Mathematical Statistics (ST2132)

Bennett Clement

Bird's eye view of what we are trying to do.

1. Model Choice. From experience / prior knowledge, we pick a suitable underlying distribution of a phenomenon.
2. Model Calibration. Given the data (samples), and the model from stage 1, we proceed to **estimate** the parameters of the model
3. Model Validation. We test the goodness of fit
4. Other models?
5. Model Selection. We **test our hypotheses**
6. Modify for future

Some Mathematical Preliminaries

Let $U = \max\{X_1, \dots X_n\}$ and $V = \min\{X_1, \dots X_n\}$, then

1. For a given u, $U \leq u$ iff $X_i \leq u$ for all i.
2. For a given v, $V \geq v$ iff $X_i \leq v$ for all i.

$\int_{-\infty}^{\infty} e^{-x^2/2} \mathrm{d}x = \sqrt{2\pi}$.
$\int_{-\infty}^{\infty} xe^{-x^2/2} \mathrm{d}x = 0$.
$\int_{-\infty}^{\infty} x^2 e^{-x^2/2} \mathrm{d}x = \sqrt{2\pi}$.

$\sum (X_i - \mu_0)^2 = \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$.
$\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$

RV (Chapter 2)

Discrete RV

Bernoulli RV

$p(x) = p^x (1-p)^{1-x}$ if $x = 0$ or $x = 1$ and 0 otherwise.

Binomial Distribution

n trial with k successes

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

- $E(X) = np; \quad Var(X) = np(1-p); \quad M(t) = (1 - p + pe^t)^n$

Negative Binomial Distribution

k trial until there are r successes

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

- **Geometric distribution** is a special case when $r=1$
- $E(X)=\frac{r}{p}; \quad Var(X)=\frac{r(1-p)}{p^2}; \quad M(t)=\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$

Poisson Distribution

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, \quad k=0,1,2,\dots$$

- Can be used to **approximate binomial probabilities** for large n and small p. Usually when $n \geq 20$ and $np \leq 5$ or $nq \leq 5$
- Model assumption:
  1. Underlying rate (λ) at which the events occur is constant in space / time
  2. Events in disjoint intervals of space / time occur independently
  3. There are no multiple events (in very short time intervals).
- Used in analysis of telephone systems, modeling number of alpha particles emitted from radioactive source during a period of time, model number of freak accidents for large population during a period of time
- If $X \sim \text{Poisson}(\alpha), Y \sim \text{Poisson}(\beta)$ and X, Y are independent, then $X+Y \sim \text{Poisson}(\alpha+\beta)$
- $E(X)=\lambda=Var(X); \quad M(t)=exp\{\lambda(e^t-1)\}$

Continuous RV

Density function $f(x)$ satisfies $f(x) \geq 0$, f is piecewise continuous and $\int_{-\infty}^{\infty} f(x)dx=1$

Exponential Distribution

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x \leq 0\end{cases}$$

- Used to model lifetimes or waiting times
- $P(T>t)=e^{-\lambda t}$
- Has the memoryless property: $P(T>t+s \mid T>s)=P(T>t)$
- Special case of Gamma with $\alpha=1$

Chi-square with n degree of freedom

Special case of Gamma distribution with $\alpha=n/2$ and $\lambda=1/2$

Gamma Distribution

The gamma density function depends on 2 parameters, $\alpha > 0$ (shape) and $\lambda > 0$ (scale)

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

where $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du, t \geq 0$

- $E(X) = \frac{\alpha}{\lambda}; \quad Var(X) = \frac{\alpha}{\lambda^2}; \quad M(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, \, t < \lambda$
- For positive integer arguments, $\Gamma(n) = (n-1)!$. For positive half integers, $\Gamma(n/2) = \sqrt{\pi}(n-2)!!/2^{(n-1)/2}$. In general, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$.
- (Problem 12) If $U \sim Gamma(\alpha, \lambda)$ and $c > 0$, then $cU \sim Gamma(\alpha, \lambda/c)$.
- (Ex 9. MGF) If $X \sim Gamma(\alpha_1, \lambda)$ and $Y \sim Gamma(\alpha_2, \lambda)$, then $X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$
- (Tut 5) $E(X^r) = \frac{\Gamma(\alpha+r)}{\lambda^r \Gamma(\alpha)}$ if $r > -\alpha$.
- Quite flexible to model **non-negative** RV

Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 \,/\, 2\sigma^2}, \qquad -\infty < x < \infty$$

- $M(t) = e^{t^2/2}$ for $X \sim N(0,1)$

Functions of RV

(Problem 13). If $Y = aX + b$, then $f_Y(y) = \frac{1}{|x|} f_X\left(\frac{y-b}{a}\right)$.

The above result can be used to prove $E(aX + b) = aE(X) + b$.

More generally we can say if g is a differentiable, strictly monotonic function on an interval I, then $Y = g(X)$ has the density function

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

for y such that $y = g(x)$ for some x

Joint Distribution (Chapter 3)

Motivating example

A model for the joint distribution of age and length in population of fish. Length distribution can be used to estimate the age distribution so as to set reasonable harvesting policies.

Definition

If $X_1, \ldots, X_n$ are jointly distributed RV, their joint cdf is

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n).$$

Discrete RV

Joint frequency function

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

Marginal frequency function

$$p_X(x) = \sum_j p(x, y_j) \quad \text{and} \quad p_Y(y) = \sum_i p(x_i, y)$$

Multinomial Distribution

A generalization of binomial distribution, with each independent trials can result in one of r types of outcomes, each with a probability of p_i

$$p(x_1, x_2, \ldots, x_r) = \binom{n}{x_1 \ldots x_r} p_1^{x_1} p_2^{x_2} \ldots p_r^{x_r} \quad \text{s.t.} \quad x_1 + x_2 + \cdots + x_r = n$$

Furthermore, the marginal frequence function for any particular X_i is given by the binomial distribution $B(n, p_i)$

Cont RV

$$P((X, Y) \in A) = \iint_A f(x, y) \, dy \, dx$$

Independent RV

Definition 3. RV $X_1, X_2, \ldots X_n$ are independent if their joint cdf factors into the product of their marginal cdf's:

$$F(x_1, x_2, \ldots x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_n}(x_n) \quad \forall x_1, x_2 \ldots x_n$$

If X and Y are independent, then

- $P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$
- $g(X)$ and $h(Y)$ are independent for functions g and h.

Quotient of 2 Cont RV

Suppose that X and Y are continuous with joint density function f and that $Z = \frac{Y}{X}$. Then

$$
\begin{aligned}
F_Z(z) = P(Z \le z) = P(X \le 0, Y \ge Xz) + P(X \ge 0, Y \le Xz) \\
= \int_{-\infty}^{z} \int_{-\infty}^{\infty} |x| f(x, xv)\, dx\, dv
\end{aligned}
$$

Differentiating $F_Z(z)$ under an assumption of continuity, we find

$$f_Z(z) = \int_{-\infty}^{\infty} |x| f(x, xz)\, dx$$

In particular, if X and Y are independent,

$$f_Z(z) = \int_{-\infty}^{\infty} |x| f_X(x) f_Y(xz)\, dx$$

Extrema

Assume $X_1, X_2, \dots, X_n$ are independent RVs with common cdf F and density f. Let $U = max\{X_1, X_2, \dots, X_n\}$

$$F_U(u) = P(U \le u) = P(X_1 \le u, X_2 \le u, \dots X_n \le u) = F(u)^n$$

Differentiating, we get the density $f_U(u) = nf(u)F(u)^{n-1}$

Example: n system components connected in series with lifetimes independent exponential(λ) RV. The minimum lifetime V is exponentially distributed with parameter $n\lambda$. See Textbook pp 105.

Expected Value, Variance (Chapter 4)

Expected Value

If the E.V diverges, the E.V is said to be undefined.
Properties

- If X and Y are independent RV, $E[g(X)h(Y)] = E[g(X)] \times E[h(Y)]$
- $E(a + \sum_{i=1}^{n} b_i X_i) = a + \sum_{i=1}^{n} b_i E(X_i)$ regardless of whether X_i are independent.

Variance

- $Var(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$
- $Var(a + bX) = b^2\, Var(X)$
- $Var(aX + bY) = a^2\, Var(X) + b^2\, Var(Y) + 2ab\, Cov(X, Y)$ where $Cov(X, Y) = E(X - \mu_X)E(Y - \mu_Y) = E(XY) - E(X)E(Y)$

Chebyshev's Ineq

Thm 7. Let X be a RV with mean μ and variance σ^2, then for any $t > 0$, $P(|X - \mu| > t) \le \frac{\sigma^2}{t^2}$

Moment Generating Function

Definition: $M(t) = E[e^{tX}]$
Properties:

1. If the mgf exists for t in an open interval containing zero, it uniquely determines the probability distribution.
2. If the mgf exists in an open interval containing zero then $M^{(r)}(0) = E(X^r)$, the rth moment of the distribution.
3. If $Y = a + bX$, then Y has the mgf $M_Y(t) = e^{at}M_X(bt)$
4. If X and Y are independent RV, and $Z = X + Y$, $M_Z(t) = M_X(t)M_Y(t)$ on the common interval where both mgf's exist.

Limit Theorems (Chapter 5)

Weak Law of Large Numbers. Let (X_i) be a sequence of independent RV with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \longrightarrow_P \mu$$

Definition. **Convergence in Distribution**. Let $X, X_1, X_2, \ldots$ be a sequence of RV with cdf's $F, F_1, F_2, \ldots$. We say that X_n converges in distribution to X if $F_n(x) \to F(x)$ as $n \to \infty$

Central Limit Theorem (CLT): Let (X_i) be a sequence of independent RV with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, a common c.d.f F and m.g.f M defined in a neighbourhood of zero. Let $S_n = \sum_{i=1}^{n} X_i$. Then, the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the standard normal distribution.

Personal remark: How fast the approximation becomes good depends on the distribution of the summands X_i. This is the reason why we can approximate the value of $\bar{X_n}$ using the normal distribution only when $n \geq 30$, otherwise use t-dist.

Sampling Dist (Chapter 6)

chi-square dist

Definition 8. If Z is a std normal RV, the distribution $U = Z^2$ is the $\chi_1^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$

Definition 9. If $U_1, U_2, \ldots, U_n$ are independent χ_1^2 RV, the distribution of $U_1 + U_2 + \cdots + U_n = V \sim \chi_n^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$.

t dist

Definition 10. If $Z \sim N(0,1)$ and $U \sim \chi_n^2$ are independent, then

$$T = \frac{Z}{\sqrt{U/n}}$$

is called the t-distribution with n degrees of freedom.

Proposition 3. The density function of the t_n distribution is given as

$$f(t) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty$$

Properties:

1. Symmetric about 0. $f(t) = f(-t)$.
2. The t_1 distribution is called the **Cauchy distribution**
3. $t_n \to N(0,1)$ as $n \to \infty$.

F dist

Definition 11. If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then the distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom.

Proposition 4. The density function of th $F_{m,n}$ distribution is

$$f(w) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2} w^{\frac{m}{2}-1}\left(1+\frac{mw}{n}\right)^{-(m+n)/2}, \quad w \geq 0$$

Properties:

- $E(W) = \frac{n}{n-2}$ for $n > 2$.
- Let $T \sim t_n$, then $T^2 \sim F_{1,n}$

Sample Mean and Variance

Let $X_1, \ldots X_n$ be independent $N(\mu, \sigma^2)$ RV. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Thm 10.The RV $\bar{X}$ and the vector of RV $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ are independent.

Corollary 4. $\bar{X}$ and S^2 are independently distributed.

Thm 11. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Corollary 5. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Estimation (Chapter 8)

Background

Many families of probability laws depend on a small number of parameters. Unless the values of the parameters are known in advance, they must be **estimated** from the observed data. After parameter values have been chosen, the model should be compared to the data to see if the **fit** is reasonable.

We regard the observed data as realizations of random variables $X_1, X_2, \ldots, X_n$ whose joint distribution depends on an unknown parameter θ (possibly vector valued).

If the X_i are independent RVs having the same distribution $f(x|\theta)$ -- **independent and identically distributed (i.i.d)** -- their joint distribution $f(x_1, \ldots, x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$

An estimator x of θ will be a function of $X_1, X_2, \ldots, X_n$, a RV with a **sampling distribution**.

The idea is that given a choice of many estimation procedures, we would like to use that estimate whose sampling distribution is most concentrated around the true parameter value.

Unbiased estimator $\hat{\theta}$ iff $E(\hat{\theta}) = E(\theta)$.

Method of Moments (MoM)

Definition 12.

1. The k-th moment of a probability law $\mu_k = E(X^k)$ where X is a RV following that law.
2. If $X_1, X_2, \ldots, X_n$ are **i.i.d** RV from that distribution, the k-th **sample moment** is defined as $\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

We can view $\widehat{\mu}_k$ as an estimate of μ_k.

To construct a method of moments estimate:

1. Express low-order moments (usually μ_1 and μ_2 is sufficient) in terms of the parameters
2. Invert the previous expression(s) to express parameters in terms of moments
3. Insert sample moments to obtain estimates of the parameters

Consistency

Definition 13. Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is **consistent** in probability if $\hat{\theta}_n$ converges to θ as $n \to \infty$.

Consistency of MoM: The Weak Law of Large Numbers ensures that the sample moments converge in probability to the population moments.

Summary

The method of moments can provide estimates of parameters of a probability distribution based on a "sample" (an i.i.d. collection) of RV from that distribution. To address the reliability of estimates, we can approximate the **std error** of the estimate (a.k.a. std dev of the sampling distribution) by (in some cases) substituting our estimates for unknown parameters.

Method of Maximum Likelihood (MLE)

Definition 14. Suppose RVs $X_1, \ldots, X_n$ have a joint density function $f(x|\theta) = f(x_1, \ldots, x_n|\theta)$. Given observed values $X_i = x_i$, the **likelihood** of θ is defined as $lik(\theta) = f(x_1, \ldots, x_n|\theta)$.

Comment: likelihood = joint density function.

The **maximum likelihood estimate (mle)** of θ is that value of θ that maximizes the likelihood -- that is, makes the observed data "most probable".

It might be easier to maximize the **log likelihood** $l(\theta) = \log[lik(\theta)]$. If the X_i are assumed to be i.i.d, $l(\theta) = \sum\limits_{i=1}^{n} \log[f(X_i|\theta)]$

Comment: This method is more precise than the Method of Moments

MLE of Multinomial Cell Probabilities

Suppose that X_i are the counts in cells $i = 1, \ldots, m$, that follows a multinomial distribution with total count of n and cell probabilities $p_i(\theta)$ s.t. $\sum\limits_{i=1}^{m} p_i(\theta) = 1$. Recall that the X_i are not independent, and the joint frequency function of $X_1, \ldots, X_m$ is $f(x \mid p_1, \ldots, p_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}$.

So, the log likelihood of θ is

$$l(\theta) = n! - \sum_{i=1}^{m} \log X_i! + \sum_{i=1}^{m} X_i \log p_i(\theta)$$

Next, we find the value of θ where $l(\theta)$ achieves its minimum. We will get $\hat{p}_j = X_j/n$. (When taking the first derivative, don't forget to take into account that $\sum p_i(\theta) = 1$)

Large Sample Theory for MLEs

We define the **Fisher information** (in one observation) $I(\theta)$ by

$$I(\theta) = E\left\{ \left[\frac{\partial}{\partial\theta} \log f(X \mid \theta) \right]^2 \right\}$$

Lemma: Under appropriate smoothness conditions on f, $I(\theta)$ may also be expressed as

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$$

Theorem 12. Under appropriate smoothness conditions on f,

1. The mle $\hat{\theta}$ from an i.i.d sample is consistent.
2. The probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution

Comments:

1. The large sample distribution of an mle is approximately normal with mean θ_0 and variance $1/[nI(\theta_0)]$. The result also holds for non-iid samples if we replace the Fisher information for i.i.d samples, i.e. $nI(\theta_0)$, with the Fisher information for a general sample (see comment 4).
2. For a large sample, the mle is said to be **asymptotically unbiased** and $1/[nI(\theta_0)]$ is called the **asymptotic variance**.
3. IMPORTANT: The above theorem does not hold if the support [1] of f depends on θ (because the lemma holds only if differentiation and integration can be interchanged)
4. In general, the Fisher information of a sample (which need not be i.i.d) is given directly by $E[l'(\theta)^2]$ or $-E[l''(\theta)]$ where $l(\theta)$ is the log likelihood, and the asymptotic variance is given by $1/E[l'(\theta)^2]$ or $-1/E[l''(\theta)]$.

Confidence intervals from MLEs

Definition 15. A $100(1-\alpha)\%$ confidence interval for a population parameter θ is a random interval, calculated from the sample that contains θ with $1-\alpha$ probability (also called coverage probability).

Based on the large sample theory, $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim N(0,1)$ approximately. Since θ_0 is unknown, we will use $I(\hat{\theta})$ in place of $I(\theta_0)$. An **approximate** $100(1-\alpha)\%$ confidence interval for θ is $\hat{\theta} \pm z_{\alpha/2}/\sqrt{nI(\hat{\theta})}$

Obtaining an **exact** confidence interval requires detailed knowledge of the sampling distribution of the point estimates. Such knowledge may allow us to identify a pivotal quantity or **pivot**[2]. Some examples: (Thm 11) $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Efficiency and the Cramer-Rao Lower Bound

Definition 16.

1. The **mean squared error** of $\hat{\theta}$ as an estimate of θ_0 is $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2] = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta_0]^2$ (variance + bias squared)
2. Given two estimates α and β of a parameter θ, the **efficiency** of α relative to β is defined as $eff(\alpha, \beta) = Var(\beta)/Var(\alpha)$

Comments:

1. If the estimate $\hat{\theta}$ is unbiased, $MSE(\hat{\theta}) = Var(\hat{\theta})$
2. The second comparison is most meaningful when both α and β are unbiased or when both have the same bias.

Theorem 13 (Cramer-Rao Inequality). Let $X_1, \ldots X_n$ be i.i.d samples with density function $f(x \mid \theta)$. Let $T = t(X_1, \ldots X_n)$ be an unbiased estimate of θ. Under appropriate smoothness assumptions on f, we have $Var(T) \geq \frac{1}{nI(\theta)}$

Comments:

- The theorem gives a **lower bound** on the variance of **any** unbiased estimate
- An unbiased estimate whose variance achieves this lower bound is said to be **efficient**. In particular, mle's are said to be **asymptotically efficient**.

Sufficiency

Motivation: (TLDR: data compression).

Given a sample $X_1, \ldots, X_n$ from a probability distribution indexed by θ, is there a statistic $T(X_1, \ldots, X_n)$ which contains all the information in the sample about θ ? If so, a reduction of the original data to this statistic without loss of information is possible.

Definition 17. A statistic $T(X_1, \ldots X_n)$ is said to be **sufficient** for θ if $P(X = (x_1, \ldots, x_n) \mid T = t)$ does not depend on θ for any value of t. T is called a **sufficient statistic**.

Example. Given a sequence of n independent Bernoulli RV with $P(X_i = 1) = p$, then $T = \sum\limits_{i=1}^{n} X_i$ is a sufficient statistic.

Remark: We regard 2 sufficient statistics as equivalent if one can be achieved via a series of monotone transformations of the other.

Thm 14. (Factorization Theorem). A necessary and sufficient condition for $T(X)$ to be sufficient for a parameter θ is that the joint probability function factors in the form $f(\mathbf{x} \mid \theta) = g[T(\mathbf{x}), \theta] \cdot h(\mathbf{x})$

Corollary 6. If T is sufficient for θ, then the mle is a function of T.

Note: The number of sufficient statistics need not be the same as the number of parameters regardless of sample size, e.g. the Cauchy distribution with location parameter θ (t dist with 1 degree of freedom with $f(x \mid \theta) = 1/(\pi[1 + (x - \theta)^2])$.

Exponential family of distributions

We have noted previously that the number of sufficient statistics need not be the same as the number of parameters.

We call the set of probability distribution in which the sufficient statistics have the same dimension (k) as the parameter space regardless of sample size as the k-parameter **exponential family** of probability distributions.

A k-parameter member of the exponential family has a density or frequency function of the form

$$f(x \mid \theta) = \begin{cases} \exp\left[\sum_{i=1}^{k} c_i(\theta)T_i(x) + d(\theta) + S(x)\right] & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

where the set A does not depend on θ.

Example: The Bernoulli, Binomial and Poisson distributions are member of the 1-parameter exponential family. The Normal and Gamma distributions are member of the 2-parameter exponential family.

Thm 15. (Rao-Blackwell Theorem). Let α be an estimator of θ with $E(\alpha^2) < \infty$ for all θ. Suppose T is sufficient for θ and let $\beta = E(\alpha \mid T)$. Then, for all θ, $E[(\beta - \theta)^2] \leq E[(\alpha - \theta)^2]$.

The inequality previous is strict except when $\alpha = \beta$.

Comment: The Rao-Blackwell theorem gives a strong rationale for basing estimators on sufficient statistics if they exist. If an estimator is not a function of a sufficient statistic, then it can be improved.

Testing Hypotheses (Chapter 9)

Motivation: Bayesian Approach

Given an observation $\mathbf{x}$ and 2 hypotheses H_0, H_1. We would like to assess the evidence for each of the hypotheses. We would accept H_0 if H_0 is more likely, i.e.

$$\frac{P(H_0 \mid \mathbf{x})}{P(H_1 \mid \mathbf{x})} = \frac{P(H_0)}{P(H_1)} \frac{P(\mathbf{x} \mid H_0)}{P(\mathbf{x} \mid H_1)} > 1$$

Equivalently, we accept H_0 if $\frac{P(\mathbf{x}|H_0)}{P(\mathbf{x}|H_1)} > c$ where the critical value c depends on the ratio between $P(H_1)$ and $P(H_0)$.

However, this approach requires the specification of **prior probabilities**[3] $P(H_0)$ and $P(H_1)$ before observing any data.

Neyman-Pearson Paradigm

Neyman and Pearson formulated their theory of hypothesis testing as a decision problem. One hypothesis is singled out as **null hypothesis** (H_0) and the other as the **alternative hypothesis** (H_1).

There is an asymmetry in the Neyman-Pearson's approach between the null and alternative hypotheses because, unlike the Bayesian approach, which requires the distribution under H_0 and H_1, as well as prior probabilities to be known, this approach only requires the distribution under H_0 in order to construct a test.

The decision as to which is the null is not a mathematical one, and depends on the context, custom and convenience, e.g. H_0 is simpler, the consequence of incorrectly rejecting H_0 may be graver than those incorrectly rejecting H_1, etc.

The following terminology is standard:

1. **Type I error** -- rejecting H_0 when it is true
2. The probability of a type I error (α) is called the **significance level** of the test.
3. **Type II error** -- accepting H_0 when it is false. Its probability is denoted by β.
4. The probability that H_0 is rejected when it is false is called the **power** of the test; it equals $1 - \beta$.
5. The **test statistic** is the thing that we test (obvious?!). E.g. the likelihood ratio $\frac{P(x|H_0)}{P(x|H_1)}$, number of heads in a coin toss, etc.
6. The set of value of the test statistic that leads to rejection of H_0 is called the **rejection region**, and the set of values that leads to acceptance is called the **acceptance region**.
7. The probability distribution of the test statistic when H_0 is true is called the **null distribution**.
8. The p**-value** is the smallest significance level at which H_0 would be rejected. **The smaller the p-value, the stronger the evidence against H_0.** Indeed, the p-value can be thought of as the probability under H_0 of a result as or more extreme than that actually observed.
9. A hypotheses that completely specifies the probability distribution (distribution and parameter) is called **simple hypotheses**. A hypotheses that does not completely specifies the probability distribution is called **composite hypothesis**.
10. The optimal test is the test that is at least as powerful as any other test for a same significance level α.

Thm 16. (Neyman-Pearson Lemma) Suppose that H_0 and H_1 are **simple** hypotheses and the test that rejects H_0 for small values of likelihood has significance level α. Then any other test for which the significance level is $\leq \alpha$ has $\leq$ power than that of the likelihood ratio test.

In other words, Neyman-Pearson Lemma states that the likelihood ratio test minimizes P(type II error) among all tests with a given P(type I error).

Likelihood ratio test

1. We write down the likelihood ratio : $f_0(\mathbf{x})/f_1(\mathbf{x}) = f(\mathbf{x} \mid H_0)/f(\mathbf{x} \mid H_1)$
2. Observe that small values of LR corresponds in a 1-to-1 manner with extreme values of a test statistic. (We reject statistic with likelihood ratio less than a value c)
3. Knowing the null distribution of the test statistic and the desired significance level α , we can derive the rejection region (i.e. finding the value of an unknown **critical level** x_0).

Uniformly most powerful tests

The optimality result of the Neyman-Pearson Lemma requires that both hypotheses be simple. In some cases, the theory can be extended to include composite hypotheses.

If H_1 is composite, we say a test to be **uniformly most powerful** if it is the most powerful for every simple alternative in H_1, (i.e. the test do not depend on the value given to H_1)

Remark: In typical composite situations, there is no uniformly most powerful test.

Duality of CI and Hypo Test

Idea: The confidence interval consists *precisely* of all those values of μ_0 for which $H_0 : \mu = \mu_0$ is accepted (proved by the following 2 theorems).

Thm 17. Suppose we have a test for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. As we vary θ_0, the values θ_0, for which H_0 is not rejected at level α form a $100(1-\alpha)\%$ confidence region.

Thm 18. If θ_0 lies in a $100(1-\alpha)\%$ confidence region for θ, then the hypothesis $H_0 : \theta = \theta_0$ is accepted at level α.

Generalized Likelihood Ratio Tests

We generalize the likelihood ratio test for situations in which the hypotheses are not simple. Although such tests are not optimal for every situation, they are typically nonoptimal in situations for which no optimal test exists, and they usually perform reasonably well.

We can use generalized likelihood ratio tests to choose the most powerful among different tests.

(Likelihood Ratio Test). Suppose the observations $\mathbf{X} = (X_1, \dots, X_n)$ have a joint density or frequency function $f(\mathbf{X} \mid \theta)$. Let ω_0 and ω_1 be partition of the set Ω of all possible values of θ. For testing $H_0 : \theta \in \omega_0$ vs $H_1 : \theta \in \omega_1$, we use the test statistic

$$\Lambda = \frac{\max\limits_{\theta \in \omega_0} \operatorname{lik}(\theta)}{\max\limits_{\theta \in \Omega} \operatorname{lik}(\theta)}$$

and reject for small values of Λ, i.e. $\Lambda \leq \lambda_0$, where λ_0 is chosen s.t. $P(\Lambda \leq \lambda_0 | H_0) = \alpha$.

Thm 19. (Large Sample Theory for Likelihood Ratio Tests) Under smoothness conditions on the p.d.f functions involved, the null distribution (under H_0) of $-2\log\Lambda$ tends to a χ^2 distribution with $\dim\Omega - \dim\omega_0$ d.f. as the sample size tends to infinity.

Note: $\dim\Omega$ and $\dim\omega_0$ refers to the **number of free parameters** (unspecified parameters) under Ω and ω_0 respectively.

Likelihood Ratio Tests for the Multinomial Distribution

We wish to judge the plausibility of a model H_0 with the vector of cell probabilities p relative to another model H_1 in which the cell probabilities are free except for the constraint that they are nonnegative and sum to 1.

Below we show 2 **goodness-of-fit tests**:

G-test

Using the likelihood for multinomial distribution and unrestricted mle $\hat{p}_i = x_i/n$, we obtain the likelihood ratio to be $\prod\limits_{i=1}^{m} [p_i(\hat{\theta})/\hat{p}_i]^{x_i}$.

Observing that $x_i = n\hat{p}_i$ and denoting by $O_i = n\hat{p}_i$ and $E_i = np_i(\hat{\theta})$ the observed and expected counts respectively, we can write the large sample distribution $-2\log\Lambda = 2\sum\limits_{i=1}^{m} O_i \log\left(\frac{O_i}{E_i}\right) \sim \chi^2_{m-1-(\dim\,\omega_0)}$, and we reject for large values of $-2\log\Lambda$.

Note: Under Ω, the cell probabilities are allowed to be free, with the constraint that they sum to 1, so $\dim\Omega = m - 1$.

(Pearson's χ^2 statistic) $X^2 = \sum\limits_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$ is asymptotically equivalent with $-2\log\Lambda$ under H_0. We reject for large values of X^2.

Comparing Two Samples

Methods based on the normal distribution

An i.i.d sample $X_1, \ldots, X_n$ is drawn from a $N(\mu_X, \sigma^2)$ distribution and a i.i.d sample is drawn from a $Y_1, \ldots, Y_n$ is drawn from a $N(\mu_Y, \sigma^2)$. We wish to make inferences about $\mu_X - \mu_Y$.

Some facts:

1. The mle of $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y}$.
2. $\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma^2[1/n + 1/m])$.
3. If σ^2 is known then $Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{1/n+1/m}} \sim N(0, 1)$.
4. If σ^2 is unknown, it can be estimated from the data using **pooled sample variance**.

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

where S_X^2 is sample variance of $X_1, \ldots X_n$ and S_Y^2 is sample variance of $Y_1, \ldots Y_n$. The **estimated std error** of $\bar{X} - \bar{Y}$ is then $s_{\bar{X}-\bar{Y}} = s_p\sqrt{1/n + 1/m}$.

Thm 20. The statistic $t = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s_{X-Y}}$ follows a t dist with df = $m + n - 2$.

Hypothesis Testing for two-sample problem

For the two-sample problem, let the null hypothesis asserts that there is no difference between the distributions of the X's and Y's. Three common alternative hypotheses are: $H_1 : \mu_X \neq \mu_Y$, $H_2 : \mu_X > \mu_Y$, $H_3 : \mu_X < \mu_Y$

Assuming the variance of X and Y are equal, the test statistic that will be used to make a decision is $t = \frac{(\bar{X}-\bar{Y})}{s_{X-Y}}$. The rejection region for the 3 alternative hypotheses are: $H_1 : |t| > t_{m+n-2}(\alpha/2)$, $H_2 : t > t_{m+n-2}(\alpha)$, $H_3 : t < -t_{m+n-2}(\alpha)$

We can prove that the test H_0 vs H_1 is indeed equivalent to a likelihood ratio test.

If the two variances are **not assumed to be equal**, a natural estimate of $Var(\bar{X} - \bar{Y})$ is $S_X^2/n + S_Y^2/m$. Hence, the appropriate test statistic is then $\frac{\bar{X}-\bar{Y}}{\sqrt{S_X^2/n+S_Y^2/m}}$ which can be closely approximated by the t-distribution with df (rounded to nearest integer) calculated from

$$\frac{[(S_X^2/n) + (S_Y^2/m)]^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}}$$

Non-normality

If the underlying distributions are not normal, and the sample sizes are large, the CLT justifies the use of the t distribution or normal distribution. Note that the probability levels of CIs and hypothesis tests are now **approximately valid**. (In such a case, however, there is little difference between the t and normal distributions).

However, if the sample sizes are small and the distributions are not normal, the conclusions based on the assumption of normality may not be valid and we need to use nonparametric methods.

---

1. Support is the set of values that a RV can take with positive probability.↩
2. A pivot is a function of observations and unobservable parameters whose **distribution does not depend on the unknown parameters**↩
3. the **prior probability** of a random event or an uncertain proposition is the unconditional probability that is assigned before any relevant evidence is taken into account. (prior expresses one's beliefs about this quantity)↩