• Course website: http://web.cs.ucla.edu/~sriram/courses/cm226.fall-2016/html/index.html

• Course format changes
  • Homeworks: 50%
  • Exam: 50%. Nov 16 (note date change)
• Previous class: intro to genomics
• This class: intro to statistics
• \((X_1, \ldots, X_n)\) is the data drawn from a distribution \(P\)
• \(\theta\) is an unknown population parameter, \textit{i.e.}, a function of \(P\).
• Inference: Learn \(\theta\) from \((X_1, \ldots, X_n)\).
• \(P\) links \(\theta\) and data.
Three broad and related inference tasks

- Estimation
- Hypothesis testing
- Prediction

For each of these tasks, we have procedures for finding an estimator or test or predictor and procedures for evaluating them. We will focus on the first two tasks.
Outline

Point estimation

Hypothesis testing

Interval estimation
A statistic is a function of the data

Goal of estimation: Find statistic $t$ such that $\hat{\theta}_n = t(X_1, \ldots, X_n)$ is “close” to $\theta$.

$\hat{\theta}_n$ is an estimator of $\theta$.

- Find estimators
- Evaluate an estimator
Point estimations

Procedures to find estimators

- Maximum likelihood
- Method of moments
- Bayes
Maximum likelihood estimation

Given \( X_1, \ldots, X_n \overset{iid}{\sim} P(x|\theta) \), the likelihood function
\[
\mathcal{L}(\theta|x) \equiv P(x_{1:n}|\theta) = \prod_{i=1}^{n} P(x_i|\theta).
\]

A maximum likelihood estimator
\[
\hat{\theta}(x) \triangleq \arg\max_{\theta} \mathcal{L}(\theta|x) = \arg\max_{\theta} \mathcal{L}(x|\theta) = \arg\max_{\theta} \mathcal{L}(x|\theta).
\]

\[
\mathcal{L}(x|\theta) \triangleq \log \mathcal{L}(x|\theta) = \sum_{i=1}^{n} \log P(x_i|\theta)
\]
Maximum likelihood estimation

Given $X_1, \ldots, X_n \overset{iid}{\sim} P(x|\theta)$, the likelihood function
$L(\theta|x) \equiv P(x_{1:n}|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$.

For a differentiable log likelihood function, possible candidates for MLE are
$\theta$ that solve

$$\frac{d}{d\theta} \mathcal{L}(\theta) = 0$$

Why only possible candidates?

- Not sufficient condition for a maximum
- Maximum could also be on the boundary
- Often, no closed-form solution
Example 1
Bernoulli likelihood

Model

\[ X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p), X_i \in \{0, 1\} \]

Likelihood

\[ \mathcal{L}(p) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1-x_i} \]

Log likelihood

\[ \mathcal{LL}(p) = \sum_{i=1}^{n} x_i \log(p) + (1 - x_i) \log(1 - p) \]
Example 1
Bernoulli likelihood

Log likelihood

\[ \mathcal{L}(p) = \sum_{i=1}^{n} x_i \log(p) + (1 - x_i) \log(1 - p) \]

MLE

\[ \hat{p} = \bar{x} \]
Example 1
Bernoulli likelihood

5 coin flips, 2 heads

**Likelihood**

\[ \mathcal{L}(p) = (p)^2(1 - p)^3 \]

\[ \mathcal{LL}(p) = 2 \log(p) + 3 \log(1 - p) \]

\[ \frac{d\mathcal{LL}}{dp}(p) = 2 \frac{1}{p} - 3 \frac{1}{1 - p} \]

**MLE**

\[ \hat{p} = \frac{2}{5} \]
Example 2
Normal likelihood

Model

\[ X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \]

Likelihood

\[ \mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \]

Log likelihood

\[ \mathcal{L}\mathcal{L}(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2) \]
Example 2
Normal likelihood

Log likelihood

$$\mathcal{L}(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)$$

MLE

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}$$
Evaluating estimators

- Finite sample
- Asymptotic
Evaluating estimators

Finite sample

**Bias of an estimator**

\[ bias(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta \]

An estimator is unbiased if \( bias = 0 \).

**Variance of an estimator**

\[ var(\hat{\theta}_n) = \mathbb{E} \left[ \left( \hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] \right)^2 \right] \]

**Mean Square Error of an estimator**

\[ mse(\hat{\theta}_n) = \mathbb{E} \left[ (\hat{\theta}_n - \theta)^2 \right] \]

The \( mse \) decomposes as \( mse(\hat{\theta}_n) = bias(\hat{\theta}_n)^2 + var(\hat{\theta}_n) \)
Comparing estimators not straightforward

- Depends on the evaluation criterion
- Even with a single criterion, performance can depend on the true parameter.
Evaluating estimators

Finite sample

Normal likelihood

\[ X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2) \]

- \( X_1 \) is an unbiased estimator of \( \mu \).
- \( \hat{\mu} = \bar{X} \) is also an unbiased estimator of \( \mu \).
- \( \text{Var} [X_1] = \sigma^2 \) while \( \text{Var} [\hat{\mu}] = \frac{\sigma^2}{n} \).
Normal likelihood

\[ X_1, \ldots, X_n \sim iid \mathcal{N}(\mu, \sigma^2) \]

• \( \hat{\sigma}^2 \) is not an unbiased estimator of \( \sigma^2 \). Its bias is \( \frac{\sigma^2}{n} \).
Evaluating estimators

Asymptotics

- Challenging to compute finite-sample properties.
- What happens when sample size goes to infinity?
Theorem (Weak Law of Large Numbers) Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2 < \infty$. $\overline{X}_n \xrightarrow{p} \mu$.

Theorem (Central Limit Theorem) Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2 < \infty$.

$$\sqrt{n} \frac{(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

. We will informally write $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
**Evaluating estimators**

**Asymptotics**

**Consistency** A sequence of estimators $\hat{\theta}_n$ is a consistent sequence of estimators of a parameter $\theta$ if $\hat{\theta}_n \xrightarrow{p} \theta$.

**Normal likelihood**

$\hat{\mu}_n = \bar{x}_n$ is a consistent estimator of $\mu$.

$\hat{\sigma^2}_n$ is a consistent estimator of $\sigma^2$. 
Theorem (Consistency of MLE) Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ and let $\hat{\theta}_n$ denote the MLE of $\theta$.

\[
\hat{\theta}_n \overset{p}{\to} \theta
\]

If the bias of $\hat{\theta} \to 0$ and its variance $\to 0$, i.e., its MSE goes to 0, then it is consistent.
Theorem (Asymptotic Normality of MLE) Under “regularity conditions”

\[ \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta)) \]

Here

\[ I(\theta) = -\mathbb{E} \left[ \frac{\partial^2 \log P(X|\theta)}{\partial^2 \theta} \right] \]

is the Fisher information associated with the model. What is Fisher information?
Focus on the mean.
For a single observation $X \sim \mathcal{N}(\mu, \sigma^2)$

\[
\frac{\partial \log P(X|\theta)}{\partial \mu} = \frac{x - \mu}{\sigma^2}
\]

\[
\frac{\partial^2 \log P(X|\theta)}{\partial^2 \mu} = -\frac{1}{\sigma^2}
\]

\[
I(\mu) = \frac{1}{\sigma^2}
\]

Makes intuitive sense:

- Inverse Fisher information is the variance
- Smaller the variance, larger the Fisher information, and the estimates of the mean should be more precise.
Fisher information

Normal model
Fisher information
Bernoulli model

\[ \frac{\partial \log P(X|p)}{\partial p} = \frac{X}{p} - \frac{1 - X}{1 - p} \]

\[ \frac{\partial^2 \log P(X|p)}{\partial^2 p} = -\frac{X}{p^2} - \frac{X}{(1 - p)^2} \]

\[ I(p) = \frac{1}{p(1 - p)} \]

- Again inverse Fisher information is the variance.
Outline

Point estimation

Hypothesis testing

Interval estimation
Hypothesis testing
Hypothesis

- Offspring has equal chance of inheriting either of two alleles from a parent (Mendel’s first law holds)
- Two genes segregate independently (Mendel’s second law holds)
- A genetic mutation has no effect on disease
- A drug has no effect on blood pressure

Null hypothesis: \( H_0 : \theta \in \Theta_0 \)
Alternate hypothesis: \( H_1 : \theta \in \Theta_1 \)
Mendel’s first law/Test parameter of Bernoulli

- We observe \( n \) gametes from a diploid individual.
- Probability of transmitting maternal allele is \( p \).
- We can tell for each gamete \( i \) if it is the paternal or the maternal allele. \( X_i \sim \text{Ber}(p) \).
- After observing data from \( n \) gametes, accept or reject \( H_0 \).

\[
H_0 : p = \frac{1}{2} \\
H_1 : p \neq \frac{1}{2}
\]
A rule that specifies values of the sample for which $H_0$ should be accepted or rejected.

Specify a test statistic $t(X_1, \ldots, X_n)$ and a rejection/acceptance region.

**Mendel’s first law**

One test:

Statistic: $t(X_1, \ldots, X_n) = \bar{X}$.

Reject $H_0$ if $\bar{X} > 0.9$ or $\bar{X} < 0.1$.

Rejection region $\{(x_1, \ldots, x_n) : \bar{X} > 0.9 \text{ or } \bar{X} < 0.1\}$. 
Hypothesis testing
Procedures to find tests

- Likelihood Ratio Test (LRT)
- Wald test
- Score test
- Bayesian test
LRT statistic

\[ \lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta | x)}{\sup_{\theta \in \Theta} L(\theta | x)} \]

\[ = \frac{L(\hat{\theta}_0 | x)}{L(\hat{\theta} | x)} \]

\[ \Theta = \Theta_0 \cup \Theta_1 \]

Reject \( H_0 \) is \( \lambda(x) < c. \)
Test mean of normal model

Given \( X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1) \), \( H_0 : \mu = 0 \) versus \( H_1 : \mu \neq 0 \).
Test mean of normal model

\[ \lambda(x) = \frac{\exp\left(-\frac{\sum_{i=1}^{n} x_i^2}{2}\right)}{\exp\left(-\frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{2}\right)} = \exp\left(-\frac{n \bar{x}^2}{2}\right) \]

Reject \( H_0 \) if \( \lambda(x) < c \) equivalent to reject \( H_0 \) if \( |\bar{x}| > \sqrt{\frac{2 \log(\frac{1}{c})}{n}} \).
Mendel’s first law

\[
\lambda(x) = \frac{\frac{1}{2} \sum_i x_i 1^n - \sum_i x_i}{\hat{p} \sum_i x_i (1 - \hat{p})^n - \sum_i x_i} = \frac{\frac{1}{2}}{\overline{x} n \overline{x} (1 - \overline{x}) n (1 - \overline{x})} = 2n H(\overline{x}) - n
\]

Here \( H(x) = -(x \log(x) + (1 - x) \log(1 - x)) \), \( x \in [0, 1] \) is the entropy function.

Reject \( H_0 \) is \( \lambda(x) < c \) equivalent to reject \( H_0 \) if \( H(\overline{x}) < 1 - \frac{\log_2(\frac{1}{c})}{n} \).
Mendel’s first law

How do we choose $c$?
## Hypothesis testing

### Evaluating tests

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>$H_0$</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
</tr>
<tr>
<td>$H_1$</td>
<td>$H_0$</td>
</tr>
<tr>
<td></td>
<td>$H_1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TN</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FP: False positive (type-I error)

FN: False negative (type-II error)

TP: True positive

TN: True negative

For $\theta \in \Theta_0$, probability of false positive for a test $R$: $= P_{\theta}(\text{Test rejects})$.

For $\theta \in \Theta_1$, probability of false negative: $P_{\theta}(\text{Test accepts})$. 
To evaluate (or compare tests), we need to tradeoff false positives and false negatives. For tests with false positive probability below some threshold, what is the false negative threshold?

A test is a level \( \alpha \) test if \( \sup_{\theta \in \Theta_0} P_{\theta}(\text{Test rejects}) \leq \alpha \).

- \( \alpha \) is an upper bound on the false positive probability.
- Often set \( \alpha = 0.05 \)
• Choose $c$ to control $\alpha$: $\sup_{\theta \in \Theta_0} P_{\theta}(\lambda(X) \leq c) = \alpha$.
• Trivial test is never reject.
• We would like to maximize the probability of rejecting the null (i.e., power which we denote $\beta(\theta)$) if it is false while controlling the false positive probability.
For test of the normal mean, $\Theta_0 = \{0\}$.
For $\theta = \theta_0$ (under the null), the test statistic $\overline{X} \sim \mathcal{N}(0, \frac{1}{n})$.

$$P_0(\lambda(X) \leq c) = P_0(|\overline{X}| \geq \sqrt{\frac{2 \log(\frac{1}{c})}{n}})$$

$$= P(|Z| \geq \sqrt{2 \log(\frac{1}{c})})$$

Here $Z \sim \mathcal{N}(0, 1)$. $P(|Z| \geq z_{1-\frac{\alpha}{2}}) = \alpha$ where $z_\alpha$ is the $\alpha$-quantile of the standard normal distribution.
Normal mean

Null

Alternative

$1 - \beta$

$\alpha$

$\mu_0$

$\mu_1$
• Level is a crude measure. If we reject the null hypothesis at level $\alpha = 0.05$, it does not tell us how strong the evidence against $H_0$ is.

• $p$-value of a test is the smallest $\alpha$ at which the test rejects $H_0$.

• Small values of $p$-value implies stronger evidence against $H_0$.

For a statistic $\lambda(X)$ such that small values of $\lambda$ give evidence against $H_0$, the $p$-value $p(x)$:

$$ p(x) = \sup_{\theta \in \Theta_0} P_\theta(\lambda(X) \leq \lambda(x)) $$
• Observe 10 samples and mean is 0.8. Reject $H_0$ at $\alpha = .05$. 
  p-value=$0.01$.

• Observe 10 samples and mean is 2. Reject $H_0$ at $\alpha = .05$. 
  p-value=$2 \times 10^{-10}$.

• Observe 100 samples and mean is 0.8. Reject $H_0$ at $\alpha = .05$. 
  p-value=$1 \times 10^{-15}$.
\( \sup_{\theta \in \Theta_0} P_{\theta}(\lambda(X) \leq c) = \alpha. \)

- Choose \( c \) so that the false positive probability is bounded.

- To compute false positive probability, we need to know how the statistic is distributed under the null (sampling distribution of the statistic).

- Easy for Normal mean test: \( \bar{X} \).

- More complicated statistics?
  - Asymptotics
  - Permutation test
Asymptotics of the LRT

**Theorem** Let \( X_1, \ldots, X_n \sim iid \ p(x|\theta) \). For testing \( \theta = \theta_0 \) versus \( \theta \neq \theta_0 \), under regularity conditions on \( p(x|\theta) \)

\[
-2 \log \lambda(X_n) \xrightarrow{d} \chi^2_1
\]

\( \chi^2_1 \) is the distribution of the square of a normal random variable.

For testing \( \theta \in \Theta_0 \) versus \( \theta \neq \Theta_0 \), under regularity conditions on \( p(x|\theta) \), if \( \nu \) is the difference in the number of free parameters between \( \theta \in \Theta \) and \( \theta \in \Theta_0 \):

\[
-2 \log \lambda(X_n) \xrightarrow{d} \chi^2_\nu
\]

Reject \( H_0 \) iff \( -2 \log \lambda(X) \geq \chi^2_{\nu,1-\alpha} \) gives us an asymptotic (hence approximate) level \( \alpha \) test.
Using the asymptotic distribution of the LRT, the p-value for the LRT statistic:

\[
p(x) = P(\lambda(X) \leq \lambda(x)) \\
= P(-2 \log \lambda(X) \geq -2 \log \lambda(x)) \\
\approx P(\chi^2_1 \geq -2 \log \lambda(x))
\]

Reject \( H_0 \) iff \( p(x) \leq \alpha \) gives us approximate level \( \alpha \) test.
• On many instances, even asymptotic results are difficult. Also not clear, if sample size is big enough.

• **Key idea:** We don’t know the sampling distribution but can sample from this distribution.
Permutation test example
Test if distribution of two alleles is the same in disease vs healthy

\[ X_1, \ldots, X_n \sim F \]
\[ Y_1, \ldots, Y_n \sim G \]
\[ H_0 : F = G = P_0 \]

Statistic: \( t((X_1, \ldots, X_n), (Y_1, \ldots, Y_n)) = \bar{X} - \bar{Y} \).

Key idea:
Under \( H_0 \), all permutations of the data are equally likely.
Use this idea to compute the sampling distribution.
Permutation test example
Test if distribution of two alleles is the same in disease vs healthy

- Randomly choose \( n \) out of \( 2n \) elements to belong to group 1. The remaining belong to group 2.
- Compute the difference in means between the two groups.
- How often does the difference in means on the permuted data exceed the observed difference?

In many cases, the number of permutations is too large to exhaustively enumerate. Then use a random sample of permutations (Monte-Carlo approximation).
Permutation test example

Test if distribution of two alleles is the same in disease vs healthy
Outline

Point estimation

Hypothesis testing

Interval estimation
Point estimates do not give you a measure of confidence

For a scalar parameter $\theta$, given data $X$, an interval estimator is a pair of functions $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, $L \leq U$, such that $[L(X), U(X)]$ covers $\theta$. The interval is random while $\theta$ is fixed.
Evaluating interval estimators

The coverage probability of an interval estimator $[L(X), U(X)]$ is the probability that this interval covers the true parameter $\theta$, denoted $P_\theta(\theta \in [L(X), U(X)])$.

An interval estimator is termed a $1 - \alpha$ confidence interval if $\inf_{\theta} P_\theta(\theta \in [L(X), U(X)]) \geq 1 - \alpha$. 
Finding interval estimators

Inverting a test statistic

Given \( X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1) \), \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu \neq 0 \).

One test is to reject \( H_0 \) if \( \{ x : |\bar{x} - \mu_0| > \frac{z_{1-\alpha}}{\sqrt{n}} \} \) So \( H_0 \) is accepted for

\[ \{ x : |\bar{x} - \mu_0| \leq \frac{z_{1-\alpha}}{\sqrt{n}} \} \]

or equivalently:

\[ \bar{x} - \frac{z_{\alpha/2}}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + \frac{z_{\alpha/2}}{\sqrt{n}} \]

Since the test has size \( \alpha \),

\[ P(H_0 \text{ is accepted}|\mu = \mu_0) = 1 - \alpha, \]

\[ P(\bar{x} - \frac{z_{1-\alpha}}{2 \sqrt{n}} \leq \mu_0 \leq \bar{x} + \frac{z_{1-\alpha}}{2 \sqrt{n}}|\mu = \mu_0) = 1 - \alpha \]

Since this is true for all \( \mu_0 \), we have

\[ P_\mu(\bar{x} - \frac{z_{1-\alpha}}{2 \sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{1-\alpha}}{2 \sqrt{n}}) = 1 - \alpha \]

So \([ \bar{x} - \frac{z_{1-\alpha}}{2 \sqrt{n}}, \bar{x} + \frac{z_{1-\alpha}}{2 \sqrt{n}} ]\) is a \( 1 - \alpha \) confidence interval.
Finding interval estimators
Inverting a test statistic

- A general method
- Methods for constructing tests can be used for getting confidence intervals
Summary

- Two major inferential tasks
- General procedures for each task and to evaluate the quality.
- Can evaluate finite sample and large sample properties.
  - Testing: Likelihood Ratio Test: False positive and false negative probabilities (or power). Permutation is a useful tool in many settings.
  - Interval estimation: Can use tests to get at confidence intervals.