Topic 5: Generalized Likelihood Ratio Test

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN

5.1 Introduction

We already started studying *composite* hypothesis problems of the form:

$$\begin{aligned} H_0: \ x &\sim \mathsf{p}_0(\mathrm{x}|\theta_0^\star), \quad \theta_0^\star \in \Theta_0, \\ H_1: \ x &\sim \mathsf{p}_1(\mathrm{x}|\theta_1^\star), \quad \theta_1^\star \in \Theta_1. \end{aligned} \tag{5.1}$$

where the parameters $\theta_0^\star$ and/or $\theta_1^\star$ are unknown.

Example 5.1. In Example 3.3 we wanted to determine whether two meteorites came from the same asteroid in space using the difference x of their magnesium composition. In Example 3.4 we wanted to determine whether a certain gene was associated with a disease, using the difference x of the gene's average activation levels between healthy and sick people. Both of these problems can be modeled as

$$\begin{aligned} H_0: \ x &\sim \mathcal{N}(0, \sigma^2) && \Rightarrow \text{same asteroid/gene unrelated to disease,} \\ H_1: \ x &\sim \mathcal{N}(\mu_1^\star, \sigma^2) && \Rightarrow \text{different asteroids/gene related to disease,} \end{aligned}$$

where $\mu_1^\star$ is unknown.

Example 5.2. In Example 4.1 we use an array of sensors to record small voltages generated by your brain and store them in a *signal* vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$. A machine (phone, computer, server, etc.) should

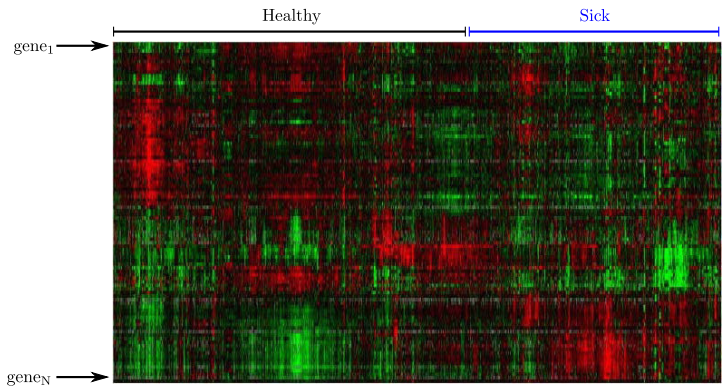

Figure 5.1: Gene microarrays are data matrices indicating gene *activation levels*. Each row corresponds to one gene, and each column corresponds to one individual. We want to know which genes are related to a disease. See Example 5.1.

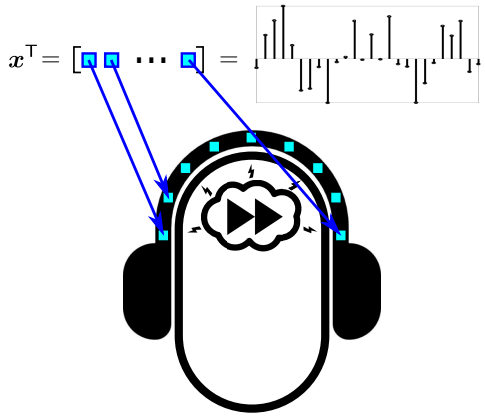

Figure 5.2: Sensors record small voltages generated by your brain and store them in a *signal* vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}}$. A machine (phone, computer, server, etc.) should interpret $\boldsymbol{x}$ and *skip* the song if that is what you thought about. See Example 5.2.

interpret $\boldsymbol{x}$ and *skip* the song if that is what you thought about. This can be modeled as

$$\begin{aligned} H_0: \ \boldsymbol{x} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) && \Rightarrow \text{do nothing}, \\ H_1: \ \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{\mu}_{\blacktriangleright\blacktriangleright}^\star, \sigma^2 \mathbf{I}) && \Rightarrow \text{skip song}, \end{aligned}$$

where $\boldsymbol{\mu}_{\blacktriangleright\blacktriangleright}^\star$ is unknown.

In Section 3.7 we already studied the hypothesis problems in Example 5.1, which led us to the intuitive answer of Wald's test in Example 3.14. We now generalize this using our knowledge from estimation theory to obtain the *generalized likelihood ratio test* (GLRT).

Definition 5.1 (Generalized likelihood ratio test (GLRT))**.** Consider a hypothesis problem as in (5.1), where $\theta_0^\star$ and/or $\theta_1^\star$ are unknown. The *generalized likelihood ratio statistic* is defined as:

$$\hat{\Lambda}(\mathrm{x}) \ := \ \frac{\max\limits_{\theta_1 \in \Theta_1} \ \mathsf{p}_1(\mathrm{x}|\theta_1)}{\max\limits_{\theta_0 \in \Theta_0} \ \mathsf{p}_0(\mathrm{x}|\theta_0)},$$

and the *generalized likelihood ratio test* is defined as $\hat{\Lambda}(\mathrm{x}) \ \gtrless_{H_0}^{H_1} \ \tau$.

The idea behind the GLRT is actually quite simple: if you don't now a parameter, then first estimate it using maximum likelihood, and then use it as if it were the *true* parameter in a *likelihood ratio test* (LRT).

Example 5.3 (Derivation of Wald's test as a GLRT)**.** Let us show that Wald's test is just one particular case of the GLRT. Consider The setup in Example 5.1. Here θ_0 is known, and $\theta_1^\star = \mu_1^\star$, so $\Theta_1 = \mathbb{R}$.

Then

$$\begin{aligned}\arg\max_{\theta_1\in\Theta_1} \mathsf{p}_1(\mathrm{x}|\theta_1) &= \arg\max_{\mu_1\in\mathbb{R}} \mathsf{p}_1(\mathrm{x}|\mu_1) = \arg\max_{\mu_1\in\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu_1}{\sigma}\right)^2} \\ &= \arg\max_{\mu_1\in\mathbb{R}} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu_1}{\sigma}\right)^2}\right) = \arg\max_{\mu_1\in\mathbb{R}} \underbrace{\log\frac{1}{\sqrt{2\pi}\sigma}}_{\text{constant}} - \frac{1}{2}\left(\frac{\mathrm{x}-\mu_1}{\sigma}\right)^2 \\ &= \arg\max_{\mu_1\in\mathbb{R}} -\underbrace{\frac{1}{2\sigma^2}}_{\text{constant}} (\mathrm{x}-\mu_1)^2 = \arg\max_{\mu_1\in\mathbb{R}} (\mu_1-\mathrm{x})^2\end{aligned}$$

To find this maximum we use our usual tricks: take derivative with respect to (w.r.t.) μ_1, set to zero and solve for μ_1.

$$\frac{\partial}{\partial\mu_1}(\mu_1-\mathrm{x})^2 = 2(\mu_1-\mathrm{x}) = 0$$

Which yields $\hat{\mu}_1 = \mathrm{x}$. Then

$$\hat{\Lambda}(\mathrm{x}) = \frac{\max_{\theta_1\in\Theta_1} \mathsf{p}_1(\mathrm{x}|\theta_1)}{\max_{\theta_0\in\Theta_0} \mathsf{p}_0(\mathrm{x}|\theta_0)} = \frac{\max_{\mu_1\in\mathbb{R}} \mathsf{p}_1(\mathrm{x}|\mu_1)}{\mathsf{p}_0(\mathrm{x}|\theta_0^\star)} = \frac{\mathsf{p}_1(\mathrm{x}|\hat{\mu}_1)}{\mathsf{p}_0(\mathrm{x}|\theta_0^\star)} = \frac{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\hat{\mu}_1}{\sigma}\right)^2}}{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{\mathrm{x}}{\sigma}\right)^2}} = \frac{e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mathrm{x}}{\sigma}\right)^2}}{e^{-\frac{1}{2}\left(\frac{\mathrm{x}}{\sigma}\right)^2}} = e^{\frac{\mathrm{x}^2}{2\sigma^2}},$$

and our GLRT is $e^{\frac{\mathrm{x}^2}{2\sigma^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \tau$. Taking log on both sides this becomes $\frac{\mathrm{x}^2}{2\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} \log\tau$, or equivalently

$$|\mathrm{x}| \underset{H_0}{\overset{H_1}{\gtrless}} \tau',$$

where $\tau' = \sqrt{2\sigma^2\log\tau}$, which is precisely Wald's test from Example 3.14.

5.2 Asymptotics

Recall that in hypothesis testing we often want to bound the probability p_{10} of deciding H_1 when H_0 was true, and so we select our threshold test τ' accordingly. For instance, in Wald's test, $\mathsf{p}_{10} = \mathsf{P}(|\mathrm{x}| > \tau')$ given that H_0 is true. Since $H_0 : x \sim \mathcal{N}(0,\sigma^2)$, if we want $\mathsf{p}_{10} < \alpha$, all we need to do is find the τ' such that the probability of the tails of a $\mathcal{N}(0,\sigma^2)$ is α (see Figure 5.3 to build some intuition). We can do this because we know the distribution of $|x|$. Similarly, if we had a test like:

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}^2 \underset{H_0}{\overset{H_1}{\gtrless}} \tau', \tag{5.2}$$

where $H_0 : x_\mathrm{i} \overset{iid}{\sim} \mathcal{N}(0,1)$, we would also know that $\sum_{\mathrm{i}=1}^{\mathrm{N}} x_\mathrm{i}^2 \sim \chi^2(\mathrm{N})$. Again, if we wanted $\mathsf{p}_{10} < \alpha$, all we would need to do is find the τ' such that the probability of the tail of a $\chi^2(\mathrm{N})$ is α (see Figure 5.3 for more intuition).

However, we don't always know the distribution of our test. For example, if we had the same test in (5.2), but with $H_0 : x_i \overset{iid}{\sim}$ Poisson, or Cauchy, or Weibull, then what is the distribution of $\sum_{\mathrm{i}=1}^{\mathrm{N}} x_\mathrm{i}^2$? The following theorem states that if N is large enough, we don't need to worry about it, because under H_0, the GLRT is

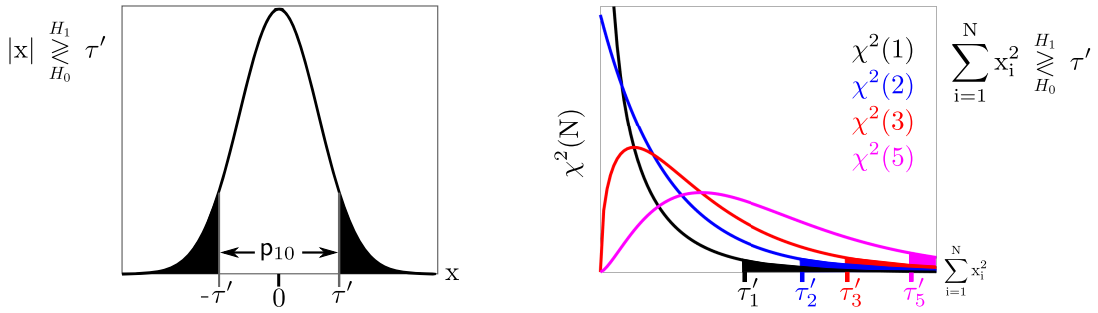

Figure 5.3: We want our test to have $\mathsf{p}_{10} < \alpha$. **Left:** If our test is $|\mathrm{x}| \gtrless_{H_0}^{H_1} \tau'$ and $x \sim \mathcal{N}(0, \sigma^2)$, then we need to find the τ' such that the probability of the tails of a $\mathcal{N}(0, \sigma^2)$ is α. **Right:** if our test is $\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}^2 \gtrless_{H_0}^{H_1} \tau'$ and $x_\mathrm{i} \overset{iid}{\sim} \mathcal{N}(0,1)$, then we need to find the τ' such that the probability that the tail of a $\chi^2(\mathrm{N})$ is α. We can do this because we know the distribution of our test. What happens if we don't? Wilks' Theorem gives us an answer.

asymptotically distributed $\chi^2(\mathrm{N})$.

Theorem 5.1 (Wilks' Theorem — Asymptotic distribution of the GLRT)**.** Consider a composite hypothesis problem of the form:

$$H_0: \ x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathsf{p}(\mathrm{x}|\boldsymbol{\theta}_0^\star), \quad \boldsymbol{\theta}_0^\star \in \mathbb{R}^{\mathrm{K}_0},$$
$$H_1: \ x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathsf{p}(\mathrm{x}|\boldsymbol{\theta}_1^\star), \quad \boldsymbol{\theta}_1^\star \in \mathbb{R}^{\mathrm{K}_1},$$

where p has the same form in both hypotheses, and the K_0 unknown parameters in $\boldsymbol{\theta}_0^\star$ are a subset of the K_1 unknown parameters in $\boldsymbol{\theta}_1^\star$ (these are called *nested hypotheses*). Suppose that for every $\mathrm{k}, \ell \in \{1, \ldots, \mathrm{K}\}$, $\frac{\partial \mathsf{p}(\mathrm{x}|\boldsymbol{\theta})}{\partial \theta_\mathrm{k}}$ and $\frac{\partial^2 \mathsf{p}(\mathrm{x}|\boldsymbol{\theta})}{\partial \theta_\mathrm{k} \partial \theta_\ell}$ exist, and that $\mathsf{E}\left[\frac{\partial \log \mathsf{p}(\mathrm{x}|\boldsymbol{\theta})}{\partial \theta_\mathrm{k}}\right] = 0$ (this guarantees that the MLE $\hat{\boldsymbol{\theta}}_{ML}$ converges to the true $\boldsymbol{\theta}^\star$). Then under H_0,

$$2 \log \hat{\Lambda}(\boldsymbol{x}) \overset{\mathrm{N}\to\infty}{\sim} \chi^2(\mathrm{K}_1 - \mathrm{K}_0).$$

Proof. The proof follows as a consequence of the central limit theorem. For a detailed proof see Theorems 10.3.1 and 10.3.3 in *Statistical Inference* by George Casella and Roger L. Berger, second edition. □

Example 5.4. Consider

$$H_0: \ x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0),$$
$$H_1: \ x_1, \ldots, x_\mathrm{N} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_1^{\star 2}), \quad \sigma_1^\star > 0.$$

The MLE of $\sigma_1^\star$ is

$$\begin{aligned}\underset{\theta_1\in\Theta_1}{\arg\max}\ \mathsf{p}_1(\mathbf{x}|\theta_1) &= \underset{\sigma_1\in\mathbb{R}_+}{\arg\max}\ \mathsf{p}_1(\mathbf{x}|\sigma_1) = \underset{\sigma_1\in\mathbb{R}_+}{\arg\max}\ \frac{1}{(\sqrt{2\pi}\sigma_1)^{\mathrm{N}}}e^{-\frac{1}{2\sigma_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}\\ &= \underset{\sigma_1\in\mathbb{R}_+}{\arg\max}\ \log\Big(\frac{1}{(\sqrt{2\pi}\sigma_1)^{\mathrm{N}}}e^{-\frac{1}{2\sigma_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}\Big)\\ &= \underset{\sigma_1\in\mathbb{R}_+}{\arg\max}\ \underbrace{\log\frac{1}{(\sqrt{2\pi})^{\mathrm{N}}}}_{\text{constant}} - \log\sigma_1^{\mathrm{N}} + \log e^{-\frac{1}{2\sigma_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}\\ &= \underset{\sigma_1\in\mathbb{R}_+}{\arg\max}\ -\frac{\mathrm{N}}{2}\log\sigma_1^2 - \frac{1}{2\sigma_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})\end{aligned}$$

Taking derivative w.r.t. σ_1^2 and setting to zero we have:

$$\frac{1}{\sigma_1^4}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1}) - \frac{\mathrm{N}}{\sigma_1^2} = \frac{1}{\sigma_1^2}\Big(\frac{1}{\sigma_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1}) - \mathrm{N}\Big) = 0.$$

Solving for σ_1^2 we obtain the MLE:

$$\hat{\sigma}_1^2 = \frac{1}{\mathrm{N}}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i}-\mu)^2.$$

Then

$$\begin{aligned}\hat{\Lambda}(\mathbf{x}) &= \frac{\max_{\theta_1\in\Theta_1}\ \mathsf{p}_1(\mathbf{x}|\theta_1)}{\max_{\theta_0\in\Theta_0}\ \mathsf{p}_0(\mathbf{x}|\theta_0)} = \frac{\max_{\sigma_1\in\mathbb{R}_+}\ \mathsf{p}_1(\mathbf{x}|\sigma_1)}{\mathsf{p}_0(\mathbf{x}|\theta_0^\star)} = \frac{\mathsf{p}_1(\mathbf{x}|\hat{\sigma}_1)}{\mathsf{p}_0(\mathbf{x}|\sigma_0^\star)} = \frac{\frac{1}{(\sqrt{2\pi}\hat{\sigma}_1)^{\mathrm{N}}}e^{-\frac{1}{2\hat{\sigma}_1^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}}{\frac{1}{(\sqrt{2\pi}\sigma_0^\star)^{\mathrm{N}}}e^{-\frac{1}{2\sigma_0^{\star2}}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}}\\ &= \Big(\frac{\sigma_0^\star}{\hat{\sigma}_1}\Big)^{\mathrm{N}} e^{(\frac{1}{2\sigma_0^{\star2}}-\frac{1}{2\hat{\sigma}_1^2})\overbrace{(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}^{\mathrm{N}\hat{\sigma}_1^2}} = \Big(\frac{\sigma_0^\star}{\hat{\sigma}_1}\Big)^{\mathrm{N}} e^{\frac{\mathrm{N}}{2}(\frac{\hat{\sigma}_1^2}{\sigma_0^{\star2}}-1)}.\end{aligned}$$

Taking log on both sides we obtain:

$$\log\hat{\Lambda}(\mathbf{x}) = -\overbrace{\frac{\mathrm{N}}{2}\log\frac{1}{\mathrm{N}}\underbrace{\sum_{\mathrm{i}=1}^{\mathrm{N}}\Big(\frac{\mathrm{x_i}-\mu}{\sigma_0^\star}\Big)^2}_{H_0:\sim\chi^2(\mathrm{N})}}^{\frac{\mathrm{N}}{2}\log\left(\frac{\hat{\sigma}_1}{\sigma_0^\star}\right)^2} + \overbrace{\frac{1}{2}\underbrace{\sum_{\mathrm{i}=1}^{\mathrm{N}}\Big(\frac{\mathrm{x_i}-\mu}{\sigma_0^\star}\Big)^2}_{H_0:\sim\chi^2(\mathrm{N})}}^{\frac{\mathrm{N}}{2}\left(\frac{\hat{\sigma}_1}{\sigma_0^\star}\right)^2} - \frac{\mathrm{N}}{2}.$$

Now the question is: do you know the distribution of $-\log\chi^2(\mathrm{N})+\chi^2(\mathrm{N})$? Sure as hell *I* don't! Luckily, there are no unknown parameters under H_0, and one unknown parameter under H_1 (namely $\sigma_1^\star$), so these are nested hypotheses, and hence can use Theorem 5.1 to know that under H_0,

$$2\log\hat{\Lambda}(\boldsymbol{x}) \overset{\mathrm{N}\to\infty}{\sim} \chi^2(1).$$

Hence, if we want $\mathsf{p}_{10} < \alpha$, it suffices to select the threshold τ for which the tail probability of a $\chi^2(1)$ random variable is α.

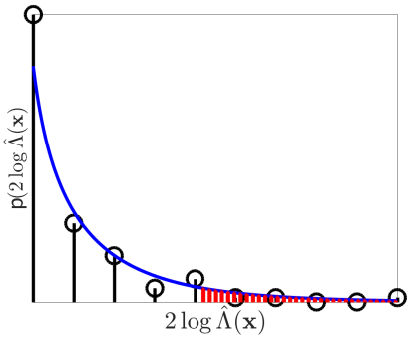

Figure 5.4: Results of the simulation of Example 5.4. In **black** is the histogram of $2\log\hat{\Lambda}(\mathbf{x})$ over $\mathrm{T} = 100$ trials. In **blue** is the asymptotic distribution of $2\log\hat{\Lambda}(\boldsymbol{x})$. In **red** is the probability of error p_{10} set to $\alpha = 0.05$. The code for this simulation is in the appendix.

5.3 Experiments

Let us now run some simulations to verify Theorem 5.1 and our results from Example 5.4. We will generate $x_1, \ldots, x_{\mathrm{N}} \overset{iid}{\sim} \mathcal{N}(\mu, \sigma_0^{\star 2})$ and we will compute $2\log\hat{\Lambda}(\mathbf{x})$ as in Example 5.4. We will repeat this T trials, plot a histogram, and compare it with the $\chi^2(1)$ distribution predicted by Theorem 5.1. Figure 5.4 shows some results. The code for this simulation is in the appendix.

A Code for Simulation of Example 5.4

```
clear all; close all; clc; warning('off','all');

% === Code to simulate the asymptotic distribution of log Lambda_hat ===
% ===== See Example 5.4.

T = 100;            % Number of trials.
N = 100;            % Number of samples.
mu = 5;             % Known mean under both hypotheses.
sigma0_star = 1;    % True standard deviation of X_i under H0.
sigma1_star = 2;    % True standard deviation of X_i under H1.
alpha = 0.05;       % Allowed probability of error (1|0).

log_Lambda_hat = zeros(T,1);    %log_Lambda_hat over trials.
for t=1:T,
    X = (randn(N,1)+mu) * sigma0_star;

    log_Lambda_hat(t) = -N/2 * log( 1/N * sum(((X-mu)/sigma0_star).^2) ) ...
        + 1/2 * sum(((X-mu)/sigma0_star).^2) ...
        - N/2;
end

% Plot results.
figure(1);
axes('Box','on');
hold on;
[h,rangeh] = hist(2*log_Lambda_hat);
stem(rangeh,h/T,'k','LineWidth',4,'MarkerSize',20);

% Asymptotic distribution of 2*log Lambda_hat.
rangeh = min(rangeh):0.01:max(rangeh);
chi2 = chi2pdf(rangeh,1);
plot(rangeh,chi2,'b','LineWidth',4);

% Highlight the probability alpha.
tau = chi2inv(1-alpha,1);
rangeh = tau:0.125:max(rangeh);
chi2 = chi2pdf(rangeh,1);
stem(rangeh,chi2,'r','LineWidth',5,'MarkerSize',1);

% Make figure look sexy.
axis tight;
set(gca,'XTick',[],'xticklabel',[]);
set(gca,'YTick',[],'yticklabel',[]);
ylabel('$\mathsf{p}(2\log\hat{\Lambda}($\textbf{x}$)$','Interpreter','latex','fontsize',25);
xlabel('$2\log\hat{\Lambda}($\textbf{x}$)$','Interpreter','latex','fontsize',30);

% Save figure.
set(gcf, 'renderer','default');
figurename = 'wilks.pdf';
saveas(gcf,figurename);
```