Point Estimation

- Suppose we observe a random variable $X$ that belongs to an unknown member of a family of distributions $f_{\theta}(X)$. That is, each $\theta \in \Theta$ determines a single distribution, but we do not know which value of $\theta$ produced the $X$ that we actually observed. We call this unknown value the *true* value of $\theta$.

- A basic problem in statistics is to estimate $\theta$ based on $X$. That is, we wish to construct a function $\hat{\theta}(X)$ that can be used to estimate $\theta$, and we wish to understand the performance of $\hat{\theta}(X)$ as an estimator of $\theta$.

- Estimation problems roughly fall into two categories.

  1. *Parameter estimation problems* begin with the specification of a family of density or mass functions $f_{\theta}(X)$, and the problem is to estimate the parameter $\theta$ based on $X$. For instance we may take the geometric mass function $f_{\theta}(X) = (1-\theta)^{X-1}\theta$ or the binomial mass function $f_{\theta}(X) = \binom{n}{X}\theta^{X}(1-\theta)^{n-X}$.

  2. The second class of estimation problems involves estimating properties of a distribution that do not depend on the specification of a family of distribution functions. For example, we may want to estimate the mean or variance of a distribution.

- More generally, we may observe iid $X_1, \ldots, X_n$ according to $f_{\theta}(X)$, in which case we would estimate $\theta$ using all of the $X_i$, so our estimator would have the form $\hat{\theta}(X_1, \ldots, X_n)$.

- As an example, suppose we want to estimate $EX$ based on $X_1, \ldots, X_n$. We could select any of the following estimators
\[ \hat{\theta}_1(X_1, \ldots, X_n) = X_1 \]
\[ \hat{\theta}_2(X_1, \ldots, X_n) = X_2 \]
\[ \hat{\theta}_3(X_1, \ldots, X_n) = (X_1 + \cdots + X_n)/n. \]

- The performance of an estimator may be evaluated by many criteria. Three of the most common are bias, variance, and mean squared error (MSE). We desire each of these three quantities to be close to zero.

1. The bias is defined by bias(\(\theta\)) = \(E\hat{\theta}(X_1, \ldots, X_n) - \theta\). That is, observing that \(\hat{\theta}(X_1, \ldots, X_n)\) is a random variable, we take its expectation and determine the difference relative to the true value of \(\theta\). If this difference is zero we say that the estimator is unbiased. Since \(EX_1 = EX_2 = EX\bar{X}_n\), each of \(\hat{\theta}_1, \hat{\theta}_2,\) and \(\hat{\theta}_3\) is unbiased.

2. The variance is simply \(\text{Var}(\theta) = \text{Var}(\hat{\theta})\).

3. The mean squared error is \(\text{MSE}(\theta) = E(\hat{\theta} - \theta)^2\).

- A key fact is \(\text{Var}(\theta) + \text{bias}^2(\theta) = \text{MSE}(\theta)\). This follows from

\[
\text{MSE}(\theta) = E(\hat{\theta} - \theta)^2 \\
= E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2 \\
= E(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2 + 2E(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta) \\
= E(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2 + 2(E\hat{\theta} - \theta)E(\hat{\theta} - E\hat{\theta}) \\
= E(\hat{\theta} - E\hat{\theta})^2 + (E\hat{\theta} - \theta)^2.
\]

- For a given \(\theta\), there is a least MSE that can be achieved by any estimator. Denote this value \(\text{MSE}_*\). If \(\hat{\theta}\) is an estimator with
MSE equal to \( \text{MSE}_* \), the bias of \( \hat{\theta} \) can not be improved without making the variance worse, and the variance can not be improved without making the bias worse. This is called the bias/variance trade off.

- When estimating \( \theta = EX \) using \( \hat{\theta} = (X_1 + \cdots + X_n)/n \), \( \hat{\theta} \) is unbiased (since \( E\hat{\theta} = \theta \)), and the variance and MSE are both equal to \( \text{Var}(X_1)/n \).

- The following functions, known as the sample mean and variance, are standard for estimating the mean (\( \mu \)) and the variance (\( \sigma^2 \)) from an iid sequence \( X_1, \ldots, X_n \):

\[
\hat{\mu} = (X_1 + \cdots X_n)/n \\
\hat{\sigma}^2 = \frac{\sum_i (X_i - \hat{\mu})^2}{n - 1}.
\]

- Both \( \hat{\mu} \) and \( \hat{\sigma}^2 \) are unbiased. The variance of \( \hat{\mu} \) is \( \sigma^2/n \), where \( \sigma^2 = \text{Var}(X_1) \), and this is the smallest variance that can be achieved for any unbiased estimate of \( \mu \). The variance of \( \hat{\sigma}^2 \) follows a complicated formula that is rarely used.

- There are a number of generic procedures that can be used to construct parameter estimates in a variety of problems. Three that we will discuss are (i) the method of moments, (ii) maximum likelihood, and (iii) posterior mean.

- Suppose we observe \( X_1, \ldots, X_n \) iid from a density of mass function \( f_\theta(X) \). It is usually simple to work out the mean and variance of \( X \) as functions of \( \theta \): \( \mu(\theta) \), \( \sigma^2(\theta) \). For instance in the geometric distribution, \( \mu(p) = 1/p \) and \( \sigma^2(p) = (1-p)/p^2 \). For the Poisson distribution \( \mu(\lambda) = \lambda \) and \( \sigma^2(\lambda) = \lambda \). The method
of moments (MOM) involves equating $\mu(\theta)$ to $\hat{\mu}$ or equating $\sigma^2(\theta)$ to $\hat{\sigma}^2$, and solving the equation for $\theta$.

- In the geometric distribution, the MOM estimate of $p$ based on the mean is

$$\hat{p} = 1/\hat{\mu}.$$ 

The MOM estimate based on the variance is

$$\hat{p} = \frac{-1 + \sqrt{1 + 4\hat{\sigma}^2}}{2\hat{\sigma}^2}.$$ 

- MOM estimates can be very useful, but they often do not achieve the minimum MSE, $\text{MSE}^*$.

- If $X_1, \ldots, X_n$ are iid with distribution $f_\theta(X)$, the joint distribution is given by

$$p(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i),$$

and the logarithm of this is given by

$$L(\theta|X_1, \ldots, X_n) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

This is called the log likelihood function, and it is the basis of the maximum likelihood estimate (MLE).

- The MLE is defined to be the value of $\theta$ that maximizes the log likelihood function:

$$L(\hat{\theta}_{\text{MLE}}|X_1, \ldots, X_n) \geq L(\theta|X_1, \ldots, X_n)$$
for all $\theta \in \Theta$.

- If the log-likelihood is a convex and differentiable function of $\theta$, then the MLE is equal to the root of the derivative of the log-likelihood function:

$$\frac{d}{d\theta} L(\hat{\theta}_{MLE}|X_1, \ldots, X_n) = 0.$$ 

- If $X_1, \ldots, X_n$ have a geometric distribution with parameter $p$, then

$$L(p|X_1, \ldots, X_n) = (X. - n) \log(1 - p) + n \log(p).$$

The derivative of the log likelihood is

$$\frac{d}{dp} L(p|X_1, \ldots, X_n) = \frac{n - X.}{1 - p} + \frac{n}{p},$$

and the root is

$$\hat{p}_{MLE} = X./n,$$

where $X. = \sum_i X_i$.

- Under fairly general conditions, for iid observations, as $n \to \infty$, the MSE of the MLE becomes lower than the MSE of any other estimator. However for small to moderate values of $n$ it is often possible to beat the MLE.

- In the case of the normal distribution $N(\mu, \sigma^2)$, the MLE for $\mu$ is the same as the MOM estimate of $\mu$, and the MSE for $\sigma^2$ is
\[ \hat{\sigma}_{\text{MLE}}^2 = \frac{\sum_{i}(X_i - \hat{\mu})^2}{n}, \]

which is nearly identical to the MOM estimate.

- A posterior mean estimate is a type of Bayesian estimate, in contrast to the MOM and MLE estimates which are frequentist estimates. Bayesian estimates are constructed by treating the parameter \( \theta \) as a random variable. It has a distribution \( p(\theta) \) that is called the prior distribution.

- We observe \( X_1, \ldots, X_n \) as iid realizations from the observation distribution \( f_\theta(X) \), which in the Bayesian setting is more often written \( p(X|\theta) \). Estimation of \( \theta \) is carried out using the posterior distribution of \( \theta \) given \( X \), namely

\[
p(\theta|X_1, \ldots, X_n).
\]

- The posterior distribution can be computed using Bayes’ theorem:

\[
p(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)p(\theta)}{p(X_1, \ldots, X_n)}.
\]

- There are a variety of different Bayesian estimators, but they have in common that they are functions of the posterior distribution. We will focus on the posterior mean:

\[
\hat{\theta} = E(\theta|X_1, \ldots, X_n) = \int \theta p(\theta|X_1, \ldots, X_n).
\]

- Suppose we observe a binomial random variable \( X \) based on a sample of size \( n \), so that the observation distribution is
\[ p(X|\theta) = \binom{n}{X} \theta^X (1 - \theta)^{n-X}. \]

Suppose we take as the prior distribution a uniform distribution, so \( p(\theta) \equiv I(0 \leq \theta \leq 1) \). Then the posterior distribution is given by

\[ p(\theta|X) \propto \theta^X (1 - \theta)^{n-X}. \]

As in this case, it is very common to begin by computing the posterior distribution without its \textit{normalizing constant}. That is, we drop all factors that do not involve \( \theta \). The normalizing constant is uniquely determined by the constraint \( \int p(\theta|X)d\theta = 1 \). Therefore the strategy is to try to recognize \( p(\theta|X) \) (without its normalizing constant) as a familiar density, and then use the known structure of that density to obtain the correct constant.

- In this example, we recognize the posterior density as a Dirichlet distribution with parameters \( \alpha_1 = X + 1 \) and \( \alpha_2 = n - X + 1 \). Therefore the posterior distribution (with its normalizing constant) is

\[ p(\theta|X) = \frac{\Gamma(X+1)\Gamma(n-X+1)}{\Gamma(n+2)} \theta^X (1 - \theta)^{n-X}. \]

The posterior mean (which is our estimate of \( \theta \)) is

\[ \hat{\theta} = \frac{X + 1}{n + 2}, \]

which is not much different from the MLE, which is \( \hat{\theta}_{\text{MLE}} = X/n \).
Uncertainty in Point Estimates

- Since $\hat{\theta}(X)$ is generally different from $\theta$, it is useful to have a means for assessing how far $\theta$ may lie from $\hat{\theta}$. This leads to the ideas of error bars or confidence intervals.

- The standard approach is to select a level $0 \leq \alpha \leq 1$ and then construct upper and lower limits $\hat{\theta}_1(X)$ and $\hat{\theta}_2(X)$ so that

$$P(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X)) \geq \alpha.$$ 

Standard choices for $\alpha$ are .9, .95, and .99.

- In general, $P(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X))$ varies with $\theta$. Therefore to be more correct we should require

$$\inf_{\theta} P(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X)) \geq \alpha.$$ 

- Ideally, we would like to find a scaling function $\hat{S}$ so that $(\hat{\theta} - \theta)/\hat{S}$ is pivotal, meaning that its distribution does not depend on $\theta$.

- Let $q^*$ be such that $P(|(\hat{\theta} - \theta)/\hat{S}| < q^*) = \alpha$. This equation holds for all values of $\theta$. Therefore

$$\alpha = P(-q^* \leq (\hat{\theta} - \theta)/\hat{S} \leq q^*)$$

$$P(\hat{\theta} - q^*\hat{S} \leq \theta \leq \hat{\theta} + q^*\hat{S})$$

In our original notation we can set $\hat{\theta}_1 = \hat{\theta} - q^*\hat{S}_n$ and $\hat{\theta}_2 = \hat{\theta} + q^*\hat{S}_n$ and achieve a level $\alpha$ confidence interval.
• For example, if $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then setting $\hat{S} = \hat{\sigma}/\sqrt{n-1}$ we have

$$\frac{\bar{X}_n - \mu}{S(\theta)} = \sqrt{n-1}\frac{\bar{X}_n - \mu}{\hat{\sigma}},$$

which has the same distribution regardless of the values of $\mu$ and $\sigma$ (specifically it has a $t$ distribution with $n - 1$ degrees of freedom). This gives the confidence interval

$$\hat{\mu}_1 = \bar{X}_n - q^\ast\hat{\sigma}/\sqrt{n-1}$$
$$\hat{\mu}_2 = \bar{X}_n + q^\ast\hat{\sigma}/\sqrt{n-1}.$$

• In practice it is difficult to find pivotal estimators, so it is common to fall back on one of two approximations.

• The first approximation is to pretend that the random variable is normal and use the pivotal construction given above.

• The second approximation is known as the plug-in method. There are several variations on this approach. Here we will search for a constant $C > 1$ such that the following holds:

$$P_{\theta = \hat{\theta}}(\hat{\theta}/C \leq \theta \leq C\hat{\theta}) \geq \alpha.$$ 

In words, we assume that the estimate $\hat{\theta}$ is exact for purposes of constructing the confidence interval. Then we determine a factor by which we must scale up and scale down the estimate so that the desired coverage is achieved. Since $\hat{\theta}$ will usually differ from $\theta$, the probability $P_{\theta}(\hat{\theta}/C \leq \theta \leq C\hat{\theta})$ may be smaller than $\alpha$ for certain values of $\theta$. In this context, $\alpha$ is referred to as
the nominal level, and the true probability that the confidence interval contains $\theta$ is called the coverage probability.

• For example, suppose we observe 2 *iid* geometric random variables $X = 37, Y = 43$, and we want to construct a confidence interval for $p$ based on the MLE of $p$. It can be shown that the distribution of $Z = X + Y$ has mass function $P(Z = k) = (k - 1)p^2(1 - p)^{k-2}$ for $k \geq 2$. Recall that the MLE is $\hat{p} = 2/Z = 1/40$. Next we find $C > 1$ such that

$$P_{p=1/40}(\hat{p}_{\text{MLE}}/C \leq p \leq \hat{p}_{\text{MLE}}C) = P_{p=1/40}(2/Cp < Z < 2C/p) \geq \alpha.$$ 

For instance if $\alpha = .95$, we find $C = 5.4$, so the plug-in confidence interval for $p$ is $[.005, 0.135]$.

**Hypothesis Testing**

• Hypothesis testing refers to the act of making a binary decision based on data. For example, we may observe *iid* $X_1, X_2, X_3$ and pose the question “is $E X_i = 0$?” We can not answer this question with certainty, but we can formulate decision procedures that have desirable statistical properties.

• We will consider decisions involving two hypotheses, where a hypothesis can be defined as a statement about a distribution. For example, if we are observing geometrically distributed data, a hypothesis might be that the parameter $p$ is less than .1. In a parametric setting, the hypothesis may be identified with the
set of parameter values that are allowed under the hypothesis, e.g. $H = [0, 1]$.

- The conventional frequentist framework for hypothesis testing is called the Neyman-Pearson framework. This is an asymmetrical framework, in which one hypothesis (the null hypothesis) is presumed to be true unless there is overwhelming evidence that the other hypothesis (the alternative hypothesis) is true. The null hypothesis represents conventional wisdom that should not be overturned based on a single experiment unless the evidence is very strong.

- The first step in hypothesis testing is always to construct a test statistic $T = T(X_1, \ldots, X_n)$ that measures the evidence in the data in favor of the alternative hypothesis. Usually the test statistic is formulated so that large values of $T$ indicate greater evidence for the alternative hypothesis ($H_1$) while values of $T$ close to 0 indicate greater evidence for the null hypothesis ($H_0$).

- To ensure that $H_0$ is rejected only in the face of “overwhelming evidence”, we specify a level $\alpha$ (usually .05 or .01) and use it to determine a cut point $T_0$ such that $P_{H_0}(T > T_0) \leq \alpha$. We decide in favor of $H_0$ if $T < T_0$ and we decide in favor of $H_1$ if $T \geq T_0$.

- For example, suppose we observe a binomial observation $X$ based on $n = 30$ Bernoulli trials with a common, unknown success probability $p$. We want to test $H_0 : p \leq .5$ against $H_1 : p > .5$. We use $T \equiv X - 15$ as the test statistic. The cut point $T_0$ is defined to be the smallest $T_0$ such that
\[
\sum_{k=T_0+15}^{30} \binom{30}{k} 2^{-30} \leq \alpha.
\]
For example if \(\alpha = .05\) then \(T_0 = 5\), and if \(\alpha = .01\) then \(T_0 = 7\).

- A more informal way of carrying out the hypothesis test is to state the \textit{p-value} for the observed data. The p-value is defined to be \(P_{H_0}(T > T_{\text{obs}})\). A small p-value suggests that the observed value \(T_{\text{obs}}\) would have been very unlikely to occur if the null hypothesis were true. This can be taken as a measure of evidence in favor of the alternative hypothesis. When working with p-values, it is not necessary to state a level \(\alpha\) or a cut point \(T_0\).

- In the binomial example, suppose we observed \(T_{\text{obs}} = 4\). The corresponding p-value is

\[
\sum_{k=19}^{30} \binom{30}{k} 2^{-30} \approx 0.10.
\]
This is not considered to be strong evidence in most situations. If we had observed \(T_{\text{obs}} = 7\), the p-value would be \(\approx .003\), which is considered to be strong evidence for \(H_1\).

- Hypotheses \(H_0\) and \(H_1\) are said to be \textit{nested} if \(H_0 \subset H_1\). For example \(H_0: p = .5\) and \(H_1: p \geq .5\) are nested. When working with nested hypotheses, the \textit{likelihood ratio method} provides a means for constructing test statistics with good properties. The log likelihood ratio is defined to be

\[
\text{LLR} \equiv \max_{\theta \in H_0} \log f_\theta - \max_{\theta \in H_1} f_\theta(X).
\]
Since the hypotheses are nested, \(\text{LLR} < 0\). We take the test statistic to be \(T = -2\text{LLR}\).
It is a fact that under general conditions, $-2\text{LLR}$ approximately has a $\chi^2_d$ distribution under $H_0$, where $d$ is the dimension of the parameter set for $H_1$ minus the dimension for the parameter set for $H_0$. Since $\chi^2_d$ tables are readily available, this fact makes it easy to construct approximate p-values and cut-points.

- For example, suppose we observe trinomial data according to the probability vector $(p_1, p_2, p_3)$. That is, $X_1, \ldots, X_n$ are iid with $P(X_i = k) = p_k$ for $k = 1, 2, 3$. We observe $Z_1, Z_2, Z_3$, where $Z_k = \sum I(X_i = k)$. The mass function for $Z_1, Z_2, Z_3$ is

$$\frac{n!}{Z_1! Z_2! Z_3!} p_1^{Z_1} p_2^{Z_2} p_3^{Z_3}.$$ 

Now suppose we want to test $H_0 : p_1 = p_2$ against the alternative where the $p_i$ vary freely. The likelihood ratio is

$$\max_p \{(Z_1 + Z_2) \log(p) + Z_3 \log(1 - 2p)\}$$

$$- \max_{p_1, p_2, p_3} \{Z_1 \log(p_1) + Z_2 \log(p_2) + Z_3 \log(p_3)\}.$$ 

It is not hard to show that the two maxima occur when $p = (Z_1 + Z_2)/2n$ and $p_k = Z_k/n$ for $k = 1, 2, 3$. The null hypothesis has 1 dimension and the alternative has two dimensions, so $T = -2 \cdot \text{LLR}$ has approximately a $\chi^2_1$ distribution under $H_0$.

Suppose $n = 20$ and we observe $Z_1 = 7$, $Z_2 = 4$, and $Z_3 = 9$. Taking $p = 11/40$, $p_1 = 7/20$, $p_2 = 4/20$, and $p_3 = 9/20$, we get $T = .83$. From a $\chi^2$ table we find that the p-value associated with $T$ is $\approx 0.37$, so there is not much evidence that $p_1$ and $p_2$ are different.
Hypothesis Testing in 2-way Tables

• Suppose we observe two random variables $X$ and $Y$ that have finite sample spaces. The following two questions are commonly of interest:

1. If $X$ and $Y$ have the same sample space, we may ask “do $X$ and $Y$ have the same distribution?” That is, does $P(X = z) = P(Y = z)$ for every point in the sample space?
2. If $(X, Y)$ are observed jointly, we may ask “are $X$ and $Y$ independent.”

• As an example of question 1, we may wish to compare two DNA sequences $S_1$ and $S_2$ of length 1000. Suppose we base our comparison on the triplet distribution. For each of the 64 possible triplets, let $Z_1, \ldots, Z_{64}$ denote the number of occurrences in $S_1$, and let $W_1, \ldots, W_{64}$ denote the number of occurrences in $S_2$.

• Another example of question 1 would occur if we were to observe 30 sequences of DNA having length 12 that are thought to be target sequences for a certain DNA binding protein. It may be of interest to know whether the base at position 3 has the same distribution as the base at position 5.

• As an example of question 2, continuing with the previous point, we may observe wish to know whether the base at position 3 varies independently of the base at position 7.

• All of these questions can be answered using the likelihood ratio test. For example, suppose that among the 30 sequences, the number of occurrences of A, T, G, C at position 3 is (12, 11, 3, 4), and the number of occurrences at position 5 are (5, 4, 11, 10). It
appears that position 5 may be more GC rich, but is this merely sampling variation?

Let \( H_0 \) denote the hypothesis that the probabilities are equal, so a probability vector \( \pi \) describes both distributions. To construct the alternative \( H_1 \), we pretend that positions 3 and 5 are independent, which is likely not true, but it is a common assumption that makes the calculation much easier. Under the \( H_1 \), \( \pi_1 \) denotes the distribution for position 3, and \( \pi_2 \) denotes the distribution for position 5. Next we get the MLE’s of \( \pi \), \( \pi_1 \), and \( \pi_2 \). They are as follows:

\[
\hat{\pi} = (17/60, 15/60, 14/60, 14/60) \\
\hat{\pi}_1 = (12/30, 11/30, 3/30, 4/30) \\
\hat{\pi}_2 = (5/30, 4/30, 11/30, 10/30).
\]

The likelihood ratio is:

\[
17 \log(17/60) + 15 \log(15/60) + 14 \log(14/60) + 14 \log(14/60) - 12 \log(12/30) - 11 \log(11/30) - 3 \log(3/30) - 4 \log(4/30) - 5 \log(5/30) - 4 \log(4/30) - 11 \log(11/30) - 10 \log(10/30) \\
\approx -6.94.
\]

Thus \(-2 \text{LLR} = 13.8\), which if \( H_0 \) is true has a \( \chi^2 \) distribution. From a \( \chi^2 \) table, the p-value is .003, so there is strong evidence that the distributions at position 3 and 5 differ.

- A likelihood ratio test of this type is very similar to a \( \chi^2 \) test of homogeneity. These \( \chi^2 \) tests have the same \( \chi^2_d \) reference
distribution as the likelihood ratio test, but they use a different test statistic, called the \( \chi^2 \) test statistic, which has the general form:

\[
\sum_{\text{cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}.
\]

The expected count for cell \( i \) is just \( n \hat{p}_i \), where \( \hat{p}_i \) is the MLE for \( P(X = v_i) = P(Y = v_i) \) under \( H_0 \) (\( v_1, \ldots, v_m \) is the sample space), and \( n \) is the number of observations. For the current example, the expected cell counts are \((8.5, 7.5, 7, 7)\), and the \( \chi^2 \) statistic is

\[
(12 - 8.5)^2/8.5 + (11 - 7.5)^2/7.5 + (3 - 7)^2/7 + (4 - 7)^2/7 + \\
(5 - 8.5)^2/8.5 + (4 - 7.5)^2/7.5 + (11 - 7)^2/7 + (10 - 7)^2/7 \\
\approx 13.29.
\]

Now we turn to question 2, namely, whether \( X \) and \( Y \) are independent. Suppose that the bases at positions 3 and 5 in the 30 sequences have the following counts:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

We let \( H_0 \) be the hypothesis in which position 3 and 5 have their own marginal distributions, but are independent of each other (we could also force them to have the same marginal distribution). Thus \( H_0 \) is described by two probability vectors \( \pi_1 \) and \( \pi_2 \).
Hypothesis $H_1$ has no restrictions, so it is specified by a probability distribution $Q$ on the product sample space containing 16 points. The MLE’s for these parameters are:

$$\hat{\pi}_1 \approx (0.10, 0.07, 0.43, 0.40)$$
$$\hat{\pi}_2 \approx (0.40, 0.37, 0.10, 0.13)$$

$$\hat{Q} \approx \begin{pmatrix}
0.03 & 0.07 & 0.00 & 0.00 \\
0.00 & 0.07 & 0.00 & 0.00 \\
0.13 & 0.13 & 0.07 & 0.10 \\
0.23 & 0.10 & 0.03 & 0.03
\end{pmatrix}$$

The likelihood ratio test statistic is

$$3 \log(0.10) + 2 \log(0.07) + 13 \log(0.43) + 12 \log(0.40) + 12 \log(0.40) + 11 \log(0.37) + 3 \log(0.10) + 4 \log(0.13) - \log(0.03) - 2 \log(0.07) - 2 \log(0.07) - 4 \log(0.13) - 4 \log(0.13) - 2 \log(0.07) - 3 \log(0.10) - 7 \log(0.23) - 3 \log(0.10) - \log(0.03) - \log(0.03) \approx -4.29$$

This gives $-2LLR \approx 8.6$, which has a p-value under a $\chi^2_9$ distribution of 0.48, so there is no evidence that the two positions are independent.

Recall that $-2LLR$ is not exactly $\chi^2_9$ distributed under the $H_0$, this is only an approximation that becomes better as the number
of observations grows. In this case there aren’t very many observ-
ations, so the $\chi^2$ approximation may not be very close. Thus
the p-value should come with a warning, In general it is com-
mon to say that the expected counts (under the null hypothesis
should all be 5 or greater in order for the $\chi^2$ approximation to
be accurate.

- There is also a $\chi^2$ counterpart to the likelihood ratio test for
  independence. We use the same basic formula as with the ho-
mogeneity test, but here we derive the expected cell counts under
$H_0$ as specified in the previous point. We get

$$
\begin{pmatrix}
1.2 & 1.2 & 0.3 & 0.3 \\
1.2 & 1.2 & 0.3 & 0.3 \\
4.8 & 4.8 & 1.2 & 1.2 \\
4.8 & 4.8 & 1.2 & 1.2 \\
\end{pmatrix}
$$

Thus the $\chi^2$ test statistic is

$$(1 - 1.2)^2/1.2 + (2 - 1.2)^2/1.2 + (2 - 1.2)^2/1.2 + (4 - 4.8)^2/4.8 +
(4 - 4.8)^2/4.8 + (2 - 1.2)^2/1.2 + (3 - 1.2)^2/1.2 + (7 - 4.8)^2/4.8 +
(3 - 4.8)^2/4.8 + (1 - 1.2)^2/1.2 + (1 - 1.2)^2/1.2
\approx 6.35.$$ 

Relative to a $\chi^2_9$ reference distribution the p-value is 0.70, so
there is no evidence of dependence.

**Comparison of Two Distributions**

- There are many situations in which one wants to compare whether
two distributions on a common sample space are identical. That
is, given mass or density functions $f$ and $g$, is it true that $f(z) = g(z)$ for every $z$ in the sample space?

- If the sample space is finite (and not too large), then the $\chi^2$ test of homogeneity can be applied.

- More generally, there are a number of formal strategies such as the *Kolmogorov-Smirnov* test. Suppose we observe $N_X$ observations $X_1, X_2, \ldots$, from distribution $f$ and $N_Y$ observations $Y_1, Y_2, \ldots$, from distribution $g$. The test statistic is defined by

$$T = \max_Z |\#\{X_i : X_i < Z\}/N_X - \#\{Y_i : Y_i < Z\}/N_Y|.$$ 

The p-value is given by

$$Q(T(\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e})), $$

where $N_e = N_X N_Y / (N_X + N_Y)$, and

$$Q(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 \lambda^2).$$

The Kolmogorov-Smirnov statistic can also be used to compare the distribution of observed data to a fixed distribution (a normal distribution or a geometric distribution, for instance). In this case use $N_e = N_X$, and replace $\#\{Y_i : Y_i < Z\}/N_Y$ with the exact probability $P(Y < Z)$ under the distribution.

- A more informal strategy is to compute the empirical quantiles of the two distributions and compare them in a scatterplot. This is called a Q-Q plot. If $x_1, \ldots, x_n$ are observed, then the *order statistics* are just the data values sorted:
For $0 \leq q \leq 1$, the $q^{th}$ quantile is

$$x_{([1+q(n-1)]^\alpha)} + (1 - \alpha)x_{([1+q(n-1)])},$$

with $\alpha = [1 + q(n - 1)] - 1 - q(n - 1)$.

Now if we observe $Z_1, X_2, \ldots$, and $Y_1, Y_2, \ldots$ and want to compare their distributions, we compute the quantiles for $X$ and the quantiles for $Y$ (on a suitable grid, say, 0.01, 0.02, $\ldots$, 0.99, 1), and plot them against each other in a scatterplot. If the distributions are equal then the scatterplots should fall along the 45° diagonal.

An important application is to the clustering of matches to a given short DNA sequence (e.g. a DNA binding site template). It is often of interest to know whether the matches to the template are uniformly distributed, or if they tend to be packed in clusters, or if they are even more evenly spaced than would be expected by chance. Let $x_1, x_2, \ldots, x_n$ be the number of nucleotides between consecutive matches. If the binding sites are iid, then the $x_i$ should follow a geometric distribution with parameter $p = \#\text{matches/target sequence length}$. Using a Q-Q plot of the $x_i$ quantiles against the geometric quantiles, we can determine whether (i) the binding sites are uniformly distributed (the plot appears as a diagonal line), (ii) the binding sites cluster (most of the points are below the diagonal), or (iii) the binding sites are exceptionally evenly dispersed (most of the points are above the diagonal).