# Supplementary Handout for Dis 8: Sampling Distributions

## 1   Motivation

- Prior to this week, we discussed random variables, and the possible probability distributions that such random variables could follow.

- As we learned from Dis 5 and 6,

    - A random variable assign a number to each possible outcome.
    - A discrete probability distribution describes the point probability at all possible values for a discrete random variable.
    - A continuous probability distribution describes the density (PDF) at all possible values for a continuous random variable.

    Thus, these measures are related to the population.

- However, in reality, what we get to work with is often the sample data, which means we need to relate statistics obtained from samples to the population ($\Rightarrow$ process of statistical inference).

- This is why we need to look at the distribution of sample statistics, i.e. **sampling distributions**

## 2   Examples of Sampling Distributions

### 2.1   Sampling distribution of the mean

- Statistic of interest: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, obtained from simple random sampling

- How $\bar{X}$ is distributed depends on the distribution of $X_i$:

    - If each $X_i$ is normally distributed, then $\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$ with certainty
    - If each $X_i$ is NOT normally distributed, we might be able to approximate $\bar{X}$ using a normal distribution (i.e. $\bar{X} \overset{a}{\sim} N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$) based on central limit theorem.

    > **Definition 1** (Central limit theorem (CLT)).  The mean of a random variable drawn from any population is approximately normal for a sufficiently large sample size.

    In practice, we use $n \geq 30$ as the cutoff:

      * For non-normally distributed $X_i$, if $n \geq 30$, then CLT can be invoked, and $\bar{X} \overset{a}{\sim} N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$
      * For non-normally distributed $X_i$, if $n < 30$, then CLT cannot be invoked, so the distribution of $\bar{X}$ is undetermined.

- To summarize, for random variable $X$, the sampling distribution of the mean is the following:

| | $X$ **is normally distribution** | $X$ **is NOT normally distribution** |
|---|---|---|
| **Sample size is small ($n < 30$)** | | |
| **Sample size is large ($n \geq 30$)** | | |

- What is $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}^2$?

  - $\mu_{\bar{X}}$ is the expected value of $\bar{X}$:

  $$\mu_{\bar{X}} = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} \times n \times \mu_X = \mu_X$$

  - $\sigma_{\bar{X}}^2$ is the variance of $\bar{X}$, and it depends on the population size:
    * If population size is infinitely large (in practice, if $N \geq 20n$),

    $$\sigma_{\bar{X}}^2 = V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \times n \times \sigma_{X}^2 = \frac{\sigma_{X}^2}{n}$$

    * If population size is not infinitely large (in practice, if $N \geq 20n$), then a **finite population correction factor** needs to be added:

    $$\sigma_{\bar{X}}^2 = \frac{\sigma_{X}^2}{n} \cdot \frac{N - n}{N - 1}$$

## 2.2 Sampling distribution of the proportion (from a binomial experiment)

- Say that we have a random variable $X \sim \text{Binomial}(n, p)$ recording the number of successes in $n$ trials where the probability of success in each trial is $p$.

- Turns out, under certain conditions, $X$ can be well approximated by a normal distribution.

---

Conditions for normal approximation of a binomial random variable $X$:

1. $np \geq 5$, and
2. $n(1 - p) \geq 5$

---

If the aforementioned conditions are satisfied, then

$$X \overset{a}{\sim} N(\mu_X, \sigma_X^2)$$

where, based on binomial distribution properties,

$$\mu_X = E(X) = np$$
$$\sigma_X^2 = V(X) = np(1 - p)$$

---

Aside: A binomial $X$ is a discrete random variable. However, the approximation approximates $X \overset{a}{\sim} N(np, np(1 - p))$, which is a continuous distribution.

Thus, a **correction factor for continuity** is needed when calculating probability using the normal approximation.

---

Exercise. Accounting for the correction factor for continuity, how should the following probabilities be expressed for a binomial random variable $X$?

1. $P(X = 3) =$

2. $P(X \geq 3) =$

3. $P(X > 3) =$

4. $P(X \leq 3) =$

5. $P(X < 3) =$

- Why is this needed? $\Rightarrow$ helps us approximate the sampling distribution of the proportion!

  - As long as a binomial distributed $X$ can be approximated using a normal distribution (i.e. $np \geq 5$ and $n(1 - p) \geq 5$), then the proportion of successes ($\hat{p}$) can be approximated using a normal distribution:

  $$\hat{p} = \frac{X}{n} \overset{a}{\sim} N(\mu_{\hat{p}}, \sigma_{\hat{p}}^2)$$

  - What is $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}^2$?

  $$\mu_{\hat{p}} = E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = p$$

  $$\sigma_{\hat{p}}^2 = V(\hat{p}) = V\left(\frac{X}{n}\right) = \left(\frac{1}{n}\right)^2 V(X) = \frac{p(1-p)}{n}$$

## 2.3 Sampling distribution of the difference between two means

- Statistic of interest: $\bar{X} - \bar{Y}$, where $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$, and $X$ is independent of $Y$
- From subsection 2.1, assuming that the population sizes are sufficiently large, we know that

$$\bar{X} \sim N(\mu_X, \frac{\sigma_X^2}{n_X}) \qquad \bar{Y} \sim N(\mu_Y, \frac{\sigma_Y^2}{n_Y})$$

- Since the sum of two normal distributions is still a normal distribution, we have

$$\bar{X} - \bar{Y} \sim N(\mu_{\bar{X}-\bar{Y}}, \sigma_{\bar{X}-\bar{Y}}^2)$$

where

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

$$\sigma_{\bar{X}-\bar{Y}}^2 = V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) - 2\underbrace{Cov(\bar{X}, \bar{Y})}_{=0 \text{ by indep}} = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$