INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 9, 2023

Recap and Today

Recap -
Confidence Interval for the case of Proportion
Sample Size Determination Ideas

Today -
Discussion on Sampling (Literary Digest Example)
Confidence Interval when Population distribution is normal
Hypothesis Testing

Summary of results for 100(1-α)% C.I.

n	σ^2	C.I. Type	Symmetric C.I.
Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ, s is sample standard deviation

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

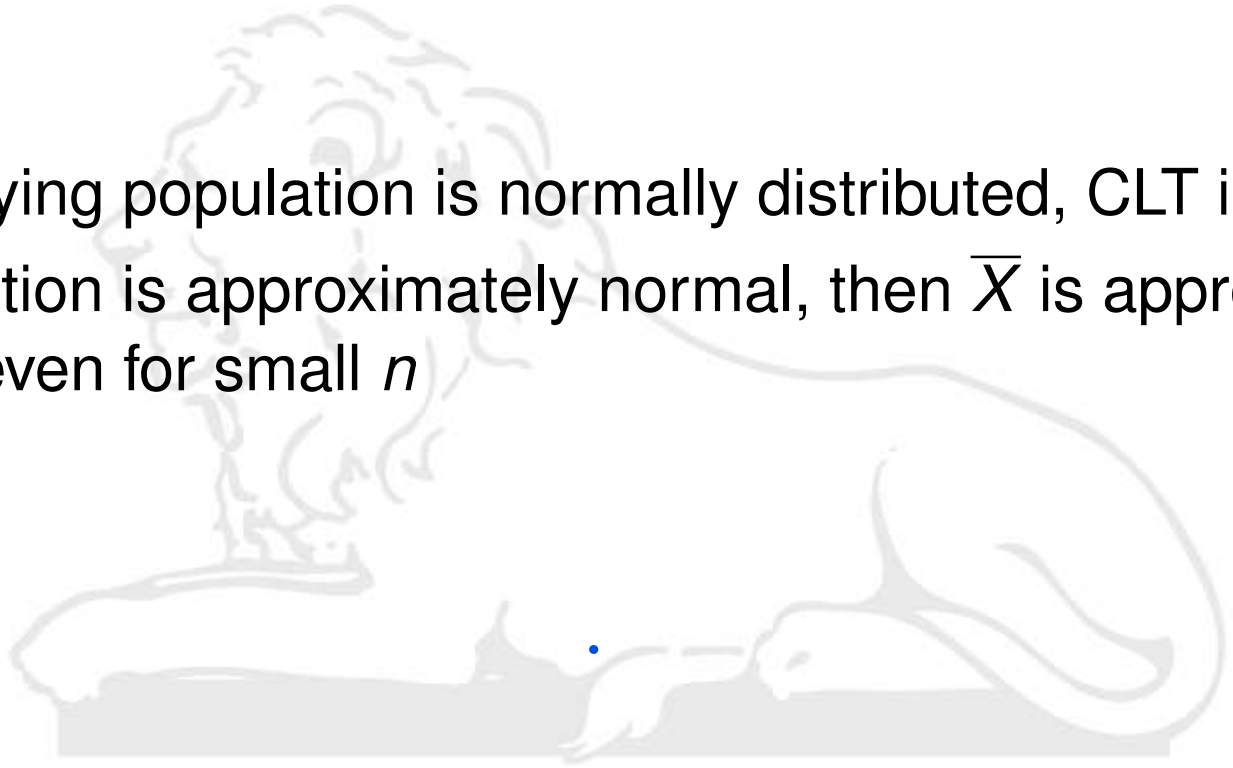
Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

Sample Size Determination

- A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- We want to repeat this study, how many students should you survey so that the 99% confidence interval's width is no more than 2 minutes?

CLT when population is normally distributed

1. If underlying population is normally distributed, CLT is not required
2. If population is approximately normal, then $\overline{X}$ is approximately normal even for small n

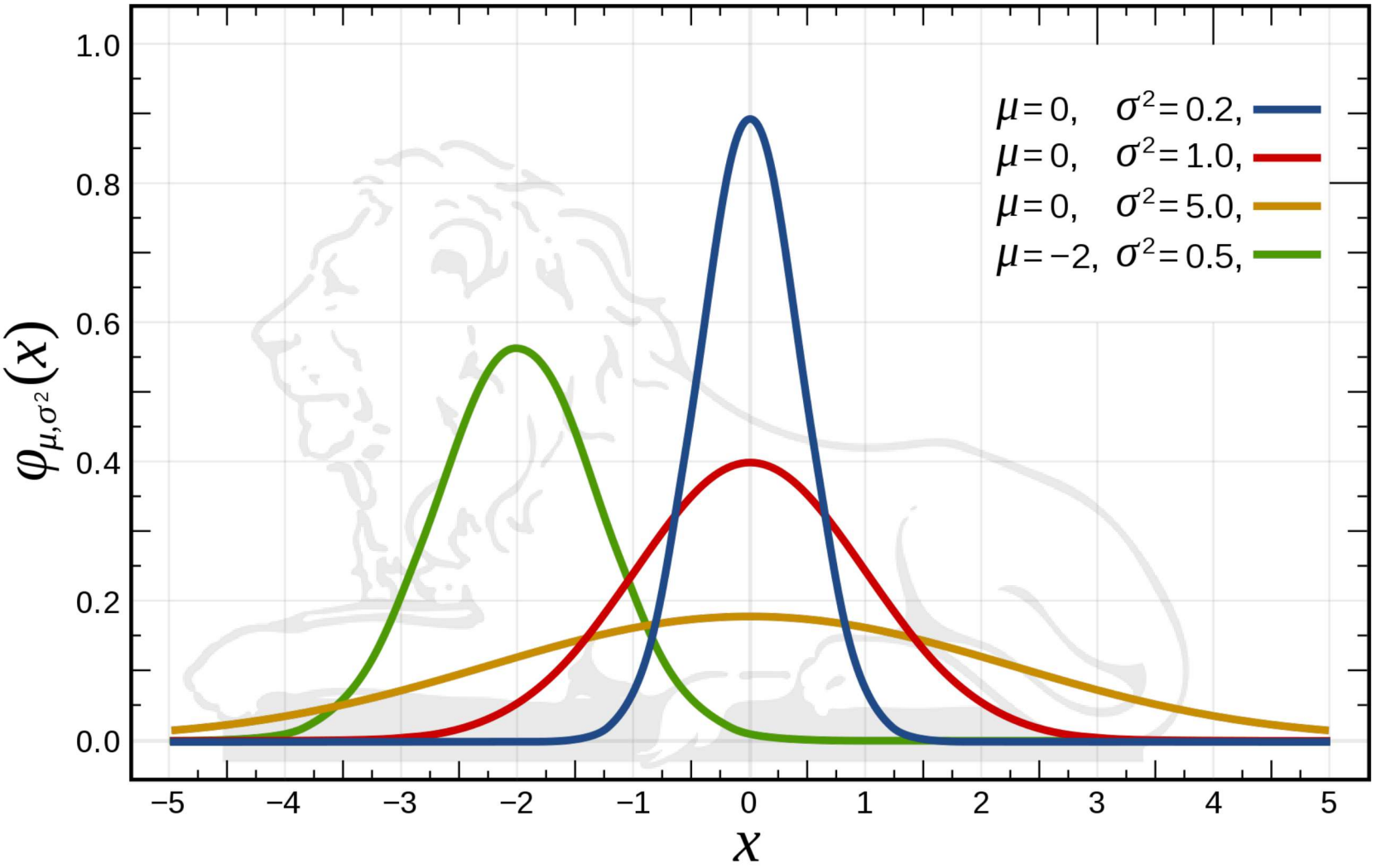


Basics about Random Variable - Recap

- A variable whose value depends on outcome of a random phenomenon
- A random variable is characterized by its distribution
- Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- Similarly, any other algebraic operation of two or more random variables also remain a random variable
- If X and Y are random variables, $X + Y, X - Y, XY, \frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



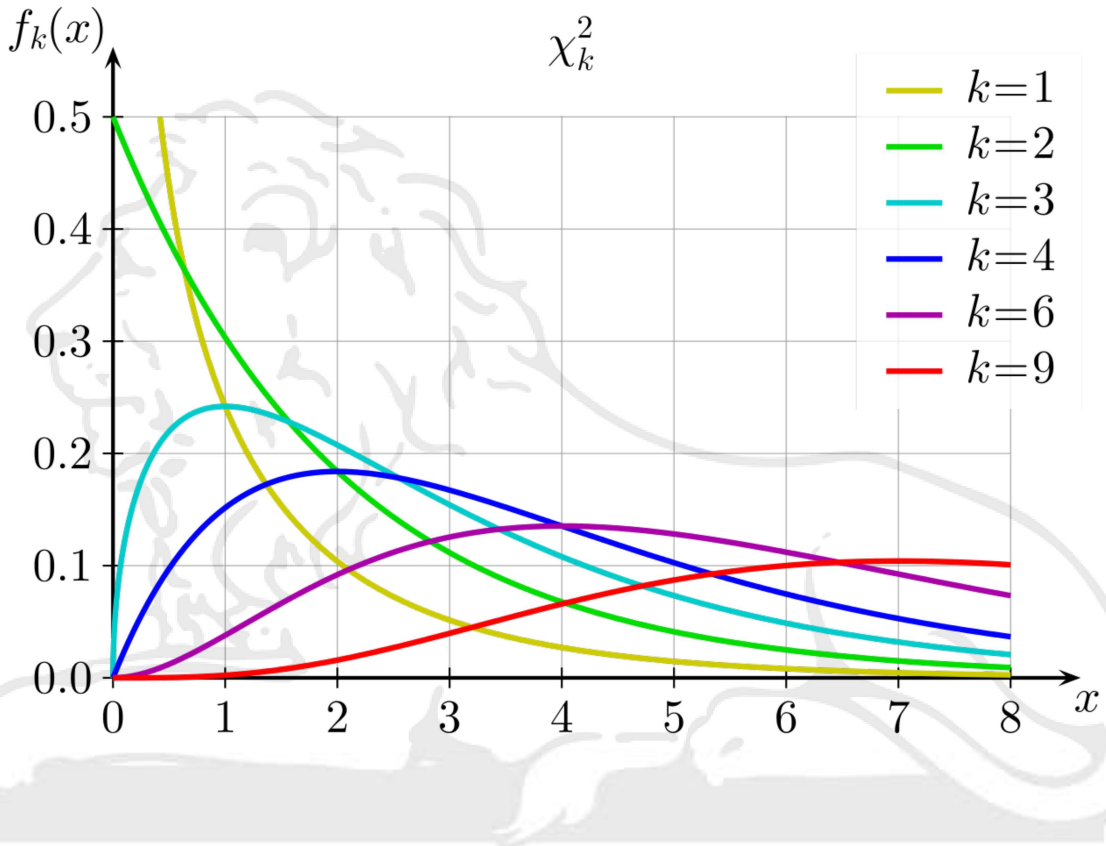
Source - https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution

Chi-square distribution

If $Z_1, Z_2, \ldots, Z_n$ are all independent and normally distributed with mean 0 and variance 1, then $U = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

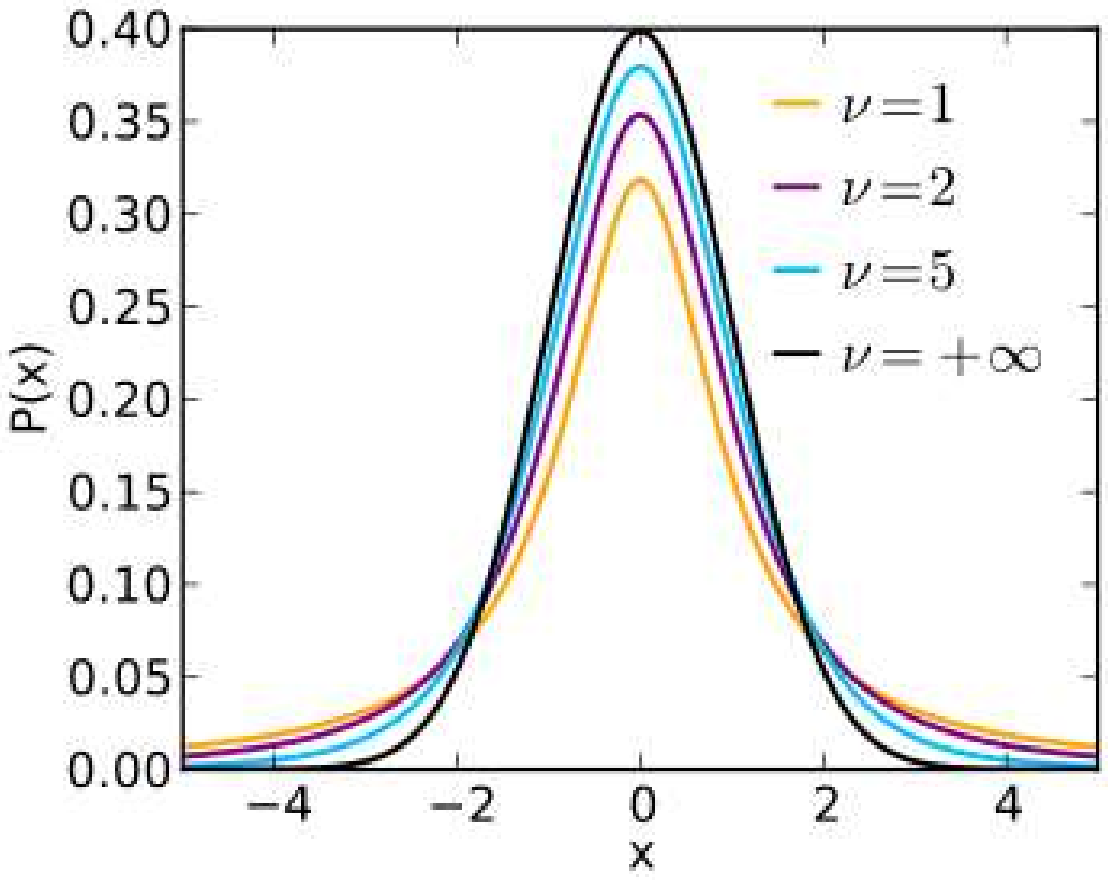
Chi-square distribution



Source - https://en.wikipedia.org/wiki/Chi-squared_distribution

t-distribution

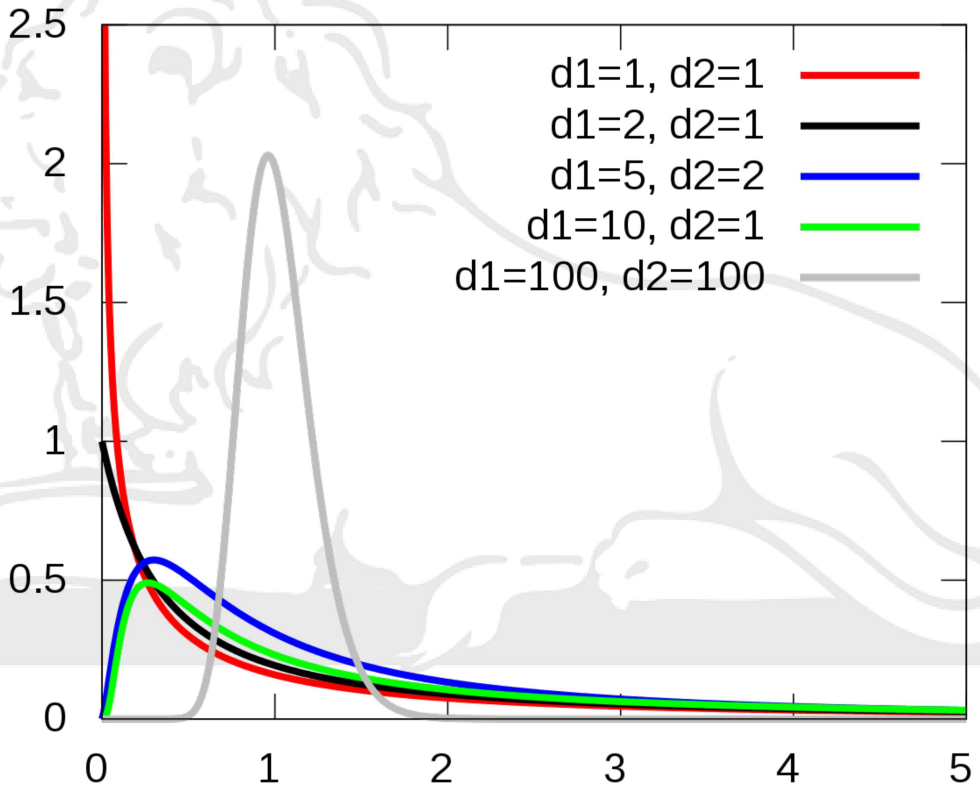
If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom



Source - https://en.wikipedia.org/wiki/File:Student_t_pdf.svg

F distribution

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then the distribution of $\frac{\frac{U}{m}}{\frac{V}{n}}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



Source - https://en.wikipedia.org/wiki/F-distribution

Application of distributions that we just saw

- Does having the idea of population distribution itself a useful information?
- If yes, how do we make use of it?
- Let us concern ourselves with sample mean
- Assume you have the information that the population distribution is normal.
- How do you use this??
- Is CLT required??

Case of normal population

- Sample mean is denoted by $\overline{X}$ and defined as follows
  $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
- If the sampling scheme is WITH REPLACEMENT, sample variance equals
  $$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$
- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$.
- Can you guess the distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence

Case of normal population

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X}-\mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Case of normal population

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X}-\mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Moral of the story - results for 100(1-α)% C.I. for Mean

Population distribution	n	σ^2	C.I. Type	Symmetric C.I.
Any	Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Any	Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$
Normal	Any	known	Exact	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Normal	Any	unknown	Exact	$\left(\overline{X} - \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

Table: C.I. for population mean μ, s is sample standard deviation

Typically, $t_{n-1,\frac{\alpha}{2}}$ is used only for small n, because for large n, $z_{\frac{\alpha}{2}}$ gives a good approximation

Moral of the story - results for 100(1-α)% C.I. for Proportion

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

For case of proportion, it is advised to use these formulae only when apart from n being large, $n\hat{p} \geq 10$ and also $n(1-\hat{p}) \geq 10$