ell''(\hat{\theta}) \simeq \frac{1}{n}\ell''(\theta_0) = -I(\theta_0) \simeq -I(\hat{\theta})$$

$$\ell(\theta) = \ell(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^2 \cdot nI(\hat{\theta})$$

$\ell(\hat{\theta})$ depends on $\mathbb{X}$ so it can be absorbed into proportionality

4/4: MLE under misspecified models

$X_1, \dots X_n \overset{iid}{\sim} f(x|\theta_0)$

Estimate θ, quantify uncertainty

- Bias, variance, MSE
- Consistency, asymptotic normality, efficiency

Kullback-Leibler divergence

for discrete distributions w/ PMFs f and g on the sample space $\mathcal{X}$

$$D_{KL}(g\|f) = \sum_{x\in\mathcal{X}} g(x) \log \frac{g(x)}{f(x)}$$

For continuous distributions w/ PDFs f and g

$$D_{KL}(g\|f) = \int g(x) \log \frac{g(x)}{f(x)}\, dx$$

Equivalent to

$$D_{KL}(g\|f) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right] = \mathbb{E}_g[\log g(x)] - \mathbb{E}_g[\log f(x)]$$

Asymmetric definition

expectation under true distribution $X \sim g(x)$

Properties:

if $f(x) = g(x)$ for all x, then $D_{KL}(g\|f) = 0$ since $\log \frac{g(x)}{f(x)} = \log 1 = 0$

For any f and g, $D_{KL}(g\|f) \geq 0$

Follows from applying Jensen's inequality

$$D_{KL}(g\|f) = \mathbb{E}_g\left[-\log \frac{f(x)}{g(x)}\right] \geq -\log E_g\left[\frac{f(x)}{g(x)}\right] = -\log \int \cancel{g(x)} \cdot \frac{f(x)}{\cancel{g(x)}}\, dx = -\log 1 = 0$$

f and g don't need to come from the same family

Ex. Let f be $N(\mu_0, \sigma^2)$ and g be $N(\mu_1, \sigma^2)$. What is $D_{KL}(g\|f)$?

$$\log \frac{g(x)}{f(x)} = \log\left[\frac{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}\, e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}}\, e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}}\right] = \frac{-(x-\mu_1)^2}{2\sigma^2} + \frac{(x-\mu_0)^2}{2\sigma^2} = \frac{2(\mu_1-\mu_0)x - (\mu_1^2-\mu_0^2)}{2\sigma^2}$$

$$D_{KL}(g\|f) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right] = \frac{2(\mu_1-\mu_0)\,\mathbb{E}_g[X] - (\mu_1^2-\mu_2^2)}{2\sigma^2} \quad \text{where } \mathbb{E}_g[X] = \mu_1$$

$$= \frac{1}{2\sigma^2}(\mu_1-\mu_2)^2$$

KL divergence is the squared difference of means, normalized by variance

In this example KL divergence is symmetric

Ex. Let f be bernouilli(p) What is $D_{KL}(g\|f)$

g be Bernoulli (q)

Then
$$\log \frac{g(x)}{f(x)} = \begin{cases} \log q/p & x=1 \\ \log \frac{1-q}{1-p} & x=0 \end{cases}$$

$$D_{KL}(g\|p) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right] = q \log \frac{q}{p} + (1-q)\log \frac{1-q}{1-p}$$

If $p \simeq q$, we can approximate this by taylor expansion

$$\log p \doteq \log q + (p-q)\cdot\frac{1}{q} + \frac{1}{2}(p-q)^2\cdot\left(-\frac{1}{q^2}\right)$$

$$\log(1-p) \doteq \log(1-q) + (p-q)\left(-\frac{1}{1-q}\right) + \frac{1}{2}(p-q)^2\cdot\left(-\frac{1}{(1-q)^2}\right)$$

$$D_{KL}(g\|f) = q(\log q - \log p) + (1-q)(\log(1-q) - \log(1-p))$$

$$\approx q\cdot\left(-(p-q)\frac{1}{q} + \frac{1}{2}(p-q)^2\cdot\frac{1}{q^2}\right) + (1-q)\left((p-q)\cdot\frac{1}{1-q} + \frac{1}{2}(p-q)^2\cdot\frac{1}{(1-q)^2}\right.$$

$$= \frac{1}{2}(p-q)^2\cdot\left(\frac{1}{q}+\frac{1}{1-q}\right) = \frac{(p-q)^2}{2q(1-q)}$$

Ex. f is binomial (n,p)

g is binomial (n,q)

Then
$$\log \frac{g(x)}{f(x)} = \log\left(\binom{n}{x} q^x (1-q)^{n-x} \Big/ \binom{n}{x} p^x (1-p)^{n-x}\right) = x \log \frac{q}{p} + (n-x)\log\frac{1-q}{1-p}$$

$$D_{KL}(g\|f) = \mathbb{E}_g\left[\log \frac{g(x)}{f(x)}\right] = \mathbb{E}_g[X]\cdot \log\frac{q}{p} + (n - \mathbb{E}_g[X])\log\frac{1-q}{1-p}$$

$$= n\left(q\cdot\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}\right)$$

↑ exactly n times KL divergence in Bernoulli example

For $p \approx q$

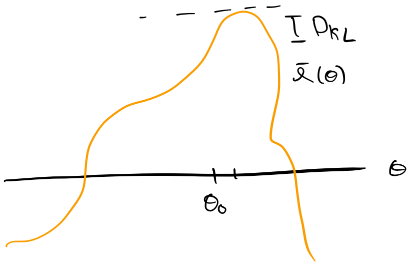
$$D_{KL}(g\|f) = n\cdot\frac{(p-q)^2}{2q(1-q)}$$

Generally, for $f(x) = f(x|\theta)$ and $g(x) = f(x|\theta_0)$

$$D_{KL}(f(x|\theta_0)\|f(x|\theta)) = \mathbb{E}_{\theta_0}\left[\log\frac{f(x|\theta_0)}{f(x|\theta)}\right]$$

$$= \mathbb{E}_{\theta_0}[\log f(x|\theta_0)] - \mathbb{E}_{\theta_0}[\log f(x|\theta)]$$

$$= \bar{\ell}(\theta_0) - \bar{\ell}(\theta)$$



$$\bar{\ell} = \mathbb{E}_{\theta_0}\left[\log f(X|\theta)\right]$$

Using a taylor expansion we find

$$D_{KL}\left(f(X|\theta_0) \,\|\, f(x|\theta)\right) \approx \frac{1}{2}(\theta-\theta_0)^2 \cdot I(\theta_0)$$

Interpretation of $I(\theta_0)$: scale factor that relates $(\theta-\theta_0)^2$ in parameter space to the theoretical "differences" D_{KL}

Consistency of MLE in misspecified model

Suppose $X_1 \dots X_n \overset{iid}{\sim} g(x)$, the pdf of the true distribution

We fit a parametric model $f(x|\theta)$ that doesn't contain $g(x)$

What happens to the MLE?

By definition the mle $\hat{\theta}$ maximizes

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

By LLN, $\frac{1}{n}\ell(\theta) \to \bar{\ell}(\theta)$ in probability

$$\bar{\ell}(\theta) = \mathbb{E}_g\left[\log f(X|\theta)\right]$$

Here, if g doesn't belong to our model

$$\bar{\ell}(\theta) = \underbrace{\mathbb{E}_g\left[\log g(x)\right]}_{\text{independent of } \theta} - \underbrace{\mathbb{E}_g\left[\log \frac{g(x)}{f(x|\theta)}\right]}_{D_{KL}(g\,\|\, f(x|\theta))}$$

therefore we maximize $\bar{\ell}(\theta)$ by maximizing $D_{KL}(g\|f(x|\theta))$

Suppose $\theta \mapsto D_{KL}(g(x)\,\|\, f(x|\theta))$ has a unique maximizer θ^*

Under some smoothness assumptions as $n \to \infty$, the MLE $\hat{\theta}$ converges to θ^* in probability

Example: Suppose we observe $X_1 \dots X_n > 0$ and fit a model Exponential λ

$$f(x|\lambda) = \lambda e^{-\lambda x} \text{ for } x > 0$$

In reality $X_1 \dots X_n \overset{iid}{\sim}$ Gamma $(\alpha, 1)$, $g(x) \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$ only exponential if $\alpha = 1$

Thm tells us that maximizing $D_{KL}(g(x)\,\|\, f(x|\lambda^*))$ estimates λ^*

$$\log \frac{g(x)}{f(x|\lambda)} = \log\left(\frac{\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}}{\lambda e^{-\lambda x}}\right) = -\log \Gamma(\alpha) + (\alpha-1)\log x - \log \lambda - (\lambda-1)x$$

Minimizing λ

$$0 = -\frac{1}{\lambda} + \mathbb{E}_g[X] \Rightarrow \lambda^* = \frac{1}{\mathbb{E}_g[X]} = \frac{1}{\alpha}$$

MLE $\hat{\lambda} \to \frac{1}{\alpha}$ in probability

Could solve this more directly through traditional methods

Ex. What is the asymptotic variance of $\hat{\lambda}$?

We have explicit $\hat{\lambda} = \frac{1}{\bar{X}}$ so we can apply the delta method

$$\sqrt{n}\left(\bar{X} - \alpha\right) \to N\left(0, \mathrm{Var}_g[X]\right) \text{ where } \mathrm{Var}_g[X] = \alpha$$

Apply delta method w/ $h(\alpha) = \frac{1}{\alpha} \Rightarrow h'(\alpha)^2 = \frac{1}{\alpha^4}$

$$\text{So, } \sqrt{n}\left(\hat{\lambda} - \tfrac{1}{\alpha}\right) = \sqrt{n}\left(h(\bar{x}) - h(\alpha)\right)$$

$$\to N\left(0, \alpha \cdot \tfrac{1}{\alpha^4}\right) = N\left(0, \tfrac{1}{\alpha^3}\right)$$

Variance $\hat{\lambda}$ for large n is $\frac{1}{n\alpha^3}$

Fisher information estimate would be incorrect

$$I(\lambda) = \frac{1}{\lambda^2}$$

$\frac{1}{nI(\hat{\lambda})}$ to estimate variance

$$\hat{\lambda} \approx \lambda^* = \tfrac{1}{\alpha}$$

$I(\hat{\lambda}) \approx I(\lambda^*) = \alpha^2$ and $\frac{1}{nI(\hat{\lambda})} \approx \frac{1}{n\alpha^2}$ instead of correct variance $\frac{1}{n\alpha}$

4/6: The bootstrap

Simulation based approach to quantify uncertainity of statistical estimates

Can estimate standard error or a confidence interval

Given $X_1 \ldots X_n \overset{iid}{\sim} f(X|\theta)$, what is the standard error for an estimator $\hat{\theta}$ for θ

Idea of the bootstrap is to simulate new data and compute $\hat{\theta}$ from each dataset

Unfortunately, you can't simulate $f(X|\theta)$ without knowing θ

Bootstrap method involves simulations from an estimate of the true distribution

Parametric Bootstrap

- Assume $X_1 \ldots X_n \overset{iid}{\sim} f(X|\theta)$
- Estimate θ by $\hat{\theta}$ and simulate $X_1^* \ldots X_n^* \overset{iid}{\sim} f(X|\hat{\theta})$
- Analagous to the plug-in principle

Non-parametric bootstrap

- No assumption of a parametric model
- Instead, we sample $X_1^* \ldots X_n^*$ independently with replacement from the original $X_1 \ldots X_n$
  - Sample size is n
  - Likely to have repeated values
  - Some samples will be lost (unsampled)
    - 63.2% of samples are expected to be present

Rationale for the nonparametric bootstrap

- Estimated distribution is the empirical distribution
  - each observation places a mass of $\frac{1}{n}$
- Draw new data from this estimated distribution

Key differences between true distribution and empirical distribution

- Empirical distribution is always discrete
- Some statistics don't make sense to compare
  - mode, max value, min value

The empirical CDF very closely resembles the true CDF

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le t\}$$

as $n \to \infty$, this converges to CDF

Mean translates well between empirical and true distributions

Bootstrap and Misspecified Models

Suppose $X_1, \dots X_n \overset{iid}{\sim} g(x)$. We fit $g(x)$ with Poisson λ

Fischer Information:

$$I(\lambda) = 1/\lambda$$

$$\sqrt{1/nI(\hat{\lambda})} = \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\bar{X}}{n}}$$

so true standard error is $\sqrt{\frac{\sigma^2}{n}}$

If poisson were correct then $\sigma^2 = \lambda$ so $\sqrt{\frac{\hat{\lambda}}{n}}$ is an accurate estimate of standard error

Non parametric bootstrap guards against model misspecification

Bootstrap Confidence Intervals

Let $\hat{\theta}$ be an estimator of θ and $\hat{se}$ be the bootstrap standard error estimate of $\hat{\theta}$

$\theta \pm z(\alpha/2)\,\hat{se}$ for 100 - α confidence interval ← Normal

Percentile bootstrap interval

Let $\hat{\theta}_1^*, \dots \hat{\theta}_B^*$ be the values of $\hat{\theta}$ computed in B simulations

Let $\hat{\theta}^{*(\alpha/2)}$ and $\hat{\theta}^{*(1-\alpha/2)}$ be the empirial $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated values

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

Requires symmetry?

Basic Bootstrap Interval

Let $q^{(\alpha/2)}$ and $q^{(1-\alpha/2)}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of $\hat{\theta}^* - \hat{\theta}, \dots \hat{\theta}_B^* - \hat{\theta}$

Use this to approximate true distribution of $\hat{\theta} - \theta$

$$\left[\hat{\theta} - q^{(1-\alpha/2)}, \hat{\theta} - q^{(\alpha/2)}\right] \quad \text{since} \quad \hat{\theta} - \theta \in \left[q^{(\alpha/2)}, q^{(1-\alpha/2)}\right] \iff$$

Same as percentile confidence interval if it is symmetric

Advantages and Disadvantages of bootstrap

- Easy to apply
- Obtains estimates valid under misspecification
- Can be computationally prohibitive
- Validity of bootstrap requires analytic proofs

4/11: Generalized Likelihood Ratio Test

① Simple Null Hypothesis
② Sub-model null hypothesis

Generalized Likelihood Ratio Test for Simple null hypothesis

Observe: $X_1, \dots X_n \overset{iid}{\sim} f(x|\theta)$

$H_0 : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

The GLRT rejects H_0 for small values of Λ

$$\Lambda = \frac{\text{lik}(\theta_0)}{\max_\theta \text{lik}(\theta)} \quad \text{where } \text{lik}(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

H_1 is composite, so you replace $\text{lik}(\Theta_1)$ with max-$\text{lik}(\Theta)$

$\max_\Theta \text{lik}(\Theta) = \text{lik}(\hat{\Theta})$ where $\hat{\Theta}$ is the MLE

By definition of MLE

$\text{lik}(\hat{\Theta}) \geq \text{lik}(\Theta_0)$ so $\Lambda \leq 1$

If Ho were true, we expect $\hat{\Theta} \approx \Theta_0$ for large n

Rejecting Ho for small Λ is the same as rejecting Ho for large $-2\log\Lambda$

$$-2\log\Lambda = 2\ell(\hat{\Theta}) - 2\ell(\Theta_0)$$

$$\ell(\Theta) = \log \text{lik}(\Theta) = \sum_{i=1}^{n} \log(f(x_i|\Theta))$$

Since $\Lambda \leq 1$, $-2\log\Lambda \geq 0$

We reject Ho when

$-2\log\Lambda \geq \chi^2_K(\alpha)$ ← upper α point of χ^2_K

K is the dimension of the parameter space

Example: Let $X_1 \ldots X_n \overset{iid}{\sim} N(\Theta,1)$. Test null hypothesis $\Theta=0$ and $H_1: \Theta \neq 0$

MLE for Θ is $\hat{\Theta} = \bar{X}$. We compute

$$\text{lik}(0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-x_i^2}{2}} \qquad \Lambda = \frac{\text{lik}(0)}{\text{lik}(\hat{\Theta})} = \exp\left(-\sum_{i=1}^{n}\frac{x_i^2}{2} + \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2}\right)$$

$$\text{lik}(\hat{\Theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\bar{x})^2}{2}}$$

$$= \exp\left(-\frac{n}{2}\bar{x}^2\right)$$ ← expand square and combine terms

$$-2\log\Lambda = n\bar{X}^2 = (\sqrt{n}\bar{X})^2$$

Under $H_0: \Theta = 0$ — $\bar{X} \sim N(0, 1/n)$ so $\sqrt{n}\bar{X} \sim N(0,1)$

so $-2\log\Lambda = (\sqrt{n}\bar{X})^2 \sim \chi^2_1$

For more general non-parametric models

Thm: Let $X_1 \ldots X_n \overset{iid}{\sim} f(x|\Theta)$, a parametric model with parameter space of dimension K. Let Θ_0 be in the interior of the parameter space. Under smoothness conditions

$$H_0: \Theta = \Theta_0 \quad \text{vs.} \quad H_1: \Theta \neq \Theta_0$$

$-2\log\Lambda \to \chi^2_K$ in distribution as $n \to \infty$ under H_0

So, the GLRT that rejects $-2\log\Lambda \geq \chi^2_K(\alpha)$ is an asymptotic level-α test

Proof sketch: Suppose K=1 when $\Theta \in \mathbb{R}$ is a single parameter

$$-2\log\Lambda = 2\ell(\hat{\Theta}) - 2\ell(\Theta_0)$$

Taylor Expansion: $\ell(\Theta_0)$ around $\hat{\Theta}$

$$\ell(\Theta_0) \approx \ell(\hat{\Theta}) + (\Theta_0 - \hat{\Theta})\cdot\ell'(\hat{\Theta}) + \frac{1}{2}(\Theta_0-\hat{\Theta})^2\cdot\ell''(\hat{\Theta})$$

Recall: $\ell'(\hat{\Theta}) = 0$ since it is at a maximizing point

$\ell''(\hat{\Theta}) \approx -n I(\hat{\Theta}) = -n I(\Theta_0)$

↑ from earlier

$$\Rightarrow \ell(\Theta_0) \approx \ell(\hat{\Theta}) - \frac{n}{2} I(\Theta_0)(\Theta_0 - \hat{\Theta})^2$$

$$\Rightarrow -2\log\Lambda \approx n I(\Theta_0)\cdot(\Theta_0 - \hat{\Theta})^2$$

Under H_0: $\sqrt{n}(\hat{\Theta} - \Theta_0) \longrightarrow N(0, 1/I(\Theta_0))$ in distribution, by asymptotic normality for MLE

$$\sqrt{n I(\Theta_0)}\,(\hat{\Theta} - \Theta_0) \longrightarrow N(0,1)$$

$$n I(\Theta_0)(\hat{\Theta} - \Theta_0)^2 \longrightarrow \chi^2_1$$

Dimension is the number of parameters minus the number of equality constraints

Example: Consider $(X_1, \dots X_n) \sim \text{Multinomial}(n, (p_1, p_2, \dots p_k))$

$$H_0: (p_1, \dots, p_k) = (p_{0_1}, \dots p_{0_k})$$
$$H_1: (p_1, \dots, p_k) \neq (p_{0_1}, \dots p_{0_k})$$

Likelihood Function:

$$\text{lik}(p_1 \dots p_k) = \binom{n}{x_1 \dots x_k} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

$$\ell(p_1 \dots p_k) = \log \binom{n}{x_1 \dots x_k} + X_1 \log p_1 + \dots + X_k \log p_k$$

The MLE is $(\hat{p}_1 \dots \hat{p}_k) = \left(\frac{X_1}{n}, \dots, \frac{X_k}{n}\right)$

$$\Rightarrow -2 \log \Lambda = 2\ell(\hat{p}_1 \dots \hat{p}_k) = -2\ell(p_{0_1}, \dots p_{0_k})$$
$$= 2X_1\left(\log \frac{X_i}{n} - \log p_{0_1}\right) + \dots + 2X_k\left(\log \frac{X_k}{n} - \log p_{0_k}\right)$$

Dimension is $k-1$ (1 equality constraint)

GLRT rejects when $-2 \log \Lambda \geq \chi^2_{k-1}(\alpha)$

GLRT for a sub-model

Let $X_1 \dots X_n \overset{iid}{\sim} f(x|\theta)$. Let Ω be the parameter space.

$H_0: \theta \in \Omega_0$

$H_1: \theta \notin \Omega_0$

$\Omega_0 \subset \Omega$ is a lower dimensional space in Ω

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

$\Omega = \{(\mu, \sigma^2) : \sigma^2 > 0\}$ is a region of the plane $\mathbb{R}^2$

$H_0: \mu = 0$

$H_1 \ \mu \neq 0$

Then $\Omega_0 = \{(\mu, \sigma^2) : \mu = 0, \sigma^2 > 0\}$ ← a line inside Ω

Ω_0 has dimension 1 while Ω has dimension 2

GLRT rejects H_0 for small values of

$$\Lambda = \frac{\max_{\theta \in \Omega_0} \text{lik}(\theta)}{\max_{\theta \in \Omega} \text{lik}(\theta)}$$

In other words, $\Lambda = \dfrac{\text{lik}(\hat{\theta}_0)}{\text{lik}(\hat{\theta})}$ $\quad$ $\hat{\theta}_0$ is MLE in Ω_0, $\hat{\theta}$ is MLE in Ω

Once again we consider $-2 \log \Lambda$

GLRT rejects for

$$-2 \log \Lambda \geq \chi^2_k(\alpha)$$

where k is the difference in dimension between Ω and Ω_0

If $\Omega_0 = \Theta_0$ we have dimension $= 0$, so the submodel generalization simplifies to the simple null example

Example: (One-sample t-test)

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

log-likelihood:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Full model MLE:

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$$

$$\Rightarrow \ell(\hat{\mu}, \hat{\sigma}^2) = \frac{-n}{2}\left(\log(2\pi\hat{\sigma}^2) + 1\right)$$

Sub-model MLE:

$\mu = 0$

MLE is $(\mu_0^2, \sigma_0^2) = \left(0, \frac{1}{n}\sum_{i=1}^{n} x_i^2\right)$ ← from HW6

$$\Rightarrow \ell(\mu_0^2, \sigma_0^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}$$

$$-2\log\Lambda = 2\ell(\hat{\mu}, \hat{\sigma}^2) - 2\ell(\hat{\mu}_0, \hat{\sigma}_0^2)$$

$$= n\log\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)$$

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{x})^2$ $\quad$ $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2$ $\quad$ so $\hat{\sigma}^2 = \hat{\sigma}_0^2 - \bar{X}$

$$\Rightarrow -2\log\Lambda = n\log\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = n\log\left(1 + \frac{\bar{X}^2}{\hat{\sigma}^2}\right) = n\log\left(1 + \frac{1}{n-1}T^2\right)$$

recall: $T = \dfrac{\sqrt{n}\bar{X}}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{x})^2}$

This is increasing in $|T|$ so rejecting for large $-2\log\Lambda$ is the same as rejecting large $|T|$

Except, for 2 sided t-test we reject H_0 when $|T| \geq t_{n-1}(\alpha/2)$

for GLRT we reject H_0 when $-2\log\Lambda \geq \chi_1^2(\alpha)$

for large n $\quad -2\log\Lambda \approx \chi_1^2$

$$-2\log\Lambda = n\log\left(1 + \frac{1}{n-1}T^2\right)$$

$$\approx n \cdot \frac{1}{n-1}T^2 \approx T^2 = Z^2 \text{ where } Z \sim N(0,1$$

4/13: Chi-squared tests for categorical data

Recap: GLRT

$H_0: \Theta \in \Omega_0$ vs. $H_1: \Theta \notin \Omega_0$

Generalized Neyman Pearson Lemma

$$\Lambda = \frac{\text{lik}(\hat{\Theta}_0)}{\text{lik}(\hat{\Theta})}$$ where: $\hat{\Theta}_0$ the MLE in Ω_0

$\hat{\Theta}$ the MLE in full model

Test statistic: $-2\log\Lambda$

for large sample distribution is $\approx \chi_k^2$ where k is the difference in dimension between Ω and Ω_0

Example: Hardy-Weinberg Equilibrium

Genotypes at single locus $\in \{AA, Aa, aa\}$ in n individuals with observed counts N_{AA}, N_{Aa}, and N_{aa}

Full model: $(N_{AA}, N_{Aa}, N_{aa}) \sim \text{Multinomial}(n, (P_{AA}, P_{Aa}, P_{aa}))$

H_0: $P_{AA} = (1-\Theta)^2$, $P_{Aa} = 2\Theta(1-\Theta)$, $P_{aa} = \Theta^2$ for some $\Theta \in (0,1)$

$$\text{lik}(P_{AA}, P_{Aa}, P_{aa}) = \binom{n}{N_{AA}, N_{Aa}, N_{aa}} \times P_{AA}^{N_{AA}} P_{Aa}^{N_{Aa}} P_{aa}^{N_{aa}}$$

$$\Rightarrow \log \binom{n}{N_{AA}, N_{Aa}, N_{aa}} + N_{AA} \log P_{AA} + N_{Aa} \log P_{Aa} + N_{aa} \log P_{aa}$$

$$\Rightarrow -2 \log \Lambda = 2\ell(\hat{p}_{AA}, \hat{p}_{Aa}, \hat{p}_{aa}) - 2\ell(\hat{P}_{0AA}, \hat{P}_{0Aa}, \hat{P}_{0aa})$$

$$= 2N_{AA}\left(\log \frac{\hat{P}_{AA}}{\hat{P}_{0AA}}\right) + 2N_{Aa}\left(\log \frac{\hat{P}_{Aa}}{\hat{P}_{0Aa}}\right) + 2N_{aa}\left(\log \frac{\hat{p}_{aa}}{\hat{p}_{0aa}}\right)$$

Full Model MLE: $\hat{P}_{AA} = \frac{N_{AA}}{n}$ $\quad \hat{P}_{Aa} = \frac{N_{Aa}}{n}$ $\quad \hat{P}_{aa} = \frac{N_{aa}}{n}$

Sub-model MLE: $\hat{\theta} = \dfrac{2N_{aa} - N_{Aa}}{2n}$

$P_{0AA} = (1-\hat{\theta})^2$, $P_{Aa} = 2\hat{\theta}(1-\hat{\theta})$, $P_{aa} = \hat{\theta}^2$

Dimensions : Full model : 2 (3 params w/ 1 constraint)

sub-model : 1 (single param θ)

$k = 1$

Compare $-2\log \Lambda$ to $\chi_1^2(\alpha)$

Test of Independence

Example: General Social Survey

Random sample of 1972 people

	Dem	Repub	Indep
M	422	381	273
F	299	365	232

Want to test whether gender is independent of party affiliation

Model : $\begin{pmatrix} N_{11} & N_{12} & N_{13} \\ N_{21} & N_{22} & N_{23} \end{pmatrix} \sim \text{Multinomial}\left(n, \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \end{pmatrix}\right)$

Let $P_{i\cdot} = \sum_j P_{ij}$ (row sums)

$P_{j\cdot} = \sum_i P_{ij}$ (column sums)

H_0: $P_{ij} = P_{i\cdot} \times P_{j\cdot}$ for every i,j vs. H_1:

Dimensions :

Full model : 5 (6 params, 1 constraint $\sum_{ij} P_{ij} = 1$)

Sub model : 3 (5 params, 2 constraints)

Parameters : $P_{1\cdot}$ $P_{2\cdot}$ $P_{\cdot 1}$ $P_{\cdot 2}$ $P_{\cdot 3}$

Constraints : $P_{1\cdot} + P_{2\cdot} = 1$ $\quad P_{\cdot 1} + P_{\cdot 2} + P_{\cdot 3} = 1$

$k = 2$ so $-2 \log \Lambda$ compare against $\chi_2^2(\alpha)$

Consider more generally,

$$(N_1, \dots, N_K) \sim \text{Multinomial}(n, (p_1 \dots p_K))$$

Test $H_0 : (p_1 \dots p_K) \in \Omega_0$

Multinomial likelihood

$$\text{lik}(p_1 \dots p_K) = \binom{n}{N_1 \dots N_K} \times \prod_{i=1}^{K} p_i^{N_i}$$

Let $\hat{p}_i$ and $\hat{p}_{0i}$ be the full model and submodel MLEs.

$$-2 \log \Lambda = 2 \log \text{lik}(\hat{p}_1 \dots \hat{p}_K) - 2 \log \text{lik}(\hat{p}_{0_1} \dots \hat{p}_{0_K})$$

$$= 2 \sum_{i=1}^{K} N_i \cdot \log \frac{\hat{p}_i}{\hat{p}_{0,i}}$$

Recall: $\hat{p}_i = \frac{N_i}{n}$

Expected counts in sub-model $E_i = n \cdot \hat{p}_{0,i}$

$$\Rightarrow -2 \log \Lambda = 2 \sum_{i=1}^{K} N_i \log \frac{N_i}{E_i}$$

Return to $k=6$ example

$$\text{lik}(p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}, p_{\cdot 3}) = \binom{n}{N_{11}, N_{12} \dots N_{23}} \times \prod_{i=1}^{2} \prod_{j=1}^{3} (p_{i\cdot} \dots p_{\cdot j})^{N_{ij}}$$

$$= \binom{n}{N_{11}, N_{12} \dots N_{23}} \times \prod_{i=1}^{2} p_{i\cdot}^{N_{i\cdot}} \times \prod_{j=1}^{3} p_{\cdot j}^{N_{\cdot j}}$$

where $N_{i\cdot} = N_{i1} + N_{i2} + N_{i3}$

$N_{ij} = N_{1j} + N_{2j}$

Take a log and maximize subject to $\sum_i p_{i\cdot} = 1$, $\sum_j p_{\cdot j} = 1$

$$\mathcal{L} = \text{Lagrangian}: \log \binom{n}{N_{11} \dots N_{23}} + \sum_{i=1}^{2} N_{i\cdot} \log p_{i\cdot} + \sum_{j=1}^{3} N_{\cdot j} \log p_{\cdot j} + \lambda \left(\sum_{i=1}^{2} p_{i\cdot} - 1 \right) + \mu \left(\sum_{j=1}^{3} p_{\cdot j} - 1 \right)$$

$$0 = \frac{\partial \mathcal{L}}{\partial p_{i\cdot}} = \frac{N_{i\cdot}}{p_{i\cdot}} + \lambda \Rightarrow p_{i\cdot} = \frac{-N_{i\cdot}}{\lambda}$$

$$0 = \frac{\partial \mathcal{L}}{\partial p_{\cdot j}} = \frac{N_{\cdot j}}{p_{\cdot j}} + \mu \Rightarrow p_{\cdot j} = \frac{-N_{\cdot j}}{\mu}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \lambda} = p_{1\cdot} + p_{2\cdot} - 1 \Rightarrow -\frac{N_{1\cdot}}{\lambda} - \frac{N_{2i}}{\lambda} = -\frac{n}{\lambda} = 1$$

$$\Rightarrow \lambda = -n$$

$$\Rightarrow \hat{p}_{i\cdot} = \frac{N_{i\cdot}}{n}$$

$$0 = \frac{\partial \mathcal{L}}{\partial \mu} = p_{\cdot 1} + p_{\cdot 2} + p_{\cdot 3} - 1 = \frac{-N_{\cdot 1}}{\mu} - \frac{N_{\cdot 2}}{\mu} - \frac{N_{\cdot 3}}{\mu} = \frac{-n}{\mu} = 1$$

$$\Rightarrow \mu = -n$$

$$\Rightarrow \hat{p}_{ij} = \frac{N_{\cdot j}}{n}$$

$$\Rightarrow \hat{p}_{0_{ij}} = \hat{p}_{i\cdot} \times \hat{p}_{\cdot j} = \frac{N_{i\cdot} N_{\cdot j}}{n^2}$$

Expected counts: $E_{ij} = n \cdot \hat{p}_{0\,ij} = \frac{N_{i\cdot} \times N_{\cdot j}}{n}$

Pluggin into $-2\log\Lambda = 2\sum_{i=1}^{2}\sum_{j=1}^{3} N_{ij}\log\frac{N_{ij}}{E_{ij}} = 8.31$

Compare to χ^2_2 so we find a p-value of 0.016

Pearson chi-squared test

Alternative to GLRT

$$X^2 = \sum_{i=1}^{k}\frac{(N_i-E_i)^2}{E_i}$$

Test rejects Ho when X^2 exceeds $\chi^2_{dof}(\alpha)$ value

↗ difference in dimension of the models

X^2 is similar to $-2\log\Lambda$

See via a taylor expansion (4/13 lecture 22)

Test of Homogeneity

Setting: $(N_1, \dots N_k) \sim$ Multinomial $(n, (p_1 \dots p_k))$

$(M_1 \dots M_k) \sim$ Multinomial $(m, (q_1 \dots q_k))$

$H_0: p_i = q_i$ for all $i = 1 \dots k$

Example: Jane Austen + Emulator

	a	an	this	that	with	without
Ch. 1 + 6	101	11	15	37	28	10
Ch. 12 + 24	83	29	15	22	43	4

Ch 1+6 $\sim$ Multinomial $(202, (p_1 \dots p_6))$

Ch 12. + 24 $\sim$ Multinomial $(196\ (q_1 \dots q_6))$

Test $H_0: p_i = q_i$ for all $i = 1 \dots 6$

GLRT statistic

$$-2\log\Lambda = 2\sum_{i=1}^{k} N_i\log\frac{N_i}{E_i} + M_i\log\frac{M_i}{F_i}$$

N_i, M_i are the observed counts

$E_i = n \cdot \hat{p}_{0i}$ $F_i = m \cdot \hat{p}_{0i}$

MLEs are $\hat{p}_{0,i} = \frac{N_i + M_i}{n+m}$

Dimension of full model : $5+5 = 10$

Dimension of sub model : 5

$k = 10 - 5 = 5$ so compare against $\chi^2_5(\alpha)$

Find p-value to be 0.0014 ← very different!

4/18: The Bradley-Terry Model

Data: $\mathbb{Y} = (Y_1 \dots Y_n)$

Parametric model - $f(\mathbb{Y}|\Theta)$ is the joint PDF/PMF of our data dependent on $\Theta \in \mathbb{R}^k$

Log-likelihood function - $\ell(\Theta) = \sum_{i=1}^{n} \log f(Y_i | \Theta)$

MLE - $\hat{\Theta}$ that maximizes $\ell(\Theta)$

Recall: If $Y_1 \dots Y_n \overset{iid}{\sim} f(y|\theta)$

$$I(\theta) = \operatorname{Var}_\theta\left[\frac{\partial}{\partial\theta} \log f(Y|\theta)\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \log f(y|\theta)\right]$$

For large n, the MLE $\hat{\theta}$ is approximately distributed as

$$N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

Define $$I_{\mathbb{Y}}(\theta) = nI(\theta) = \operatorname{Var}_\theta\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(Y_i|\theta)\right] = \operatorname{Var}_\theta[\ell'(\theta)]$$

$$= -\mathbb{E}_\theta\left[\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(Y_i|\theta)\right] = -\mathbb{E}_\theta[\ell''(\theta)]$$

Generally for $\mathbb{Y} \sim f(\mathbb{Y}|\theta)$, define fisher information of all observations

$$I_{\mathbb{Y}}(\theta) = \operatorname{Var}_\theta[\ell'(\theta)] = -\mathbb{E}_\theta[\ell''(\theta)]$$

Under regularity assumptions, the MLE $\hat{\theta}$ has approximately distribution $N\left(\theta_0, \frac{1}{I_{\mathbb{Y}}(\theta_0)}\right)$ ← Estimate when data isn't iid

For multiple parameteres $\Theta \in \mathbb{R}^k$, we define $I_{\mathbb{Y}}(\theta) \in \mathbb{R}^{k \times k}$

$$I_{\mathbb{Y}}(\theta)_{ij} = \operatorname{Cov}_\theta\left[\frac{\partial}{\partial\theta_i}\ell(\theta), \frac{\partial}{\partial\theta_j}\ell(\theta)\right]$$

$$= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right]$$

Again, $\hat{\Theta}$ is $\approx N(\Theta_0, I_{\mathbb{Y}}(\Theta_0)^{-1})$

Bradley Terry Model

Example: NBA has 30 basketball teams

Each team plays 82 games

How can we rank teams?

① Naively: Count number of wins against number of losses

However, each team plays each other team between 2 and 4 times

② Bradley-Terry: Represent strength of team i by $B_i \in \mathbb{R}$

If game played between teams i and j, outcome is random and depends on B_i and B_j

Outcome — Bernoulli (P_{ij})

$$\log \frac{P_{ij}}{1-P_{ij}} = B_i - B_j \implies P_{ij} = \frac{e^{B_i - B_j}}{1+e^{B_i - B_j}} = \frac{e^{B_i}}{e^{B_i} + e^{B_j}}$$

- Each B_i is only meaningful relative to other B_j's

  We an add any constant $C \in \mathbb{R}$ to all B_i's without changing the model

  Allows us to select a team as a standard (set to 0)

  in this case B_j is relative strength to the standardized teams

- Can specify order of each game s.t. the first team is always the home team

  Incorporate home team advantage by adding an additional intercept

$$\log \frac{P_{ij}}{1-P_{ij}} = \alpha + B_i + B_j$$

Estimation and Inference

$K = 30$ (number of teams)

$n = 1580$ (number of total games)

Questions: Estimate α and $B_1, B_2, \dots B_K$ (Constraining $B_{nets} = B_1 = 0$)

Test the null hypothesis $H_0 : \alpha = 0$ (no home team advantage)

Obtain confidence interval

We observe n games $(i_1, j_1), (i_2, j_2), \dots (i_n, j_n)$ where i is always the home team

Outcomes: $Y_1, Y_2, \dots Y_n \in \{0,1\}$

$$Y_m = \begin{cases} 1 & \text{if } i \text{ beat } j \text{ in the mth game} \\ 0 & \text{otherwise} \end{cases}$$

Log-likelihood:

$$\text{lik}(\alpha, B_2, \dots B_K) = \prod_{m=1}^{n} P_{i_m j_m}^{Y_m} (1 - p_{i_m j_m}^{Y_m}) = \prod_{m=1}^{n} (1 - P_{i_m j_m}) \left(\frac{P_{i_m j_m}}{1 - P_{i_m j_m}} \right)^{Y_m}$$

$$\ell(\alpha, B_2 \dots B_K) = \sum_{m=1}^{n} Y_m \underbrace{\log \frac{P_{i_m j_m}}{1 - P_{i_m j_m}}}_{\substack{\text{log-odds} \\ = \alpha + B_{i_m} + B_{j_m}}} + \log (1 - P_{i_m j_m})$$

$$= \sum_{m=1}^{n} Y_m \left[\alpha + B_{i_m} + B_{j_m} \right] - \log \left(1 + e^{\alpha + B_{i_m} - B_{j_m}} \right)$$

① Common approach to estimate $\Theta\ (\alpha, B_2, \dots B_K)$ via MLE

Solve:

$$0 = \frac{\partial \ell}{\partial \alpha} = \sum_{m=1}^{n} \left[Y_m - \frac{e^{\alpha + B_{i_m} - B_{j_m}}}{1 + e^{\alpha + B_{i_m} - B_{j_m}}} \right]$$

$$0 = \frac{\partial \ell}{\partial B_i} = \sum_{m : i_m = i} \left[Y_m - \frac{e^{\alpha + B_{i_m} - B_{j_m}}}{1 + e^{\alpha + B_{i_m} - B_{j_m}}} \right] + \sum_{m : j_m = i} \left[-Y_m + \frac{e^{\alpha + B_{i_m} - B_{j_m}}}{1 + e^{\alpha + B_{i_m} - B_{j_m}}} \right]$$

No closed form so we solve it numerically

Gradient: $\nabla \ell(\Theta) = \left(\frac{\partial \ell}{\partial \alpha}, \dots \frac{\partial \ell}{\partial B_K} \right)$

Hessian: $$\nabla^2 \ell(\Theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \alpha^2} & \cdots & \frac{\partial^2 \ell}{\partial \alpha \partial B_K} \\ \vdots & \ddots & \\ \frac{\partial \ell^2}{\partial \alpha \partial B_K} & & \frac{\partial \ell^2}{\partial B_K^2} \end{pmatrix}$$

Newton - Raphson: $\Theta^{(t+1)} = \Theta^{(t)} - \left(\nabla^2 \ell(\Theta^{(t)}) \right)^{-1} \nabla \ell(\Theta^{(t)})$

② To test $H_0 : \alpha = 0$

Use the GLRT sub-model where $\alpha = 0$

$$0 = \frac{\partial \ell}{\partial B_i} = \sum_{m : i_m = i} \left[Y_m - \frac{e^{B_{i_m} - B_{j_m}}}{1 + e^{B_{i_m} - B_{j_m}}} \right] + \sum_{m : j_m = i} \left[-Y_m + \frac{e^{B_{i_m} - B_{j_m}}}{1 + e^{B_{i_m} - B_{j_m}}} \right]$$

Solve numerically using Newton-Raphson

Compute $-2\log\Lambda = 2\ell(\hat{\alpha}, \hat{B}_2, \dots \hat{B}_K) - 2\ell(0, \hat{B}_{0_2}, \dots \hat{B}_{0_K})$

Compare against $\chi^2_1(\alpha)$

③ Confidence Interval for $B_i - B_j$:

Center the interval at $\hat{B}_i - \hat{B}_j$ with the MLE's

Estimate standard error of $\hat{B}_i - \hat{B}_j$

$$\begin{aligned}
\mathrm{Var}[\hat{B}_i - \hat{B}_j] &= \mathrm{Cov}[\hat{B}_i - \hat{B}_j, \hat{B}_i - \hat{B}_j] \\
&= \mathrm{Cov}[\hat{B}_i, \hat{B}_i] - \mathrm{Cov}[\hat{B}_i, \hat{B}_j] - \mathrm{Cov}[\hat{B}_j, \hat{B}_i] + \mathrm{Cov}[\hat{B}_j, \hat{B}_j] \\
&= \mathrm{Var}[\hat{B}_i] + \mathrm{Var}[\hat{B}_j] - 2\,\mathrm{Cov}[\hat{B}_i, \hat{B}_j] \\
&\approx (I_y(\theta)^{-1})_{ii} + (I_y(\theta)^{-1})_{jj} - 2\,(I_y(\theta)^{-1})_{ij}
\end{aligned}$$

Holds for large n

$I_y(\theta) = -\mathbb{E}_\theta\left[\nabla^2 \ell(\theta)\right]$ where $\nabla^2 \ell(\theta)$ is the hessian

Can estimate standard erro of $\hat{B}_i - \hat{B}_j$ as

$$\hat{se} = \sqrt{(I_y(\theta)^{-1})_{ii} + (I_y(\theta)^{-1})_{jj} - 2\,(I_y(\theta)^{-1})_{ij}}$$

We expect $\hat{B}_i - \hat{B}_j$ to be approx. normal for large n,

$$\hat{B}_i - \hat{B}_j \pm \hat{se} \cdot z(\alpha/2)$$

②' Test $H_0: \alpha = 0$

Randomly permute (i_m, j_m) for each game

Compute a test statistic T on the permuted data

$T = -2\log\Lambda$

③' Non parametric bootstrap

Resample (i_m^*, j_m^*) with replacement

Estimate $\hat{B}_i - \hat{B}_j$ using bootstrap samples

Avoid model misspecification

4/20: Logistic Regression

Model for binary classification

Example: Predict whether a user will click on an ad

Have n ad impressions

- Binary response $Y_i = \begin{cases} 1 & \text{if clicked} \\ 0 & \text{otherwise} \end{cases}$

- Other features
  - size of ad
  - User age
  - etc.

  $\Big\}$ represented by a collection of p covariates

Logistic Regression: n responses are independent

$$Y_i \sim \text{Bernoulli}(p)$$

log odds: $\log \frac{p_i}{1-p_i} = \alpha + B_1 X_{i1} + \dots B_p X_{ip}$

α is the intercept or baseline log-odds

B_j represents the change in log odds for a one unit change in X_j

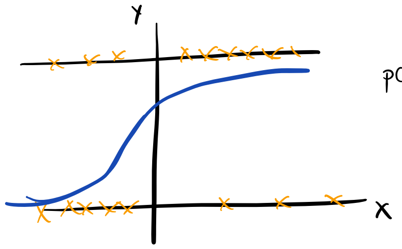
Parameters: $(\alpha, B_1, \dots B_p)$

$$\mathbb{P}[Y_i = 1] = p_i = \frac{e^{\alpha + B_1 X_{i1} + \dots B_p X_{ip}}}{1 + e^{\alpha + B_1 x_{i1} + \dots B_p x_{ip}}}$$

Assume one covariate $p=1$ for simplicity

$B = B_1$ and $X_i = X_{i1}$

Data = $(X_1, Y_1), \dots (X_n, Y_n)$



$$p(x) = \frac{e^{\alpha + Bx}}{1 + e^{\alpha + Bx}}$$

Estimation and Inference

① Estimating the regression coefficients

② Estimating the "conversion" probability $p(x) = \frac{e^{\alpha + Bx}}{1 + e^{\alpha + Bx}}$ for a new impression with covariate x

③ Provide a 95% Confidence interval for B

④ Test a null hypothesis $H_0 : B = 0$

① Estimating (α, B) via MLE

$$\text{lik}(\alpha, B) = \prod_{i=1}^{n} p_i^{Y_i} (1-p_i)^{1-Y_i} \quad \text{where} \quad p_i = \frac{e^{\alpha + Bx_i}}{1 + e^{\alpha + BX_i}}$$

$$= \prod_{i=1}^{n} (1-p_i) \left(\frac{p_i}{1-p_i} \right)^{Y_i}$$

$$\Rightarrow \ell(\alpha, B) = \sum_{i=1}^{n} Y_i \underbrace{\log \frac{p_i}{1-p_i}}_{\alpha + BX_i} + \underbrace{\log (1-p_i)}_{-\log(1+e^{\alpha + BX_i})}$$

$$= \sum_{i=1}^{n} \left(Y_i (\alpha + BX_i) - \log\left(1 + e^{\alpha + BX_i}\right) \right)$$

MLE: Set derivatives to 0

$$0 = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} (Y_i - p_i)$$

$$0 = \frac{\partial \ell}{\partial B} = \sum_{i=1}^{n} (Y_i - p_i) X_i$$

No closed form solution so we once again use Newton - Raphson Method

$$\begin{pmatrix} \alpha^{(t+1)} \\ B^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ B^{(t)} \end{pmatrix} - \left(\nabla^2 \ell (\alpha^{(t)}, B^{(t)}) \right)^{-1} \cdot \nabla \ell (\alpha^{(t)}, B^{(t)})$$

$$\nabla \ell (\alpha, B) = \begin{pmatrix} \frac{\partial \ell}{\partial \alpha} \\ \frac{\partial \ell}{\partial B} \end{pmatrix} = \sum_{i=1}^{n} (Y_i - p_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

$$\nabla^2 \ell (\alpha, B) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \alpha \partial B} \\ \frac{\partial^2 \ell}{\partial \alpha \partial B} & \frac{\partial^2 \ell}{\partial B^2} \end{pmatrix} \underset{\text{Skipped Algebra}}{\underset{\longrightarrow}{=}} \sum_{i=1}^{n} -p_i (1 - p_i) \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

Interpretation : The update $(\alpha^{(t+1)}, B^{(t+1)})$ solves a least - squares problem

$$\underset{\alpha, B}{\arg\min} \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) \left(z_i^{(t)} - (\alpha + B X_i) \right)^2$$

$$z_i^{(t)} = \alpha^{(t)} + B^{(t)} X_i + \frac{Y_i - p_i^{(t)}}{p_i^{(t)} (1 - p_i^{(t)})}$$

Check: To minimize, we would set derivatives α, B to 0

$$0 = \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)} \cdot 2 (\alpha + B X_i - z_i^{(t)})$$

$$= 2 \left[(\alpha - \alpha^{(t)}) \cdot \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) + (B - B^{(t)}) \cdot \sum_{i=} p_i^{(t)} (1 - p_i^{(t)}) X_i - \sum_{i=1}^{n} (Y_i - p_i^{(t)} \right]$$

$$0 = \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) \cdot 2 X_i (\alpha + B X_i - z_i^{(t)})$$

$$= 2 \left[(\alpha - \alpha^{(t)}) \cdot \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) X_i + (B - B^{(t)}) \sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) X_i^2 - \sum_{i=1}^{n} (Y_i - p_i^{(t)}) X_i \right]$$

$$\Leftrightarrow \sum_{i=1}^{n} (Y_i - p_i^{(t)}) \cdot \begin{pmatrix} 1 \\ X_i \end{pmatrix} = \left[\sum_{i=1}^{n} p_i^{(t)} (1 - p_i^{(t)}) \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix} \right] \cdot \begin{pmatrix} \alpha - \alpha^{(t)} \\ B - B^{(t)} \end{pmatrix}$$

Also called Iterative Reweighted Least Squares (glm in R)

② Estimate $p(x) = \frac{e^{\alpha + Bx}}{1 + e^{\alpha + Bx}}$: Use plug-in estimates

$$\hat{p}(x) = \frac{e^{\hat{\alpha} + \hat{B} x}}{1 + e^{\hat{\alpha} + \hat{B} x}}$$

③ Confidence Interval for B

Model Based Approach: Compute Fisher information to estimate std error of $\hat{B}$

Fisher Information for all n observations is

$$I_y(\alpha,B) = -\mathbb{E}\left[\nabla^2 \ell(\alpha,B)\right] = \sum_{i=1}^{n} p_i(1-p_i)\begin{pmatrix}1 & x_i\\ x_i & x_i^2\end{pmatrix} \qquad p_i = \frac{e^{\alpha+Bx_i}}{1+e^{\alpha+Bx_i}}$$

Covariance of $(\hat\alpha,\hat B)$ is approximately $I_y(\alpha,B)^{-1}$ for large n

Estimate p_i by $\hat{p}_i = \frac{e^{\hat\alpha+\hat B x_i}}{1+e^{\hat\alpha+\hat B x_i}}$ and estimate $I_y(\alpha,B)$ by

$$I_y(\hat\alpha,\hat B) = \sum_{i=1}^{n} \hat{p}_i(1-\hat{p}_i)\begin{pmatrix}1 & x_i\\ x_i & x_i^2\end{pmatrix}$$

$$\hat{se} = \sqrt{\left(I_y(\alpha,B)^{-1}\right)_{22}}$$

95% Confidence interval is $\hat{B} + z(0.025)\cdot\hat{se}$

④ To test $H_0: B=0$

GLRT: Need to compute MLE $\hat\alpha_0$ for α in the null model where $B=0$

log-likelihood:

$$\ell(\alpha) = \sum_{i=1}^{n} Y_i \underbrace{\log\frac{p_i}{1-p_i}}_{\alpha} + \underbrace{\log(1-p_i)}_{-\log(1+e^{\alpha})}$$

$$= \sum_{i=1}^{n} \left(Y_i\alpha - \log(1+e^{\alpha})\right.$$

$$0 = \frac{\partial\ell}{\partial\alpha} = \sum_{i=1}^{n}\left(Y_i - \frac{e^{\alpha}}{1+e^{\alpha}}\right)$$

$$\iff \hat\alpha_0 = \log\frac{\bar{Y}}{1-\bar{Y}}$$

GLRT test statistic is $-2\log\Lambda = 2\ell(\hat\alpha,\hat B) - 2\ell(\hat\alpha_0, 0)$

Compare against χ^2_1 null distribution

⑤ Diagnostic of model-fit is based on residual

Pearson Residual $\dfrac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$

If logistic regression is correct, we expect

mean 0

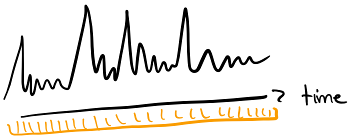
variance 1 ← Common to not follow this

uncorrelated w/ covariate X_i

Overdispersion is a common mis-specified model outcom
variance greater than 1

4/25: Poisson Log-Linear Model + Generalized Linear Model

Example: Neuron spiking overtime



Basic Model: Poisson Process

# spikes in the ith bin = $Y_i \sim \text{Poisson}(\lambda_i, \Delta)$ independent for $i=1,2,\dots n$

- WLOG $\Delta = 1$
- λ_i is the spiking rate in the ith bin (Depends on external Stimuli)
- Encode stimuli using p covariates

Poisson Log-linear model

$$\log \lambda_i = \alpha + B_1 X_{i_1} + \dots B_p X_{i_p}$$

α is the intercept, $B_1 \dots B_p$ are the regression coefficients

Equivalently,

$$\lambda_i = e^{\alpha + B_1 X_{i_1} + B_p X_{i_p}}$$

The responses are then $Y_i \sim \text{Poisson}(\lambda_i)$ independent for $i=1,\dots,n$

To simplify notation, $p=1$

Data: $(X_1, Y_1), \dots, (X_n, Y_n)$

Consider fixed X's with random responses Y_i

Model: $Y_i \sim \text{Poisson}(\lambda_i)$ $\lambda_i = e^{\alpha + BX_i}$

Estimation and Inference

- Estimate regression coefficients
- Provide 95% Confidence interval for B

Likelihood Function:

$$lik(\alpha, B) = \prod_{i=1}^{n} \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!} \quad \text{where} \quad \lambda_i = e^{\alpha + Bx_i}$$

$$\ell(\alpha, B) = \sum_{i=1}^{n} Y_i \underbrace{\log(\lambda_i)}_{\alpha+BX_i} - \lambda_i - \log(Y_i!)$$

$$= \sum_{i=1}^{n} Y_i(\alpha + BX_i) - e^{\alpha+BX_i} - \log(Y_i!)$$

Solving MLE via derivatives

$$0 = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \left(Y_i - e^{\alpha+BX_i}\right) = \sum_{i=1}^{n} (Y_i - \lambda_i)$$

$$0 = \frac{\partial \ell}{\partial B} = \sum_{i=1}^{n} (Y_i - \lambda_i) X_i$$

No closed form so we can apply Newton-Raphson Method

Gradient of $\ell(\alpha,B)$:

$$\nabla\ell(\alpha,B) = \begin{pmatrix} \frac{\partial\ell}{\partial\alpha} \\ \frac{\partial\ell}{\partial B} \end{pmatrix} = \sum_{i=1}^{n} (Y_i - \lambda_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

Hessian of $\ell(\alpha,B)$

$$\nabla^2\ell(\alpha,B) = \begin{pmatrix} \frac{\partial^2\ell}{\partial\alpha^2} & \frac{\partial^2\ell}{\partial\alpha\partial B} \\ \frac{\partial^2\ell}{\partial\alpha\partial B} & \frac{\partial^2\ell}{\partial B^2} \end{pmatrix} = -\sum_{i=1}^{n} w_i \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix} \qquad w_i = \lambda_i = e^{\alpha + BX_i}$$

=> Newton Raphson Iterations

$$\begin{pmatrix} \alpha^{(t+1)} \\ B^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ B^{(t)} \end{pmatrix} - \left(\nabla^2\ell\left(\alpha^{(t)}, B^{(t)}\right)\right)^{-1} \cdot \nabla\ell\left(\alpha^{(t)}, B^{(t)}\right)$$

Newton - Raphson update solves a weighted least- squares regression

Also called iterative reweighted least square

To make a model based 95% Confidence Interval for B

- Fischer Information of all n observations

$$I_y(\alpha,B) = -\mathbb{E}\left[\nabla^2\ell(\alpha,B)\right]$$

← full log-likelihood

(Can't use n·I since each is different)

$$= \sum_{i=1}^{n} \lambda_i \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

Use plug-in estimate

$$I_y(\hat\alpha,\hat B) = \sum_{i=1}^{n} \hat\lambda_i \begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}$$

Estimate standard erro of $\hat B$ by

$$\hat{se} = \sqrt{\left(I_y(\hat\alpha,\hat B)\right)^{-1}_{22}}$$

Model Diagnostics:

- Based on normalized Residuals

Pearson Residual : $\dfrac{Y_i - \hat\lambda_i}{\sqrt{\hat\lambda_i}}$

If model is correctly specified we expect

mean 0

variance 1

uncorrelated with X_i

Generalized Linear Model

A GLM is a model for responses $Y_i \dots Y_n$ where $Y_i \sim f(y|\theta)$, $\theta \in \mathbb{R}$

θ_i is the value for observation i

For observation i we observe p covariates $x_i \dots x_{ip}$

GLM assumes

$$g(\theta_i) = \alpha + B_1 x_{i1} + \dots B_p x_{ip}$$

$g: \mathbb{R} \to \mathbb{R}$ is the link function

In logistic regression, g -> log odds

poisson log-linear, g -> log d

Alternative link functions

$$g(p) = \Phi^{-1}(p)$$

Φ is the standard normal quantile function

Choice of link function

- Goodness of model fit to data
- Interpretation of model / parameters
- Mathematical convinience

Change of variable $\theta \to \eta(\theta)$ so pdf/pmf has the form

$$f(y|\eta) = e^{\eta y - A(y)} \cdot h(y)$$ ← Canonical or natural link

- For Bernoulli(p): The PMF is

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

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

Set $\eta(p) = \log\frac{p}{1-p}$, $A(\eta) = -\log(1-p) = \log\left(1+e^{\eta}\right)$ and $h(y)=1$

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$

- For poisson: The PMF is

$$f(y) = \frac{\lambda^y e^{-y}}{y!} = e^{(\log\lambda) y - \lambda} \cdot \frac{1}{y!}$$

Set $\eta(\lambda) = \log\lambda$ $A(\eta) = \lambda$ and $h(y) = \frac{1}{y!}$

$$\lambda = e^{\eta}$$

$f(y|\eta) = e^{\eta y - A(\eta)} h(y)$ is called the exponential family form of the model

η is the natural parameter of the model

For a GLM based on a parametric model $f(Y|\theta)$, the natural/canonical link is the choice $g(\theta) = \eta(\theta)$ where η is natural parameter

If we use the natural link:

Log-likelihood function for $(X_1, Y_1) \dots (X_n, Y_n)$ is

$$\ell(\alpha, \beta) = \log \prod_{i=1}^{n} e^{\eta_i Y_i - A(\eta_i)} h(Y_i)$$

$$= \sum_{i=1}^{n} Y_i \underset{\substack{\uparrow \\ \alpha + \beta x_i \text{ when we use the natural link}}}{\eta_i} - A(\eta_i) + \log(h(Y_i))$$

$$= \sum_{i=1}^{n} Y_i(\alpha + \beta x_i) - A(\alpha + \beta x_i) + \log h(y_i)$$

Computing MLE:

$$0 = \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} Y_i - A'(\eta_i)$$

$$\eta_i = \alpha + \beta x_i$$

$$0 = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} Y_i x_i - x_i A'(\eta_i)$$

$$\Rightarrow \quad 0 = \sum_{i=1}^{n} (Y_i - A'(\eta_i)) \begin{pmatrix} 1 \\ x_i \end{pmatrix}$$

What is $A'(\eta)$

$$1 = \sum_y f(y|\eta) = \sum_y e^{\eta y - A(\eta)} h(y)$$

differentiate w.r.t y

$$0 = \sum_y (y - A'(\eta)) \cdot \underbrace{e^{\eta y - A(\eta)} h(y)}_{f(y|\eta)}$$

$$= \mathbb{E}[Y] - A'(\eta)$$

$$\Rightarrow A'(\eta) = \mathbb{E}[Y]$$

For GLM, $A'(\eta) = \mathbb{E}[Y_i]$ is the model prediction for mean of Y_i and $Y_i - A'(\eta)$ is a residual

Fisher Information is

$$I_{y}(\alpha, B) = -\mathbb{E}\left[\nabla^2 \ell(\alpha, B)\right]$$

$$= \sum_{i=1}^{n} w_i \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \quad \text{where } x_i = A''(\mu_i)$$

4/27: Proportional Hazards Model

Example: Clinical trial studying the effect of a cancer drug produces data

$T_1, T_2, \dots T_n$

where T_i is the time of remission

For the ith patient we have p covariates

ex. Treatment vs. Control

Stage of Cancer

Age

Family history

Goal: Model T_i via $x_{i1}, \dots x_{ip}$

Let T be a continuous variable

Hazard Function

$$\lambda(t) = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}\left[T \in [t, t+\delta] \ T \geq t\right] \quad \leftarrow \text{Instantaneous Risk}$$

$$\lambda(t) = \lim_{\delta \to 0} \frac{\frac{1}{\delta}\mathbb{P}\left[T \in [t, t+\delta] \ T \geq t\right]}{\mathbb{P}[T \geq t]} = \frac{f(t)}{1 - F(t)}$$

Conditional probability

Example: When $T \sim \text{Exponential}(\theta)$

$$f(t) = \theta e^{-\theta t} \qquad \lambda(t) = \frac{\theta e^{-\theta t}}{1 - (1 - e^{-\theta t})} = \theta$$

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

Hazard function is constant in time ← Memoryless Property!!

In general the hazard may depend on time and covariates for each patient

Model T_i via

$$\lambda_i(t) = \underset{\substack{\uparrow \\ \text{time dependence}}}{\lambda(t)} \exp(\underset{\substack{\uparrow \\ \text{covariates}}}{B_1 x_{i1} + \dots B_p x_{ip}})$$

Same time dependence for each patient

Shape of hazard function will have the same shape over time but it will be scaled based on the covariates

Origin of the name proportional hazards

This model is semi-parametric

parametric component: Regression Coefficient

Non-parametric component: hazard function

Usually we care more about $B_1 \dots B_p$ than $\lambda(t)$

Idea: Condition on the set of all observed reoccurance terms

$t_{(1)} < t_{(2)} < \dots < t_{(n)}$

Fixes times at which the n reocurrence events occured but not the patient

Inference for $B_1 \dots B_p$ will then be based on only the information about which patient had reoccurance

Akin to permutation tests

For each $t_{(k)}$, let $R_{(k)}$ be the at risk set immediately before time $t_{(k)}$ (Patients still in remission before t_k)

Conditional on some patient in $R_{(k)}$ having reoccurence at time $t_{(k)}$, the probability it is patient i for $i \in R_{(k)}$

$$\frac{\lambda_i(t_{(k)})}{\sum_{j \in R_{(k)}} \lambda_j(t_{(k)})}$$

← Ratio of the instantaneous rate of reoccurence in patient i to the sum of risk for all patients

The baseline $\lambda(t)$ cancels and we find

$$\frac{\lambda_i(t_{(k)})}{\sum_{j \in R_{(k)}} \lambda_j(t_{(k)})} = \frac{\exp(B_1 x_{i_1} + \dots B_p x_{ip})}{\sum_{j \in R_{(k)}} \exp(B_1 x_{j_1} + \dots B_p x_{jp})}$$

For each time $t_{(k)}$ $k = 1, \dots, n$

Partial likelihood function

$$\text{plik}(B_1 \dots B_p) = \prod_{k=1}^{n} \frac{\exp(B_1 x_{i_{k_1}} + \dots B_p x_{i_k p})}{\sum_{j \in R_{(k)}} \exp(B_1 x_{j_1} + \dots B_p x_{jp})}$$

Use partial likelihood in place of actual likelihood

Note: Typically the responses $T_1 \dots T_n$ are right-censored

If the i^{th} patient is still in remission by the end of the trial we do not know T_i

just that $T_i \geq l_i$

l_i is the duration of the trial

If cancer never reccurs, then $T_i = \infty$

When some responses are right-censored, the at risk set $R_{(k)}$ is defined as the set of patients

- still in remission
- Not right censored

Maximum Likelihood Estimation

Assume $p=1$ for simplicity

Data: $(X_1, T_1) \dots (X_n, T_n)$

log-partial-likelihood

$$\ell(B) = \log \prod_{k=1}^{n} \frac{\exp(BX_{i_k})}{\sum_{j \in R_{(k)}} \exp(BX_j)} = \sum_{k=1}^{n} \left(BX_{i_k} - \log \sum_{j \in R_{(k)}} \exp(BX_j) \right)$$

$$\hat{B} = \arg\max \ell(B)$$

Solve via Newton - Raphson

Variance is estimated by

$$I(\hat{B})^{-1} \simeq \left(\frac{-\partial^2 \ell(\hat{B})}{\partial B^2} \right)^{-1}$$

Test of $H_0 : B = 0$ based on GLRT statistic

$$-2 \log \Lambda = 2\ell(\hat{B}) - 2\ell(0)$$

w/ distribution χ^2_1

Course Review

Hypothesis Testing: Deciding whether a particular null hypothesis about the underlying distribution is true/false

Estimation: Estimating quantities/parameters related to this distribution

Hypothesis Testing

Binary decision to accept/reject H_0 based on data $X_1 \dots X_n$ using a test statistic $T(X_1 \dots X_n)$

Step 1: How to choose T?

Step 2: Deciding whether H_0 is true/false based on T?

Neyman- Pearson Lemma: Maximize power via likelihood ratio statistic

If monotonic increasing/ decreasing we can use its input in its place

Simple Alternatives

Composite Alternatives: