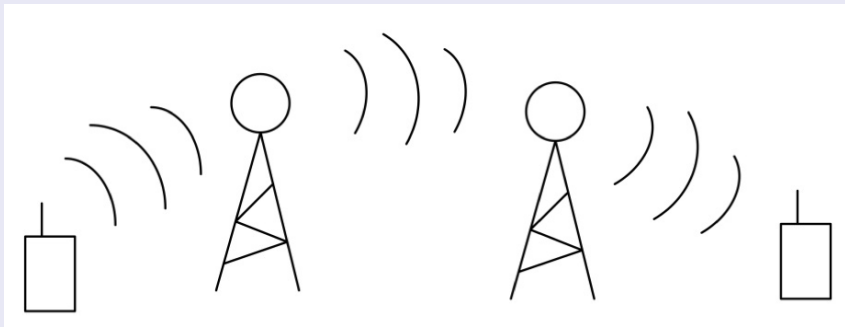
10. Composite Hypothesis Testing

ECE 830, Spring 2014

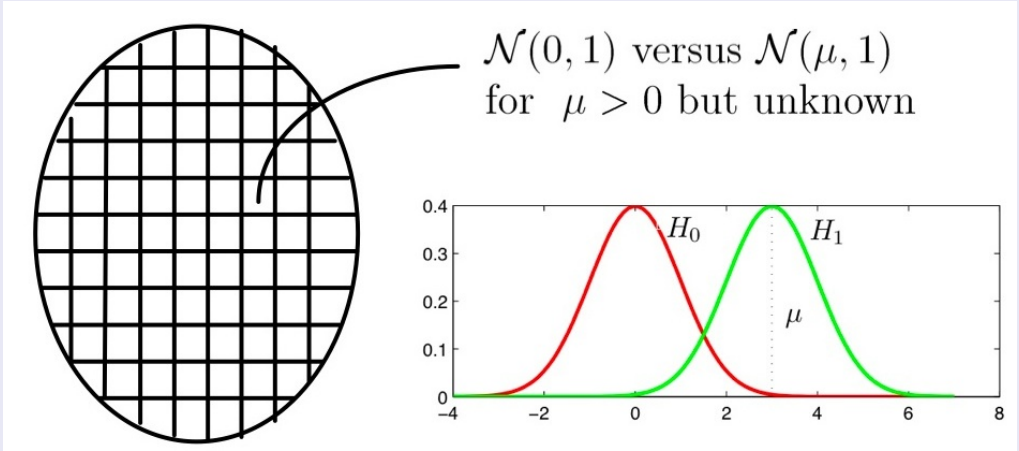
In many real world problems, it is difficult to precisely specify probability distributions. Our models for data may involve unknown parameters or other characteristics. Here are a few motivating examples.

Example: Unknown amplitudes/delays in wireless communications.

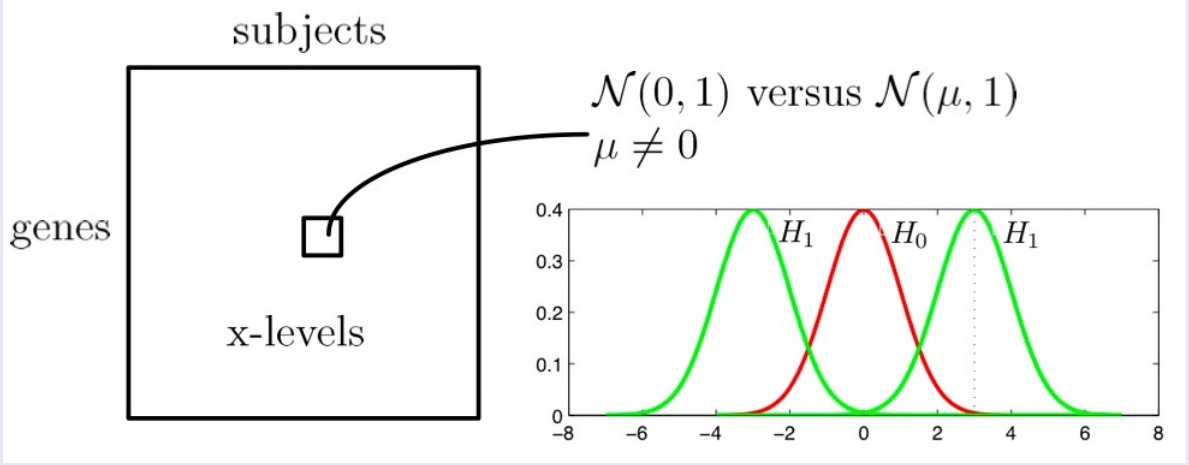


We don't always know how many relays a signal will go through, how strong the signal will be at each receiver, the distance between relay stations, etc.

Example: Unknown signal amplitudes in functional brain imaging.



Example: Unknown expression levels in gene microarray experiments.



Composite Hypothesis Tests

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

$$H_0 : X \sim p_0(x|\theta_0),\ \theta_0 \in \Theta_0$$
$$H_1 : X \sim p_1(x|\theta_1),\ \theta_1 \in \Theta_1$$

- 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.

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

Example: Brain imaging

Recall the brain imaging problem.

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

In this example, H_0 is and H_1 is

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:

Test statistic:

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

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.**

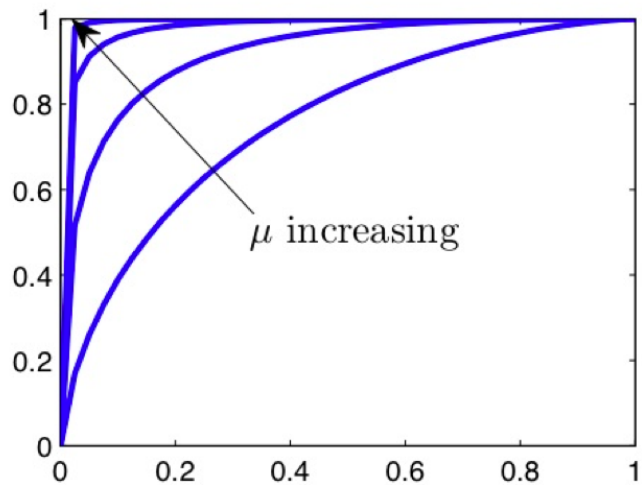
Let $t = \sum_{i=1}^{n} x_i$ denote the test statistic. It is easy to determine its distribution(s) under each hypothesis (a composite in the case of H_1).

$$H_0: \quad t \sim$$
$$H_1: \quad t \sim \qquad\qquad \mu > 0 \text{ unknown}$$

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

$$P_{FA} = \qquad\qquad \Rightarrow \gamma =$$

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.



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).

Karlin-Rubin Theorem

Karlin-Rubin Theorem

Let t be a scalar test statistic whose density, under both hypotheses, is parameterized by a **scalar** parameter θ. Assume that the likelihood ratio statistic

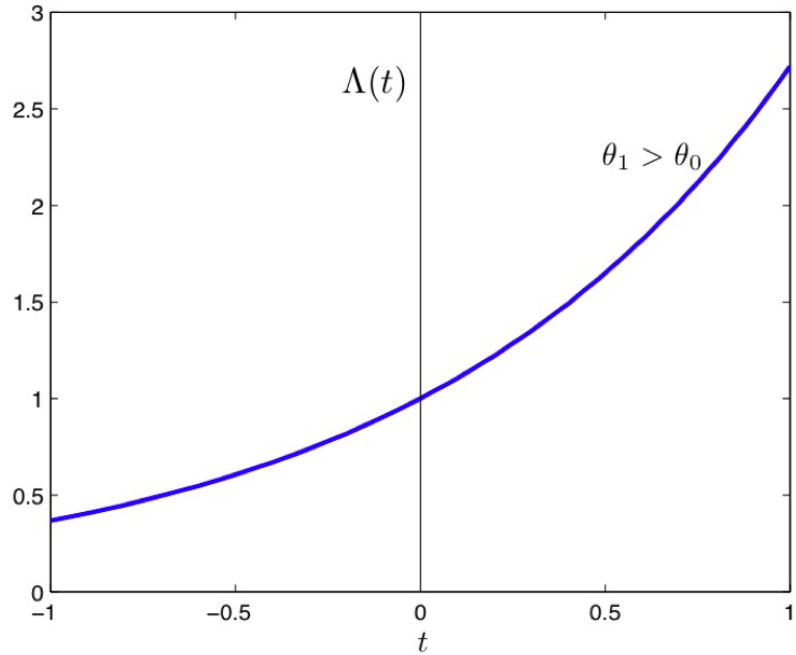
$$\Lambda(t) \;=\; \frac{p(t|\theta_1)}{p(t|\theta_0)}$$

is a non-decreasing function of t for every pair $(\theta_0, \theta_1 > \theta_0)$. In this case, the threshold test

$$t \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

is the test that maximizes P_D for a given P_{FA} (both depend on γ) for all $(\theta_0, \theta_1 > \theta_0)$. We say that this test is uniformly most powerful (UMP) among all tests with this P_{FA}.

We say that t has a monotone likelihood ratio, and the idea is depicted in the figure below.



$\Lambda(t)$ as a non-decreasing function of t for a pair $\theta_1 > \theta_0, \theta_0$.

The interpretation is simple: the larger t is, the more probable H_1 looks compared to H_0.

Example: Poisson rates

$$H_0 : \quad x \sim \text{Poisson}(\lambda_0)$$
$$H_1 : \quad x \sim \text{Poisson}(\lambda_1),\ \lambda_1 > \lambda_0$$

$$\Lambda(x) = e^{-(\lambda_1 - \lambda_0)} \left(\frac{\lambda_1}{\lambda_0} \right)^x$$

is a non-decreasing function of x for any $(\lambda_1 > \lambda_0, \lambda_0)$ pair.

Example:

$$H_0 : x \sim \mathcal{N}(0, \Sigma), \qquad (x = w)$$
$$H_1 : x \sim \mathcal{N}(As, \Sigma), \qquad (x = As + w)$$

$w \sim \mathcal{N}(0, \Sigma)$ is noise, $s \in \mathbb{R}^n$ is a known waveform, $A > 0$ is an unknown amplitude. The log LRT is

$$\log \Lambda(x) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma'$$
$$\underset{H_0}{\overset{H_1}{\gtrless}} \gamma'$$
$$t(x) = \underset{H_0}{\overset{H_1}{\gtrless}} \gamma, \quad \text{since } A > 0.$$

Example: (cont.)

$$p(t|A) = \mathcal{N}(As^\top \Sigma, s^\top \Sigma^{-1} s)$$
$$p(t|0) = \mathcal{N}(0, s^\top \Sigma^{-1} s)$$

$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{s^\top \Sigma^{-1} s}}\right)$$

Suppose $\Sigma = \sigma^2 I$; then the log LRT is

$$\gtrless \gamma$$

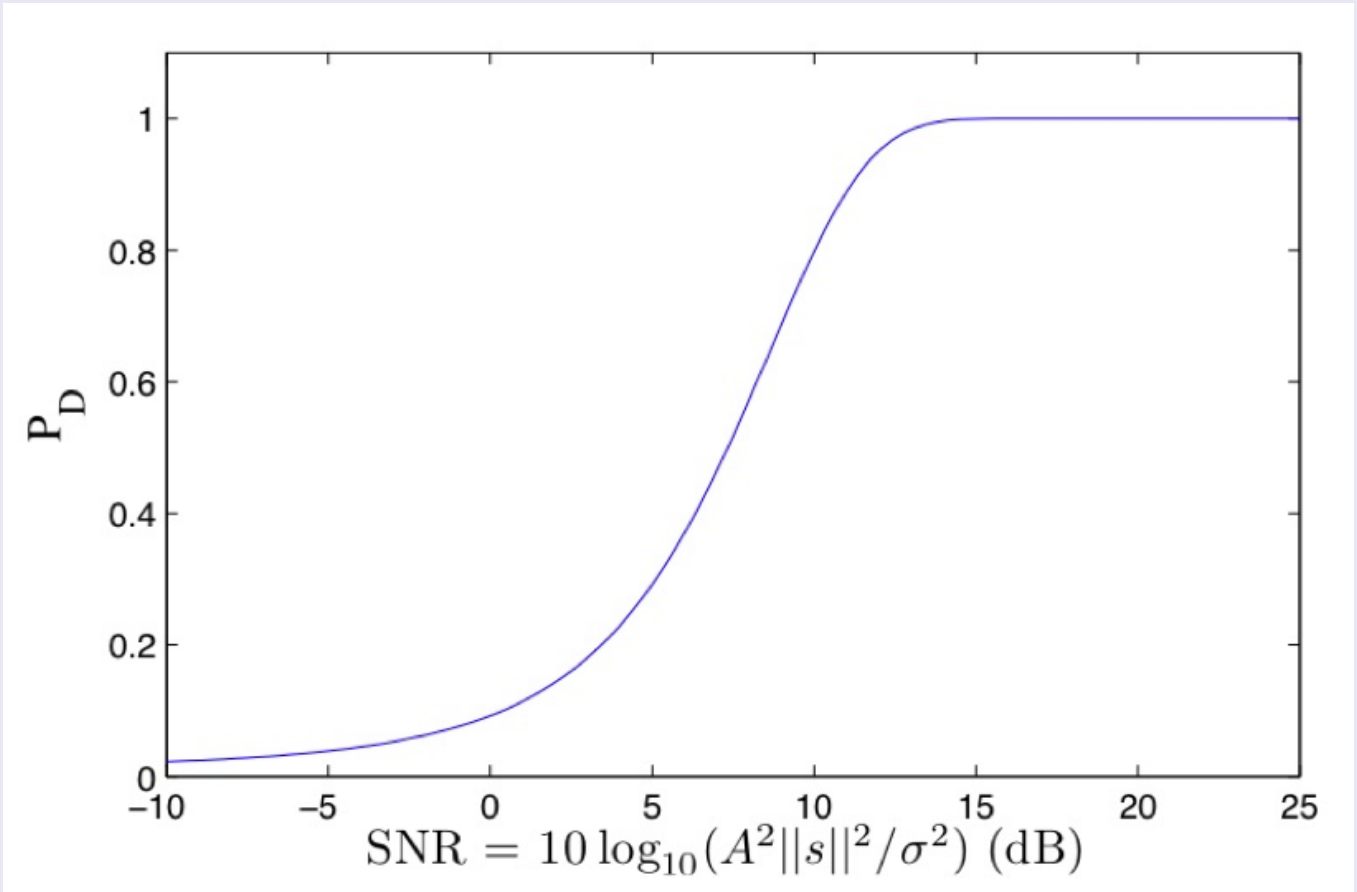
$$P_{FA} = Q\left(\frac{\gamma}{\sqrt{\sigma^2 s^\top s}}\right), \ \gamma = \sqrt{||s||^2 \sigma^2} Q^{-1}\left(P_{FA}\right)$$

$$P_D = Q\left(\frac{\gamma - A||s||^2}{\sqrt{\sigma^2 ||s||^2}}\right) = Q\left(Q^{-1}\left(P_{FA}\right) - \sqrt{\frac{A^2||s||^2}{\sigma^2}}\right)$$

where we term SNR $= A^2||s||^2/\sigma^2$.

Example: (cont.)

If $s_\ell = A\sin(\frac{2\pi}{10}\ell), \ell = 1, \cdots, 100$ and we set $P_{FA} = 10^{-2}$, then we have the below relationship between P_D and SNR.



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

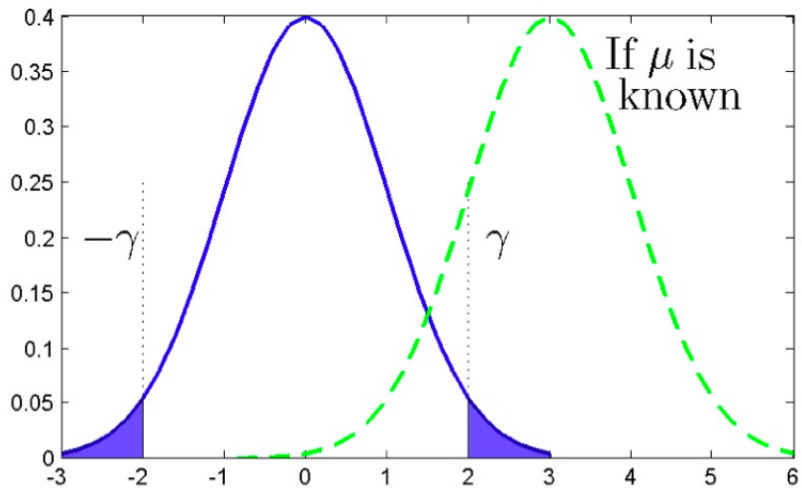
$$\mu\, x - \mu^2/2 \ \underset{H_0}{\overset{H_1}{\gtrless}} \ \gamma' \ .$$

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

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

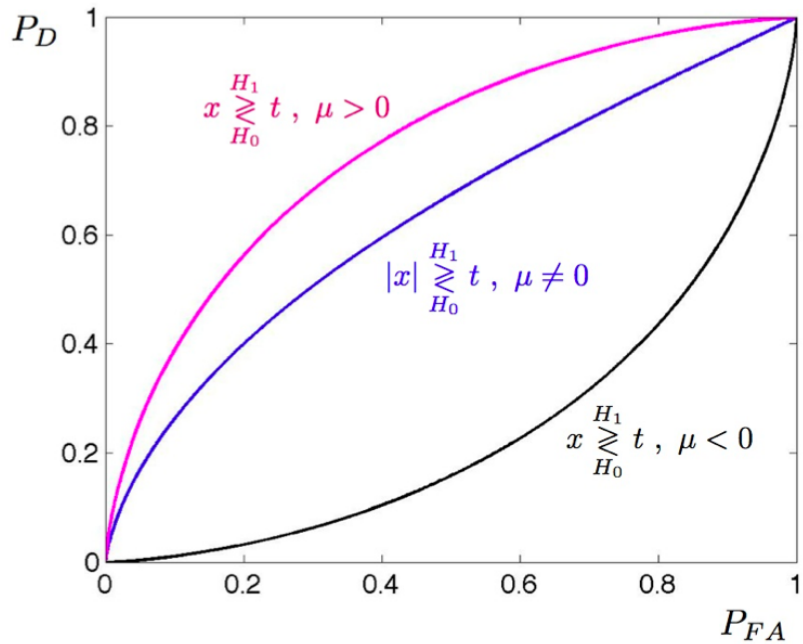
Since μ is unknown (and not necessarily positive) the test is uncomputable.

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.



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

The P_D depends on μ, which is unknown.



Two "Derivations" of the Wald Test

(1) Generalized Likelihood Ratio Test (GLRT) 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 likelihod 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

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

which is equivalent to the Wald test.

Two "Derivations" of the Wald Test

(2) Bayes Factor Model μ as an independent random parameter with prior probability distribution $p(\mu)$. The alternative hypothesis is that $\mu \neq 0$, and with no other prior information it is natural to take $p(\mu)$ to be symmetric about the origin. In particular, the prior probability distirbution $\mu \sim \mathcal{N}(0, \sigma^2)$ is symmetric and models a prior belief that smaller values of μ are more probable that larger values. The Gaussian form is also convenient to use with the Gaussian likelihood. The **Bayes Factor** is the ratio

$$\Lambda_{BF}(x) \;=\; \frac{\int p(x|\mu)p(\mu)d\mu}{p(x|0)} \;.$$

This ratio **compares the average model in H_1 (with respect to the prior $p(\mu)$) with the H_0 model.**

The integral in the numerator is easy to compute. Note that $X = \mu + W$, where $W \sim \mathcal{N}(0, 1)$. So X is the sum of two independent Gaussian random variables, and its distribution is $X \sim N(0, 1 + \sigma^2)$. The Bayes Factor is therefore

$$\Lambda_{BF}(x) \;=$$

Taking the log and absorbing constant factors and terms into the threshold yields the test

$$x^2 \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \;,$$

which again is equivalent to the Wald test.

GLRT and Bayes Factors

Consider a composite hypothesis test of the form

$$H_0 : X \sim p_0(x|\theta_0),\ \theta_0 \in \Theta_0$$
$$H_1 : X \sim p_1(x|\theta_1),\ \theta_1 \in \Theta_1$$

The general forms for the GLRT and Bayes Factor are as follows.

GLRT:

$$\frac{\max_{\theta_1 \in \Theta_1} p_1(x|\theta_1)}{\max_{\theta_0 \in \Theta_0} p_0(x|\theta_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma\ .$$

Bayes Factor Assume $\theta_0 \sim p(\theta_0)$ and $\theta_1 \sim p(\theta_1)$, two different prior probability distributions. The Bayes Factor is

$$\frac{\int_{\Theta_1} p_1(x|\theta_1)p(\theta_1)d\theta_1}{\int_{\Theta_0} p_0(x|\theta_0)p(\theta_0)d\theta_0}\ .$$

The GLRT compares the best model in H_1 to the best in H_0, and the Bayes Factor compares the average model in H_1 to the average model in H_0, with respect to the specified prior probability distributions.

Example:

$$H_0 : \mathcal{N}(0, \sigma^2 I)$$
$$H_1 : \mathcal{N}(H\theta, \sigma^2 I).$$

where $\sigma^2 > 0$ is known, $\theta \in \mathbb{R}^k$ is unknown, and $H \in \mathbb{R}^{n \times k}$ is known. Then the log LRT is

$$\begin{aligned}
\Lambda(x) = & -\frac{1}{2\sigma^2}\left((x - H\theta)^\top (x - H\theta) - x^\top x\right) \\
= & -\frac{1}{\sigma^2}(-2\theta^\top H^\top x + \theta^\top H^\top H\theta) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma' \\
\iff & \theta^\top H^\top x \underset{H_0}{\overset{H_1}{\gtrless}} \gamma,
\end{aligned}$$

and this test is not computable without knowledge of θ

Example: (cont.)

Recall that

$$\begin{aligned} & H_1 : x \sim \mathcal{N}(H\theta, \sigma^2 I), && \theta \in \mathbb{R}^k \\ \iff & H_1 : x \sim p_1, && p_1 \in \{\mathcal{N}(H\theta, \sigma^2 I)\}_{\theta \in \mathbb{R}^k}, \end{aligned}$$

We want to pick p_1 in $\{\mathcal{N}(H\theta, \sigma^2 I)\}$ that matches x the best.

$$p(x|\theta, H_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - H\theta)^\top (x - H\theta)\right\}$$

Find θ that maximizes the likelihood of observing x

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^k} \underbrace{(x - H\theta)^\top (x - H\theta)}_{||x - H\theta||^2} = (H^\top H)^{-1} H^\top x$$

Example: (cont.)

Plugging $\hat{\theta}$ into the test statistic $\theta^\top H^\top x$, we have

$$\hat{\theta}^\top H^\top x = \underset{H_0}{\overset{H_1}{\gtrless}} \gamma$$

This is the so-called Generalized LRT (GLRT) and its distribution is chi-squared with k degrees of freedom (k being the dimension of the subspace spanned by the columns of H). This distribution is denoted χ^2_k. **The test computes the energy in the signal subspace and if the energy is large enough, then H_1 is accepted.**

Exercise: show that using the prior $\theta \sim \mathcal{N}(0, \alpha^2 I)$ and computing the Bayes Factor yields the same test.