Lecture 26: Distribution-free Tests and Other Stuff

Relevant textbook passages:
Larsen–Marx [14]: Chapter 14

26.1 ★ Principal components

According to Ted Anderson [1, p. 272],

In many exploratory studies, the number of variables under con-
sideration is too large to handle. Since it is the deviations in these
studies that are of interest, a way of reducing the number of variables
to be treated is to discard the linear combinations which have small
variances and study only those with large variances.

The first principal component of a random vector $X$ is the normalized
linear combination $w_1$ that maximizes the variance of $w'X$. The normalization
is that $w_1'w_1 = 1$. The $k^{th}$ principal component is the linear combination that
maximizes the variance subject to normalization and being uncorrelated with the
first $k-1$ principal components. Thus if $X$ is a vector in $\mathbb{R}^n$ it has $n$ principal
components.

Assume that $X$ has variance-covariance matrix $\Sigma$ and mean 0. (The assump-
tion of zero mean is simply a normalization.) Then

$$\text{Var} w'X = E(w'X)^2 E w'XX'w = w'\Sigma w.$$ 

I thought it was well-known that the $w$ that maximized $w'\Sigma w$ subject to $w'w = 1$
is an eigenvector of $\Sigma$ corresponding to the largest eigenvalue $\lambda_1$ and that $\lambda_1$ is
also the Lagrange multiplier associated with the maximization and the value of
the maximum. I even use this my linear algebra notes to prove the Principal
Axis Theorem. Recent correspondence has led me to believe that this result may
not be as widely-known as I thought, so I am reproducing it here. For the proof,
see [3], Anderson [1, pp. 273–275], Carathéodory [5, §195], Franklin [9, Section 6.2,
pp. 141–145], or Rao [17, 1f.2.iii, p. 62].

26.1.1 Proposition (Extrema of quadratic forms on the sphere) Let $A$
be an $n \times n$ symmetric matrix. Define $x^1, \ldots, x^n$ recursively so that $x^{k+1}$ maximizes
the quadratic form $Q(x) = x \cdot Ax$ over $S_k = S \cap M_{k\perp}$, where $S$ is the unit sphere
in $\mathbb{R}^n$, and $M_k$ denotes the span of $x^1, \ldots, x^k$, with $M_0 = \{0\}$. Then each $x^k$,
k = 1, \ldots, $n$ is an eigenvector of $A$, and $\lambda_k = Q(x^k)$ is its corresponding eigenvalue.
Recall from Lecture 22 that the Principal Axis Theorem 22.4.2 states there is an orthogonal $C$, whose columns are eigenvectors of $\Sigma$, such that the diagonal matrix $\Lambda$ of eigenvalues of $\Sigma$ satisfies

$$\Sigma C \Lambda C'.$$ 

Since $C$ is orthogonal, it is an isometry of $R^n$ (Proposition 22.3.2), that is, a combination of rotations and reflections. The transformation

$$U = CX$$

expresses $X$ in terms of the new basis, and $\text{Var} U = \Lambda$. Moreover, we can arrange the columns of $C$ so that the $k^{th}$ column $C^k$ gives $C^k X$ is the $k^{th}$ principal component of $X$. That is, $U = CX$ is the vector of principal components of $X$.

### 26.2 MLE of principal components

Principal Components are defined in terms of eigenvectors of the variance-covariance matrix $\Sigma$, which is unknown. Thus they must be estimated. Since the MLE of a function of parameters is the function of MLE of the parameter, we must compute the eigenvectors of the MLE of $\Sigma$. So if $\text{seq} XN$ is a sample of independent draws from a multivariate Normal($\mu, \Sigma$) distribution,

$$\hat{\Sigma}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})'.$$

There are now algorithms that work quite well at computing eigenvectors, but this course is not the right one to get into the subtleties. Remember, Trust, but verify.

### 26.3 Distribution-free tests

All of the significance testing we have discussed so far has been based on likelihood functions. That is we assume we know the function $f(x; \theta)$. There are hypothesis test that do not rely such knowledge. Instead they rely on the fact (Glivenko–Cantelli Theorem 8.9.3) that the empirical cdf is approximately the same as the cdf of the data generating process. Since the empirical cdf is a step function with jumps at the sample values, the jumps occur at the order statistics of the sample, and the next set of tests rely heavily on order statistics. This material is covered in Larsen–Marx [14, Chapter 14], but Breiman [4, Chapters 8–9], Hogg and Craig [10, § 9.6], and. van der Waerden [18, § 63–64] provide more detail.
26.4 A test for the median

If we have a sample from a continuous pdf, there is a simple test for the null hypothesis

$$H_0: \text{median } f = \theta_0, \quad \text{against the alternative } H_1: \text{median } f \neq \theta_0.$$ 

By definition the probability of exceeding the median is 1/2, so for an independent sample $$X_1, \ldots, X_n$$, the statistic

$$T = |\{i : x_i > \theta_0\}|$$

has a Binomial$$\left(n, \frac{1}{2}\right)$$ distribution, which has mean $$n/2$$ and variance $$n/4$$.

For the two sided alternative the test takes the form choose a critical value $$k^*$$ and reject the null hypothesis if

$$T \leq k^* \text{ or } T \geq n - k^*.$$ 

Unless you want to use a randomized rule, you are unlikely to find to find a $$k^*$$ that will give you a significance level, of say 0.05, but you can compute the Binomial$$\left(n, \frac{1}{2}\right)$$ probability of $$P (T \leq k^* \text{ or } T \geq n - k^*)$$ to get the size of the test.

Or if $$n$$ is large enough ($$\geq 10$$) you could use the Normal approximation and treat $$z = (T - n/2)/\sqrt{n}/4$$ as a standard Normal and reject the null if $$|z| \geq z_{\alpha/2}$$, to get a test of size $$\alpha$$.

One-sided tests are analogous.

26.5 Testing the equality of two distributions

Suppose we have two random samples $$x_1, \ldots, x_n$$ and $$y_1, \ldots, y_m$$, and want to know if they came from the same continuous distribution. The null hypothesis and the alternative are

$$H_0: f_X = f_Y \quad H_1: f_X \neq f_Y.$$ 

This exposition is based on Breiman [4, pp. 290–298].

26.5.1 A simple test

If $$n + m$$ is even, a simple idea to test the null hypothesis is this: Let $$\bar{m}$$ be the sample median. If the null hypothesis is true, then the number $$N$$ of the $$x_i$$s that are less than $$\bar{m}$$ should be about half of the $$x_i$$s, namely $$n/2$$. In fact, the exact distribution of $$N$$ is given by

$$P (N = k) = \binom{n}{k} \binom{m}{n+m-k} \binom{n+m}{k}.$$
This is known as the **hypergeometric distribution**. It is equivalent to the following experiment. An urn contains \( n \) red balls and \( m \) white balls. A sample of size \( (n + m)/2 \) is drawn at random without replacement. Then \( P(N = k) \) is the probability that \( k \) of them are red. See, e.g., Pitman [16, p. 125].

For large \( n, m \) this is approximately normal with mean \( n/2 \) and variance \( (nm/(n + m - 1))/4 \). This can be used as a basis for a test in the usual fashion. However this test is very inefficient.

To be fair, this is really only a test of the hypotheses 

\[ H_0: \text{median } f_X = \text{median } f_Y, \quad \text{against the alternative } \quad H_1: \text{median } f_X \neq \text{median } f_Y. \]

### 26.6 The Wilcoxon–Mann–Whitney test

A better test is the following test, based on the order statistics, known as the **Wilcoxon test** or the **Mann–Whitney test**. (According to William Kruskal [11], this test was independently discovered at least seven times, dating back to Deuchler [6] in 1914.) Array the combined sample from smallest to largest, sort of like this:

\[
x_3 < x_7 < y_4 < \cdots < y_2 < x_5,
\]

and assign them ranks from 1 to \( n + m \) starting. Let

\[
s_i = |\{j : y_j > x_i\}|, \quad r_i = |\{j : x_j > x_i\}|.
\]

Then the rank \( t_i \) of \( x_i \) is just \( n + m - s_i - r_i \). The Wilcoxon test statistic is

\[
T = \sum_{i=1}^{n} t_i.
\]

If the null hypothesis is true, then \( T \) has the same distribution as the sum of \( n \) numbers drawn at random without replacement from the set \( 1, \ldots, n + m \). The Mann–Whitney test statistic\(^1\) is

\[
U = \sum_{i=1}^{n} s_i.
\]

\[
T = \sum_{i=1}^{n} (n + m - s_i - r_i), \quad U = \sum_{i=1}^{n} s_i,
\]

so

\[
T + U = n(n + m) - \sum_{i=1}^{n} r_i,
\]

but a moment’s reflection should tell you that

\[
\sum_{i=1}^{n} r_i = (n - 1) + (n - 2) + \cdots + 1 + 0 = n(n - 1)/2.
\]

\(^1\)Van der Waerden [18, p. 275] defines the Mann–Whitney statistic by \( U = \sum_{i=1}^{n} (m - s_i) \).
Therefore
\[ U + T = n(n + m) - \frac{n(n - 1)}{2} = nm + \frac{n(n + 1)}{2}, \]
so knowing one statistic tells us the other.

It turns out the Mann–Whitney statistic $U$ is more convenient to work with. Let $h(x, y)$ be the indicator of $x < y$. That is
\[ h(x, y) = \begin{cases} 1 & \text{if } x < y \\ 0 & \text{otherwise} \end{cases}. \]

Then, if the null hypothesis is true, $P(X < Y) = 1/2$, so
\[ U = \sum_{i=1}^{n} \sum_{j=1}^{m} h(x_i, y_j), \quad \text{so} \quad E U = \sum_{i=1}^{n} \sum_{j=1}^{m} E h(x_i, y_j) = nm P(X < Y) = nm/2. \]

A tedious computation along the same lines shows that
\[ \text{Var} U = \frac{mn(n + m + 1)}{12}. \]

Moreover, it can be shown that the standardized $U$,
\[ \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(n + m + 1)}{12}}} \]
is approximately Normal$(0, 1)$ when $m$ and $n$ are both large. [18, p. 277].

It is possible to get a recursive formula for the exact distribution of $U$ and it can be used when $n$ and $m$ are small. Here is the argument [15]: Let $\varphi(u; m, n)$ denote the probability that $U = u$ under the null hypothesis. There are two ways $U = u$ can occur. Case 1: The largest value is an $x$, so it contributes nothing to the sum that is $U$. The probability of this is just $n/(n + m)$. Case 2: The largest value is a $y$, so that if we were to drop it, each $s_i$ would decrease by 1. This happens with probability $m/(n + m)$. Thus
\[ \varphi(u; m, n) = \frac{n}{n + m} \varphi(u, n - 1, m) + \frac{m}{n + m} \varphi(u - n, n, m - 1). \]

There are some simple to compute boundary cases: $\varphi(u; 0, k) = \varphi(u; k, 0) = 1$ if $u = 0$ and $u = 0$ if $u > 0$ and $k > 1$. And $\varphi(u; n, m) = 0$ if $u < 0$.

Note that $U/mn$ is an unbiased estimate of $P(X < Y)$. Mann and Whitney [15] proposed it as a test of the following hypotheses:
\[ H_0: f_X = f_Y \quad H_1: f_Y \text{ stochastically dominates } f_X, \]
where, as you may recall, $Y$ dominates $X$ if for all $t$, $PY > t \geq P(X > t)$. For such hypotheses, a one-sided test is appropriate, and the null should be rejected for large values of $U$. 

KC Border
26.7 The Kruskal–Wallis test

The Wilcoxon rank test can be extended to several samples. It is best suited to testing location of distributions.

Start with a continuous density \( f \). Define

\[
f_\theta(x) = f(x - \theta).
\]

Then \( f_\theta \) and \( f_{\theta'} \) differ only in their location parameter \( \theta \).

For example, the mean \( \mu \) of a normal distribution is a location parameter (for \( \sigma^2 \) fixed).

Consider \( k \geq 2 \) independent samples of sizes \( n_1, \ldots, n_k \) from distributions \( f_{\theta_1}, \ldots, f_{\theta_k} \).

How can we test the hypothesis

\[ H_0: \theta_1 = \cdots = \theta_k \]

against the alternative

\[ H_1: \text{not all } \theta_j \text{s are equal.} \]

The idea behind the Kruskal–Wallis test is that if the location parameters are all the same, then the values of \( x_1, \ldots, x_n \) from each sample ought to be “evenly distributed” among the set of values. So arrange the \( n = n_1 + \cdots + n_k \) values in order from smallest to largest, and assign each its rank in the list (average out ties). Let \( R_{ij} \) denote the rank of \( x_{ij} \), the \( i \)th observation of the \( j \)th group, in the overall list. Define

\[
R_{\bullet j} = \sum_{i=1}^{n_j} R_{ij}.
\]

Under the null hypothesis, we would expect the average rank \( R_{\bullet j}/n_j \) to be about the same for each \( j \). In fact, under \( H_0 \) the test statistic

\[
B = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{\bullet j}^2}{n_j} - 3(n+1)
\]

is approximately \( \chi^2 \) with \( k-1 \) degrees of freedom. The null hypothesis \( H_0 \) should be rejected at the \( \alpha \)-level of significance if \( B > \chi^2_{1-\alpha,k-1} \). [14, Theorem 14.4.1]

26.7.1 Example (Case study 14.4.1, pp. 678, [14]) The first Vietnam war draft lottery was held on Dec 1, 1969. The lottery was conducted by putting capsules with birthdays in an urn (actually a plexiglass cylinder), mixing it up, and then drawing them out. The first birthday drawn got rank 1, the next 2, and down to rank 366. Men who were born from 1944 through 1950 were subject to the draft in order of their rank. (Low ranks first.) The last birthday called for service was the 195th.

An unusual pattern emerged. The later months in the year had much lower average ranks than the early months. Only five birthdays in December had
ranks greater than 195. It turns out the urn had been loaded with January on
the bottom, then February, etc., and not very thoroughly mixed.

A Kruskal–Wallis test of the hypothesis that the month averages are equal
yielded a $\chi^2(11)$ statistic of 25.95, which has a $p$-value of 0.006602.

The lotteries conducted in later years, for later birth cohorts were better
designed.

26.8 The bootstrap

Bootstrapping is a relatively new\footnote{I have found that academics view anything that was published after they started graduate
school as recent, and anything published before then as classic.} statistical technique that was developed by
Bradley Efron \cite{Efron1979} in 1979. It is based on a Monte Carlo method of resampling. The idea is this. Suppose you have an unknown distribution, or a
statistic whose distribution is quite difficult to determine. How do you construct
confidence intervals for the value of interest?

If you have a reasonably large data set, the Glivenko–Cantelli Theorem tells
you that the empirical cdf of your data is very close to the distribution function
that your data is drawn from. So the quantiles of your sample should be close to
being uniformly distributed. What this means is that drawing a point uniformly
at random from your sample is almost like drawing from the distribution.

So if you start with sample of size $n$, you can create a new sample of size $n$ by
drawing $n$ points independently with replacement from a uniform distribution on
your sample. With this new sample, you can recompute your statistic. Do this over
and over again and you get $m$ values for your statistic. The empirical cdf of this
sample should be close to the actual cdf of the statistic. Thus you can compute
confidence intervals for your test statistic, without actually knowing anything
about its true distribution. Well, that’s the idea, and it works for many problems.
For a further discussion, see, e.g., Efron and Gong \cite{Efron1983}, and for some theory, and
some cases where bootstrapping is ineffective, see Bickel and Freedman \cite{Bickel1981}.

26.8.1 Example (The World Series) We computed the maximum likelihood
estimate of the probability that the better team wins the world series as $\hat{p}_{MLE} = 0.594$. How confident are we in this estimate?

Recall that the log-likelihood function for $p$, the probability that the better
team wins, is (up to a constant),

$$
\ln L(p; n_4, n_5, n_6, n_7) = \sum_{k=4}^7 n_k \ln(p^4(1-p)^{k-4} + (1-p)^4p^{k-4})
$$

where $n_k$ is the number of Series that lasted $k$ games.

If I was really good at math, I could derive the distribution of the estimator
and use that to find a confidence interval. Or I could ask my computer to resample
the data 10,000 times and recomputes $\hat{p}_{MLE}$ each time. Since it only takes about 0.1 seconds to compute the estimator, in about 1,000 seconds I will have my answer.\(^3\) I’m pretty sure that’s less time than it would take me to derive the distribution. Plus I can go out for a *caffe latte* while the program is running.

Figure 26.1 shows the histogram from 10,000 simulated 102-Series samples. As you can see, the distribution is bimodal, with the most common estimate being 0.5. The MLE was equal to 0.5 for 1164 simulations. The two-sided 95% confidence interval (running from the 0.025 quantile to the 0.975 quantile) is

$$[0.5000, 0.6666]$$

The mean $\hat{p}_{MLE}$ is 0.585918 and the standard deviation is 0.0503606.

Figure 26.1: The results of bootstrapping the MLE for the World Series problem 10,000 times.

The bootstrap has made it possible to test incredibly complicated models, such as neural nets or decision trees, and is a staple of *machine learning*.

**Bibliography**


\(^3\)It actually took 945.135 seconds for Mathematica 8 on a one-year old MacBook Air. Last year, with 101 Series, it took 1083.25 seconds using Mathematica 9 on a five-year old Mac Pro.
http://www.jstor.org/stable/2240410

http://www.hss.caltech.edu/~kcb/Notes/QuadraticForms.pdf


This was originally published in 1935 in two volumes by B. G. Teubner in Berlin as *Variationsrechnung und Partielle Differentialgleichungen erster Ordnung*. In 1956 the first volume was edited and updated by E. Hölder. The revised work was translated by Robert B. Dean and Julius J. Brandstatter and published in two volumes as *Calculus of variations and partial differential equations of the first order* by Holden-Day in 1965–66. The Chelsea second edition combines and revises the 1967 edition.


http://www.jstor.org/stable/2958830

http://www.jstor.org/stable/2685844


http://www.jstor.org/stable/2280906

http://www.jstor.org/stable/2280779

http://www.jstor.org/stable/2281082


