Class 0x0C: Hypothesis testing and significance tests

Different cases of hypothesis and significance testing

- Comparison of two hypotheses H_0 and H_1:
  - The classic "simple hypothesis test".
  - Characterized by the *significance level* of the test.
- Comparison of hypotheses in some model for various values of some unknown parameters of the model:
  - Confidence intervals or confidence regions.
  - Characterized by the *confidence level* of the region.
- Consistency of a single hypotheses H with data:
  - Goodness-of-fit test or significance test.
  - Returns a so-called "p-value" which, if small, can be interpreted as the *significance* of any disagreement.

Simple hypothesis tests

Test statistic: one or more statistic(s) t, a function of the observations x.

Critical region: a region defined by some limits on the test statistic(s). This is the "rejection region" for H_0.

Significance level α of a critical region: Probability for a result to be in critical region if H_0 is true. ("Type I error")

False negative probability β of a critical region: Probability for a result to be outside critical region if H_1 is true. ("Type II error") $1 - \beta$ is called the statistical power of the test.

- The critical region is more generally defined as a region in the full space of measurements x. This is equivalent to the definition here in the extreme case $t = x$.

- The "reject/accept" terminology is just terminology: "reject H_0" just means "in the critical region".
- H_0 is also called the "null hypothesis". A result outside the critical region is also called "negative", a result inside is called "positive". (Think of a medical test where H_0 is "healthy", H_1 is a diagnosis of a particular disease.)
- α is a probability depending only on the hypothesis H_0 and the critical region, not on any measurements. It is not a random variable or an observable as such. Similarly, β depends only on H_1 and the critical region.

Constructing a good test statistic

- We can't simultaneously minimize α and β.
- We can fix α and find the test that minimizes β.
- The Neyman-Pearson lemma says that a test that always achieves the lowest possible β for a given α has a critical region R of the following form:
$$R = \underline{x} : L_x(H_0) \leq kL_x(H_1)$$
That is, the critical region is defined by the region where the likelihood of the observation $\underline{x}$ assuming hypothesis H_0 is not greater than k times the likelihood assuming hypothesis H_1.
- Written another way, the critical region is defined by the test statistic
$$t = \frac{L_x(H_0)}{L_x(H_1)},$$
and the critical region is defined by $t \leq k$.
- There is a 1-1 mapping between k and α. Choose the k that gives the α you want. (Obviously, you need to know the p.d.f.s of both hypotheses to do this.)

Example 1: unrealistic light bulb models

Suppose we have one model H_0 that the p.d.f. for light bulb lifetime T is given by
$$\frac{dP}{dT} = f_0(T) = \mu_0^{-1} e^{-T/\mu_0}$$
for some known μ_0, and another model H_1 which is the same except that the mean is μ_1, also known.

What's the Neyman-Pearson test statistic?

$$t = \frac{\prod_i f_0(T_i)}{\prod_i f_1(T_i)}.$$

Since we're just going to compare it to a value k, we can just as well use

$$\log t = \sum_i \log f_0(T_i) - \sum_i \log f_1(T_i).$$

In this case, this is simply

$$\log t = N \log(\mu_1/\mu_0) + (\mu_1^{-1} - \mu_0^{-1}) \sum T_i.$$

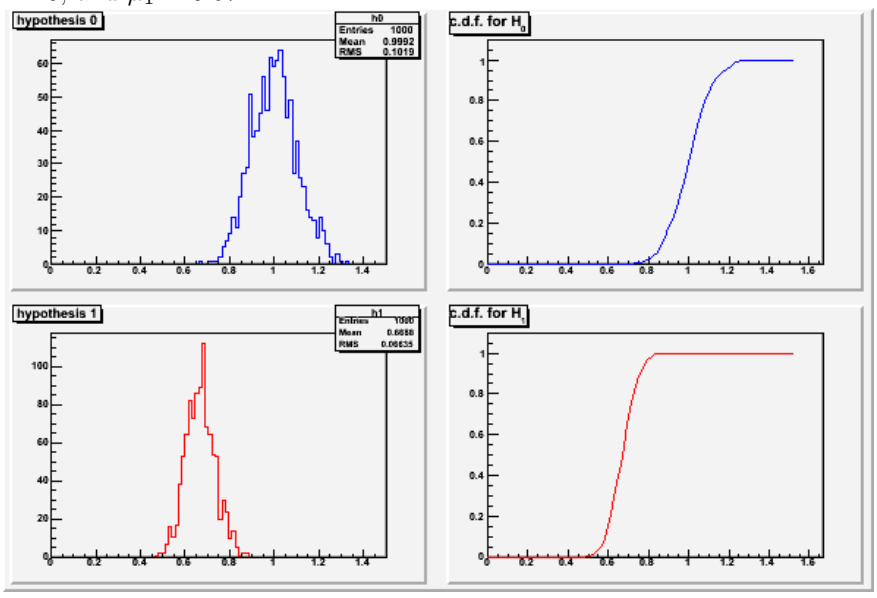
Let's take $\mu_1 < \mu_0$. The critical region defined by $\log t \leq k$ can be rewritten as

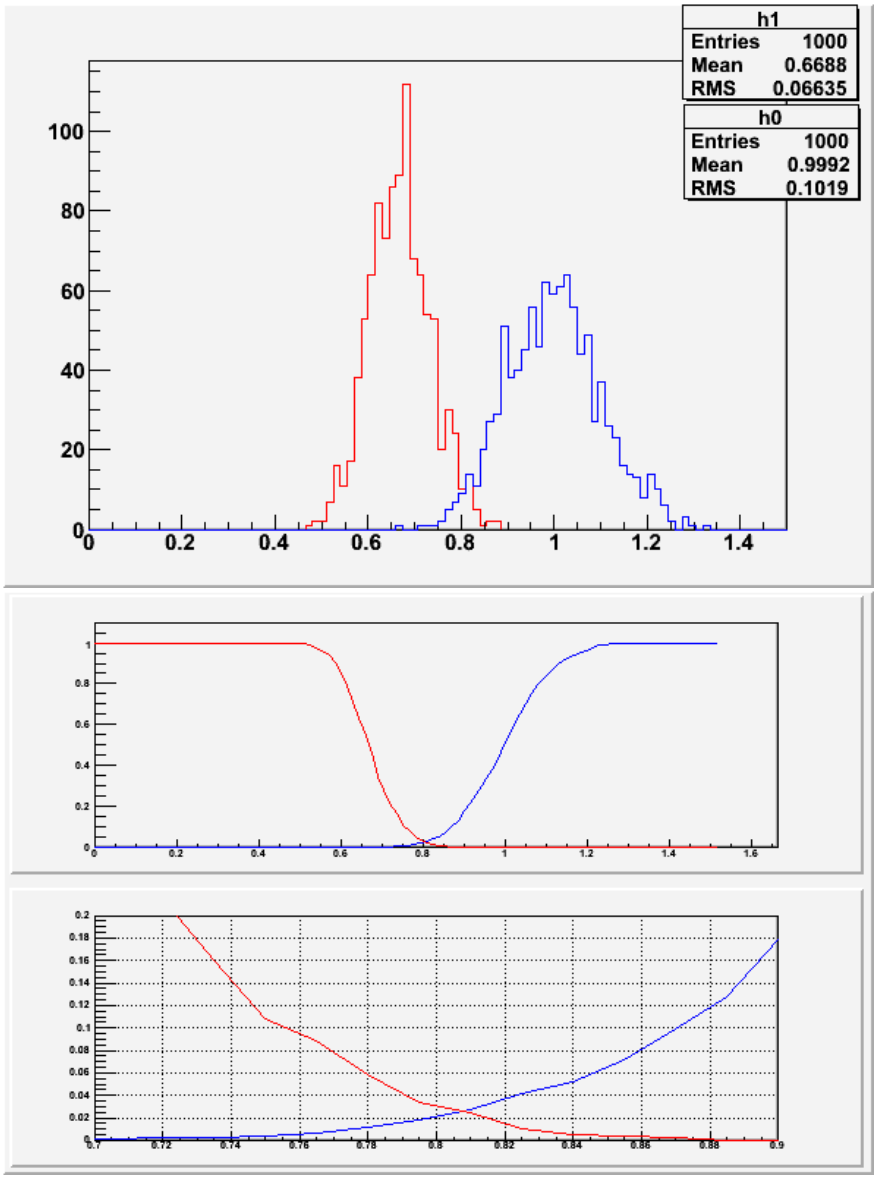
$$\frac{1}{N} \sum T_i \leq \frac{N^{-1} \log k - \log(\mu_1/\mu_0)}{\mu_1^{-1} - \mu_0^{-1}}.$$

Stated in words: "reject" the larger lifetime hypothesis if the observed mean is smaller than some amount. Adjust that amount to get the desired α. This can be done assuming a gaussian distribution for the mean if N is large; otherwise, evaluate it analytically or using MC methods.

Histograms of the test statistics from example 1

These plots were made with the code in `classCh_example1.cc` with $N = 100$, $\mu_0 = 1.0$, and $\mu_1 = 0.67$.





For any given value x chosen as the decision criteria on the test statistic, the value of α is given by the c.d.f. of the test statistic assuming H_0 (red curve above), and β is given by 1-c.d.f. of the test statistic assume H_1 (blue curve above). Generally one chooses α first and then finds the necessary value of x for the decision. The Neyman-Pearson lemma says that β is as low as it can be for that α. Note there is in general no particular advantage to setting $\alpha = \beta$, although there may be reason to do so in some cases.

Example 2: More realistic light bulb models

Suppose we have one model H_0 that the p.d.f. for light bulb lifetime T is given by

$$\frac{dP}{dT} = f_0(T) = \mu^{-1} e^{-T/\mu}$$

for some unknown μ, and another model H_1

$$\frac{dP}{dT} = f_1(T) = \begin{cases} (h - |T-\mu|)/h^2 & \text{if } |T-\mu| < h, \\ 0 & \text{otherwise} \end{cases},$$

for some unknown μ and h. Construct the test statistic as before, and compare the best fit for H_0 to the best fit for H_1.

Here you might want to evaluate the significance levels α for a given k using a MC simulation.

Example 3: applying a signal/background cut

(... see discussion in hypothesis test section of [PDG-Stat] ...)

Two-hypothesis significance test

- You don't *have* to decide in advance at what significance level you will accept or reject a hypothesis.
- Depending on what you are doing, it may not even be appropriate to do so. It might be more appropriate to report the *significance* of the result: the value α' such that the observed data would be in the critical region for $\alpha > \alpha'$, out of the critical region of $\alpha < \alpha'$.

  > For example, you might be investigating a specific alternative to Einstein's theory of general relativity in light of some new data. Rather than report just "hypothesis accepted" or "hypothesis rejected" according to your personal, pre-chosen α, the world would like to know what α' is. Then every person can know, for her/his *own* personal α, whether they want to accept or reject the null hypothesis.

- Despite the fact that the significance is often reported as a percentage, it is a random variable. It is definitely *not* the probability the hypothesis is *really* right or wrong.

Why the statistical significance isn't the probability you'd like

- The significance level, α, is a number you (or someone) chooses. You adjust your test so it has that probability of giving you a false positive (type 1 error), on average, over many data sets.
- α' is random variable determined by one measurement or set of measurements, numerically equal to $P(t \geq t'|H_0)$, where t' is the value of the test statistic corresponding to the α significance level.
- What you often most want is $P(H_0|\alpha')$, the probability that the null hypothesis is true given one measurement or set of measurements. Bayes' theorem tells us

$$P(H_0|\alpha') = \frac{P(\alpha'|H_0)P(H_0)}{P(\alpha')},$$

  if the truth/falseness of H_0 is itself a random variable. This might be possible in the case of a medical diagnosis, but not for a law of nature.

  - In the case of medicine, we can (perhaps) know all three probabilities or p.d.f.s for a very well studied diagnostic test:
    - the p.d.f. of diagnosis significances for healthy people $P(\alpha'|H_0)$;
    - the p.d.f. of the diagnosis significance for the entire population $P(\alpha')$, and
    - the probability of someone in the population being healthy $P(H_0)$.
  - In the case of a physical law, we have only one universe to observe, so this becomes meaningless. Even if we adopt the many universes idea, we have no way of knowing $P(H_0)$ over the many universes.*

Goodness-of-fit or one-hypothesis significance test

Again, we have a test statistic, which I'll call T.

- The value of T should reflect how compatible the data is with the hypothesis. (E.g., higher values of χ^2 indicate less compatibility.)
- We should be able to derive the probability $P_T(t) = \text{Prob}(T \leq t)$ for any hypothesis.
- Examples of statistics for which $P_T(t)$ is well known include the χ^2 for gaussian data and the Kolmogorov-Smirnov statistic for histogram data. (See discussion in *Numerical Recipes* [NumRecip].)

* See Comment on Bayesian statistics.

- The log-likelihood $\log L$ is also used, although it requires derivation or numerical simulation to determine $P_T(t)$.
- As far as I know, there is no equivalent to the Neyman-Pearson theorem for this kind of test, probably because of the lack of an alternative hypothesis with which to compare. There's no universal, general purpose "best approach".
  - For example, a test based on $\log L$ of the measurements might miss an inconsistency that is readily apparent in a comparison of the histogram of the data points with the expectation values from the model. (An example of this in assignment "option B" at the end of this lecture.)

The *p*-value

The p-value is what the hypothetical model H says should be the probability to find the statistic T in a region of equal or lesser compatibility than the observed t: that is, $p = P_T(t)$ assuming H is true.

- If H is in fact true, p will be a random variable with *uniform* distribution between 0 and 1.[†]
- If H is "significantly wrong", then p will be a small number.
- The value of $1 - p$ is often reported as the *significance* with which a hypothesis has been rejected. (For an example, see the claimed rejection of the no-oscillation hypothesis in the abstracts of KamLAND2004 and KamLAND2002.)

Why this statistical significance isn't the probability you'd like (II)

- The significance $1 - p$ of a single-hypothesis significance test is not the same thing as the *signifiance level* α previously defined, but it is closely related to the statistical significance α' of a two-hypothesis test. Similar comments apply.
- Because p is a uniform random variable when H is true, if you reject hypotheses every time they have a p-value less than some personal threshold α, you'll eventually end up rejecting a fraction α of whatever true hypotheses you examined.

[†] The proof is the reverse of the derivation of the inverse-distribution-function method of generating a random variable.

- You'll also end up rejecting a lot of false hypotheses, but the probability for that will depend on what the true model really is, which this kind of test doesn't consider.
- It's not reasonable to use p to select hypotheses to accept, since p can take on any value from 0 to 1 with equal probability when H is true.

Example 4: exponential plus background

How good is the fit of the exponential-plus-background model to the data in the last assignment? Let's use the best-fit likelihood $L_{\max}$ as our test statistic. We'll get the p.d.f. for $L_{\max}$ using MC simulation.

Procedure for building up p.d.f. of $L_{\max}$:

    make a histogram to store the p.d.f.
    loop M times:
        simulate a dataset using the hypothesis
        fit the dataset
        "fill" histogram using $L_{\max}$

Procedure for simulating a dataset:

    Loop N times:
        generate random variable x according to the model p.d.f. for x
            (see class notes on MC simulation, use inverse distribution method)
        store x in vector of doubles to be used as dataset
            (instead of reading x from a file)

Now just read off the p-value from this histogram: according to the simulation, if the hypothesis is true, what fraction of $L_{\max}$ would be worse than what you got for the actual data?

Assignment

Choose *either* "option A" *or* "option B" below -- you do not have to do both.

Option A: Complete example 4 above.

Option B: See below.

Assignment option B: globular clusters

Are the 119 globular clusters in the Arp 1965 catalog uniformly distributed in $\cos\theta$, where θ is galactic latitude?

- The likelihood of the $\cos\theta$ values is not a good choice in this case: the hypothesized distribution is uniform, $L_{\cos\theta}$ doesn't depend on the data values, it's always $N\log(1/2)$!
- Instead, try using the likelihood of the θ values: $dP/d\theta = \sin\theta/2$.
- Make a simple simulation along the lines of the exercise:
  - Generate 1000 simulated data sets of 119 clusters distributed uniformly in $\cos\theta$. Transform to θ.
  - Accumulate histogram of $\log L_\theta$.
- Download the data from Vizier copy of Arp 1965. (Suggestion: download in tab-separated-value format.)
- Read galactic longitude from file, calculate $\log L_\theta$ for this data set.
- What fraction of simulated data sets have less consistent data? (Lower values of $\log L$.) If there are almost none, then the hypothesis can be rejected with some significance.
- Option B' (very optional): instead of (or in addition to) using the likelihood as the goodness-of-fit statistic, try using the Kolmogorov-Smirnov test (see [NumRecip]) or some other test.

Comment on Bayesian statistics

Even though $P(H_0|\text{observation})$ is meaningless as a probability in the sense of "the fraction of possible universes in which H_0 would turn out to be true given that we made these observations", some people like to use Bayes' law anyway to characterize and update what they call their "subjective degree of belief". Rather than using objective data for $P(H_0)$, they use that term in Bayes' law to reflect their "prior subjective beliefs" ("priors" for short), deliberately introducing this as something that can only be changed by statistically significant evidence to the contrary. This approach has caused a lot of controversy over the years. In my opinion, there is nothing wrong with this as long as one evaluates the resulting p.d.f.s as carefully as possible and keeps in mind the limitations. However, it can go badly wrong if the "prior" pre-assigns very low probability to what the data actually ends up indicating, even if the "prior" is based on little or no relevant data: for an extreme case, see "The Logic of Intelligence Failure" by Bruce G. Blair [Blair2004], actually written by a proponent of this way of thinking. I won't talk about that further today.

References

In the following, (R) indicates a review, (I) indicates an introductory text, and (A) indicates an advanced text.

Probability:

PDG-Prob: (R) "Probability", G. Cowan, in *Review of Particle Physics*, C. Amsler et al., PL B667, 1 (2008) and 2009 partial update for the 2010 edition (http://pdg.lbl.gov).

See also general references cited in PDG-Prob.

Statistics:

PDG-Stat: (R) "Statistics", G. Cowan, in *Review of Particle Physics*, C. Amsler et al., PL B667, 1 (2008) and 2009 partial update for the 2010 edition (http://pdg.lbl.gov).

See also general references cited in PDG-Stat.

Larson: (I) *Introduction to Probability Theory and Statistical Inference*, 3rd ed., H.J. Larson, Wiley (1982).

NumRecip: (A) *Numerical Recipes*, W.H. Press, *et al.*, Cambridge University Press (2007).

Other cited works:

Blair2004: B.G. Blair, "The Logic of Intelligence Failure", Forum on Physics and Society Newsletter, April, 2004; http://www.aps.org/units/fps/newsletters/2004/april/article3.html browsed 2010/06/01.

KamLAND2002: KamLAND collaboration, Phys.Rev.Lett.90:021802,2003; arXiv:hep-ex/0212021.

KamLAND2004: KamLAND collaboration, Phys.Rev.Lett.94:081801,2005; arXiv:hep-ex/0406035.