Sampling Distributions

Lecture 5

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

1 Outline

- Importance of Sampling Distributions
- Distributions of the Sample Mean
- Distributions of the Difference Between Two Sample Means
- Distributions of the Sample Proportion
- Distributions of the Difference Between Two Proportions

2 Importance of Sampling Distributions

2.1 Activity

Let us use the "Sampling Words" applet through this link. Suppose we are interested in calculating the average word length of different samples of 10 words from Beyonce's Crazy in Love.

- Did we get the same average length for all samples?

2.2 Sampling Distributions

Repeated samples will yield different values for the statistics.

> **Note**
>
> We have to view statistics as the sample mean $\bar{x}$ and sample proportion $\hat{p}$ as random variables. The probability distribution of these statistics are called sampling distributions. These distributions enable us to:
>
> - Answer probability questions about sample statistics
> - Provide the necessary theory for some statistical inference tests.

2.3 Sampling Distributions

The sampling distribution of a statistic is the distribution of *all possible values* that can be assumed by some statistic computed from samples of the same size from the population.

> **Note**
>
> We are usually interested in knowing the functional form (refer to distributions discussed in Chapter 4), mean, and variance.

3 Distributions of the Sample Mean

3.1 Sampling from Gaussian-distributed populations

For a sample with size n that comes from a Gaussian-distributed population with mean μ and variance σ^2, i.e. the samples $X_1, X_2, ..., X_n$ are all independent and identically distributed such that $X_i \sim N(\mu, \sigma^2)$, the sample mean can be defined as:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

The probability distribution of the sample mean, also known as the sampling distribution of the sample mean can be expressed as:

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

3.2 Implications

> **Important**
>
> The mean of the sampling distribution, $\mu_{\bar{X}}$ is equal to the population mean, while the variance, $\sigma^2_{\bar{X}}$ is reduced by a factor of the sample size n.
> The standard deviation of the sampling distribution can be calculated by taking the square root of the variance of the distribution, $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$. This is referred to as the **standard error** of the sample mean.

3.3 Standardization

The sample mean can also be standardized based on the properties of the sampling distribution. The sample mean $\bar{X}$ can be transformed into a random variable Z such that $Z \sim N(0, 1)$.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

In addition,

$$P(\bar{X} \leq \bar{x}) = P(Z \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}) = P(Z \leq z)$$

3.4 Sampling from Non-Gaussian Populations

The sampling distribution previously derived could also apply to samples from non-Gaussian populations under certain conditions. These conditions are provided by the **central limit theorem**.

> **Central Limit Theorem (CLT)**
>
> Given a population of any non-Gaussian functional form with a mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$ computed from samples of size n from this population, will have mean μ and variance σ^2/n and will be approximately Gaussian distributed **when the sample size is large**.

> **Tip**
>
> The rule of thumb for sample means is that a sample size of 30 is satisfactory for the central limit theorem, but this maybe too small for skewed distributions. The Gaussian

approximation provided by the CLT becomes better as the sample size increases.

3.5 Example

Suppose it is known that in a certain large human population cranial length is approximately Gaussian distributed with a mean of 185.6 mm and a standard deviation of 12.7 mm.

3.5.1 Question

What is the probability that a random sample of size 10 from this population will have a mean greater than 190?

3.5.2 Answer

We can use the sampling distribution because we know the sample is from a Gaussian population. The mean of the sampling distribution, $\mu_{\bar{X}}$, should be equal to the population mean. Hence, $\mu_{\bar{X}} = 185.6$.

The standard error, $\sigma_{\bar{X}}$, can be calculated using the following formula: $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 12.7/\sqrt{10} = 4.0161$

Using the knowledge that $\bar{X} \sim N(185.6, 12.7/\sqrt{10})$, we can use `pnorm` to calculate the probability that the sample mean is greater than 190.

```
pnorm(190,mean=185.6,sd=12.7/sqrt(10),lower.tail=F)
```

```
[1] 0.1366286
```

```
z <- (190-185.6)/(12.7/sqrt(10))
pnorm(z,0,1,lower.tail=F)
```

```
[1] 0.1366286
```

3.6 Example 2

Suppose the hourly number of customer arrivals at a hospital ED has a Poisson distribution with rate parameter $\lambda = 6$. Every hour for 48 hours, the number of customer arrivals is counted and recorded.

3.6.1 Question

Use the CLT to approximate the probability that the average number of arrivals is between 5 and 8.

3.6.2 Answer

The data is non-Gaussian, but we will assume that the CLT holds. Recall that for a Poisson distribution, the mean is equal to the variance. Specifically, $\mu = 6$, $\sigma^2 = 6$, $n = 48$. Hence, the mean and the standard error can be expressed as:

$$\mu_{\bar{X}} = 6; \sigma_{\bar{X}} = \sqrt{6/48}$$

The probability that the average number of arrivals in the 48-hour period is between 5 and 8 is:

```
pnorm(8,6,sqrt(6/48)) - pnorm(5,6,sqrt(6/48))
```

```
[1] 0.9976611
```

3.7 Exercise

The daily average screen time of elementary school students is assumed to follow a non-Gaussian distribution with 2.6 hours with a standard deviation of 5.3.

3.7.1 Question

Use the central limit theorem to calculate the probability that a sample of 250 elementary school students will yield an average screen time between 1.5 hours and 2 hours?

3.7.2 Answer

The mean of the sampling distribution is 2.6 hours and the standard error of the mean is $5.3/\sqrt{250} = 0.3352014$.

```
pnorm(2,2.6,5.3/sqrt(250)) - pnorm(1.5,2.6,5.3/sqrt(250))
```

```
[1] 0.03621341
```

3.8 Finite Population Correction

The resulting sampling distribution assumes that sampling was done with replacement (those who were sampled can be sampled again) or sampling was done from an infinite population.

> **Important**
>
> When dealing with smaller populations, it is possible for these assumptions to be broken. We need to account for the type of sampling and the finiteness of the population.

> **Important**
>
> This correction is often ignored when the sample size is less than 5% of the population.

3.9 Finite Population Correction Formula

> **Tip**
>
> When sampling is without replacement from a finite population, the sampling distribution of $\bar{x}$ will have mean μ and variance
>
> $$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$
>
> The standard error can be expressed as the square root of this variance.
>
> $$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

> **Important**
>
> It also follows that the standardization with the finite population correction can be expressed as
>
> $$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}\sqrt{\left(\frac{N-n}{N-1}\right)}}$$

3.10 Example

Suppose 30 prototypes of a newly invented suture thread have a mean tensile strength of 140 psi and standard deviation of 10.5 psi. Assuming the tensile strengths are normally distributed,

3.10.1 Question

What is the probability that a sample of 7 prototypes will have a mean tensile strength lower than 137 psi?

- Assuming sampling with replacement + infinite populations
- Using the finite population correction

3.10.2 Answer

- Infinite population: standard error $= \sqrt{\sigma^2/n} = \sqrt{10.5^2/7} = 3.968627$

```
mu <- 140
se <- sqrt(10.5^2/7)

pnorm(137,mu,se)
```

```
[1] 0.2248459
```

- Finite Population Correction

standard error $= \sqrt{\sigma^2/n}\sqrt{\left(\frac{N-n}{N-1}\right)} = \sqrt{10.5^2/7}\sqrt{\frac{30-7}{30-1}} = 7.7701351$

```
mu <- 140
se <- sqrt(10.5^2/7)*sqrt((30-7)/(30-1))

pnorm(137,mu,se)
```

```
[1] 0.1979905
```

3.11 Exercise

Suppose **300** prototypes of a newly invented suture thread have a mean tensile strength of 140 psi and standard deviation of 10.5 psi. Assuming the tensile strengths are normally distributed,

3.11.1 Question

What is the probability that a sample of 7 prototypes will have a mean tensile strength lower than 137 psi?

- Assuming sampling with replacement + infinite populations
- Using the finite population correction

3.11.2 Answer

- Infinite population: standard error $= \sqrt{\sigma^2/n} = \sqrt{10.5^2/7} = 3.968627$

```
mu <- 140
se <- sqrt(10.5^2/7)

pnorm(137,mu,se)
```

```
[1] 0.2248459
```

- Finite Population Correction

standard error $= \sqrt{\sigma^2/n}\sqrt{\left(\frac{N-n}{N-1}\right)} = \sqrt{10.5^2/7}\sqrt{\frac{300-7}{300-1}} = 3.9286062$

```
mu <- 140
se <- sqrt(10.5^2/7)*sqrt((300-7)/(300-1))

pnorm(137,mu,se)
```

```
[1] 0.222544
```

4 Distribution of the Difference Between Two Sample Means

4.1 Sample Mean vs. Difference Between Two Sample Means

Consider Populations A and B with the following possible measurements:

A: {0,1,2}; B:{1,2,3}.

If we take a sample with size 2, we can have the following scenario:

Sample A: {0,2}; Sample B: {1,2}

> **Warning**
>
> The sample mean of Sample A is 1, and the mean of Sample B is 1.5. However, the difference between the sample means is -0.5, which is not a possible value for the sample means of A and B.

4.2 Sampling from Gaussian-distributed populations

When sampling from two Gaussian-distributed populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, the distribution of the difference between sample means $\bar{X}_1 - \bar{X}_2$ is a Gaussian distribution with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Mathematically,

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

4.3 Sampling from non Gaussian-distributed populations

When sampling from two Gaussian-distributed populations with finite means and variances, we can impose the Central Limit Theorem such that the distribution of the difference between sample means $\bar{X}_1 - \bar{X}_2$ can be approximated by the following Gaussian distribution.

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

4.4 Standardization

Similar to the sample mean, the difference of the sample means can also be standardized.

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

In addition,

$$P((\bar{X}_1 - \bar{X}_2) \leq (\bar{x}_1 - \bar{x}_2)) = P\left(Z \leq \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

4.5 Example

Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes.

4.5.1 Question

If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between client type one and client type two by 20 or more minutes?

4.5.2 Answer

```
# Calculate the mean and variance of the sampling distribution first.
diff_means <- 45-30                                                    ①
diff_var <- 15^2/35 + 20^2/40                                          ②

pnorm(20,diff_means,sqrt(diff_var),lower.tail=F)

# OR An alternative solution

z <- (20 - diff_means)/sqrt(diff_var)                                  ③

pnorm(z,0,1,lower.tail=F)
```

① mu_1-mu_2
② sigma_1^2/n1 + sigma_2^2/n2
③ Standardization method

```
[1] 0.1086783
[1] 0.1086783
```

4.6 Exercise

Suppose the age of two student organizations were normally distributed. Organization A has an average age of 20.5 and a standard deviation of 3.7, while Organization B has an average of 19.7 and a standard deviation of 4.5.

4.6.1 Question

If 50 students were sampled from each organization, what is the probability that the sample from Organization A is older than Organization B by a value between one and two years?

4.6.2 Answer

```
diff_means <- 20.5-19.7
diff_var <- 4.5^2/50 + 3.7^2/50

pnorm(2,diff_means,sqrt(diff_var))-pnorm(1,diff_means,sqrt(diff_var))
```

```
[1] 0.3314723
```

```
# OR An alternative solution

z1 <- (2 - diff_means)/sqrt(diff_var)
z2 <- (1 - diff_means)/sqrt(diff_var)

pnorm(z1,0,1) - pnorm(z2,0,1)
```

```
[1] 0.3314723
```

5 Distribution of the Sample Proportion

5.1 Sampling Distribution for the Sample Proportion

Oftentimes, we are interested in the distribution of the sample proportion when dealing with frequency data.

> **Note**
>
> Unlike the means where we measure values, we create the sampling distribution for sample proportions by taking all possible samples of a given size and calculating the proportion of the variable of interest for each sample. The variable is usually dichotomous (yes/no, 0/1)

5.2 Sampling Distribution for the Sample Proportion

Suppose the true population/long-run proportion is π. Invoking the Central Limit Theorem, the sampling distribution of the sample proportion, $\hat{P}$ when the sample size is large follows a Gaussian distribution with mean π and variance $\frac{\pi(1-\pi)}{n}$.

Formally,

$$\hat{P} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

It follows that the standard error of the sample proportion is $\sigma_{\hat{p}} = \sqrt{\frac{\pi(1-\pi)}{n}}$.

5.3 Standardization

The sample proportion can also be standardized based on the properties of the sampling distribution. The sample proportion $\hat{P}$ can be transformed into a random variable Z such that $Z \sim N(0,1)$.

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

In addition,

$$P(\hat{P} \leq \hat{p}) = P\left(Z \leq \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\right) = P(Z \leq z)$$

5.4 Example

In a town with 23,673 adult residents, 63.7% of the adult residents were against the new proposed zoning maps.

5.4.1 Questions

What is the probability that in a random sample of 200, less than 125 will be against the new proposed zoning maps?

5.4.2 Answer

The population proportion is $\pi = 0.637 = \mu$, the sample size $n = 200$. The variance can be calculated using the formula $\pi(1-\pi)/n = (0.637)(1-0.637)/200 = 0.0011562$.

```
phat <- 125/200
pi <- 0.637
variance <- 0.637 * (1-0.637)/200

pnorm(phat,pi,sqrt(variance))
```

```
[1] 0.3620751
```

5.5 Exercise

According to the CDC, 1 in 31 children aged 8 years has been identified with Autism Spectrum Disorder (ASD) in the United States.

5.5.1 Question

What is the probability that more than 50 out of 1000 randomly sampled children aged 8 years old were identified with ASD?

5.5.2 Answer

The population proportion is $\pi = 1/31 = \mu$, the sample size $n = 1000$. The variance can be calculated using the formula $\pi(1-\pi)/n = (1/31)(1-1/31)/1000 = 3.1217482 \times 10^{-5}$.

```
phat <- 50/1000
pi <- 1/31
variance <- (1/31) * (1-1/31) / 1000
1-pnorm(phat,pi,sqrt(variance))
```

```
[1] 0.0007480821
```

```
pnorm(phat,pi,sqrt(variance),lower.tail=F)
```

```
[1] 0.0007480821
```

5.6 Continuity Correction

> **Warning**
>
> Recall that proportions often come from discrete distributions such as the binomial distribution. However, we're using the Gaussian distribution, which is continuous, to approximate a discrete distribution.

> **Tip**
>
> We can employ a continuity correction to account for the adjustment from discrete to continuous distributions. The correction is implemented during the standardization of the sample proportion. Given the sample proportion $\hat{p}$, sample size n, and the population proportion π,
>
> $$z = \begin{cases} \frac{\hat{p}+(0.5/n)-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}, & \hat{p} < \pi \\ \frac{\hat{p}-(0.5/n)-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}, & \hat{p} > \pi \end{cases}$$

> **Important**
>
> The effect of the continuity correction on the resulting standardized z-score is lower for high sample sizes.

5.7 Example

In a town with 23,673 adult residents, 63.7% of the adult residents were against the new proposed zoning maps.

5.7.1 Questions

Use the continuity correction to calculate the probability that in a random sample of 20, less than 12 will be against the new proposed zoning maps?

5.7.2 Answer

The population proportion is $\pi = 0.637 = \mu$, the sample size $n = 20$. The variance can be calculated using the formula $\pi(1-\pi)/n = (0.637)(1-0.637)/20 = 0.0115616$. Note that $\hat{p} = 12/20 = 0.6$ is less than $\pi = 0.637$.

```r
phat <- 12/20
pi <- 0.637
variance <- 0.637 * (1-0.637)/20
sampsize <- 20

z <- (phat + (0.5/sampsize) -pi)/sqrt(variance)
pnorm(z,0,1)
```

```
[1] 0.4555694
```

Without continuity correction:

```r
phat <- 12/20
pi <- 0.637
variance <- 0.637 * (1-0.637)/20

pnorm(phat,pi,sqrt(variance))
```

```
[1] 0.3653829
```

5.8 Exercise

According to the CDC, 1 in 31 children aged 8 years has been identified with Autism Spectrum Disorder (ASD) in the United States.

5.8.1 Question

Use the continuity correction to calculate the probability that more than 50 out of 1000 randomly sampled children aged 8 years old were identified with ASD?

5.8.2 Answer

The population proportion is $\pi = 1/31 = \mu$, the sample size $n = 1000$. The variance can be calculated using the formula $\pi(1-\pi)/n = (1/31)(1-1/31)/1000 = 3.1217482 \times 10^{-5}$. Note that $\hat{p} > \pi$.

```r
phat <- 50/1000
pi <- 1/31
variance <- (1/31) * (1-1/31) / 1000
sampsize<- 1000

z <- (phat - (0.5/sampsize) -pi)/sqrt(variance)
pnorm(z,0,1,lower.tail=F)
```

```
[1] 0.001014558
```

```r
1-pnorm(z,0,1)
```

```
[1] 0.001014558
```

6 Distribution of the Difference Between Two Sample Proportions

6.1 Difference Between Two Sample Proportions

Similar to the difference between two sample means, some values of the difference in proportions cannot be found in the sampling distribution of single proportions from two different groups.

> **!** Important
>
> For instance, it is possible to get a negative difference between proportions, but not a negative proportion.

6.2 Sampling Distribution of the Difference Between Two Sample Proportions

Suppose the respective true population/long-run proportion for groups 1 and 2 are π_1 and π_2 and the respective samples from these groups are n_1 and n_2. Invoking the Central Limit Theorem, the sampling distribution of the difference in sample proportions, $\hat{P}_1 - \hat{P}_2$ when the sample size is large follows a Gaussian distribution with mean $\pi_1 - \pi_2$ and variance $\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$.

Formally,

$$\hat{P}_1 - \hat{P}_2 \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

It follows that the standard error of the sample proportion is $\sigma_{\hat{P}_1-\hat{P}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$.

6.3 Standardization

Similar to the single proportion case, the difference of two proportions can also be standardized to $Z \sim N(0,1)$.

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

In addition,

$$P((\hat{P}_1 - \hat{P}_2) \leq \hat{p}_1 - \hat{p}_2) = P\left(Z \leq \frac{\hat{p}_1 - \hat{p}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}\right) = P(Z \leq z)$$

6.4 Example

Researchers investigated food insecurity in the urban and rural areas of their country. It was previously reported by the country's census that the proportion of adults who experienced food insecurity in the past year was 33% in rural areas and 23% in urban areas. The researchers sampled 450 adults from urban areas and 300 adults from rural areas.

6.4.1 Question

What is the probability that the samples will yield a result such that the proportion of food insecure adults from the rural sample was **at least** 5% higher than that of the urban sample?

6.4.2 Answer

Remember that $\pi_{rural} - \pi_{urban} = .10, \pi_{rural} = 0.33, \pi_{urban} = 0.23$. We are interested in $P((\hat{P}_1 - \hat{P}_2) \geq 0.05)$.

```r
pi_1 <- 0.33
n_1 <- 300
pi_2 <- 0.23
n_2 <- 450
variance <- (pi_1 *(1-pi_1))/n_1 + (pi_2 *(1-pi_2))/n_2

pnorm(0.05,pi_1-pi_2,sqrt(variance),lower.tail = F)
```

```
[1] 0.9314985
```

OR through standardization

```r
z <- (0.05-(pi_1-pi_2))/sqrt(variance)

pnorm(z,0,1,lower.tail=F)
```

```
[1] 0.9314985
```

6.5 Exercise

The prevalence of the Influenza virus in State A was reported to be 30%, while the prevalence of Influenza in State B is 24%.

6.5.1 Question

If we randomly sample 100 constituents from each state, what is the probability that there will be a higher proportion of individuals infected with Influenza in State B compared to State A?

6.5.2 Answer

We are interested in $P(\hat{P}_A - \hat{P}_B < 0)$, where $\hat{P}_A$ is the sample proportion of infected individuals in State A and $\hat{P}_B$ is the sample proportion for State B. It follows that $\pi_A = 0.30$, $\pi_B = 0.24$, and $n_1 = n_2 = 100$.

```r
pi_1 <- 0.30
n_1 <- 100
pi_2 <- 0.24
n_2 <- 100
variance <- (pi_1 *(1-pi_1))/n_1 + (pi_2 *(1-pi_2))/n_2

pnorm(0,pi_1-pi_2,sqrt(variance))
```

```
[1] 0.1690752
```

OR through standardization

```r
z <- (0-(pi_1-pi_2))/sqrt(variance)

pnorm(z,0,1)
```

```
[1] 0.1690752
```