MASTER
Economics & Psychology

Universitè Paris 1 Panthèon-Sorbonne
Universitè Paris Descartes

Exercise Class - Statistics Review

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

Overview on integration

This is meant to be a VERY simplified overview on integration to give you the instruments to solve basic exercises as the ones below (for more on this, you can refer to William Neilson's Overview on Math for Economists). So, two most important things on integration:

- integration is the opposite of differentiation,
- integration finds the area under a curve.

To see this denote with $F(x)$ a function whose derivative is $f(x)$. The following two statements provide the fundamental relationship between derivatives and integrals:

$$\int_a^b f(x)dx = F(b) - F(a) \tag{1}$$

$$\int f(x)dx = F(x) + c \tag{2}$$

where c is a constant. The integral in (1) is a definite integral, and its distinguishing feature is that the integral is taken over a finite interval. The integral in (2) is an indefinite integral, and it has no endpoints. The reason for the names is that the solution in (1) is unique, or definite, while the solution in (2) is not unique. This occurs because when we integrate the function $f(x)$, all we know is the slope of the function $F(x)$, and we do not know anything about its height. If we choose one function that has slope $f(x)$, call it $F(x)$, and we shift it upward by one unit, its slope is still $f(x)$. The role of the constant c in (2), then, is to account for the indeterminacy of the height of the curve when we take an integral. The two equations (1) and (2) are consistent with each other. To see why, notice that

$$\int f(x)dx = \int_{-\infty}^{\infty} f(x)dx \tag{3}$$

so an indefinite integral is really just an integral over the entire real line $(-\infty, \infty)$. Some important properties to remember:

- Additive properties: $\int_a^a f(x)dx = 0$; $\int_a^b f(x)dx = -\int_b^a f(x)dx$; $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$
- Scaling by a constant and integral of a sum: $\int_a^b cf(x)dx = c\int_a^b f(x)dx$; $\int_a^b \big(f(x) + g(x)\big)dx = \int_a^b f(x)dx + \int_a^b g(x)dx$

The former point says that a constant inside of the integral can be moved outside of the integral and that the integral of the sum of two functions is the sum of the two integrals. Together they say that integration is a linear operation.
The most straightforward integrals, those that you can find in future exercises, are the follows (note that you can easily check by differentiating the right hand side)

- $\int x^n dx = \frac{x^{n+1}}{n+1} + c$
- $\int \frac{1}{x} dx = ln(x) + c$
- $\int e^{rx} dx = \frac{e^{rx}}{r} + c$

Ex.0: First moments of a continuous RV

1. *Let $X \sim Uniform(0,1)$. Find and draw the pdf. Find the cdf. Compute E(X) and Var(X).*

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)*

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.*
2. *Show that $\hat{X}$ is an unbiased estimator.*
3. *Compute the variance of the estimator.*

Consider the following weighted estimator:

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

5. *Show that also $\hat{\hat{X}}$ is an unbiased estimator.*
6. *Compute the variance of the estimator. Is this more efficient than $\hat{X}$?*
7. *If $\sigma^2 = 9$, calculate the probability that each estimator is within 1 unit on either side of μ. Compare and comment.*
8. *What is the most efficient with the data at hand? Why?*

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?

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 solving on average 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.*

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

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.*

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.*
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 equal.*

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

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.*

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

Ex.5: Sample dimension

The firm 'Pippo' is now evaluating the possibility to offer to its employees a free access to a new canteen and has to determine the average days a year that his workers are in their office, knowing that this is normally distributed and the standard deviation is $\sigma = 50$ days. Since the corporation has thousands of workers, a sample is to be taken.

1. *How large should the sample be to ensure that a 95% interval estimate of mean days at office is no more than 4 days wide?*
2. *Without computing it, do you expect the sample size to be larger or smaller to ensure what before with a 90% interval?*