Topic 2: Sampling distribution of the difference of 2 means

Background for inference of location parameter in location/scale families

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Sampling distributions

(Blue text = links!)

Review:

1. How does inference relate to analytics?
2. What is a sampling distribution?
3. Sampling distribution of the mean: in previous lecture, we applied the Central Limit Theorem to the one-paramter location problem.

Sampling distribution of the difference in means

Setup

- Previous case study:
  - statistical inference about a single population mean μ.
- What if you have a comparative experiment in which two methods are compared?
  - Two 'methods'? $\rightarrow$ manufacturing methods; statistical/machine learnig methods; etc.

$$\mu_1 \text{ VS. } \mu_2$$

... compare... how? $\mu_1 ? = \mu_2 ?$

Sampling distribution of the difference in means

Setup

- Previous case study:
  - statistical inference about a single population mean μ.
- What if you have a comparative experiment in which two methods are compared?
  - Two 'methods'? $\rightarrow$ manufacturing methods; statistical/machine learnig methods; etc.

$$\mu_1 \text{ VS. } \mu_2$$

... compare... how? $\mu_1 ? = \mu_2 ?$ Does $\mu_1 - \mu_2 = 0$?

Sampling distribution of the difference in means

Setup

- Previous case study:
  - statistical inference about a single population mean μ.
- What if you have a comparative experiment in which two methods are compared?
  - Two 'methods'? $\rightarrow$ manufacturing methods; statistical/machine learnig methods; etc.

$$\mu_1 \text{ VS. } \mu_2$$

  ... compare... how? $\mu_1 ? = \mu_2 ?$ Does $\mu_1 - \mu_2 = 0$?
- Compare the means of the two populations (one representing each method), denoted by μ_1 and μ_2, or sometimes μ_A and μ_B.

Underlying question:

What is the sampling distribution of $\mu_1 - \mu_2$??

Theoretical distribution

Suppose that two independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1 and σ_2, respectively. Then:

Theorem

The sampling distribution of the differences of means, $\overline{X}_1 - \overline{X}_2$, is approximately normally distributed with:

$$\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2, \quad \sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Hence,

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

is approximatley a standard normal random variable.

Exploring the meaning of the theorem: R

Suppose we just take one sample each, X and Y, of size 100 from two populations like this:

```
x <- rnorm(n=100,mean=25,sd=10)
Xbar <- mean(x)
y <- rnorm(n=100,mean=35,sd=10)
Ybar <- mean(y)
Xbar-Ybar
```

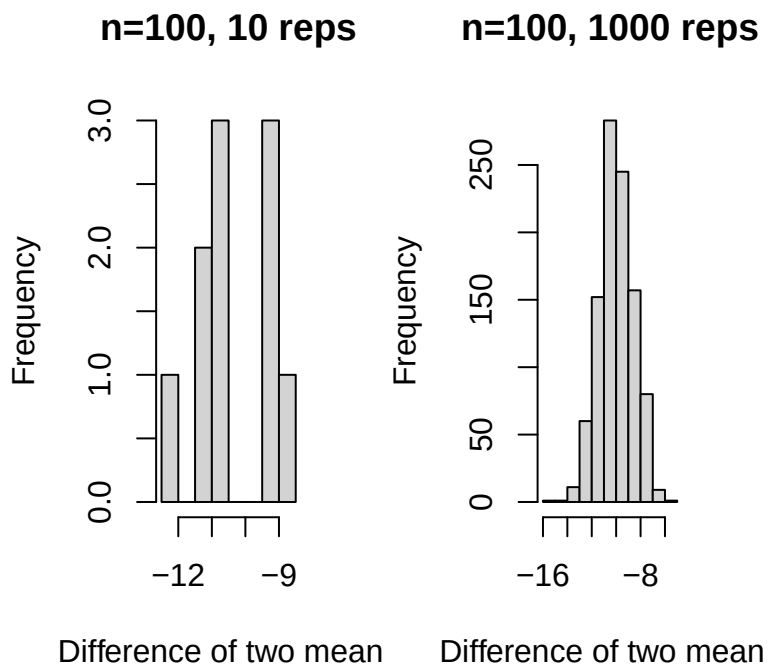
```
[1] -9.852698
```

Of course this is just *one* value of the staistic $\overline{X} - \overline{Y}$. . . . Now, repeat!

Exploring the meaning of the theorem: R (cont'd)

Values of $\overline{X} - \overline{Y}$ in repeated sampling, $n_1 = n_2 = n$:

```
sample.size=100
diff.in.means <- replicate(10, mean(rnorm(n=sample.size,mean=2
```



Aha!

Are you able to infer anything from the histograms?

Exploring the meaning of the theorem: Python

Python code

```
from scipy.stats import norm
import numpy as np
x = norm.rvs(loc=25,scale=10,size=100)
Xbar = np.mean(x)
y = norm.rvs(loc=35,scale=10,size=100)
Ybar = np.mean(y)
Xbar-Ybar
```

```
-9.636742241173565
```

Case Study 2: paint drying time

Problem Two independent experiments are run in which two different types of paint are compared. 18 specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Question:

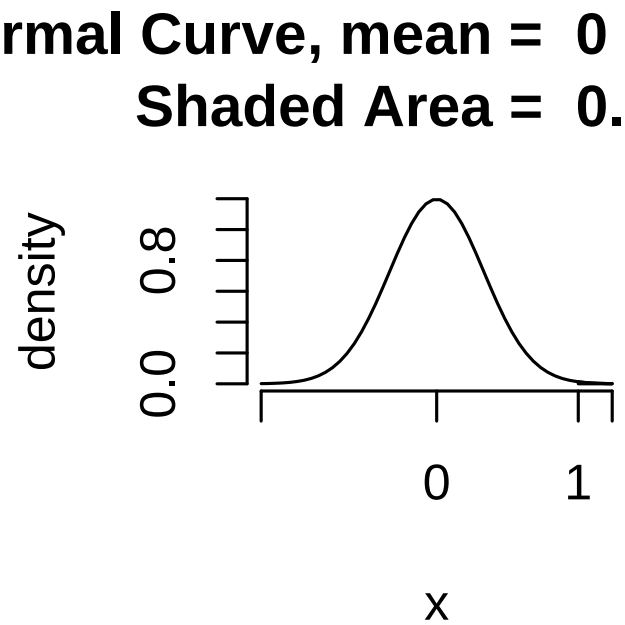
Assuming that the mean drying time is equal for the two types of paint, find $P(\overline{X}_A - \overline{X}_B > 1)$, where $\overline{X}_A$ and $\overline{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Case study 2: Solution - code

The sampling distribution of $\overline{X}_A - \overline{X}_B$ is ...

Case study 2: Solution - code

The sampling distribution of $\overline{X}_A - \overline{X}_B$ is . . . approximately normal, mean $\mu_{\overline{X}_A - \overline{X}_B} = \mu_A - \mu_B = 0$ and variance $\sigma^2_{\overline{X}_A - \overline{X}_B} = 1/18 + 1/18 = 1/9$.




```
[1] 0.001349898
```

Case study 2: Solution - compute

The probability that we compute is given by:

$$P(\overline{X}_A - \overline{X}_B > 1) = P\left(\frac{\overline{X}_A - \overline{X}_B - 0}{\sqrt{1/9}} \geq \frac{1-0}{\sqrt{1/9}}\right) = P(Z > 3) = 0.0013.$$

What do you conclude?

$\rightarrow$ Discussion: what we learned from this case study.

What is next?

There are other results on sampling distributions of other statistics, and we will cover them as needed. For now, look out for the use of these theorems for:

- constructing interval estimators of unknown population parameters, μ or $\mu_1 - \mu_2$;
- designing hypothesis tests for the same parameters;
- a hands-on case study and how these are executed in R/Python.

Note: both languages have libraries that implement these tests and estimators for you; you don't have to do it from scratch. But you should know the result that is being used, because computers don't check if theorems are applicable- they apply them when you instruct them to do so!

License

This document is created for ITMD/ITMS/STAT 514, Spring 2021, at Illinois Tech.