Math 481

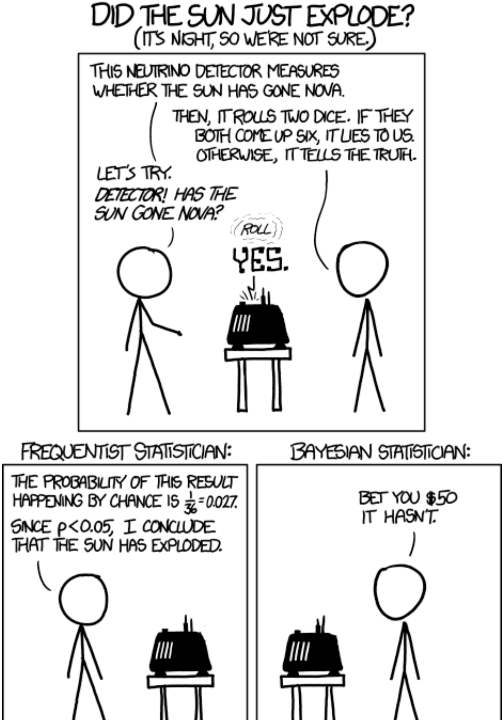
- Instructor: Chenxi Wu wuchenxi2013@gmail.com
- Office: Hill 434, Office hours: 10-11 am Tu, Wed or by appointment, starting from Jan 28.
- Grading policy: 10% weekly homework (lowest dropped), 20% each of the two midterms, 50% final exam.
- Prerequisite: Probability. Will finish review of basic probability on Feb 12.
- Weekly assignments: 2-3 homework problems a week, grade for correctness, similar to exams. There will also be questions from textbook assigned for practice which you don't need to hand in.
- No late homework or make up midterms.

Main topics we will cover:

- Review of probability
- Point estimate
- p-values and hypothesis testing
- Confidence intervals
- Bayesian statistics

Bayesian and non-Bayesian approaches to statistics

- Non-Bayesian approach: Set up a null hypothesis and try to show that observation is highly unlikely if null hypothesis is true.
- Bayesian approach: Assume prior distribution of some parameter, calculate posterior via Bayes formula


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

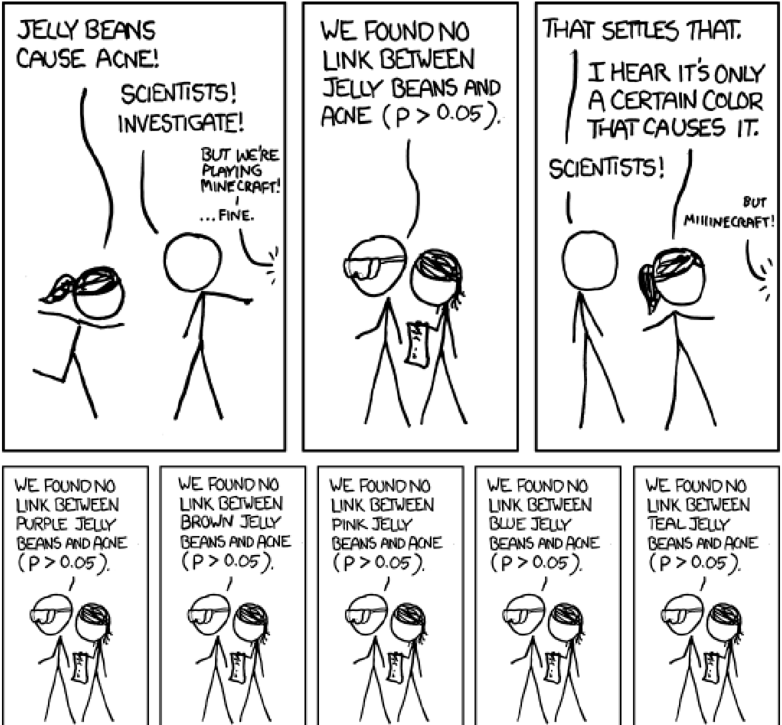
Some review of basic probability

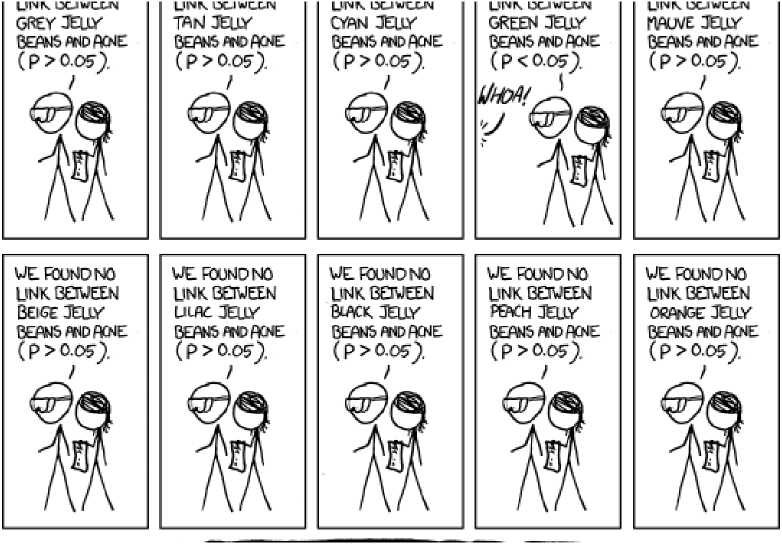
- Two random events A and B are called **independent** if $P(A \cap B) = P(A)P(B)$
- If A and B are two random events, $P(A) > 0$. The conditional probability of B when A is given is $P(B|A) = P(A \cap B)/P(A)$.

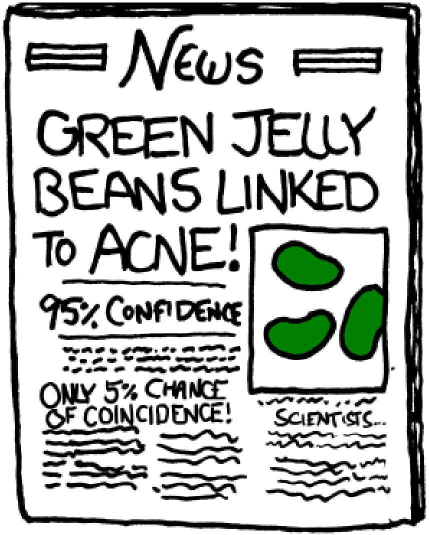
Example

Suppose you are given a coin, you flip it 5 times and get head on all 5 of them.

- Suppose the coin is fair, what is the odds that it gets head for 5 times in 5 flips?
- **Null hypothesis**
- **p-value**


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


GREY JELLY BEANS AND ACNE (P > 0.05).
TAN JELLY BEANS AND ACNE (P > 0.05).
CYAN JELLY BEANS AND ACNE (P > 0.05).
GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

- Suppose the coin is biased and gets head at probability p.
  - What is the probability that it gets head for 5 times in 5 flips?
  - What is the p that maximizes this probability?
  - What is the range of p such that the probability for 5 heads in 5 flips is no less than 0.05?
- **Maximum likelihood estimate (MLE)**
- **Confidence interval**

- Suppose you pick the coin among a pile of 100 coins, 99 of which is fair and 1 has head on both sides. What is the chance of the coin being unfair given the results of the 5 flips?
- **Prior and posterior**

- Suppose the odds for getting a head is uniformly distributed in $[0, 1]$, given the results of the 5 flips, what do you think is the most likely value for p? How about the expectation?
- **Maximum a posteriori (MAP) estimate**

Basic definitions in probability

A **Probability** is a triple (S, F, P) where S is called the **sample space** denoting all possible states of the world, $F \subset \mathcal{P}(S)$ the **event space** and $P : F \to \mathbb{R}$ a real-valued function on F, such that:

1. F is closed under complement and countable union.
2. P is non negative.
3. $P(S) = 1$
4. If $\{E_i\}$ is a countable sequence of disjoint events in F, $P(\bigcup_i E_i) = \sum_i P(E_i)$.

Random variables

- A (real valued) **random variable** X is a function $S \to \mathbb{R}$ such that the preimage of any open interval is in F. Multivariant random variables can be defined similarly.
- The **cumulative distribution function (cdf)** of a random variable X is $F(x) = P(X \leq x)$.
- If $F(x) = \int_{-\infty}^{x} f(t)dt$ we call f the **probability density function (pdf)**
- If there is a countable set C and $g : C \to \mathbb{R}$ such that $F(x) = \sum_{y \in C, y \leq x} g(y)$ we call X **discrete** and g the **probability distribution**
- The **expectation** of a random variable X is defined as $E[X] = \int_S XdP$.

For those who know analysis

- A probability is a measure $P : F \to \mathbb{R}$, where F is a σ-algebra on sample space S and $P(S) = 1$.
- A random variable X is a P-measurable function on S.
- The expectation of a random variable X is the integral $\int_S X dP$.

Some questions

- Must the cdf of a random variable be left or right continuous?
- X is the number of heads in 2 fair coin flips. What is the cdf of X? What is the expectation of X? What is the expectation of $(X - E[X])^2$?
- Can you write down a random variable that is neither discrete nor has a pdf?
- Can you write down a random variable which has no expectation?

Independence and conditional probability

- X and Y are 2 random variables, X and Y are independent iff $F_{X,Y}(s,t) = P(X \leq s \cap Y \leq t) = F_X(s)F_Y(t)$.
- If A is some event with non zero probability, $F_{X|A}(s) = P(X \leq s|A) = P(X \leq s \cap A)/P(A)$.
- If X and Y has joint p.d.f. $f_{X,Y}$ with non zero marginal density f_Y, then $f_{X|Y=a}(s) = f_{X,Y}(s,a)/f_Y(a)$.
- If A_i are disjoint events with non zero probabilities, $B \subset \mathbb{R}$, $P(X \in B| \cup_i A_i) = \sum_i (P(A_i)P(X \in B|A_i))/\sum_i P(A_i)$.
- If Y has p.d.f. f_Y, $A \subset \mathbb{R}$such that $P(Y \in A) > 0$, B is a random event, then $P(B|Y \in A) = \int_A f_Y(s)P(B|Y = s)ds/P(Y \in A)$.

Special random variables

- **Discrete**: Takes on countably values, has p.d.
- **Continuous**: has p.d.f.

2 random variables X and Y has the same distribution iff they have the same c.d.f., or for any $A \subset \mathbb{R}$, $P(X \in A) = P(Y \in A)$. Random variables with the same distribution are NOT necessarily the same.

Special Probability distributions

- **Bernoulli distribution**: $f(1) = \theta$, $f(0) = 1 - \theta$.
- **Binomial distribution** (sum of iid Bernoulli):
  $f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$, $x = 0, 1, \ldots, n$.
- **Negative Binomial distribution** (waiting time for the k-th success of iid trials): $f(x) = \left(\begin{array}{c} x-1 \\ k-1 \end{array} \right) \theta^k (1-\theta)^{x-k}$, $x = k, k+1, \ldots$. When $k = 1$ it is the **geometric distribution**.
- **Hypergeometric distribution** (randomly pick n elements at random from N elements, the number of elements picked from a fixed subset of M elements)
  $f(x) = \left(\begin{array}{c} M \\ x \end{array} \right) \left(\begin{array}{c} N-M \\ n-x \end{array} \right) \left(\begin{array}{c} N \\ n \end{array} \right)^{-1}$.

- **Poisson distribution** (limit of binomial as $n \to \infty$, $n\theta \to \lambda$) $f(x) = \lambda^x e^{-\lambda}/x!$.
- **Multinomial distribution**
  $f(x_1, \ldots x_k) = \left(\begin{array}{c} n \\ x_1, \ldots, x_k \end{array} \right) \theta_1^{x_1} \ldots \theta_k^{x_k}$, $\sum_i x_i = n$, $\theta_i \theta_i = 1$.
- **Multivariate Hypergeometric distribution**
  $f(x_1, \ldots, x_k) = \prod_i \left(\begin{array}{c} M_i \\ x_i \end{array} \right) \cdot \left(\begin{array}{c} N \\ n \end{array} \right)^{-1}$. $\sum_i x_i = n$, $\sum_i M_i = N$.

Special Probability Density Functions

$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. $\Gamma(k) = (k-1)!$ when $k = 1, 2, \ldots$.

- **Uniform distribution**: $f(x) = \begin{cases} 1/(b-a) & x \in (a,b) \\ 0 & x \notin (a,b) \end{cases}$.
- **Normal distribution**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- **Multivariate Normal distribution**: $x \in \mathbb{R}^d$, Σ positive definite $d \times d$ symmetric matrix, $f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$.
- χ^2 **distribution** d: degrees of freedom. Squared sum of d normal distributions: $f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$.

- **Exponential distribution** $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}.$
- **Gamma-distribution**: $f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \leq 0 \end{cases}$
- **Beta distribution**: (conjugate prior of Bernoulli distribution) $f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & x \in (0,1) \\ 0 & x \notin (0,1) \end{cases}.$

Example: If the bias of a coin p has a uniform **prior** in $[0,1]$, after n flips there are a heads and b tails, the **posterior** will be Beta distribution with $\alpha = a+1$, $\beta = b+1$.

Sample mean and sample variance

X_i i.i.d. (independent with identical distribution)

- **Sample mean**: $\overline{X} = \frac{1}{n}\sum_i X_i$
- **Sample variance**:
  $S^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_i X_i^2 - n\overline{X}^2)$.

Properties:

- $E[\overline{X}] = E[X_1]$
- $Var(\overline{X}) = \frac{1}{n} Var(X_1)$
- $\sqrt{\frac{n}{Var(X_1)}}(\overline{X} - E[X_1]) \to \mathcal{N}(0,1)$ (Central Limit Theorem)
- $E[S^2] = Var(X_1)$

Assuming $X_i \sim \mathcal{N}(\mu, \sigma^2)$:

- $\overline{X}$ and S^2 are independent.
- $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

Proof of $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$$(n-1)S^2 = \sum_i (X_i - \overline{X})^2 = \sum_i ((X_i^2 - E[X_i]) - (\overline{X} - E[\overline{X}]))^2$$

$$= \sum_i (X_i^2 - E[X_i])^2 - n(\overline{X} - E[\overline{X}])^2$$

Now divide by σ^2, the first term is $\chi^2(n)$ and second $\chi^2(1)$.

χ^2 distribution

Definition: X_i independent, $\mathcal{N}(0,1)$, then $\sum_{i=1}^{n} X_i = \chi^2(n)$
PDF:

$$f(x) = \begin{cases} \frac{1}{n/2\Gamma(n/2)} x^{\frac{n-2}{2}} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Calculation of PDF:

$$f_{\chi^2(n)}(r) = \frac{d}{dr} \int_{\sum_i x_i^2 \leq r} (2\pi)^{-n/2} e^{-\sum_i x_i^2/2} dx_1 \ldots dx_n$$

$$= (2\pi)^{-n/2} e^{-r/2} \frac{d}{dr} Vol(B(\sqrt{r}))$$

Where $B(x)$ is the ball of radius x.

t distribution

Definition: X and Y independent, $X \sim \mathcal{N}(0,1)$, $Y \sim \chi^2(n)$, then $\frac{X}{\sqrt{Y/n}} \sim t(n)$.
By LLN, when $n \to \infty$ this converges to $\mathcal{N}(0,1)$.
PDF:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

Calculation of PDF of t

$$f_{t(n)}(s) = \frac{d}{ds} P(X \leq s\sqrt{Y/n}) = \frac{d}{ds}\int_0^\infty dy \int_{-\infty}^{s\sqrt{y/n}}$$

$$dx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n-2}{2}} e^{-y/2}$$

$$= \int_0^\infty dy \sqrt{y/n} \frac{1}{\sqrt{2\pi}} e^{-s^2 y/2n} \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n-2}{2}} e^{-y/2}$$

$$= \frac{1}{\sqrt{2\pi n} 2^{n/2}\Gamma(n/2)} \int_0^\infty dy y^{\frac{n-1}{2}} e^{-y(1+\frac{s^2}{n})/2}$$

Now let $z = y(1 + \frac{s^2}{n})/2$ and it's done.

F-distribution

Definition: U and V independent, $U \sim \chi^2(m)$, $V \sim \chi^2(n)$, then
$\frac{U/m}{V/n} \sim F(m,n)$
CDF:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}(\frac{m}{n})^{m/2} x^{m/2-1}(1+\frac{m}{n}x)^{-\frac{m+n}{2}} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Strategy for calculating the PDF of $Y = g(X_i)$:

1. Find joint pdf of X_i
2. Write down the CDF of Y as a probability, hence, some integral of the pdf of X_i
3. Differentiate the CDF of Y.

Probability Review

- Probability, cdf and pdf for continuous random variables:
  - **Probability to cdf**: $F_X(t) = P(X \leq t)$
  - **cdf to pdf**: $f_X(t) = \frac{d}{dt} F_X(t)$
  - **pdf to probability**: $P(X \in A) = \int_A f_X(s) ds$
- Probability, cdf and pd for discrete random variables:
  - **Probability to cdf**: $F_X(t) = P(X \leq t)$
  - **cdf to pd**: $F_X(t) = \sum_{s \leq t} g_X(s)$
  - **pd to probability**: $P(X \in A) = \sum_{s \in A} g_X(s)$
- Joint cdf/pdf/pd, independence, conditional probability.
- Expectation, variance, covariance
- LLN and CLT
- Special distributions: binomial, uniform, normal, χ^2, etc.

Point estimates

Basic setting:

- $\mathcal{F}$: a family of possible distributions (represented by a family of cdf, pdf, or pd)
- $\theta : \mathcal{F} \to \mathbb{R}$ population parameter
- $X_1, \ldots X_n$ i.i.d. with distribution $F \in \mathcal{F}$
- $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a function of X_i, which is an estimate of $\theta(F)$, is called a point estimate.

Example: $\mathcal{F}$: all distributions with an expectation, then $\overline{X}$ is a point estimate of the expectation.

$\hat{\theta}$ is a point estimate of θ.

- The **bias** is $E[\hat{\theta}] - \theta$. $\hat{\theta}$ is called unbiased if $E[\hat{\theta}] = \theta$.
- The **variance** is $Var(\hat{\theta})$.
- $\hat{\theta}$ is called **minimum variance unbiased estimate** if it has the smallest variance among all unbiased estimates.
- $\hat{\theta}_1$ and $\hat{theta}_2$ are two unbiased estimates, the relative efficiency is the ratio of their variance. When they are biased, one can use the mean squared error $E[(\hat{\theta} - \theta)^2]$ instead.
- $\hat{\beta}$ is called **asymptotically unbiased** if bias converges to 0 as $n \to \infty$.
- $\hat{\beta}$ is called **consistent** if $\hat{\beta}$ converges to β in distribution.

Review of definitions regarding point estimates

$\hat{\theta}$ is a point estimate of θ

- Unbiased
- Minimal Variance Unbiased
- Asymptotically unbiased
- Consistent

Properties:

- Minimal Variance Unbiased can be verified via Cramer-Rao
- Mean squared error
  $E[(\hat{\theta}-\theta)^2] = E[((\hat{\theta}-E[\hat{\theta}])+(E[\hat{\theta}]-\theta))^2] = Var(\hat{\theta})+(E[\hat{\theta}]-\theta)^2$
- Mean squared error $\to 0$ implies consistence:

$$P(|\hat{\theta}-\theta| > \epsilon) < \frac{E[(\hat{\theta}-\theta)^2]}{\epsilon^2}$$

  But consistence does not imply mean squared error $\to 0$.

Maximal Likelihood Estimate (MLE)

Suppose $X_i \sim F(\theta)$, i.i.d., observation is $x_1, \ldots, x_k$, then $\hat{\theta} = \arg\max_\theta L(x_1, \ldots x_k, \theta)$.

- When F is a continuous distribution with p.d.f. $f(x, \theta)$, let $L(x_1, \ldots, x_k, \theta) = \prod_i f(x_i, \theta)$
- When F is a discrete distribution with p.d. $g(x, \theta)$, let $L(x_1, \ldots, x_k, \theta) = \prod_i g(x_i, \theta)$

When there are multiple parameters, we can get their MLE by taking $\arg\max$ to all of them altogether.
Sometimes we maximize $log(L)$ (log likelihood) instead of L, which is equivalent.

The basic idea of Bayesian statistics

- Input:
  - Some (possibly vector valued) random variable Θ with given distribution (**prior**)
  - Some (possibly vector valued) random variable X with known conditional distribution conditioned at a value of Θ, $X \sim F(X|\Theta)$. (**observable**)
- Output: the conditional distribution of Θ conditioned at a value of X (**posterior**) $\Theta \sim F(\Theta|X)$.

Example:

- **Prior** $Y \sim Bernoulli(\frac{1}{100})$
- **Observable** X_1, X_2 conditionally i.i.d. when $Y = y$, and their conditional distribution is Bernoulli with $p = \frac{1+8Y}{10}$.

Calculation of the posterior:

$$P(Y=1|X_1,X_2) = \frac{P(Y=1,X_1,X_2)}{P(X_1,X_2)}$$

$$= \frac{P(X_1,X_2|Y=1)P(Y=1)}{P(X_1,X_2|Y=0)|P(Y=0) + P(X_1,X_2|Y=1)|P(Y=1)}$$

$$= \frac{(9/10)^{X_1+X_2}(1/10)^{2-X_1-X_2} \times \frac{1}{100}}{(9/10)^{X_1+X_2}(1/10)^{2-X_1-X_2} \times \frac{1}{100} + (1/10)^{X_1+X_2}(9/10)^{2-X_1-X_2} \times \frac{99}{100}}$$

$$= \frac{9^{X_1+X_2}}{9^{X_1+X_2} + 99 \times 9^{2-X_1-X_2}}$$

So, for example, if we know both X_i takes a value of 1, then the probability of $Y = 1$ is $9/20$.

We can answer many questions using posterior, for example:

- What is the probability of Θ taking value in A given X?
- What is the "most likely" value of Θ? $\hat{\Theta}_{MAP} = \arg\max_s f_{\Theta|X}(s)$, where f is p.d.f. when $\Theta|X$ is continuous and p.d. when it is discrete. This is called the **maximum a posteriori (MAP)** estimate.
- What is the average value of Θ? $\hat{\Theta} = E[\Theta|X]$. This is called the **Bayesian point estimate with L^2 lost**.
- In general, let $l(\cdot, \cdot)$ be a lost function (a positive function such that $l(a, a) = 0$), then $\hat{\Theta} = \arg\min_\theta E[l(\Theta, \theta)|X]$ is called the **Bayesian point estimate**.

MLE vs. Point estimate using Bayesian statistics

MLE:

- Input: Assumption on the distribution of X: $X \sim F(\alpha)$. A likelihood function $L(X, \alpha)$.
- Output: $\hat{\alpha}_{MLE} = \arg\max_{\alpha} L(X, \alpha)$.

Bayesian statistics:

- Input: Prior: $\alpha \sim F_0$, Conditional distribution: $X|\alpha \sim F(\alpha)$.
- Calculated output: Posterior: $\alpha|X \sim F'(X)$
- MAP Point estimate: $\hat{\alpha} = \arg\max_{\alpha} f_{alpha|X}(\alpha)$
- L^2-Bayesian Point estimate: $\hat{\alpha} = E[\alpha|X]$.

Input:

- $\mu \sim \mathcal{N}(0,1)$
- $X_i|\mu$ cond. i.i.d., $\sim \mathcal{N}(\mu, 1)$

Posterior:

$$f_{\mu|X_i}(s) = \frac{f_{\mu,X_i}(s, X_1, \ldots, X_n)}{f_{X_i}(X_1, \ldots X_n)} = \frac{f_{\mu,X_i}(s, X_1, \ldots, X_n)}{\int_{\mathbb{R}} f_{\mu,X_i}(t, X_1, \ldots X_n) dt}$$

$$= \frac{\prod_i f_{X_i|\mu=s}(X_i) f_\mu(s)}{\int_{\mathbb{R}} \prod_i f_{X_i|\mu=t}(X_i) f_\mu(t) dt} = \frac{(2\pi)^{-\frac{n+1}{2}} e^{-\sum_i (X_i - s)^2/2 - s^2/2}}{\int_{\mathbb{R}} (2\pi)^{-\frac{n+1}{2}} e^{-\sum_i (X_i - t)^2/2 - t^2/2} dt}$$

So

$$\mu|X_i \sim \mathcal{N}\left(\frac{\sum_i X_i}{n+1}, \frac{1}{n+1}\right)$$

The MAP and L^2 Bayesian estimate of μ are both $\hat{\mu} = \frac{\sum_i X_i}{n+1}$.

Formula for Posterior

$$f_{\mu|X}(s) \propto f_{X|\mu=s}(X) f_\mu(s)$$

This works for discrete μ or X as well!

Example: P uniform on $[0,1]$, $X|P \sim \text{Binomial}(5, P)$, then $f_{P|X}(s) \propto s^X(1-s)^{5-X} \cdot 1$, hence $P|X \sim Beta(X+1, 6-X)$.

Often in practice we build "hierarchical models" by stacking multiple layers of Bayesian and non Bayesian models together. For example:

$$\sigma_i^2 \sim \Gamma(\alpha, \beta)$$

$$\sigma^2 \sim \Gamma(\alpha', \beta')$$

$$\mu_i \sim \mathcal{N}(0, \sigma^2)$$

$$X_{ij} \text{ ind. } \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

How would you estimate σ_i and μ_i from the values of X_{ij}?

We will talk about models like this if we have more time at the end of the semester.

More examples

1. t has p.d.f. $f_t(x) = \begin{cases} 0 & x < 0 \\ e^{-x} & x > 0 \end{cases}$.
$P(Y = n|t) = (1 - e^{-t})e^{-nt}$. Knowing Y, find $\hat{t}_{MAP}$ and $E[t|Y]$.
2. a, t indep. $\sim$ Uniform$([0, 1])$. $X_i|a, t$ i.i.d. $\sim$ Uniform$([a, a + t])$, find $\hat{t}_{MAP}$.

Answer: $M = \max(X_i)$, $m = \min(X_i)$, then:

$$f_{a,t|X_i} \propto \begin{cases} t^{-n} & 0 \leq a \leq m \leq M \leq a + t \leq a + 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$f_{t|X_i} \propto \begin{cases} t^{-n} \cdot (\min(1, m) - (M - t)) & M - \min(1, m) \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{t}_{MAP} = \min(1, \frac{n}{n-1}(M - \min(1, m)))$$

Review: Point estimate

- Problem: $X \sim F(\Theta)$, want to know unknown parameter Θ.
- Solution: Build a random variable $\hat{\Theta}$ depending on X via:
  - MOM
  - MLE
  - Bayesian-based methods like MAP or Bayesian point estimate
  - Other methods

Hypothesis testing

- Problem: want to know if the distribution of X satisfy certain propositions (**null hypothesis**), for example:
  - Will anyone be infected by covid-19 2 years from now?
  - Will the expectation of our midterm 2 grade be better than midterm 1?
  - Is the performance of a machine learning algorithm better than random chance?
- Solution: Find a random variable Z (**test statistics**) depending on X and a set A (**critical region**), and reject the hypothesis when $Z \in A$.

- (Z, A) is called a **statistical test** to null hypothesis H_0.
- If $Z \in A \iff Z' \in A'$ we consider (Z, A) and (Z', A') to be the same test.
- If H_0 completely determines $P(Z \in A)$ (**simple hypothesis**), $p = P(Z \in A | H_0)$ is called the **significance level**.

Example 1: Suppose your grade for midterm 1 is X_1, your grade for midterm 2 is X_2, $Y = X_2 - X_1$ satisfies normal distribution with variance 25. How do we test the null hypothesis $E[Y] = 0$?

- Answer 1: $Z = Y$, $A = (-\infty, -M) \cup (M, \infty)$.

$$p = P(Y < -M \cup Y > M | H_0)$$

$$= P(Y < -M | Y \sim \mathcal{N}(0, 25))$$

$$+ P(Y > M | Y \sim \mathcal{N}(0, 25))$$

$$= 2 \int_M^\infty \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$$

- Answer 2: $Z = Y$, $A = (M, \infty)$, $p = \int_M^\infty \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$
- Answer 3: $Z = Y$, $A = (-M, M)$, $p = \int_{-M}^M \frac{1}{\sqrt{50\pi}} e^{-t^2/50} dt$

Which of the three is more reasonable?

Ways to evaluate a test

- **Alternative hypothesis**: an alternative to the null hypothesis H_0, called H_1.
- $P(Z \in A|H_0)$ is called **Significance level** or **type I error**.
- If H_1 is a simple hypothesis, $P(Z \notin A|H_1)$ is called **type II error**.
- If H_1 is a simple hypothesis, $1 - P(Z \notin A|H_1) = P(Z \in A|H_1)$ is called **(statistical) power**
- If $X \sim F(\theta)$, $\pi(\theta) = P(Z \in A|\theta)$ is called the **power function**. If $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, then significance is $\pi(\theta_0)$ and power is $\pi(\theta_1)$.

In Example 1, let $Y = \mathcal{N}(\theta, 25)$, what is the power function of the three tests?

Example 2: Y_i i.i.d. $\sim \mathcal{N}(\theta, 25)$, $H_0 : \theta = 0$.
Example 3: Y_i i.i.d. Bernoulli distribution with parameter θ, $H_0 : \theta = 1/2$.

Review

- $X \sim F(\theta)$. Null hypothesis: $H_0 : \theta = \theta_0$, alternative hypothesis $H_1 : \theta = \theta_1$.
- Statistical test: (Z, A), Z: test statistics, A: critical region
- Type I error: $P(Z \in A | H_0)$
- Type II error: $P(Z \notin A | H_1)$
- Power: $P(Z \in A | H_1)$
- Power function: $\pi(t) = P(Z \in A | \theta = t)$

Intuition behind statistical tests

- If (Z, A) is a test such that the significance level is very small.
- Suppose H_0 is true.
- It must mean that $P(Z \in A)$ is very small.
- However, in an experiment we get $Z \in A$
- Hence the assumption earlier is probably untrue.
- Hence H_0 is probably false.

Example 2

X_i $i = 1, \ldots 6$ $i.i.d.$, Bernoulli with $P(X_i = 1) = p$.
$H_0 : p = 0.5$, $H_1 : p = 0.9$.
Test statistics: $Z = \sum_i X_i$. $A = [M, 6]$, M is an integer.
Then power function is:

$$\pi(p) = P(Z \geq M|p) = \sum_{i=M}^{6} \binom{6}{i} p^i (1-p)^{6-i}$$

Significance is $\pi(0.5) = \frac{1}{64} \sum_{i=M}^{6} \binom{6}{i}$.
Power is $\pi(0.9) = \sum_{i=M}^{6} \binom{6}{i} (0.9)^i (0.1)^{6-i}$.

- $M = 6$: significance=0.0156, power=0.531
- $M = 5$: significance=0.109, power=0.886
- $M = 4$: significance=0.344, power=0.984

There is trade-off between significance and power. Which M to choose depends on the purpose of the test, in particular whether false positive or false negative would be more costly.

Neyman-Pearson test

Recall that the likelihood function is $L(x, \theta) = f_{X|\theta}(x)$, which is the p.d.f. when X is continuous and p.d. when X is discrete.
The Neyman-Pearson test for $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ is:

$$(X, \{x : L(x, \theta_0)/L(x, \theta_1) \leq k\})$$

Example 2, Neyman-Pearson test

$$p_0 = 0.5, p_1 = 0.9$$

$$L(X_1, \ldots, X_6, p_0) = \prod_i p_0^{X_i}(1 - p_0)^{1-X_i} = \frac{1}{4^6}$$

$$L(X_1, \ldots, X_6, p_1) = \prod_i p_1^{X_i}(1 - p_1)^{1-X_i}$$

$$= 0.9^{\sum_i X_i} \cdot 0.1^{6-\sum_i X_i} = 0.1^6 \cdot 9^{\sum_i X_i}$$

Sometimes we need to consider **composite hypothesis**, i.e. cases when H_0 and H_1 does not completely determine the distribution of X. Suppose $H_0 : \theta \in D_0$, $H_1 : \theta \in D_1$, the likelihood ratio test becomes:

$$(X, \{x : \frac{\sup_{\theta \in D_0} L(x, \theta)}{\sup_{\theta \in D_0 \cup D_1} L(x, \theta)} \leq k\})$$

How would you do likelihood ratio test for the following examples:

- X_i i.i.d. Bernoulli(p). $H_0 : p = 0.5$, $H_1 : p \neq 0.5$.
- X_i i.i.d. $\mathcal{N}(\mu, 1)$. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.

Review

- Because (Z, A) and (Z', A') are the same test if $Z \in A \iff Z' \in A'$, we sometimes don't specify test statistics and critical region and just call the proposition $Z \in A$ a statistical test.
- Neyman-Pearson test: $f_{X|H_0}(X)/f_{X|H_1}(X) \leq k$
- Likelihood ratio test: $H_0 : \theta \in D_0$, $H_1 : \theta \in D_1$.

$$\frac{\sup_{\theta \in D_0} f_{X|\theta}(X)}{\sup_{\theta \in D_0 \cup D_1} f_{X|\theta}(X)} \leq k$$

- Correction: type I error should be called the **significance level** of a test.

Neyman-Pearson Lemma

Neyman-Pearson test has the highest power for given significance, and lowest significance level for given power.
Proof in continuous case: Let X taking value in $\mathbb{R}^n$, k be the threshold of the Neyman-Pearson test with significance α. In other words,

$$\int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)}\leq k} f_{X|H_0}(x)dx = \alpha$$

Then its power is $\beta_0 = \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)}\leq k} f_{X|H_1}(x)dx$.
Suppose another test (Z, A) has significance α, then by definition of conditional p.d.f.,

$$\int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_0}(x)dx = \alpha$$

While the power is

$$
\begin{aligned}
&\int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_1}(x)dx \\
&= \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \in A|X) f_{X|H_1}(x)dx + \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_1}(x)dx \\
&= \beta_0 - \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \notin A|X) f_{X|H_1}(x)dx + \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_1}(x)dx \\
&\leq \beta_0 - \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} P(Z \notin A|X) f_{X|H_0}(x)dx + \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} > k} P(Z \in A|X) f_{X|H_0}(x)dx \\
&= \beta_0 - \frac{1}{k} \int_{\frac{f_{X|H_0}(x)}{f_{X|H_1}(x)} \leq k} f_{X|H_0}(x)dx + \frac{1}{k} \int_{\mathbb{R}^n} P(Z \in A|X) f_{X|H_0}(x)dx \\
&= \beta_0
\end{aligned}
$$

Significance and p-value

$X \sim F(\theta)$, $H_0 : \theta \in D_0$.
Suppose a family of statistical tests with parameter k is $X \in A(k)$.
Then:

- The significance level of the test $X \in A(k)$ is $\alpha = \sup_{\theta \in D_0} P(X \in A(k)|\theta)$. $k \leq k' \implies A(k) \leq A(k')$.
- The p-value for x, which is an observed value of X, is

$$p = \inf_{k \in \{k : x \in A(k)\}} \sup_{\theta \in D_0} P(X \in A(k))$$

- Suppose the test $X \in A(k_0)$ has significance level α_0. Then $x \in A(k_0)$ (i.e. $X = x$ results in rejection of H_0 under this test) implies that x has a p-value no larger than α_0, and x has p-value less than α_0 implies that $x \in A(k_0)$.

Relationship between significance and p-value

Proof: Let $\alpha(k) = \sup_{\theta \in D_0} P(X \in A(k)|\theta)$, then because $P(X \in A(k)|\theta)$ is non-increasing, $k \mapsto \alpha(k)$ is non increasing. Furthermore, by assumption, $\alpha(k_0) = \alpha_0$, and $\alpha(k) > \alpha_0 \implies k > k_0$, and the p-value for x is

$$p = \inf_{k \in \{k : x \in A(k)\}} \alpha(k)$$

Suppose $x \in A(k_0)$, then the p-value of x is $p = \inf_{k \in \{k : x \in A(k)\}} \alpha(k) \leq \alpha(k_0) = \alpha_0$.
Now suppose the p-value of x is less than α_0, then there is some k' such that $x \in A(k')$ and $\alpha(k') < \alpha_0$. Hence, $k' \leq k_0$, $x \in A(k') \subset A(k_0)$.

Example 1: Normal approximation for large sample

X_i i.i.d., Bernoulli distribution with parameter p. $H_0 : p = p_0$, $H_1 : p \neq p_0$. Likelihood ratio test:

$$\frac{\prod_i p_0^{X_i}(1-p_0)^{1-X_i}}{\sup_p \prod_i p^{X_i}(1-p)^{1-X_i}} \leq k$$

$$\frac{p_0^{\sum_i X_i}(1-p_0)^{n-\sum_i X_i}}{(\frac{1}{n}\sum_i X_i)^{\sum_i X_i}(1-\frac{1}{n}\sum_i X_i)^{n-\sum_i X_i}} \leq k$$

$$\log(LHS) = n\overline{X}(\log(p_0)-\log(\overline{X}))+n(1-\overline{X})(\log(1-p_0)-\log(1-\overline{X}))$$

Which is non positive and 0 iff $\overline{X} = p_0$. So for k close to 1 the test should be of the form:

$$|\overline{X} - p_0| > \epsilon$$

From CLT, if $n >> 1$, under H_0, $\sqrt{\frac{n}{p_0(1-p_0)}} \cdot (\overline{X} - p_0)$ has distribution close to $\mathcal{N}(0,1)$, so the test with significance level α is roughly $|\overline{X} - p_0| \geq \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}$ where Φ is the cdf of $\mathcal{N}(0,1)$.
And the p-value for given $\overline{X} = \overline{x}$ is

$$p = \inf\{\alpha : |\overline{x} - p_0| \geq \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}\}$$

$$= 2(1 - \Phi(\sqrt{\frac{n}{p_0(1-p_0)}} \cdot |\overline{x} - p_0|))$$

Suppose $n = 100$, $p_0 = 0.5$, 60 of the X_i has a value of 1 and 40 has a value of 0. We want to test if $H_0 : p = p_0$ is true with a significance level 0.05.

- **Method 1**: The test with significance level 0.05 is roughly $|\overline{X} - p_0| \geq \Phi^{-1}(1 - 0.05/2)\sqrt{\frac{p_0(1-p_0)}{n}} = 0.0980$. $\overline{X} - p_0 = 0.1$ which is larger than the threshold, hence we should reject H_0.
- **Method 2**: Calculate the p-value, we get $p = 2(1 - \Phi(\sqrt{\frac{n}{p_0(1-p_0)}} \cdot |\overline{X} - p_0|)) = 0.0455 \leq 0.05$, so we should reject H_0.

Review

- Neyman-Pearson test: $f_{X|H_0}(X)/f_{X|H_1}(X) \leq k$
- Likelihood ratio test: $H_0 : \theta \in D_0$, $H_1 : \theta \in D_1$.

$$\frac{\sup_{\theta \in D_0} f_{X|\theta}(X)}{\sup_{\theta \in D_0 \cup D_1} f_{X|\theta}(X)} \leq k$$

- Significance level of a test: highest possible probability of false positive under H_0. It is a increasing function of the threshold k.
- p-value of a possible value of X: the significance level of the test with the lowest threshold that rejects H_0.
- How to test H_0 with given significance level α:
  - Method I: Find the threshold k corresponding to α, test the observed value of X using threshold k.
  - Method II: Find the p-value corresponding to the observed value of X, compare it with α.

Example 2: single sample t-test

X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, here μ and σ^2 are both unknown. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.
Likelihood ratio test:

$$\frac{\sup_{\sigma^2} (2\pi\sigma^2)^{-n/2} \prod_i e^{-X_i^2/2\sigma^2}}{\sup_{\mu,\sigma^2} (2\pi\sigma^2)^{-n/2} \prod_i e^{-(X_i-\mu)^2/2\sigma^2}} \leq k$$

Do the optimization we get the optimal μ is $\overline{X}$, the optimal σ^2 in denominator is $\frac{1}{n}\sum_i X_i^2$, and the optimal σ^2 in the numerator is $\frac{1}{n}\sum_i (X_i - \overline{X})^2 = \frac{1}{n}\sum_i X_i^2 - \overline{X}^2$. (Recall examples we did in MLE).

Hence

$$\log(LHS) = -\frac{n}{2}(\log(\frac{1}{n}\sum_i X_i^2) - \log(\frac{1}{n}\sum_i X_i^2 - \overline{X}^2)) + \frac{n}{2} - \frac{n}{2}$$

$$= \frac{n}{2}\log(1 - \frac{\overline{X}^2}{\frac{1}{n}\sum_i X_i^2}) = h(|\frac{\overline{X}}{\sqrt{S^2/n}}|)$$

Where $h(t) = \frac{n}{2}\log(1 - \frac{1}{1+\frac{1}{(n-1)t^2}})$ is a decreasing function of t^2.

So the LRT must be of the form $\left|\frac{\overline{X}}{\sqrt{S^2/n}}\right| \geq M$. From the definition of t-distribution, we know that if

$$X_i \sim \mathcal{N}(0, \sigma^2)$$

Then

$$(n-1)S^2/\sigma^2 \sim \chi(n-1)$$
$$\overline{X}/\sqrt{\sigma^2/n} \sim \mathcal{N}(0,1)$$

So

$$\frac{\overline{X}}{\sqrt{S^2/n}} = \frac{\overline{X}/\sqrt{\sigma^2/n}}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} \sim t(n-1)$$

For any observed value x_i, let $\overline{x}$ and s^2 be the sample mean and sample variance, then the largest threshold M which yield positive result (which corresponds to the smallest k) is:

$$M_0 = \left| \frac{\overline{x}}{\sqrt{s^2/n}} \right|$$

The p-value, which is the significance level of the test with threshold M_0, is:

$$p = P(\left| \frac{\overline{X}}{\sqrt{S^2/n}} \right| \geq M_0 | \frac{\overline{X}}{\sqrt{S^2/n}} \sim t(n-1))$$

$$= 2(1 - T(\left| \frac{\overline{x}}{\sqrt{s^2/n}} \right|)$$

Where T is the cdf of $t(n-1)$.

Example 3: one sided single sample t-test

X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, here μ and σ^2 are both unknown. $H_0 : \mu \leq 0$, $H_1 : \mu > 0$.
Likelihood ratio test:

$$\frac{\sup_{\mu \leq 0, \sigma^2} (2\pi\sigma^2)^{-n/2} \prod_i e^{-(X_i-\mu)^2/2\sigma^2}}{\sup_{\mu, \sigma^2} (2\pi\sigma^2)^{-n/2} \prod_i e^{-(X_i-\mu)^2/2\sigma^2}} \leq k$$

The likelihood ratio is 1 if $\sum_i X_i \leq 0$, and the same as Example 2 if $\sum_i X_i > 0$. Hence, the LRT is of the form:

$$\left| \frac{\overline{X}}{\sqrt{S^2/n}} \right| \geq M \text{ and } \overline{X} > 0$$

Hence

$$\frac{\overline{X}}{\sqrt{S^2/n}} \geq M$$

Hence, for given significant level α we let

$$M = T^{-1}(1-\alpha)$$

For given value x_i we can calculate the p-value as

$$p = 1 - T(\frac{\overline{x}}{\sqrt{s^2/n}})$$

Where $\overline{x}$ and s^2 are the calculated sample mean and sample variance.

Some conceptual questions

- Suppose a statistical test with significance level 0.05 is used to test covid-19, null hypothesis being not having covid-19. If your test come out positive, what do you know about your probability of getting covid-19?
- Let p be a function that sends observed value X to a p-value. What can you say about the c.d.f. of random variable $p(X)$ when H_0 is true?

Midterm 2 Review

- Regular OHs: 10-11 am Tu Wed Fr, Extra OH: 5-8 pm April 6.
- Please make sure you understand the examples fully before doing homework.
- If you find a homework problem too challenging, write down your thought process and where you get stuck, and make sure to read the posted solution after it is due!
- All homework grades lower than your final grades will be replaced by your final grades.
- Please tell me to stop if there is anything you do not understand.
- April 10 is the last day to drop the class.

Midterm 2 review

- MOM
- Bayesian-based point estimates: expectation of posterior, MAP, etc.
- Neyman-Pearson test (the proof that it is optimal will not be tested in the exam)
- Likelihood ratio test
- Significance, power, and p-value

How to read examples and do homework problems

When reviewing the examples, please do not focus on the calculation part and focus on the concepts and ideas.
For example, this is part of the HW7 due yesterday:

X_i, $i = 1, 2, 3$ are i.i.d. with p.d.f. $f_{X_i}(x) = \begin{cases} 0 & x < 0 \\ ce^{-cx} & x > 0 \end{cases}$.

- Let $H_0 : c = 1$, $H_1 : 0 < c < 1$ or $c > 1$. Find the likelihood ratio test.
- Find the threshold in the likelihood ratio test above that makes type I error α equals 0.01.

Relevant examples from the lectures

LRT for X_i i.i.d. $\mathcal{N}(\mu, 1)$. $H_0 : \mu = 0$, $H_1 : \mu \neq 0$.
Likelihood under H_0 is

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi}} e^{-X_i^2/2} = (2\pi)^{-n/2} e^{-\frac{\sum_i X_i^2}{2}}$$

maximum likelihood under H_0 or H_1 is

$$L_1 = \sup_\mu \prod_i \frac{1}{\sqrt{2\pi}} e^{-(X_i-\mu)^2/2}$$

$$= \sup_\mu (2\pi)^{-n/2} e^{-\frac{\sum_i (X_i-\mu)^2}{2}}$$

$$= (2\pi)^{-n/2} e^{-\frac{\sum_i X_i^2 - (\sum_i X_i)^2/n}{2}}$$

So

$$L_0/L_1 = e^{-\frac{(\sum_i X_i)^2}{2n}}$$

So the likelihood ratio test must be of the form $|\sum_i X_i| \geq C$.

Strategy for the HW problem

So, to find the LRT, find the maximal likelihood (here we are dealing with continuous random variables, so just the joint p.d.f.) under H_0 and H_0 or H_1 respectively as $L_0(X_1, X_2, X_3)$ and $L_1(X_1, X_2, X_3)$, and the test is $L_0(X_1, X_2, X_3)/L_1(X_1, X_2, X_3) \leq k$. For each k, the type I error is by definition

$$\alpha = P(L_1(X_1, X_2, X_3)/L_2(X_1, X_2, X_3) \leq k | H_0)$$

Recall that to get probability of a continuous random variable on certain range one integrate its pdf. So here integrate the joint pdf of X_1, X_2 and X_3 on the region defined by the LRT.

Solution to this HW problem

LRT:

$$\frac{L_0}{L_1} = \frac{e^{-X_1} \cdot e^{-X_2} \cdot e^{-X_3}}{\sup_c ce^{-cX_1} \cdot ce^{-cX_2} \cdot ce^{-cX_3}} \leq k$$

So

$$3 + 3(\log(\overline{X}) - \overline{X}) \leq \log(k)$$

Let $a < b$ be the two numbers such that $ae^{-a} = be^{-b}$, and $\int_{x_1,x_2,x_3 \geq 0, x_1+x_2+x_3 \leq a} e^{-(x_1+x_2+x_3)} dx_1 dx_2 dx_3 + \int_{x_1,x_2,x_3 \geq 0, x_1+x_2+x_3 \geq b} e^{-(x_1+x_2+x_3)} dx_1 dx_2 dx_3 = 0.01$, then the threshold k is $a^3 e^{3-3a}$. You will get full credit if you write up to this or something equivalent to this.

One can further simplify this statement by doing the integration, for instance, and get something like:

$$\frac{1}{2} e^{-3a}(9a^2 + 6a + 2) - \frac{1}{2} e^{-3b}(9b^2 + 6b + 2) = 0.99$$

$$k = a^3 e^{3-3a} = b^3 e^{3-3b}$$

Practice Midterm 2

1. X is a random variable with uniform distribution on $[0, 1]$, Y_i, $i = 1, 2$ i.i.d. conditioned at any value of X, and are of the distribution $\mathcal{N}(0, 1 + X)$.

- Write down the joint p.d.f. of X, Y_1, Y_2.
- Find the conditional distribution of X conditioned at $Y_1 = 1$, $Y_2 = 2$.
- Find the conditional expectation of X when $Y_1 = 1$, $Y_2 = 2$.

Answer:

- $f_{X,Y_1,Y_2}(x, y_1, y_2) =$
$$\begin{cases} 0 & x \notin [0,1] \\ (2\pi(1+x))^{-1} e^{-(y_1^2+y_2^2)/(2+2x)} & x \in [0,1] \end{cases}.$$
- $$f_{X|Y_1=1,Y_2=2}(x) = \begin{cases} 0 & x \notin [0,1] \\ \frac{(2\pi(1+x))^{-1} e^{-5/(2+2x)}}{\int_0^1 (2\pi(1+s))^{-1} e^{-5/(2+2s)} ds} & x \in [0,1] \end{cases}$$
- $$\frac{\int_0^1 (2\pi(1+s))^{-1} s e^{-5/(2+2s)} ds}{\int_0^1 (2\pi(1+s))^{-1} e^{-5/(2+2s)} ds}.$$

2. X_i, $i = 1, \ldots, n$ i.i.d. with p.d.f. $f(x) = ae^{-2a|x-b|}$. Find the estimate of a and b using method of moments.
Answer:

$$\hat{b} = \frac{1}{n}\sum_i X_i$$

$$\hat{b}^2 + \frac{1}{2\hat{a}^2} = \frac{1}{n}\sum_i X_i^2$$

So

$$\hat{a} = \sqrt{\frac{1}{2(\frac{1}{n}\sum_i X_i^2 - \frac{1}{n^2}(\sum_i X_i)^2)}}$$

3. X_i, $i = 1, 2, 3$ i.i.d., H_0 is that they are standard normal, H_1 is that they are uniform on $[0, 1]$.

- Find the Neyman-Pearson test.
- What is the smallest possible type I error for a Neyman-Pearson test that has non-zero power?

Answer: The Neyman-Pearson test is:

$$(2\pi)^{-3/2} e^{-\frac{1}{2}\sum_i X_i^2} \leq k, X_i \in [0, 1]$$

To make sure that the power is non-zero, we must let

$$k > \min_{X_i \in [0,1]} (2\pi)^{-3/2} e^{-\frac{1}{2}\sum_i X_i^2} = (2\pi)^{-3/2} e^{-3/2}$$

Hence the type I error

$$\alpha = \int_{x_i \in [0,1], \sum_i x_i^2 \geq -2\log((2\pi)^{3/2}k)} (2\pi)^{-3/2} e^{-\frac{1}{2}\sum_i x_i^2} dx_1 dx_2 dx_3$$

decreases as k decreases. The function being integrated is bounded, and the region of integration has area that goes to 0 as k goes to $(2\pi)^{-3/2}e^{-3/2}$, hence the type I error can be as close to 0 as one wants.

4. X_i, $i = 1, 2, \ldots n$ i.i.d. and are discrete random variables taking value on $\{-2, -1, 1, 2\}$. H_0: $P(X_i = n) = P(X_i = -n)$ for all n, H_1: $P(X_i = n) \neq P(X_i = -n)$ for some n.

- Find the likelihood ratio test.
- Find the p-value for the observation: $X_1 = -1$, $X_2 = -1$, $X_3 = -2$, $X_4 = 2$.
- Find a sequence X_i with the smallest possible n and a p-value less than 0.05.

Answer: Let n_{-2}, n_{-1}, n_1 and n_2 be the number of X_i taking value at -2, -1, 1, and 2 respectively. The likelihood ratio test is:

$$\frac{\sup_{p+q=1}(p/2)^{n_{-2}+n_2}(q/2)^{n_{-1}+n_1}}{\sup_{a+b+c+d}\, a^{n_{-2}}b^{n_{-1}}c^{n_1}d^{n_2}} \le k$$

In other words,

$$(n_{-2}+n_2)\log(\frac{n_{-2}+n_2}{2}) + (n_{-1}+n_1)\log(\frac{n_{-1}+n_1}{2})$$

$$-n_{-2}\log(n_{-2}) - n_{-1}\log(n_{-1}) - n_1\log(n_1) - n_2\log(n_2) \le \log k$$

Here $0\log 0 = 0$.

When $n = 4$, $n_{-2} = 1$, $n_2 = 1$, $n_{-1} = 2$, $n_1 = 0$, the left-hand-side of the inequality above becomes $-2\log 2$. So the smallest possible k is $1/4$. Now we find out the possible cases where the likelihood ratio is no larger than $1/4$: Assuming
$p/2 = P(X_i = 2) = P(X_i = -2)$,
$q/2 = P(X_i = 1) = P(X_i = -1)$.

1. If $n_1 + n_{-1} = n_2 + n_{-2} = 2$, the likelihood ratio is $1/4$ if one of the n_i is 2, $1/16$ if two of them are 2. Total probability is $(p/2)^2(q/2)^2 \frac{4!}{1!1!2!} \cdot 2 \cdot 2 + 2^2 \cdot \frac{4!}{2!2!} = 72(p/2)^2(q/2)^2$.
2. If $n_1 + n_{-1} = 1$, $n_2 + n_{-2} = 3$, the likelihood ratio is no larger than $1/4$ iff one of the n_j is 3. Total probability is $(p/2)^3(q/2) \cdot 2 \cdot 2 \cdot 4$.
3. Similarly, if $n_1 + n_{-1} = 3$, $n_2 + n_{-2} = 1$, we get $(p/2)(q/2)^3 \cdot 2 \cdot 2 \cdot 4$.
4. Lastly, if $n_2 + n_{-2} = 4$ or $n_1 + n_{-1} = 4$, the only possibility for getting likelihood ratio less than $1/4$ is if one of the n_j is 4. So, total probability is $((p/2)^4 + (q/2)^4) \cdot 2$

So, total probability is $\frac{9p^2q^2}{2} + (p^3q + pq^3) + \frac{p^4+q^4}{8}$. The minimum is taken at $p = q = 1/2$, so the p-value is $9/32 + 1/8 + 1/64 = 27/64$.

For every n, it is evident that the smallest k is 2^{-n} and it is obtained when either n_1 or n_{-1} is 0, either n_2 or n_{-2} is 0. Hence, the total probability for that is $\sum_{i=1}^{n-1} 2^2\binom{n}{i}(p/2)^i(q/2)^{n-i} + 2(p/2)^n + 2(q/2)^n = 2^2(p/2+q/2)^n - 2(p/2)^n - 2(q/2)^n = 2^{-n+2} - 2^{-n+1}(p^n + q^n)$. So the maximum is obtained when $p = q = \frac{1}{2}$, and is $2^{-n+2} - 2^{-2n+2}$. Hence the smallest n is 6. We can pick this sequence $1, 1, 1, 1, 1, 1$, the p-value is $2^{-4} - 2^{-6} = \frac{3}{64} < 0.05$.

Example: X_i, $i = 1, 2$ independent and normal, with same variance and expectations μ and 2μ respectively.

- If variance is 1 and μ has $\mathcal{N}(0, 1/\lambda)$ prior, what is its posterior?
- $H_0 : \mu = 0$, and $H_1 : \mu \neq 0$. Find the likelihood ratio test and p-value.

Answer:

- $f_{\mu|X_i}(t) \propto e^{-t^2\lambda/2} e^{-\sum_k (X_k - kt)^2/2}$, so $\mu|X_i \sim \mathcal{N}(\frac{X_1+2X_2}{5+\lambda}, \frac{1}{5+\lambda})$. (this prior is call the prior for ridge regression or L^2 regularization)
- LRT:

$$\frac{\sup_\sigma (2\pi\sigma^2)^{-1} e^{-\sum_k X_k^2/2\sigma^2}}{\sup_{\sigma,\mu} (2\pi\sigma^2)^{-1} e^{-\sum_k (X_k - k\mu)^2/2\sigma^2}} \leq k$$

$$\log(LHS) = (-\log(\frac{\sum_k X_k^2}{2}) - 1)$$

$$-(-\log(\frac{\sum_k (X_k - k\left(\frac{\sum_k kX_k}{5}\right))^2}{2}) - 1)$$

So the LRT is of the form:

$$\frac{\sum_k (X_k - k\left(\frac{\sum_k kX_k}{5}\right))^2}{\sum_k X_k^2} = \frac{(2X_1 - X_2)^2}{(X_1 + 2X_2)^2 + (2X_1 - X_2)^2} \leq C$$

Which is equivalent to

$$\frac{(2X_1 - X_2)^2}{(X_1 + 2X_2)^2} \leq M$$

Where $M/(1+M) = C$. It is easy to see that under H_0, the test statistics $\frac{(2X_1 - X_2)^2}{(X_1+2X_2)^2} \sim F(1,1)$. So the p-value when $X_1 = x_1$, $X_2 = x_2$ is $p = F_{F(1,1)}(\frac{(2x_1-x_2)^2}{(x_1+2x_2)^2})$.

Midterm 2

Mean and Median: about 70

1. X_i, $i = 1, 2, \ldots, n$ are i.i.d. random variables that satisfies normal distribution with expectation λ and variance $1 + \lambda$.

- Find the MOM estimate for λ. (10 points)
- Is the MOM estimate for λ biased or unbiased? (10 points)

Answer: $\hat{\lambda}_{MOM} = \frac{1}{n}\sum_i X_i$. Yes.

2. c has uniform distribution on $[1,2]$, X_i, $i = 1, 2$, are conditionally i.i.d. for given value of c, and has conditional p.d.f. of the form $f_{X_i|c}(x) = ce^{-2c|x|}$.

- Find the conditional p.d.f. of c when $X_1 = 1$, $X_2 = -2$. (10 points)
- Find the MAP estimate for c when $X_1 = 1$, $X_2 = -2$. In other words, find the value $\hat{c}_{MAP}$ that maximizes the conditional p.d.f. you calculated above. (10 points)
- Find the conditional expectation of c when $X_1 = 1$, $X_2 = -2$. (10 points)

Answer: $f_{c|X_1=1,X_2=-2}(x) = \begin{cases} 0 & x < 1 \text{ or } x > 2 \\ \frac{108x^2 e^{-6x}}{25e^{-6} - 85e^{-12}} & 1 \leq x \leq 2 \end{cases}$.

$\hat{c}_{MAP} = 1$, and the conditional expectation is $\frac{61e^{-6} - 373e^{-12}}{50e^{-6} - 170e^{-12}}$.

3. Random variable X has p.d.f.
$f(x) = \frac{1}{\sqrt{2\pi}} \left(ce^{-(x+1)^2/2} + (1-c)e^{-(x-1)^2/2} \right)$. Here $0 \le c \le 1$.
Let $H_0 : c = 0$, $H_1 : c > 0$.

- Find the likelihood ratio test. (10 points)
- If the threshold for likelihood ratio in the test above is set to be 0.5, calculate the significance level. (10 points)

Answer: LRT is $\frac{e^{-(x-1)^2/2}}{\sup_c \left(ce^{-(x+1)^2/2} + (1-c)e^{-(x-1)^2/2} \right)} \le k$. In the denominator, the optimal c is 1 if $x > 0$ and 0 if $x < 0$. Hence, if $k = 1$ then x can be anything, if $0 < k < 1$ then $x \le \frac{1}{2}\log(k)$. If $k = 0.5$, the significance level is $F(-1 - \log(2)/2)$ where F is the c.d.f. of standard normal.

4. X_i, $i = 1, 2$ are i.i.d. random variables taking values in $\{1, 2, 3\}$. Null hypothesis H_0 is $P(X_i = 1) = P(X_i = 2) = P(X_i = 3) = \frac{1}{3}$, and alternative hypothesis H_1 is $P(X_i = k) = k/6$ for $k = 1, 2, 3$.

- Write down the Neyman-Pearson test for this problem. (10 points)
- Calculate the p-value for $X_1 = X_2 = 3$. (10 points)
- Find the threshold for the Neyman-Pearson test that minimizes that sum of false positive (probability of rejecting H_0 when H_0 is true) and false negative (probability of not rejecting H_0 when H_1 is true). (10 points)

Answer: Likelihood ratio for choices of possible values of X_i are

$X_2 \backslash X_1$	1	2	3
1	4	2	4/3
2	2	1	2/3
3	4/3	2/3	4/9

So the N-P test for different threshold r, as well as the type I and II errors, are:

threshold	test	type I error	type II error
$r < 4/9$	$\emptyset$	0	1
$4/9 \leq r < 2/3$	$X_1 = X_2 = 3$	1/9	3/4
$2/3 \leq r < 1$	$X_1 + X_2 \geq 5$	1/3	5/12
$1 \leq r < 4/3$	$X_1 > 1, X_2 > 1$	4/9	11/36
$4/3 \leq r < 2$	$X_1 + X_2 \geq 4$	2/3	5/36
$2 \leq r < 4$	$X_1 + X_2 \geq 3$	8/9	1/36
$r \geq 4$	everything	1	0

The p-value for $X_1 = X_2 = 3$ is $1/9$, and the threshold that minimize the sum of two types of errors is in the range $[2/3, 4/3)$.

How to use a given statistical test

Some common hypothesis testing problems have well known tests, which are usually either LRT or approximated LRT. We will illustrate via examples how to use some of the tests in Chapter 13 of the textbook.
Usually a statistical test is stated as follows:
Testing H_0 against H_1, test statistics $z = z(X)$, critical region of size (significance level) α is $z \in D_\alpha$.

For example, for the **One sample, One sided t-test**:
$X_1, \dots X_n$, **i.i.d.** $\sim \mathcal{N}(\mu, \sigma^2)$. **Testing** $\mu \leq 0$ **against** $\mu > 0$.
Test statistics $t = \frac{\overline{X}}{\sqrt{S^2/n}}$ **Critical region** $t \geq T^{-1}(1-\alpha)$,
where T **is the cdf of** $t(n-1)$.
To make use of it, say $n = 5$ and X_i are $-1, 0, 1, 2, 1$. The t statistics can be calculated as 1.1767. $T^{-1}(1-0.05) = 2.1318$, so we can not reject H_0 when significance level is chosen to be 0.05. The minimal α such that 1.1767 is in the critical region is $1 - T(1.1767) = 0.1523$, so the p-value is 0.1523.
If X_i are $0, 1, 2, 3, 4$ however, $t = 2.8284 \geq 2.1318$, so reject H_0 under significance level 0.05. The p-value is 0.0237.

Sometimes we make use of a test indirectly by transforming the observed random variables: from some observed random variables X, we build random variables Y, and use a known test on Y. For example: X_i $i = 1, \dots 10$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, Y_i, $i = 1, \dots 10$, i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, X_i and Y_j are all independent. Want to test if $\mu_1 = \mu_2$. One way to do so would be to consider $Z_i = X_i - Y_i$, which are i.i.d.normal, and test if their expectation is 0.

This approach usually won't give us the most powerful test as we are losing information during the transformation. However in many situations this is good enough.

Some commonly used statistical tests

X_i i.i.d. $i = 1, \ldots, n$, $\sim \mathcal{N}(\mu, \sigma^2)$.

- Test $\mu = \mu_0$ against $\mu \neq \mu_0$. $t = \frac{\overline{X} - \mu_0}{\sqrt{S^2/n}}$, critical region $|t| \geq F^{-1}_{t(n-1)}(1 - \alpha/2)$, where $F_{t(n-1)}$ is the c.d.f. of $t(n-1)$.
- Test $\mu \leq \mu_0$ against $\mu > \mu_0$, same t as above, critical region $t \geq F^{-1}_{t(n-1)}(1 - \alpha, n-1)$.
- Test $\sigma^2 = \sigma_0^2$: $\chi^2 = (n-1)S^2/\sigma_0^2$, critical region $\chi^2 \in (-\infty, F^{-1}_{\chi(n-1)}(\alpha/2)] \cup [F^{-1}_{\chi(n-1)}(1 - \alpha/2), \infty)$
- Test $\sigma^2 \leq \sigma_0^2$ against $\sigma^2 > \sigma_0^2$: χ same as above, critical region $\chi^2 \geq F^{-1}_{\chi(n-1)}(1 - \alpha)$.
- Test $\sigma^2 \geq \sigma_0^2$ against $\sigma^2 < \sigma_0^2$: χ same as above, critical region $\chi^2 \leq F^{-1}_{\chi(n-1)}(\alpha)$.

$X_i,\ i = 1, \dots n_1$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, $Y_i\ i = 1, \dots n_2$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, $X_i,\ Y_j$ indep.

- Test for $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$, knowing σ_1^2 and σ_2^2. $z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_2^2/n_1 + \sigma_2^2/n_2}}$. Critical region $|z| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)$.
- If σ_i^2 unknown but number of samples is large, can approximate them with S^2.
- $\sigma_1^2 = \sigma_2^2$ but unknown, test $\mu_1 = \mu_2$ against $\mu_1 \neq \mu_2$:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(1/n_1 + 1/n_2) \cdot \left(\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2} \right)}}$$

  Critical region $|t| \geq F^{-1}_{t(n_1+n_2-2)}(1 - \alpha/2)$.
- One sided tests are similar.

$X_i,\ i = 1, \ldots n_1$ i.i.d. $\mathcal{N}(\mu_1, \sigma_1^2)$, $Y_i\ i = 1, \ldots n_2$ i.i.d. $\mathcal{N}(\mu_2, \sigma_2^2)$, $X_i,\ Y_j$ indep.

- Testing $\sigma_1^2 = \sigma_2^2$ against $\sigma_1^2 \neq \sigma_2^2$. $f = S_X^2 / S_Y^2$. Critical region $f \in (0, F^{-1}_{F(n_1-1, n_2-1)}(\alpha/2)] \cup [F^{-1}_{F(n_1-1, n_2-1)}(1 - \alpha/2), \infty)$.
- One sided tests are similar.

One can check by calculation that all these tests have the significance level α.

Pearson's χ^2 test

X_i i.i.d. taking values at $\{1, 2, \ldots m\}$. Test for null hypothesis: $P(X = j) = e_j$, where $e_j = f(j, \theta_1, \ldots, \theta_k)$. Let n_j be the number of X_i taking value j. Then likelihood ratio test gives:

$$\frac{\sup_{\theta_1,\ldots,\theta_k} \prod_j e_j^{n_j}}{\sup_{p_j, \sum_j p_j = 1} \prod_j p_j^{n_j}} \leq k$$

The optimal p_j is n_j/n where $n = \sum_j n_j$. So

$$\log(LHS) = \sum_j n_j(-\log(\frac{n_j/n}{\hat{e}_j})) = \sum_j n_j(-\log(1 + \frac{n_j - n\hat{e}_j}{n\hat{e}_j}))$$

Here $\hat{e}_j$ is the MLE of e_j.
Taylor expansion at $n_j = ne_j$, we get approximated LRT:

$$\sum_j \frac{(n_j - n\hat{e}_j)^2}{n\hat{e}_j} \geq m$$

When n is large, and with some additional assumptions, the test statistics $\sim \chi^2(m - k - 1)$.

Examples of Pearson's χ^2 test:

- X_i takes value in $\{1, 2, 3, 4\}$. To test if $P(X = 1) = P(X = 2) = P(X = 3) = P(X = 4) = 1/4$, consider $\chi^2 = \sum_{j=1}^{4} \frac{(n_j - n/4)^2}{n/4}$ satisfies $\chi^2(3)$, so critical region is $\chi^2 \geq F^{-1}_{\chi^2(3)}(1 - \alpha)$.
- X_i, Y_i taking values in $\{0, 1\}$, (X_i, Y_i) i.i.d. Want to test if X_i and Y_i are independent. Consider the random variable $Z_i = 2X_i + Y_i + 1$, then Z_i takes value at $1, 2, 3, 4$, and this is the same as testing if

$$P(Z_i = k) = \begin{cases} ab & k = 1 \\ a(1-b) & k = 2 \\ (1-a)b & k = 3 \\ (1-a)(1-b) & k = 4 \end{cases}$$

MLE: $\hat{a} = \frac{n_1 + n_2}{n}$, $\hat{b} = \frac{n_1 + n_3}{n}$. Use Pearson's χ^2 test, there is 1 degrees of freedom.

The final exam is open book, so there is no need to memorize the tests! You just need to know how to use a given statistical test.

Confidence interval

Setting: X has p.d.f. (or p.d.) $f(x, \theta)$, where θ is unknown.

- **Point estimate**: find a random variable $\hat{\theta}$ based on X, which is close to θ.
- **Hypothesis testing**: given θ_0, we can tell how unlikely it is to get the observed value of X if $\theta = \theta_0$.
- **Confidence interval** is related to both of these concepts:
  - Conceptually, confidence interval is an extension of point estimate: this is a random variable taking value in the set of sets, such that θ is in it with probability $1 - \alpha$.
  - Mathematically, confidence intervals are equivalent to certain types of statistical tests.

Definition of confidence interval

X has p.d.f. (or p.d.) $f(x, \theta)$, where θ is unknown.
The $1-\alpha$-**confidence interval** of θ is a set $I(X)$ depending on X, such that for any possible value of θ, $P(\theta \in I(X)|\theta) = 1-\alpha$.
Here, as in hypothesis testing, $P(\theta \in I(X)|\theta)$ does not necessarily mean conditional probability. It means the probability after we fix the value of θ.
Equivalence between confidence intervals and statistical tests:

- If $X \in D(\theta_0)$ is a statistical test of the null hypothesis $H_0 : \theta = \theta_0$, which has significance level α. Then $I(X) = \{\theta_0 : X \notin D(\theta_0)\}$ is a $1-\alpha$ confidence interval for θ.
- If $I(X)$ is a $1-\alpha$ confidence interval for X, then $\theta_0 \notin I(X)$ is a statistical test of the null hypothesis $H_0 : \theta = \theta_0$.

In some textbooks the CI is defined as $P(\theta \in I(X)|\theta) \geq 1 - \alpha$, then, they should correspond to statistical tests of significance level $\leq \alpha$. They will not be the focus of this course, but in case we need to mention them in examples, let's call them *CI with confidence level at least* $1 - \alpha$.

Proof of equivalence

- Suppose $P(X \in D(\theta)|\theta) = \alpha$. Let

$$I(X) = \{\theta : X \notin D(\theta)\}$$

then

$$P(\theta \in I(X)|\theta) = P(X \notin D(\theta)|\theta)$$

$$= 1 - P(X \in D(\theta)|\theta) = 1 - \alpha$$

- Suppose $P(\theta \in I(X)|\theta) = 1 - \alpha$. Let

$$D(\theta_0) = \{X : \theta_0 \notin I(X)\}$$

Then

$$P(X \in D(\theta)|\theta) = P(\theta \notin I(X)|\theta)$$

$$= 1 - P(\theta \in I(X)|\theta) = \alpha$$

Example 1

X normal distribution with expectation μ and variance 1. Find the 0.95 confidence interval for μ.
Likelihood ratio test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$:

$$\frac{e^{-(X-\mu_0)^2/2}}{\sup_\mu e^{-(X-\mu)^2/2}} \leq k$$

The optimal μ is X, so the LRT is

$$|X - \mu_0| \geq \sqrt{-2\log(k)}$$

Let Φ be the c.d.f. of standard normal distribution. The significance level is the probability of success under null hypothesis, and under null hypothesis, $X - \mu_0$ is standard normal. So,

$$\alpha = 2(1 - \Phi(\sqrt{-2\log(k)}))$$

So the test

$$|X - \mu_0| \geq \Phi^{-1}(0.975)$$

Is a test with significance level α, the confidence interval is

$$I(X) = \{\mu : |X - \mu| \leq \Phi^{-1}(0.975)\} = [X - \Phi^{-1}(0.975), X + \Phi^{-1}(0.975)]$$

One sided confidence interval

Sometimes we want the confidence interval to be one sided, like $I = [a(X), \infty)$. The statistical test associated to it should be $mu < a(X)$, in other words, it should only reject null hypothesis $\mu = \mu_0$ if μ_0 is too small. Hence, let's consider $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$, then the LRT becomes

$$\frac{e^{-(X-\mu_0)^2/2}}{\sup_{\mu \geq \mu_0} e^{-(X-\mu)^2/2}} \leq k$$

So the optimal μ is μ_0 if $X \leq \mu_0$, X if $X > \mu_0$. So the test is

$$X - \mu_0 \geq \sqrt{-2\log(k)}$$

When $k < 1$, and everything when $k = 1$. So

$$\alpha = 0.05 = 1 - \Phi(\sqrt{-2\log(k)})$$

$$X - \mu_0 \geq \Phi^{-1}(0.95)$$

$$I(X) = \{\mu : X - \mu \leq \Phi^{-1}(0.95)\} = [X - \Phi^{-1}(0.95), \infty)$$

- As an exercise, read Chapter 11 and Chapter 13. For every statistical test in 13.2-13.6, find the corresponding confidence interval, if there are any, from 11.2-11.7.
- True or false: suppose based on the statistics up to today, the reproductive number R_0 of covid-19 has a 95% confidence interval $[2.1, 2.5]$. Then the probability of R_0 being between 2.1 and 2.5 is 0.95.
- True or false: suppose after the covid-19 outbreak we found a very good model for estimating the R_0 of an epidemic, and this model gives a 95% confidence interval. Then, the probability of R_0 lying in this confidence interval is 0.95.

Example 2

X_1, X_2 i.i.d. with uniform distribution on $[a-1/2, a+1/2]$. Find d such that $[\overline{X}-d, \overline{X}+d]$ is a 95% confidence interval, and find the corresponding statistical tests.
$\overline{X}$ has p.d.f.

$$f_{\overline{X}}(x) = \begin{cases} 0 & x \notin [a-1/2, a+1/2] \\ 4|x-a| & x \in [a-1/2, a+1/2] \end{cases}$$

Because $a \in [\overline{X}-d, \overline{X}+d]$ iff $\overline{X} \in [a-d, a+d]$,

$$0.95 = P(a \in [\overline{X}-d, \overline{X}+d]|a) = P(\overline{X} \in [a-d, a+d]|a) = \int_{a-d}^{a+d} f_{\overline{X}}(s)ds$$

So $d = 1/2 - \sqrt{1/80}$. The test for $H_0 : a = a_0$ is
$a_0 \leq \overline{X} - 1/2 + \sqrt{1/80}$ or $a_0 \geq \overline{X} + 1/2 - \sqrt{1/80}$.

Remark: Bayesian analogy of hypothesis testing and confidence interval

One can create some analogy of hypothesis testing and confidence interval under Bayesian statistics as well, which is conceptually much simpler but completely different from the ones we learn in the non-Bayesian setting:

- Recall that the output of Bayesian statistics is the posterior, i.e. conditional distribution of θ conditioned at X.
- For a hypothesis $H : \theta \in D$, we can calculate its probability under this posterior $P(\theta \in D|X)$, and reject it when this probability is small.
- The $1-\alpha$-**credible interval** is $J(X)$ such that $P(\theta \in J(X)|X) = 1-\alpha$.

This slide will not be in the exam.

Review of LRT, significance level, p value

Likelihood ratio test:
X has p.d.f. (or p.d.) $f(x, \theta)$, $H_0 : \theta \in D_0$, $H_1 : \theta \in D \backslash D_0$.
The LRT is:

$$L_0(X)/L_1(X) \leq k$$

Where

$$L_0(X) = \sup_{\theta \in D_0} f(X, \theta), L_1(X) = \sup_{\theta \in D} f(X, \theta)$$

And k is an arbitrary threshold parameter. The significance level is

$$\alpha = \sup_{\theta \in D_0} P(L_0/L_1 \leq k|\theta)$$

For $X = x$, we find the smallest k that rejects H_0, which is $k_m = L_0(x)/L_1(x)$. The significant level for the LRT with threshold k_m is called the p-value for x:

$$p = \sup_{\theta \in H_0} P(L_0(X)/L_1(X) \leq k_m|\theta)$$

Example: X has p.d.f. $f(x) = \begin{cases} 0 & x \notin [0,1] \\ 1 + c(x - 1/2) & x \in [0,1] \end{cases}$.
$c \in [-2, 2]$. $H_0 : c = 0$, $H_1 : c \neq 0$.

$$L_0(X) = \begin{cases} 1 & X \in [0,1] \\ 0 & X \notin [0,1] \end{cases}, L_1(X) = \begin{cases} 1 + 2|X - 1/2| & X \in [0,1] \\ 0 & X \notin [0,1] \end{cases}$$

LRT:

$$1/(1 + 2|X - 1/2|) \leq k$$

$$|X - 1/2| \geq 1/(2k) - 1/2$$

If $k = 2/3$, the above becomes $|X - 1/2| \geq 1/4$. Under null hypothesis X is uniform on $[0, 1]$, so the significance level is $\alpha = P(|X - 1/2| \geq 1/4) = P(X \in [0, 1/4] \cup [3/4, 1]) = 1/2$.

If $X = 2/3$, the minimal k, k_m, satisfies

$$1/6 = 1/(2k_m) - 1/2$$

So the LRT with threshold k_m is:

$$|X - 1/2| \geq 1/6$$

The significance level for this test is

$$\alpha_{p_m} = 2/3$$

And the p-value is

$$p = \alpha_{p_m} = 2/3$$

Review for CI

$X \sim F(\theta)$.

- Definition: An CI with $1-\alpha$ confidence level is a random set $I(X)$ depending on X, such that $P(\theta \in I(X)|\theta) = 1-\alpha$.
- If there is a statistical test for $\theta = \theta_0$ with significance level α of the form $X \in D(\theta_0)$, then $I(X) = \{\theta : X \notin D(\theta_0)\}$ is a $1-\alpha$ CI.
- If $I(X)$ is a $1-\alpha$ CI, $\theta_0 \notin I(X)$ is a test for $\theta = \theta_0$ with significance level α.
- Sometimes we want $I(X)$ to be one-sided, e.g. of the form $[A(X), \infty)$. Hence the corresponding statistical test must be of the form $\theta_0 \leq \alpha(X)$. In other words, the null hypothesis can be rejected is only if θ_0 is too small, i.e. when $\theta_0 < \theta$. Hence the power function of the test must be no more than α on $(-\infty, \theta_0]$, and one can pick the alternative hypothesis as $\theta_0 < \theta$.

In practice we often use the following definitions, which will NOT be in the HW or exam:

$X \sim F(\theta)$.

- Definition: An CI with confidence level bounded by $1-\alpha$ is a random set $I(X)$ depending on X, such that $P(\theta \in I(X)|\theta) \geq 1-\alpha$.
- If there is a statistical test for $\theta = \theta_0$ with significance level $\leq \alpha$ of the form $X \in D(\theta_0)$, then $I(X) = \{\theta : X \notin D(\theta_0)\}$ is a CI with CL bounded by $1-\alpha$.
- If $I(X)$ is a CI with CL bounded by $1-\alpha$, $\theta_0 \notin I(X)$ is a test for $\theta = \theta_0$ with significance level $\leq \alpha$.

Example 1: normal approximation of binomial distribution

$X \sim B(n, p)$, $n >> 1$, p not too close to 0 or 1. Want CI of p.
From what we learned some weeks ago, we have an approximated LRT based on CLT which says that the test for $p = p_0$ against $p \neq p_0$ with significance level α is

$$|X/n - p_0| \geq \sqrt{\frac{p_0(1-p_0)}{n}} \cdot \Phi^{-1}(1-\alpha/2)$$

Where Φ is the cdf of standard normal.
So the approximated $1-\alpha$ CI is

$$\{p : |X/n - p| \leq \sqrt{\frac{p(1-p)}{n}} \cdot \Phi^{-1}(1-\alpha/2)\} = [p_1, p_2]$$

Where

$$X/n - p_1 = \sqrt{\frac{p_1(1-p_1)}{n}} \cdot \Phi^{-1}(1-\alpha/2)$$

$$p_2 - X/n = \sqrt{\frac{p_2(1-p_2)}{n}} \cdot \Phi^{-1}(1-\alpha/2)$$

Because $n >> 0$, $p_1, p_2 \approx X/n$, we have

$$[p_1, p_2] = [X/n - \sqrt{\frac{X(n-X)}{n^3}} \cdot \Phi^{-1}(1-\alpha/2),$$

$$X/n + \sqrt{\frac{X(n-X)}{n^3}} \cdot \Phi^{-1}(1-\alpha/2)]$$

Example 2: Exponential distribution

X has p.d.f. $f(x) = \begin{cases} ce^{-cx} & x \geq 0 \\ 0 & x \leq 0 \end{cases}$. Find the one sided CI of the form $(0, A]$.
LRT with $H_0 : c = c_0$ and $H_1 : c < c_0$.

$$\frac{c_0 e^{-c_0 X}}{\sup_{c \leq c_0} ce^{-cX}} \leq k$$

If $X \leq 1/c_0$ the LHS is 1, if $X > 1/c_0$, the optimal c in denominator is $1/X$, and we get

$$\log(c_0) - Xc_0 \leq \log(k) - \log(X) - 1$$

$$c_0 X - \log(X) \geq \log(c_0) + 1 - \log(k)$$

The LHS is an increasing function, so the test must be of the form $X \geq M$. If we want the significance level to be α,

$$\alpha = P(X \geq M | c = c_0) = \int_M^\infty f(s) ds$$

So $M = -\log(\alpha)/c_0$. The one sided CI is now

$$\{c : X \leq -\log(\alpha)/c\} = (0, -\log(\alpha)/X]$$

Example 3: Making use of the t-, χ^2-, F- ... tests

Suppose there are 2 independent i.i.d. normal samples X_i, $i = 1, \dots n_1$, Y_j, $j = 1, \dots n_2$, with variance σ_1^2 and σ_2^2 respectively. Want the one sided CI of σ_1^2/σ_2^2 of the form $(0, A]$.
$H_0 : \sigma_1^2/\sigma_2^2 = r$, $H_1 : \sigma_1^2/\sigma_2^2 < r$. Let $Y_j' = r^{1/2} Y_j$, then the test is for $Var(X_i) = Var(Y_j')$ against $Var(X_i) \le Var(Y_j')$, use one sided F-test with significance level α is:

$$S_X^2/S_{Y'}^2 = S_X^2/(rS_Y^2) \le F^{-1}_{F(n_1-1, n_2-1)}(\alpha)$$

So CI with CL $1-\alpha$ is

$$\{r : S_X^2/(rS_Y^2) \geq F^{-1}_{F(n_1-1,n_2-1)}(\alpha)\}$$

$$= (0, \frac{S_X^2}{S_Y^2 F^{-1}_{F(n_1-1,n_2-1)}(\alpha)}]$$

In the textbook they used the relationship
$F^{-1}_{F(n_1-1,n_2-1)}(\alpha) = (F^{-1}_{F(n_2-1,n_1-1)}(1-\alpha))^{-1}$
The CI for other tests are analogous.

More approximated CI via CLT

When $n_1 >> 1$, $n_2 >> 1$, CLT allow us to do normal approximation for the χ^2 distribution. This can also be used to derive approximated CI for the ratio of variance:
By definition, $\chi^2(k)$ is the squared sum of k standard normal, so CLT tells us, if $X \sim \chi^2(k)$, when $k \to \infty$, $\frac{X-k}{\sqrt{2k}} \to$ standard normal.

$$\frac{(n_1-1)S_X^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

So

$$\frac{(n_1-1)(S_X^2-\sigma_1^2)}{\sigma_1^2\sqrt{2n_1-2}} \to \mathcal{N}(0,1)$$

Similarly

$$\frac{(n_2-1)(S_Y^2-\sigma_2^2)}{\sigma_2^2\sqrt{2n_2-2}} \to \mathcal{N}(0,1)$$

Hence the distribution of $S_X^2 - rS_Y^2$ is approximately

$$\mathcal{N}(\sigma_1^2 - r\sigma_2^2, \frac{2\sigma_1^4}{n_1-1} + \frac{2\sigma_2^4}{n_2-1})$$

$$\approx \mathcal{N}(\sigma_1^2 - r\sigma_2^2, \frac{2S_X^4}{n_1-1} + \frac{2r^2S_Y^4}{n_2-1})$$

So the test is

$$S_X^2 - rS_Y^2 \leq \Phi^{-1}(1-\alpha)\sqrt{\frac{2S_X^4}{n_1-1} + \frac{2r^2S_Y^4}{n_2-1}}$$

You can now use this to get a corresponding approximated CI $(0, r_m]$, where

$$S_X^2 - r_m S_Y^2 = \Phi^{-1}(1-\alpha)\sqrt{\frac{2S_X^4}{n_1 - 1} + \frac{2r_m^2 S_Y^4}{n_2 - 1}}$$

Review for statistical testing and CI:

- Find statistical test and finding CI are equivalent.
- Common ways to find a statistical test:
  - Neyman-Pearson Lemma
  - Likelihood ratio test
  - Use known tests
  - Transform the random variables then use known tests

Resampling techniques

If X_i i.i.d., distribution has some parameter θ, $n >> 1$. Suppose, via CLT or some other means, we can get a point estimate $\hat{\theta}$ such that its distribution converges to some $\mathcal{N}(\theta, \sigma^2)$ as $n \to \infty$. Then, one can get CI for θ by estimating θ using $\{X_1, \dots X_m\}$, $\{X_{m+1}, \dots, X_{2m}\}$, $\dots$ and do t-test for the resulting i.i.d. normal random variables.

Some commonly used resampling techniques:

- Bootstraping, bagging
- Jackknife
- Cross validation
- U-statistics
- ...

Linear Regression

Setting: $x_1, \dots x_n$ real numbers, $Y_1, \dots Y_n$ independent, $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$. How do we estimate c and σ^2?

(1) MLE for c and σ^2

Likelihood function:

$$L = \prod_i f_{Y_i}(Y_i) = (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}$$

$$\log(L) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (Y_i - cx_i)^2$$

$$\frac{\partial}{\partial c}\log(L) = -\frac{1}{2\sigma^2}\sum_i (2x_i Y_i - 2cx_i^2)$$

$$\hat{c}_{MLE} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$$

$$\frac{\partial}{\partial \sigma^2}\log(L) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (Y_i - cx_i)^2$$

$$\hat{\sigma^2}_{MLE} = \frac{1}{n}\sum_i (Y_i - \hat{c}_{MLE} x_i)^2 = \frac{1}{n}\left(\sum_i Y_i^2 - \left(\sum_i x_i Y_i\right)^2 / \sum_i x_i^2\right)$$

(2) Prior on c, knowing $\sigma^2 = 1$

Suppose $\sigma^2 = 1$, c has a prior $\mathcal{N}(0, \lambda)$.
Posterior will be proportional to

$$g(c) = \frac{1}{\sqrt{2\pi\lambda}} e^{-c^2/(2\lambda)} (2\pi)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/2}$$

So

$$c|Y_i \sim \mathcal{N}\left(\frac{\sum_i X_i Y_i}{\sum_i x_i^2 + 1/\lambda}, \left(\sum_i x_i^2 + 1/\lambda\right)^{-1}\right)$$

(3) Prior on c and σ^2

Suppose σ^2 has a prior $f(s) = \begin{cases} \alpha e^{-\alpha s} & s \geq 0 \\ 0 & s < 0 \end{cases}$, c has a prior $\mathcal{N}(0, \lambda\sigma^2)$.

Posterior will be proportional to

$$g(c, \sigma^2) = \frac{\alpha}{\sqrt{2\pi\lambda}} e^{-c^2/(2\lambda\sigma^2)} e^{-\alpha\sigma^2} (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}$$

MAP estimate:

$$\hat{c}_{MAP} = \frac{\sum_i X_i Y_i}{\sum_i x_i^2 + 1/\lambda}$$

$$\hat{\sigma^2}_{MAP} = \frac{2(\sum_i (Y_i - \hat{c}_{MAP} x_i)^2 + \hat{c}_{MAP}^2/\lambda)}{n + \sqrt{n^2 + 8\alpha(\sum_i (Y_i - \hat{c}_{MAP} x_i)^2 + \hat{c}_{MAP}^2/\lambda)}}$$

Similarly we can calculate the expectation of c and σ^2 under posterior distribution. It is evident that $E[c|Y_i] = \hat{c}_{MAP}$.

$$E[\sigma^2|Y_i] = \frac{\int_0^\infty d\sigma^2 \int_{-\infty}^\infty \sigma^2 g(c, \sigma^2)}{\int_0^\infty d\sigma^2 \int_{-\infty}^\infty g(c, \sigma^2)}$$

(4) Test for hypothesis $H_0 : c = 0$ against $H_1 : c \neq 0$, knowing $\sigma^2 = 1$

LRT:

$$\frac{(2\pi)^{-n/2} e^{-\sum_i Y_i^2/2}}{\sup_c (2\pi)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/2}} \leq k$$

The optimal c is $\frac{\sum_i x_i Y_i}{\sum_i x_i^2}$ from (1), so

$$-\sum_i Y_i^2 + \sum_i (Y_i - \frac{\sum_i x_i Y_i}{\sum_i x_i^2} \cdot x_i)^2 \leq 2 \log k$$

$$\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} \geq -2 \log k$$

So the test should be

$$|\sum_i x_i Y_i| \geq M$$

Under null hypothesis $\sum_i x_i Y_i \sim \mathcal{N}(0, \sum_i x_i^2)$, so significance level is

$$\alpha = 2(1 - F_{\mathcal{N}(0,1)}(\frac{M}{\sqrt{\sum_i x_i^2}}))$$

The test with significance level α should be

$$|\sum_i x_i Y_i| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)\sqrt{\sum_i x_i^2}$$

p-value for $Y_i = y_i$ is

$$p = 2(1 - F_{\mathcal{N}(0,1)}(\frac{|\sum_i x_i y_i|}{\sqrt{\sum_i x_i^2}}))$$

(5) CI for c, knowing $\sigma^2 = 1$

We need statistical test for $H_0 : c = c_0$ against $H_0 : c \neq c_0$. Let $Z_i = Y_i - c_0 x_i$, then $Z_i \sim \mathcal{N}(c - c_0)x_i, 1)$. Now make use of the test in (4), we get

$$|\sum_i x_i Z_i| = |\sum_i x_i Y_i - \sum_i c_0 x_i^2| \geq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)\sqrt{\sum_i x_i^2}$$

So the corresponding $1 - \alpha$ CI for c is

$$\{c : |\sum_i x_i Y_i - \sum_i c x_i^2| \leq F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)\sqrt{\sum_i x_i^2}\}$$

$$= [\frac{\sum_i x_i Y_i}{\sum_i x_i^2} - \frac{F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)}{\sqrt{\sum_i x_i^2}}, \frac{\sum_i x_i Y_i}{\sum_i x_i^2} + \frac{F^{-1}_{\mathcal{N}(0,1)}(1 - \alpha/2)}{\sqrt{\sum_i x_i^2}}]$$

(6) Test for hypothesis $H_0 : c = 0$ against $H_1 : c \neq 0$, with unknown σ^2

LRT:

$$\frac{\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2}e^{-\sum_i Y_i^2/(2\sigma^2)}}{\sup_{c,\sigma^2}(2\pi\sigma^2)^{-n/2}e^{-\sum_i(Y_i-cx_i)^2/(2\sigma^2)}} \leq k$$

$$\sup_{\sigma^2}(2\pi\sigma^2)^{-n/2}e^{-\sum_i Y_i^2/(2\sigma^2)} = (2\pi \cdot \frac{\sum_i Y_i^2}{n})^{-n/2}e^{-n/2}$$

$$\sup_{c,\sigma^2}(2\pi\sigma^2)^{-n/2}e^{-\sum_i(Y_i-cx_i)^2/(2\sigma^2)} = (2\pi\cdot\frac{\sum_i(Y_i - \hat{c}x_i)^2}{n})^{-n/2}e^{-n/2}$$

Where $\hat{c} = \frac{\sum_i x_i Y_i}{\sum_i x_i^2}$. So the test becomes

$$\frac{\sum_i(Y_i - \hat{c}x_i)^2}{\sum_i Y_i^2} \leq k^{2/n}$$

$$\sum_i (Y_i - \hat{c}x_i)^2 = \sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$$

So we can rewrite the test as

$$\left|\frac{\sum_i x_i Y_i / \sqrt{\sum_i x_i^2}}{\sqrt{\frac{1}{n-1} \cdot \left(\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}\right)}}\right| \le M$$

By calculation (using multivariable calculus and linear algebra) we can see that, under null hypothesis, $\frac{\sum_i Y_i^2}{\sigma^2} \sim \chi^2(n)$, $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$ is independent from $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$, and $\frac{1}{\sigma^2} \cdot \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} \sim \chi^2(1)$. So,

$$\frac{\sum_i x_i Y_i / \sqrt{\sum_i x_i^2}}{\sqrt{\frac{1}{n-1} \cdot \left(\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}\right)}} \sim t(n-1)$$

The M for significance level α is $F_{t(n-1)}^{-1}(1-\alpha/2)$.

(7) CI for c, unknown σ^2

Use the same technique as in (5), and the test in (6), we get

$$[\frac{\sum_i x_i Y_i}{\sum_i x_i^2} - F^{-1}_{t(n-1)}(1-\alpha/2) \cdot \frac{\sqrt{\frac{1}{n-1} \cdot (\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}}{\sqrt{\sum_i x_i^2}},$$

$$\frac{\sum_i x_i Y_i}{\sum_i x_i^2} + F^{-1}_{t(n-1)}(1-\alpha/2) \cdot \frac{\sqrt{\frac{1}{n-1} \cdot (\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2})}}{\sqrt{\sum_i x_i^2}}]$$

Review:

- Point estimate: MLE, MAP and Bayesian point estimate.
- Hypothesis testing: LRT.
- Confidence interval.

Setting: $x_1, \ldots x_n$ real numbers, $Y_1, \ldots Y_n$ independent, $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$. How do we estimate c and σ^2?

Examples we will do today:

- CI of σ^2.
- Logistic regression.
- Higher dimensional models.

Independence of residue and regression coefficient

This slide is just for those who remember linear algebra and multivariable calculus.
Last week we made the claim:
If Y_i i.i.d. normal, $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$ is independent from $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2}$.
Proof: Let $c_1 = [x_i / \sqrt{\sum_i x_i^2}]^T \in \mathbb{R}^n$. $|c_1| = 1$, so we can find an orthonormal basis $\{c_1, \ldots, c_n\}$ of $\mathbb{R}^n$. Let $C = [c_1 \ldots . c_n]^T$, $Y = [Y_1, \ldots Y_n]^T$, $Z = CY$. Because Y_i are i.i.d. normal, the p.d.f. of Y is $f(y) = 2\pi\sigma^{2^{-n/2}} e^{-\frac{1}{2} y^T y}$, so for any set $A \subset \mathbb{R}^n$,

$$P(Z \in A) = P(CY \in A) = \int_{C^{-1}A} (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} y^T y} dy$$

$$= \int_A (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} z^T z} dz$$

So Z_i i.i.d. $\mathcal{N}(0, \sigma^2)$.

By calculation it is easy to verify that $\sum_i Y_i^2 - \frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} = \sum_{i=2}^n Z_i^2$ and $\frac{(\sum_i x_i Y_i)^2}{\sum_i x_i^2} = Z_1^2$, hence they must be independent. The same calculation works for $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ as well by change of variable $Y_i = cx_i + Y_i'$.

(8) Test for $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$

Likelihood ratio test:

$$\frac{\sup_c (2\pi\sigma_0^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma_0^2)}}{\sup_{c,\sigma^2} (2\pi\sigma^2)^{-n/2} e^{-\sum_i (Y_i - cx_i)^2/(2\sigma^2)}} \leq k$$

The optimal c is $\frac{\sum_i Y_i x_i}{\sum_i x_i^2}$. Let $r^2 = \sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2$, log of LHS is

$$-\frac{n}{2}\log(\sigma_0^2) - \frac{r^2}{2\sigma_0^2} + \frac{n}{2}\log(r^2/n) + \frac{n}{2} \leq \log(k)$$

Hence the critical region should be of the form $r^2/n \geq \sigma_0^2 A$ or $r^2/n \leq \sigma_0^2 B$ for some positive numbers $0 < B < 1 < A$. By similar argument as in the previous slides, under H_0, $\frac{1}{\sigma_0^2} r^2 \sim \chi^2(n-1)$, so significance level

$$\alpha = F_{\chi^2(n-1)}(nB) + 1 - F_{\chi^2(n-1)}(nA)$$

$$\log(A) - A = \log(B) - B$$

In practice, we usually just ignore the second equation and let $F_{\chi^2(n-1)}(nB) = 1 - F_{\chi^2(n-1)}(nA) = \alpha/2$, hence the test is

$$\frac{1}{\sigma_0^2}\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2 \notin [F^{-1}_{\chi^2(n-1)}(\alpha/2), F^{-1}_{\chi^2(n-1)}(1-\alpha/2)]$$

(9) CI for σ^2

Using the test on the previous slide, we have the CI:

$$\left[\frac{\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2}{F^{-1}_{\chi^2(n-1)}(1-\alpha/2)}, \frac{\sum_i (Y_i - \frac{\sum_i Y_i x_i}{\sum_i x_i^2} \cdot x_i)^2}{F^{-1}_{\chi^2(n-1)}(\alpha/2)}\right]$$

Logistic regression

Materials from this slide on will be beyond the scope of final exam.
Setting Y_i independent, $Y_i \sim Bernoulli(\frac{e^{cx_i}}{1+e^{cx_i}})$.
Likelihood function

$$L = \prod_i \frac{y_i e^{cx_i} + (1 - y_i)}{1 + e^{cx_i}}$$

It is easy to see that $\log(L)$ is concave w.r.t. c, hence any local maximum is the MLE, and we can use convex optimization to calculate the optimal c.
This is a first example of Generalized Linear Models (GLM).

Higher dimensional linear regression

Setting: $x_1, \ldots x_n \in \mathbb{R}^d$, $Y_1, \ldots Y_n$ independent, $\beta \in \mathbb{R}^d$, $Y_i \sim \mathcal{N}(\beta^T x_i, \sigma^2)$. How do we estimate β and σ^2?
MLE: Log likelihood is

$$\log(L) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}\sum_i (Y_i - \beta^T x_i)^2$$

So

$$\hat{\beta}_{MLE} = \arg\min_{\beta} \sum_i (Y_i - \beta^T x_i)^2$$

Take derivative, we get:

$$2\sum_i (Y_i - \hat{\beta}^T x_i) x_i = 0$$

$$\hat{\beta} = (\sum_i x_i x_i^T)^{-1} (\sum_i Y_i x_i)$$

The MLE for σ^2 is the same as the univariant case.

Linear regression with constant term

$x_1, \ldots x_n \in \mathbb{R}$, $Y_1, \ldots Y_n$ independent, $\beta \in \mathbb{R}^d$,
$Y_i \sim \mathcal{N}(d + cx_i, \sigma^2)$. Find MLE for c and d.
Let $x_i' = [1, x_i]^T$, $\beta = [d, c]$, then use the formula on the previous slide, we get

$$[\hat{d}, \hat{c}]^T = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

So

$$\hat{d} = \frac{\sum_i x_i^2 \sum_i Y_i - \sum_i x_i \sum_i x_i Y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$\hat{c} = \frac{-\sum_i x_i \sum_i Y_i + n \sum_i x_i Y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

Ridge Regression

Suppose $\sigma = \sigma_0$, and we add a prior to β as $\beta \sim \mathcal{N}(0, \lambda\sigma_0^2 I_d)$, log of posterior will be, up to a constant,

$$-\frac{\beta^T\beta}{2\lambda\sigma^2} - \frac{1}{2\sigma^2}\sum_i (Y_i - \beta^T x_i)^2$$

So the MAP estimate for β is

$$\hat{\beta} = \arg\min \beta \sum_i (Y_i - \beta^T x_i)^2 + \frac{1}{\lambda}\beta^T\beta$$

$$\hat{\beta} = (\sum_i x_i x_i^T + I_d/\lambda)^{-1}(\sum_i Y_i x_i)$$

This works even when $n < d$.

Alternative interpretation of Ridge Regression

The idea from the previous slide has an alternative formulation as follows: $x_i, x \in \mathbb{R}^d$, Y_i, Y satisfies joint distribution $\mathcal{N}(0, \sigma^2(K + \delta I))$, with known σ^2 and δ, and where $K = [x_1, \dots x][x_1, \dots x]^T$. Find the conditional expectation of Y with known $Y_1, \dots Y_n$. The log of joint p.d.f. of $[Y_1, \dots, Y_n, Y]^T$ is, up to a constant, proportional to

$$-\frac{1}{2}[y_1, \dots y_n, y](K + \delta I)^{-1}[y_1, \dots y_n, y]^T$$

Let $K_0 = [x_1, \dots x_n][x_1, \dots x_n]^T$, $b = [x_1^T x, \dots x_n^T x]$, then
$K + \delta I = \begin{bmatrix} K_0 + \lambda I & b^T \\ b & x^T x + \lambda \end{bmatrix}$, hence

$$(K + \delta I)^{-1} = \begin{bmatrix} * & B^T \\ B & C \end{bmatrix}$$

Where

$$C = (x^T x - b(K_0 + \delta I)^{-1} b^T)^{-1}$$

$$B = -(x^T x - b(K_0 + \delta I)^{-1} b^T)^{-1} b(K_0 + \delta I)^{-1}$$

So the conditional distribution for y is normal, and the expectation is

$$\hat{y} = -\frac{1}{C} B \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = b(K_0 + \delta I)^{-1} \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$$

Let $X = [x_1, \dots x_n]$, $Y = [y_1, \dots y_n]^T$, then this equals

$$x^T X (X^T X + \delta I)^{-1} Y = x^T (X X^T + \delta I)^{-1} X Y$$

$$= x^T (\sum_i x_i x_i^T + \delta I_d)^{-1} (\sum_i y_i x_i)$$

So δ takes the role of $\frac{1}{\lambda}$ earlier. This model allows us to get a value for $\frac{1}{\lambda}$, by setting it as $\hat{\delta}_{MLE}$. This is the simplest case of a family of statistical models called mixed models.

Questions to think about

- Suppose $x_i \in \{1, 2, 3\}$, how do you check $H_0 : Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ against $H_1 : Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$ where f is an arbitrary function?
- How do you check that $Y_i \sim \mathcal{N}(cx_i, \sigma^2)$ in general?

Final review: Probability prerequisites

- Random variables
- c.d.f., p.d.f., p.d.
- Conditional p.d.f
- Expectation
- LLN, CLT
- sample mean and sample variance
- order statistics

Example

X_i i.i.d., $i = 1, \ldots n$, with p.d.f.

$$f(x) = \begin{cases} ce^{-cx} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The joint p.d.f. is

$$f_{X_1,\ldots X_n}(x_1, \ldots x_n) = \prod_i f(x_i) = \begin{cases} c^n e^{-c\sum_i x_i} & \min\{x_i\} \geq 0 \\ 0 & \min\{x_i\} < 0 \end{cases}$$

The joint c.d.f. is

$$F_{X_1,\ldots X_n}(x_1, \ldots x_n) = \prod_i F_{X_i}(x_i)$$

$$= \prod_i \int_{-\infty}^{x_i} f(s)ds = \begin{cases} \prod_i (1 - e^{-cx_i}) & \min\{x_i\} \geq 0 \\ 0 & \min\{x_i\} < 0 \end{cases}$$

Example

$$E[X_i] = \int_0^\infty sce^{-cs} ds = \frac{1}{c}$$

$$Var(X_i) = E[X_i^2] - E[X_i]^2 = \int_0^\infty s^2 ce^{-cs} ds - \frac{1}{c^2} = \frac{1}{c^2}$$

$$E[\overline{X}] = E[X_i] = \frac{1}{c}$$

$$Var(\overline{X}) = \frac{1}{n} Var(X_i) = \frac{1}{nc^2}$$

$$E[S_X^2] = E[\frac{1}{n-1}(\sum_i X_i^2 - n\overline{X}^2)] = \frac{1}{n-1}(\frac{2n}{c^2} - n(Var(\overline{X}) + E[\overline{X}]^2))$$

$$= \frac{1}{n-1}(\frac{2n}{c^2} - \frac{1}{c^2} - \frac{n}{c^2}) = \frac{1}{n-1} \cdot \frac{n-1}{c^2} = \frac{1}{c^2} = Var(X_i)$$

The p.d.f. of $\min\{X_i\} = Y_1$ is

$$f_{Y_1}(x) = \frac{n!}{(1-1)!(n-1)!} F(x)^{1-1} f(x)(1-F(x))^{n-1}$$

Where f is the p.d.f. of X_i and F is the c.d.f. By calculation, the answer is

$$f_{Y_1}(x) = \begin{cases} 0 & x < 0 \\ nce^{-ncx} & x \geq 0 \end{cases}$$

LLN for X_i implies that as $n \to \infty$,

$$\overline{X} \to \frac{1}{c}$$

While CLT implies that as $n \to \infty$,

$$\sqrt{nc^2}(\overline{X} - \frac{1}{c}) \to \mathcal{N}(0,1)$$

Point estimate

Basic setting: X has a distribution with parameter Θ.

- Point estimate: a random variable $\hat{\Theta}$ which we use to estimate Θ.
- Bias: $E[\hat{\Theta}] - \Theta$.
- Variance: $Var(\hat{\Theta})$.
- Consistency: $X = [X_1, \ldots, X_n]$, X_i i.i.d., $\hat{\Theta} \to \Theta$ as $n \to \infty$.
- Some ways to show consistency
  - Definition.
  - Variance and bias goes to 0 (due to Chebyshev's theorem)
  - LLN.
- Ways to find point estimate
  - MLE
  - MOM
  - MAP
  - Bayesian point estimate

Example, continued

- MLE for c

$$L(X_i, c) = f_{X_1,\ldots X_n}(X_1, \ldots X_n, c) = c^n e^{-c\sum_i X_i}$$

So

$$\hat{c}_{MLE} = \frac{n}{\sum_i X_i}$$

- MOM for c: First empirical moment is $\overline{X}$, first moment is $\frac{1}{c}$, so

$$\frac{1}{\hat{c}_{MOM}} = \overline{X}$$

$$\hat{c}_{MOM} = \frac{n}{\sum_i X_i} = \hat{c}_{MLE}$$

- Suppose prior of c has p.d.f. $f_c(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$. Then posterior is:

$$f_{c|X_i}(c) = Cf_c(c)L(X_1, \ldots X_n, c) = Cc^n e^{-c(\sum_i X_i + 1)}$$

  So

$$\hat{c}_{MAP} = \frac{n}{\sum_i X_i + 1} = \frac{1}{\overline{X} + 1/n}$$

$$2(\sum_i X_i + 1)c|X_i \sim \chi^2(2n+2)$$

  So the Bayesian point estimate with L^2 lost is

$$\hat{c}_{L^2, Bayesian} = \arg\min_{c'} E[(c - c')^2 | X_i] = E[c|X_i] = \frac{n+1}{\sum_i X_i + 1}$$

- All these point estimates are consistent due to LLN.

Hypothesis testing

Setting: X has a distribution with parameter Θ. $H_0 : \Theta \in D$, $H_1 : \Theta \in D'$, $D \cap D' = \emptyset$.

- Statistical test: a random event $Z \in A$, where Z is a random variable defined using X, A the critical region (usually the "tail" of the distribution of Z).
- Significance level (bound on type I error): $\sup_{\Theta \in D} P(Z \in A|\Theta)$.
- Power (one minus type II error): $P(Z \in A|\Theta)$ for some specific $\Theta \in D'$.
- p-value: the lowest significance level that result in rejection of H_0.
- Intuition of statistical tests: suppose H_0 is true, then a test with small significance level is unlikely to be true. So, if we observed that it is true, probably H_0 isn't.
- Neyman-Pearson test: D and D' both consists of a single point Θ_0 and Θ_1, then test is $L(X, \Theta_0)/L(X, \Theta_1) \leq k$.
- LRT: $\sup_{\Theta_0 \in D} L(X, \Theta_0)/\sup_{\Theta_1 \in D \cup D'} L(X, \Theta_1) \leq k$.
- How to use known statistical tests.

Example, Continued

- It is easy to see that $2cX_i \sim \chi^2(2)$, so $2c\sum_i X_i \sim \chi^2(2n)$.
- $H_0 : c = 1$, $H_1 : c = 2$. Neyman-Pearson test:

$$\frac{e^{-\sum_i X_i}}{2^n e^{-2\sum_i X_i}} \leq k$$

So the test should be of the form $\sum_i X_i \leq M$, significance level is

$$\alpha = P(\sum_i X_i \leq M | c = 1) = F_{\chi^2(2n)}(2M)$$

So the test with significance level α is

$$\sum_i X_i \leq \frac{1}{2} F^{-1}_{\chi^2(2n)}(\alpha)$$

The p-value for $X_i = x_1$ is

$$p = \min\{\alpha : \sum_i x_i \leq \frac{1}{2} F^{-1}_{\chi^2(2n)}(\alpha)\} = F_{\chi^2(2n)}(2\sum_i x_i)$$

- Suppose a test on $H_0 : Z \sim \chi^2(k)$ is $Z \notin (F^{-1}_{\chi^2(k)}(\alpha/2), F^{-1}_{\chi^2(k)}(1-\alpha/2))$, then we can apply it to $2c\sum_i X_i$, and get a test for $H_0 : c = c_0$ as

$$\sum_i X_i \notin \left(\frac{F^{-1}_{\chi^2(2n)}(\alpha/2)}{2c_0}, \frac{F^{-1}_{\chi^2(2n)}(1-\alpha/2)}{2c_0}\right)$$

$$\sum_i X_i \leq \frac{F^{-1}_{\chi^2(2n)}(\alpha/2)}{2c_0} \text{ or } \sum_i X_i \geq \frac{F^{-1}_{\chi^2(2n)}(1-\alpha/2)}{2c_0}$$

Confidence Intervals

Setting: X has a distribution with parameter Θ.

- A $1-\alpha$-CI is a set $I(X)$ depending on X such that $P(\Theta \in I(X)|\Theta) = 1-\alpha$.
- CI with CL $1-\alpha$ are related to statistical tests of significance level α.
- One sided CIs are usually related to tests where the alternative hypothesis is one sided as well.

Two types of statistical inference:

- Bayesian approach: Θ has assumed prior distribution, use successive observation to estimate the posterior, eventually converging to the true value.
- Non-Bayesian, or frequentist approach: Θ is a constant with unknown value. Use observation to rule out more and more unlikely values of Θ, until we have an estimate of its true value.

Example, continued

- The CI from the test on $H_0 : c = c_0$ earlier is

$$\{c_0 : \sum_i X_i \in [\frac{F^{-1}_{\chi^2(2n)}(\alpha/2)}{2c_0}, \frac{F^{-1}_{\chi^2(2n)}(1-\alpha/2)}{2c_0}]\}$$

$$= [\frac{F^{-1}_{\chi^2(2n)}(\alpha/2)}{2\sum_i X_i}, \frac{F^{-1}_{\chi^2(2n)}(1-\alpha/2)}{2\sum_i X_i}]$$

- To get one sided CI, use LRT for $H_0 : c = c_0$ against $H_1 : c > c_0$.

$$\frac{c_0^n e^{-c_0 \sum_i X_i}}{\sup_{c \geq c_0} c^n e^{-c \sum_i X_i}} \leq k$$

So the test is

$$\sum_i X_i \leq M$$

for some $M < n/c_0$.

To make significance level α, we must let $M = \frac{F^{-1}_{\chi^2(2n)}(\alpha)}{2c_0}$, so the test is

$$\sum_i X_i \leq \frac{F^{-1}_{\chi^2(2n)}(\alpha)}{2c_0}$$

The one sided $1-\alpha$ CI for c is

$$[\frac{F^{-1}_{\chi^2(2n)}(\alpha)}{2\sum_i X_i}, \infty)$$

Review Examples

Suppose $X \sim B(n,p)$, $n >> 1$.

1. Find the LRT for $H_0 : 1/3 \leq p \leq 2/3$
2. Use the CLT to calculate its approximated significance level.
3. Find the approximated p-value for $n = 9000$, $X = 2900$.

LRT:

$$\frac{\sup_{p\in[1/3,2/3]} \binom{n}{X} p^X p^{n-X}}{\sup_p \binom{n}{X} p^X p^{n-X}} \leq k$$

So

$$\log(LHS) = \begin{cases} 0 & X \in [n/3, 2n/3] \\ X\log(\frac{n}{3X}) + (n-X)\log(\frac{2n}{3(n-X)}) & X < n/3 \\ X\log(\frac{2n}{3X}) + (n-X)\log(\frac{n}{3(n-X)}) & X > 2n/3 \end{cases}$$

It is easy to see that this function is increasing when $X < n/3$, decreasing when $X > 2n/3$, and takes the same value at $n/3 - a$ and $2n/3 + a$ for any $0 < a < n/3$, so the LRT is of the form $X \notin [n/3 - M, 2n/3 + M]$

CLT says that as $n \to \infty$, $\frac{X - pn}{\sqrt{np(1-p)}} \sim \mathcal{N}(0, 1)$. Hence

$$\alpha = \sup_{p \in [1/3, 2/3]} P(X \leq n/3 - M \text{ or } X \geq 2n/3 + m)$$

$$\approx \sup_{p \in [1/3, 2/3]} P\left(\frac{X - p}{\sqrt{np(1-p)}} \leq \frac{n/3 - M - pn}{\sqrt{np(1-p)}} \text{ or } \frac{X - p}{\sqrt{np(1-p)}} \geq \frac{2n/3 + M - pn}{\sqrt{np(1-p)}}\right)$$

If $p \in (1/3, 2/3)$, as $n \to \infty$ the two bounds go to infinity and probability goes to 0. So we can only use $p = 1/3$ or $p = 2/3$. In both cases, the significance level is

$$\alpha = 1 - \Phi\left(\frac{M}{\sqrt{2n/9}}\right)$$

Where Φ is the c.d.f. of standard normal. The p-value for $n = 9000$, $X = 2900$ is

$$1 - \Phi\left(\frac{100}{\sqrt{2000}}\right) \approx 0.0127$$

Suppose X_i i.i.d. and $\sim \mathcal{N}(0, \sigma^2)$, $X \sim \mathcal{N}(a, \sigma^2)$ is independent from X_i, $H_0 : a = 0$, $H_1 : a \neq 0$.

1. If $Z - b$ satisfies $t(d)$ distribution (which is the distribution of $\frac{X}{\sqrt{Y/d}}$ where X and Y are independent, $X \sim \mathcal{N}(0, 1)$, $Y \sim \chi^2(d)$), $|Z| \geq M$ is a test for $H_0 : b = 0$. Find the significance level of this test.
2. Find C_n such that $\frac{C_n X}{\sqrt{\sum_i X_i^2}} \sim t(n)$.
3. Use the test in 1. to find a test for $H_0 : a = 0$ with significance level α.

Answer:

1. $\alpha = P(|Z| \geq M | Z \sim t(d)) = 2(1 - F_{t(d)}(M))$
2. Use definition we know that $C_n = \sqrt{n}$.
3. $|X \sqrt{\frac{n}{\sum_i X_i^2}}| \geq F_{t(d)}^{-1}(1 - \alpha/2)$.

X_1, X_2 i.i.d. uniform on $[a, a+l]$.

1. Find a CI for l of the form $[C|X_1 - X_2|, \infty)$ with CL $1-\alpha$
2. Use this CI to derive a test for $H_0 : l = 1$ with significance level α.

Answer:

1. $1-\alpha = P(l \geq C|X_1 - X_2|) = P(|X_1 - X_2| \leq \frac{l}{C})$, so $C = 1/(1-\sqrt{\alpha})$.
2. $|X_1 - X_2| \geq 1 - \sqrt{\alpha}$.

True or false:

- Let Z be a test statistics, H_0, H_0' two disjoint null hypothesis, if the significance level of $Z \in C$ as a test for H_0 is 0.05, the significance level of $Z \in C'$ as a test for H_0' is 0.05, then $Z \in C \cap C'$ as a test for $H_0 \cup H_0'$ is no more than 0.05
- If X_i i.i.d., θ is a parameter of the distribution of X_i, I_n a $1-\alpha$-CI for θ such that the maximal length goes to 0 as n goes to infinity, then the midpoint of I_n is a consistent estimator for θ.

HW 10

Suppose $X_1 \sim \mathcal{N}(0, \sigma^2)$, $X_2 \sim \mathcal{N}(0, 2\sigma^2)$, X_1, X_2 independent, $H_0 : \sigma^2 = c$, $H_1 : \sigma^2 > c$, where $c > 0$. Recall that the one sided χ^2 test for the null hypothesis $Z \sim \chi^2(d)$ is $Z \geq F^{-1}_{\chi^2(d)}(1-\alpha)$, here F is the c.d.f. of $\chi^2(d)$, and α is the significance level.

1. Find a numbers a and b such that under H_0, $aX_1^2 + bX_2^2 \sim \chi^2(2)$.
2. Use the one-sided χ^2 test described above to write down a statistical test of H_0 against H_1.
3. Find the one-sided confidence interval using the test you found above.

Answer:

1. $a = \frac{1}{c}$, $b = \frac{1}{2c}$.
2. $\frac{X_1^2}{c} + \frac{X_2^2}{2c} \geq F^{-1}_{\chi^2(2)}(1-\alpha)$.
3. $[\frac{X_1^2 + X_2^2/2}{F^{-1}_{\chi^2(2)}(1-\alpha)}, \infty)$, or $(\frac{X_1^2 + X_2^2/2}{F^{-1}_{\chi^2(2)}(1-\alpha)}, \infty)$.