EXERCISE CLASS 2 (Part 1/3)

Review of basic statistical concepts - Hypothesis testing

Chapter 3-4, D.C. Montgomery: "Statistical Quality Control - an introduction", 7th Ed., Wiley

Statistical Inference

We want to infer properties of the source population by analysing data that are sampled from that distribution

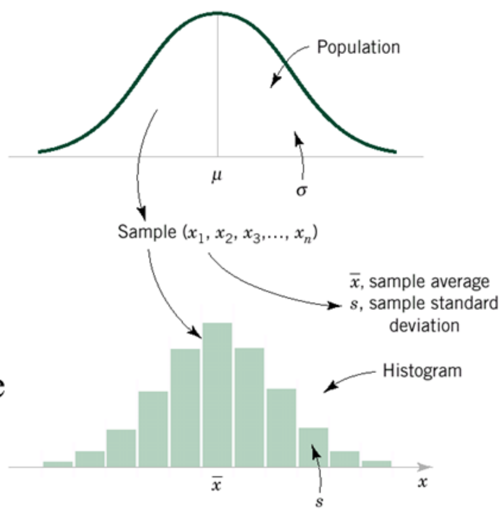
Point estimators

A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\widehat{\Theta}$

The **point estimator** $\widehat{\Theta}$ is an unbiased estimator of the parameter θ if:

$$E\left(\widehat{\Theta}\right) = \theta$$

If the estimator is not unbiased, then the difference $E\left(\widehat{\Theta}\right) - \theta$ is called **bias** of the estimator $\widehat{\Theta}$



REMIND

Unknown Parameter θ	Statistic $\hat{\Theta}$	Point Estimate $\hat{\theta}$
μ	$\overline{X} = \frac{\Sigma X_i}{n}$	$\overline{x}$
σ^2	$S^2 = \frac{\Sigma (X_i - \overline{X})^2}{n - 1}$	s^2
$\mu_1 - \mu_2$	$\overline{X}_1 - \overline{X}_2 = \frac{\Sigma X_{1i}}{n_1} - \frac{\Sigma X_{2i}}{n_2}$	$\overline{x}_1 - \overline{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

EXERCISE T1

Given a sample of n independent and identically distributed observations, demonstrate that the sample mean $\bar{X}$ and the sample variance S^2 are unbiased estimators

Exercise T1 (solution) (1/2)

Remind: $x_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$

$$E(x_i - \mu)^2 = E[(x_i)^2] - \mu^2$$

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} n\mu = \mu$$

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2x_i\bar{x})\right) =$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right]$$

Exercise T1 (solution) (2/2)

$$E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(x_i^2) - nE(\bar{x}^2)\right]$$

Remind:

$$Var(x_i) = \sigma^2 = E(x_i - \mu)^2 = E(x_i^2) - \mu^2 \Rightarrow \underline{E(x_i^2)} = \mu^2 + \sigma^2$$

Analogously:

$$Var(\bar{x}) = \sigma^2/n = E(\bar{x}^2) - \mu^2 \Rightarrow \underline{E(\bar{x}^2)} = \mu^2 + \sigma^2/n$$

Thus:

$$E(S^2) = \frac{1}{n-1}[n(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n)] =$$

$$E(S^2) = \frac{1}{n-1}[(n-1)\sigma^2] = \sigma^2$$

EXERCISE T2

A synthetic fiber used in manufacturing industry has an ultimate tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi.

a) Compute the probability that a random sample of 6 observations has a sample mean larger than 75.75 psi.

b) How does the standard deviation of the mean estimator change by passing from a sample of 6 observations to a sample of 49 observations?

In [ ]:
```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [ ]:
```
# Input data
mu = 75.5        # Mean
sigma = 3.5      # Standard deviation
```

Point a

Compute the probability that a random sample of 6 observations has a sample mean larger than 75.75 psi.

$$\mu = 75.5$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{6}}$$

$$P(\bar{X} \geq \mu_0) = P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \geq \frac{\mu_0-\mu}{\sigma_{\bar{X}}}) = P(Z \geq \frac{75.75-75.5}{1.429}) = 1 - P(Z \leq 0.175)$$

In [ ]:
```
n = 6           # Number of samples
mu0 = 75.75     # Hypothesized mean

# Under the assumption of normality, the probability of observing a sample mean lar
Z_0 = (mu0 - mu)/(sigma/np.sqrt(n))
prob = 1 - stats.norm.cdf(Z_0)
print('The probability of observing a sample mean larger than mu0 is: %.3f' % prob
```

```
The probability of observing a sample mean larger than mu0 is: 0.431
```

Point b

How does the standard deviation of the mean estimator change by passing from a sample of 6 observations to a sample of 49 observations?

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{X}}(n=6) = \frac{3.5}{\sqrt{6}} = 1.429$$

$$\sigma_{\bar{X}}(n = 49) = \frac{3.5}{\sqrt{49}} = 0.5$$

In [ ]:
```
n_new = 49      # Number of samples

sigma_n = sigma/np.sqrt(n)               # Standard deviation of the mean with n = (
sigma_n_new = sigma/np.sqrt(n_new)       # Standard deviation of the mean with n = 4

print('The standard deviation of the mean with n = 6 samples is: %.3f psi' % sigma_
print('The standard deviation of the mean with n = 49 samples is: %.3f psi' % sigma

print('The difference between the two standard deviations is: %.3f psi' % (sigma_n_
```

```
The standard deviation of the mean with n = 6 samples is: 1.429 psi
The standard deviation of the mean with n = 49 samples is: 0.500 psi
The difference between the two standard deviations is: -0.929 psi
```

EXERCISE T3

A random sample of size 16 is drawn from a normal population with mean 75 and standard deviation 8. A second sample of size 9 is drawn from a normal population with mean 70 and standard deviation 12.

a) Compute the probability that the sample mean difference between the first and the second sample is greater than 4 (assume that the two populations are independent).

b) Compute the probability that the sample mean difference between the first and the second sample ranges between 3.5 and 5.5 (same assumption).

In [ ]:
```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [ ]:
```
# Input data
n1 = 16          # Number of samples
mu1 = 75         # Mean
sigma1 = 8       # Standard deviation

n2 = 9           # Number of samples
mu2 = 70         # Mean
sigma2 = 12      # Standard deviation
```

Point a

Compute the probability that the sample mean difference between the first and the second sample is greater than 4 (assume that the two populations are independent).

a)

$n_1 = 16 \qquad n_2 = 9$

$\mu_1 = 75 \qquad \mu_2 = 70$

$\sigma_1 = 8 \qquad \sigma_2 = 12$

Remind:

$V(x_1 - x_2) = \sigma_1^2 + \sigma_2^2 - 2Cov(x_1, x_2)$

If they are independent, cov=0

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_{\bar{X}_1} - \mu_{\bar{X}_2}, \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}) = N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) = N(75 - 70, \frac{8^2}{16} + \frac{12^2}{9})$$

$$\bar{X}_1 - \bar{X}_2 \sim N(5, 20)$$

$$P(\bar{X}_1 - \bar{X}_2 > 4)$$

$$P(Z > \tfrac{4-5}{\sqrt{20}}) = P(Z > -0.2236) = 1 - P(Z \le -0.2236)$$

$$= 1 - 0.4115 = 0.5885$$

In [ ]:
```
# Answer to point a
# Compute the mean and the variance of the difference between the two populations
mu_diff = mu1 - mu2
sigma_diff = np.sqrt(sigma1**2/n1 + sigma2**2/n2)   # the operator ** stands for ^

mu0 = 4        # Difference between the means

# P(X1 - X2 > mu0) = P(Z > (mu0 - mu_diff)/sigma_diff)
prob = 1 - stats.norm.cdf((mu0 - mu_diff)/sigma_diff)

print('Probability of the difference between the means being greater than %.1f is %
```

```
Probability of the difference between the means being greater than 4.0 is 0.5885
```

Point b

Compute the probability that the sample mean difference between the first and the second sample ranges between 3.5 and 5.5 (same assumption).

> ## Solution
>
> We can use he following formula to compute the probability:
>
> $$Pr(3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5) = Pr(\frac{3.5-5}{\sqrt{20}} \leq Z \leq \frac{5.5-5}{\sqrt{20}}) = Pr(Z \leq \frac{5.5-5}{\sqrt{20}}) -$$

In [ ]:
```
# Answer to point b
lower_bound = 3.5       # Lower bound of the interval
```

```
upper_bound = 5.5      # Upper bound of the interval

# P(lower_bound < X1 - X2 < upper_bound) = P(X1 - X2 < upper_bound) - P(X1 - X2 < l
prob = stats.norm.cdf((upper_bound - mu_diff)/sigma_diff) - stats.norm.cdf((lower_l

print('Probability of the difference between the means being between %.1f and %.1f
```

```
Probability of the difference between the means being between 3.5 and 5.5 is 0.175
9
```