STA 501 Week #5: Sampling Distributions

Cheng Peng

2/18/2021

Contents

1 Introduction **1**

2 Concepts of the sampling distribution **2**

3 Sampling Distribution of Sampling Means **2**
3.1 Working Data Set: Plant Diversity 2
3.2 Normal Population with given mean (μ) and Variance (σ^2) 4
3.3 Normal Population with a given mean (μ) and an **Unknown** Variance (σ^2) 6
3.4 Unspecified Population with a given mean (μ)and an unspecifed variance (σ^2) 7

4 Sampling distribution of sample proportions **9**

5 Summary **11**

1 Introduction

In this section, we study the random behavior of quantities obtained from random samples using the probability distributions introduced in the previous weeks.

Some technical terms:

- **Population** - set of all subjects of interest. For example, if we want the average height of WCU students, then the set of all WCU students is the **population**.
- **Sample** - a subset of subjects of the population. For example, all students in the department of Biology form a subset of all WCU students. In other words, the set of all students in Biology is a **sample**. **However**, this **sample** does not represent the **population** since WCU has many other majors.
- **Random Sample** - a subset of subjects that represent the population. For example, we can use one of the methods introduced in week #2 to collect a subset from the WCU student population - to obtain a random sample.
- **Parameter** - numerical characteristic of the population. For example, the average height of the WCU student population, denoted by μ. Apparently, μ is unknown but fixed.
- **Statistic** - numerical characteristic of population calculated from the random sample. For example, we select a random sample of 150 students from WCU and find the average height, denoted by $\bar{X}$. Apparently, $\bar{X}$ is random since its value is dependent on the random sample.

Since $\bar{X}$ is a random variable, we use probability to characterize the random behavior of $\bar{X}$.

2 Concepts of the sampling distribution

We see from the examples in the previous section that the sample mean is a random variable. In fact, all population parameters evaluated at a random sample taken from the population are random variables. To characterize the behavior, we need to use probability distributions.

- **A sampling distribution** is the distribution of a **sample statistic** such as sample mean, sample variance, sampling coefficient of variation, sample correlation coefficient, etc.

A population has many different numerical characteristics that require different probability distributions to characterize them. In this course, we focus on the mean and proportion and some coefficients of regression that affect the mean and proportion. In the next few modules, we introduce procedures for constructing confidence intervals and testing hypotheses inferences for population means and proportions.

In the next sections, we introduce sampling distributions of sample means and proportions under different assumptions.

3 Sampling Distribution of Sampling Means

Inferential statistics is all about making the inference about population parameters by using the information of individual subjects. The estimated population parameters could be used to make a prediction at the individual level.

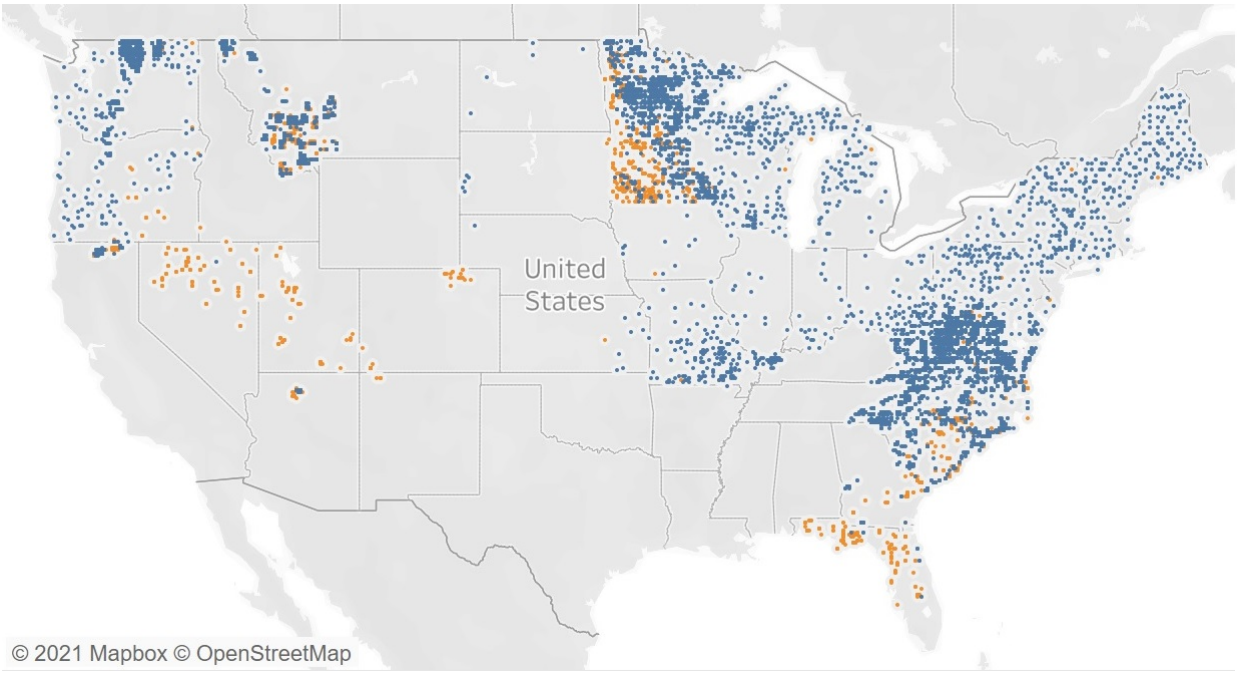
The information in the estimate is dependent on the amount of information in the sample and the population as well. In the next few sections, we introduce the sampling distribution of sample means under different assumptions.

3.1 Working Data Set: Plant Diversity

This data set includes the geographic location (lat/lon) for 15,136 plots, as well as the herbaceous species richness, climate, soil pH, and other variables related to the plots.

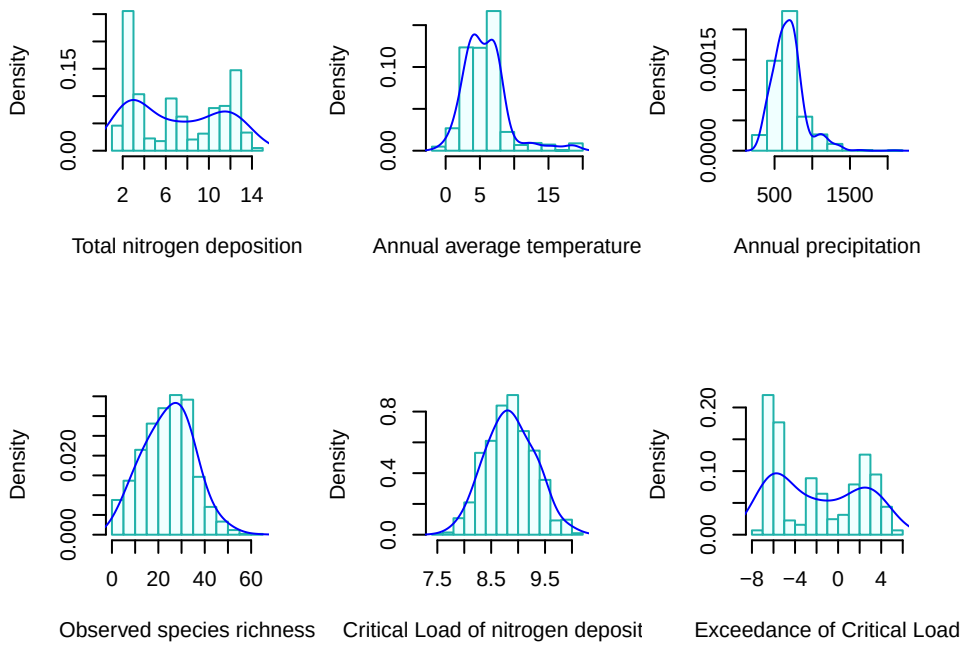
This data set is associated with the following publication: Simkin, S., C. Clark , W. Bowman, E. Allen, J. Belnap, and L. Pardo. The conditional vulnerability of plant diversity to atmospheric nitrogen deposition across the United States. PNAS (PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES). National Academy of Sciences, WASHINGTON, DC, USA, 113(15): 4086-4091, (2016).

```
include_graphics("w05-site-map.jpg")
```



We choose a special subset that contains data associated with **Perennial graminoid vegetation**. That has 1152 records. The distributions of 6 numerical variables are given in the following histograms.

```
plant.diversity = read.csv("https://stat501.s3.amazonaws.com/w05-plant-diversity.csv", header = TRUE)
```



We can see from the above histograms that the **Critical Load of nitrogen deposition** is close to the normal distribution. Other variables are **not** normally distributed. In the rest of this note, I will assume the above data sets to be "populations" and take random samples from appropriate populations listed above in

different examples. Some of the following examples will be based on the above populations. You may want to use the distributional information (the shapes of histograms) to choose appropriate sampling distributions of sample statistics based on the sample taken from one of the above populations.

3.2 Normal Population with given mean (μ) and Variance (σ^2)

Result #1: Assume that X is a normal random variable with an unknown mean μ and the known variance σ^2. That is,

$$X \to N(\mu, \sigma^2)$$

Let random sample $\{x_1, x_2, \cdots, x_n\} \to N(\mu, \sigma_0^2)$, where population mean μ is unknown and variance σ_0^2. Then the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n}}{n}$$

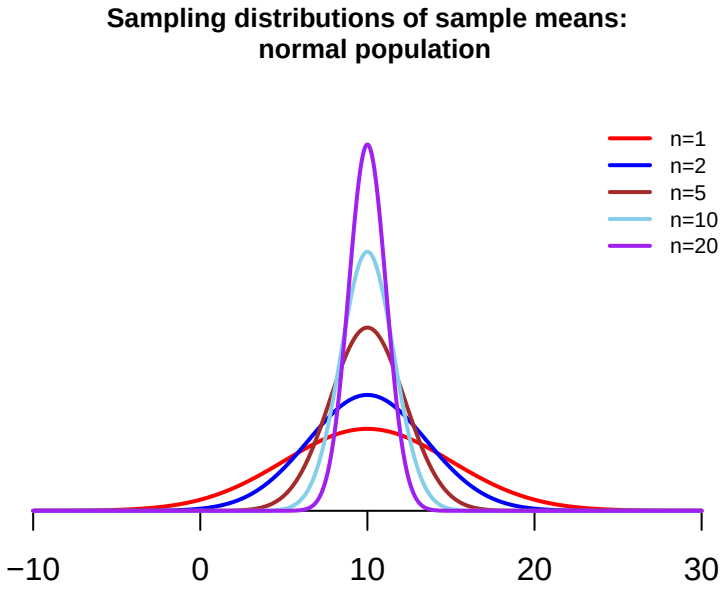
is normally distributed with mean μ and variance σ^2/n. In other words,

$$\bar{X} \to N(\mu, \sigma_0^2/n)$$

Remarks: The simple comparison of X and $\bar{X}$.

- The means of the distributions of X and $\bar{X}$ are equal.
- The variances of the distributions of X and $\bar{X}$ are **not** equal. In fact, the variance of $\bar{X}$ is less than the variance of X.
- As the sample size increases, the variance of $\bar{X}$ decreases since the denominator of the variance of $\bar{X}$ contains the sample size n.
- Since the R functions **pnorm()** and **qnorm()** require the standard deviation as an argument.

The following figure shows the variance of the sample means with different sample sizes.



The above figure shows that, as sample size increases, the variance of $\bar{X}$ decreases (the density curve becomes skinnier as sample size increases).

Example 1. We use **Critical Load (CL) of nitrogen deposition** in the **plant diversity** data set. Assume that **Critical Load (CL) of nitrogen deposition** follows a normal distribution with a mean of $\mu = 10$ and known variance $\sigma^2 = 0.2$. Use this information to answer the following questions.

1. If a random of 20 Critical Load (CL) of nitrogen deposition sample values are taken from the population, what is the probability that the mean critical load nitrogen deposition ($\bar{X}$) is great than 10.1 kg N/ha/yr?
2. What is the level of critical load (CL) of nitrogen deposition that is higher than the 95% mean level of critical load (CL) of nitrogen deposition with the same sample size 20?

Solution: Since critical load (CL) of nitrogen deposition is a normal population, therefore, the sampling distribution of $\bar{X} \to N(10, .2/20)$. The solutions to problems are based on this sampling distribution.

1. $P(\bar{X} > 10.05)$ is equal to the right tail area of $N(10, (\sqrt{0.2/20})^2$. Note that, by default, R function **pnorm()** only gives the left-tail area. The desired probability is equal to $1 - left.tail.area$. The R code is given in the following.

```
1-pnorm(10.1, mean = 10, sd = sqrt(0.2/20))
```

```
## [1] 0.1586553
```

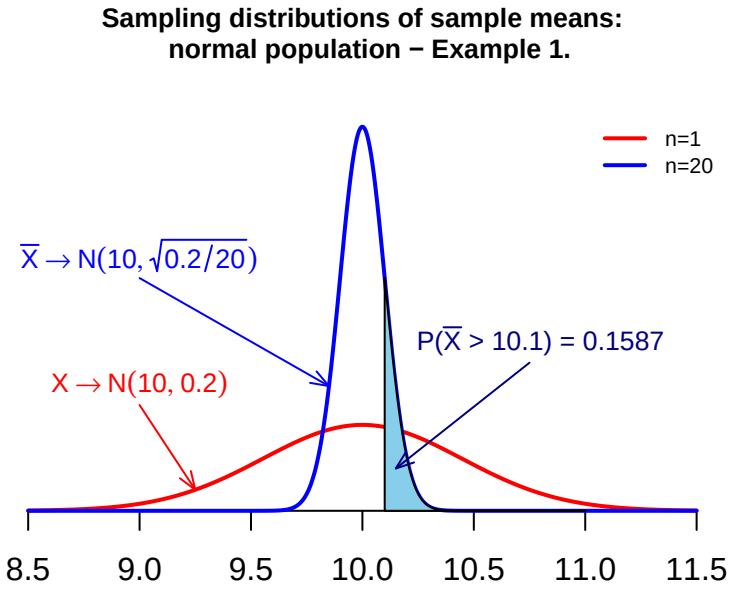
Therefore, $P(\bar{X} > 11.5) = 0.1587$. The tail probability is labeled in the following figure.

2. the desired cut-off if actually the 95% quantile of the distribution of $\bar{X}$ which can be found by **qnorm()**.

```
qnorm(0.95, mean = 10, sd = sqrt(0.2/20))
```

```
## [1] 10.16449
```

Therefore the 95% quantile is 10.16449.



3.3 Normal Population with a given mean (μ) and an Unknown Variance (σ^2)

If the population variance is unknown, then we have to use sample variance to characterize the distribution $\bar{X}$. Unlike in the case of the normal population with known variance in which we can specify the sampling distribution $\bar{X}$ directly, we cannot but have the following result based on the standardized statistic.

Result #2: Let random sample $\{X_1, X_2, \cdots, X_n\} \to N(\mu, \sigma^2)$ and $\bar{X}$ and s^2 be the sample mean and sample variance, respectively. then we have

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

Where t_{n-1} is a t-distribution with $n-1$ degrees of freedom. The two basic types of questions associated with $t-$distribution were discussed in the previous module.

Example 2. Assume that the sample of 11 levels of level of critical load (CL) of nitrogen deposition $\{8.67, 9.38, 9.27, 8.56, 8.72, 8.83, 9.72, 8.36, 8.76, 8.91, 9.49\}$ is randomly selected from all sites in the study. Its **sample standard deviation** is 0.430. What is the percent of the sample means based on the same sample size 11 will be bigger than 10.2? Note that the population mean is $\mu = 10$.

Solution The question to answer is about the sample mean $\bar{X}$ with size $n = 11$. Since the population variance is unknown, $\bar{X}$ is not a normal distribution.

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \to t_{n-1}$$

Therefore,

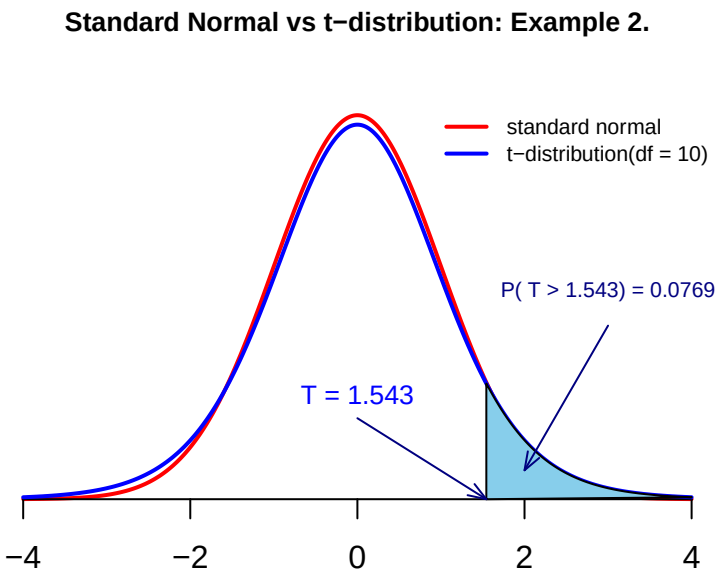
$$P(\bar{X} > 10.2) = P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} > \frac{10.2 - 10}{0.43/\sqrt{11}}\right) = P(T > 1.543)$$

As we know, $P(T > 1.543)$ is the right-tail area of $t_{11-1} = t_{10}$. The desired probability is 1− left-tailed area. The following R code finds the above probability.

```
1- pt(1.543, df = 11-1)
```

```
## [1] 0.07693013
```

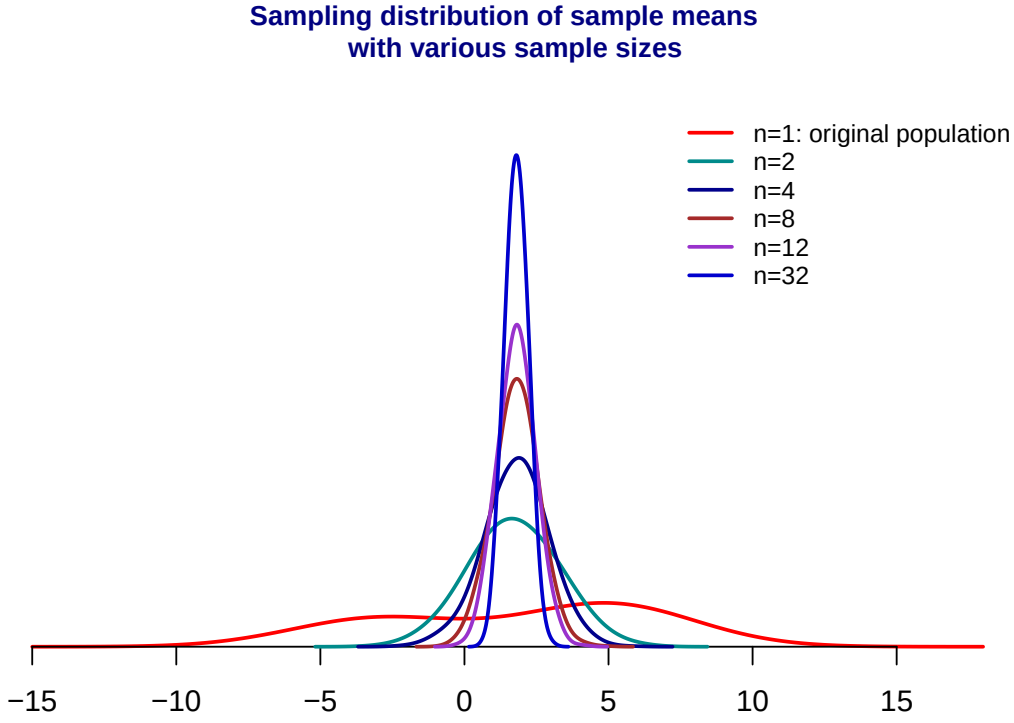
Therefore, $P(T > 1.543) = 0.0769$.



3.4 Unspecified Population with a given mean (μ)and an unspecifed variance (σ^2)

In previous two cases, we assume the population to be normal. Quite frequently we need to deal with a population without a specified distribution real-world applications. The means the population distribution is unknown. As usual, we assume the population mean is known. The population variance is either given or unknown. If we still want to use the sampling distribution of sample means to make inferences about the population mean, what result(s) we can use to characterize the sampling distribution of the sample mean?

In this section, we discuss this types sampling distribution of sample means under certain conditions. Before we present the result, we perform a simulations to show the patterns of the sampling distributions of sample means using various sample sizes. We simulate a non-normal distribution with two peaks and take samples from that population with different sample sizes, and use these sample means to estimate the corresponding density curves.



The figure shows that, as the sample size increases, the sampling distribution of the sampling mean approaches a normal distribution regardless of the distribution of the original population. The following theorem explicitly specifies the sampling distribution.

Results #3: **Central Limit Theorem:** Let $X \to (\mu, \sigma^2)$ (**caution**: the population is not necessarily to be a normal population.). Let $\bar{X}$ be the sample mean with size n. If n is large, the sampling $\bar{X}$ is **approximately** normally distributed with a mean $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$. That is,

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

Comments: The following are comments related to the normal approximations.

- the baseline population is not specified in the theorem. The distribution could be discrete or continuous.
- when the sample size is large, the sample mean is approximately normally distributed. The question is how large is called "large"? As a convention, we call the sample size is large if $n > 30$.

Example 3. The blood cholesterol levels of a population of workers have mean 202 and standard deviation 14. If a sample of 36 workers is selected, approximate the probability that the sample mean of their blood cholesterol levels will lie between 198 and 206.

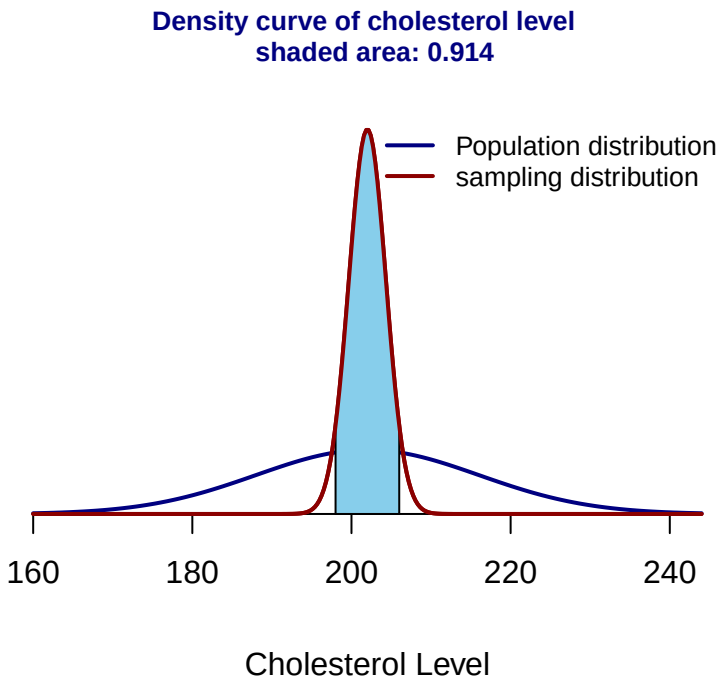
Solution Let X be the blood cholesterol levels of a population of workers. Then $X \to (\mu = 202, \sigma = 14)$. Since $n = 36 > 30$, using the C.L.T, we have $\bar{X} \to N(202, 14/\sqrt{36})$. Therefore,

$$P(198 < \bar{X} < 206) = P(\bar{X} < 206) - P(\bar{X} < 198) = 0.9135237$$

```r
pnorm(206, mean = 202, sd = 14/sqrt(36)) - pnorm(198, mean = 202, sd = 14/sqrt(36))
```

```
## [1] 0.9135237
```

The above probability is the area of the shaded area in the following figure.



4 Sampling distribution of sample proportions

For a population of binary data that only takes on exactly two possible values such as "success" vs "failure", "diseased" vs "disease-free", etc., its distribution is uniquely determined by the proportion of one of the categories.

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a random sample taken from a population of binary data taking on only two possible values "success" or "failure". That is, x_i = "success" or "failure", for $i = 1, 2, \cdots, n$. If we are interested in the proportion of "success" of the population, we can do the following numerical coding on the sample: 1 = "success" and 0 = "failure". With this numerical coding, we calculate the sample mean

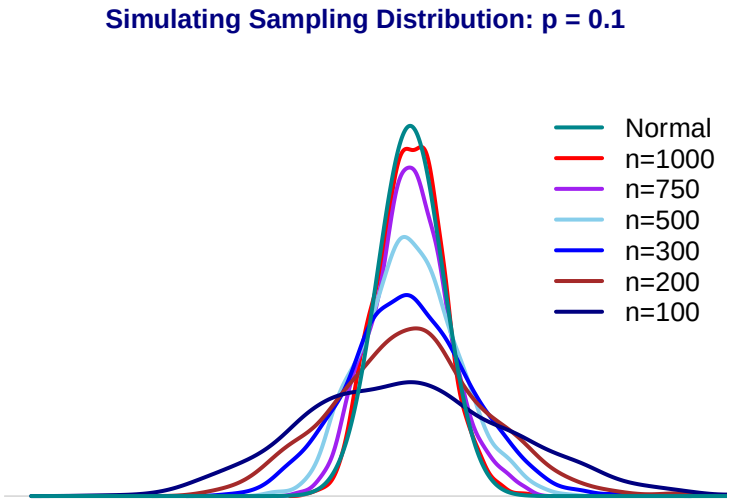
$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\#1s}{n} = \frac{\#succeses}{n} = proportion.of.successes.$$

Therefore, the sample proportion is actually a **sample mean**. Therefore, according to the central limit theorem, $\hat{p}$ is approximately normally distributed under certain conditions. This idea is formalized in the following result.

Result #4. Let p be the proportion of one of the categories, say "success" and $\hat{p}$ the sample proportion based on a random sample with size n. If $np > 5$ and $n(1-p) > 5$, then

$$\hat{p} \to N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

In some applications, the standard deviation $\sqrt{p(1-p)/n}$ is estimated by $\sqrt{\hat{p}(1-\hat{p})/n}$.



The above simulated sampling distributions with samples taken from a binary population with true $p = 10\%$ use different sample sizes. We see that as the sample size increases, the sampling distribution approaches the normal distribution.

Example 4 The frequency of color blindness (dyschromatopsia) in the Caucasian American male population is estimated to be about 8%. We take a random sample of size 125 from this population. What is the probability that more than 9% of the 125 subjects have color blindness?

Solution Since $np = 125 \times 0.08 = 10 > 5$, $n(1-0.08) = 115 > 5$, the C.L.T can be used to approximate the sampling distribution of the sample proportion to the normal distribution as outlined in the **result #3**. That is,

$$\hat{p} \to N\left(0.08, \sqrt{\frac{0.08(1-0.08)}{125}}\right) = N\left(0.08, 0.0243\right).$$

The probability we want to find is $P(\hat{p} > 0.09) = 1 - P(\hat{p} < 0.09) = 0.34$

```
1 - pnorm(0.09, mean = 0.08, sd = 0.0243)
```

```
## [1] 0.3403447
```

Example 5 Suppose 60% of seniors who get flu shots remain healthy, independent from one person to the next. If we selected a random sample from the complex of 100, what is the probability that the sample proportion will be greater than 50%?

Solution: We are given that $p = 0.6$. Since $np = 200 \times 0.6 = 120 > 5$ and $n(1-p) = 200 \times 0.4 = 80 > 5$. Using **result #3**, we have

$$\hat{p} \to N\left(0.6, \sqrt{\frac{0.6(1-0.6)}{200}}\right) = N(0.6, 0.0346).$$

Therefore, $P(\hat{p} > 0.5) = 1 - P(\hat{p} < 0.5) = 0.2859$

```
1-pnorm(0.6, 0.0346)
```

```
## [1] 0.2859009
```

5 Summary

In this module, we introduced the sampling distribution of sample means and proportions under various assumptions. Let random sample $\{x_1, x_2, \cdots, x_n\}$ be taken from a population with sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Th sampling distribution of $\bar{X}(\hat{p})$ under various conditions is summarized in the following .

- **Sampling distribution of sampling means** $\bar{X}$
  - Population is normal with known variance (σ_0^2).

$$\bar{X} \to N\left(\mu, \frac{\sigma_0^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \to N(0,1).$$

  - Population is normal with unknown variance (σ^2).

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1).$$

    However, if the sample variance s^2 is used,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \to t_{n-1}.$$

  - Population is unspecified: **if the sample size is large**,

$$\bar{X} \to N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1) \quad \text{and} \quad \frac{\bar{X} - \mu}{s/\sqrt{n}} \to N(0,1).$$

where s is the sample standard deviation.

- **Sampling distribution of proportions** $(\hat{p})$: if $np > 5$ and $n(1-p) > 5$, then

$$\hat{p} \to N\left(p, \frac{p(1-p)}{n}\right) \Rightarrow \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \to N(0,1).$$

Note also that

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \to N(0,1).$$