Lecture 7

Hypothesis testing:-

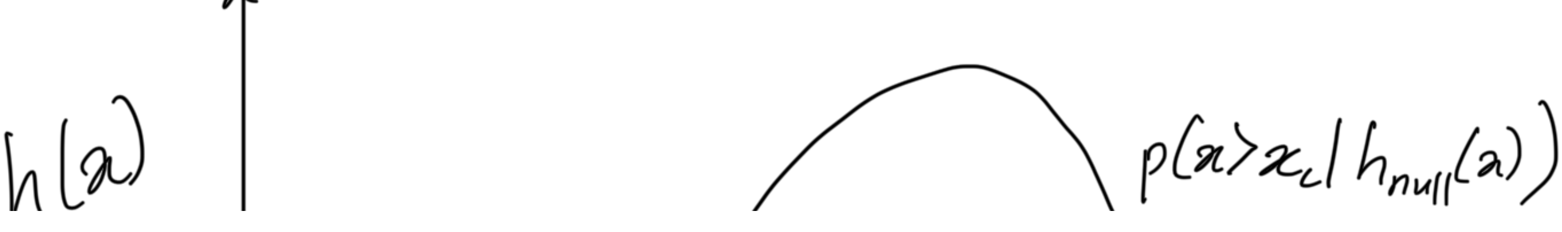
- There are 2 key questions any time we are trying to "detect" something:

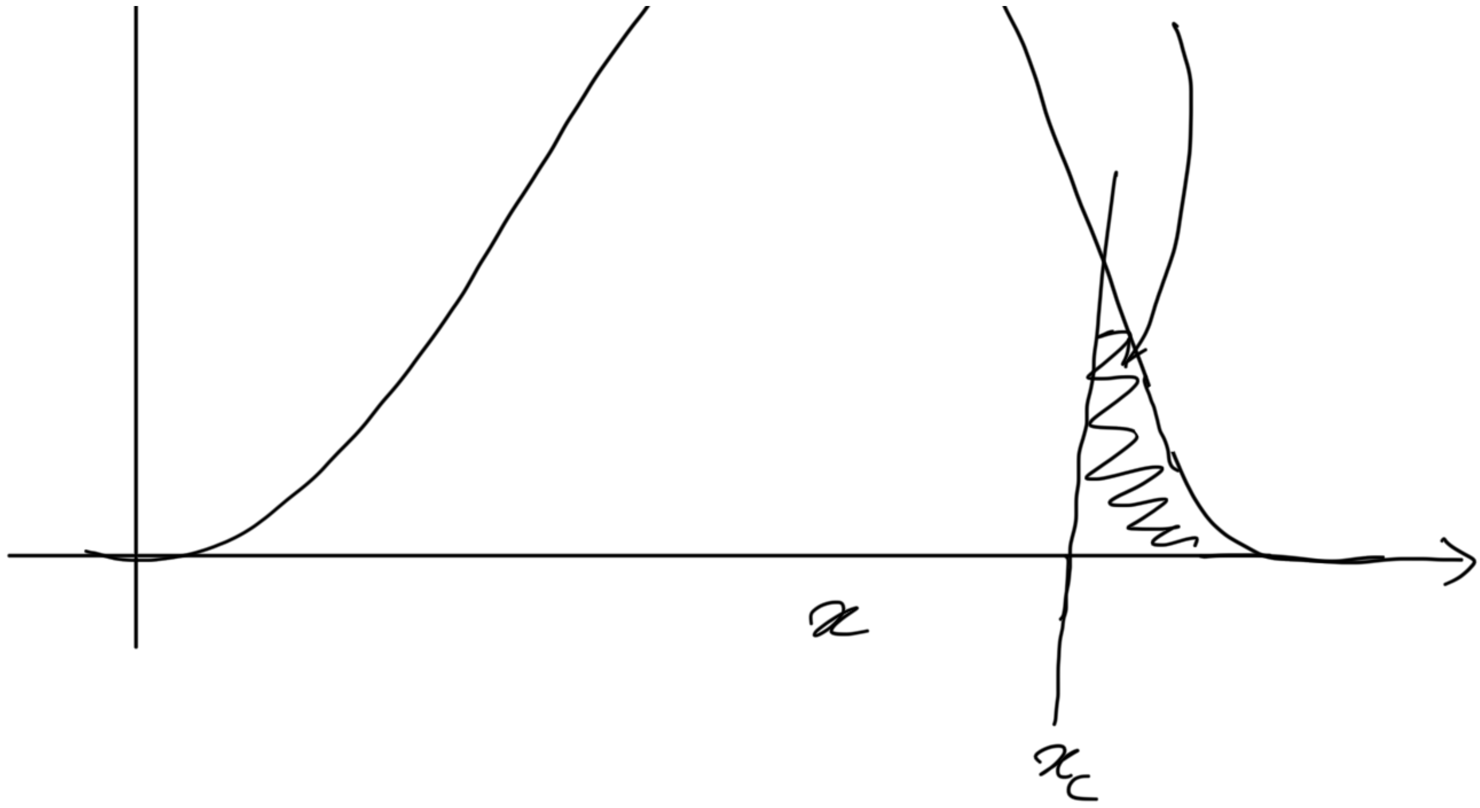
① With my data, can I reject a null (noise/background) hypothesis?
② Is my data consistent with my signal hypothesis?

- An incorrect conclusion on Qn. ①, i.e. rejection of null hypothesis when it is inappropriate/not true, is called a Type I error, aka False Positives.

- An incorrect conclusion on Qn. ②, i.e. an acceptance of the null hypothesis when it is not true is a Type II error, i.e. False Negative.

- It is generally assumed that we can calculate any event probability under our null distribution.





- The p-value for the observed data/statistic is

$$\boxed{p(x > x_c) = 1 - H(x_c)}$$

where $H(x_c)$ is the CDF of the null distribution at value x_c.

- The p-value quantifies how "unlikely" the observation is under the null hypothesis.

- Typically, a threshold p-value, called the significance level α, is adopted.

- The null hypothesis can then be rejected when p-value $< \alpha$.

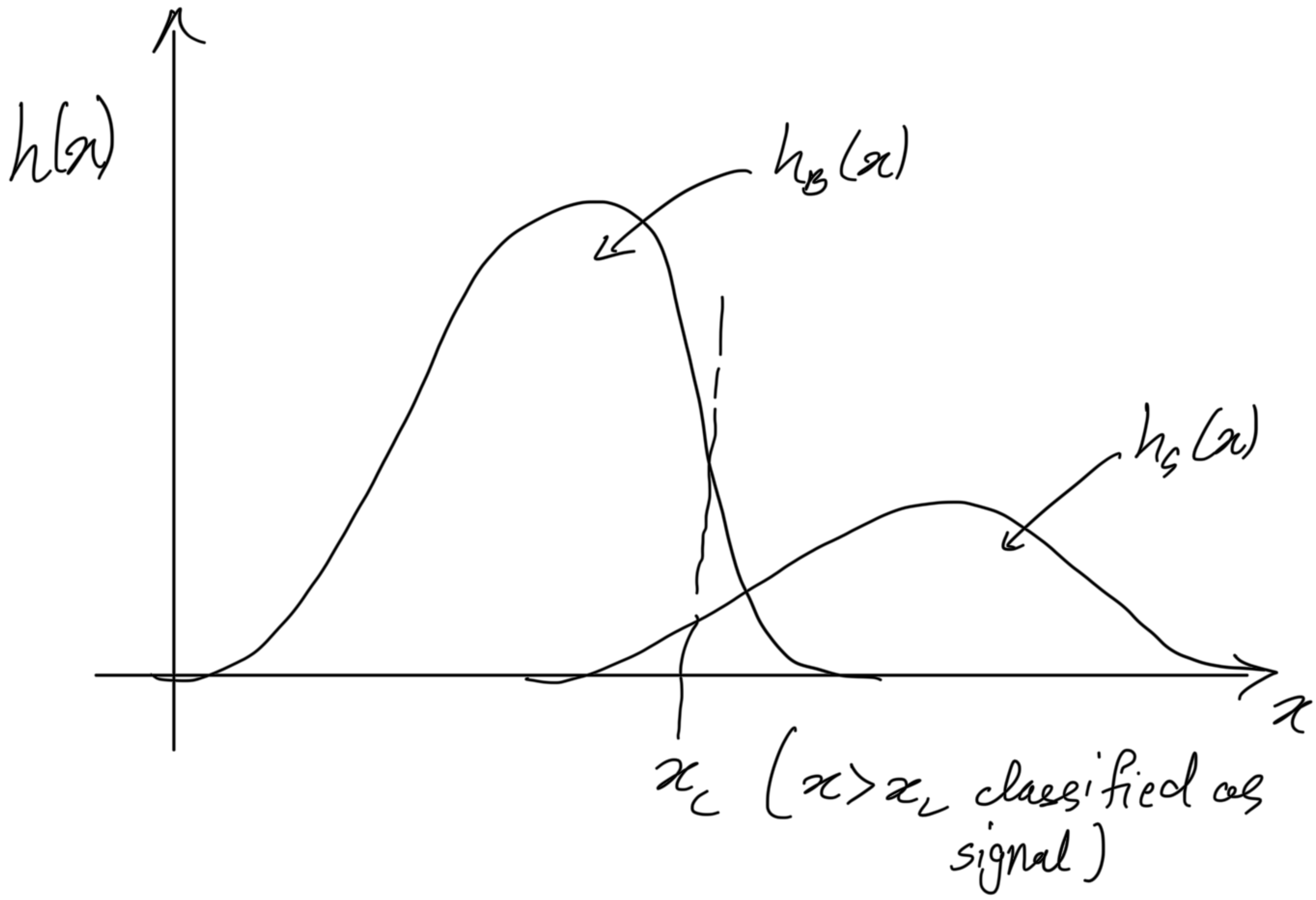
* Note that if we fail to reject a hypothesis, it does not mean that we proved its correctness, since our sample may not be large enough to detect an effect.

- Type I errors are thus limited by the
adopted significance level α

adopted significance level α.

- Type II errors are characterised by a probability, β, which is related to the "power of a test" as $(1-\beta)$.

- However, we typically do not have data from just the null hypothesis, $h_B(x)$, but also from some signal hypothesis, $h_S(x)$.



- The net effect is that there could be more Type I (false positives) errors than α would suggest.

- The above distribution of data points can be written as,

$$h(x) = (1-a)\,h_B(x) + a\,h_S(x)$$

where "a" is a relative normalization factor, quantifying how "more frequent" the source/signal is relative to the background.

- Thus, one can think of this as a classification problem, assigning data to $h_B(x)$, with probability $p_B(x)$, or $h_S(x)$, with probability $p_S(x)$.
- Note,

$$p_B(x_i) = 1 - p_S(x_i)$$

since we only have two classes.

Comparison of distributions

- We often want to know if two sets of measurements are drawn from the same distribution.
eg: are my observed AGN consistent with quasars or Seyferts?

- The idea of this statistical test is to use the data to compute an appropriate statistic & compare it to its expected value.
- The expected distribution is evaluated by assuming the null distribution is true.

- If the data-based p-value for the statistic is small, the null hypothesis is rejected with some significance, α.

- Lots of tests in literature to do this comparison of distributions.
  - See notebook for details.

Nonparametric modeling & Histograms

- When there is no strong justification for adopting a parametrized description (i.e. an analytical function with free parameters), nonparametric methods offer an alternative approach.

- Note: nonparametric does not mean there are no parameters, just that there is a general set of distribution-free models, that satisfy some smoothness condition.
  - An example of such a condition is the Sobolev space,

$$\int [h''(x)]^2 dx < \infty$$

i.e. no functions with infinite spikes.

- These non-parametric models play a central role in machine learning.
  - They provide the highest possible accuracy, as they can model any shape of the distribution
  - But, they come at a higher computational cost.

- The most common examples of nonparametric distributions are histograms & Kernel Density

distributions are histograms & Kernel Density
Estimators (KDE).

- See notebook for more details.