Lecture 5: Bayesian Estimation & Hypothesis Testing

May 15, 2012

GENOME 560, Spring 2012

Su-In Lee, CSE & GS

suinlee@uw.edu

Homework Assignment

- Exercises designed to help you get familiar with statistical concepts and practices
- The more you struggle now, the more you will learn and the better for your research career
    - You can learn statistics only by doing statistics!
- I would encourage to work with other students on the homework problems
    - However, each student has to write his or her own solution

Goals

- Parameter estimation
    - Maximum likelihood estimation
    - Bayesian inference
- Hypothesis testing
    - Overview of key elements of hypothesis testing
    - Review of common one and two sample tests
    - The *t* statistic
- R instruction
    - Maximum Likelihood Estimation (MLE)

Joint Probability Distribution

- Consider two RVs X and Y
    - X represents a genotype of a certain locus: {AA, CC, AC}
    - Y indicates whether to have T2D or not: {normal, disease}
- Individuals are *instantiations* (or *realization*) of RVs X and Y
- **Joint probability P(X, Y)**
    - It actually refers to the following 6 probabilities:
        - P(X=AA, Y=normal), P(X=CC, Y=normal), P(X=AC, Y=normal)
        - P(X=AA, Y=disease), P(X=CC, Y=disease), P(X=AC, Y=disease)

Interpretation of P(X=AA, Y=normal)

- Frequency of observing individuals with X=AA and Y=normal

Joint Probability Distribution

- Consider two RVs X and Y
  - X represents a genotype of a certain locus: {AA, CC, AC}
  - Y indicates whether to have T2D or not: {normal, disease}
- **Conditional probability P(X | Y)**
  - It actually refers to the following 6 probabilities:
    - P(X=AA|Y=normal), P(X=CC|Y=normal), P(X=AC|Y=normal)
    - P(X=AA|Y=disease), P(X=CC|Y=disease), P(X=AC|Y=disease)

Interpretation of P(X=AA|Y=normal)

- Frequency of observing individuals with X=AA within the pool of individuals having Y=normal

$$P(X=AA \mid Y=\text{normal}) = \frac{P(X=AA, Y=\text{normal})}{\mathrm{P}(Y=\text{normal})}$$

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- **Discrete** $P(B) = \sum_{i=1}^{n} P(B \mid A = a_i)P(A = a_i)$

$$= \sum_{i=1}^{n} P(B, A = a_i) = P(B)$$

- **Continuous** $P(B) = \int P(B \mid A)P(A)dA$

Bayesian Estimation

- In order to make probability statements about θ given some observed data, D, we make use of Bayes' rule

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)} = \frac{P(\theta)P(D \mid \theta)}{\int P(\theta)P(D \mid \theta)d\theta}$$

Not a function of θ!

$$\boxed{P(\theta \mid D) \propto P(\theta)P(D \mid \theta)}$$

Posterior ∝ Prior × Likelihood

- The prior is the probability of the parameter and represents what was thought ***before*** observing the data
- The likelihood is the probability of the data given the parameter and represents the data now available
- The posterior represents what is thought given both prior information and the data just **observed**

Bayesian Estimation

- **Find θ such that the posterior P(θ|D) is maximized**
- **MLE: Find θ that maximizes log P(D|θ)**
- **BE: Find θ that maximizes log P(D|θ) + log P(θ)**

$$\boxed{P(\theta \mid D) \propto P(\theta)P(D \mid \theta)}$$

Posterior ∝ Prior × Likelihood

- The prior is the probability of the parameter and represents what was thought ***before*** observing the data
- The likelihood is the probability of the data given the parameter and represents the data now available
- The posterior represents what is thought given both prior information and the data just **observed**

Simple Example

- Say that we want to estimate the recombination fraction (θ) between locus A and B from 5 heterozygous (AaBb) people. We examined 30 gametes for each and observed 4,3,5,6 and 7 recombinants gametes in the five parents. What is the MLE of the recombination fraction θ?
- Let's simplify and ask what the recombination fraction (θ) is for subject # 3, who had 5 observed recombinant gametes.

Specifying the Posterior Density

$$P(\theta \mid D) = P(\theta \mid n=30, r=5) = \frac{P(\theta)P(r=5 \mid \theta, n=30)}{\int_0^{0.5} P(r=5 \mid \theta, n=30)P(\theta)d\theta}$$

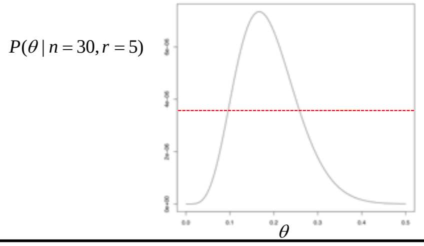
- **Prior** $P(\theta) = \text{uniform}[0,0.5] = 0.5$
- **Likelihood** $P(r=5 \mid \theta, n=30) = \binom{30}{5}\theta^5(1-\theta)\theta^{30-5}$
- **Normalizing constant** $\int_0^{0.5} P(r=5 \mid \theta, n=30)P(\theta)d\theta$

$$= 0.5 \cdot \binom{30}{5}\int_0^{0.5} \theta^5(1-\theta)^{25}d\theta \approx 6531$$

Specifying The Posterior Density

$$P(\theta \mid D) = P(\theta \mid n=30, r=5) = \frac{P(\theta)P(r=5 \mid \theta, n=30)}{\int_0^{0.5} P(r=5 \mid \theta, n=30)P(\theta)d\theta}$$

$$= \frac{0.5 \cdot \binom{30}{5}\theta^5(1-\theta)^{25}}{6531}$$



Goals

- Parameter estimation
  - Maximum likelihood estimation
  - Bayesian inference
- Hypothesis testing
  - Overview of key elements of hypothesis testing
  - Common one and two sample tests
- R session
  - Generating random numbers
  - T-test

Hypothesis Testing

- Formally examine two opposing conjectures (hypotheses), H_0 and H_A
- These two hypotheses are mutually exclusive and exhaustive so that one is true to the exclusion of the other
- We accumulate evidence – collect and analyze sample information – for the purpose of determining which of the two hypotheses is true and which of the two hypotheses is false

Example

- Consider a genome-wide association study (GWAS) for T2D and you measure the blood glucose level of the case/control groups
- **The null hypothesis, H_0:**
  - There is no difference between the case/control groups in the mean blood glucose levels
  - H_0: $\mu_1 - \mu_2 = 0$
- **The alternative hypothesis, H_A:**
  - The mean blood glucose levels in the case/control groups are "different"
  - H_A: $\mu_1 - \mu_2 \neq 0$

The Null and Alternative Hypothesis

- **The null hypothesis, H_0:**
  - States the assumption (numerical to be tested)
  - Begin with the assumption that the null hypothesis is TRUE
  - Always contains the "=" sign
- **The alternative hypothesis, H_A:**
  - Is the opposite of the null hypothesis
  - Challenges the status quo
  - Never contains just the "=" sign
  - Is generally the hypothesis that is believed to be true by the researcher

One and Two Sided Tests

- Hypothesis tests can be one or two sided (tailed)
- One tailed tests are directional:

  H_0: $\mu_1 - \mu_2 \leq 0$

  H_A: $\mu_1 - \mu_2 > 0$
- Two tailed tests are not directional:

  H_0: $\mu_1 - \mu_2 = 0$

  H_A: $\mu_1 - \mu_2 \neq 0$

P-values

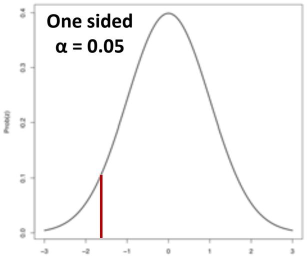
- Calculate a **test statistic** in the sample data that is relevant to the hypothesis being tested
  - e.g. In our GWAS example, the test statistic can be determined based on μ_1, μ_2 and σ_1, σ_2 computed from the GWAS data
- After calculating a test statistic we convert this to a **P-value** by comparing its value to distribution of test statistic’s under the null hypothesis
- Measure of how likely the test statistic value is under the null hypothesis

  P-value ≤ α → Reject H_0 at level α

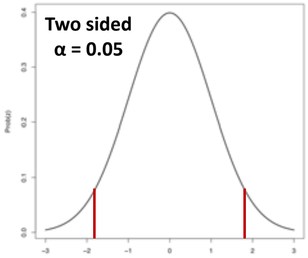
  P-value > α → Do not reject H_0 at level α

When To Reject H_0

- **Level of significance, α:** Specified before an experiment to define rejection region
- **Rejection region:** set of all test statistic values for which H_0 will be rejected



Critical Value = -1.64



Critical Value = -1.96 and +1.96

Some Notation

- In general, critical values for an α level test denoted as:

  One sided test: X_α

  Two sided test: $X_{\alpha/2}$

  where X depends on the distribution of the test statistic

- For example, if $X \sim$ N(0,1):

  One sided test: z_α (i.e., $z_{0.05} = 1.64$)

  Two sided test: $z_{\alpha/2}$ (i.e., $z_{0.05/2} = z_{0.05/2} = +-1.96$)

Errors in Hypothesis Testing

Actual Situation “Truth”

Decision	**H_0 True**	**H_0 False**
Don Not Reject H_0		
Reject H_0		

Errors in Hypothesis Testing

Decision	Actual Situation “Truth” H_0 True	 H_0 False
Don Not Reject H_0	Correct Decision 1-α	Incorrect Decision Type II Error B
Reject H_0	Incorrect Decision Type I Error α	Correct Decision 1-β

Type I and II Errors

Decision	Actual Situation “Truth” H_0 True	 H_0 False
Don Not Reject H_0	Correct Decision 1-α	Incorrect Decision Type II Error B
Reject H_0	Incorrect Decision Type I Error α	Correct Decision 1-β

α = P(Type I Error) β = P(Type II Error)
Power = 1 - β

Parametric and Non-Parametric Tests

- **Parametric Tests:** Relies on theoretical distributions of the test statistic under the null hypothesis and assumptions about the distribution of the sample data (i.e., normality)
- **Non-Parametric Tests:** Referred to as “Distribution Free” as they do not assume that data are drawn from any particular distribution

Whirlwind Tour of One and Two Sample Tests

	Type of Data		
Goal	Gaussian	Non-Gaussian	Binomial
Compare one group to a hypothetical value	One sample t-test	Wilcoxon test	Binomial test
Compare two paired groups	Paired t-test	Wilcoxon test	McNemar’s test
Compare two unpaired groups	Two sample t-test	Wilcoxon-Mann-Whitney test	Chi-square or Fisher’s exact test

Normality

- Use Gaussian (normal) distribution to explain a sample of *n* data points

$$x_1,\ x_2,\ \ldots,\ x_n$$

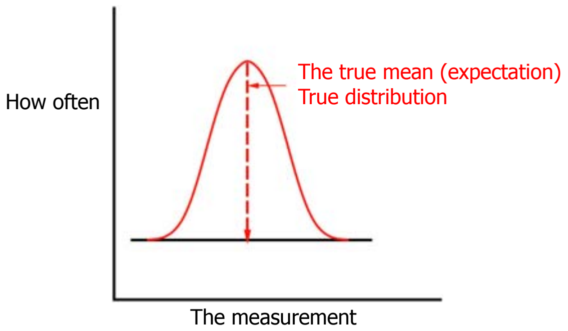
- The best estimate of the true mean μ is the average of the samples (called the *sample mean*)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

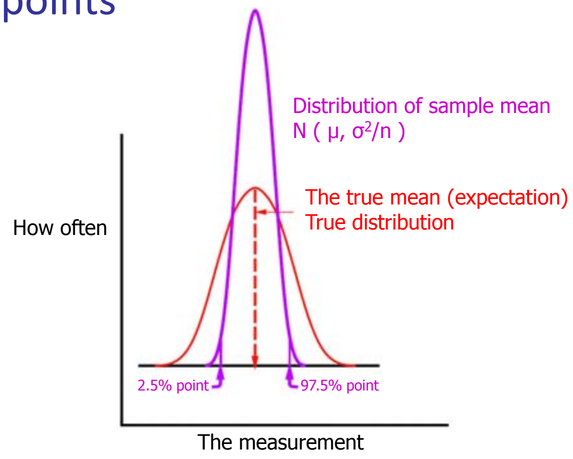
- How noisy the estimate will be?
- Can we make an interval estimate?

A Normal Distribution

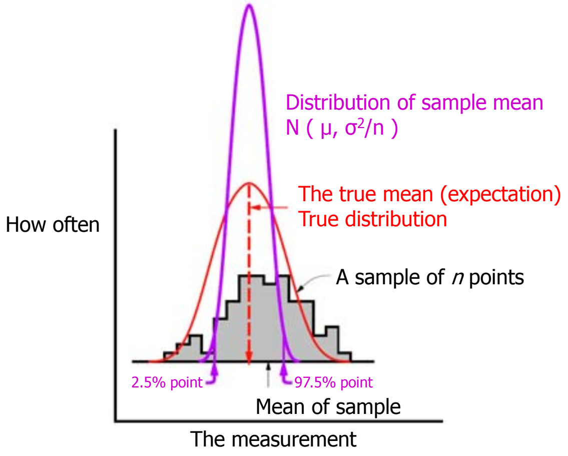
- Say that the (unknown) standard deviation of the true distribution is σ
- The variance of the sample mean (average of a sample of *n* points) is σ^2/n



The distribution of the sample mean of *n* points



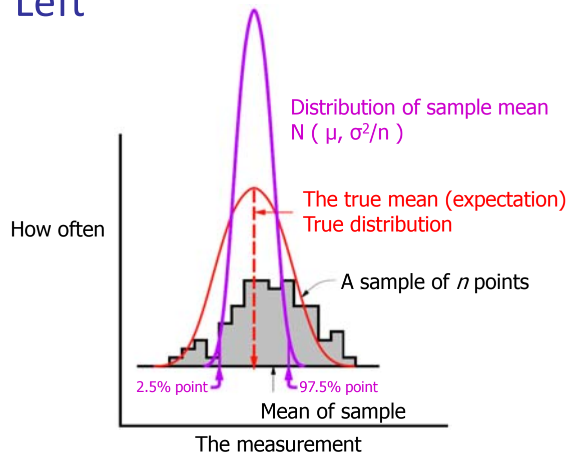
A Particular Sample



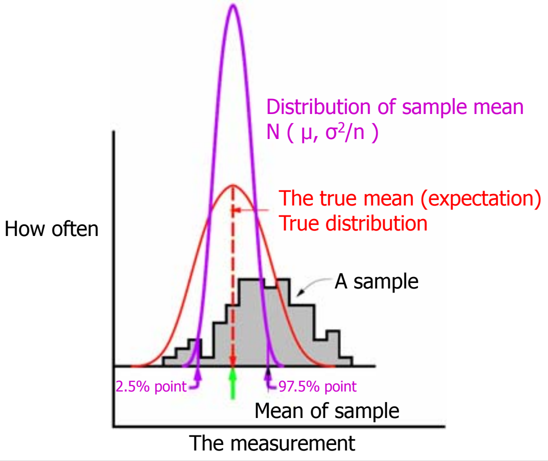
Confidence Interval

- So, that solves it, right?
- No! We don't know μ which is what we want to know!
- But, we can say that, 95% of the time, the sample mean $\bar{x}$ that we calculate is below that upper limit, and above that lower limit.

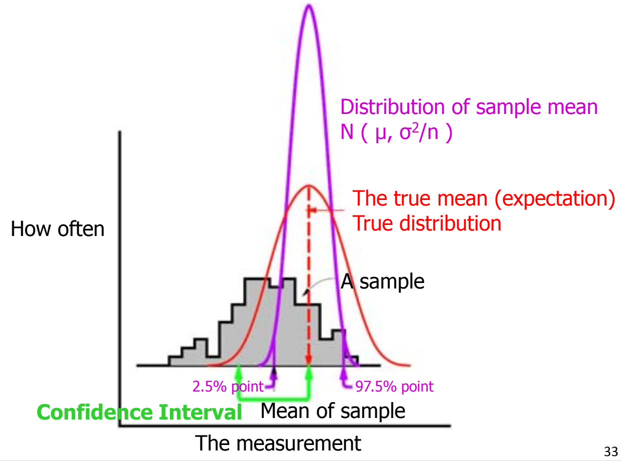
Let's Get Ready to Slide the True Stuff Left



Not Any Lower Than This …



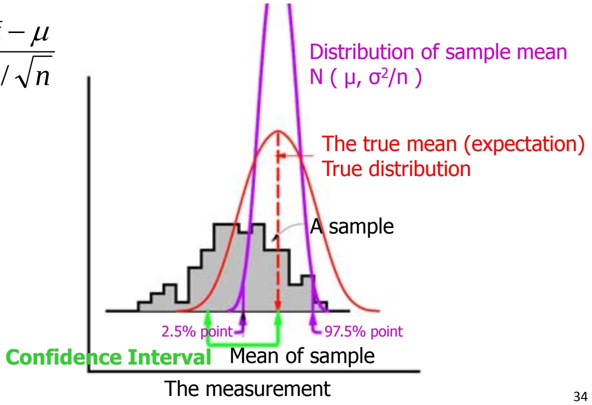
Not Any Higher Than This …



The *t* Statistic

- The number of (estimated) standard deviations of the sample mean that it deviates from its expected value μ

$$t = \frac{\overline{x} - \mu}{\hat{s} / \sqrt{n}}$$



The *t* Statistic

- The number of (estimated) standard deviations of the sample mean that it deviates from its expected value μ

$$t = \frac{\overline{x} - \mu}{\hat{s} / \sqrt{n}}$$

- where $\hat{s}$ is the estimated standard deviation, from a sample of *n* values, and $\overline{x}$ is the average of the sample
- This does not have a normal distribution but it is closer to normal the bigger *n* is. The quantity (*n*-1) is called the degrees of freedom of the *t* value