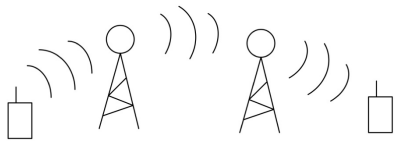
Decision making with uncertainty

Rebecca Willett, 2017

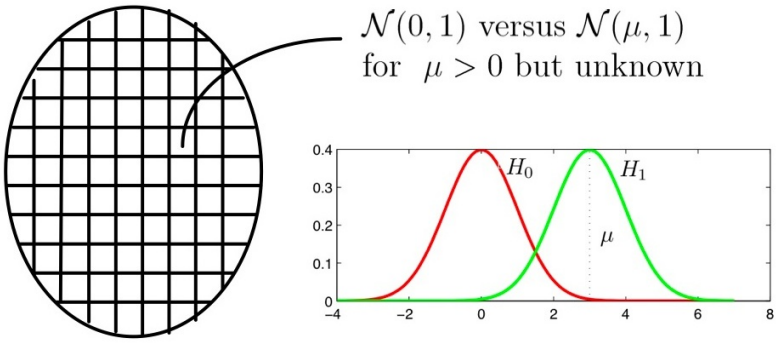
In many real world problems, it is difficult to precisely specify probability distributions. Our models for data may involve unknown parameters or other characteristics. Here are a few motivating examples.

Example: Unknown amplitudes/delays in wireless communications.

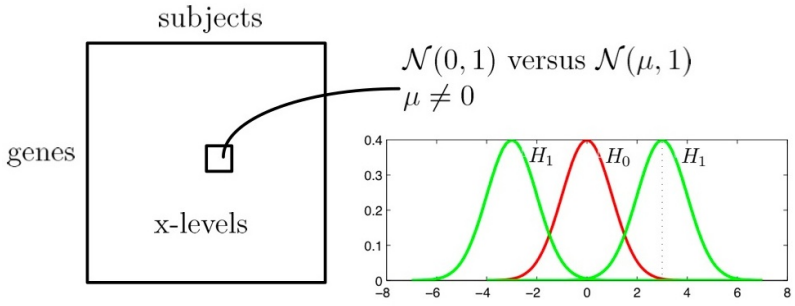


We don't always know how many relays a signal will go through, how strong the signal will be at each receiver, the distance between relay stations, etc.

Example: Unknown signal amplitudes in functional brain imaging.



Example: Unknown expression levels in gene microarray experiments.



1 Composite Hypothesis Tests

We can represent uncertainty by specifying a collection of possible models for each hypothesis. The collections are indexed by a parameter.

$$\begin{aligned} H_0 : X &\sim p_0(x|\theta_0),\ \theta_0 \in \Theta_0 \\ H_1 : X &\sim p_1(x|\theta_1),\ \theta_1 \in \Theta_1 \end{aligned}$$

- In general, the distributions p_0 and p_1 may have different parametric forms.
- The sets Θ_0 and Θ_1 represent the possible values for the parameters.
- If a set contains a single element (i.e., a single value for the parameter), then we have a simple hypothesis, as discussed in past lectures. When a set contains more than one parameter value, then the hypothesis is called a composite hypothesis, because it involves more than one model.

The name is even clearer if we consider the following equivalent expression for the hypotheses above.

$$\begin{aligned} H_0 : X &\sim p_0,\ p_0 \in \{p_0(x|\theta_0)\}_{\theta_0 \in \Theta_0} \\ H_1 : X &\sim p_1,\ p_1 \in \{p_1(x|\theta_1)\}_{\theta_1 \in \Theta_1} \end{aligned}$$

Example: Brain imaging

Recall the brain imaging problem.

$$\begin{aligned} &H_0 : X \sim \mathcal{N}(0,1) \\ &H_1 : X \sim \mathcal{N}(\mu,1),\ \mu > 0 \text{ but otherwise unknown} \\ &\qquad \text{equivalently } X \sim p, p \in \{\mathcal{N}(\mu,1)\}_{\mu>0} \end{aligned}$$

In this example, H_0 is simple and H_1 is composite.

2 Uniformly Most Powerful Tests

Let us begin by considering special cases in which the usual likelihood ratio test is computable and optimal. Here is an example.

$$\begin{aligned} H_0 :&\quad x_1, \cdots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ H_1 :&\quad x_1, \cdots, x_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu,1),\ \mu > 0 \end{aligned}$$

Log LRT:

$$\begin{aligned} \log \left(\frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i)^2/2}} \right) &= \sum_{i=1}^n -\frac{(x_i-\mu)^2}{2} + \frac{x_i^2}{2} \\ &= \mu \sum_{i=1}^n x_i - \frac{n\mu^2}{2} \end{aligned}$$

Test statistic:

$$\mu \sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \iff \sum_{i=1}^{n} x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'/\mu = \gamma$$

We were able to divide both sides by μ since $\mu > 0$. We do not need to know the exact value of μ in order to compute the test $\sum_i x_i \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$ for any value of γ.

Let $t = \sum_{i=1}^{n} x_i$ denote the test statistic. It is easy to determine its distribution(s) under each hypothesis (a composite in the case of H_1).

$$\begin{aligned} H_0 : &\quad t \sim \mathcal{N}(0, n) \\ H_1 : &\quad t \sim \mathcal{N}(n\mu, n) \quad \mu > 0 \text{ unknown} \end{aligned}$$

Since distribution of t under H_0 is known, we can choose threshold to control P_{FA}.

$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{n}}\right) \Rightarrow \gamma = \sqrt{n} Q^{-1}\left(P_{FA}\right)$$

This is optimal detector (most powerful) according to NP lemma. Several ROC curves corresponding to different values of the unknown parameter $\mu > 0$ are depicted below. We cannot know which curve we are operating on, but we can choose a threshold for a desired P_{FA} and the resulting P_D is the best possible (for the unknown value of μ). In such cases we say that the test is uniformly most powerful, that is most powerful no matter what the value of the unknown parameter.

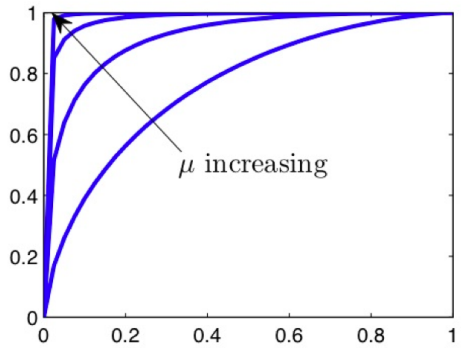


Figure 1: ROC for various $\mu > 0$ for the simple case.

Definition: Uniformly Most Powerful Test

A uniformly most powerful (UMP) test is a hypothesis test which has the greatest power (i.e. greatest probability of detection) among all possible tests yielding a given false alarm rate regardless of the underlying true parameter(s).

3 Two-sided Tests

To see how special the UMP condition is, consider the following simple generalization of the testing problems above.

$$\begin{aligned} H_0: &\quad x \sim \mathcal{N}(0,1) \\ H_1: &\quad x \sim \mathcal{N}(\mu,1),\ \mu \neq 0 \end{aligned}$$

The log-likelihood ratio statistic is

$$\log \Lambda(x) \;=\; -\frac{(x-\mu)^2}{2} + \frac{x^2}{2} = \mu x - \mu^2/2$$

and the log-LRT has the form

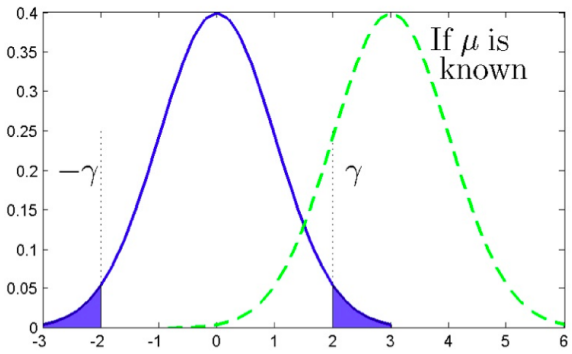
$$\mu\, x - \mu^2/2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \ .$$

We can move the term $\mu^2/2$ to the other side and absorb it into the threshold, but this leaves us with a test of the form

$$\mu\, x \underset{H_0}{\overset{H_1}{\gtrless}} \gamma.$$

Since μ is unknown (and not necessarily positive) the test is uncomputable.

How can we proceed? Look at two densities in the microarray experiment. Intuitively the test $|x| \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$ seems reasonable. This is called the Wald Test. The P_{FA} of the Wald test can be seen below.



$$\begin{aligned} P_{FA} &= 2Q(\gamma) \Rightarrow \gamma = Q^{-1}(P_{FA}/2) \\ P_D &= \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx + \int_{-\infty}^{-\gamma} \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} dx \quad y = x - \mu \\ &= \int_{\gamma-\mu}^{\infty} \mathcal{N}(0,1) dy + \int_{-\infty}^{-\gamma-\mu} \mathcal{N}(0,1) dy \\ &= Q(\gamma-\mu) + Q(\gamma+\mu)\ . \end{aligned}$$

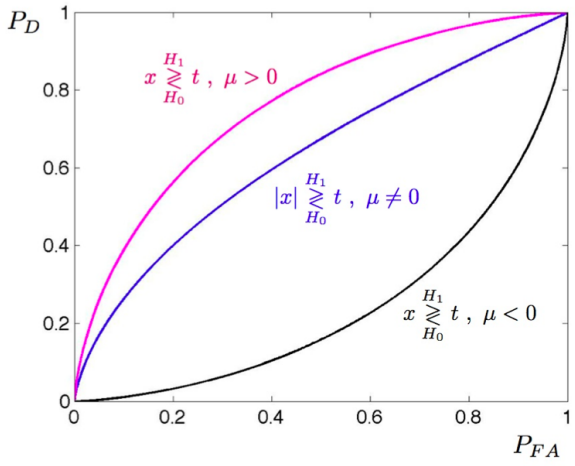


Figure 2:

The P_D depends on μ, which is unknown.

Model μ as a deterministic, but unknown, parameter. Estimate μ from the data and plug the estimate into the LRT. Under H_1 the distribution is $X \sim \mathcal{N}(\mu, 1)$, so a natural estimate for μ is $\widehat{\mu} = x$, the observation itself. The plugging this into the likelihood ratio yields

$$\widehat{\Lambda}(x) \;=\; \frac{p(x|\widehat{\mu})}{p(x|0)} \;=\; \frac{\exp(-(x-\widehat{\mu})^2/2)}{\exp(-x^2/2)} \;=\; e^{x^2/2}\ .$$

This is the generalized likelihood ratio. In effect, this compares the best fitting model in the composite hypothesis H_1 with the model H_0. Taking the log yields the test

$$\log \widehat{\Lambda}(x) \;=\; x^2 \;\gtrless_{H_0}^{H_1}\; \gamma\ ,$$

which is equivalent to the Wald test.

4 The Generalized Likelihood Ratio Test (GLRT)

Consider a composite hypothesis test of the form

$$\begin{aligned} H_0 : X &\sim p_0(x|\theta_0),\ \theta_0 \in \Theta_0 \\ H_1 : X &\sim p_1(x|\theta_1),\ \theta_1 \in \Theta_1 \end{aligned}$$

The parametric densities p_0 and p_1 need not have the same form.

The generalized likelihood ratio test (GLRT) is a general procedure for composite testing problems. The basic idea is to compare the best model in class H_1 to the best in H_0, which is formalized as follows.

Definition: Generalized Likelihood Ratio Test (GLRT)

The GLRT based on an observation x of X is

$$\widehat{\Lambda}(x) = \frac{\max\limits_{\theta_1 \in \Theta_1} p_1(x|\theta_1)}{\max\limits_{\theta_0 \in \Theta_0} p_0(x|\theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma,$$

or equivalently

$$\log \widehat{\Lambda}(x) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma.$$

5 Wilks' Theorem

Wilk's Theorem (1938)

Consider a composite hypothesis testing problem

$$\begin{aligned} H_0 \quad &: \quad X_1, X_2, ..., X_n \overset{iid}{\sim} p(x|\theta_0), \\ &\quad \text{where } \theta_{0,1}, \ldots, \theta_{0,\ell} \in \mathbb{R} \text{ are free parameters and} \\ &\quad \theta_{0,\ell+1} = a_{\ell+1}, \ldots, \theta_k = a_k \text{ are fixed at the values} \\ &\quad a_{\ell+1}, \ldots, a_k \\ H_1 \quad &: \quad X_1, X_2, ..., X_n \overset{iid}{\sim} p(x|\theta_1), \theta_1 \in \mathbb{R}^k \text{are all free parameters} \end{aligned}$$

and the parametric density has the same form in each hypothesis.

In this case family of models in H_0 is a subset of those in H_1, and we say that the hypotheses are nested. (This is a key condition that must hold for this theorem.)

If the 1^{st} and 2^{nd} order derivatives of $p(x|\theta_i)$ with respect to θ_i exist and if $\mathbb{E}\left[\frac{\partial \log p(x|\theta_i)}{\partial \theta_i}\right] = 0$ (which guarantees that the MLE $\widehat{\theta}_i \to \theta_i$ as $n \to \infty$), then the generalized likelihood ratio statistic, based on an observation $X = (X_1, \ldots, X_n)$,

$$\widehat{\Lambda}_n(X) = \frac{\max\limits_{\theta_1} p(x|\theta_1)}{\max\limits_{\theta_0} p(x|\theta_0)} \tag{1}$$

has the following asymptotic distribution under H_0:

$$2 \log \widehat{\Lambda}(x) \overset{n\to\infty}{\sim} \chi^2_{k-\ell} \qquad i.e., \qquad 2 \log \widehat{\Lambda}(x) \overset{D}{\to} \chi^2_{k-\ell}$$

Proof: (Sketch) under the conditions of the theorem, the log GLRT tends to the log GLRT in a Gaussian setting according to the Central Limit Theorem (CLT).

Example: Nested Condition

$$H_0 : x_i \overset{iid}{\sim} \mathcal{N}(0,1)$$
$$H_1 : x_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2), i = 1,2,\ldots,n, \sigma^2 > 0 \text{ unknown}$$

log LR:

$$\sum_{i=1}^{n}\left(-\frac{1}{2}\log\sigma^2 - x_i^2\left(\frac{1}{2\sigma^2}-\frac{1}{2}\right)\right)$$

MLE of σ^2:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

log GLR under H_0:

$$2\left[\sum -\frac{1}{2}\log\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \frac{x_i^2}{2}\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}x_i^2}-1\right)\right] \overset{n\to\infty}{\sim} \chi_1^2$$

6 Student's t distribution and t-tests

Consider the following hypothesis testing problem:

$$H_0 : x_i \overset{iid}{\sim} \mathcal{N}(\mu_0,\sigma^2),\ i = 1,\ldots,n$$
$$H_1 : x_i \overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2),\ i = 1,\ldots,n,\ \mu > \mu_0 \text{ but otherwise unknown}$$

We have discussed how to handle this test when σ^2 is known. But how should we proceed if it is unknown?

One option is the GLRT, discussed above. However, (a) we must estimate μ and (b) Wilk's theorem only tells us that the test statistic corresponding to maximum likelihood estimates of σ^2 and μ is *asymptotically* chi-squared. For small n, then, it can be difficult to set a threshold to achieve a desired probability of false positives or type I error.

As alternative is the celebrated t-test. Specifically, let

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

and note that under H_0, $\overline{x} \sim \mathcal{N}(\mu_0,\sigma^2/n)$. So if we knew σ^2, we could compute the statistic $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ and set a threshold as discussed in previous units. Since we do not know σ^2, we can estimate it from our data; specifically, let $s_n := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$ be the sample standard deviation. Then $s/\sqrt{n}$ is called the standard error of the mean and is an estimate of $\sigma/\sqrt{n}$. This leads us to the t-statistic:

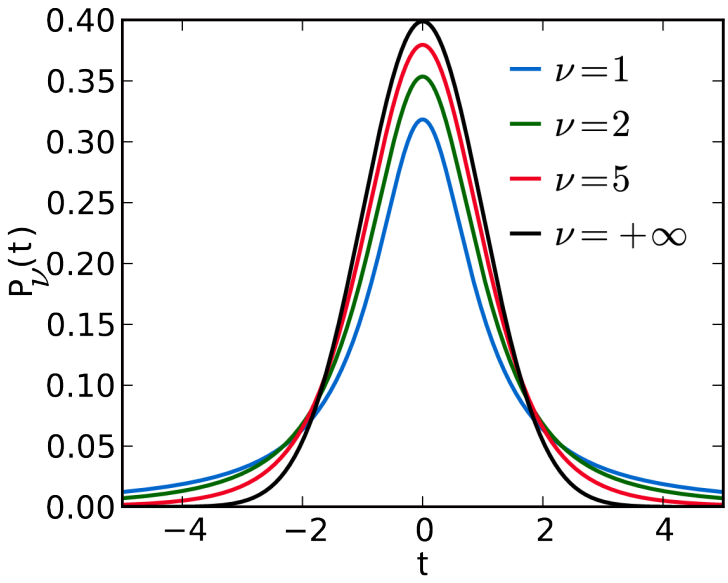
$$t^* = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}.$$

Ultimately we will perform our hypothesis test by thresholding t^*, and to set a threshold guaranteed to yield a certain probability of false positives or type I error we must undertand the distribution of t^*.

In 1908, Guinness statistician William Gosset published a paper characterizing this distribution under the pseudonym "Student", and subsequently the distribution has been dubbed Student's t-distribution. It is parameterized by ν, the number of degrees of freedom in the distribution, and takes the form

$$p_\nu(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}.$$

The test statistic t^* above is drawn from the t-distribution with $\nu = n-1$ degrees of freedom.



As $\nu \longrightarrow \infty$, $p_\nu(t) \longrightarrow \mathcal{N}(0,1)$. For smaller ν corresponding to smaller sample sizes, though, the t-distribution has heavier tails, and its tail probabilities can be used to determine appropriate thresholds for t-statistics.

7 p-values

So far we have considered making decisions or performing hypothesis testing by computing a test statistic and thresholding it. Our aim is the answer the key question

Note: Does our data provide enough evidence for us to reject the null hypothesis H_0?

We saw that we can choose a threshold to minimize the probability of error or probability of false positives or other measures of error. However, the result of such a test is always a binary decision (H_0 or H_1) and not a measure of how strong our evidence is again H_0. p-values bridge this gap.

Specifically, for a given test statistic t^*, we could perform the test

$$t^* \underset{H_0}{\overset{H_1}{\gtrless}} \tau_\alpha$$

where τ_α is a threshold associated with a type I error or false positive rate of α (the value of τ_α depends on the distribution of t^* under the null hypothesis). One can easily imagine that there is a range of values of α which would all lead us to reject H_0. The p-value is essentially the smallest α (corresponding to the largest threshold τ_α) for which we would reject H_0 with our test statistic. More formally

Definition: p-value

The p-value is the smallest level at which we can reject H_0:

$$\text{p-value} = \inf\{\alpha : t^* > \tau_\alpha\}.$$

More generally, if R_α is the rejection region associated with a test at level α, then

$$\text{p-value} = \inf\{\alpha : t^* \in R_\alpha\}.$$

Note: Notes on the p-value

- it measures the strength of the evidence against H_0: a small p-value (e.g., below 0.05, ideally below 0.01) indicates strong evidence against H_0.
- a large p-value is NOT evidence in favor of H_1 (it's possible we just have a low-power test)
- the p-value should NOT be thought of as $\mathbb{P}(H_0|\text{data})$.

Theorem: Computation of the p-value

Let p_0 denote the distribution of the test statistic under H_0. If we have a test of the form *reject* H_0 *if and only if* $t^* \geq \tau_\alpha$, then

$$\text{p-value} = \mathbb{P}(T \geq t^* | T \sim p_0).$$

In other words, the p-value is the probability under H_0 of observing a test statistic at least as extreme as what was observed.

Distribution of the p-value

If the test statistic has a continuous distribution, then under H_0 the p-value is uniformly distributed between 0 and 1. Thus if we reject H_0 whenever a p-value is less than α, that test as a type I error or probability of false positives of α.