The Center of Applied Data Science

python™

Statistical Data Analysis

Day 2.1

Content outline

In [ ]:

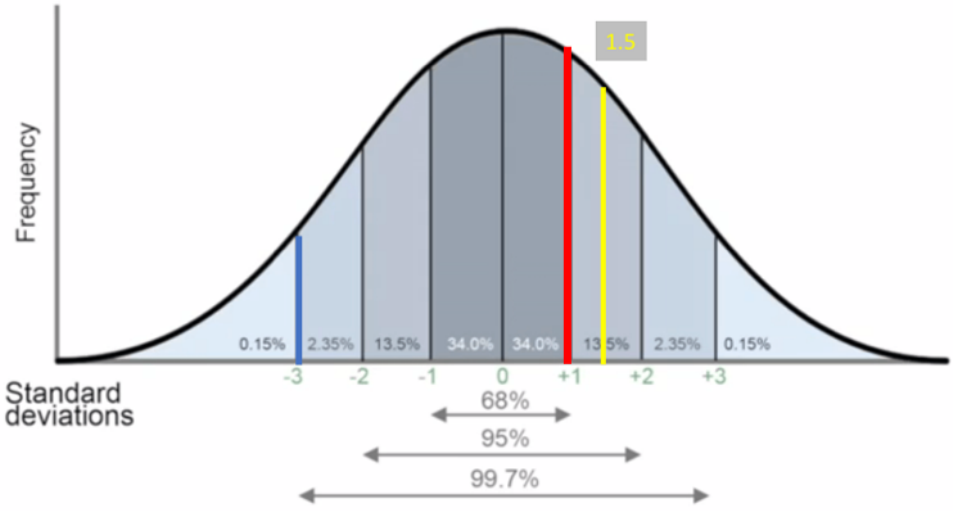
```
import matplotlib.pyplot as plt
import numpy as np
```

1. Invoking the normal distribution

In Python, samples from the normal distribution can be generated by the function `np.random.normal`. The PDF of the normal distribution can be calculated by the function `scipy.stats.norm.pdf`, the CDF by the function `scipy.stats.norm.cdf`, and the inverse of the CDF (the quantile function) by the function `scipy.stats.norm.ppf`.

Example

What is the probability that the random variable x with a standard normal distribution $Normal(\mu = 0, \sigma = 1)$ lies within the following intervals of values? Answer the following questions if we randomly choose x **from a population** with this probability distribution.



a\. $P(x > -3) = ?$ b\. $P(x \leq -3) = ?$ c\. $P(-3 \leq x \leq 1) = ?$ d\. $P(x \geq 1) = ?$ e\. $P(x \leq 1.5) = ?$ f\. Find m such that $P(x \leq m) = 0.975$

In [ ]:

```
from scipy.stats import norm

# a- P(x > -3) = 1-P(x < -3) = 1-0.0015 = 0.9985
print('P(x > -3) = 1 - P(x < -3) = 1 - 0.0015 = ', round(1-0.0
015,3))
print('\nP(x > -3) = 1 - P(x < -3) = 1 - norm.cdf(-3, 0, 1) =
', 1-round(norm.cdf(-3, 0, 1),3))
print('\n****************************************\n')
# b- P(x <= -3) = 0.0015
print('P(x <= -3) = ',0.0015)
print('\nP(x <= -3) = norm.cdf(-3, 0, 1) = ',round(norm.cdf(-3
, 0, 1),4))
print('\n****************************************\n')
# c- P(-3 <= x <= 1) = (2.35+13.5+34+34)/100 = 0.838
print('P(-3 <= x <= 1) = (2.35 + 13.5 + 34 + 34)/100 =',round(
(2.35 + 13.5 + 34 + 34)/100,3))
print('\nP(-3 <= x <= 1) = P(x <= 1) - P(x <= -3) = norm.cdf(1
,0,1) - norm.cdf(-3,0,1) =',
                  round(norm.cdf(1,0,1) - norm.cdf(-3,0,1), 3)
)
print('\n****************************************\n')
# d- P( x >= 1) = (13.5+2.35+0.15)/100 = 0.16
print('P( x >= 1) = (13.5 + 2.35 + 0.15)/100 = ', (13.5 + 2.35
+ 0.15)/100)
print('\nP( x >= 1) = 1 - P( x < 1) = 1 - norm.cdf(1,0,1) = ',
round(1 - norm.cdf(1,0,1), 3))
print('\n****************************************\n')
# e- P( x <= 1.5)
print('P( x <= 1.5) = norm.cdf(1.5, 0, 1) = ',round(norm.cdf(1
.5, 0, 1), 4))
print('\n****************************************\n')
#f- find m such that P( x <= m) = 0.975
print('P( x <= m) = 0.975, so m = norm.ppf(.975, 0, 1), 3) = '
,round(norm.ppf(.975, 0, 1), 3))
```

Exercise

For a random variable distributed as a $\text{Normal}(\mu = 1, \sigma = 2)$, use Python to solve for the following: P(x > 2.5) = ? P(x <= 3.25) = ? P(-1 <= x <= 1) = ? P(x >= 1) = ? P(x <= 1) = ?

In [ ]:

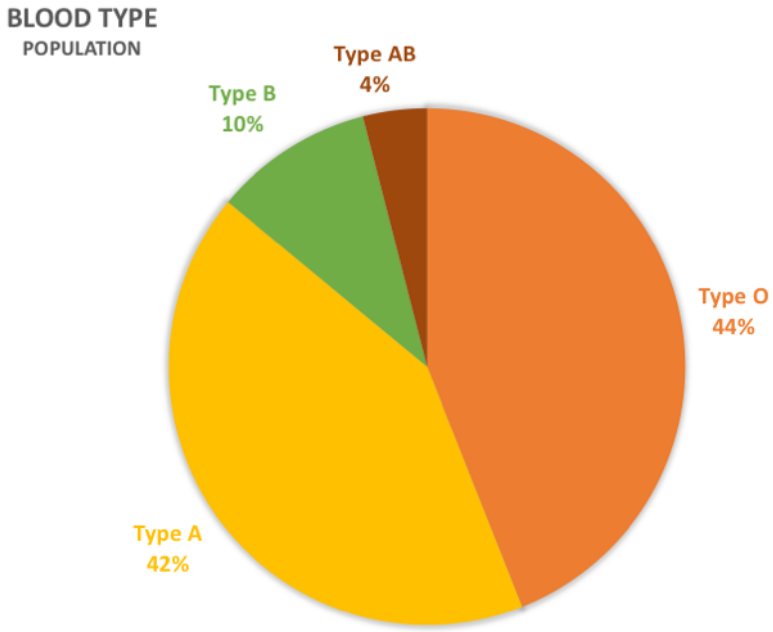
2. Population vs. Sample

In **descriptive statistics**, we try to describe the statistics of samples that we have measured. In **inferential statistics**, we make use of the samples we have measured to draw conclusions for the population from which we draw these samples. We cannot calculate exact population statistics unless we know every single element of the population, and it would cost too much to collect such an extensive sample. Thus, there will be some **uncertainty** if we try to perform inferential statistics, for example in estimating the mean of a population.

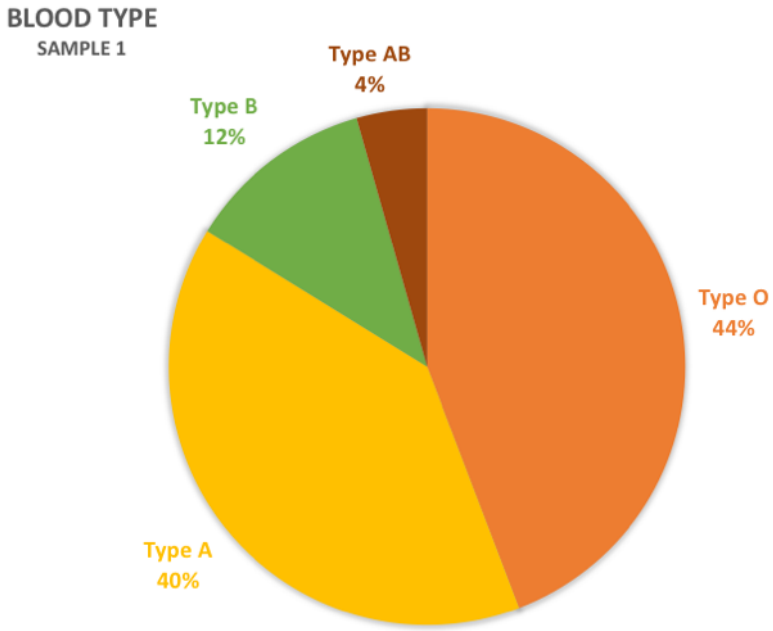
		Population (parameter)	Sample (Statistic)
Categorical Variable	**Proportion**	**P=**population proportion	$\hat{p}$=sample proportion
Numerical Variable	**Mean**	μ=population mean	$\overline{x}$=sample mean
	Standard Deviation	σ=population standard deviation	**s=**sample standard deviation

Example

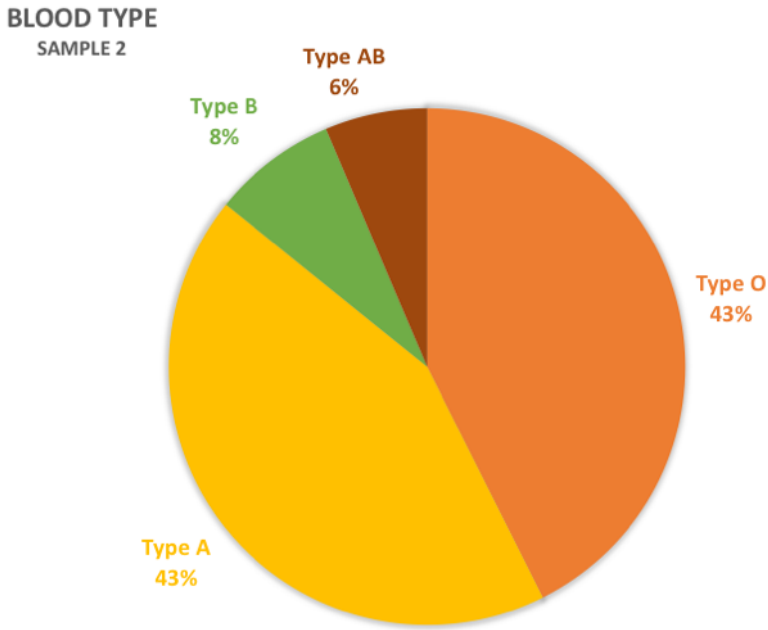
We presented the distribution of blood types in the entire U.S. population as follows:



Assume we have randomly chosen 500 people in the U.S. as sample 1 and recorded their blood type. The following pie chart shows the result. Note that the proportion (percentage) of the people in each group (blood type) is slightly different.



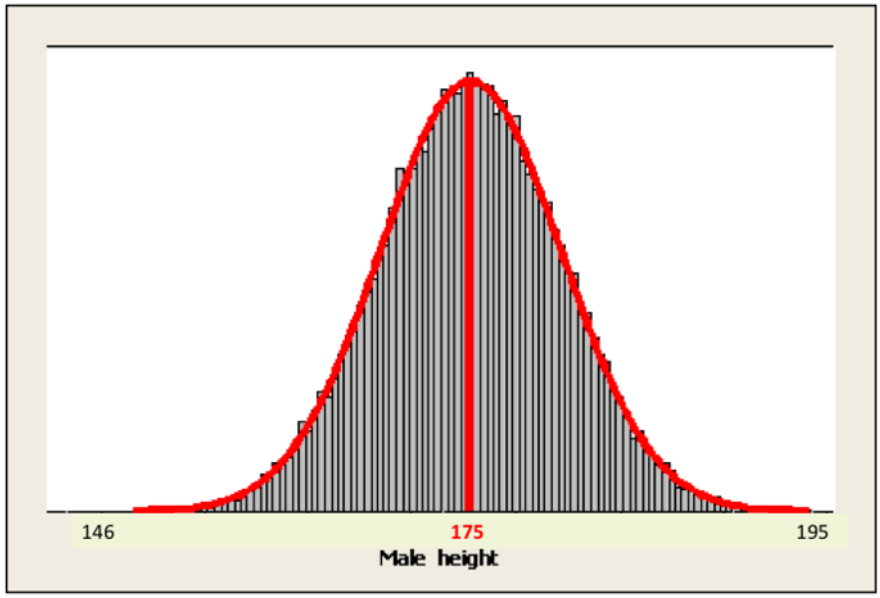
Sample 2 again includes the blood type of another 500 people in the U.S. and the results are summarized in the following pie chart:



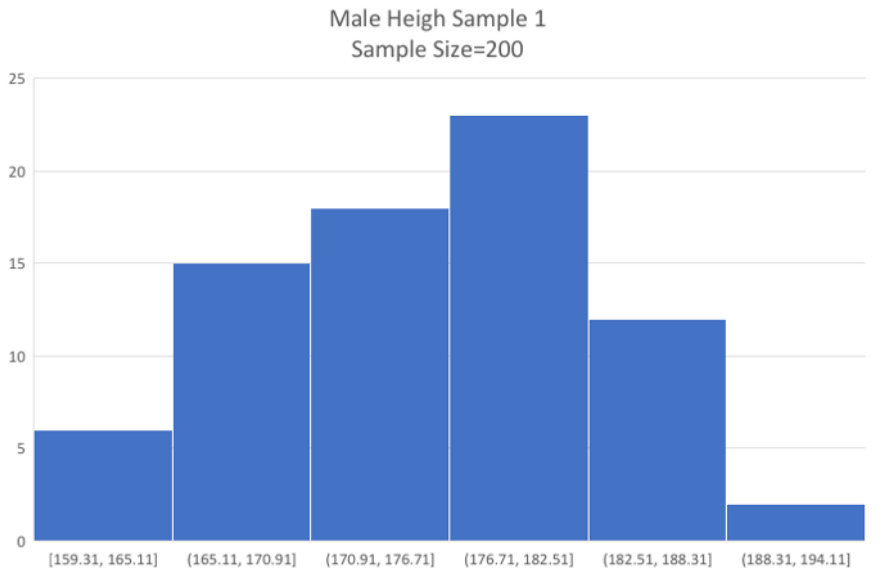
Again the sample result is slightly different from the population proportions. We call this **sampling variability**.

Example

This example shows sampling variability for a numeric variable. Heights among the population of all adult males follow a normal distribution with a mean $\mu = 175cm$ and standard deviation $\sigma = 7.11cm$. Here is a probability display of this population distribution:



A sample of 200 males were collected and their height has been recorded. The results are summarized in the following histogram:



The mean of the sample is $\bar{x} = 174.85$ and its standard deviation is $s = 7.24$.

In this section our goal is to draw conclusions about population parameters based on sample statistics. We first focus on the behavior of sample proportion with respect to the population proportion, where the variable is categorical. Then, we will explore the behavior of the sample mean relative to population mean where the variable is numerical.

3. Behavior of Sample Proportion

Assume approximately 60% of all part-time college students in the U.S. are female. (In other words, the population proportion of females among part-time college students is p = 0.6). What would you expect to see in terms of the behavior of a sample proportion of females $\hat{p}$ if random samples of size 100 were taken from the population of all part-time college students?

As we saw before, due to sampling variability, the sample proportion in random samples will take varying numerical values. Therefore, the sample proportion is a random variable. So, it will have a probability distribution function. This distribution function has center (mean), spread (standard deviation), and shape.

Intuitively, we would expect the following:

- **Center (mean)**: Sample proportions should be **centered around population proportion**. In some cases, the sample proportion might be slightly below or above the population proportion. In other words, the mean of the distribution of $\hat{p}$ should be p.
- **Spread**: We expect most of the sample proportions stay **close to the population proportion**. We also expect that we observe **less sampling variation in larger samples**. Therefore, for larger samples we will have less spread in sample proportion.
- **Shape**: Considering the intuition above, we expect a **bell-shape distribution** for the sample proportion $\hat{p}$. We would expect $\hat{p}$ to follow some normal distribution.

These videos check our intuition:

- https://www.youtube.com/watch?reload=9&v=2bIC4EmejkQ (https://www.youtube.com/watch?reload=9&v=2bIC4EmejkQ)
- https://www.youtube.com/watch?v=tUvXeJ3A3_s (https://www.youtube.com/watch?v=tUvXeJ3A3_s)

We can summarize all of the above by the following: $\hat{p}$ has a normal distribution with a mean of $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where **p is the population proportion** (and as long as np and $n(1-p)$ are at least 10).

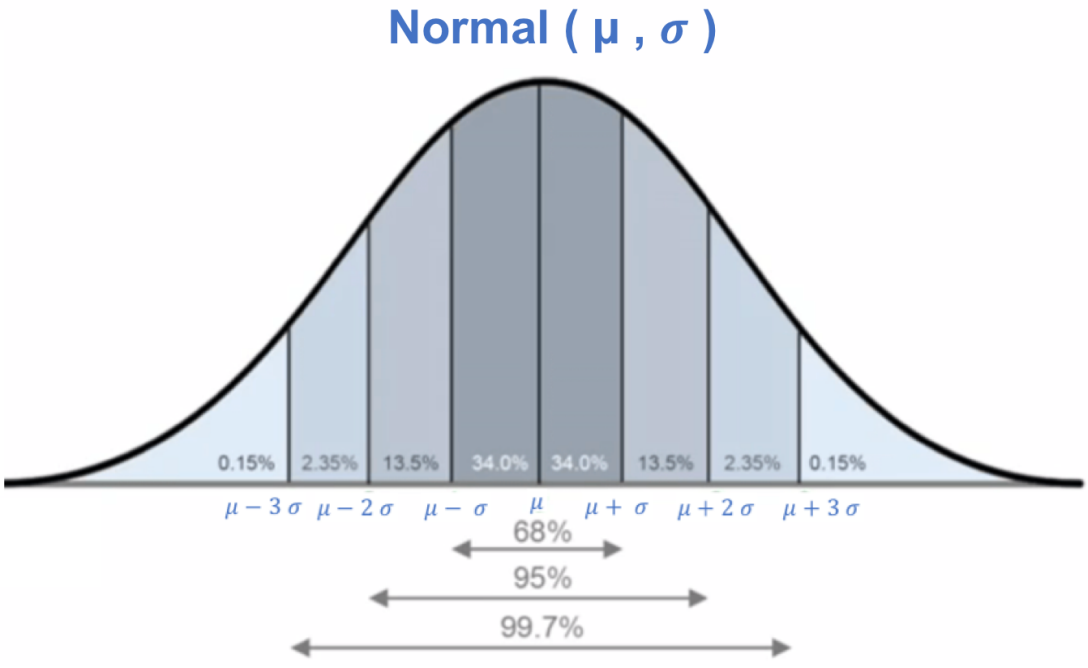
$$\hat{p} \sim Normal(mean = p, sd = \sqrt{\frac{p(1-p)}{n}})$$

Example

According to the National Postsecondary Student Aid Study conducted by the U.S. Department of Education in 2008, 62% of graduates from public universities had student loans. We randomly sample college graduates from public universities and determine the proportion in the sample with student loans. For which of the following sample sizes is a normal distribution a good fit for the distribution of sample proportions? Check all that apply.
$n = 10, n = 20, n = 30$ > Recall that the normal approximation for proportions is accurate if $np > 10$ and $n(1 - p) > 10$.
$p = 0.62$.
$n * 0.62 > 10$ and $n * (1 - 0.62) > 10$
$n > \frac{10}{0.62}$ and $n > \frac{10}{0.38}$.
$n > 16.1$ and $n > 26.3$.
Therefore, $\mathbf{n = 30}$ is a good sample size. If we randomly sample 50 students at a time, what will be the mean of the distribution of sample proportions? What will be the standard deviation of the sample proportions? > $n = 50$,
$p = 0.62$.
Therefore, $\mu_{\hat{p}} = p = 0.62$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.62(0.38)}{50}} = 0.072$

Exercise

A random sample of 100 students is taken from the population of all part-time students in the United States, for which the overall proportion of females is 0.6. a. What is the probability distribution function of the proportion of females of all the samples of size 100 from this population? b. There is a 95% chance that the sample proportion $\hat{p}$ falls between what two values? c. What is the probability that sample proportion $\hat{p}$ is less than 0.55? d. What is the probability that sample proportion $\hat{p}$ is less than or equal to 0.58?



In [ ]:

In [ ]:

4. Behavior of Sample Mean

Example: Birth weight

Birth weights are recorded for all babies in a town. The mean birth weight is 3,500 grams, $\mu = 3,500g$. If we collect many random samples of 9 babies at a time, how do you think sample means will behave?

Here again, we are working with a random variable, since random samples will have means that vary unpredictably in smaller samples but exhibit patterns as the sample gets larger.

Based on our intuition and what we have learned about the behaviour of sample proportions, we might expect the distribution of sample means to have the following properties:

- **Center**: Some sample means will be on the low side—say 3,000 grams or so—while others will be on the high side—say 4,000 grams or so. In repeated sampling, we might expect that the random samples will average out to the underlying population mean of 3,500 g. In other words, the mean of the sample means will be μ, just as the mean of sample proportions was p.
- **Spread**: For large samples, we might expect that sample means will not stray too far from the population mean of 3,500. Sample means lower than 3,000 or higher than 4,000 might be surprising. For smaller samples, we would be less surprised by sample means that varied quite a bit from 3,500. In others words, we might expect greater variability in sample means for smaller samples. So sample size will again play a role in the spread of the distribution of sample measures, as we observed for sample proportions.
- **Shape**: Sample means closest to 3,500 will be the most common, with sample means far from 3,500 in either direction progressively less likely. In other words, the shape of the distribution of sample means should bulge in the middle and taper at the ends with a shape that is somewhat normal. This, again, is what we saw when we looked at the sample proportions.

Links:

- https://www.youtube.com/watch?v=fqOOownnkA4 (https://www.youtube.com/watch?v=fqOOownnkA4)
- https://www.youtube.com/watch?v=cyNqdostWzk&t=101s (https://www.youtube.com/watch?v=cyNqdostWzk&t=101s)

5. Central Limit Theorem

The central limit theorem concerns the means of samples drawn from a population. It specifies that this mean follows a normal distribution. We will introduce the Central Limit Theorem with this video: https://www.youtube.com/watch?v=JNm3M9cqWyc (https://www.youtube.com/watch?v=JNm3M9cqWyc)

Using the theoretical results of the **Central Limit Theorem (CLT)**, we can quantify the *uncertainty* about the real population parameters. For distributions with finite mean and standard deviation **(no matter what the distribution is)**, according to the CLT, among all the possible samples (of the same size) drawn, the means of these samples:

$$\bar{x} = \frac{\sum_i x_i}{n}$$

are distributed as

$$\bar{x} \sim \text{Normal}(\text{mean} = \mu, \text{sd} = \sigma/\sqrt{n}),$$

where x_i refers to individual observations drawn from the sample, n is the **sample size**, and μ and σ are the **population mean** and **standard deviation** respectively.

$\frac{\sigma}{\sqrt{n}}$ is commonly referred to as the **standard error** - the standard deviation of the sampling distribution.

Example

Household size in the United States has a mean of 2.6 and standard deviation of 1.4. What is the probability that the mean size of a random sample of 100 households is more than 3? > The mean of samples of household size is normally distributed with mean $\mu = 2.6$ and the standard deviation=$\frac{\sigma}{\sqrt{n}}$, where n is the sample size and σ is the population standard deviation.
> The distribution of the mean of the household size in samples of size 100 follows a Normal distribution:

$$Normal(\mu = 2.6, \sigma = \frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{100}} = 0.14)$$

>

$$P(X > 3) = 1 - P(X < 3) = 1 - norm.cdf(3, mean = 2.6, sd = 0.14)$$

In [ ]:

```
mu, sigma, n = 2.6, 1.4, 100
sd=sigma/(n ** .5)
print('mu = ', mu, ', sd = ', sd)
print('P( X > 3 ) = 1 - P( X < 3) = 1 - norm.cdf(3, mu, sd) =
', round(1 - norm.cdf(3, mu, sd),4))
```

Exercise

The annual salary of teachers in a certain state X has a mean of MYR 54,000 and standard deviation of MYR 5,000. What is the probability that the mean annual salary of a random sample of 64 teachers from this state is less than MYR 52,000?

In [ ]:

Exercise

Scores on the math portion of the SAT (SAT-M) in a recent year follow a normal distribution with mean $\mu = 507$ and standard deviation $\sigma = 111$. What is the probability that the mean SAT-M score of a random sample of 40 students who took the test that year is more than 600?

In [ ]: