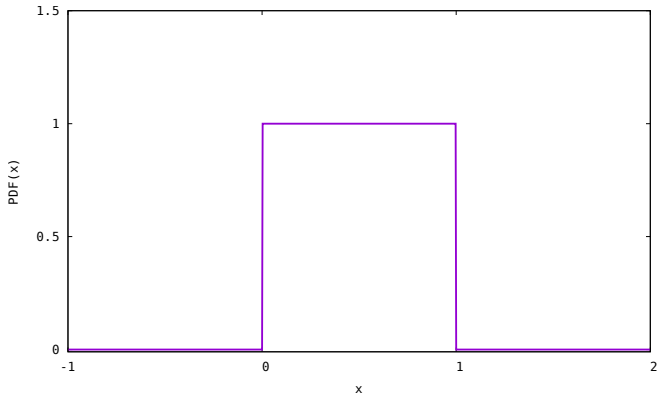
Exercise Class - Statistics Review

Instructor: Irene Iodice
Email: irene.iodice@malix.univ-paris1.fr

Ex.0: First moments of a continuous RV

1. *Let $X \sim Uniform(0,1)$. Find and draw the pdf. Find $P(x_i)$, the cdf for $X = x_i$. Compute E(X) and Var(X).*
   X has range [0, 1] and since it is a Uniform distribution, which means that each value in the range is equally probable, its pdf(x) is an horizontal line, which identifies a rectangular with the x-axis. Of which height? Recall that the area below the curve of the pdf has to be 1. Then, since the basis of the rect. is 1, also its height has to be 1: $\text{p(x)}=\begin{cases}1 & if \quad 0 \leq x \leq 1\\ 0 & otherwise\end{cases}$

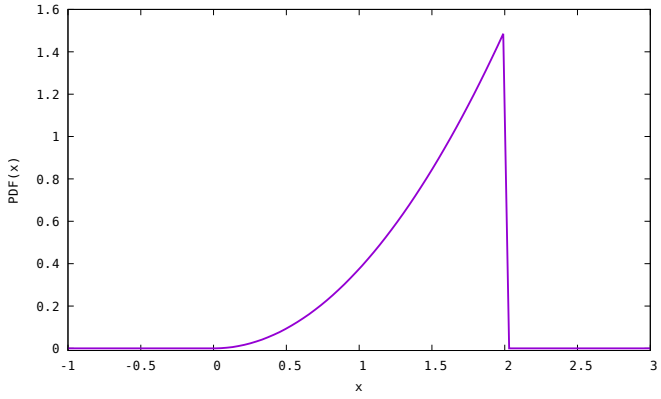


$$P(x_i) = Pr(X \leq x_i) = \int_0^{x_i} p(x)dx = \int_0^{x_i} 1dx = x\Big|_0^{x_i} = x_i$$

$$E(X) = \int_0^1 p(x)xdx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

$$Var(X) = \int_0^1 (x-\mu_x)^2 dx = \frac{1}{3}\left(x-\frac{1}{2}\right)^3\Big|_0^1 = \frac{1}{12}$$

2. *Let X have range $[0,2]$ and density $\frac{3}{8}x^2$. Sketch the pdf. Find E(X) and discuss its position with respect to the previous point)*



$$E(X) = \int_0^2 \frac{3}{8}x^3 dx = \frac{3}{32}x^4\Big|_0^2 = \frac{3}{2}$$

Since the probability density increases as x increases over the range, the average value of x should be in the right half of the range. In other words, E(X) is pulled to the right of the midpoint 1 because there is more mass to the right.

Ex.1: Properties of estimators

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population from a $N(\mu, \sigma^2)$. Consider the following estimator of μ:

$$\hat{X} = \frac{X_1 + X_2 + X_3}{3}$$

1. *Show that $\hat{X}$ a linear estimator.*

   Recall that a linear estimator is a linear combination of the data at hand, which means that it can be rewritten as $\sum w_i X_i$, where w_i are constants. In our case we have:
   $\hat{X} = \frac{X_1+X_2+X_3}{3} = \frac{X_1}{3} + \frac{X_2}{3} + \frac{X_3}{3} = \sum_{i=1}^{3} \frac{1}{3} X_i$.
   This means that our estimator is linear with $w_i = \frac{1}{3}$ for $i = 1, 2, 3$ and $w_i = 0$ for $3 < i \leq n$.

2. *Show that $\hat{X}$ is an unbiased estimator.*

   Recall that an estimator is unbiased if $E(\hat{X}) = \mu$, which is if the expected value of the estimator equates the true population parameter it estimates. In our case it is unbiased since: $E(\hat{X}) = E\big(\frac{X_1+X_2+X_3}{3}\big) = \frac{E(X_1)}{3} + \frac{E(X_2)}{3} + \frac{E(X_3)}{3} = \frac{\mu}{3} + \frac{\mu}{3} + \frac{\mu}{3} = \mu$

3. *Compute the variance of the estimator.*
   Recall that

   - $Var(w_1X_1+w_2X_2+w_3X_3) = w_1^2 Var(X_1)+w_2^2 Var(X_2)+w_3^2 Var(X_3)+2w_1w_2Cov(X_1, X_2)+ 2w_1w_3Cov(X_1, X_3) + 2w_2w_3Cov(X_2, X_3)$

   Then, $Var(\hat{X}) = Var\big(\frac{X_1+X_2+X_3}{3}\big) = \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{2}{9}\sigma_{X_1,X_2} + \frac{2}{9}\sigma_{X_1,X_3} + \frac{2}{9}\sigma_{X_2,X_3}$ since the variables are obtained from random sampling they are independent, and thus their covariances are zero, thus: $Var(\hat{X}) = \frac{1}{3}\sigma^2$

Consider the following weighted estimator:

$$\hat{\hat{X}} = \frac{3X_1}{8} + \frac{X_2}{2} + \frac{X_3}{8}$$

1. *Show that also $\hat{\hat{X}}$ is an unbiased estimator.*

   $E(\hat{\hat{X}}) = \frac{3E(X_1)}{8} + \frac{E(X_2)}{2} + \frac{E(X_3)}{8} = \frac{3\mu}{8} + \frac{\mu}{2} + \frac{\mu}{8} = \mu$

2. *Compute the variance of the estimator. Is this more efficient than $\hat{X}$?*

   $Var(\hat{\hat{X}}) = Var\big(\frac{3X_1}{8} + \frac{X_2}{2} + \frac{X_3}{8}\big) = \frac{9}{64}\sigma^2 + \frac{1}{4}\sigma^2 + \frac{1}{64}\sigma^2 = \frac{13}{32}\sigma^2$
   Since $Var(\hat{\hat{X}}) > Var(\hat{X})$, $\hat{\hat{X}}$ is less efficient.

3. *If $\sigma^2 = 9$, calculate the probability that each estimator is within 1 unit on either side of μ. Compare and comment.*

$$Pr\big(\mu-1\leq\hat{X}\leq\mu+1\big)=Pr\left[\frac{-1}{\sqrt{Var(\hat{X})}}\leq\frac{\hat{X}-\mu}{\sqrt{Var(\hat{X})}}\leq\frac{1}{\sqrt{Var(\hat{X})}}\right]$$
$$=P\left[\frac{-1}{3/\sqrt{3}}\leq Z\leq\frac{1}{3/\sqrt{3}}\right]=Pr\big[-0.57\leq Z\leq 0.57\big]$$
$$=2\Big[Pr\big[Z\leq 0.57\big]-0.5\Big]=2(0.7157-0.5)=0.4314$$

$$Pr\big(\mu-1\leq\hat{\hat{X}}\leq\mu+1\big)=Pr\left[\frac{-1}{\sqrt{Var(\hat{\hat{X}})}}\leq\frac{\hat{\hat{X}}-\mu}{\sqrt{Var(\hat{\hat{X}})}}\leq\frac{1}{\sqrt{Var(\hat{\hat{X}})}}\right]$$
$$=Pr\left[\frac{-1}{3\sqrt{13/32}}\leq Z\leq\frac{1}{3\sqrt{13/32}}\right]=Pr\big[-0.52\leq Z\leq 0.52\big]$$
$$=2\Big[Pr\big[Z\leq 0.52\big]-0.5\Big]=2(0.6985-0.5)=0.3947$$

Since the variance of $\hat{X}$ is smaller, the probability that $\hat{X}$ is within an unit of distance from the real population mean μ is higher.

4. *What is the most efficient with the data at hand? Why?* The most efficient estimator is the sample average that uses all n information. In that case the variance would be $\frac{\sigma^2}{n}$ which is smaller than that of the other two estimators. In general, increasing sample information reduces sampling variation.

Ex.2: Introduction to the concept of p-value

The hourly production of screws in a factory is normally distributed with mean 2,000 pieces and standard deviation 500 pieces. What is the probability that in a eight-hour day more than 17776 pieces will be sold?

Let X be the random variable denoting the hourly production of screws in the factory which is normally distributed; $X\sim N(2000,500^2)$.

The probability that in a 8 hour day, more than 17776 pieces will be sold is the same as the probability that average hourly production is greater than $17776/8=2,222$ pieces. Since the population mean is known, then the standard deviation of the sampling distribution of $\overline{X}$ is $\sqrt{Var(\overline{X})}=\frac{\sigma}{\sqrt{N}}$.

$$Pr[\overline{X}>2222]=Pr\left[\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{N}}}>\frac{2222-2000}{500/\sqrt{8}}\right]=Pr\left[Z>\frac{2222-2000}{500/\sqrt{8}}\right] \quad (1)$$
$$=P\left[\frac{222\sqrt{8}}{500}\right]=P[Z>1.25]=1-P[Z<1.25]=1-0.8944=0.1056 \quad (2)$$

The probability that in a 8 hour day more than 17776 pieces will be sold is around 10%.

Ex.3: Performing a test

The same Instructor of Statistics you met last week, is now wondering whether his students are as diligent as he would expect. In particular, he expects them to perform more than three exercises by themselves for each exercise he did in class. This means that for six exercises he performs in class, the students are expected to do more than 18 by themselves. The instructor randomly select eight students

from the class and asks how many exercises they did. The samples values are 3,9,12,12,18,18,24,36

1. *Assuming that the population is normally distributed, can the professor conclude at the 0.05 level of significance that the students are doing on average more than 18 exercises per class?*
   The instructor first computes the sample mean and the sample variance:

   - $\overline{X} = \frac{\sum_i x_i}{n} = \frac{3+9+12+12+18+18+24+36}{8} = 16.5$
   - $s_X^2 = \frac{\sum_i (x_i - \overline{X})^2}{n-1} = \frac{(-13.5)^2+(-7.5)^2+2(-4.5)^2+2(-1.5)^2+(7.5)^2+(19.5)^2}{7} = 102.857$

   The Instructor set the following test structure:

   - Assumptions:
     - random sample
     - normal distribution of the population
   - Hypotheses:
     H_0 : $\mu = 18$
     H_1 $\mu > 18$
   - Test statistic, given H_0 is true, is:

   $$t = \frac{\overline{X} - \mu_0}{\sqrt{\frac{s_x^2}{n}}} = \frac{\overline{X} - \mu_0}{\sqrt{SE(\overline{X})}}$$

   - Distribution of test statistic:
     If the assumptions are met and H_0 is true, the test statistic is distributed as Student's t distribution with 7 (which is n-1, 1 df less from estimating sample mean) degrees of freedom.
   - Testing procedure:
     - Decision rule: with $\alpha = .05$ and $df = 7$, the critical value of $t_{\alpha,7}$ is 1.895. He rejects H_0 if $t > 1.895$.
     - Calculation of test statistic:

     $$t^{act} = \frac{16.5 - 18}{\sqrt{\frac{102.85}{8}}} = \frac{1.5}{3.58} = -0.418 \tag{3}$$

     - Statistical decision: Since $t^{act} = -0.418 < t = 1.895$, we do not reject H_0 and therefore cannot conclude that, at the 0.05 level of significance, the students are doing more than 18 exercises per class.

2. *Construct a 95% confidence interval for the population mean number of exercises. Comment the statistical idea behind a CI*

   Recall that a $100(1-\alpha)\%$ confidence interval for the mean of a normal population is that interval between two values, that we call *ub* and *lb*, upper bound and lower bound, such that:

   $$Pr\big(lb \leq \frac{\overline{X} - \mu}{SE(\overline{X})} \leq ub\big) = 95\% \tag{4}$$

   because of normalization around zero and small sample size then $|lb| = |up| = |t_{\alpha/2}|$,where $t_{\alpha/2}$ is taken from a t-Student with 7df. For $\alpha = 0.05$, then $t_{\alpha/2}$ is 2.365. Thus our interval is $\overline{X} \pm t_{\alpha/2,7} \times SE(\overline{X}) = 16.5 \pm 2.365 \times 3.58$ which identifies the interval [8,24.97]. If the same population of students is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases. Thus, a confidence stated at a $1-\alpha$ level can be thought of as the inverse of a significance level, α.

3. *What would be a Type 1 error in this example? and a Type 2? Which one might interest your statistical decision.*

   Recall that a Type I error is the error we make when we incorrectly reject H_0 when H_0 is true (FALSE POSITIVE). In this case it would mean that the average student does not more than 18 exercises exercises but we draw the most diligent ones and we concluded that students are doing enough homeworks.

   A type II error is the error you make when we fail to reject the H_0 when it is actually false (FALSE NEGATIVE). In our case this would mean that the average student actually does more than 18 exercises per class, but we draw a sample of particularly lazy students and we therefore concluded that students are not diligent as expected.

   Since in our case we do not reject the null hypothesis, we might be committing an error of type 2, with a probability given by the size of the test α.

4. *Since the Instructor is good in statistics, he knows that he can tighten the assumption that the number of exercises done by the students are normally distributed, how? Explain two alternatives.*

   He can either increases the sample size or perform a parameter free test. In the first case, the instructor can apply the CLT if he samples more than 30 individuals, and thus has:

$$\frac{\overline{X} - \mu}{SE(\overline{X})} \xrightarrow{d} N(0,1) \tag{5}$$

   In the second case he might perform a Wilcoxon free parameter test, by just preserving the assumption that the distribution of the number of exercises is symmetric.

Ex.4: Testing differences in population means

We wish to know if we may conclude, at the 95% confidence level, that smokers, in general, have similar lung cancer spread than do non-smokers. A laboratory provides us with the following data, where X represents some measure on the speed at which Non-Small Cell Lung spreads:

	$\overline{X}$	n	σ_x^2
Smokers	17.5	18	4
Non-smokers	15.5	9	2

1. *Perform a test assuming that the populations are normally distributed.*

   - Assumptions:
     - two independent random samples
     - each drawn from a normally distributed population
   - Hypotheses:
     $H_0: \quad \mu_{X_S} = \mu_{X_N}$
     $H_1: \quad \mu_{X_S} \neq \mu_{X_N}$
   - Test statistic, given H_0 is true, is:

$$z = \frac{(\overline{X_S} - \overline{X_N}) - (\mu_{X_s} - \mu_{X_N})_0}{\sqrt{\frac{\sigma_{x_S}^2}{n_S} + \frac{\sigma_{x_N}^2}{n_N}}} = \frac{\overline{X_S} - \overline{X_N} - d_0}{\sqrt{\frac{\sigma_{x_S}^2}{n_S} + \frac{\sigma_{x_N}^2}{n_N}}}$$

   - Distribution of the test statistic:
     If the assumptions are correct and H_0 is true, the test statistic is distributed as the normal distribution.
   - Testing procedure:

- Decision rule: with alpha = .05, the critical values of z are -1.96 and +1.96. We reject $\mathrm{H}_0 if \mathrm{z}^{act} < -1.96$ or $z^{act} > 1.96$.
- Calculation of test statistic:

$$z^{act} = \frac{17.5 - 15.5 - 0}{\sqrt{\frac{4}{18} + \frac{2}{9}}} = \frac{2}{2/3} = 3$$

- Statistical decision: we reject H_0 because $z^{act} > 1.96$.
- Conclusion: from these data, it can be concluded that the population means are not equal. A 99% confidence interval would give the same conclusion since $z_{0.01/2} = 2.58$, this means that the p-value is smaller than $\alpha/2 = 0.01/2 = 0.005$.

2. *Unfortunately, the lab tells you that the population variances were added by mistake, and that instead they are unknown and, for now, they can just provide you with the sample variances, as in Table below. However, you found some previous research on the same topic, that shows that the population variances are equal. Perform again the test assuming that the populations are normally distributed and with equal variance.*

	$\overline{X}$	n	s_x^2
Smokers	17.5	18	7
Non-smokers	15.5	9	4

In this case we have few observations, but since we assume equal population variances, we can obtain a pooled value from the sample variances:

$$s_P^2 = \frac{(n_S - 1)s_{X_S}^2 + (n_N - 1)s_{X_N}^2}{n_S + n_N - 2} = \frac{17 \times 7 + 8 \times 4}{25} \approx 6$$

With respect to the test before what changes is:

- Assumptions:
  - independent random samples
  - normal distribution of the populations
  - population variances are equal
- Test statistic, given H_0 is true, is:

$$t = \frac{(\overline{X_S} - \overline{X_N}) - (\mu_{X_s} - \mu_{X_N})_0}{\sqrt{\frac{s_p^2}{n_S} + \frac{s_p^2}{n_N}}}$$

- Distribution of the test statistic:
  If the assumptions are correct and H_0 is true, the test statistic is distributed as Student's t distribution with 25 degrees of freedom.
- Testing procedure:
  - Decision rule: with alpha = .05, the critical values of t are -2.06 and +2.06. We reject H_0 if $t^{act} < -2.06$ or $t^{act} > 2.06$.
  - Calculation of test statistic:

$$t^{act} = \frac{17.5 - 15.5 - 0}{\sqrt{\frac{6}{18} + \frac{6}{9}}} = \frac{2}{\sqrt{1}} = 2$$

  - Statistical decision: we fail to reject H_0 because $|t^{act}| < |2.06|$.
  - Conclusion: from these data, we cannot conclude that the population means are different.

3. *After a while, the lab calls you back and inform you that they have collected more info for your research as in the table below. Perform the new test.*
   With respect to the test before what changes is:

	$\overline{X}$	n	s_x^2
Smokers	16.5	180	2
Non-smokers	15.5	90	1

   - Assumptions:
     - independent random samples
   - Test statistic: Because of the large samples, the central limit theorem permits calculation of the z score as opposed to using t. The z score is calculated using the given sample standard deviations.

$$z = \frac{(\overline{X_S} - \overline{X_N}) - (\mu_{X_s} - \mu_{X_N})_0}{\sqrt{\frac{s_{x_S}^2}{n_S} + \frac{s_{x_N}^2}{n_N}}} = \frac{\overline{X_S} - \overline{X_N} - d_0}{\sqrt{SE(\overline{X_S} - \overline{X_N})}}$$

   - Distribution of test statistic: if the assumptions are correct and H_0 is true, the test statistic is approximately normally distributed
   - Testing procedure:
     - Decision rule: the same as point 1).
     - Calculation of test statistic:

$$z = \frac{16.5 - 15.5 - 0}{\sqrt{\frac{2}{180} + \frac{1}{90}}} = \frac{1}{\sqrt{0.02}} = 6.7$$

     - Statistical decision: we reject H_0 because $|z| > |1.96|$.
     - Conclusion: from these data, we can state that the population means are different.

Ex.5: Sample dimension

The firm 'Pippo' is now evaluating the possibility to offer to its employees a free access to a new canteen. To plan the costs she need to know the yearly number of days spent in the office by its workers. She knows that this is normally distributed and the standard deviation is $\sigma = 50$ days. Since the company has thousands of workers, a sample is to be taken.

1. *How large has to be the sample size to ensure that the 95% coinfidence interval on the population mean is no more than four days wide?*
   The interval estimate of a normally distributed random variable is given by $\overline{X} \pm z_{\alpha/2} \times \sigma/\sqrt{N}$, where $z_{\alpha/2}$ is the corresponding critical value with $\alpha = 0.05$. The length of the interval is therefore $2(z_{\alpha/2} \times \sigma/\sqrt{N})$.
   To ensure that the length of the interval is less than 4, derive N as follows:

$$\begin{aligned} 2(z_{\alpha/2} \times \sigma/\sqrt{N}) &< 4 \\ (z_{\alpha/2} \times \sigma) &< 2\sqrt{N} \\ (1.96 \times 50)^2 &< 4N \end{aligned}$$

   Which is true for $N > 2401$. A sample size of at least 2401 employees is needed.

2. *Do you expect the sample size to be larger or smaller to ensure what before with a 90% interval?*
   The confidence interval at 90% is, *ceteris paribus*, smaller than that at 95%.