Sampling Distributions

Dr Wan Nor Arifin
Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.
wnarifin@usm.my

Last update: 25 September, 2018

Outlines

Introduction

Sampling distribution

- It is "the distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population" (Daniel, 1995)
- Most of the time, we are interested in **mean**, **variance** and **functional form** of sampling distribution.

Constructing a sampling distribution

1. For a population size *N*, calculate population statistics of interest (e.g. mean, variance).

2. Randomly draw all possible samples of size *n* from the population. Calculate sample statistics from each sample.

3. List down all distinct values of sample statistic obtained. For each value, count the frequency and calculate the relative frequency. Plot a graph of frequency vs distinct values of sample statistic (e.g. mean).

One Sample Mean

Steps

1. We start we a population of N = 3 below, for variable X

$$X = \{3, 6, 9\}$$

with mean μ

$$\mu=\sum_{i=1}^{N} x_i=\frac{3+6+9}{3}=6$$

and variance σ^2

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu^2=\frac{1}{3}(3^2+6^2+9^2)-6^2=6$$

2. For a sample size n = 2, there will be

$$N^n=3^2=9 \text{ possible samples with replacement}$$

and

$$\binom{N}{n}=\binom{3}{2}=\frac{3!}{(3-2)!\,2!}=3 \text{ possible samples without replacement.}$$

Calculate sample mean $\bar{x}$ for each sample,

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{2}(x_i+x_2) \text{ for each sample}$$

followed by mean of the sample means $\mu_{\bar{x}}$

$$\mu_{\bar{x}}=\frac{1}{N^n}\sum_{i=1}^{N^n}\bar{x}_i \text{ for sampling with replacement}$$

$$\mu_{\bar{x}}=\frac{1}{\binom{N}{n}}\sum_{i=1}^{\binom{N}{n}}\bar{x}_i \text{ for sampling without replacement}$$

and variance of the sample means $\sigma_{\bar{x}}^2$

$$\sigma_{\bar{x}}^2=\frac{1}{N^n}\sum_{i=1}^{N^n}(\bar{x}_i-\mu_{\bar{x}})^2=\frac{1}{N^n}\sum_{i=1}^{N^n}\bar{x}_i^{\,2}-\mu_{\bar{x}}^2 \text{ for sampling with replacement}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} (\bar{x}_i - \mu_{\bar{x}})^2 = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \bar{x}_i{}^2 - \mu_{\bar{x}}^2 \quad \text{for sampling without replacement}$$

For sampling **with replacement** with 9 possible samples,

Sample no.	Sample		Mean, $\bar{x}$
	x_1	x_2	
1	3	3	3.0
2	3	6	4.5
3	3	9	6.0
4	6	3	4.5
5	6	6	6.0
6	6	9	7.5
7	9	3	6.0
8	9	6	7.5
9	9	9	9.0

Mean of sampling distribution,

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{N^n}\sum_{i=1}^{N^n}\bar{x}_i \\ &= \frac{1}{3^2}(3.0+4.5+\ldots+9.0) \\ &= \frac{1}{9}\times 54 \\ &= 6\end{aligned}$$

thus,

$$\mu_{\bar{x}}=\mu=6$$

Variance of sampling distribution,

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{1}{N^n}\sum_{i=1}^{N^n}\bar{x}_i^{\,2}-\mu_{\bar{x}}^2 \\ &= \frac{1}{3^2}(3.0^2+4.5^2+\ldots+9.0^2)-6^2 \\ &= \frac{1}{9}\times 351-36 \\ &= 3\end{aligned}$$

thus,

$$\sigma_{\bar{x}}^2=\frac{\sigma^2}{n}=\frac{6}{2}=3$$

For sampling **without replacement** with 3 possible samples,

Sample no.	Sample		Mean, $\bar{x}$
	x_1	x_2	
1	3	6	4.5
2	3	9	6.0
3	6	9	7.5

Mean of sampling distribution,

$$\begin{aligned}\mu_{\bar{x}} &= \frac{1}{\binom{N}{n}}\sum_{i=1}^{\binom{N}{n}} \bar{x}_i \\ &= \frac{1}{\binom{3}{2}}(4.5+6.0+7.5) \\ &= \frac{1}{3}\times 18 \\ &= 6\end{aligned}$$

thus,

$$\mu_{\bar{x}}=\mu=6$$

Variance of sampling distribution,

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{1}{\binom{N}{n}}\sum_{i=1}^{\binom{N}{n}} \bar{x}_i^{\,2}-\mu_{\bar{x}}^2 \\ &= \frac{1}{\binom{3}{2}}(4.5^2+6.0^2+7.5^2)-6^2 \\ &= \frac{1}{3}\times 112.5-36 \\ &= 1.5\end{aligned}$$

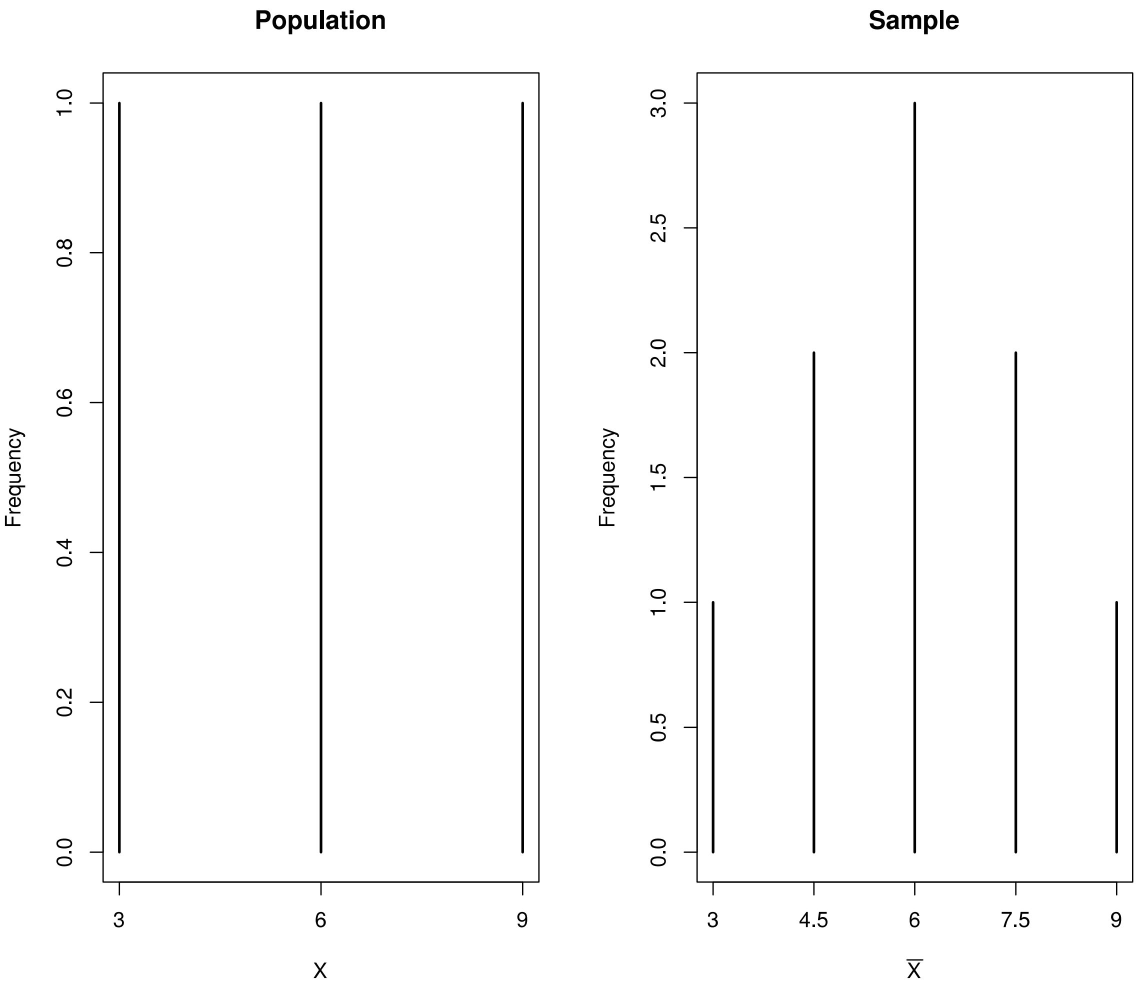
thus,

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{\sigma^2}{n}\times\frac{N-n}{N-1} \\ &= \frac{6}{2}\times\frac{3-2}{3-1} \\ &= \frac{3}{2}=1.5\end{aligned}$$

3. List distinct values of sample mean, here we show for sampling with replacement,

Mean, $\bar{x}$	Frequency	Relative frequency
3.0	1	0.111
4.5	2	0.222
6.0	3	0.333
7.5	2	0.222
9.0	1	0.111

and plot the graph,



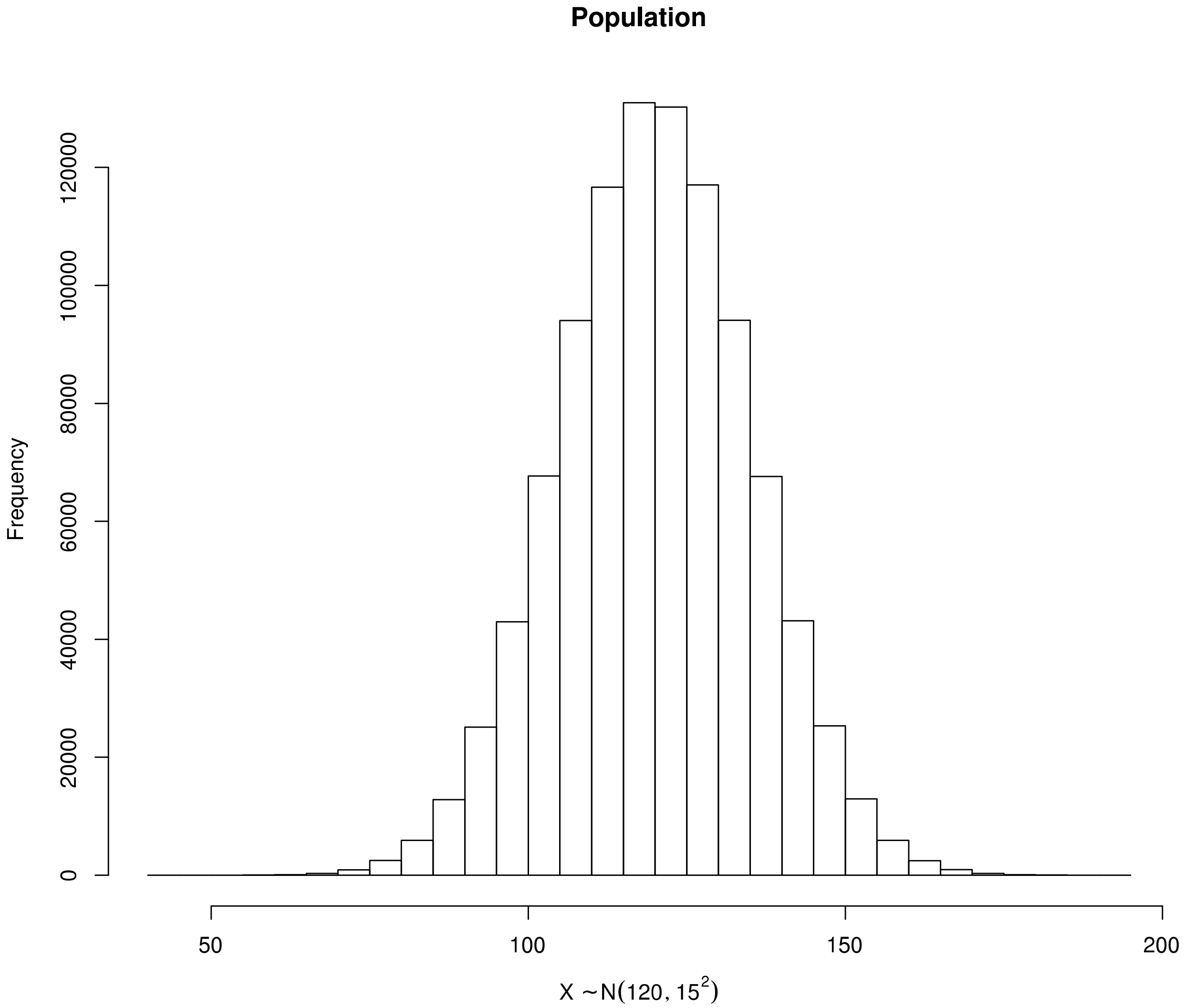
Using R
Open `samp_dist.R` and we will see how this work for large population of size N, with k samples and n sample size each. For example,

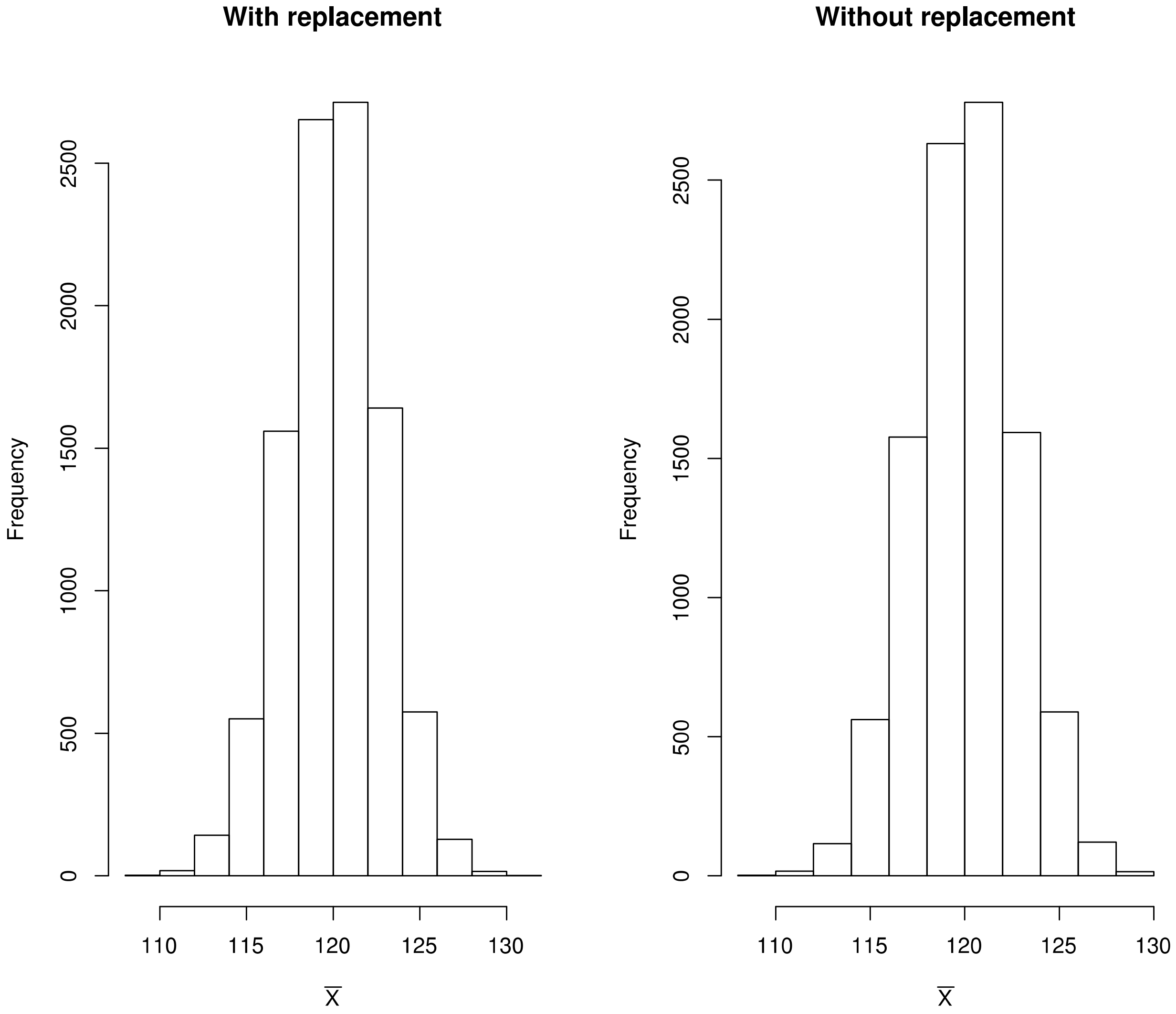
$$X \sim N(120, 15^2)$$

we simulate for

$$
\begin{aligned}
N &= 10^6 \\
n &= 30 \\
k &= 10^4
\end{aligned}
$$

Plots,





Summary

- The mean and variance of sampling distribution of one sample mean are given by,

$$
\begin{aligned}
\mu_{\bar{x}} &= \mu \\
\sigma_{\bar{x}}^2 &= \sigma^2/n \quad \text{(with replacement)} \\
\sigma_{\bar{x}}^2 &= \frac{\sigma^2}{n}\times\frac{N-n}{N-1} \quad \text{(without replacement)} \\
\sigma_{\bar{x}} &= \sigma/\sqrt{n} \quad \text{(with replacement)} \\
\sigma_{\bar{x}} &= (\sigma/\sqrt{n})\times\sqrt{\frac{N-n}{N-1}} \quad \text{(without replacement)}
\end{aligned}
$$

Application

- By plugging in the mean and standard deviation of sampling distribution into our z formula, we get,

$$z = \frac{x - \mu}{\sigma} \quad \rightarrow \quad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example 1:

In large human population it is known that systolic blood pressure is normally distributed with mean of 120mmHg and a standard deviation of 15mmHg.

Using R

Set `sd` argument value in `pnorm()` function to `sd = sd/sqrt(n)`.

a) What is the probability that a random sample of 30 subjects from the population will have a mean of less than 100mmHg?

Answer:

b) What is the probability that a random sample of 30 subjects from the population will have a mean between 110mmHg and 130mmHg?

Answer:

c) What is the probability that a random sample of 30 subjects from the population will have a mean more than 140mmHg?

Answer:

One Sample Proportion

- To show sampling distribution of one sample proportion, consider proportion as special case of mean of outcomes consisting of 0 and 1.
- Let say, to find proportion of drug addicts in a population of N = 10,

population = {1, 1, 1, 0, 1, 0, 0, 1, 0, 0}

drug addict = 1, non-addict = 0

proportion = frequency / n = 5/10 = 0.5

mean = total / n = 1+1+1+0+1+0+0+1+0+0 / 10 = 5/10 = 0.5

- Can use one sample mean formula.
- Replace the mean and variance of one sample mean formulas with mean and variance of binomial distribution.
- Remember binomial distribution is concerned with total number of success, *x*.
- Proportion/mean of binomial outcomes = *x*/*n*.
- Thus, in place of binomial,

$$\mu = np$$
$$\sigma = np(1-p)$$

by dividing with *n*, we have for proportion,

$$\mu = p$$
$$\sigma^2 = p(1-p)$$
$$\sigma = \sqrt{p(1-p)}$$

thus for sampling distribution of one sample proportion (using formula for one sample mean),

$$\mu_{\hat{p}} = p$$

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

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

and standard normal distribution, *z*

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

*Please take note that, to use *z* distribution, must make sure that *np* and *n(1-p)* is more than 5,

e.g. $n = 10$, $p = 0.5$, $1 - p = 0.5$ → both *np* and *n(1 – p)* equal to 5.

$n = 10$, $p = 0.2$, $1 - p = 0.8$ → $np = 2$, $n(1 - p) = 8$, cannot use *z* distribution to find probability.

Example 2:

In a population of drug addicts, it is known that 85% of them are HIV positive. In a random sample of 150 drug addicts, find

a) $P(\hat{p} \geqslant .9)$

Answer:

b) $P(\hat{p} \leqslant .8)$

Answer:

c) $P(.83 \leqslant \hat{p} \leqslant .88)$

Answer:

The Difference Between Two Sample Means

- The mean and variance of sampling distribution of difference between two means are given by,

$$\mu_{\bar{x}_1-\bar{x}_2}/\bar{x}_1-\bar{x}_2=\mu_1-\mu_2$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2}=\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

- Plugging in the mean and standard deviation of sampling distribution into our z distribution formula, we get,

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1^2}+\frac{\sigma_2^2}{n_2}}}$$

Example 3:

In a study among a population with active lifestyle and a population of with sedentary life style, systolic blood pressure SBP was measured among a sample of 40 subjects among those who are active and 50 subjects among those with sedentary life style. Given the standard deviation for active and sedentary populations was 15mmHg, and there is no difference between the population means,

a) What is the probability that the difference between these two sample means was less than 2mmHg?

Answer:

b) What is the probability that the difference between these two sample means was more than 3mmHg?

Answer:

c) What is the probability that the difference between these two sample means was between 2mmHg and 4mmHg?

Answer:

The Difference Between Two Sample Proportions

- Modify formulas for two sample means, and we have

$$\mu_{\bar{p}_1-\bar{p}_2}/\bar{p}_1-\bar{p}_2=p_1-p_2$$

$$\sigma^2_{\bar{p}_1-\bar{p}_2}=\sigma^2_{\bar{p}_1}+\sigma^2_{\bar{p}_2}=\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}$$

$$\sigma^2_{\bar{p}_1-\bar{p}_2}=\sqrt{\sigma^2_{\bar{p}_1}+\sigma^2_{\bar{p}_2}}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

and standard normal distribution, *z*

$$z=\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}$$

Example 4:

In a population of drug addicts, it is known that 65% of them are HIV positive. Suppose random samples of 150 drug addicts in Kelantan and 200 drug addicts in Selangor were taken, and it is assumed that the percentages of HIV positive in their populations are similar, find

a) $P(\hat{p}_K-\hat{p}_S\geq 5\%)$

Answer:

b) $P(\hat{p_K} - \hat{p}_S \leqslant 1\%)$

Answer:

c) $P(2\% \leqslant \hat{p_K} - \hat{p}_S \leqslant 4\%)$

Answer:

Central Limit Theorem

- "if $X_{1,} X_{2,}..., X_n$ are independent random variables each having the same distribution with expected value *μ* and standard deviation *σ*, then the sample mean

$\bar{X}=\frac{1}{n}(X_1+X_2+...+X_n)$ approximately has a normal distribution with expected value *μ* and standard deviation $\frac{\sigma}{\sqrt{n}}$ when *n* is sufficiently large" (Tijms, 2007).

- In other words, for a population of any distributional form with mean *μ* and variance σ^2, the sampling distribution of $\bar{x}$ formed from samples of size *n* from this population will be approximately distributed with mean *μ* and variance σ^2/n when *n* is large (Daniel, 1995).
- In short, for $X_{1,} X_{2,} \ldots, X_n$,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

- From the construction of sampling distribution before, we can see how a sampling distribution of one mean is normally distributed.
- Simulation → http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Topics for Self-study

Using R, simulate the sampling distributions of

- One Sample Proportion
- The Difference Between Two Sample Means
- The Difference Between Two Sample Proportions

References

Daniel, W. W. (1995). *Biostatistics: A foundation for analysis in the health sciences* (6th ed.). USA: John Wiley & Sons.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). USA: Duxbury Press.

Tijms, H. (2007). *Understanding probability: Chances rules in everyday life* (2nd ed.). New York, USA: Cambridge University Press.

Online Resources

MIT's Introduction to Probability and Statistics: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/

Online Statistics Education: An Interactive Multimedia Course of Study: http://onlinestatbook.com/

PennState University's Probability Theory and Mathematical Statistics: https://onlinecourses.science.psu.edu/stat414/