CS 8850: Advanced Machine Learning **Fall 2017**

Topic 3: Hypothesis Testing

Instructor: Daniel L. Pimentel-Alarcón

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:

$$\begin{aligned} H_0 : \ x \ &\sim \ \mathcal{N}(95, \sigma^2) \qquad \Rightarrow \text{healthy}, \\ H_1 : \ x \ &\sim \ \mathcal{N}(140, \sigma^2) \qquad \Rightarrow \text{diabetic}. \end{aligned}$$

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:

$$\begin{aligned} H_0 : \ x \ &\sim \ \mathcal{N}(0, \sigma^2) \qquad\qquad \Rightarrow \text{nothing there}, \\ H_1 : \ x \ &\sim \ \mathcal{N}(\mu, \sigma^2), \quad \mu > 0 \qquad \Rightarrow \text{something there}. \end{aligned}$$

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

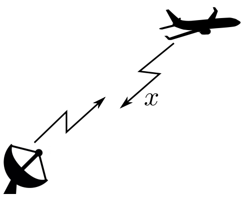

Figure 3.1: A radar is constantly receiving a signal, and needs to decide whether an object is present or not. See Example 3.2.

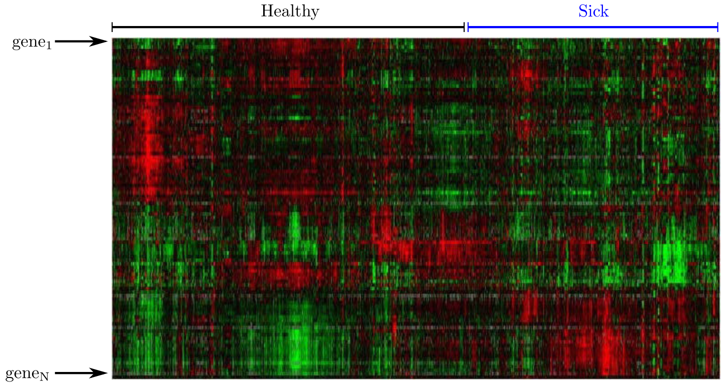

Figure 3.2: Gene microarrays are data matrices indicating gene *activation levels*. Each row corresponds to one gene, and each column corresponds to one individual. We want to know which genes are related to a disease. See Example 3.4.

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:

$$\begin{aligned} H_0 : \ x &\sim \mathcal{N}(0, \sigma^2) && \Rightarrow \text{same asteroid,} \\ H_1 : \ x &\sim \mathcal{N}(\mu, \sigma^2), \quad \mu \neq 0 && \Rightarrow \text{different asteroids.} \end{aligned}$$

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:

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

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

$$\begin{aligned} H_0 : \ x_1, \dots, x_N &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2) && \Rightarrow \text{treatment is ineffective,} \\ H_1 : \ x_1, \dots, x_N &\overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \mu < 0 && \Rightarrow \text{treatment is effective.} \end{aligned}$$

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

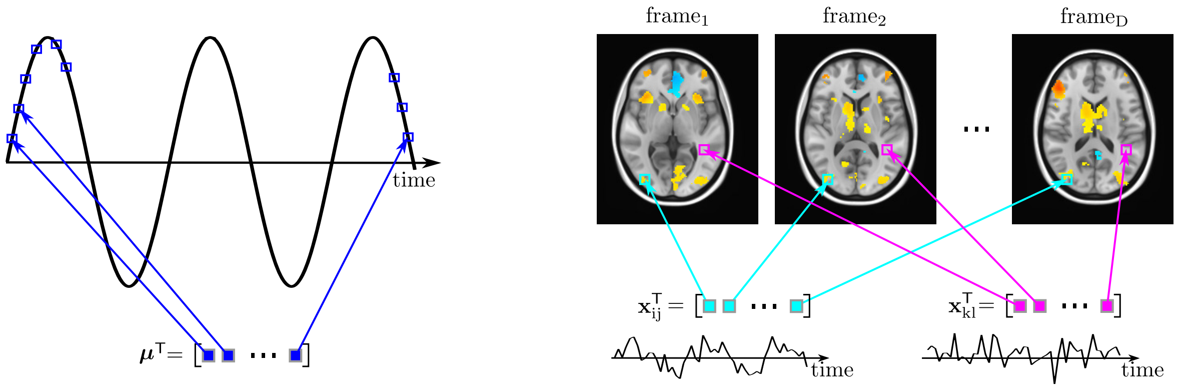
time
$\boldsymbol{\mu}^{\mathsf{T}} = [\cdots]$
frame$_1$
frame$_2$
frame$_{\mathrm{D}}$
$\cdots$
$\mathbf{x}_{\mathrm{ij}}^{\mathsf{T}} = [\cdots]$
$\mathbf{x}_{\mathrm{kl}}^{\mathsf{T}} = [\cdots]$
time
time

Figure 3.3: **Left:** Signal of the thumb $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{D}}$, usually a sinusoid with the periodicity of the thumb movement. **Right:** Each pixel produces one signal vector $\boldsymbol{x}_{\mathrm{ij}} \in \mathbb{R}^{\mathrm{D}}$ containing the brain measurements in that pixel over time. Some pixels may show neural activity correlated with the signal of the thumb. We want to find such pixels. See Example 3.6.

periodically, and take an fMRI video of her brain. Then scientists analyze one pixel at a time. The $(\mathrm{i},\mathrm{j})^{\mathrm{th}}$ pixel will produce a signal vector $\boldsymbol{x}_{\mathrm{ij}} \in \mathbb{R}^{\mathrm{D}}$ containing the brain measurements in that pixel over time. Some pixels will show neural activity correlated with the signal of the thumb $\boldsymbol{\mu} \in \mathbb{R}^{\mathrm{D}}$, usually a sinusoid with the periodicity of the thumb movement (see Figure 3.3). Then $\boldsymbol{x}_{\mathrm{ij}}$ can be modeled as $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ if the pixel is uncorrelated to the thumb movement, and $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ if the pixel is correlated ($\mathbf{I}$ denotes the identity matrix of compatible size, in this case $\mathrm{D} \times \mathrm{D}$). Hence for each pixel (i,j) they have to decide:

$$\begin{aligned} H_0 :\ \boldsymbol{x}_{\mathrm{ij}} &\overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) && \Rightarrow (\mathrm{i},\mathrm{j})^{\mathrm{th}} \text{ pixel is uncorrelated,} \\ H_1 :\ \boldsymbol{x}_{\mathrm{ij}} &\overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}) && \Rightarrow (\mathrm{i},\mathrm{j})^{\mathrm{th}} \text{ pixel is correlated.} \end{aligned}$$

Definition 3.1 (Hypothesis test)**.** A *hypothesis test* is a function $t : \Omega \to \{H_0, H_1\}$.

3.2 The Likelihood Ratio Test

In general, hypothesis testing is all about deciding between two options. We observe a random variable x, and want to decide whether

$$\begin{aligned} H_0 :\ x &\sim \mathsf{p}_0(\mathrm{x}), \\ H_1 :\ x &\sim \mathsf{p}_1(\mathrm{x}). \end{aligned}$$

If your hunch is to simply pick whichever is larger between $\mathsf{p}_0(\mathrm{x})$ and $\mathsf{p}_1(\mathrm{x})$, your intuition is correct. That is essentially the *likelihood ratio test* (LRT) in its most elemental form:

$$\frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})} \underset{H_0}{\overset{H_1}{\gtrless}} 1,$$

which in words means: let's make a *test*: if the *likelihood ratio* $\Lambda(\mathrm{x}) := \frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})}$ is larger than 1 (meaning p_1 is larger), then pick H_1. Similarly, if $\Lambda(\mathrm{x}) < 1$ (meaning p_0 is larger), then pick H_0.

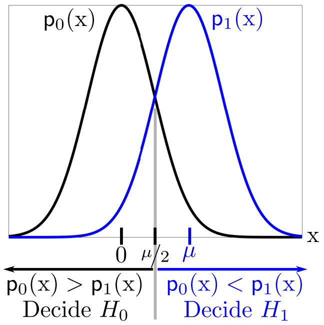

Figure 3.4: Likelihood ratio test x $\underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mu}{2}$ in Example 3.7.

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 $x = \mathrm{x}$, 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 $\mathsf{p}_0(\mathrm{x})$ [evaluated at $x = \mathrm{x}$], and similarly for p_1.

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

$$\Lambda(\mathrm{x}) \;=\; \frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})} \;=\; \frac{\cancel{\frac{1}{\sqrt{2\pi}\sigma}}e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu}{\sigma}\right)^2}}{\cancel{\frac{1}{\sqrt{2\pi}\sigma}}e^{-\frac{1}{2}\left(\frac{\mathrm{x}-0}{\sigma}\right)^2}} \;=\; \frac{e^{-\frac{\cancel{\mathrm{x}^2}-2\mathrm{x}\mu+\mu^2}{2\sigma^2}}}{e^{-\frac{\cancel{\mathrm{x}^2}}{2\sigma^2}}} \;=\; e^{\frac{\mu(2\mathrm{x}-\mu)}{2\sigma^2}} \underset{H_0}{\overset{H_1}{\gtrless}} 1.$$

Since both sides are positive, taking log we obtain:

$$\frac{\mu(2\mathrm{x}-\mu)}{2\sigma^2} \underset{H_0}{\overset{H_1}{\gtrless}} 0.$$

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

$$\mathrm{x} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mu}{2}.$$

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 $t : \Omega \to \{H_0, H_1\}$ are the inverse images of H_0, and H_1, i.e.,

$$\begin{aligned} R_0 &:= \{\mathrm{x} \in \Omega \ : \ t(\mathrm{x}) = H_0\}, \\ R_1 &:= \{\mathrm{x} \in \Omega \ : \ t(\mathrm{x}) = H_1\}. \end{aligned}$$

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 $\mathrm{x} \underset{H_0}{\overset{H_1}{\gtrless}} {}^{\mu}\!/_{2}$ are $R_0 = (-\infty, {}^{\mu}\!/_{2})$ and $R_1 = ({}^{\mu}\!/,\infty)$.

Decision regions determine the probability of each outcome as follows:

$$\mathsf{p}_{\mathrm{gh}} \ := \ \int_{R_{\mathrm{g}}} \mathsf{p}_{\mathrm{h}}(\mathrm{x}) d\mathrm{x}, \qquad \text{for g, h} \in \{0, 1\}.$$

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

$$\begin{aligned} \mathsf{p}_{00} &= Q_{0,\sigma^2}(-{}^{\mu}\!/_{2}), \\ \mathsf{p}_{01} &= Q_{\mu,\sigma^2}(-{}^{\mu}\!/_{2}), \\ \mathsf{p}_{10} &= Q_{0,\sigma^2}({}^{\mu}\!/_{2}), \\ \mathsf{p}_{11} &= Q_{\mu,\sigma^2}({}^{\mu}\!/_{2}), \end{aligned}$$

		Truth	
		H_0	H_1
Decision	H_0	**(0|0)** True Negative No-alarm Accept	**(0|1)** False negative Miss Type 2 error
	H_1	**(1|0)** False positive False alarm Type 1 error	**(1|1)** True Positive Detect Reject

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.

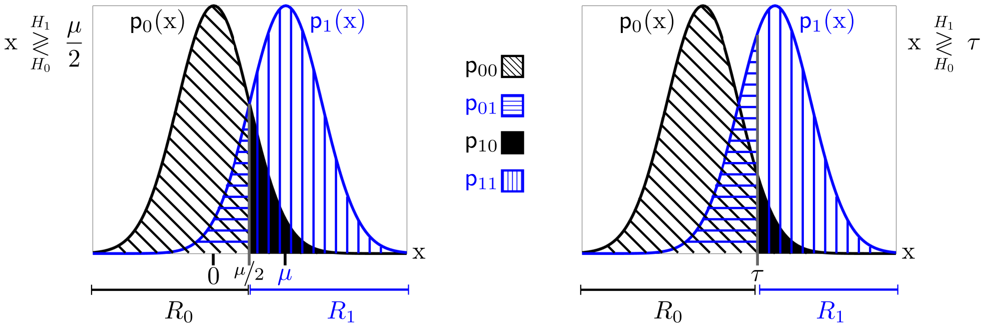

Figure 3.5: Outcomes probabilities p_{gh}. **Left**: test $\mathrm{x} \gtrless^{H_1}_{H_0} {}^{\mu}/_{2}$ from Example 3.7. **Right**: test $\mathrm{x} \gtrless^{H_1}_{H_0} \tau$ from Example 3.10; we can pick τ such that $\mathsf{p}_{10} < \alpha$.

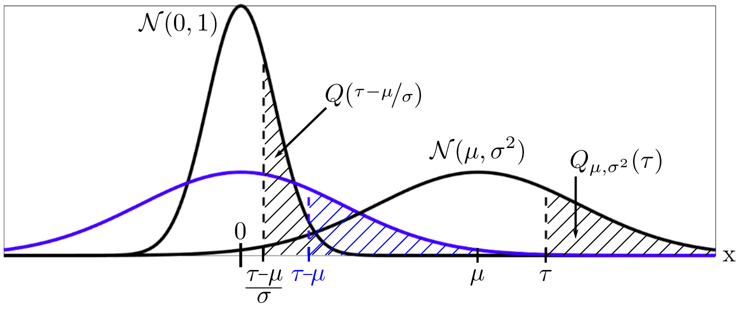

Figure 3.6: $Q_{\mu,\sigma^2}(\tau,\infty) = Q({}^{\tau-\mu}/_{\sigma},\infty)$, where Q is shorthand for $Q_{0,1}$.

where $Q_{\mu,\sigma^2}(\tau)$ is the tail probability of the $\mathcal{N}(\mu,\sigma^2)$ distribution, i.e.,

$$Q_{\mu,\sigma^2}(\tau) \;:=\; \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu}{\sigma}\right)^2} d\mathrm{x}.$$

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

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

where τ is a threshold selected to make p_{10} smaller than the desired probability of error α (see Figure 3.5). This new test has $\mathsf{p}_{10} = Q_{0,\sigma^2}(\tau)$. Furthermore, a simple change of variable shows that

$$Q_{\mu,\sigma^2}(\tau) \;=\; Q\left(\frac{\tau-\mu}{\sigma}\right), \tag{3.1}$$

where we use Q as shorthand for $Q_{0,1}$ (see Figure 3.6 to build some intuition), so $\mathsf{p}_{10} = Q\left({}^{\tau}/_{\sigma}\right)$. Finally, since Q is invertible, if we want $\mathsf{p}_{10} \leq \alpha$, 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

$$\frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \tau,$$

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

Proof. For any region $R \subset \Omega$, let $\mathsf{P}_\mathrm{h}(R)$ be the cumulative probability of $\mathsf{p}_\mathrm{h}(\mathrm{x})$ over R, i.e.,

$$\mathsf{P}_\mathrm{h}(R) \;=\; \int_R \mathsf{p}_\mathrm{h}(\mathrm{x}) d\mathrm{x}.$$

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

$$\begin{aligned} \mathsf{p}_{1\mathrm{h}} &= \mathsf{P}_\mathrm{h}(R_1 \cap R'_1) + \mathsf{P}_\mathrm{h}(R_1 \cap R'_0), \\ \mathsf{p}'_{1\mathrm{h}} &= \mathsf{P}_\mathrm{h}(R_1 \cap R'_1) + \mathsf{P}_\mathrm{h}(R_0 \cap R'_1). \end{aligned} \tag{3.2}$$

Now suppose $\mathsf{p}'_{10} \leq \alpha$. We need to show that $\mathsf{p}_{11} \geq \mathsf{p}'_{11}$. From (3.2), this is equivalent to showing that $\mathsf{P}_1(R_1 \cap R'_0) \;\geq\; \mathsf{P}_1(R_0 \cap R'_1)$, so write

$$\mathsf{P}_1(R_1 \cap R'_0) \;=\; \int_{R_1 \cap R'_0} \mathsf{p}_1(\mathrm{x}) d\mathrm{x} \;\geq\; \int_{R_1 \cap R'_0} \tau \mathsf{p}_0(\mathrm{x}) d\mathrm{x} \;=\; \tau \mathsf{P}_0(R_1 \cap R'_0), \tag{3.3}$$

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

$$(3.3) \;\geq\; \tau \mathsf{P}_0(R_0 \cap R'_1) \;=\; \int_{R_0 \cap R'_1} \tau \mathsf{p}_0(\mathrm{x}) d\mathrm{x} \;\geq\; \int_{R_0 \cap R'_1} \mathsf{p}_1(\mathrm{x}) d\mathrm{x} \;=\; \mathsf{P}_1(R_0 \cap R'_1),$$

where the last inequality follows because in R_0, $\mathsf{p}_1(\mathrm{x}) \leq \tau \mathsf{p}_0(\mathrm{x})$. □

Example 3.11. Consider

$$\begin{aligned} H_0 : \; x \;&\sim\; \mathcal{N}(0, \sigma_0^2), \\ H_1 : \; x \;&\sim\; \mathcal{N}(0, \sigma_1^2), \end{aligned}$$

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

$$\Lambda(\mathrm{x}) \;=\; \frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})} \;=\; \frac{\frac{1}{\cancel{\sqrt{2\pi}}\sigma_1} e^{-\frac{\mathrm{x}^2}{2\sigma_1^2}}}{\frac{1}{\cancel{\sqrt{2\pi}}\sigma_0} e^{-\frac{\mathrm{x}^2}{2\sigma_0^2}}} \;=\; \frac{\sigma_0}{\sigma_1} e^{\frac{1}{2}\left(\frac{\mathrm{x}^2}{\sigma_0^2} - \frac{\mathrm{x}^2}{\sigma_1^2}\right)} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

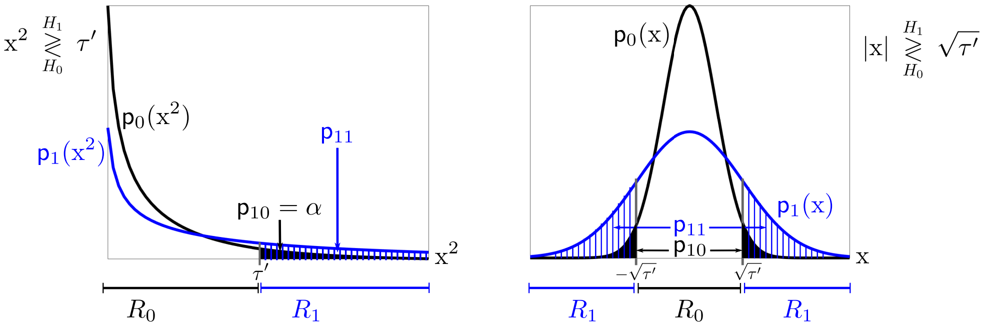

Figure 3.7: **Left:** Threshold τ' selected to achieve probability of error $\mathsf{p}_{10} = \alpha$ in test $\mathrm{x}^2 \gtrless_{H_0}^{H_1} \tau'$ of Example 3.11, where $H_0 : {x^2}/{\sigma_0^2} \sim \chi^2$. This is equivalent to the test $|\mathrm{x}| \gtrless_{H_0}^{H_1} \sqrt{\tau'}$ with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

Or equivalently,

$$\begin{aligned} e^{\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\mathrm{x}^2} &\underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma_1}{\sigma_0}\tau, \\ \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\mathrm{x}^2 &\underset{H_0}{\overset{H_1}{\gtrless}} \log\left(\frac{\sigma_1}{\sigma_0}\tau\right), \\ \mathrm{x}^2 &\underset{H_0}{\overset{H_1}{\gtrless}} \underbrace{\frac{2\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\log\left(\frac{\sigma_1}{\sigma_0}\tau\right)}_{\tau'}. \end{aligned}$$

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

$$\begin{aligned} H_0 : \ ({x}/{\sigma_0})^2 &\sim \chi^2, \\ H_1 : \ ({x}/{\sigma_1})^2 &\sim \chi^2. \end{aligned}$$

Then $\mathsf{p}_{1\mathrm{h}}$ (with $\mathrm{h} \in \{0,1\}$) is simply the probability that a σ_{h}^2-scaled χ^2 random variable is larger than τ' (see Figure 3.7), i.e.,

$$\mathsf{p}_{1\mathrm{h}} = Q_{\chi^2}\left({\tau'}/{\sigma_{\mathrm{h}}^2}\right),$$

where Q_{χ^2} is the tail probability of the χ^2 distribution. Since Q_{χ^2} is invertible, if we want $\mathsf{p}_{10} \leq \alpha$, we can pick $\tau' = \sigma_0^2 Q_{\chi^2}^{-1}(\alpha)$, and then $\mathsf{p}_{11} = Q_{\chi_1^2}({\sigma_0^2}/{\sigma_1^2} Q_{\chi_0^2}^{-1}(\alpha))$. 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:

$$\begin{aligned} H_0 : \ \boldsymbol{x} \in \mathbb{R}^{\mathrm{N}} &\sim \mathcal{N}(0, \sigma^2\mathbf{I}), \\ H_1 : \ \boldsymbol{x} \in \mathbb{R}^{\mathrm{N}} &\sim \mathcal{N}(\mu\mathbf{1}, \sigma^2\mathbf{I}), \quad \mu < 0, \end{aligned}$$

where $\mathbf{1}$ denotes the all ones vector of compatible size, in this case $\mathrm{N} \times 1$. The likelihood ratio is given by

$$\Lambda(\mathbf{x}) \;=\; \frac{\mathsf{p}_1(\mathbf{x})}{\mathsf{p}_0(\mathbf{x})} \;=\; \frac{\cancel{\frac{1}{(\sqrt{2\pi\sigma})^{\mathrm{N}}}} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu\mathbf{1})^{\mathsf{T}}(\mathbf{x}-\mu\mathbf{1})}}{\cancel{\frac{1}{(\sqrt{2\pi\sigma})^{\mathrm{N}}}} e^{-\frac{1}{2\sigma^2}\mathbf{x}^{\mathsf{T}}\mathbf{x}}} \;=\; \frac{e^{-\frac{1}{2\sigma^2}\left(\cancel{\mathbf{x}^{\mathsf{T}}\mathbf{x}}-2\mu\mathbf{x}^{\mathsf{T}}\mathbf{1}+\mu^2\mathbf{1}^{\mathsf{T}}\mathbf{1}\right)}}{e^{-\frac{1}{2\sigma^2}\cancel{\mathbf{x}^{\mathsf{T}}\mathbf{x}}}} \;=\; e^{\frac{\mu}{2\sigma^2}\left(2\mathbf{x}^{\mathsf{T}}\mathbf{1}-\mathrm{N}\mu\right)}.$$

Taking log we obtain the log-likelihood ratio test:

$$\frac{\mu}{2\sigma^2}\left(2\mathbf{x}^{\mathsf{T}}\mathbf{1}-\mathrm{N}\mu\right) \underset{H_0}{\overset{H_1}{\gtrless}} \log\tau,$$
$$\mathbf{x}^{\mathsf{T}}\mathbf{1} \underset{H_0}{\overset{H_1}{\lessgtr}} \underbrace{\frac{\sigma^2}{\mu}\log\tau + \frac{\mathrm{N}\mu}{2}}_{\tau'}.$$

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

$$\begin{aligned} H_0 :\ m \;&\sim\; \mathcal{N}(0, \mathrm{N}\sigma^2), \\ H_1 :\ m \;&\sim\; \mathcal{N}(\mathrm{N}\mu, \mathrm{N}\sigma^2), \quad \mu < 0, \end{aligned}$$

Then our log-likelihood ratio test becomes $\mathrm{m} \underset{H_0}{\overset{H_1}{\lessgtr}} \tau'$, and since $\tau' < 0$,

$$\begin{aligned} \mathsf{p}_{10} \;&=\; \Phi_{0,\mathrm{N}\sigma^2}(\tau'), \\ \mathsf{p}_{11} \;&=\; \Phi_{\mathrm{N}\mu,\mathrm{N}\sigma^2}(\tau'). \end{aligned}$$

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

$$\begin{aligned} \mathsf{p}_{10} \;&=\; \Phi\Big(\frac{\tau'}{\sqrt{\mathrm{N}}\sigma}\Big), \\ \mathsf{p}_{11} \;&=\; \Phi\Big(\frac{\tau' - \mathrm{N}\mu}{\sqrt{\mathrm{N}}\sigma}\Big). \end{aligned} \tag{3.4}$$

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

$$\mathsf{p}_{11} \;=\; \Phi\left(\frac{\sqrt{\mathrm{N}}\sigma\Phi^{-1}(\alpha) - \mathrm{N}\mu}{\sqrt{\mathrm{N}}\sigma}\right) \;=\; \Phi\left(\Phi^{-1}(\alpha) - \frac{\sqrt{\mathrm{N}}\mu}{\sigma}\right).$$

$\frac{\sqrt{\mathrm{N}\mu^2}}{\sigma}$ 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 σ.

3.6 Multiple Testing

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

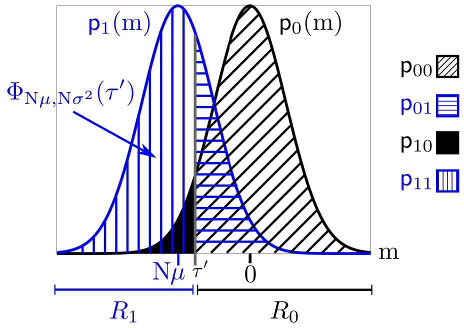

Figure 3.8: Outcomes probabilities p_{gh} of the test $m \underset{H_0}{\overset{H_1}{\lessgtr}} \tau'$. See Example 3.12.

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,

$$FWER \;=\; \mathsf{P}\left(\bigcup_{\mathrm{k}=1}^{\mathrm{K}} \{1_{\mathrm{k}} \mid 0_{\mathrm{k}}\}\right),$$

where $\{1_{\mathrm{k}} \mid 0_{\mathrm{k}}\}$ 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 $\mathsf{p}_{10} = {}^{\alpha}/_{\mathrm{K}}$ for each test achieves FWER $< \alpha$.

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

$$FWER \;=\; \mathsf{P}\left(\bigcup_{\mathrm{k}=1}^{\mathrm{K}} \{1_{\mathrm{k}} \mid 0_{\mathrm{k}}\}\right) \;\leq\; \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathsf{P}\left(\{1_{\mathrm{k}} \mid 0_{\mathrm{k}}\}\right) \;=\; \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathsf{p}_{10} \;=\; \cancel{\mathrm{K}}\frac{\alpha}{\cancel{\mathrm{K}}} \;=\; \alpha.$$

□

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 $\{\mathcal{N}(\mu,\sigma^2)\}_{\mu>0}$. More generally, a composite problem has the form:

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

where the notation $\mathsf{p}_\mathrm{h}(\mathrm{x}|\theta_\mathrm{h})$ means that θ_h is a parameter of the distribution p_h, and Θ_h is a set of all possible values of the parameter θ_h. In general, p_0 and p_1 may be entirely different distributions, and the sets Θ_0, Θ_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 μ 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

$$\Lambda(\mathrm{x}) \;=\; \frac{\mathsf{p}_1(\mathrm{x})}{\mathsf{p}_0(\mathrm{x})} \;=\; \frac{\cancel{\frac{1}{\sqrt{2\pi\sigma}}} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-\mu}{\sigma}\right)^2}}{\cancel{\frac{1}{\sqrt{2\pi\sigma}}} e^{-\frac{1}{2}\left(\frac{\mathrm{x}-0}{\sigma}\right)^2}} \;=\; \frac{e^{-\frac{\cancel{\mathrm{x}^2}-2\mathrm{x}\mu+\mu^2}{2\sigma^2}}}{e^{-\frac{\cancel{\mathrm{x}^2}}{2\sigma^2}}} \;=\; e^{\frac{\mu(2\mathrm{x}-\mu)}{2\sigma^2}} \underset{H_0}{\overset{H_1}{\gtrless}} 1.$$

Taking log and with some minor algebra we obtain:

$$\mathrm{x}\mu \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mu^2}{2}.$$

However, since we don't know the sign of μ, we cannot continue as in Example 3.7, as dividing by μ 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:

$$\mathrm{x} \underset{H_0}{\overset{H_1}{\gtrless}} \tau.$$

It could happen that we are lucky and $\mu > 0$. Then our test will be optimal (as shown by Neyman-Pearson's Lemma) with $\mathsf{p}_{10} = Q(\tau/\sigma)$ and $\mathsf{p}_{11} = Q(\tau-\mu/\sigma)$ (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 $\mathsf{p}_{11} = Q(\tau+|\mu|/\sigma)$; see Figure 3.9 to build some intuition. One good compromise is to use *Wald's test*:

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

which has

$$\begin{aligned} \mathsf{p}_{10} &= 2Q(\tau/\sigma), \\ \mathsf{p}_{11} &= Q\Big(\frac{\tau-\mu}{\sigma}\Big) + Q\Big(\frac{\tau+\mu}{\sigma}\Big). \end{aligned}$$

Wald's test is not optimal, but is sensible. It has higher probability of error p_{10} than if we are lucky and guess μ correctly, but also has higher accuracy p_{11} than if we are unlucky and guess μ incorrectly.

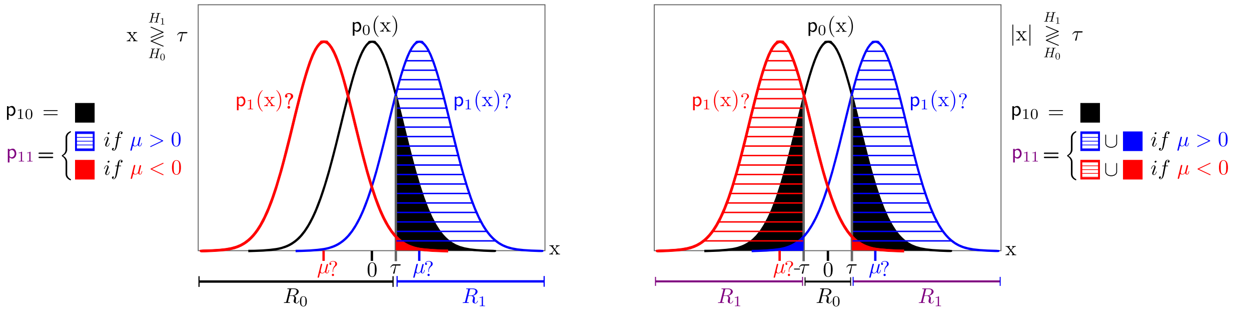
$x \underset{H_0}{\overset{H_1}{\gtrless}} \tau$
p_{10} = ■
$p_{11} = \begin{cases} ▤ \ \textit{if}\ \mu > 0 \\ 🟥 \ \textit{if}\ \mu < 0 \end{cases}$
$p_0(x)$
$p_1(x)$?
$p_1(x)$?
x
μ?
0
τ
μ?
R_0
R_1
$p_0(x)$
$p_1(x)$?
$p_1(x)$?
x
μ?
-τ
0
τ
μ?
R_1
R_0
R_1
$|x| \underset{H_0}{\overset{H_1}{\gtrless}} \tau$
p_{10} = ■
$p_{11} = \begin{cases} ▤ \cup 🟦 \ \textit{if}\ \mu > 0 \\ ▤ \cup 🟥 \ \textit{if}\ \mu < 0 \end{cases}$

Figure 3.9: Composite hypothesis test where $\mu \neq 0$ is unknown. **Left:** Probabilities p_{10} and p_{11} of the test $x \gtrless_{H_0}^{H_1} \tau$. If we are lucky and $\mu > 0$, this test test will be optimal, but if $\mu < 0$, our test will be terrible. **Right:** Wald's test $|x| \gtrless_{H_0}^{H_1} \tau$. Wald's test is not optimal, but is sensible. It has higher probability of error p_{10} than if we are lucky and guess μ correctly, but also has higher accuracy p_{11} than if we are unlucky and guess μ incorrectly. See Example 3.14 and Figure 3.10.

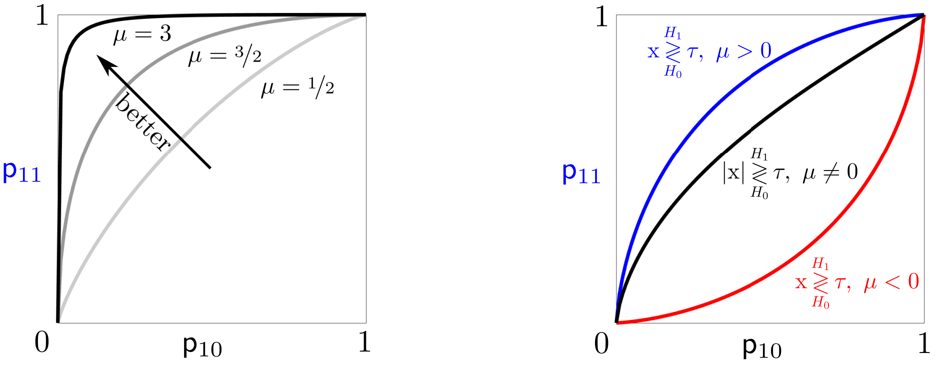

Figure 3.10: **Left:** ROC curves of the test $x \gtrless_{H_0}^{H_1} \tau$ for different values of μ. Consistent with our analysis from Example 3.12, we can see that p_{11} grows with μ. **Right:** ROC curves for the test $x \gtrless_{H_0}^{H_1} \tau$ when $\mu > 0$ (optimal), when $\mu < 0$ (terrible), 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 device 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* μ 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.