22 Nonparametric Methods.

In parametric models one assumes apriori that the distributions have a specific form with one or more unknown parameters and one tries to find the best or atleast reasonably efficient procedures that answer specifc questions regardng the parameters. If the assumptions are violated our procedures might become faulty. Often the procedures are still valid even if they are not the most efficient, and these are the stable or robust situations. Sometimes we could be way off. Let us discuss this by means of two examples. We have n observations $x_1, \cdots, x_n$ from some population and we want to test that the mean is 0. If we assume that the observations come from the normal population with mean μ and unknown variance σ^2, we would naturally use the t test. The statistic would be

$$t = \frac{\bar{x}}{s}\sqrt{n-1}$$

where $\bar{x}$ is the sample mean $\frac{\sum_i x_i}{n}$ and s^2 is the sample variance $\frac{\sum_i x_i^2}{n} - \bar{x}^2$. The staistic t has a t distribution with $n-1$ degrees of freedom. Far large n it is nearly normal with mean 0 and variance 1. If in reality the observations came from an exponential distribution with density $ae^{-ax}dx$ for $x \geq 0$, while for small n t is nolonger distributed as a "t" asymptotically it is still distrbuted like a standard normal with mean 0 and variance 1. Using the sample mean to do the t-test is robust for questons regarding the population mean. Let us look at the problem testing that the variance is 1. If we use the statistic based on the sample variance ns^2 and do the χ^2 test, which is natural for the Normal model, asymptotically

$$U_n = \frac{ns^2 - (n-1)}{\sqrt{2}n}$$

will be standard normal. But if the model were exponential and the observations are drawn from $e^{-x}dx$, although $E[ns^2] = n-1$, its variance is different and it is only

$$V_n = \frac{ns^2 - (n-1)}{\sqrt{5}n}$$

that is asymptotically normal. We are way off.

The nonparametric odels avoids these issues and makes no assumption or atleast only very general assumptions concerning the model. For instance if we want to test that $x_1, \ldots, x_n$ are drawn from a population with median 0, we do it simply by counting the number of the number of observations that are above 0. This random variable X, is a Bnomial with probability $\frac{1}{2}$ and far large n

$$\frac{X - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$$

is asymptotically normal.

One fairly general assumption that is often made is that the probabilty distribution from which the samples are drawn are continuous i.e. the distribution function

$$F(x) = P[X \leq x]$$

is a continuous functon of x. Then it is easy to check that the random variable $Y = F(X)$ which lies between 0 and 1 has the uniform distribution on $[0, 1]$. To see this let us suppose for simplicity that F is strictly increasing. Then

$$P[F(X) \leq y] = P[X \leq F^{-1}(y)] = F(F^{-1}(y)) = y$$

proving that the distribution of $Y = F(X)$ is uniform.

If we have n observations and we want to test if F is the true underlying distribution we may want to compare the empirical distribution

$$F_n(x) = \frac{[\#i : x_i \leq x]}{n}$$

with $F(x)$ and use the Kolmogorov=-Smirnov statistic

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$$

It turns out that if we employthe transformation $F(x_i) = y_i$ and calculate

$$D_n^* = \sqrt{n} \sup_{0 \leq y \leq 1} |\frac{[\#i : y_i \leq y]}{n} - y|$$

The distribution of D_n, under the assumption that the observations come from F is the same as that of D_n^* under the assumption that y_i come from the uniform distribution on $[0, 1]$. The asymptotics of this statistic has been worked out. The distribution of

$$D_n^*(t) = \sqrt{n}\Big[\frac{[\#i : y_i \leq t]}{n} - t\Big]$$

is asymptotically normal with variance $t(1-t)$. One can see this easily from the fact that $[\#i : y_i \leq t]$ is a binomial $B(n,t)$. The joint distribution of $\{D_n^*(s), D_n^*(t)\}$ is bivariate normal with covariance $\min(t,s) - ts$. From these considerations one can deduce that asymptotically the distribution of D_n^* is that of

$$\sup_{0 \leq t \leq 1} |Z(t)|$$

where $Z(t)$ is a Normal random function with mean zero and covariance $\min(s,t) - st$.

If one wants to test if two sets of samples $x_1, x{-}2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$ come from the same population F against the alternative while the x's come from F, the y's come from a shifted distrbution $F(x-a)$ for some $a > 0$. A test called rank test is used for this. Let us group the $n+m$ observations and arrange them in increasing order. The rans of the y's are some numbers $1 \leq k_1 \leq k_2 \leq \cdots \leq k_m \leq n+m$. Under the null hypothesis we expect them to uniformly spaced in $[0, m+n]$, while under the alternative they should bunch up to the right end. We want to use the statistic

$$U_{n,m} = \sum_i k_i$$

and compute its mean and variance. It is known that

$$V_{n,m} = \frac{U_{n,m} - E[U_{n,m}]}{\sqrt{\mathrm{Var}\ U_{n,m}}}$$

is asymptotically normal.

Let us compute the mean and variance. The following trick is often useful in similar contexts. Let us define $Z_i = 1$ if the i-th smallest observation is a y and 0 otherwise. Then

$$\sum_i k_i = \sum j Z_j$$

Let us compute $E[Z_j]$ and $E[Z_i Z_j]$.

$$E[Z_i] = \frac{m}{n+m}$$

is the probability that the j-th observation is a y. Note that under the null hypothesis they are all from the same population so that all possible arrangement have the same probability. Similarly

$$E[Z_i Z_j] = \frac{m(m-1)}{(n+m)(n+m-1)}$$

$$E[U_{n,m}] = \frac{n}{(n+m)}\big[\sum_j j\big] = \frac{n+m+1}{2}$$

$$\text{Var } U_{n,m} = \text{Var } [Z_j][\sum_j j^2] + \text{Cov } [Z_j Z_k][\sum_{j \neq k} jk]$$

The varinace of Z_j is computed easily to be $\frac{nm}{(n+m)^2}$ while the covariance between Z_i and Z_j equals with $N = m + n$

$$\begin{aligned}
\frac{m(m-1)}{(n+m)(n+m-1)} - \frac{m^2}{(n+m)^2} &= \frac{m}{N^2(N-1)}\Big[N(m-1) - (N-1)m\Big] \\
&= -\frac{m\ n}{N^2(N-1)}
\end{aligned}$$

The variance can now be computed as

$$\begin{aligned}
\text{Var } U_{m,n} &= \frac{mn}{N^2}\frac{N(N+1)(2N+1)}{6} - \frac{m\ n}{N^2(N-1)}\Big[(\sum_i i)^2 - \sum_i i^2\Big] \\
&= \frac{mn}{N}\frac{(N+1)(2N+1)}{6} - \frac{m\ n}{N}\frac{(N+1)}{12}(3N+2) \\
&= \frac{m\,n(N+1)}{12}
\end{aligned}$$

Finally suppose we have a finite population from which we draw a sample without replacement. The population is $a_1, \ldots, a_N$ and we draw a sample $x_1, x_2, \ldots, x_n$ of size n. We want to compute the mean and variance of the sample mean $\bar{x}$. It is better to work with $S = \sum a_j Z_j$ where Z_j is 1 if a_j is included in the sample.

$$E[Z_j] = \frac{n}{N} \quad \text{Var } [Z_j] = \frac{n(N-n)}{N^2}$$

$$\text{Cov } [Z_i Z_j] = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)}$$

From this it is easy to deduce that

$$E[\bar{x}] = \bar{a} = \frac{\sum a_j}{N}$$

and

$$\begin{aligned}
\text{Var } \bar{x} &= \frac{1}{n^2}\left[\frac{n(N-n)}{N^2}\sum_i a_i^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{i,j} a_i a_j\right] \\
&= \frac{N-n}{nN}\frac{1}{N}\sum_i (a_i - \bar{a})^2 \\
&= \frac{N-n}{nN}\text{Var } a
\end{aligned}$$