Lecture 9
Outline

Review of Previous Lecture

Estimation when the Variance is Unknown
   Estimating Standard Deviation
   t-distribution
   Non-normal Data
   Bootstrap

Brief Review for Exam 1
Part I

Review of Previous Lecture
Interpretation of Confidence Interval

For a population parameter of interest, suppose we have an estimate $\hat{\theta}$ based on a random sample.

The 100\% confidence interval for $\theta$ has the form $(\hat{\theta} - a, \hat{\theta} + b)$.

If the experiment is repeated many times, then in the long run, about 100\% of the CIs constructed this way will include the true population parameter.
The construction of the CI depends solely on the sampling distribution of $\hat{\theta}$.

There is no requirement for the CI to be symmetric about $\hat{\theta}$ (though CLT-type CI’s always are).

Be careful in interpreting statements of the type:

$$\bar{x} \pm \text{margin of error}$$

In our book, the “margin of error” is always $z^* \times \sigma / \sqrt{n}$ where $z^*$ is usually 1.96 (95% confidence), but in newspaper or other less scientific sources, this is less clear. It is probably safest to write CI’s explicitly as $(a, b)$, along with their percent confidence (90%, 95%, or whatever).
Review of Previous Lecture

Computing CI for Sample Mean

- When the population variance $\sigma^2$ is known, the sample mean $\bar{X}$ has a $\mathcal{N} (\mu, \sigma / \sqrt{n})$ distribution, where $\mu$ is the population mean and $n$ is the sample size.

- The 100$C\%$ confidence interval for $\mu$ is given by:

$$\left( \bar{X} - z^* \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + z^* \frac{\sigma}{\sqrt{n}} \right),$$

where $z^*$ is the upper $(1 - C)/2$-point of the standard normal distribution.

- Again, for a 95% CI, we have $z^* = 1.96 \approx 2$. 

PH 6450 Fall 2005
Sample Size Estimation

For a desired CI margin of error (half-width) $m$, and a given confidence level $100C\%$, the sample size required is:

$$n = \left( \frac{z^* \sigma}{m} \right)^2.$$
Part II

Estimation when the Variance is Unknown
Variance Estimation

The sample variance $s^2$ can be used to estimate the population variance $\sigma^2$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$ 

- $\bar{x}$ is the sample mean.
- The factor $1/(n - 1)$ is used so that $s^2$ will be an unbiased estimator of $\sigma^2$. 

Estimation when the Variance is Unknown

Estimating Standard Deviation
t-distribution
Non-normal Data
Bootstrap


Standard Error

Definition

For \( n \) i.i.d. samples with population variance \( \sigma^2 \), the variance of the sample mean is \( \sigma^2/n \), which can be estimated by \( s^2/n \).

\( s/\sqrt{n} \) is often referred as the standard error (of the sample mean).
Sample Distribution of the Mean

Recall that when $n$ is large, the Central Limit Theorem says

$$\bar{X} \xrightarrow{\text{approx}} \mathcal{N} \left( \mu, \sigma^2 / n \right).$$

Therefore, the so-called z-score:

$$Z \equiv \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{\text{approx}} \mathcal{N}(0, 1).$$
z-score and CI

The 100C% CI for the sample mean is derived as follows:

\[
\Pr(-z^* < Z < z^*) = C
\]

\[
\Rightarrow \Pr \left( -z^* < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z^* \right) = C
\]

\[
\Rightarrow \Pr \left( -z^* \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z^* \frac{\sigma}{\sqrt{n}} \right) = C
\]

\[
\Rightarrow \Pr \left( -\bar{X} - z^* \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z^* \frac{\sigma}{\sqrt{n}} \right) = C
\]

\[
\Rightarrow \Pr \left( \bar{X} + z^* \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z^* \frac{\sigma}{\sqrt{n}} \right) = C
\]
Student’s $t$ distribution

Definition
When $X_i \overset{iid}{\sim} N(\mu, \sigma)$ with $\sigma$ unknown,

$$t \equiv \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

the Student’s $t$ distribution with $n - 1$ degrees of freedom.

This distribution was discovered in 1908 by William S. Gosset, a statistician who worked for the Guinness Brewing Company. The brewery regarded the $t$ distribution as a trade secret, forcing Gosset to publish under the pen name “Student”!
Normal and $t$ distributions

As the number of degrees of freedom increase, the $t$ distribution becomes closer and closer to a standard normal distribution, i.e.,

$$t_{n-1} \rightarrow Z \text{ as } n \rightarrow \infty$$

Java Applet to play with various $t$ distributions:
http://www.stat.vt.edu/~sundar/java/applets/TNormal.html

or just see plot on p.494 of M&M!
Percentiles of $t$ distribution

Definition

The $100 \times u$th percentile of a $t$ distribution with $d$ degrees of freedom is denoted by $t_{d,u}$, that is,

$$\Pr(t_d < t_{d,u}) \equiv u.$$ 

A few values from Table D (inside back cover of book):

<table>
<thead>
<tr>
<th>$d$</th>
<th>$t_{d,0.975}$</th>
<th>$z_{0.975}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.776</td>
<td>1.960</td>
</tr>
<tr>
<td>9</td>
<td>2.262</td>
<td>1.960</td>
</tr>
<tr>
<td>29</td>
<td>2.045</td>
<td>1.960</td>
</tr>
<tr>
<td>60</td>
<td>2.000</td>
<td>1.960</td>
</tr>
</tbody>
</table>
Confidence Interval when the Variance is Unknown

Definition
A $100(1 - \alpha)\%$ confidence interval (CI) for the mean $\mu$ of a population with unknown population variance is given by:

$$\bar{X} \pm t^* \frac{s}{\sqrt{n}}. \quad (2)$$

where $t^* = t_{(n-1,1-\alpha/2)}$, the upper $\alpha/2$-point of a $t_{n-1}$ distribution (from Table D).

This interval is exact when the $X_i$ are normally distributed, and approximately correct for large $n$ in other cases.

**Awkward difficulty:** Hard to check normality of the data when $n$ is small enough to make us want to use the $t$ in the first place! (Remember $t_{n-1} \approx Z$ for large $n$.)
Computing the $t$ interval with SAS

```
DATA arter;
  INPUT sbp @@;
CARDS;
  194 126 130 98 136 145 110 108 102 126
PROC MEANS DATA = arter N NMISS MEAN CLM;
  VAR sbp;
RUN;
```
### Computing CI with SAS

**Analysis Variable**: SBP

<table>
<thead>
<tr>
<th>N</th>
<th>Nmiss</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>127.5000000</td>
</tr>
</tbody>
</table>

**Lower 95.0% CLM** | **Upper 95.0% CLM**

| 107.5233035 | 147.4766965 |
Normality and Approximation

\[ t \equiv \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1} \]

only when

1. The population has a normal distribution, or,
2. The sample size is large (say, \( n > 15 \)) and the population distribution is not too far away from normal (e.g., no huge outliers).
Non-normality

- When the sample size is small and the distribution is skewed to the right (i.e., heavy right tail), we can use a one-to-one transformation of the data (e.g., logarithm or square root) and work on the transformed scale instead.
- That is, we compute the CI for the mean of the log-transformed data, then convert back to the original scale by taking the inverse transformation (e.g., exponential or square) of the lower and upper interval endpoints.
- While this will produce a more exact CI on the transformed scale, this may not be a great CI on the original scale...
  - The exponential (anti-log) of the mean of the log-transformed data is not a good estimator of the mean of the original data (since the log of the means $\neq$ mean of the logs).
Variance Estimation via the Bootstrap

- Computing a CI based on $\hat{\theta}$ requires knowledge of its sampling distribution (in particular, its variance).
- If $\hat{\theta} = \bar{X}$, we know this sampling distribution (at least for moderate $n$) by the Central Limit Theorem: it’s roughly normal with variance $\sigma^2/n$.
- But what if we don’t know the sampling distribution?
- The bootstrap is a technique where we treat our sample as (an estimate of) the entire population, and resample from it to get an estimate of the sampling distribution.
A Bootstrap Example

- Here is a data set with ten systolic blood pressure measurements (mmHg):
  98 102 108 110 126 126 130 136 145 194

- The sample median is 126.

- We are interested in estimating the population median, and a corresponding confidence interval.
Bootstrap Step 1: Resampling

- First we draw random samples of the same size *with replacement* from our original sample. Here is one such sample of size $n = 10$:
  
  98 110 110 126 126 126 130 130 130 145

- Note that some of the original observations appear multiple times, while some of the original observations are not in the sample.

- The sample median for this bootstrap sample is also 126.
Bootstrap Step 2: Computing the Statistic of Interest

Having drawn 1000 such samples, we compute the median of each boostrapped sample, resulting in 1000 boostrapped medians:
Bootstrap Step 3: Inference

- What we have now is an estimate of the sampling distribution of the original sample median.

- We can compute the mean of the sample medians, which turns out to be 123, as a (better) point estimate of the population median.

- More importantly, we can use the (0.025, 0.975) percentiles of the sample distribution as a 95% confidence interval for the population median: (108, 141).

- “Pulling ourselves up by our own bootstraps”!

- See MM pp.427-428, and your TA’s SAS demo in HW 5...
Part III

Brief Review for Exam 1
Lectures 1-2

- *Descriptive Statistics*: mean, median, mode, range, interquartile range (IQR), variance, standard deviation
- *Graphical Displays*: Boxplot, stemplot, scatterplot, multivariate plots
- *Exploratory Data Analysis*: tables, pie and bar charts, histograms, symmetry and modality
- *Normal Distribution*: standard normal \( Z \sim N(0, 1) \), “forward” and “backward” use of Table A (computing normal and inverse normal probabilities), normal q-q plot, transformations to improve normality
Lectures 3-4

- **Central Limit Theorem:**
  \[
  X_i \overset{iid}{\sim} ? (\mu, \sigma) \implies \bar{X} \overset{approx}{\sim} N(\mu, \sigma / \sqrt{n})
  \]

- **Data Transformations:** We often take the log or square-root to improve the symmetry, normality of a sample of data, and reduce the impact of outliers

- **Experimental Design:** anecdotal evidence, observational study, experiment, census

- **Sampling:** simple random sample (SRS), bias, regression to the mean, Berkson’s Fallacy

- **Observation vs. Experimentation:** Confounders, cohort study, case-control study, placebo effect, blinding, randomization
Lectures 5-6

- **Elements of Probability:** Outcome space, long-run frequency interpretation
- **Rules of Probability:**
  1. $0 \leq P(A) \leq 1$
  2. $P(S) = 1$
  3. complement rule
  4. additive rule for *mutually exclusive* events
  5. multiplicative rule for *independent* events
- **Random Variables:** Bernoulli (binary), nominal, ordinal, discrete quantitative, continuous (e.g. normal and uniform)
- **Population Moments:** mean and variance, rules for computing (e.g., $E[aX] = aE[X]$, etc.)
- **Bayes’ Rule:** for finding the probability of events where the conditioning is in the reverse order
Lectures 7-8

- **Binomial Distribution**: characteristics, sampling with and without replacement, mean and variance, representation as $S = \sum X_i$ for $X_i \sim Bernoulli$, sample mean $\hat{p} = S/n$, unbiasedness, normal approximations for $S$ and $\hat{p}$

- **Normal Distribution**: Distribution of $\bar{X}$ when $X_i \overset{iid}{\sim} Normal$, or when $X_i \overset{iid}{\sim} anything$ and $n \geq 30$ (Central Limit Theorem)

- **Confidence Intervals**: Definition, interpretation (long run versus any single data set), calculation, margin of error, sample size estimation