Math 481

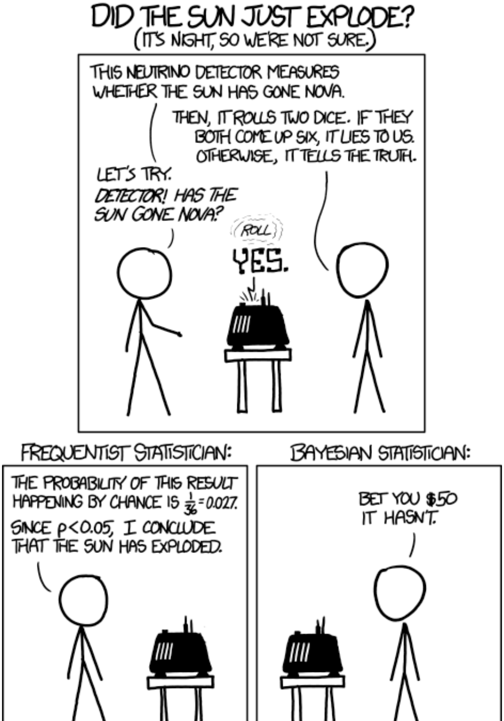
- Instructor: Chenxi Wu wuchenxi2013@gmail.com
- Office: Hill 434, Office hours: 10-11 am Tu, Wed or by appointment, starting from Jan 28.
- Grading policy: 10% weekly homework (lowest dropped), 20% each of the two midterms, 50% final exam.
- Prerequisite: Probability. Will finish review of basic probability 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 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

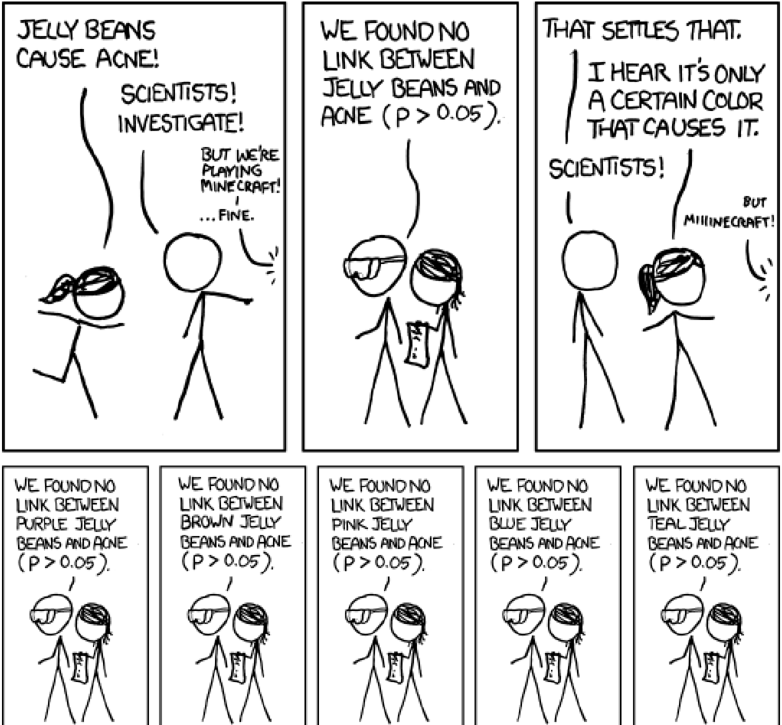
Some review of basic probability

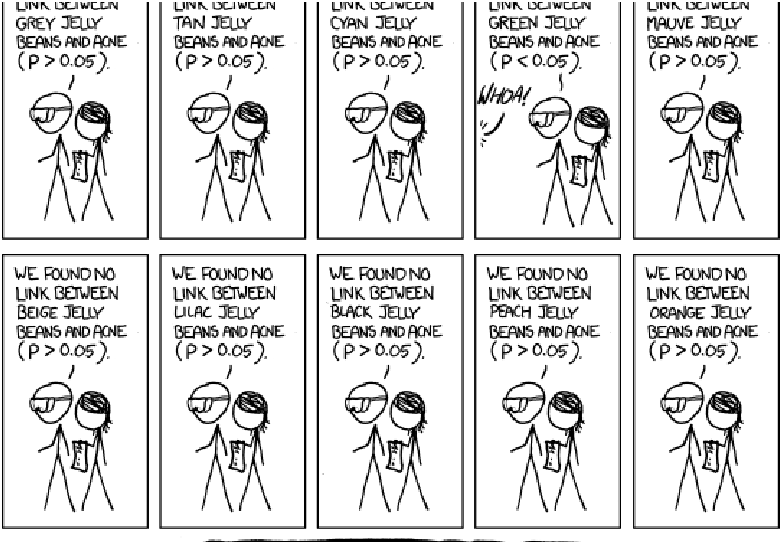
- Two random events A and B are called **independent** if $P(A \cap B) = P(A)P(B)$
- If A and B are two random events, $P(A) > 0$. The conditional probability of B when A is given is $P(B|A) = P(A \cap B)/P(A)$.

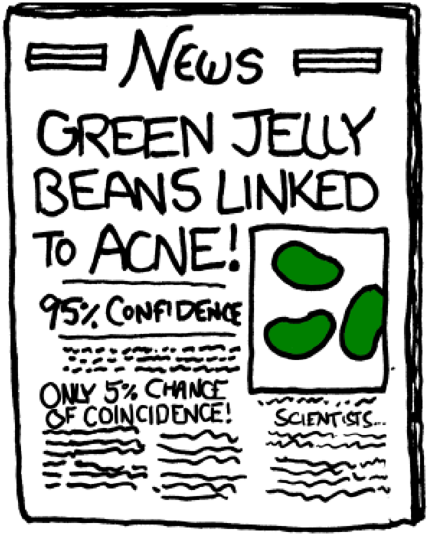
Example

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 $E[X] = \int_S X dP$.

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 $\int_S XdP$.

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): $f(x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}$, $x = 0, 1, \ldots, n$.
- **Negative Binomial distribution** (waiting time for the k-th success of iid trials): $f(x) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} \theta^k (1-\theta)^{x-k}$, $x = k, k+1, \ldots$. 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) $f(x) = \begin{pmatrix} M \\ x \end{pmatrix} \begin{pmatrix} N-M \\ n-x \end{pmatrix} \begin{pmatrix} N \\ n \end{pmatrix}^{-1}$.

- **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** $f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \le 0 \end{cases}$.
- **Gamma-distribution**: $f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & x \le 0 \end{cases}$
- **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**: $F_X(t) = P(X \le t)$
  - **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**: $F_X(t) = P(X \le t)$
  - **cdf to pd**: $F_X(t) = \sum_{s \le t} g_X(s)$
  - **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 $th\hat{e}ta_2$ 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 $\rightarrow 0$ implies consistence:

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

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

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

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

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?

Commonly used test statistics (These and what follows will NOT be in Midterm 2

X_i i.i.d. normal.

- Test for expectation:

$$t = \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}}$$

- Test for variance:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

X_i i.i.d. normal, Y_i i.i.d. normal, independent from X_i

- Test for same expectation, knowing variance:

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}}$$

  This has normal distribution under null hypothesis.

- Test for same expectation, unknown variance: replace σ^2 with s^2. This works if n_i are large.
- Test for same expectation, same but unknown variance:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(1/n_1 + 1/n_2) \cdot \left(\frac{(n_1-1)s_X^2 + (n_2-1)s_Y^2}{n_1+n_2-2}\right)}}$$

  This is $t(n_1 + n_2 - 2)$ under null hypothesis.

- Test for same variance:

$$f = s_X^2/s_Y^2$$

Example 4: goodness of fit

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

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

Taylor expansion at $n_j = ne_j$, we get approximated LRT:

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

When n is large degree of freedom is $m - k - 1$.

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, $i = 1, 2, \dots n$ 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:

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

In other words,

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

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

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 $p = F_{F(1,1)}(\frac{(2x_1 - x_2)^2}{(x_1 + 2x_2)^2})$.