3.1 Introduction

One of the simplest inference problems is that of deciding between two options (hypotheses).

**Example 3.1** (Healthy vs. Diabetic). The blood glucose level (in mg/dL) of a healthy person can be modeled as $\mathcal{N}(95, \sigma^2)$, while that of a diabetic can be modeled as $\mathcal{N}(140, \sigma^2)$. Given a new patient with glucose level $x$, you want to decide between two hypotheses:

- $H_0 : x \sim \mathcal{N}(95, \sigma^2) \implies$ healthy,
- $H_1 : x \sim \mathcal{N}(140, \sigma^2) \implies$ diabetic.

$H_0$ and $H_1$ are often called null and alternative hypotheses.

**Example 3.2** (Radar). A radar is constantly emitting a signal and monitoring to see if it bounces back (see Figure 3.1). The signal $x$ that the radar receives can be modeled as $\mathcal{N}(0, \sigma^2)$ if there is nothing (hence the signal doesn’t bounce back) and $\mathcal{N}(\mu, \sigma^2)$ for some $\mu > 0$ if an object is present (hence signal bounces back). Thus it needs to decide between:

- $H_0 : x \sim \mathcal{N}(0, \sigma^2) \implies$ nothing there,
- $H_1 : x \sim \mathcal{N}(\mu, \sigma^2), \ \mu > 0 \implies$ something there.

**Example 3.3** (Astrophysics). The NASA wants you to determine whether two meteorites — one that fell in Roswell, New Mexico, and one that fell in Chelyabinsk, Russia — came from the same asteroid in space. With help from the materials expert in your interdisciplinary team, you are able to determine that if two meteorites come from the same asteroid, the difference $x$ of their magnesium composition

Figure 3.1: A radar is constantly receiving a signal, and needs to decide whether an object is present or not. See Example 3.2.
can be modeled as \( \mathcal{N}(0, \sigma^2) \), and \( \mathcal{N}(\mu, \sigma^2) \) otherwise for some unknown \( \mu \neq 0 \). Hence you need to decide between:

\[
H_0 : x \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \text{same asteroid,}
\]
\[
H_1 : x \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \neq 0 \quad \Rightarrow \text{different asteroids.}
\]

**Example 3.4 (Genetics).** Have you wondered how geneticists determine which genes are associated to which diseases? Essentially, they compare the average activation levels of a gene in healthy and sick individuals (see Figure 3.2). The difference \( x \) between these activation levels can be modeled as \( \mathcal{N}(0, \sigma^2) \) if the gene is unrelated to the disease, and \( \mathcal{N}(\mu, \sigma^2) \) for some unknown \( \mu \neq 0 \) if the gene is related to the disease. We thus have to decide between:

\[
H_0 : x \sim \mathcal{N}(0, \sigma^2) \quad \Rightarrow \text{gene unrelated to disease,}
\]
\[
H_1 : x \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \neq 0 \quad \Rightarrow \text{gene related to disease.}
\]

**Example 3.5 (Treatment design).** Scientists often want to design a treatment (e.g., a drug or procedure) for a disease (e.g., diabetes or cancer). To this end they measure the disease presence (e.g., glucose level or tumor size) before and after treatment in \( N \) patients. The differences \( x_i \) can be modeled as independent and identically distributed (i.i.d.) \( \mathcal{N}(0, \sigma^2) \) if the treatment is ineffective, and \( \mathcal{N}(\mu, \sigma^2) \) for some \( \mu < 0 \) if the treatment is effective. Hence we have

\[
H_0 : x_1, \ldots, x_N \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad \Rightarrow \text{treatment is ineffective,}
\]
\[
H_1 : x_1, \ldots, x_N \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \mu < 0 \quad \Rightarrow \text{treatment is effective.}
\]

**Example 3.6 (Neural activity).** Scientists want to determine which regions of the brain are related to certain tasks using functional magnetic resonance imaging (fMRI), which essentially creates a video of the brain using magnetic fields that map hydrogen density. For example, say they want to know which region of the brain controls the thumb. Then they take an individual, ask her to move her thumb
periodically, and take an fMRI video of her brain. Then scientists analyze one pixel at a time. The 
(i,j)th pixel will produce a signal vector $x_{ij} \in \mathbb{R}^D$ containing the brain measurements in that pixel over time. Some pixels will show 
neural activity correlated with the signal of the thumb (see Figure 3.3). Then $x_{ij}$ can be modeled as 
$N(\mu, \sigma^2 I)$ if the pixel is correlated to the thumb movement, and $N(0, \sigma^2 I)$ if the pixel is uncorrelated 
($I$ denotes the identity matrix of compatible size, in this case $D \times D$). Hence for each pixel $(i,j)$ they 
have to decide:

$$H_0 : x_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma^2 I) \quad \Rightarrow (i,j)\text{th pixel is uncorrelated},$$

$$H_1 : x_{ij} \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2 I) \quad \Rightarrow (i,j)\text{th pixel is correlated}.$$
Remark 3.1 (Likelihood). The term likelihood is often a source of confusion. To be more precise, in hypothesis testing we observe an instance of a random variable, i.e., we observe data $x = x$, and we want to decide which of two distributions ($p_0$ or $p_1$) is more likely to have generated this data. The likelihood that $p_0$ generated $x$ is essentially $p_0(x)$ [evaluated at $x = x$], and similarly for $p_1$.

Example 3.7 (Radar). Consider the hypothesis problem in Example 3.2. Then

$$
\Lambda(x) = \frac{p_1(x)}{p_0(x)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x^2 - \mu^2}{2\sigma^2} \right)}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x^2 - 0^2}{2\sigma^2} \right)}} = e^{\frac{\mu^2 - 2x\mu + x^2}{2\sigma^2}} \leq \frac{1}{n_0}.
$$

Since both sides are positive, taking log we obtain:

$$
\frac{\mu(2x - \mu)}{2\sigma^2} \geq \frac{x}{n_0}.
$$

and since $\mu > 0$, this further simplifies into the following test:

$$
x \geq \frac{\mu}{\frac{n_0}{2}}.
$$

In words, this LRT tells us to decide $H_1$ if our observed data $x$ is larger than $\mu/2$, and $H_0$ otherwise (see Figure 3.4).

3.3 Outcomes and Decision Regions

A test has four possible outcomes, depending on what we decide and the truth: $(0|0), (0|1), (1|0), (1|1)$. See Table 3.1. Sometimes it is desirable to reduce the probability of one particular kind of error.
Example 3.8. In Example 3.5, scientists want to avoid a (1|0) error, which would mean that they believe their treatment cures a disease, when it really doesn’t.

The probabilities of the four outcomes are determined by the regions of the test.

Definition 3.2 (Decision regions). The decision regions $R_0$ and $R_1$ of a test $t : \Omega \rightarrow \{H_0, H_1\}$ are the inverse images of $H_0$, and $H_1$, i.e.,

\[
R_0 := \{ x \in \Omega : t(x) = H_0 \}, \\
R_1 := \{ x \in \Omega : t(x) = H_1 \}.
\]

In words, $R_0$ is the region of the domain of $x$ where we will decide $H_0$, and similarly for $R_1$.

Example 3.9. The decision regions of the test $x \overset{H_1}{\underset{H_0}{\gtrless}} \mu/2$ are $R_0 = (-\infty, \mu/2)$ and $R_1 = (\mu/2, \infty)$.

Decision regions determine the probability of each outcome as follows:

\[
p_{gh} := \int_{R_g} p_h(x) \, dx, \text{ for } g, h \in \{0, 1\}.
\]

Example 3.10. The test $x \overset{H_1}{\underset{H_0}{\gtrless}} \mu/2$ has the following outcomes probabilities (see Figure 3.5):

\[
P_{00} = Q_{0, \sigma^2}(-\mu/2), \\
P_{01} = Q_{\mu, \sigma^2}(-\mu/2), \\
P_{10} = Q_{0, \sigma^2}(\mu/2), \\
P_{11} = Q_{\mu, \sigma^2}(\mu/2).
\]

\[
\begin{array}{c|cc|c|cc|c}
& H_0 & H_1 \\
\hline
H_0 & (0|0) & False negative & (0|1) & False positive & (1|0) & Type 2 error \\
& True Negative & Miss & & False alarm & & \\
& No-alarm & & & & Type 1 error & \\
& Accept & & & & Detect & \\
\hline
H_1 & (1|0) & False positive & (1|1) & True Positive & & \\
& False alarm & & & & Detect & \\
& & Type 1 error & & & & \\
\end{array}
\]

Table 3.1: Four possible outcomes of a test. Depending on the field, they might come under different names. We will use (0|0), (0|1), (1|0), (1|1) for simplicity.
where \( Q_{\mu,\sigma^2}(\tau) \) is the tail probability of the \( N(\mu, \sigma^2) \) distribution, i.e.,

\[
Q_{\mu,\sigma^2}(\tau) := \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.
\]

If we want to bound \( p_{10} \), we could modify our test to

\[
x \geq \frac{\mu}{\bar{r}_n} \tau,
\]

where \( \tau \) is a threshold selected to make \( p_{10} \) smaller than the desired probability of error \( \alpha \) (see Figure 3.5). This new test has \( p_{10} = Q_{\mu,\sigma^2}(\tau) \). Furthermore, a simple change of variable shows that

\[
Q_{\mu,\sigma^2}(\tau) = Q\left(\frac{\tau-\mu}{\sigma}\right),
\]

(3.1)

where we use \( Q \) as shorthand for \( Q_{0,1} \) (see Figure 3.6 to build some intuition), so \( p_{10} = Q(\tau/\sigma) \). Finally, since \( Q \) is invertible, if we want \( p_{10} \leq \alpha \), we can pick \( \tau = \sigma Q^{-1}(\alpha) \).

### 3.4 Neyman-Pearson Lemma

As mentioned in Example 3.8, there are some cases where we want to bound a certain probability of error, say \( p_{10} \). One way to do this is by increasing \( R_0 \). However, as \( R_0 \) grows, our accuracy \( p_{11} \) decreases (see
Figure 3.5 to build some intuition). Neyman-Pearson’s Lemma tells us that the LRT is optimal in the sense that there exists no other test that has lower probability of error $p_{10}$ and higher accuracy $p_{11}$.

**Lemma 3.1** (Neyman-Pearson). Consider the likelihood ratio test $t$ given by

$$\frac{p_1(x)}{p_0(x)} \overset{H_1}{\gtrless} \tau,$$

with $\tau$ chosen such that $p_{10} = \alpha$. Then there exists no other test $t'$ with $p'_{10} \leq \alpha$ and $p'_{11} > p_{11}$.

**Proof.** For any region $R \subset \Omega$, let $P_h(R)$ be the cumulative probability of $p_h(x)$ over $R$, i.e.,

$$P_h(R) = \int_R p_h(x)dx.$$

Then for $h \in \{0, 1\}$, let

$$p_{1h} = P_h(R_1 \cap R'_1) + P_h(R_1 \cap R'_0),$$

$$p'_{1h} = P_h(R_1 \cap R'_1) + P_h(R_0 \cap R'_1).$$

(3.2)

Now suppose $p'_{10} \leq \alpha$. We need to show that $p_{11} \geq p'_{11}$. From (3.2), this is equivalent to showing that $P_1(R_1 \cap R'_0) \geq \tau P_0(R_1 \cap R'_0)$, and so write

$$P_1(R_1 \cap R'_0) = \int_{R_1 \cap R'_0} p_1(x)dx \geq \int_{R_1 \cap R'_0} \tau p_0(x)dx = \tau P_0(R_1 \cap R'_0),$$

(3.3)

where the inequality follows because in $R_1$, $p_1(x) \geq \tau p_0(x)$. By assumption, $p_{10} = \alpha \geq p'_{10}$. This, together with (3.2) imply $P_0(R_1 \cap R'_0) \geq P_0(R_0 \cap R'_1)$, so

$$(3.3) \geq \tau P_0(R_0 \cap R'_1) = \int_{R_0 \cap R'_1} \tau p_0(x)dx \geq \int_{R_0 \cap R'_1} p_1(x)dx = P_1(R_0 \cap R'_1),$$

where the last inequality follows because in $R_0$, $p_1(x) \leq \tau p_0(x)$. \qed

**Example 3.11.** Consider

$$H_0 : x \sim N(0, \sigma_0^2),$$

$$H_1 : x \sim N(0, \sigma_1^2),$$

where $\sigma_0 < \sigma_1$ are known. The likelihood ratio test is

$$\Lambda(x) = \frac{p_1(x)}{p_0(x)} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} \overset{H_1}{\gtrless} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{x^2}{2\sigma_0^2}} = \frac{\sigma_0}{\sigma_1} \left( \frac{x^2}{\sigma_0^2} - \frac{x^2}{\sigma_1^2} \right) \overset{H_0}{\gtrless} \tau.$$
Topic 3: Hypothesis Testing

Figure 3.7: Left: Threshold $\tau'$ selected to achieve probability of error $p_{10} = \alpha$ in test $x^2 \gtrsim \frac{H_0}{H_1} \tau'$ of Example 3.11, where $H_0: x^2 / \sigma_0^2 \sim \chi^2$. This is equivalent to the test $|x| \gtrsim \frac{H_0}{H_1} \sqrt{\tau'}$ with $H_0: x \sim N(0, \sigma_0^2)$, as in the Right.

Or equivalently,

$$\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1} x^2 \gtrsim \frac{H_1}{H_0} \log \left( \frac{\sigma_1}{\sigma_0} \right),$$

$$x^2 \gtrsim \frac{H_1}{H_0} \sqrt{\tau'} \left( \frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \log \left( \frac{\sigma_1}{\sigma_0} \right) \right).$$

Now recall that if $y \sim N(0,1)$, then $y^2 \sim \chi^2$. So we can rewrite our hypotheses as

$$H_0: \ (z/\sigma_0)^2 \sim \chi^2,$$

$$H_1: \ (z/\sigma_1)^2 \sim \chi^2.$$

Then $p_{1h}$ (with $h \in \{0,1\}$) is simply the probability that a $\sigma_h^2$-scaled $\chi^2$ random variable is larger than $\tau'$ (see Figure 3.7), i.e.,

$$p_{1h} = Q_{\chi^2} (\tau'/\sigma_h^2),$$

where $Q_{\chi^2}$ is the tail probability of the $\chi^2$ distribution. Since $Q_{\chi^2}$ is invertible, if we want $p_{10} \leq \alpha$, we can pick $\tau' = \sigma_0^2 Q_{\chi^2}^{-1}(\alpha)$, and then $p_{11} = Q_{\chi^2} (\sigma_0^2/\sigma_1^2 Q_{\chi^2}^{-1}(\alpha))$. Neyman-Pearson’s Lemma tells us that there exists no other test that has lower probability of error $p_{10}$ and higher accuracy $p_{11}$.

### 3.5 Multiple Observations

We now study what happens when we have several observations instead of just one.

**Example 3.12.** Consider the hypotheses in Example 3.5, or equivalently, in vector form:

$$H_0: \ x \in \mathbb{R}^N \sim N(0, \sigma^2 I),$$

$$H_1: \ x \in \mathbb{R}^N \sim N(\mu 1, \sigma^2 I), \ \mu < 0,$$
where $\mathbf{1}$ denotes the all ones vector of compatible size, in this case $N \times 1$. The likelihood ratio is given by

$$
\Lambda(x) = \frac{p_1(x)}{p_0(x)} = \frac{\frac{1}{(\sqrt{2 \pi} \sigma)^N} e^{-\frac{1}{2\sigma^2}(x - \mu_1)^T(x - \mu_1)}}{\frac{1}{(\sqrt{2 \pi} \sigma)^N} e^{-\frac{1}{2\sigma^2}x^T x}} = e^{\frac{\mu^T (2x^T - \mu_1)}{2N\sigma^2}}.
$$

Taking log we obtain the log-likelihood ratio test:

$$
\frac{\mu}{2\sigma^2} (2x^T \mathbf{1} - N\mu) \geq \log \tau,
$$

$$
x^T \mathbf{1} \leq \frac{\sigma^2}{\mu} \log \frac{\tau}{\tau'} + \frac{N\mu}{2}.
$$

Notice that the direction of the inequalities in the test was inverted because $\mu < 0$. Next observe that $m = x^T \mathbf{1} = \sum_{i=1}^{N} x_i$, and since sums of gaussians are gaussians, we can rewrite our hypotheses as

$$
H_0: m \sim N(0, N\sigma^2),
$$

$$
H_1: m \sim N(N\mu, N\sigma^2), \quad \mu < 0,
$$

Then our log-likelihood ratio test becomes $m \leq \frac{\mu}{2\sigma^2} \tau'$, and since $\tau' < 0$,

$$
p_{10} = \Phi_{0,N\sigma^2}(\tau'),
$$

$$
p_{11} = \Phi_{N\mu,N\sigma^2}(\tau').
$$

where $\Phi_{\mu,\sigma^2}$ is the cumulative distribution function (CDF) of a $N(\mu, \sigma^2)$ random variable (see Figure 3.8). Equivalently, with a similar transformation as (3.1), we can write this in terms of the CDF $\Phi$ of the standard normal $N(0,1)$,

$$
p_{10} = \Phi\left(\frac{\tau'}{\sqrt{N}\sigma}\right),
$$

$$
p_{11} = \Phi\left(\frac{\tau' - N\mu}{\sqrt{N}\sigma}\right).
$$

(3.4)

Since $\Phi$ is invertible, if we want $p_{10} \leq \alpha$, we can pick $\tau' = \sqrt{N}\sigma\Phi^{-1}(\alpha)$. Plugging this in (3.4), we obtain

$$
p_{11} = \Phi\left(\frac{\sqrt{N}\sigma\Phi^{-1}(\alpha) - N\mu}{\sqrt{N}\sigma}\right) = \Phi\left(\Phi^{-1}(\alpha) - \frac{\sqrt{N}\mu}{\sigma}\right).
$$

$\sqrt{N}\mu/\sigma$ is often known as the signal to noise ratio. Since $\Phi(\tau) \to 1$ as $\tau \to \infty$, and since $\mu < 0$ by assumption, it is easy to see that $p_{11}$ increases with $N$ and $|\mu|$, but decreases with $\sigma$.

### 3.6 Multiple Testing

In many applications we run multiple tests, and we want to bound the probability of making one or more mistakes.
Topic 3: Hypothesis Testing

Example 3.13. In Example 3.6 we have a family of $K$ tests (where $K$ is the number of pixels), and we want to be confident that all the identified pixels are truly correlated to the thumb’s movement.

Definition 3.3 (Family-wise error rate (FWER)). The family-wise error rate is the probability of making one or more $(1|0)$ errors. More precisely, for a family of $K$ tests,

$$FWER = P\left(\bigcup_{k=1}^{K} \{1_k \mid 0_k\}\right),$$

where $\{1_k \mid 0_k\}$ denotes the event of deciding $H_1$ in the $k^{th}$ test given that $H_0$ is true.

Lemma 3.2 (Bonferroni Correction). Consider a family of $K$ tests. Setting $p_{10} = \frac{\alpha}{K}$ for each test achieves $FWER < \alpha$.

Proof. As a simple consequence of the union bound, we have:

$$FWER = P\left(\bigcup_{k=1}^{K} \{1_k \mid 0_k\}\right) \leq \sum_{k=1}^{K} P(\{1_k \mid 0_k\}) = \sum_{k=1}^{K} p_{10} = K \frac{\alpha}{K} = \alpha.$$

\[\square\]

3.7 Composite Hypotheses

So far we have studied simple hypotheses where all distributions and their parameters are known, as in Examples 3.1, 3.6 and 3.11, where the distributions and their parameters are known. However, in many practical situations this is not the case. For instance, in $H_1$ of Example 3.2, all we know is that $\mu > 0.$
In this case, \( H_1 \) is composed of the collection of distributions \( \{N(\mu, \sigma^2)\}_{\mu > 0} \). More generally, a composite problem has the form:

\[
H_0 : x \sim p_0(x|\theta_0), \quad \theta_0 \in \Theta_0, \quad H_1 : x \sim p_1(x|\theta_1), \quad \theta_1 \in \Theta_1,
\]

where the notation \( p_h(x|\theta_h) \) means that \( \theta_h \) is a parameter of the distribution \( p_h \), and \( \Theta_h \) is a set of all possible values of the parameter \( \theta_h \). In general, \( p_0 \) and \( p_1 \) may be entirely different distributions, and the sets \( \Theta_0 \) and \( \Theta_1 \) may be entirely different.

Even though Examples 3.2 and 3.5 are composite problems, since we know \( \mu > 0 \), we were still able to derive its LRT in Examples 3.7 and 3.12. This is not always the case. There are some more complicated cases, like Examples 3.3 and 3.4, where \( \mu \) is completely unknown, and this can complicate things.

**Example 3.14** (Wald’s test). Consider Examples 3.3 and 3.4. The likelihood ratio is

\[
\Lambda(x) = \frac{p_1(x)}{p_0(x)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}}{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{x-0}{\sigma} \right)^2}} = e^{\frac{\mu^2}{2\sigma^2}} \frac{h_1}{h_0} \geq 1.
\]

Taking log and with some minor algebra we obtain:

\[
x \mu \ \overset{h_1}{\gtrless} \ \overset{h_0}{\mu}^2.
\]

However, since we don’t know the sign of \( \mu \), we cannot continue as in Example 3.7, as dividing by \( \mu \) could reverse the direction of the inequalities in the test. Hence this test is uncomputable, or undetermined. So how can we proceed? For example, let’s say we decide to use the test from Example 3.10:

\[
x \ \overset{h_1}{\gtrless} \ \overset{h_0}{\tau},
\]

It could happen that we are lucky and \( \mu > 0 \). Then our test will be optimal (as shown by Neyman-Pearson’s Lemma) with \( p_{10} = Q(\tau/\sigma) \) and \( p_{11} = Q(\tau - \mu/\sigma) \) (see Example 3.10 and Figure 3.5). However, if we are unlucky and \( \mu < 0 \), our test would be doing something terribly insensible, and would have terrible accuracy \( p_{11} = Q(\tau + \mu/\sigma) \); see Figure 3.9 to build some intuition. One good compromise is to use Wald’s test:

\[
|x| \ \overset{h_1}{\gtrless} \ \overset{h_0}{\tau},
\]

which has

\[
p_{10} = 2Q(\tau/\sigma), \quad p_{11} = Q\left(\frac{\tau - \mu}{\sigma}\right) + Q\left(\frac{\tau + \mu}{\sigma}\right).
\]

Wald’s test is not optimal, but is sensible. It has higher probability of error \( p_{10} \) than if we are lucky and guess \( \mu \) correctly, but also has higher accuracy \( p_{11} \) than if we are unlucky and guess \( \mu \) incorrectly.
Figure 3.9: Composite hypothesis test where $\mu \neq 0$ is unknown. **Left:** Probabilities $p_{10}$ and $p_{11}$ of the test $x \gtrless \bar{X} \tau$. If we are lucky and $\mu > 0$, this test will be optimal, but if $\mu < 0$, our test will be terrible. **Right:** Wald’s test $|x| \gtrless \bar{X} \tau$. Wald’s test is not optimal, but is sensible. It has higher probability of error $p_{10}$ than if we are lucky and guess $\mu$ correctly, but also has higher accuracy $p_{11}$ than if we are unlucky and guess $\mu$ incorrectly. See Example 3.14 and Figure 3.10.

Figure 3.10: **Left:** ROC curves of the test $x \gtrless \bar{X} \tau$ for different values of $\mu$. Consistent with our analysis from Example 3.12, we can see that $p_{11}$ grows with $\mu$. **Right:** ROC curves for the test $x \gtrless \bar{X} \tau$ when $\mu > 0$ (optimal), when $\mu < 0$ (terrible), and for Wald’s test. This shows that Wald’s test is suboptimal but sensible. See Example 3.14.

### 3.8 ROC Curves

As shown in Example 3.14, it is not always possible to devise an optimal test. It is thus reasonable to ask how good a test is. For example, how good is Wald’s test? One way to do this is with Receiver Operating Characteristic (ROC) curves, which measure a test’s performance by plotting its $p_{11}$ as a function of its $p_{10}$. ROC curves are widely used in laboratories to measure a test’s ability to discriminate diseased cases from normal cases, and also to compare the performance of two or more tests.

### 3.9 Generalized Likelihood Ratio Test

Wald’s test was an intuitive solution to the simplest composite problem. However, Wald’s test has a solid statistical foundation. In fact, Wald’s test is the result of estimating $\mu$ and then using this estimate in a likelihood ratio test. We will come back to Wald’s test and its generalization, introducing the generalized likelihood ratio test (GLRT), but first we will need to learn about estimation, which is our next topic.