Hypothesis tests

The problem: a distribution parameter $\theta$ is unknown, but lies in a parameter space $\Omega$.

$\Omega_0, \Omega_1 \subseteq \Omega$ are disjoint, and

$$\Omega = \Omega_0 \cup \Omega_1.$$
The null hypothesis $H_0$ is that $\theta \in \Omega_1$:

$$H_0 : \theta \in \Omega_0.$$  

The alternative hypothesis $H_1$ is that $\theta \in \Omega_1$:

$$H_1 : \theta \in \Omega_1.$$  

One and only one of these is true. Which is it?

The process of deciding between two such alternatives (by collecting data) is called hypothesis testing.

A procedure for deciding in such a situation is called a hypothesis test.
Monday, March 4: Hypothesis tests

In some tests, $H_0$ and $H_1$ are interchangeable, but often they are not.

**Definition** If a hypothesis being tested completely specifies the distribution that it is about, it is called a *simple* hypothesis. If not, it is called a *compound* hypothesis.

More specifically, if

$$H_0 : \theta \in \Omega_0,$$

then this hypothesis is simple when $\Omega_0$ contains only a single value, and compound otherwise.

Typical compound hypotheses are of the form

- $H_1 : \theta \neq \theta_0$ \hspace{1cm} two-sided
- $H_1 : \theta > \theta_0$ \hspace{1cm} one-sided to the right
- $H_1 : \theta < \theta_0$ \hspace{1cm} one-sided to the left.
Monday, March 4: Hypothesis tests

From a decision theory point of view, the purpose of a hypothesis test is to decide what view of the null hypothesis (either true or false) to assume for future actions.

**Definition** Proceeding in the future on the assumption that $H_0$ is false is called **rejecting** the null hypothesis. Proceeding on the assumption that it is true is called **accepting** the null hypothesis.

**Definition** Suppose we observe a random sample $X_1, \ldots, X_n$ in order to test a parameter about their distribution. The **critical region** is the subset of the space of all possible values $S$ (for these variables jointly) for which the test rejects $H_0$:

$$\text{Critical region} = \{(x_1, \ldots, x_n) \in S \mid \text{reject } H_0\}.$$
Monday, March 4: Hypothesis tests

A typical type of hypothesis test uses a test statistic

$$T = r(X_1, \ldots, X_n)$$

to define its critical region, which is often given by

$$\text{Critical region} = \{(x_1, \ldots, x_n) \in S \mid T \geq c\},$$

or

$$\text{Critical region} = \{(x_1, \ldots, x_n) \in S \mid T \leq c\}.$$
Monday, March 4: Hypothesis tests

**Definition** Suppose that we are conducting a hypothesis test about parameter(s) $\theta \in \Omega$. The power function $\text{Power} : \Omega \to \mathbb{R}$ for the test is defined by

\[
\text{Power}(\theta) = P(\text{reject } H_0 | \theta).
\]

For a simple null hypothesis $H_0 : \theta = \theta_0$, the power of the test is $\text{Power}(\theta_0)$.

The ideal power function would be 0 whenever $\theta \in \Omega_0$ and 1 whenever $\theta \in \Omega_1$, but this does not generally arise in practice.
Monday, March 4: Hypothesis tests

Recall:

**Definition** A Type I error, or *false positive*, means rejecting $H_0$ when $H_0$ is true.

**Definition** A Type II error, or *false negative*, means accepting $H_0$ when $H_0$ is false.

If $H_0$ is true, so that $\theta \in \Omega_0$, then $\text{Power}(\theta)$ is the probability of a Type I error.

If $H_0$ is false, so that $\theta \in \Omega_1$, then $1 - \text{Power}(\theta)$ is the probability of a Type II error.
In general, there is a tradeoff between Type I and Type II errors.

Consider the test where we always reject $H_0$, and the test where we always accept $H_0$.

**Example** Let $X_1, \ldots, X_n$ be a random sample of a $\text{Unif}(0, \theta)$ distribution, where $\theta > 0$ is unknown. Let

$$H_0 : 3 \leq \theta \leq 4 \quad \text{and} \quad H_1 : \theta < 3 \text{ or } \theta > 4.$$ 

Suppose that your hypothesis test accepts $H_0$ when the maximum observed value is between 2.9 and 4 and rejects otherwise. Compute the power function of this test.
Monday, March 4: Hypothesis tests

We have

\[ \text{Power}(\theta) = P(\max X_n \leq 2.9 \mid \theta) + P(\max X_n \geq 4 \mid \theta). \]

If \( \theta < 2.9 \), then \( \text{Power}(\theta) = 1 + 0 \).

If \( 2.9 < \theta \leq 4 \), then

\[ \text{Power}(\theta) = \left(\frac{2.9}{\theta}\right)^n + 0. \]

If \( \theta > 4 \), then

\[ \text{Power}(\theta) = \left(\frac{2.9}{\theta}\right)^n + \left(1 - \left(\frac{4}{\theta}\right)^n\right). \]
Suppose we are testing

\[ H_0 : \theta \in \Omega_0 \quad \text{and} \quad H_1 : \theta \in \Omega_1, \]

and \( T \) is a test statistic. Also suppose that our test rejects \( H_0 \) when \( T \geq c \). If we want to limit the probability of a Type I error to \( \alpha \), we want

\[
\sup_{\theta \in \Omega_0} P(T \geq c \mid \theta) \leq \alpha.
\]

By the way that this test is defined, the power function, and so the expression above, is a nonincreasing function of \( c \).

If we want the power function to be as large as possible for \( \theta \in \Omega_1 \), then we should choose \( c \) as small as possible while still satisfying the above equation.
Tuesday, March 5: Hypothesis tests

**Definition** The maximum allowable probability $\alpha$ of a Type I error given that $H_0$ is true is called the significance level of the test.

The significance level is chosen before conducting the test.

In the current context, for continuous $T$ approximating such a $c$ is not difficult.
**Example** Suppose that $X_1, \ldots, X_n$ are a random sample of a $N(\mu, \sigma)$ distribution, where $\sigma$ is known, and that

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0.$$ 

Also suppose that our test rejects $H_0$ when $|\bar{X} - \mu_0| > c$. How should we choose $c$ to ensure that the probability of a Type I error is no more than $\alpha$, and that the probability of a Type II error is minimized given that constraint?
Wednesday, March 6: Hypothesis tests

Suppose that $X_1, \ldots, X_n$ are a random sample with Bernoulli distribution with success probability $p$. Also, suppose that

$$H_0 : p \leq p_0 \quad \text{and} \quad H_1 : p > p_0.$$ 

Let $Y = \sum_{i=1}^n X_i$. The larger that $p$ is, the larger we expect $Y$ to be, so suppose that our hypothesis test rejects the null hypothesis when $Y \geq c$.

We want the probability of a Type I error to be at most $\alpha$, and we’d like to minimize Type II errors. So we should choose $c$ to be the smallest number for which

$$P(Y \geq c \mid p) \leq \alpha.$$
Wednesday, March 6: Hypothesis tests

This can easily be done using a computer for the calculations.

For example, if $n = 10$, $p = 0.3$, and $\alpha = 0.1$, then

$$P(Y \geq 6) = 0.047349 \quad \text{and} \quad P(Y \geq 5) = 0.1502683.$$
The problem with hypothesis tests viewed as decisions: they throw away a lot of information.

For example, finding the value of the test statistic, we are merely reporting “yes” or “no”, which is just a single bit of information (0 or 1).

An approach that retains much more information about the test is to report the set of all values of $\alpha$ for which the level $\alpha$ test would lead to reject $H_0$ (given our data).
Wednesday, March 6: Hypothesis tests

**Definition** The p-value of a hypothesis test is the largest $\alpha$ for which the test at level $\alpha$ would reject $H_0$ with the given observed data.

For the simple case in which the rejection region is defined by $T \geq c$, the p-value is

$$\sup_{\theta \in \Omega_0} P(T \geq c | \theta)$$

Let’s consider an example (on the board)…
Wednesday, March 6: Hypothesis tests

The traditional definition of a $p$-value is the probability under the null hypothesis that the test statistic would take on a value at least as extreme as what we observed.

For the usual types of tests, this definition agrees with the one given above.
Definition Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega_1$. A test $t \in \mathcal{C}$ is uniformly most powerful in $\mathcal{C}$ if

$$\text{Power}_t(\theta) \geq \text{Power}_s(\theta)$$

for all $\theta \in \Omega_1$ and for all tests $s \in \mathcal{C}$.

We will often take $\mathcal{C}$ to be the class of all significance level $\alpha$ tests of these hypotheses.

A class may or may not have a uniformly most powerful test, but if it does, that test might well be considered the best one to use from that class.
Theorem (Neyman-Pearson Lemma) Suppose that $X_1, \ldots, X_n$ are a random sample whose likelihood function $L(\theta; x_1, \ldots, x_n)$, and let

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta = \theta_1,$$

meaning both hypotheses are simple. Also, suppose the test $t$ is to reject $H_0$ precisely when

$$\frac{L(\theta_0; x_1, \ldots, x_n)}{L(\theta_1; x_1, \ldots, x_n)} \leq c$$

for some $c \geq 0$. Additionally, let $\alpha$ be the probability of a Type I error given $H_0$.

Then $t$ is a uniformly most powerful test among the class of level $\alpha$ tests of these hypotheses.
Example Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, 1)$ distribution, and let

\[ H_0 : \theta = 0 \quad \text{and} \quad H_1 : \theta = 1. \]

Find a uniformly most powerful test among level $\alpha$ tests.
The likelihood function is

\[ L(\theta) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \right). \]

This implies that

\[ \frac{L(1)}{L(0)} = \exp(n(\bar{x} - \frac{1}{2})). \]

By Neyman-Pearson, we should reject when \( \bar{x} \) is sufficiently large. How large? To get \( \alpha = 0.05 \) right:

\[ k = \frac{z_{0.95}^*}{\sqrt{n}} = \frac{1.6448536}{\sqrt{n}} \]