

## Sample exam questions

1. Consider a single channel microarray experiment (such as one using radioactive labelling on filters). We are interested to distinguish between genes that are expressed and genes that are not expressed in the given mRNA sample. Let us assume that many previous experiments have shown that both expressed genes and non-expressed genes follow a normal distribution with the same  $\sigma = 10$  but with different means:  $\mu = 55$  for unexpressed genes and  $\mu = 60$  for expressed genes.
  - (a) Formulate the null and research hypotheses.
  - (b) Draw a graph of the two distributions and show on the graph the area corresponding to unexpressed genes mistakenly identified as expressed.
  - (c) Assume that we have a sample of 25 genes. Calculate the threshold for the sample mean  $\bar{X}$  that will ensure that unexpressed genes are mistakenly identified as expressed genes in at most 1 in 20 cases.
  - (d) Calculate the probability of a Type II error in the conditions above.
  - (e) Show on the graph the area corresponding to the Type II error.
  - (f) Calculate the power of this test for the situation given above.
  - (g) Recalculate the numbers above for a p-value equal to 0.01. What happens with the power of the test?

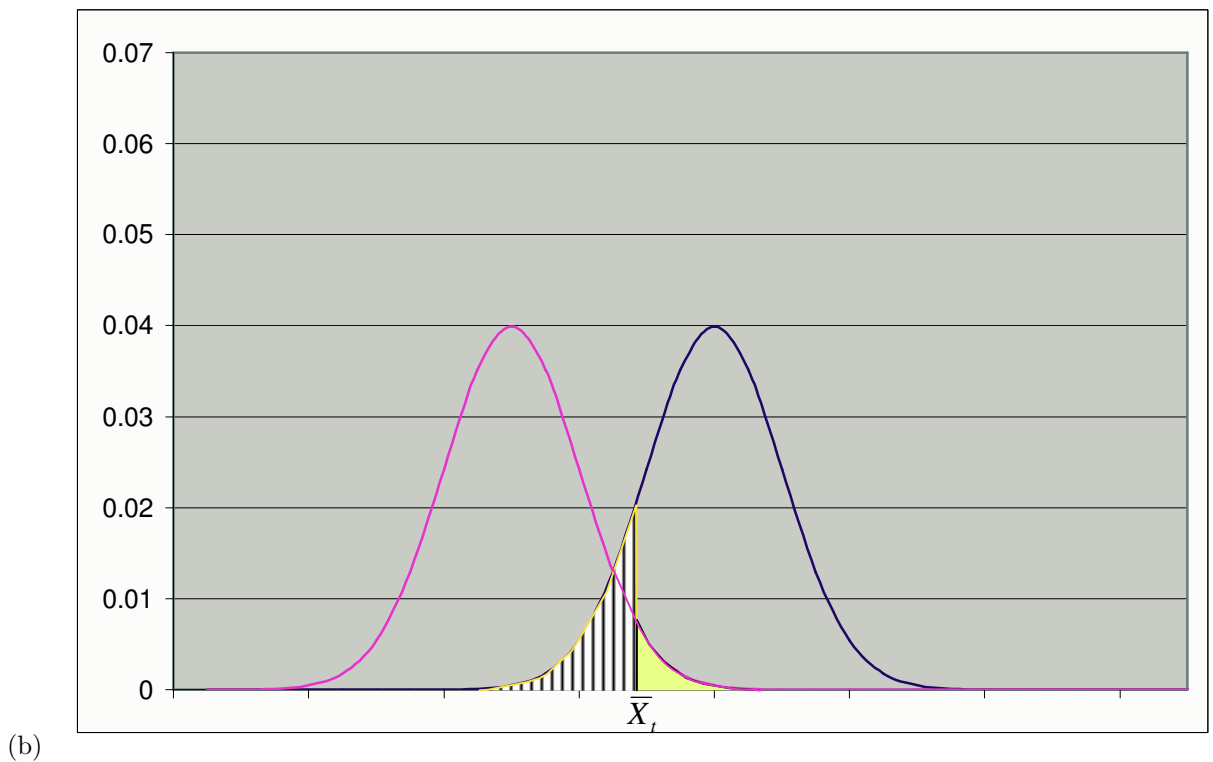
Solution

- (a) Null and research hypotheses.

There are only two possibilities: either the sample comes from the distribution of expressed genes or the sample comes from the distribution of unexpressed genes.

$H_0$  is: The sample comes from the distribution of unexpressed genes.

$H_a$  is: The sample comes from the distribution of expressed genes. The two hypotheses are mutually exclusive (only one can be true at any one time) and all inclusive (the set of the two hypotheses encompass all possibilities)



The colored area to the right of the  $\bar{X}_t$  corresponds to the unexpressed genes mistakenly identified as expressed.

(c) 1 in 20 corresponds to a p value of 0.05. We are measuring the mean of a sample. From the central limit theorem, we know that:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (1)$$

is distributed like the normal standard distribution. We read the value of Z for 0.05 and plug in our values:

$$1.645 = \frac{\bar{X}_t - 55}{\frac{10}{\sqrt{25}}} \quad (2)$$

$$\bar{X}_t = 55 + 1.645 * \frac{10}{\sqrt{25}} = 58.29 \quad (3)$$

- (d) A type II error corresponds to not rejecting the null hypothesis when the null hypothesis is in fact, false. For the given problem, this corresponds to the genes being expressed while we do not reject the null hypothesis. Above, we decided that we need to reject the  $H_0$  for values of  $\bar{X}$  more or equal to  $\bar{X}_t$ . The type II error corresponds to the area on distribution of expressed genes that is to the left of  $\bar{X}_t$ . This can be calculated as:

$$F(\bar{X}_t) = P(Z_Y < \bar{X}_t) = P(Z_Y < 58.29) = P\left(Z_Y < \frac{\bar{X}_t - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z_Y < \frac{58.29 - 60}{\frac{10}{5}}\right) \quad (4)$$

$$P\left(Z_Y < \frac{-1.71}{2}\right) = P(Z_Y < -0.855) = 0.1963 \quad (5)$$

The table only contains values for 0.85 (0.1977) and 0.86 (0.1949). We interpolate linearly as follows: 0.855 is in the middle between 0.85 and 0.86. Its corresponding value will be in the middle between 0.1977 and 0.1949. The desired value is 0.1963.

- (e) The area is dashed in the figure above.  
(f) The power of the test is:

$$power = 1 - \beta = 1 - 0.1963 \quad (6)$$