Statistics questions with answers

*P*lease consider and work through the following questions.

1. You are screening normal and healthy GOOGLE employees. Your screening test is a Chem 20. Each test has a normal range representing the 95% confidence limits of normal values. What is the likelihood that any given patient will have at least 1 abnormal lab value on this test?

   Answer

   $1 - 0.95^{20} = 64\%$

2. You are screening young patients for an illness that does not manifest until later in life. The illness has no clinical signs or symptoms in the young but can be identified by a laboratory test. The prevalence of the illness in the population is 1/1000. The test has a 100% sensitivity and a 95% specificity. What is the likelihood that a subject with a positive test has the illness?

   Answer

   less than 2%

   On average 1000 patients will include 1 with the illness and 999 healthy patients. Since the test has 100% sensitivity, the 1 patient with the illness generates 1 positive result. Since the test has 95% specificity, the healthy subjects generate 999 * 0.05 or slightly less than 50 positive results. Out of 51 positive tests, 1 will have the illness or less than 2%.
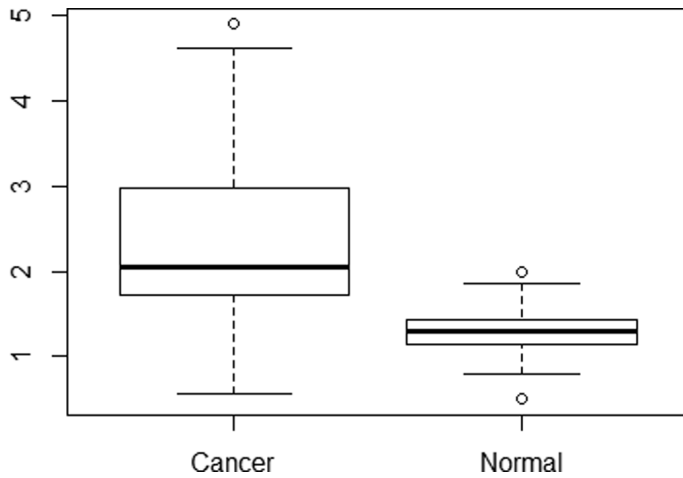
3. The BRAC1 gene expression values (normalized read count) for a group of breast cancer patients and a group of independent normal individuals are presented in the following table.

Breast cancer patients	Normal individuals
2.34	1.31
1.72	0.99
2.98	1.30
1.70	1.48
2.04	1.35
2.57	1.22
1.90	1.14
1.07	1.23
2.89	1.31
3.35	1.08
1.60	2.00
4.26	0.50
1.99	1.51
1.78	1.17
2.56	1.43
4.57	1.15
0.57	1.53
1.66	0.80
2.00	1.85
4.89	1.03
4.61	1.40

What is the representative or typical BRAC1 gene expression value for each group? Do the two groups have similar amounts of variability? What is the appropriate approach to describe the data?

Answer

Plot the data to check for potential outliers, typos, and skewness of the data. For data with symmetric distribution, mean and standard deviation can be used; otherwise, median and range can be used.

From the boxplot, we did not see any obvious outliers, and the data do not seem to be skewed. A Q-Q plot can also be generalized for visualizing data normality. The mean gene expression for the cancer patients and normal individuals are 2.5 and 1.2, respectively. The corresponding standard deviations are 1.2 and 0.33, respectively. A Bartlett's test shows that there is a significant difference in data variation between the two groups ($p < 0.001$), and a two sample t test shows that cancer patients have significant higher expression of this gene than the normal individuals ($p < 0.001$).

If you have questions or need assistance, please email Gilbert Berdine at gilbert.berdine @ttuhsc.edu or Shengping Yang at Shengping.yang@ttuhsc.edu

Submitted: 10/7/2017