1 Review

• **Binomial Approximation**: When \( X \sim Bin(n, p) \) and
  
  (1) \( np \geq 15 \), (2) \( n(1 - p) \geq 15 \), then you can approximate \( X \sim N(np, np(1 - p)) \)

In other words, new random variable,

\[
Z = \frac{X - np}{\sqrt{np(1 - p)}},
\]

is approximately \( N(0,1) \).

• Remember the fact: \( X \sim Binomial(n, p) \) (or Bin(n,p)),
  then \( E[X] = np \), \( Var[X] = np(1-p) \). In summary, under the two conditions,

• Exact: \( X \sim Bin(n, p) \)

• Approximate: \( X \sim N(np, np(1 - p)) \)

• Approximate: \( Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1) \).
2 Ch 7. Sampling dist.

• The observations $X_1, X_2, \ldots, X_n$ are called i.i.d or random sample of size $n$ from the population distribution, if they result from independent selections and each observation has the same distribution as the population. (i.i.d means independent and identically distributed.)

• The Distribution of sample mean $\overline{X}$ (i.i.d observation). Suppose $EX_1 = \mu$ and $VarX_1 = \sigma^2$, then

$$E[\overline{X}] =$$

$$Var[\overline{X}] =$$

$$SD[\overline{X}] =$$
eg1) Flip a fair coin 20 times (0 or 1). Define $X_i = \text{ith result}$. Remember,

$$E[X_1] =$$

$$Var[X_1] =$$

$$E[X] =$$

$$Var[X] =$$

$$SD[X] =$$
3 Ch7.3 $\overline{X}$ (Sample mean) distribution

- case 1: $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma)$. Then,

$$\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}), \quad (1)$$

or, equivalently,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1). \quad (2)$$

Note: This is exact.

- case 2: (Central Limit Theorem (CLT))

$X_1, X_2, \cdots, X_n \sim$ unknown AND $n \geq 30$. Then,

$$\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}), \quad (3)$$

or, equivalently,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1). \quad (4)$$

Note: This is an approximation.
We need these equations (approximation) soon. For example, **Testing of statistical hypotheses** (Ch8, p321) More details coming soon.

eg1)(p307) Example 5

Population mean $= \mu = 82$, population sd $\sigma = 12$ are given.

Q1) If 64 samples are selected randomly, what is the probability that the sample mean will lie between 80.8 and 83.2?

A1) $64 \geq 30$, use Central Limit Theorem (CLT).

$\mu = \mu$

$\sigma/\sqrt{n} =$

$P(80.8 < \bar{X} < 83.2) =$
Q2) If 100 samples are selected randomly, what is the probability that the sample mean will lie between 80.8 and 83.2?

\[ \mu = \]
\[ \sigma / \sqrt{n} = \]
\[ P(80.8 < \bar{X} < 83.2) = \]
4 Ch8 Inference about means under large samples

Here large samples means \( n > 30 \). Remember, under this condition (LEC17 p5), we have

**Central Limit Theorem (CLT)**

\[ X_1, X_2, \cdots, X_n \sim \text{unknown} \text{ AND } n \geq 30. \]

Then,

\[
\sum_{i=1}^{n} \frac{X_i}{n} = \bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}),
\]

or, equivalently,

\[
\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}(\frac{\bar{X} - \mu}{\sigma}) \sim N(0, 1).
\]

**Note:** This is an **approximation**.

Consider p321 Example 1. Table 1 showed Number of Bee Stings Per Year Reported by Beekeepers. There are 40 observations (n=40). This is a large sample case. Based on the data, we have

- sample mean \( \bar{x} = 308.9 \)
- sample sd \( s = 597.8 \)
- sample median =95
- first quartile=3.5
- third quartile =200
However, our interest lies not just in his particular set of 40 observations but in the vast population of all beekeepers and even potential beekeepers. Unknown things are:

1. **population distribution** of number of stings per year received by a beekeeper
2. **population mean** $\mu$
3. **population sd** $\sigma$.

In this course, we are especially interested in guess 2 and 3 above ($\mu$ and $\sigma$). Guessing 1 is beyond the scope of this course. You need high level math (functional analysis).

First, as the Ch8’s title says, let’s consider $\mu$. What can we say about $\mu$?

- **Statistical Inference** (p320): deals with drawing conclusions about population parameters from an analysis of the sample data.

  3 types of inference:

- **Inference 1.** Point estimation

  Use a single value to estimate the unknown mean $\mu$.

  We could estimate $\mu$ to be 308.9. That is, we guess population mean using sample mean $\bar{x} = 308.9$. 
We could estimate $\mu$ to be 95. That is, we guess population mean using sample median 95.

We could estimate $\mu$ to be 3. That is, we guess population mean using the first observation $x_1 = 3$ in the TABLE1.

We could estimate $\mu$ to be $\frac{(3 + 2000)}{2} = 1001.5$. That is, we guess population mean using the average of 1st and 40th observation $(x_1 + x_{40})/2 = 1001.5$ in the TABLE1.

There are many ways to guess $\mu$. In our course, we use sample mean to guess population mean, i.e., $\bar{x} = 308.9$ because there are nice properties.

• **Inference 2.** Interval estimation
  Determine an interval of plausible values for $\mu$. Using Section 3 method (coming soon), we conclude $\mu$ is between 123.6 and 494.2. (people write $(123.6, 494.2)$) Remember, $\mu$ is unknown. 123.6 and 494.2 are based on observations.

• **Inference 3.** Testing hypotheses
  Decide whether the $\mu = 260$ or not.
Let’s consider point estimation of a population mean. Remember (Ch7), if $X_i$ is iid (i=1,...,n), with $E[X_1] = \mu$, $Var[X_1] = \sigma$. Then

$E[\bar{X}] = \mu$

$SD[\bar{X}] = \sigma/\sqrt{n}$

Now new terminologies:

- **Standard Error (S.E.)** (p323): Standard deviation of the estimator

- **Error margin** (p323) (use p261 results): Prior to sampling,

  $P(\bar{X} \text{ is within the range } \mu \pm 1\sigma/\sqrt{n}) = 0.683.$

  $\Leftrightarrow$ When we are estimating $\mu$ by $\bar{X}$, the 68.3 % error margin is

  $P(\bar{X} \text{ is within the range } \mu \pm 2\sigma/\sqrt{n}) = 0.954.$

  $\Leftrightarrow$ When we are estimating $\mu$ by $\bar{X}$, the 95.4 % error margin is

  $P(\bar{X} \text{ is within the range } \mu \pm 3\sigma/\sqrt{n}) = 0.997.$

  $\Leftrightarrow$ When we are estimating $\mu$ by $\bar{X}$, the 99.7 % error margin is

Look at p323 Fig1.
p261 results can give you 3 specific error margins. If we want to know more general case, for example, 90 % error margin, what should we do? No problem.

**When we are estimating $\mu$ by $\bar{X}$, the 100(1-$\alpha$) % error margin is $Z_{\alpha/2} \cdot \sigma / \sqrt{n}$**

- Values of $Z_{\alpha/2}$: the upper $\alpha/2$ point of the standard normal distribution. (See p324 Fig2).

$$Z_{\alpha/2} = -qnorm(\alpha/2)$$
$$= qnorm(1 - \alpha/2)$$  \hspace{1cm} (7)

Simply speaking, $\alpha$ = area, and $Z_{\alpha/2}$=location. Remember, given location ($Z_{\alpha/2}$), we can find area ($\alpha$ or $\alpha/2$). And given area ($\alpha$ or $\alpha/2$), we can find location ($Z_{\alpha/2}$) in standard normal distribution.

<table>
<thead>
<tr>
<th>1-$\alpha$</th>
<th>0.80</th>
<th>0.85</th>
<th>0.90</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_{\alpha/2}$</td>
<td>1.28</td>
<td>1.44</td>
<td>1.645</td>
<td>1.96</td>
<td>2.58</td>
</tr>
</tbody>
</table>
So, to find 90 \% error margin,

step 1) Solve $1 - \alpha : 90 = 100(1 - \alpha)$.

This gives

$$1 - \alpha =$$

$$\alpha =$$ \hspace{1cm} (8)

step 2) Look at the table above ($1 - \alpha$ column). (or just use qnorm($1 - \alpha/2$))

$$Z_{\alpha/2} =$$

Why do we have $\alpha/2$ instead of $\alpha$? Looks unnatural! But look at p324 Fig2. $\alpha/2$ corresponds to the left area and the right area of the location $-Z_{\alpha/2}$ and $+Z_{\alpha/2}$. Hence total area is $\alpha$.

- When $\sigma$ is unknown, we can use

$$\sigma \approx sd = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}}$$ \hspace{1cm} (9)
Here, left hand side is population sd, right hand side is sample sd.

p325 Example 2: Find 80% (not 95 %) Error margin for the Bee example.
• 100(1-\(\alpha\)) Confidence Interval (C.I.) : Probability that C.I. covers the true \(\mu\) is 100(1-\(\alpha\)). See p333 Fig 4.

\((\bar{X} \text{- error margin, } \bar{X} + \text{ error margin})\)

In other words,

\((\bar{X} - Z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + Z_{\alpha/2} \cdot \sigma / \sqrt{n})\)

If \(\sigma\) is unknown,

\((\bar{X} - Z_{\alpha/2} \cdot \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} / \sqrt{n}, \bar{X} + Z_{\alpha/2} \cdot \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} / \sqrt{n})\)

eg) Calculate 95% Confidence Interval for the Bee example.

ANS:

\(\bar{x} = \)

\(s = \)

\((1 - \alpha) = \)

\(Z = \)
eg) Calculate 90% Confidence Interval for the Bee example.

ANS:

\( \bar{x} = \) 

\( s = \) 

\( (1 - \alpha) = \) 

\( Z = \)
Use Bee example, but we now have one more observation, $X_{41} = 5000$. Use same $\sigma \approx s = 597.8$. Calculate 95% Confidence Interval for the Bee example. (note: $\bar{x}$ and n are changed.)

ANS:

$\bar{x} =$

$s =$

$(1 - \alpha) =$

$Z =$