Introduction before Ch8

A. Distinguish between population and sample

The entire group of objects or people about which information is wanted is called the **population**.

A **sample** is a part of the population that is actually examined in order to gather information.

B. Distinguish between parameter and statistic

A **parameter** is a number describing the population.

A **statistic** is a number that can be computed from the data without making use of any unknown parameters.

C. Sampling Distribution

The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

In this course, we should understand what the sampling distribution of the sample mean is.

Some symbols:

$$\begin{cases} \textit{Population mean} \qquad \mu \\ \\ \textit{Population variance} \qquad \sigma^2 \end{cases}$$

$$\begin{cases} \textit{Sample mean} \qquad \bar{x} \\ \\ \textit{Sample variance} \qquad s^2 \end{cases}$$

D. Sampling distribution of the sample mean

If a population has the $N(\mu, \sigma^2)$ distribution, then the sample mean of n independent observations has the $N(\mu, \sigma^2/n)$ distribution. [Note: this is not applying CLT, and note too that n need not be large]

If the population distribution is not normal, then for **large n**, we can use the Central Limit Theorem which states that (for large n) the sampling distribution of $\bar{X}$ is **approximately** $N(\mu, \sigma^2/n)$ for any population with finite variance σ^2.

E. Other sampling distributions

For statistics other than the sample mean, there are also sampling distributions.

In STAT 241/251, we need to be familiar with

- Sampling distribution of sample mean (use normal distributions or CLT to solve)
- Sampling distribution for difference of means (Ch8)

In other introductory statistics courses, the following sampling distribution is also introduced:

- Sampling distribution for proportions (not covered in STAT 241/251)

We will work through some problems so you get comfortable with the terminology used in sampling distributions problems, and also how to use the correct symbols and write out your solutions.

Explain to class what a SRS (simple random sample) is:

A simple random sample (SRS) of size n consists of n units from the population chosen in such a way that every set of n units has an equal chance to be the sample actually selected.

Example on sampling distribution of the mean (using CLT)

The time X that a technician requires to perform preventive maintenance on an air conditioning unit is governed by the exponential distribution. The mean time $\mu = 1$ hour and the standard deviation is $\sigma = 1$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

Solution:

Let $\bar{X}$ be the random variable representing the sample mean of 70 air-conditioning units. By central limit theorem,

$$\bar{X} \overset{approx.}{\sim} N(\mu, \sigma^2/70), \quad where\ \mu = 1\ and\ \sigma^2 = 1^2$$

$$\Rightarrow \quad \bar{X} \sim N(1, 1/70)$$

$$\begin{aligned} P(\bar{X} > 0.83) &= P(Z > \frac{0.83 - 1}{\sqrt{1/70}}) \\ &= P(Z > -1.42) \\ &= 1 - P(Z \leq -1.42) \\ &= 0.9222 \end{aligned}$$

note 50 mins=0.83 hr

Sampling distribution for difference of means

Theory:

We'll revisit this again in Ch 8, but it is mentioned here to illustrate the sampling distribution for difference of means.

(Two-sample Z statistic)

Suppose that $\bar{x}_1$, is the mean of an SRS of size n_1, drawn from an $N(\mu_1, \sigma_1^2)$ population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2^2)$ population.

Then

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

[Later in Ch8, you will learn about **2-sample t** statistic which is used when the population standard deviations σ_1 and σ_2 are not known]

Example on sampling distribution for difference of means

A fourth-grade class has 12 girls and 8 boys. The children's height are recorded on their 10th birthdays. What is the chance that the girls are taller than the boys?

We translate the question into the following: what is the probability that the mean height of the girls is greater than the mean height of the boys?

Additional info needed to solve this problem: Based on information from Health

Canada, we assume that the heights (in inches) of 10-year old girls are $N(56.4, 2.7^2)$ and the heights of 10-year old boys are $N(55.7, 3.8^2)$. The heights of the students in our class are assumed to be random samples from these populations.

Solution:

Let $\bar{X}_1 - \bar{X}_2$ represent the difference in female and male mean heights.

$$E(\bar{X}_1 - \bar{X}_2) = 56.4 - 55.7 = 0.7$$

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{2.7^2}{12} + \frac{3.8^2}{8} = 2.41$$

$$\bar{X}_1 - \bar{X}_2 \sim N(0.7, 2.41)$$

$$P(\bar{X}_1 - \bar{X}_2) > 0) = P(Z > \frac{0 - 0.7}{\sqrt{2.41}})$$

$$= P(Z > -0.45)$$

$$= 0.67$$

<u>One more definition - unbiased estimator</u>

A statistic used to estimate a parameter is an unbiased estimator of the parameter if the mean of its sampling distribution is equal to the true value of the parameter.

$\hat{}$ means "estimate of"

[Explain to class using symbols: $E(\hat{\theta}) = \theta$]

Fact:

① Sample mean is an unbiased estimator of population mean. [$\bar{x} = \frac{\sum x}{n}$]

② Sample variance is an unbiased estimator of population variance [Remind class sample variance appeared in Ch 1 and the formula to calculate sample variance is

$$\underbrace{[SD(x)]^2}_{\text{sample standard deviation } s} = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}$$

Shortcut formula:

$$\begin{aligned}
\frac{\sum(x-\bar{x})^2}{n-1} &= \frac{\sum(x^2-2x\bar{x}+\bar{x}^2)}{n-1} \\
&= \frac{\sum x^2 - 2\bar{x}\sum x + n\bar{x}^2}{n-1} \\
&= \frac{\sum x^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2}{n-1} \qquad [since\ \bar{x} = \frac{\sum x}{n}] \\
&= \frac{\sum x^2 - n\bar{x}^2}{n-1} \\
&= \frac{\sum x^2 - n(\frac{\sum x}{n})^2}{n-1} \\
&= \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}
\end{aligned}$$

Learn:

$$\frac{\sum(x-\bar{x})^2}{n-1} = \underbrace{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}_{\text{use this for faster calculations}}$$

Practice:

What is the (sample) variance of: 3, 4, 5, 4, 6, 7 ?

① Learn to use your calculator if it has such a function

② By hand

$$\bar{X} = \frac{\sum x}{n} = 29/6, \quad \sum x^2 = 151, \quad n = 6$$

$$\begin{aligned} \textit{sample variance} &= \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \\ &= \frac{151 - \frac{(29)^2}{6}}{5} \\ &= 2.166667 \end{aligned}$$

③ Using R:

```
> data<-c(3,4,5,4,6,7)

> var(data)

[1] 2.166667
```