Final Review

Fall 2013
Prof. Yao Xie, yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial Systems & Engineering
Georgia Tech
Random sampling model

random samples: $x_1, \ldots, x_n$

For example, we use digital thermometer to measure body temperature for 5 times, we obtain a sequence.
If we do this experiment the next day, we get a different sequence of measures.
The result of the measurement is a sequence of random samples (also called data).
Population: all possible observations

Sample: observed measurements taken from population

Statistics: useful summary of the sample data
Descriptive statistics

- Quantitative values
  - provides simple summaries about samples
- plot

Histogram

Box plot

Stem & Leaf diagram

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>5 1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>5 8 0</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>1 0 3</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>4 1 3 5 3 5</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>2 9 5 8 3 1 6 9</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>4 7 1 3 4 0 8 8 6 8 0 8</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>3 0 7 3 0 5 0 8 7 9</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>8 5 4 4 1 6 2 1 0 6</td>
<td>10</td>
</tr>
<tr>
<td>18</td>
<td>0 3 6 1 4 1 0</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>9 6 0 9 3 4</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>7 1 0 8</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>1 8 9</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)
Linear Regression

Numerical Data

Normal Distribution

N(μ, σ²)

\[ \hat{\mu} = \bar{X} \]

\[ \hat{\sigma}^2 = S^2 \]

Categorical Data

Binomial Distribution

X ~ Bin(n, p)

\[ \hat{p} = \frac{X}{n} \]

Random Variables

Point Estimation

Confidence Interval

(L(\hat{\mu}), U(\hat{\mu}))

Hypothesis Testing

H₀: μ = μ₀

Confidence Interval

(L(\hat{p}), U(\hat{p}))

Hypothesis Testing

H₀: p = p₀

Statistical Inference

Inference on Multiple Populations

Confidence Interval

μ₁ - μ₂

Hypothesis Testing

H₀: μ₁ = μ₂

Confidence Interval

p₁ - p₂

Hypothesis Testing

H₀: p₁ = p₂

ANOVA

Statistical Modeling

Linear Regression

Contigency Tables

Logistic Regression
Data summary

- **Samples**  $x_1, x_2, \cdots, x_n$
- **Sample mean**  $\bar{x} = \frac{1}{n} x_1 + \frac{1}{n} x_2 + \cdots + \frac{1}{n} x_n$
- **Sample median**
  - 1) rank samples from smallest to largest
    $y_1, y_2, \cdots, y_n$
  - 2) odd number of samples, median = $y_{(n+1)/2}$
    even number of samples, median =
    $\frac{(y_{(n-1)/2} + y_{(n-1)/2})}{2}$
• Sample range = largest - smallest

• Sample variance

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

• Sample quartile \( x_p \): \( p \)th quartile is such that \( p \)-percent of samples are smaller than \( x_p \)

• upper quartile

• lower quartile

• Inter quartile range (IQR) = upper quartile - lower quartile
Sampling distribution

- Distribution of the **statistics** we come up (above)
- Sampling distribution extremely useful for determining
  - forms of confidence interval
  - hypothesis test
## Sampling distribution: summary

<table>
<thead>
<tr>
<th>Form</th>
<th>Sample mean</th>
<th>Sample variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample mean</td>
<td>Sample variance</td>
</tr>
<tr>
<td><strong>Form</strong></td>
<td>$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$</td>
<td>$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$</td>
</tr>
<tr>
<td><strong>Sample i.i.d. normal</strong></td>
<td>$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$</td>
<td>$\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$</td>
</tr>
<tr>
<td><strong>Known variance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unknown variance</strong></td>
<td>large $n$, approximately normal as above</td>
<td>large $n$, approximately normal</td>
</tr>
</tbody>
</table>
Other common sampling distribution

<table>
<thead>
<tr>
<th>Sample proportion</th>
<th>Standardized sample mean, known variance</th>
<th>Standardized sample mean, unknown variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{p} = \frac{X}{n} )</td>
<td>( \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} )</td>
<td>( \frac{\bar{X} - \mu}{\sqrt{S^2 / n}} )</td>
</tr>
<tr>
<td><strong>Exact:</strong></td>
<td>( n\hat{p} \sim BIN(n, p) )</td>
<td>( \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0, 1) )</td>
</tr>
<tr>
<td><strong>Large sample:</strong></td>
<td>( \hat{p} \sim N(np, np(1 - p)) )</td>
<td></td>
</tr>
</tbody>
</table>
## Two sample

<table>
<thead>
<tr>
<th>Difference in sample mean, known variance</th>
<th>Difference in sample mean, unknown (but identical) variance,</th>
<th>Proportion of sample variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \left( \frac{X_1 - X_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) \sim N(0, 1) ]</td>
<td>[ \left( \frac{X_1 - X_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right) \sim t_{n_1 + n_2 - 1} ]</td>
<td>[ \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{n_1 - 1, n_2 - 1} ]</td>
</tr>
</tbody>
</table>

\[ S_p^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2} \]
Statistical methods

• Point estimator
• Confidence interval
• Hypothesis test

• Two sample test (two populations)
• ANOVA (more than two populations)

• Linear regression
Point estimator

- Mean of estimator: unbiased
- Variance of estimator
- Mean Square Error (MSE)
  - $\text{MSE} = \text{bias}^2 + \text{variance}$

- Method of finding point estimators
  - method of moment
  - maximum likelihood
Confidence interval

• Point estimator: a single value for estimated parameter
• Confidence interval: an interval such that true parameter lies in

• \([a, b]\) contains true parameter with probability \(1 - \alpha\)
• then \([a, b]\) is the \(1 - \alpha\) confidence interval
Typical forms of $k$

- $k = \text{upper cutting point} \times \text{variance of point estimator}$

\[
\left( \bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)
\]

\[
\left( \bar{x} - \frac{s}{\sqrt{n}} t_{\alpha, n-1}, \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha, n-1} \right)
\]

\[
\left( \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} - z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \right)
\]

- width of confidence interval determined by sample size and confidence level
Tails etc

\[ \Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \, du \]

\( \Phi(z) \)

CDF

Upper cutting point (also called “percentage point” in textbook)
Forms of confidence intervals

• Two-sided interval
  \[ \text{point estimator} - k, \text{point estimator} + k \]

• One-sided interval
  \[ \text{point estimator} + k, \infty \]
  \[ -\infty, \text{point estimator} - k \]

• \( k \) specifies width of confidence interval
Hypothesis test

• Use data to test two contradicting statements
  • $H_0$: null hypothesis
  • $H_1$: alternative hypothesis

• Two approaches
  • Fixed confidence level
    • **Form**: reject $H_0$ when **test statistic** falls out of **thresholds**
  • **p-value**
    • probability of observing something more “extreme” than data
Procedure of hypothesis test (sec. 9.1.6)

1. Set the significance level (.01, .05, .1)
2. Set null and alternative hypothesis
3. Determine other parameters
4. Decide type of the test
   - test for mean with known variance (z-test)
   - test for mean with unknown variance (t-test)
   - test for sample proportion parameter
5. Use data available:
   - perform test to reach a decision
   - and report p-value
## Summary: test for mean

### Null Hypothesis

\[ H_0 : \mu = \mu_0 \]

### Test Statistic

\[ \bar{x} \]

**Significance level:** $\alpha$

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>Known Variance $H0$ is rejected if</th>
<th>Unknown Variance $H0$ is rejected if</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1 : \mu \neq \mu_0$</td>
<td>$</td>
<td>\bar{x} - \mu_0</td>
</tr>
<tr>
<td>$H_1 : \mu &gt; \mu_0$</td>
<td>$\bar{x} &gt; \mu_0 + z_{\alpha} \sigma / \sqrt{n}$</td>
<td>$\bar{x} &gt; \mu_0 + t_{\alpha,n-1} s / \sqrt{n}$</td>
</tr>
<tr>
<td>$H_1 : \mu &lt; \mu_0$</td>
<td>$\bar{x} &lt; \mu_0 - z_{\alpha} \sigma / \sqrt{n}$</td>
<td>$\bar{x} &lt; \mu_0 - t_{\alpha,n-1} s / \sqrt{n}$</td>
</tr>
</tbody>
</table>
# Test for sample proportion

## Null Hypothesis

\[ H_0 : p = p_0 \]

## Test Statistic

\[ \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \]

Significance level: \( \alpha \)

## Alternative Hypothesis

\[ H_1 : p \neq p_0 \]

<table>
<thead>
<tr>
<th>Alternative Hypothesis</th>
<th>( H_0 ) is rejected if</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_1 : p \neq p_0 )</td>
<td>( \left</td>
</tr>
</tbody>
</table>
Two sample test: mean

For the following hypothesis test

\[ H_0 : \mu_1 - \mu_2 = \Delta \]
\[ H_1 : \mu_1 - \mu_2 \neq \Delta \]

Reject \( H_0 \) when

\[
\left| \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \right| > t_{\alpha/2}
\]
Two-sample test: sample proportion

For two-sided test, \( H_0 : p_1 = p_2 \) and \( H_1 : p_1 \neq p_2 \).

Reject \( H_0 \) when

\[
\left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > z_{\alpha/2}
\]
Analysis of variance

- Multiple populations
- Analyze difference in their means

We would reject $H_0$ if

$$F_0 > F_{\alpha,a-1,a(n-1)}$$
Linear regression

• Simple linear regression

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \ldots, n \]
Fitted coefficients

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \quad (11-10)
\]

\[
S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \quad (11-11)
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

\[
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}
\]

\[
\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i
\]

Fitted (estimated) regression model
Model diagnosis

• Plot residuals
• Use R and read the output
• For simple and multiple linear regression: we are going to rely on R to do the calculations
Finally…

The Sampling Distribution…

...is the distribution of a statistic across an infinite number of samples.
Finally…

• What statistics is about?
• Fit model using data (e.g. distributions)
• Use model to make inferences
  • estimation
  • hypothesis testing
  • prediction (e.g. using linear regression)
• Why model is useful?
  • report findings from data
  • systematically quantify uncertainty
Dear News Media,

When reporting poll results, please keep in mind the following suggestions:

1. If two poll numbers differ by less than the margin of error, it's not a news story.
2. Scientific facts are not determined by public opinion polls.
3. A poll taken of your viewers/internet users is not a scientific poll.
4. What if all polls included the option "Don't care"?

Signed,

-Someone who took a basic statistics course.