Lesson: Good tests are based on the likelihood ratio.

The proof is easy in the case of simple hypotheses:

\[ H_0 : x \sim p_0(x) = f(x|\theta_0) \]

\[ H_1 : x \sim p_1(x) = f(x|\theta_1) \]

The last equality is provided so this can look like a more familiar parametric test.
Suppose we have a sample \( x = (x_1, ..., x_n) \in \mathbb{R}^n \) and we want to choose between \( H_0 \) and \( H_1 \). (Note that \( p_i \) is the likelihood function.) Define a decision function \( d: \mathbb{R}^n \to \{0, 1\} \) such that \( d(x) = 0 \) when \( H_0 \) is accepted and \( d(x) = 1 \) when \( H_1 \) is accepted. Thus, \( d \) defines a partition of the sample space. The following diagrams illustrate situations where \( n = 2 \).
Let $A$ be the region in which $d = 0$. $A^c$ is the complement of $A$ in $R^n$. Then the error probabilities are

$$\alpha = P(d = 1|H_0) = \int_{A^c} p_0(x)dx$$
$$\beta = P(d = 0|H_1) = \int_A p_1(x)dx.$$ 

Note: $\alpha$ is the size of the test - the probability of an error of the first type, and $\beta$ is the operating characteristic of the test - the probability of an error of the second type. $(1 - \beta)$ is the power of the test.
You would like to choose a test minimizing both error probabilities, but there are tradeoffs. $\alpha$ can be set to 0, its minimum, by choosing $d = 0$ always; but then $\beta = 1$. This is the only way $\alpha$ can be assured to be 0. Similarly, $\beta = 0$ if $d = 1$, but then $\alpha = 1$. Now, $\alpha = 1/2$ and $\beta = 1/2$ can be obtained by flipping a coin and ignoring the data. Thus we have 3 points on the “frontier” available without data.

The “information budget constraint” with no data is the solid line in the following figure:
Good tests using data will get a constraint like the curve (of course, \((0, 1)\) and \((1, 0)\) are always the endpoints).  \textit{(Exercise: Why does this constraint have this shape?)}
This is like an income effect - information gives a better tradeoff between the two types of errors.

**Definition:** $p_0/p_1$ is the likelihood ratio where $p_i = f(x|\theta_i)$ is the joint distribution of data.

Let $A(T) = \{x: p_0/p_1 > T\}$ (a set in $R^n$) and $\alpha^* = \int_{A^c} p_0(x)dx$; $\beta^* = \int_A p_1(x)dx$.

$A$ defines a decision rule $d = 0$ if $x \in A$ and $d = 1$ if $x \in A^c$.

Let $B$ be any other region in $R^n$ with error probabilities $\alpha$ and $\beta$. Then:

**Neyman-Pearson Lemma:**
If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

What does this say?
Proof: Define $I_A(x) = 1$ if $x \in A$ and $I_B(x) = 1$ if $x \in B$. Then $(I_A - I_B)(p_0(x) - Tp_1(x)) \geq 0$.

To check this, look at both cases: If $x \in A$, then $I_A = 1$ and $p_0/p_1 > T$.... (think about this)

Multiplication yields:

$$0 \leq I_A p_0 - I_A Tp_1 - I_B p_0 + I_B Tp_1.$$

If this holds for any given $x$, it certainly holds on the average. Thus

$$0 \leq \int_A p_0 - Tp_1 \, dx - \int_B p_0 - Tp_1 \, dx.$$

Hence (recall definitions of $\alpha$, $\beta$, $\alpha^*$, $\beta^*$),

$$0 \leq (1 - \alpha^*) - T \beta^* - (1 - \alpha) + T \beta = T(\beta - \beta^*) + (\alpha - \alpha^*).$$

Thus, if $\beta < \beta^*$, $\alpha$ must be $> \alpha^*$, and vice versa. ■
The result says that when designing tests we should look at the likelihood ratio.

*Indifference curves for error probabilities:*

Let \((\alpha^0, \beta^0)\) and \((\alpha^1, \beta^1)\) be error probabilities associated with two different tests. Suppose you are indifferent between these tests, then you do not care if the choice is made with a coin flip.

But this defines another test with error probabilities \(\alpha^2 = 1/2\alpha^0 + 1/2\alpha^1\) and \(\beta^2 = 1/2\beta^0 + 1/2\beta^1\), and you are indifferent between this new test and the others. Continuing, you derive a linear indifference curve.
Note that the practice of fixing $\alpha$ (e.g., 0.05) for all sample sizes ($\Rightarrow$ all values of $\beta$) corresponds to lexicographic preferences, which are not continuous and therefore illogical in this setting.

**Example:** Consider the following composite hypothesis:

$H_0$: $\theta = \theta_0$ (null hypothesis)
$H_1$: $\theta \neq \theta_0$ (alternative hypothesis)

Here we find the ML estimator $\hat{\theta}$ and consider the likelihood ratio $f(x|\theta_0)/f(x|\hat{\theta})$. Basically we are choosing the "best" value under the alternative hypothesis for the denominator.

**Exercise:** Consider the regression model

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. Is the $F$-test for $\beta_2 = 0$ a likelihood ratio test?
Asymptotic Testing:

In this section, we will study the three tests: Likelihood Ratio (LR), Wald and Score (Lagrange Multiplier - LM) tests.

**Background: (Asymptotics)**

\[ \ell(\theta) = \sum \ln p(x|\theta) \] is the log likelihood function. Define the score function

\[ s(\theta) = \frac{d\ell}{d\theta} \]

and

\[ i(\theta) = -E \left[ \frac{d^2 \ln p}{d\theta^2} \right] = E \left[ \left( \frac{d \ln p}{d\theta} \right)^2 \right]. \]
By CLT,

\[
\frac{1}{\sqrt{n}} s_0 \sim N(0, i_0)
\]

where \( \theta_0 \) is the true value, \( s_0 = s(\theta_0) \) and \( i_0 = i(\theta_0) \).

**Testing:**

Let \( \hat{\theta} \) be the ML estimator. Let \( d_0 = \hat{\theta} - \theta_0 \) denote the vector of deviations.

Then, \( n^{-1/2} s_0 = i_0 d_0 n^{1/2} \) asymptotically. Note that this is the same as

\[
n^{1/2} d_0 = i_0^{-1} s_0 n^{-1/2}.
\]

Further, \( 2[\ell(\hat{\theta}) - \ell(\theta_0)] = nd'_0 i_0 d_0 \) asymptotically. (*To get this result, expand \( \ell(\hat{\theta}) \) around \( \theta_0 \) and take probability limits.*)
Consider the hypothesis:

\( H_0: \theta = \theta_0 \)
\( H_1: \theta \neq \theta_0 \)

Note that the restriction is \( \theta = \theta_0 \).

**Likelihood Ratio Test:**

Likelihood ratio:

\[
LR = \frac{p(x|\theta_0)}{\max_\theta p(x|\theta)} = \frac{p(x|\theta_0)}{p(x|\hat{\theta})}
\]

The test statistic is \(-2 \ln LR = 2[\ell(\hat{\theta}) - \ell(\theta_0)]\) and it is distributed as \( \chi^2 \) (with degrees of freedom equal to the number of restrictions imposed) under the null hypothesis.
**Wald Test:**

The test statistic is \( nd_0 i(\hat{\theta})d_0 \), and it is distributed as \( \chi^2 \) under the null hypothesis.

**Score Test:**

The test statistic is \( n^{-1}s_0' i_0^{-1} s_0 \), and it is distributed as \( \chi^2 \) under the null hypothesis.

Note: \( p\lim i(\hat{\theta}) = i(\theta_0) = i_0 \) when the restriction is true and real that

\[
p\lim(n d_0' i_0 d_0 - n^{-1}s_0' i_0^{-1} s_0) = 0 \text{ since asymptotically}
\]

\[
n^{1/2}d_0 = i_0^{-1}s_0 n^{-1/2}
\]
So, the tests are asymptotically equivalent. Note that the Wald and LM tests are appealing because of their asymptotic equivalence to the LR test, which is an optimal test in the Neyman-Pearson sense.

Discussion:

- What are the computational requirements for these tests?
- Which is best?
For illustrative purposes, $\theta$ is one-dimensional.

Here, we look at the change in the log likelihood function $\ell(\theta)$ evaluated at $\hat{\theta}$ and $\theta_0$, $\ell(\hat{\theta})$ and $\ell(\theta_0)$. If the difference between is too large, we reject $H_0$. 
LR test is based on this difference
Here, we look at the deviation in parameter space.

The difference between $\hat{\theta}$ and $\theta_0$ implies a larger difference between $\ell(\hat{\theta})$ and $\ell(\theta_0)$ for the more curved log likelihood function. Evidence against the hypothesized value $\theta_0$ depends on the curvature of the log likelihood function measured by $ni(\hat{\theta})$.

Hence the test statistic is $n(\hat{\theta} - \theta_0)^2 i(\hat{\theta})$. 
* Wald test is based on this difference
Here, we look at the slope of the log likelihood function at the hypothesized value of $\theta_0$.

Since two log likelihood functions can have equal values of $s_0$ with different distances between $\hat{\theta}$ and $\theta_0$, $s_0$ must be weighed by the change in slope (i.e. curvature). A bigger change in slope implies less evidence against the hypothesized value $\theta_0$.

Hence the test statistic $n^{-1}s_0^2i_0^{-1}$. 
* Score(LM) test is based on this difference
Why is the score test also called the Lagrange Multiplier test?

The log likelihood function is maximized subject to the restriction \( \theta = \theta_0 \):

\[
\max_{\theta} \ell(\theta) - \lambda(\theta - \theta_0).
\]

This gives

\[
\hat{\theta} = \theta_0 \text{ and } \lambda = s(\theta_0) = \frac{\partial \ell}{\partial \theta_0}.
\]
* 2 likelihood functions and a test of $\Theta = \Theta_0$