Lecture 7.

**Bayesian Estimation.** Here we assume that while we do not know the exact value of the unknown parameter $\theta$ we suppose that it is chosen randomly from a set of possible values of $\theta$ and we have reason to believe that its distribution is given by some probability distribution $p_0(\theta)$ on the set of possible values of $\theta$. For simplicity we assume that the set of possible values of $\theta$ is the real line or a subset of it and $p(\theta)$ represents the density of the distribution of $\theta$. We have an observation (or a set of observations) $x$ and their distribution is given by the density $f(\theta, x)$. The joint distribution of $\theta$ and $x$ is given by $p_0(\theta)f(\theta, x)$ and the marginal of $x$ is

$$
\bar{f}(x) = \int f(\theta, x)p_0(\theta)d\theta
$$

The conditional $p_1(\theta|x)$, the posterior distribution of $\theta$ given $x$ is

$$
p_1(\theta|x) = \frac{f(\theta, x)p_0(\theta)}{\bar{f}(x)}
$$

As we gather more data, we can update by taking $p_1$ as the new $p_0$.

**Example.** Let $0 \leq \theta \leq 1$ be the probability of head in a single toss. Initially we may take $p_0(\theta) \equiv 1$. Suppose we have $n_1$ tosses resulting in $r_1$ heads.

$$
p_0(\theta)f(\theta, r_1) = \binom{n_1}{r_1}\theta^{r_1}(1-\theta)^{n_1-r_1}
$$

$$
\bar{f}(r_1) = \int_0^1 \binom{n_1}{r_1}\theta^{r_1}(1-\theta)^{n_1-r_1}d\theta
$$

$$
= \binom{n_1}{r_1}\beta(r_1 + 1, n_1 - r_1 + 1)
$$

and

$$
p_1(\theta|n_1, r_1) = \frac{1}{\beta(r_1 + 1, n_1 - r_1 + 1)}\theta^{r_1}(1-\theta)^{n_1-r_1}
$$

If we now have an additional $n_2$ tosses that resulted in $r_2$ heads, doing the Bayesian procedure again we get for $p_2(\theta|n_1, r_1, n_2, r_2)$

$$
\frac{1}{\beta(r_1 + r_2 + 1, n_1 + n_2 - r_1 - r_2 + 1)}\theta^{r_1+r_2}(1-\theta)^{n_1+n_2-r_1-r_2}
$$
Example. For estimation of the mean of a normal population with an unknown mean $\theta$ and known variance 1, it is natural to start with a prior distribution for $\theta$ which is Normal with some mean $a$ and some variance $\sigma^2$. Says some thing about our best guess $a$ for the mean and the level of our uncertainty as measured by $\sigma^2$. If we have $n$ observations with a mean of $y = \bar{x}$,

$$p_0(\theta) f(\theta, y) = \frac{1}{\sigma \sqrt{2\pi}} \sqrt{n} \exp\left[-\frac{(\theta - a)^2}{2\sigma^2} - \frac{n(y - \theta)^2}{2}\right]$$

$$\bar{f}(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{1}{n} + \sigma^2}} \exp\left[-\frac{(y - a)^2}{2\left(\frac{1}{n} + \sigma^2\right)}\right]$$

$$p_1(\theta|y) = \frac{1}{\sqrt{2\pi}} \sqrt{n + \frac{1}{\sigma^2}} \exp\left[-\frac{(n + \frac{1}{\sigma^2})(\theta) - \left(\frac{a + ny}{\sigma^2 + n}\right)^2}{2}\right]$$

It is again normal with mean $\frac{a + n\sigma^2 y}{1 + n\sigma^2}$ and variance $\frac{\sigma^2}{1 + n\sigma^2}$.

Testing of Hypotheses. Suppose we have two possible densities $f_0(x)$ and $f_1(x)$ and an observation $X$ from one of the two populations and we have to decide. What can we do?

If $X \in E_0$ we say $f_0$ and $X \in E_1$ we say $f_1$. Seems reasonable. What should $E_0$ and $E_1$ be? $E_0 = E_1^c$. Basically we just need to choose $A = E_1$. Would like $\int_A f_0(x)dx$ should be small and $\int_A f_1(x)dx$ to be big. One way is to fix $\int_A f_0(dx) = \alpha$ and maximize $\int_A f_1(dx)$.

$$A = A_\lambda = \{x : \frac{f_1(x)}{f_0(x)} \geq \lambda\}$$

will do it. Fix $\lambda$ so that $\int_{A_\lambda} f_0(x)dx = \alpha$

$f_0$ is called the null hypothesis. $f_1$ is called the alternate hypothesis. $\alpha$ is called the size of the test, or the size of type I error. $\beta = \int_A f_1(x)dx$ is called the power of the test. $1 - \beta$ is called the type II error. A hypothesis that fully specifies the distribution is called a simple hypothesis. We can have a simple null hypothesis or a composite null hypothesis and similarly a null or composite alternate.
Example 1. \( \{X_i\} \) are \( n \) i.i.d observations from \( N(\mu, 1) \). \( H_0 = \{\mu = 0\} \).
\( H_1 = \{\mu = 1\} \).

Example 2. \( \{X_i\} \) are \( n \) i.i.d observations from \( N(\mu, 1) \). \( H_0 = \{\mu = 0\} \).
\( H_1 = \{\mu = 2\} \).

Example 3. \( \{X_i\} \) are \( n \) i.i.d observations from \( N(\mu, 1) \). \( H_0 = \{\mu = 0\} \).
\( H_1 = \{\mu = -1\} \).

Example 4. \( \{X_i\} \) are \( n \) i.i.d observations from \( N(\mu, 1) \). \( H_0 = \{\mu = 0\} \).
\( H_1 = \{\mu > 0\} \).

Example 5. \( \{X_i\} \) are \( n \) i.i.d observations from \( N(\mu, 1) \). \( H_0 = \{\mu = 0\} \).
\( H_1 = \{\mu < 0\} \) The set \( A \) where the null hypothesis is rejected is called the critical region.

\[
A_\lambda = \left\{ x_1, \ldots, x_n : \frac{f_1(x_1) \cdots f_1(x_n)}{f_0(x_1) \cdots f_1(x_n)} > \lambda \right\}
\]

\[
- \sum (x_i - \mu)^2 \geq - \sum x_i^2 + \lambda
\]

\[
\mu \sum x_i \geq \lambda
\]

\[
\bar{x} > c \quad \text{if} \quad \mu > 0
\]

and

\[
\bar{x} < c \quad \text{if} \quad \mu < 0
\]

Determine \( c \) so that \( P_0[A_c] = \alpha \). Test is the same for 1, 2, 4. Uniformly most powerful tests. 2 and 5 have the same test. What if \( H_1 = \{\mu \neq 0\} \)? There is no UMP test. Need to take \( A_c = \{|\bar{x}| > c\} \).

Some times the null hypothesis can be composite. For example we may want to test that the mean of a normal population is 0, without making any assumptions about its variance. \( H_0 = \{\mu = 0, \theta > 0\} \), \( H_1 = \{\mu > 0, \theta > 0\} \).

The critical region should be of the form

\[
\exp[-\frac{1}{2\theta} \sum x_i^2] < \exp[-\frac{1}{2\theta} \sum (x_i - \mu)^2]
\]

3
Same as $\bar{x} > c$. But the distribution $\bar{x}$ depends on $\theta$ and one can not determine the value of $c$ that corresponds to a given size $\alpha$. One uses instead the $t$ test $\frac{\bar{x}}{s} > c$ for the critical region. Or equivalently

$$"t_{n-1}" = \frac{\bar{x}}{s} \sqrt{n - 1} > c$$

The distribution of $"t_{n-1}"$ being independent of $\theta$ one can determine $c$ from $\alpha$. For two sided alternatives one can do two sided tests.

Testing for variances in normal populations. $H_0 = \{\mu = 0, \theta = 1\}$ and $H_1 = \{\mu = 0, \theta > 1\}$.

$$\log p(0, x_1, \ldots, x_n) - \log p(\theta, x_1, \ldots, x_n) = \frac{n}{2} \log \theta + \frac{1}{2\theta - 1} \sum_{i=1}^{n} x_i^2$$

Reject if $\sum_{i} x_i^2 = \chi^2_n > c$. The alternative $\theta < 1$ and the two sided alternatives are handled in a similar way.

**Likelihood ratio criterion.** In general testing composite hypothesis is not easy. However far large samples there is a reasonable theory. Suppose there is a model where the population is specified by a parameter $\theta \in \Theta \subset R^d$. The null hypothesis states that $\theta \in \Theta_1 \subset \Theta$. For simplicity let us take $\theta = (\theta_1, \ldots, \theta_d)$ and $\Theta_1 = \{\theta : \theta_1 = \theta_2 = \theta_k = 0\}$. $k < d$. We have the likelihood ratio

$$\lambda = \frac{\sup_{\theta \in \Theta_1} L(\theta, x_1, \ldots, x_n)}{\sup_{\theta \in \Theta} L(\theta, x_1, \ldots, x_n)}$$

It is clear that we should reject the null hypothesis if $\lambda$ is small or $-2 \log \lambda > c$. If the null hypothesis is true then $-2 \log \lambda$ is a $\chi^2_k$ for large $n$ and that helps us to determine $c$.

**Example.** $\{x_i\}$ are $N(\mu, \theta)$. $H_0 = \{\mu = 0\}$. $d = 2$, $k = 1$.

$$\sup_{\theta} \log p(0, \theta, x) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log[\frac{1}{n} \sum x_i^2] - \frac{n}{2}$$

$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log[s^2 + \bar{x}^2] - \frac{n}{2}$$

$$\sup_{\mu, \theta} \log p(\mu, \theta, x) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log[s^2] - \frac{n}{2}$$
\[-2 \log \lambda = n \log(1 + (\bar{x}/s)^2) \simeq \chi^2_1\]

**Example.** \(\{x_i\}\) are \(N(\mu, \theta)\). \(H_0 = \{\mu = 0, \theta = 1\}\). \(d = 2, k = 2\)

\[
\log p(0, 1, \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum x_i^2
\]

\[
= -\frac{n}{2} \log(2\pi) - \frac{n}{2}[s^2 + \bar{x}^2]
\]

\[
\sup_{\mu, \theta} \log p(\mu, \theta, \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log[s^2] - \frac{n}{2}
\]

\[-2 \log \lambda = n[s^2 - \log s^2 - 1] + nx^2\]

If \(s^2 - 1 = \xi\), then, for large \(n\), \(\sqrt{n\xi} \simeq N(0, 2)\)

\[-2 \log \lambda = n(-\log(1 + \xi) + \xi) + nx^2 \simeq \left[\sqrt{\frac{n\xi}{2}}\right]^2 + \sqrt{nx^2} \simeq \chi^2_2\]

**Approximations.** MLE estimate is often the solution of (not always)

\[
\sum_{i=1}^{n} \frac{\partial \log f(\theta_1, \ldots, \theta_d, x_i)}{\partial \theta_r} = 0; \quad r = 1, 2 \ldots, d.
\]

Therefore for \(1 \leq r \leq d\), if \(\tilde{\theta}_r\) are other estimates

\[
0 = \sum_{i=1}^{n} \frac{\partial \log f(\hat{\theta}_1, \ldots, \hat{\theta}_d, x_i)}{\partial \theta_r}
\]

\[
= \sum_{i=1}^{n} \frac{\partial \log f(\tilde{\theta}_1, \ldots, \tilde{\theta}_d, x_i)}{\partial \theta_r}
\]

\[
+ \sum_{i=1}^{n} \sum_{s=1}^{d} (\hat{\theta}_s - \tilde{\theta}_s) \frac{\partial^2 \log f(\theta_1, \ldots, \theta_d, x_i)}{\partial \theta_r \partial \theta_s}
\]

Provides a method for approximation. If \(\tilde{\theta}\) is good estimate then the MLE \(\hat{\theta}\) is given by

\[
\hat{\theta} = \tilde{\theta} + [I(\tilde{\theta})]^{-1}\left[\frac{1}{n} \sum_{i=1}^{n} (\nabla \theta \log f)(\tilde{\theta}_1, \ldots, \tilde{\theta}_d, x_i)\right]
\]
Likelihood Ratio Criterion.

\[-2 \log \lambda =
\]
\[2[\log L(\hat{\theta}_1, \ldots, \hat{\theta}_d, x_1, \ldots, x_n) - \log L(\bar{\theta}_1, \ldots, \bar{\theta}_k, \theta_{k+1}, \ldots, \theta_d, x_1, \ldots, x_n)]\]
\[\geq 0\]

Here \((\theta_1, \ldots, \theta_k)\) are the true values of the parameters. \(\{\hat{\theta}_j\}\) and \(\{\bar{\theta}_j\}\) are the two sets of MLE’s. The second term is constrained optimization, whereas the first term is unconditioned and is therefore larger. Its distribution under the null hypothesis that \(\theta_{k+1}, \ldots, \theta_d\) are indeed the correct values for these parameter will be a \(\chi^2_{d-k}\) and is used to test the hypothesis.

Goodness of fit. Often we have data grouped into categories and is presented as frequencies \(\{f_i\}\) in \(k\) categories. \(N = \sum_{i=1}^{k} f_i\) being the total number of observations. We have a model that predicts the probabilities that an observation belongs to these categories are \(\{p_i\}\). The expectation is then that \(f_i \simeq Np_i\). The statistic used to test the hypothesis is

\[\sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i} = \sum_{i=1}^{k} \frac{f_i^2}{Np_i} - N\]

Its distribution is a \(\chi^2\) with \(k - 1\) degrees of freedom. If the model had a certain number \(r\) of parameters \(\{\theta_j\}\) and we used maximum likelihood method to estimate them and used \(p_i(\{\theta_j\})\) to compare then

\[\chi^2 = \sum_{i=1}^{k} \frac{(f_i - Np_i)^2}{Np_i}\]

will be a \(\chi^2_{k-1-r}\) degrees of freedom. We lose one degree of freedom for each parameter we estimate. Note that \(\{p_i\}\) although there are \(k\) of them are only \(k - 1\) parameters. All this depends on the following. We have a quadratic form \(Q = < \xi, B\xi >\) in Gaussian random variables \(\{\xi_j\}\) with mean 0 and covariance \(C_{i,j} = E[\xi_i \xi_j]\). When is the distribution of \(Q\) a \(\chi^2\) and what is its degrees of freedom? We need a calculation.

\[E[\exp\left(-\frac{\lambda}{2}Q\right)] = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{\sqrt{|C|}} \int \exp\left[-\frac{1}{2} \langle \xi, (Q + C^{-1})\xi \rangle\right] d\xi\]

\[= ||C||^{-\frac{1}{2}} CQ^{-\frac{1}{2}} + I^{-\frac{1}{2}}\]

\[= |\lambda CQ + I|^{-\frac{1}{2}}\]

\[= |\lambda Q + 1|^{-\frac{1}{2}}\]
This will be the same as $E[\exp[-\lambda \chi^2_{q}]]$ provided $Q^{\frac{1}{2}}CQ^{\frac{1}{2}}$ is projection of rank $q$. For the multinomial goodness of fit

$$Q = \begin{pmatrix}
\frac{1}{p_1} & 0 & \cdots & 0 \\
0 & \frac{1}{p_2} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \frac{1}{p_k}
\end{pmatrix}$$

$$C = \begin{pmatrix}
p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\
-p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\
\cdots & \cdots & \cdots & \cdots \\
-p_1p_k & -p_2p_k & \cdots & p_k(1-p_k)
\end{pmatrix}$$

$$Q^{\frac{1}{2}}CQ^{\frac{1}{2}} = I - P$$

where

$$P = \begin{pmatrix}
p_1 & \sqrt{p_1p_2} & \cdots & \sqrt{p_1p_k} \\
\sqrt{p_1p_2} & p_2 & \cdots & \sqrt{p_2p_k} \\
\cdots & \cdots & \cdots & \cdots \\
\sqrt{p_1p_k} & \sqrt{p_2p_k} & \cdots & p_k
\end{pmatrix}$$

$P$ is a projection of rank 1. We have a $\chi^2_{k-1}$.

**Example.** Let $f_1, \ldots, f_k$ be multinomial cell frequencies form $N = \sum f_i$ observations. The individual probabilities $p_i$ are modeled by a binomial. Is this valid?. MLE is given by maximizing

$$\frac{N!}{f_1! \cdots f_k!} p_1^{f_1} \cdots p_k^{f_k}$$

where $p_i = (\binom{k}{i}) \theta^i (1-\theta)^{k-i}$. The equation for MLE is

$$\sum_i f_i \left[ \frac{i}{\hat{\theta}} - \frac{k-i}{1-\hat{\theta}} \right] = 0$$

$$\hat{\theta} = \frac{1}{kN} \sum_{i=1}^k if_i$$

$$\hat{p}_i = \binom{k}{i} \hat{\theta}^i (1-\hat{\theta})^{k-1}$$

$$\chi^2_{k-2} = \sum_i \frac{(f_i - N\hat{p}_i)^2}{N\hat{p}_i}$$
Is the data consistent with a model of Binomial with \( \theta = \frac{1}{2} \). then we use 
\[ p_i(\theta) = \binom{k}{i} 2^{-k}. \]
The degrees of freedom is \( k - 1 \).

**Checking for Independence.** We have two classifications labeled by \( X \) and \( Y \) that can take values from \( 1, 2, \ldots, k \) and \( 1, 2, \ldots, \ell \) We have frequencies \( f_{i,j} \) the number with labels \( X = i \) and \( Y = j \). Is there dependence? The model is that the probabilities are given by

\[
P[X = i, Y = j] = \pi_{i,j} = p_i q_j
\]

The MLE are easily calculated as

\[
\hat{p}_i = \frac{1}{N} \sum_{j=1}^{\ell} f_{i,j} \quad \hat{q}_j = \frac{1}{N} \sum_{i=1}^{k} f_{i,j}
\]

\[
\chi^2_d = \sum_{i,j} \frac{(f_{i,j} - N\hat{p}_i \hat{p}_j)^2}{N\hat{p}_i \hat{p}_j}
\]

The degrees of freedom is

\[
d = (k\ell - 1) - (k - 1) - (\ell - 1) = (k - 1)(\ell - 1)
\]