My interpretation of significance testing

November 13, 2023

- Statistics is confusing as I don't think it is explained well in text books.
- Often a probability density function (PDF) is mentioned but it is rarely said what the PDF is actually of.
- The null hypothesis is mentioned usually without explicitly saying what it is. Is it a 'good' or 'bad' for the scientist's work if the null hypothesis is proved or disproved?
- Levels of significance are mentioned without interpreting what it means for the level of significance to be 'low' or 'high'.

The hypothesis and the null hypothesis

- A scientist may make a claim of a new effect and wishes to demonstrate it with measurements.
  - This is the hypothesis – that the new effect is real.
  - The null hypothesis is that there is no new effect, i.e. that the measurements behave according to existing theory. It is assumed that one can compute the PDF of making certain measurements on the basis that the null hypothesis is correct (the PDF conditioned on the null hypothesis).
- Suppose that a measurement is made, $x_{\rm m}$.
- As mentioned above suppose also that the scientist has calculated the PDF of possible measurements (Fig. 1), where the PDF is conditioned on the null hypothesis being true, $p_{\rm A}(x)$. This indicates the frequencies of observations that would be expected for each possible value according to existing (old) theory.
- The measurement, $x_{\rm m}$, provides evidence of the null hypothesis not being true if it lies in the tails of $p_{\rm A}(x)$. Assume here that $x_{\rm m}$ is made to the right of the mode of $p_{\rm A}(x)$.
  - The significance of this result is quantified by first computing the probability (according to $p_{\rm A}(x)$) that this measurement (or a higher value in the example) would be found by chance. This is estimated by computing (in the example given in the Fig.)

$$\alpha = \int_{x=x_{\rm m}}^{\infty} p_{\rm A}(x)dx. \tag{1}$$

  - It is then said that the measurement disproves the null hypothesis at the $1-\alpha$ level of significance (or $100 \times (1-\alpha)$ if one prefers percentage significance levels).
  - If $\alpha \ll 1$ (Fig. 1, left), then $x_{\rm m}$ is in the tail of the distribution, which means that the measurement is unlikely to support the null hypothesis. This statement can be made with significance level $1-\alpha$.
  - If α is not small (Fig. 1, right), then $x_{\rm m}$ is in the bulk of the distribution, which means that the measurement is likely to support the null hypothesis. The significance that the measurement is unlikely to support the null hypothesis, $1-\alpha$, is therefore small.
- This analysis obviously assumes that $x_{\rm m}$ happens to be to the right of the mode of $p_{\rm A}(x)$. If it is made to the left of the mode of $p_{\rm A}(x)$, then α is calculated instead on the basis of the following

$$\alpha = \int_{x=-\infty}^{x_{\rm m}} p_{\rm A}(x)dx \tag{2}$$

  (not shown).
- This is the one-tail significance. The two tail significance is found from

$$\alpha = \int_{x=-\infty}^{x_{\rm m}} p_{\rm A}(x)dx + \int_{x=x_{\rm m}}^{\infty} p_{\rm A}(x)dx. \tag{3}$$

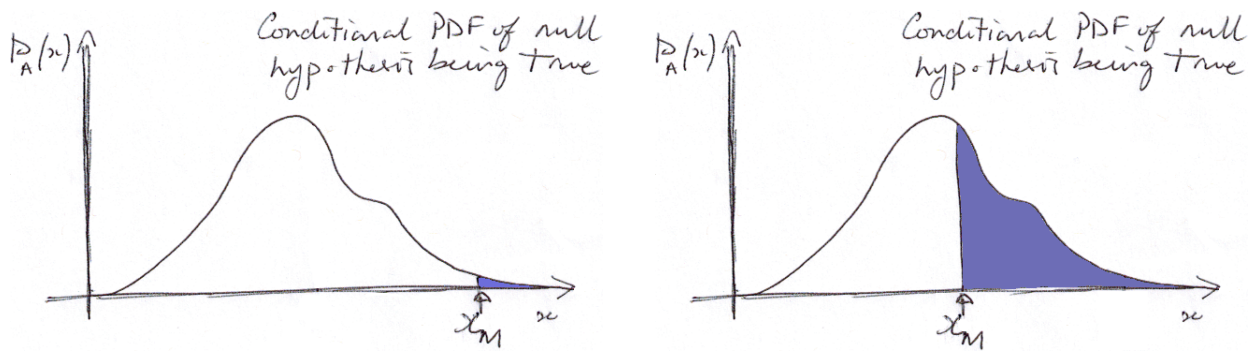

Figure 1: Conditional PDF that the null hypothesis is true with an observation in the tails (left) and in the bulk (right) of the distribution. The shading is the value of α (see Eq. (1)).

- Under what circumstances should one use the one-tail significance or the two-tail significance?
- If the conditional PDF is multi-moded, then this school of thought breaks down.

Are two distributions the same or different?

- Suppose we have two distributions, each approximated by a set of measurements. Set A of measurements is $\{x_1^{\rm A}, x_2^{\rm A}, \ldots, x_{N_{\rm A}}^{\rm A}\}$ and set B is $\{x_1^{\rm B}, x_2^{\rm B}, \ldots, x_{N_{\rm B}}^{\rm B}\}$. The PDFs can be approximated by the sum of Dirac delta functions:

$$p_{\rm A}(x) = \sum_{i=1}^{N_{\rm A}} \delta(x - x_i^{\rm A}), \text{ and } p_{\rm B}(x) = \sum_{i=1}^{N_{\rm B}} \delta(x - x_i^{\rm B}), \tag{4}$$

  (Fig. 2, top and middle).

- One can measure the 'distance' between these two distributions by comparing their cumulative density functions (CDFs). Let

$$D_{\rm m} = \max_x \left| \int_{x'=0}^{x} p_{\rm A}(x)dx - \int_{x'=0}^{x} p_{\rm B}(x)dx \right|, \tag{5}$$

  i.e. the maximum absolute difference between the CDFs, found by varying x. The value of a CDF is between 0 and 1, and so the minimum and maximum values that $D_{\rm m}$ can have are 0 and 1 respectively. If $D_{\rm m} = 0$ then the CDFs (and hence the PDFs) are measured to be identical. This is demonstrated in Fig. 2 where the top two panels are $p_{\rm A}(x)$ and $p_{\rm B}(x)$ and the bottom panel shows the two CDFs and $D_{\rm m}$.

- How statistically significant is this result? Numerical Recipes [1], Sect. 14.3 (Kolmogorov-Smirnov Test) states that the significance is found by computing the function $Q_{\rm KS}\left(\left[\sqrt{N_{\rm e}} + 0.12 + 0.11/\sqrt{N_{\rm e}}\right] D_{\rm m}\right)$, with

$$Q_{\rm KS}(\lambda) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2j^2\lambda^2\right), \tag{6}$$

$$\text{and } N_{\rm e} = \frac{N_{\rm A} N_{\rm B}}{N_{\rm A} + N_{\rm B}}. \tag{7}$$

  It states that (my comments/questions are in italics):

  - The null hypothesis is that the two distributions, $p_{\rm A}(x)$ and $p_{\rm B}(x)$, are the same.
  - $Q_{\rm KS}$ is the probability of $D > D_{\rm m}$. *This is according to what distribution – is it that calculated based on the null hypothesis being true? Let us assume this is the case.*
  - $Q_{\rm KS}$ indicates the significance that the null hypothesis is disproved.
  - Small values of $Q_{\rm KS}$ indicates that the two PDFs $p_{\rm A}(x)$ and $p_{\rm B}(x)$ are different with high significance.

- Let us work with these to try to understand what is meant by these statements.

  - Figure 3 shows two $Q_{\rm KS}$ functions calculated from (6) and (7), and the associated PDFs. The left is for a small population and the right is for a larger population.
  - As the population increases, the mode is shifted to smaller values of D, and the variance decreases.

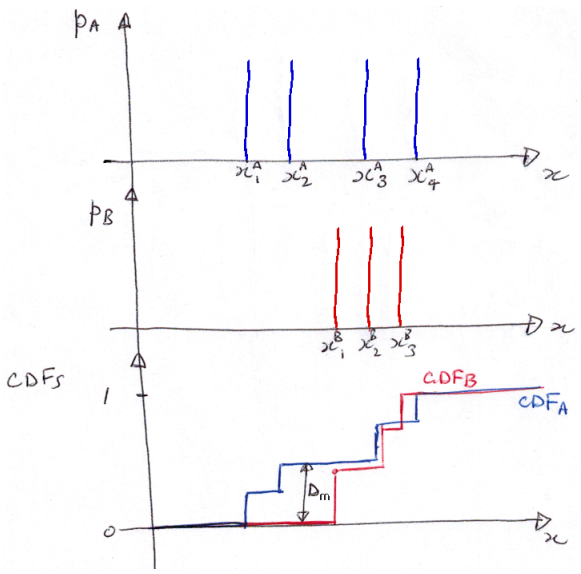

Figure 2: Measurement set A approximates the $p_{\mathrm{A}}(x)$ (top), and set B approximates the $p_{\mathrm{B}}(x)$ (middle) each by a series of Dirac delta functions, (4). The two PDFs are compared by computing the respective CDFs (bottom) and taking the maximum absolute value of the difference between them, D_{m}.

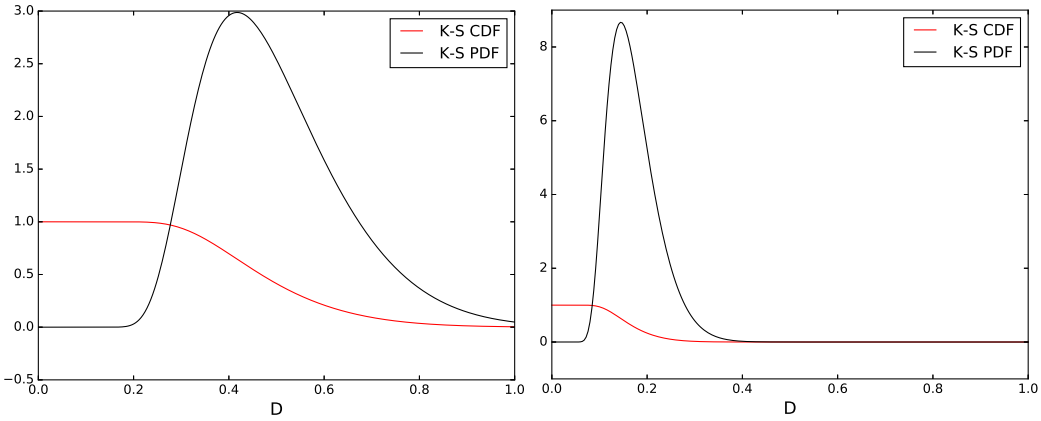

Figure 3: Two evaluations of the Q_{KS} function (red) and the PDF implied from it (minus the gradient of Q_{KS} with respect to D, black). The left is calculated based on a small population $N_{\mathrm{A}} = N_{\mathrm{B}} = 5$ (left) and a larger population $N_{\mathrm{A}} = N_{\mathrm{B}} = 50$ (right).

- Assume that what is meant by the "probability of $D > D_{\mathrm{m}}$" mentioned above is calculated from the PDF conditioned on the null hypothesis being true.
  - Assume we always measure $D_{\mathrm{m}} >$ mode. The larger the population, for such a particular D_{m}, the smaller the value of Q_{KS}, and the higher the significance that the null hypothesis is rejected (i.e. that the two distributions, $p_{\mathrm{A}}(x)$ and $p_{\mathrm{B}}(x)$ are different). The significance level therefore must be $1 - Q_{\mathrm{KS}}$.
  - Problem: for the PDFs shown in Fig. 3 are meant to be conditioned on the null hypothesis being true, then the PDFs should surely peak at $D = 0$? The actual PDFs plotted suggest that it is virtually impossible to make sets of measurements that have $D \ll 1$. Perhaps the distribution given by (6) for the K-S test is valid only for $D_{\mathrm{m}} >$ mode?
  - Problem: surely the PDF should depend (at least) on the variances of the distributions? Currently it depends on the sizes of the populations of the data only.

RNB

References

[1] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2007.