Basic Principles of Statistical Inference

Min Chen, Ph.D.
Quantitative Biomedical Research Center, Department of Clinical Sciences

Lecture 3
July 3, 2012

Some Important Concepts

Population and sample
- A data sample is taken from a population. For example:
  Measure the heights of 3 randomly chosen graduate students.
  Population = All graduate students
  Sample = 3 chosen students

Random variable
- Certain characteristics (e.g., heights) may vary among members of the population.

Parameter and statistics
- A parameter (i.e., $\mu$, $\sigma$) is a characteristic of the population—a constant number
- A statistics (i.e., $\bar{X}$, $s$) is a characteristic of the sample—a random variable

Types of Random Variables

Continuous variable
- Having a range of real values such as gene expression values

Categorical variable
- Ordinal – Obvious order to the categories, e.g., different dosages of medicine
- Nominal – No obvious order to the categories, e.g., type of cancer, gender, race
Statistical inference is about making sensible decisions under uncertainty.

- Based on probability theory.
- Make general conclusions from limited amounts of data.
- Extrapolate from a sample to a population.
- Separate objective decision making from personal opinion.
  
  *We don’t guess. We estimate!*

**Statistical Inference is based on data**

Statistical Inference is a process of drawing conclusions that are supported by the data.

- Degree of certainty.
- Account for variability in the data when generalizing beyond the immediate data to a population.

*“In God we trust; all others must bring data.”*  
- William Edwards Deming

Famous Guru of Quality Control, American Statistician, Professor, Management Consultant

**In God we trust; all others must bring data.**

- William Edwards Deming

Famous Guru of Quality Control, American Statistician, Professor, Management Consultant
Decriptive vs. Inferential Statistics

Descriptive
- Summarize data:
  - Measure of central tendency, dispersion, association, etc.
  - Usage of descriptive statistics
    - Identify pattern
    - Identify outliers
    - Leads to hypothesis generating

Inferential
- Draw conclusions from data:
  - Point estimation
  - Interval estimation
  - Hypothesis testing
  - Analysis of variance
  - Regression

Parameter Estimation

Estimate a numerical characteristic of a population.
- Measure of the central location
- Measure of the variability
- Frequencies

Basic Principles of Statistical Inference

Outline

1. Introduction of Statistical Inference
2. Statistical Inference Procedures
   - Point Estimation
   - Interval Estimation
   - Hypothesis Testing

Parameter – Numerical Characteristic

- A geneticist wants to estimate the allele frequency of certain SNP in a target population.
- A clinician wants to estimate the amount of viral shedding in patients with a certain infection.
- A biochemist wants to estimate the average concentration of a certain protein in a patient population.

Wanzhu Tu (2007); From Methods in Molecular Biology, Vol 404: Topics in Biostatistics
Q: Should one assess every single member of a population?

A: This is called census and is only feasible for a limited number of applications. For a large population, one would:

- Draw a random sample of subjects from the intended population.
- Then use sample data to estimate the unknown parameter.

Two requirements for parameter estimation:

1. Sample must be representative of the population;
2. Estimate must be good in some sense.

Goal: Estimate the mean age of this target population.

Population N=300,000

Pretend we don’t know the real mean age!
To estimate the mean age of the population $\mu$.
- We could pick any number from the observed sample data, say, 35 years or 62 years:
  \[
  \hat{\mu} = 35 \\
  \hat{\mu} = 62
  \]
- Sometimes underestimates and other times overestimates
- Is this a good estimator?
  - NO!

Finite sample properties
- Minimum Variance
  - Estimator varies little ("fluctuates little")
  - i.e., does not vary too much from sample to sample
- Unbiasedness
  - Neither overestimate nor underestimate the parameter on average

Asymptotic properties
- Consistency, efficiency, etc.
Sample mean, $\bar{X}$, is a good estimator of population mean $\mu$, (actually the "best")

- It has minimum variance and is unbiased

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

**Sample Mean is a Random Variable**

**Distribution of Sample vs. Distribution of Sample Mean**

**Outline**

1. Introduction of Statistical Inference
2. Statistical Inference Procedures
   - Point Estimation
   - Interval Estimation
   - Hypothesis Testing

**Estimation of Mean**

- Sample mean, $\bar{X}$, is a good estimator of population mean $\mu$, (actually the "best")
- It has minimum variance and is unbiased

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

**Basic Principles of Statistical Inference**
Let $\theta$ be the parameter in which we are interested, e.g., age.

We want an interval estimator from the sample data, rather than a point estimator.

The interval should reflect uncertainties of our estimation.

Choose a value $\alpha$ and $(1 - \alpha)$ is called the confidence level;

- i.e., degree of certainty

- Based on data $X$, construct lower and upper bounds $L(X)$ and $U(X)$;

- The probability that the constructed interval $(L, U)$ contains the true parameter $\theta$, is $(1 - \alpha)$:

$$Pr(L(X) \leq \theta \leq U(X)) = 1 - \alpha.$$  

Often we set $\alpha = 0.05$.

$(L(X), U(X))$ is called a 95% Confidence Interval if

$$Pr(L(X) \leq \theta \leq U(X)) = 1 - 0.05 = 0.95.$$
A Common Mistake in Interpretation of CI

- The CI:
  \[ Pr(L(X) \leq \theta \leq U(X)) = 1 - \alpha. \]
- It is tempting to state "the probability that the \( \theta \) lies between two numbers, \( L \) and \( U \), is \( (1 - \alpha) \)."
  - Wrong because \( \theta \) is a fixed number;
  - \( L(X) \) and \( U(X) \) are random variables, not numbers.

95% Confidence Interval

- Q: What does 95% confidence interval mean?
- A: If we select 100 sets of samples from the population, then on average 95 out of the 100 calculated intervals will contain the true population parameter \( \theta \).

(\( 1 - \alpha \))% Confidence Interval of the Mean

Lower Limit:
\[ L = \bar{X} - z_{\alpha/2} \times \frac{s}{\sqrt{n}} \]
Upper Limit:
\[ U = \bar{X} + z_{\alpha/2} \times \frac{s}{\sqrt{n}} \]

95% Confidence Interval:
\[ \bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \]
**Example 1**

Prostate specific antigen (PSA) is a protein produced by the cells from the prostate and is used as a biomarker of prostate cancer. Table 1 contains PSA levels of 30 patients who had radical prostatectomy, measured 6 months after surgery.

What is the mean PSA level among the general population of patients who undergo prostatectomy at 6 months post-surgery?

---

**Confidence Interval of Mean using Excel**

**Table 1: Prostate Specific Antigen (PSA) Levels (ng/mL)**

<table>
<thead>
<tr>
<th>PSA (ng/mL)</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Alpha</th>
<th>z alpha/2</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.1559 - 0.2837</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.3452</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.1999 - 0.4865</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.3381</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.1828 - 0.3834</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.3058</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.1460 - 0.3656</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2667</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0988 - 0.3346</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0391 - 0.3143</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0391 - 0.3143</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0391 - 0.3143</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0391 - 0.3143</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.2267</td>
<td>0.05</td>
<td>1.6448</td>
<td>0.0391 - 0.3143</td>
</tr>
</tbody>
</table>

**Confidence Interval from Small Sample**

As a rule of thumb, if sample size, N < 30, use the formula below.

\[
(1 - \alpha)\% \text{ Lower Limit: } \quad L = \bar{X} - t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}
\]

\[
(1 - \alpha)\% \text{ Upper Limit: } \quad U = \bar{X} + t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}
\]

\[
(1 - \alpha)\% \text{ Confidence Interval: } \quad \bar{X} \pm t_{\alpha/2, n-1} \times \frac{s}{\sqrt{n}}
\]

where \( t_{\alpha/2} \) is the \((\alpha/2)\)th quantile of the \( t \)-distribution with \((n-1)\) degrees of freedom.

**Confidence Interval of the Variance**

\[
(1 - \alpha)\% \text{ Lower Limit: } \quad L = \frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}}
\]

\[
(1 - \alpha)\% \text{ Upper Limit: } \quad U = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}
\]

\[
(1 - \alpha)\% \text{ Confidence Interval: } \quad \left( \frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}} \right)
\]

where \( \chi^2_{\alpha/2, n-1} \) and \( \chi^2_{1-\alpha/2, n-1} \) are the \((\alpha/2)\)th and \((1 - \alpha/2)\)th quantile of the \( \chi^2 \)-distribution with \((n-1)\) degrees of freedom, respectively.
Example 2

The manufacturer of a particular model of a bone density scanner randomly selected 4 machines for testing. The four scanners gave following readings: 4.1, 4.0, 3.9, 3.9.

We want to know how precise is this model of bone density scanner. That is, what is the estimated variance of this model?

<table>
<thead>
<tr>
<th>Measurement</th>
<th>4.1</th>
<th>4.0</th>
<th>3.9</th>
<th>3.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std Dev.</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Variance</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Alpha.</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Chi^2 alpha/2: 9.3484
Chi^2 (1-alpha/2): 0.2158

95% Confidence interval of Variance:
\[ (0.003, 0.127) \]

Variance of the bone density scanner is estimated to be 0.00917, with

95% Confidence Interval:
\[ (0.003, 0.127) \]

Hypothesis Testing

- Another type of statistical inference procedure;
- Pre-specify a research statement;
- Use sample data to decide whether the statement is true or false;
- Research Statement is called "Hypothesis".
Hypothesis Testing

- Approach is to specify two contradictory research statements
  - Null hypothesis ($H_0$)
  - Alternative hypothesis ($H_a$)
- Null is just the status quo.
- Alternative hypothesis
  - is the opposite of null hypothesis
  - is generally the hypothesis that is believed to be true by the researcher

Logic of Hypothesis Testing is parallel to Proof by contradiction.
- Assume null hypothesis $H_0$ is true.
- Draw sample data from population.
- If the data are unlikely to occur under the null hypothesis, the alternative hypothesis $H_a$ will be accepted.
The significance level is denoted as $\alpha$.
- It is the pre-specified maximum probability that the researcher will incorrectly reject the null hypothesis $H_0$ when it is true.
- If $\alpha = 0.05$, then the researcher specifies that there is max 5% chance that their research statement could be incorrectly shown to be true, when in fact the null is true.

A test statistic $T$ is a quantity that measures how likely or unlikely the sample data is drawn under null hypothesis.
- $T$ is a random variable that depends on sample data.
- If the value of $T$ is LARGE, then data show evidence against the null hypothesis. Hence the alternative hypothesis will be accepted.

Suppose the null hypothesis is true.
- What is the probability of observing the test statistic (or more extreme values) by chance alone?
- Set a rule to reject $H_0$ if the above probability is below $\alpha$.
- This is called rejection region.

P-value is the probability that an observed value (or more extreme values) of the test statistic could have occurred solely by chance under null hypothesis.
- Computed from test statistics.
- The actual probability of incorrectly rejecting the null hypothesis for given sample data.
- If P-value is "large", then there is a "large" risk of incorrectly rejecting the true null.
- If P-value is below a cut off $\alpha$, it is referred as "statistically significant".
  - Therefore, accept the alternative hypothesis.
**How to Think About P-Values**

- A P-value is a conditional probability—the probability of the observed statistics given that the null hypothesis is true.
- The P-value is NOT the probability that the null hypothesis is true.
- It's not even the conditional probability that null hypothesis is true given the data.

**One-sample T-test**

- Revisit example of PSA levels
- Is the mean PSA level in prostate cancer patients greater than zero after prostatectomy?

**One-sample T-test Using Excel**

- \( H_0 : \mu = 0 \text{ ng/mL} \) vs. \( H_a : \mu > 0 \text{ ng/mL} \)
- Let \( \alpha = 0.05 \)
- Test statistic
  \[
  T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{0.2267 - 0}{0.5252/\sqrt{30}} = 2.3642
  \]
- \( P = Pr(T \geq 2.3642) = 0.0125 \)
  Since \( P < 0.05 \), reject the null.
  Mean PSA levels are greater than 0.
**Types of Errors in Hypothesis Testing**

- **Actual situation**
  - No difference ($H_0$)
  - Difference ($H_1$)

- **Measured**
  - Accept $H_0$
  - Reject $H_0$

- **Actual situation**
  - FOB screening (bowel cancer)

---

**Relationship between α and Power**

- Suppose the Null Hypothesis is true.
- Suppose the Null Hypothesis is not true.

---

**Power**

Power ($1 - \beta$, aka true positive rate (TP))

- Probability of detecting a significant scientific difference when it does exist

---

**Power**

Power depends on:

- Sample size $n$
- Standard deviation $\sigma$
- Size of the difference you want to detect $\delta$
- False positive rate $\alpha$

The sample size is usually adjusted to make power equal 0.8.
relationship between power and its affecting factors

Power increases as:
- Sample size \( n \) increases
- Standard deviation \( \sigma \) decreases
- Detectable difference \( \delta \) increases
- FP rate \( \alpha \) increases

Example: Assuming a standard deviation of 0.6 ng/mL, if we want to detect a 0.20 pg/mL change in PSA level with 0.05 significance level and 80% power in a two-sided test, how many patients do we need for the test?

\[
n = \left\lceil \frac{z_{\alpha/2} + z_{\beta}}{\delta} \right\rceil^2 = \left\lceil \frac{(1.96 + 0.84) \times 0.6}{0.2} \right\rceil^2 \approx 71
\]

references

- Rafia Bhere. Lecture notes.
- Xiayu Huang, Bioinformatics Shared Resource, Sanford, Burnham Medical Research Institute. 
  http://bssweb.burnham.org/ 
- Jason A. Osborne, Notes. Department of Statistics, North Carolina State University