Statistics for the Doctoral School in Biomolecular Sciences
Academic year 2014-2015

Exercises after Lecture 3

Introductory statistics

The following are basic (and quite intuitive) properties of expectation and variance that have to be used in the next 3 exercises:

$$\mathbb{E}(cX) = c\mathbb{E}(X), \qquad \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$
$$\mathbb{V}(cX) = c^2\mathbb{V}(X) \qquad \mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) \text{ (if } X \text{ and } Y \text{ are independent)}$$

EXERCISE 3.1 A morphometric index D is the difference between the length of lower L_i and upper limbs L_u. Assume that both the measures of L_i and L_u have an error of mean 0 and standard deviation 1 cm, and that the errors of the two measures are independent. Which is the standard deviation of the error in the measure D?

Repeat now the same measurements on a sample of $n = 100$ individuals, and compute the mean (on this sample) of L_i, L_u and D. Which is the standard deviation of the error of the means?

EXERCISE 3.2 Show that, if $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \sigma^2$, then $Y = (X-\mu)/\sigma$ satisfies $\mathbb{E}(Y) = 0$ and $\mathbb{V}(Y) = 1$. On the basis of this, convince yourself[1] that, if $X \sim N(\mu, \sigma^2)$, then $Y \sim N(0,1)$ a standard normal.

The following are properties of the standard normal distribution that are useful to compute confidence intervals and to test hypotheses. If $X \sim N(0,1)$ (standard normal), then

$$\begin{array}{lcl} \mathbb{P}(|X| > 1.645) = 10\% & & \mathbb{P}(X > 1.645) = \mathbb{P}(X < -1.645) = 5\% \\ \mathbb{P}(|X| > 1.96) = 5\% & & \mathbb{P}(X > 1.96) = \mathbb{P}(X < -1.96) = 2.5\% \\ \mathbb{P}(|X| > 2.326) = 2\% & & \mathbb{P}(X > 2.326) = \mathbb{P}(X < -2.326) = 1\% \\ \mathbb{P}(|X| > 2.576) = 1\% & & \mathbb{P}(X > 2.576) = \mathbb{P}(X < -2.576) = 0.5\% \end{array}$$

EXERCISE 3.3 Assume that X follows a normal distribution with expected value μ not known and variance $\sigma^2 = 4$. We take n measures $x_1, \dots x_n$ and compute $\bar{x}$.

Show that $\sqrt{n}\frac{\bar{X}-\mu}{2}$ follows a standard normal distribution. Conclude that $\mathbb{P}(\sqrt{n}\frac{|\bar{X}-\mu|}{2} > 1.96) = 5\%$.

From the previous inequality, find the value k such that $\mathbb{P}(|\bar{X} - \mu| > k) = 5\%$. [*The interval* $(\bar{x}-k, \bar{x}+k)$ *is named a 95% confidence interval for* μ*. On the basis of the previous computations, discuss why this is an appropriate word and what exactly it implies.*]

Find the number of samples n such that the value k, obtained in the previous inequality satisfies $k < 0.1$.

EXERCISE 3.4 A researcher has measured the distance between canine and last molar in 35 upper jaws of wolves, and states that the 95% confidence interval for the mean distance is 10.17 cm $< \mu <$ 10.47 cm and that the 99% confidence interval for the mean distance is 10.21 cm $< \mu <$ 10.44 cm. Why can we say there is a mistake, even without having seen the data?

And, assuming that the 95% confidence interval is correct and that we can approximate the distribution of the distance as a normal, can we compute the correct 99% confidence interval, again without seeing the data?

EXERCISE 3.5 The quantity X has in a population variance equal to 4. Computing the sample mean $\bar{X}$ on a sample (from this population) of size 100, build a symmetric interval around μ_X (the true mean of X) in which $\bar{X}$ is contained with probability 90%. For this computation, assume that $\bar{X}$ follows a normal distribution.

[1] one may note that there is a missing mathematical step that we take for granted.

EXERCISE 3.6 Assume that $\bar{X}_1$ and $\bar{X}_2$ are the mean of two independent samples of size n from the same population that has variance σ^2 and normal distribution. Find n such that the probability that $\bar{X}_1$ and $\bar{X}_2$ differ more than σ is 1%. *[use $Y = \bar{X}_1 - \bar{X}_2$]*

EXERCISE 3.7 The tensile strength has been measure in 20 carbon fiber samples obtaining a sample mean of 5,450 and a standard deviation of 250 (in MPa).

Assuming that the tensile strength follows a normal distribution, compute a 95% confidence interval for the mean tensile strength. *[The result can be computed using quantiles from the normal distribution, instead of the appropriate t-distribution; the correct method should however be indicated]*

Assuming that the true mean and variance are equal to the sample ones, find the stress at which only 5% of fibers would 'neck'[2].

EXERCISE 3.8 We want to take a poll about the vote at a referendum, choosing a sample of n individuals. How large must the sample be for the standard deviation of the sample mean (i.e., of the fraction of respondents in favour of the referendum) be below 0.1? *[we do not know a priori the fraction in favour]*

And if we wish to be sure that the standard deviation is below 0.01?

[2]tensile strength is defined as the maximum stress that a material can withstand while being stretched or pulled before 'necking,' which is when the specimen's cross-section starts to significantly contract