KOLMOGOROV–SMIRNOV AND MANN–WHITNEY–WILCOXON TESTS

1. The Kolmogorov test

Let F be any probability distribution function on the real line $\mathbb{R}$. Recall that the distribution function of a random variable X is $F(x) := F_X(x) := \Pr(X \le x)$. For any real numbers $x_1, ..., x_n$, the corresponding *empirical distribution function* is defined by $F_n(x) := \frac{1}{n}\sum_{j=1}^n 1_{x_j \le x}$ where $1_{x_j \le x} = 1$ if $x_j \le x$ and 0 otherwise. Let $x_1, ..., x_n$ in order (order statistics) be $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Then we will have $F_n(x) = 0$ for $x < x_{(1)}$, $F_n(x) = j/n$ for $x_{(j)} \le x < x_{(j+1)}$ for each $j = 1, ..., n-1$, and $F_n(x) = 1$ for $x \ge x_{(n)}$.

Usually and here, $x_1, ..., x_n$ will be the observed values of some random variables $X_1, ..., X_n$. Suppose we want to test the hypothesis H_0 that $X_1, ..., X_n$ are i.i.d. with a given, fixed distribution function F. Let F_n be the empirical distribution function based on $X_1, \ldots, X_n$. One form of the Kolmogorov test statistic for H_0 is

$$D_n := \sup_x |(F_n - F)(x)|.$$

Any F is continuous from the right and at each x has a left limit $F(x-) := \lim_{y \uparrow x} F(y)$. Because F is nondecreasing, and F_n is constant between consecutive order statistics, we will have

$$D_n = \max_{1 \le j \le n} \max(|(F_n - F)(X_{(j)}-)|, |(F_n - F)(X_{(j)})|).$$

So to compute D_n, given that we can compute F, does not take excessive computation.

If H_0 is true, then as $n \to \infty$, D_n will approach 0 at a $1/\sqrt{n}$ rate, as will follow from Theorem 2. (Similarly, a sample mean $\overline{X}$ of i.i.d. variables $X_1, ..., X_n$ with $E(X_1^2) < \infty$ approaches the true mean μ at that rate.) So, it can be useful to normalize the statistic, giving

$$K_n := \sqrt{n} \sup_x |(F_n - F)(x)|.$$

H_0 will be rejected if K_n is too large. Note that $0 \le D_n \le 1$, so it never becomes large. About the distributions of D_n and K_n under H_0, here is one fact:

Date: 18.465, Feb. 1, 2015 .

Theorem 1. *If $X_1, ..., X_n$ are i.i.d. (F), then for each $n = 1, 2, ...$, the distribution of D_n is the same for all continuous F, and so, the same is true for K_n.*

We'll return later to a proof of the theorem. The practical importance of it is that for F continuous, one can tabulate the distribution of D_n (or K_n) for each n, or more economically give quantiles of interest such as the 0.95 and 0.99 quantiles, so that H_0 is rejected at respective levels 0.05 or 0.01 if the statistic is larger than the respective quantile.

The meaning of "continuous" in Theorem 1 is the usual mathematical one. For a nondecreasing function such as F it means it has no jumps. Some beginning probability texts define "continuous" for a distribution (function) to mean there is a density f of which F is the indefinite integral. That is the case in practice, for example, for the usual parametric families of continuous distributions in probability and statistics. There do exist, however, continuous distribution functions without densities. One is the "Cantor function," on which there is a Wikipedia article. Proofs about continuous F to be given in this handout will use only continuity, not densities.

For large n, the asymptotic distribution for K_n is useful and exists, as follows (it is a consequence of facts in Section 5):

Theorem 2. *If $X_1, X_2, ...,$ are i.i.d. with the continuous distribution function F, then for $K_n := \sqrt{n} \sup_x |(F_n - F)(x)|$, and any M with $0 < M < \infty$,*

(1)

$$\lim_{n \to \infty} Pr(K_n \geq M) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 M^2) < 2 \exp(-2M^2).$$

This is an infinite series, not summable in closed form, but for M at all large, it converges very fast. The terms alternate in sign and decrease in absolute value as j increases. The kth partial sum is given by the R function (not built in, but supplied for this course on www-math.mit.edu/∼rmd/465) supabsbt: once this is brought into one's R working folder with source("supabsbt") it can be called by supabsbt(M,k).

Dvoretzky, Kiefer and Wolfowitz (1956) proved that for some constant C with $2 < C < \infty$, for all n and F, under H_0, $\Pr(K_n > M) \leq C \exp(-2M^2)$. P. Massart (1990) proved this for the best possible constant $C = 2$:

Theorem 3 (Massart)**.** *For any distribution function F (continuous or not) and empirical distribution function F_n based on $X_1, ..., X_n$ i.i.d.*

(F), for any $M > 0$,

$$Pr(\sqrt{n}\sup_x |(F_n - F)(x)| > M) \le 2\exp(-2M^2). \tag{2}$$

Massart's inequality has a more detailed proof in Dudley (2014), §1.5.

1.1. **The one-sample Kolmogorov test: approximate quantiles by regression.** For the Kolmogorov one-sample statistic D_n (which they call D) under H_0, Hollander and Wolfe (1999) tabulate the $q = 1-\alpha$ quantiles for $n = 1, \ldots, 40$ and $\alpha = 0.01$, 0.02, 0.05, 0.1, and 0.2. For $n > 40$ H.& W. propose some coefficients over $\sqrt{n}$, where the coefficients are corresponding quantiles of the asymptotic distribution for K_n, in other words H. & W. propose to use the asymptotic distribution as $n \to \infty$ for all $n > 40$. The quantiles for $n \le 40$ for given α or q are decreasing monotonically as n increases. However, if one plugs in and gets the resulting values for $n = 40$ or 41, the decreasing pattern is violated by amounts in the third significant digit, and this remains true if the asymptotic quantiles are replaced by more precise ones such as 1.6278 in place of 1.63 or 1.3581 in place of 1.358. One can get a more accurate approximation as follows.

For $x(q, n)$ the quantile given in the table ($q = 1 - \alpha$) let $y(q, n) := \sqrt{n}x(q, n)$. Specifically, consider the the $q = 0.95$ column. For consecutive values of n, specifically $n = 35, 36, 37, 38, 39, 40$ there is a lot of noise apparently resulting from rounding error, so consider $n =$ 15, 20, 25, 30, 35, 40. For these n's, the 0.95 and 0.99 quantiles of D_n under H_0 are given in www/math.mit.edu/∼rmd/465/onesamplequants. A correlation of -0.98980 of $y(q, n)$ with $1/\sqrt{n}$ suggests regression should work well. Regressing $y(q, n)$ on $1/\sqrt{n}$ gave an intercept of 1.3591 and a slope of -0.1948. But more precisely, for the asymptotic distribution, the 0.95 quantile is 1.3581 to the given number of decimal places, so let's use that as the intercept. To fit a line $y = a + bx$ for least-squares y-on-x regression with fixed a and given data vectors X and Y, one can see that the slope b must equal $X \cdot (Y - a)/(X \cdot X)$ where $\cdot$ is the dot product and $Y - a$ means the vector with jth component $Y_j - a$ for each j. This gives $b \doteq -0.1439 \doteq -0.144$. The resulting regression predicted the 0.95 quantile accurately to three significant digits (which is all the given x's have) up to an error of 1 in the third digit, for $n = 30,\ 35,\ 40$, and for $n \to +\infty$, when the quantile $y(q, n)$ converges to a limit which equals 1.3581 to the given number of digits.

In summary, for $q = 0.95$, in place of the formula $1.36/\sqrt{n}$ for $n > 40$ (proposed by Hollander and Wolfe), a more accurate formula appears

to be

$$\frac{1.358}{\sqrt{n}} - \frac{0.144}{n}.$$

The formula should work well if n is very large since 1.3581 is asymptotically correct. Similar formulas could be found for other q's by the same method.

1.2. **The Kolmogorov test in the R system.** There are 18 parametric families of probability distributions defined in R, listed in Venables and Ripley Table 5.1 p. 108. Some are discrete and some continuous. For the Kolmogorov test we're focusing on continuous distributions. One of the parametric families, for example, is the uniform distributions $U[a, b]$ for $-\infty < a < b < +\infty$. The R code for the family is "unif." It may be preceded by any of the four letters p, d, q, or r. Here "p" is used for the (cumulative) distribution function. Thus `punif(x,a,b)` would call for R to find the value at x of the distribution function of $U[a, b]$ at x, which is simply 0 for $x \le a$, $(x - a)/(b - a)$ for $a \le x \le b$, and 1 for $x \ge b$. Next, `dunif(x,a,b)` would give the density of the distribution at x, namely $1/(b - a)$ for $a \le x \le b$ and 0 otherwise; `qunif(y,a,b)` calls for the y quantile, if $0 < y < 1$, of the $U[a, b]$ distribution, in other words the x for which `punif(x,a,b)` equals y, that is, $x = a + (b - a)y$. (For a lot of distributions, such as the beta, gamma, t, and chi-squared distributions, the quantiles don't have such simple closed forms.) Lastly, `v = runif(n,unif,a,b)` in R would generate $V_1, ..., V_n$ i.i.d. $U[a, b]$ and store them in a vector $v = (V_1, ..., V_n)$.

Conversely, to test in R whether a given data vector v, with corresponding empirical distribution function F_n, has distribution $U[a, b]$, by the Kolmogorov test, one can type

ks.test(v, "punif",a,b).

Note that one needs not just "unif" but "punif." If the parameters "a" and "b" in a uniform distribution are omitted in R, the default is the $U[0, 1]$ distribution.

Likewise for the family of normal distributions, "norm" is the basic code. The two parameters are the mean μ and the standard deviation σ, with default values 0 and 1.

1.3. **Some mathematical facts relating to the Kolmogorov test.** Probability distribution functions can converge pointwise but not uniformly: for example, as $n \to \infty$, $1_{[-1/n,+\infty)}(x) \to 1_{[0,+\infty)}(x)$ for all x but not uniformly. We have, however:

Theorem 4 (Glivenko–Cantelli, 1933)**.** *For any distribution function F (continuous or not) and corresponding empirical distribution functions F_n based on $X_1, ..., X_n$ i.i.d. (F), almost surely,*
$D_n = \sup_x |(F_n - F)(x)| \to 0$ as $n \to \infty$, in other words, with probability 1, $F_n \to F$ uniformly.

A proof will be given later. The Glivenko–Cantelli theorem implies that the Kolmogorov test is what is called consistent against all alternatives:

Corollary 1. *Suppose that $X_1, ..., X_n, ...$ are i.i.d. F but we test by Kolmogorov's test the hypothesis H_1 that they are i.i.d. for some distribution function $G \neq F$ (where F and G need not be continuous). Then with probability $\to 1$ as $n \to \infty$, H_1 will (correctly) be rejected.*

Proof. By the Glivenko–Cantelli theorem, the statistic D_n will converge to $\sup_x |(F-G)(x)| > 0$, so $K_n \to \infty$ and by Theorem 2, H_1 will be rejected. □

1.4. **Testing composite hypotheses.** If $X_1, ..., X_n$ are i.i.d. G for some unknown G, then the Kolmogorov test is of the simple hypothesis H_0: $G = F$ for a specified F against the general alternative H_1: $G \neq F$. Such a test is sometimes, but not often, useful in practice. Much more often, the hypothesis H_0 is tested against a simple alternative H_1: $G = H$. Then let F have a density f and H a density h. One forms the likelihood ratio h/f (defined as $+\infty$ if $h > 0 = f$ and 0 when $h = 0$) and the product $LR_n := \prod_{j=1}^n (h/f)(X_j)$. One decides in favor of H_1 if $LR_n > c$ and in favor of H_0 if $LR_n \leq c$ (likelihood ratio test, Neyman–Pearson lemma) where c may be selected based on the costs c_i when H_i is true and H_{1-i} is chosen, and on the prior probabilities of the two hypotheses, assuming they are given. The problem is treated in beginning statistics (e.g. 18.443).

Against a general alternative, rather than testing a simple hypothesis, one often wants to test a composite hypothesis H_0, for example, that X_j are i.i.d. with a distribution of a given parametric form, say a normal distribution $N(\mu, \sigma^2)$ for some unknown μ and unknown $\sigma > 0$, against the alternative that the distribution is not in the given family. In the normal case, $(X_j - \mu)/\sigma$ are i.i.d. $N(0, 1)$ but not observed. If we try to estimate μ by the sample mean $\overline{X} := (X_1 + \cdots + X_n)/n$ and σ by the sample standard deviation $s_X := \left(\frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2\right)^{1/2}$, then $(X_j - \overline{X})/s_X$ all have the same distribution, but they are not independent, nor are they normally distributed (recall that $\sqrt{n}(\overline{X} - \mu)/s_X$ has a t distribution, which is very different from normal for small n). The

normality hypothesis can be tested by quite a different method, not via empirical distribution functions, in the Shapiro–Wilk test, which we'll get to in a couple of weeks.

A highly composite hypothesis that can be tested using empirical distribution functions is the following:

2. The Kolmogorov–Smirnov two-sample test

Suppose we're given $X_1, ..., X_m$ i.i.d. (F), independent of $Y_1, ..., Y_n$ i.i.d. G, and we want to test the hypothesis H_0 that the two distribution functions F and G are the same, where neither of them is specified in advance.

A natural test statistic is $D_{m,n} := \sup_x |(F_m - G_n)(x)|$. As in Theorem 1 in the one-sample case we have:

Theorem 5. *For given positive integers m and n, if the variables $X_1, ..., X_m, Y_1, ..., Y_n$ are all i.i.d. with a continuous distribution function F, then the distribution of $D_{m,n}$ does not depend on F.*

By the Glivenko–Cantelli theorem 4, if H_0 holds, then $D_{m,n} \to 0$ with probability 1 as m and n both go to $+\infty$. A test of H_0 based on a suitable multiple of $D_{m,n}$, namely, one takes $KS_{m,n} := \sqrt{\frac{mn}{m+n}} D_{m,n}$, will have the same asymptotic distribution as K_n:

Theorem 6. *If H_0: $F = G$ continuous holds, with $X_1, ..., X_m, Y_1, ..., Y_n$ all i.i.d. (F), then*

$$\lim_{m,n \to \infty} Pr(KS_{m,n} \geq M) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 M^2). \tag{3}$$

If $F = G$ is not continuous, then

$$\limsup_{m,n \to \infty} Pr(KS_{m,n} \geq M) \leq 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 M^2). \tag{4}$$

This will follow from facts to be given in Subsection 5.2.

Corollary 2. *A test of $H_0 : F = G$, based on $KS_{m,n}$, is consistent against all alternatives $F \neq G$ as m and n both go to $+\infty$, i.e. H_0 will be (correctly) rejected.*

Proof. For $F \neq G$, we will have with probability 1, as $m, n \to \infty$, $D_{m,n} \to \sup_x |(F - G)(x)| > 0$ by the Glivenko–Cantelli theorem. We also have $\sqrt{mn/(m+n)} = \sqrt{1/\left(\frac{1}{m} + \frac{1}{n}\right)} \to +\infty$, and so $KS_{m,n} \to +\infty$. Thus for any $M < \infty$, $\Pr(KS_{m,n} > M)$ will approach 1, whereas by Theorem 6, for $F = G$, the probability is small for M large. $\square$

But, finding p-values in the two-sample case with a given accuracy is considerably harder than in the one-sample case. The possible values of $D_{m,n}$ are of the form $\left|\frac{i}{m} - \frac{j}{n}\right|$ where $i = 0, 1, \ldots, m$ and $j = 0, 1, \ldots, n$ are integers. If the integer $L := L(m, n)$ is the least common multiple of m and n, then all possible values of $J := L(m, n)D_{m,n}$ are integers. Under H_0 the value $J = 0$ can't occur except with probability 0.

The distribution of J under H_0 for $1 \le m \le n = 1, \ldots, 20$ is tabulated by Hollander and Wolfe (1999) pp. 606-630, a 25-page table for the $\binom{20}{2} = 190$ possible pairs of values. For $m = n = 20$, so $L = 20$, the distribution has large atoms, and relatively few p-values are possible. For example 0.0811 is the smallest possible p-value larger than 0.05 and 0.0335 is the largest possible p-value less than 0.05. The probability that $J = 8$, the borderline value, is 0.0476 under H_0. There is an even larger atom of size 0.0934 at $J = 7$. If $G_{m,n}$ is the distribution function of $KS_{m,n}$, and F is any continuous distribution function (such as the asymptotic one), then $\sup_x |(G_{m,n} - F)(x)|$ is at least half the size of the largest atom (jump of $G_{m,n}$), so it can't be small.

For a statistic that has a continuous distribution, if one finds how often H_0 is rejected at level α, one will find that for a large number N of simulations it will be rejected about $N\alpha$ times, or in a fraction about α of the simulations. But for $m = n = 20$ the fraction of times H_0 will be rejected by the two-sample K-S test at level 0.05 will converge toward 0.0335 since that is the largest possible p-value less than 0.05.

On the other hand, for $m = 19$ and $n = 20$ the least common multiple is $L = mn = 380$ (since 19 is a prime), the table has to show many more values, and the atoms are much smaller. Since $\Pr(J \ge 152) = 0.0503$ one can do a test at level 0.05 almost exactly.

How to get quantiles for m or n larger than 20 but not very large? That seems difficult to do by hand or with tables. Partly because of the discrete distribution of the statistic, and since we have two variables m and n, the kind of regression done in the one-sample case seems not feasible. The 2-sample Kolmogorov–Smirnov test, given two samples $x = (X_1, \ldots, X_m)$ and $y = (Y_1, \ldots, Y_n)$, can be done in R, ks.test(x,y), so that will be the recommended method. According to the documentation, exact p-values are computed for $mn < 10,000$. For $mn \ge 10,000$, the asymptotic approximation is used.

As the asymptotic distribution for $m \to \infty$ and $n \to \infty$ of the two-sample statistic $KS_{m,n}$ under the null hypothesis $F = G$ continuous (3), and the upper bound when $F = G$ discontinuous (4) is the same as the asymptotic distribution for the one-sample statistic K_n (1), namely, the distribution of the supremum of the absolute value of the Brownian

bridge (10), and the equal sums are each $< 2\exp(-2M^2)$ as noted in (1), it's natural to ask whether one one has a two-sample version of the Dvoretzky–Kiefer–Wolfowitz inequality, namely, for some C,

$$\Pr(KS_{m,n} > M) \le C\exp(-2M^2), \tag{5}$$

for all m and n and all $M > 0$, and if so, whether it holds with $C = 2$ as in the one-sample case (Massart's inequality (2)). Fan Wei (MIT S.B. 2012) settled the case $m = n$. Namely, she proved that (5) does hold for $C = e \doteq 2.71828$ for all n, but that it holds with $C = 2$ if and only if $n \ge 458$. The inequality with $C = 2$ also fails for $1 \le m < n \le 3$. We don't know of any other violations. The paper by Wei and Dudley (2012) states results, and the longer version by Wei and Dudley (2011) gives more details, including Wei's proofs for $m = n$.

3. The Mann–Whitney–Wilcoxon rank-sum test

This is a test of whether two samples come from the same distribution, against the alternative that members of one sample tend to be larger than those of the other sample (a location or shift alternative). No parametric form of the distributions is assumed. They can be quite general, as long as the distribution functions are continuous.

The general assumption for the test is that real random variables $X_1, \ldots, X_m$ are i.i.d. with a distribution function F, and independent of $Y_1, \ldots, Y_n$ which are i.i.d. with another distribution function G, with both F and G continuous. The hypothesis to be tested, as in the Kolmogorov–Smirnov test, is H_0: $F = G$. There are two formulations of the test, one due to Mann and Whitney and the other to Wilcoxon. R uses the Mann–Whitney form, as follows: let W be the number of pairs (i, j) with $1 \le i \le m$ and $1 \le j \le n$ such that $Y_j \le X_i$. Then W is the test statistic. H_0 will be rejected if either W is too small, indicating that the X's tend to be less than the Y's, or if W is too large, indicating that the Y's tend to be less than the X's.

For m and n not too large, one can tabulate the distribution, but as with the Kolmogorov–Smirnov tests, tabulation is rather unwieldy. One can find the mean and variance of T_X under H_0 in terms of m and n and use that it is asymptotically normal if m and n are both large. The test can be done in R via `wilcox.test(x,y)` for x= $(X_1, ..., X_m)$, y= $(Y_1, ..., Y_n)$, which will evaluate the statistic W and give a p-value for the test. Since there are mn total pairs, if one does instead `wilcox.test(y,x)` one will get for the statistic $W' = mn - W$, but with the same p-value. R gives a warning message saying p-values

are not exact if any of the variables are tied. That is because the p-values are computed assuming the distributions are continuous. Ties $X_i = Y_j$ for some i and j are a still worse problem because the value of the statistic W is uncertain, as it can be affected by arbitrarily small changes in X_i or Y_j.

4. Some facts about distribution and quantile functions

Beginning in this section, there will be some theoretical developments including proofs of statements made previously. Let F be a function from $\mathbb{R} = (-\infty, \infty)$ into itself. The "df properties" (which turn out to characterize of probability distribution functions) will be defined as the following four properties:

(i) F is nondecreasing;
(ii) $F(x) \to 1$ as $x \to +\infty$;
(iii) $F(x) \to 0$ as $x \to -\infty$;
(iv) F is right-continuous, i.e. $F(y) \to F(x)$ as $y \downarrow x$ for all x.

It follows from (i), (ii), and (iii) that F takes values in the interval $[0, 1]$. If F is continuous, it follows from (ii), (iii), and the intermediate value theorem that F takes all values in the open interval $(0, 1)$.

Each real-valued random variable X has a distribution function $F = F_X$ such that $F_X(x) = P(X \le x)$ for all x. Then F_X is easily seen to have all four df properties.

If F is continuous and strictly increasing from $(-\infty, \infty)$ onto $(0, 1)$, as is the standard normal distribution function Φ, then it has a unique inverse F^{-1} from $(0, 1)$ onto $\mathbb{R}$ such that $F(F^{-1}(u)) = u$ for $0 < u < 1$. But even if F is discontinuous or not strictly increasing, one can also define a kind of inverse as follows.

Definition. Let F be any function from $\mathbb{R}$ into $[0, 1]$ having the four df properties. For any y with $0 < y < 1$, let $F^{\leftarrow}(y) := \inf\{x : F(x) \ge y\}$.

Then we have the following fact:

Theorem 7. *Let F have the four df properties. Then:*
(a) For any y with $0 < y < 1$, $F^{\leftarrow}(y)$ is a well-defined real number.
(b) For $0 < y < 1$, $F(F^{\leftarrow}(y)) \ge y$.
(c) For any $x \in \mathbb{R}$ and $y \in (0, 1)$, $F(x) \ge y$ if and only if $x \ge F^{\leftarrow}(y)$.
(d) If U is a random variable having the uniform $U[0, 1]$ distribution, then $F^{\leftarrow}(U)$ is a random variable having the distribution function F.

Proof. Let $0 < y < 1$. Then because $F(x) \to 1$ as $x \to +\infty$, there must exist some x such that $F(x) \ge y$, so we are not taking the infimum

of the empty set. Moreover, the infimum cannot be $-\infty$, because if $F(x_n) \geq y$ for some $x_n \to -\infty$, that would contradict property (iii), which implies that such $F(x_n)$ must approach 0. So the infimum $F^{\leftarrow}(y)$ equals some finite x, proving part (a).

To prove part (b), let $x := F^{\leftarrow}(y)$. For some $x_n \downarrow x$ we have $F(x_n) \geq y$ and therefore by right-continuity of F at x, $F(F^{\leftarrow}(y)) = F(x) \geq y$ as stated.

To prove part (c), if $F(x) \geq y$ then $F^{\leftarrow}(y) \leq x$ by definition of $F^{\leftarrow}(y)$. Conversely, if $F^{\leftarrow}(y) \leq x$, then by part (b), $y \leq F(F^{\leftarrow}(y))$, and by the nondecreasing property of F, $F(F^{\leftarrow}(y)) \leq F(x)$, so $F(x) \geq y$ as stated.

Now for part (d), using part (c), for any real x,

$$\Pr(F^{\leftarrow}(U) \leq x) \;=\; \Pr(U \leq F(x)) \;=\; F(x)$$

since $0 \leq F(x) \leq 1$ and U has $U[0,1]$ distribution, having distribution function equal to the identity function on $[0,1]$. So $F^{\leftarrow}(U)$ does have distribution function F and the theorem is proved. □

In computer generation of (pseudo-) random variables, $U[0,1]$ variables are the most basic ones and are used in generating other random variables. If F is a distribution for which $F^{\leftarrow}$ is easy to compute, then taking $U_1, \ldots, U_n$ to be i.i.d. $U[0,1]$ and letting $X_j = F^{\leftarrow}(U_j)$ may be an efficient way to generate X_j i.i.d. (F). If $F^{\leftarrow}$ is relatively hard to compute there may be better ways. Later in the course we'll return to techniques of generating random variables with a given distribution.

For the present, we'll use the theorem for another purpose, to show how the distributions of Kolmogorov–Smirnov statistics don't depend on F for F continuous. Let $\mathcal{U}(x) = \max(0, \min(x, 1))$, the uniform $U[0,1]$ distribution function, which equals x for $0 \leq x \leq 1$, 0 for $x < 0$, and 1 for $x > 1$. We can take $U_1, \ldots, U_n$ i.i.d. with this distribution and form the corresponding empirical distribution function $\mathcal{U}_n$. For any two functions f and g such that f is defined on the range of g the composition $f \circ g$ is defined by $(f \circ g)(x) = f(g(x))$ for all x in the domain of g. We can, if we choose, form empirical distribution functions F_n for any distribution function F as follows.

Proposition 1. *Let F be any distribution function on $\mathbb{R}$, let $X_1, \ldots, X_n$ be i.i.d. F and let F_n be the empirical distribution function they define. Also let $\mathcal{U}_n$ be empirical distribution functions for $U[0,1]$. Then $\mathcal{U}_n \circ F$ have all the properties of F_n, so that we can assume $F_n \equiv \mathcal{U}_n \circ F$.*

Proof. By Theorem 7(d) we can assume that $X_j = F^{\leftarrow}(U_j)$ for each $j = 1, \ldots, n$. For each x and j, $F^{\leftarrow}(U_j) \leq x$ if and only if $U_j \leq F(x)$

by Theorem 7(c). Thus for each x, the number of values of j for which these inequalities occur is the same, and we have $F_n(x) = \mathcal{U}_n(F(x))$, so the conclusion follows. □

We also have, clearly, $\mathcal{U} \circ F \equiv F$ for any distribution function F and therefore for the one-sample Kolmogorov statistic we have, if the hypothesis H_0 of sampling from F holds,

$$\sup_x |(F_n - F)(x)| \;=\; \sup_x |(\mathcal{U}_n - \mathcal{U})(F(x))| \le \sup_{0\le t\le 1} |(\mathcal{U}_n - \mathcal{U})(t)|. \tag{6}$$

We have $(\mathcal{U}_n - \mathcal{U})(y) = 0$ with probability 1 for $y = 0$ or 1. Therefore if F is a continuous distribution function, so that its range includes the open interval $(0, 1)$ (and may or may not include either endpoint) it follows from (6) that under H_0.

$$\sup_x |(F_n - F)(x)| \;=\; \sup_{0<y<1} |(\mathcal{U}_n - \mathcal{U})(y)|. \tag{7}$$

In general the random variables on the two sides might have been defined on different probability spaces, but at any rate we can say that they are equal in distribution. Thus we have proved that the distribution of the one-sample Kolmogorov statistic doesn't depend on F for F continuous (Theorem 1). Moreover we can see from (6) that if F is discontinuous, the statistic will be smaller in distribution than if F is continuous. Thus we can still reject the hypothesis that F is the true distribution function when F is not continuous if we would when it is continuous, as this will tend to be overly conservative if anything.

Let $\mathcal{U}_m$ be an empirical distribution function for m i.i.d $U[0, 1]$ random variables and let $\mathcal{V}_n$ be another, independent, such empirical distribution function for n further i.i.d. $U[0, 1]$ variables. Then in the two-sample Kolmogorov–Smirnov test, if $X_1, \ldots, X_m$ are i.i.d. (F) and $Y_1, \ldots, Y_n$ are i.i.d. (G) and independent of $X_1, \ldots, X_m$, under the hypothesis H_0 that $F = G$, we will get as in (6)

$$\sup_x |(F_m - G_n)(x)| \;=\; \sup_x |(\mathcal{U}_m - \mathcal{V}_n)(F(x))| \le \sup_t |(\mathcal{U}_m - \mathcal{V}_n)(t)|, \tag{8}$$

and under the further condition that F is continuous, we will get as in (7) that

$$\sup_x |(F_m - G_n)(x)| \;=\; \sup_{0<y<1} |(\mathcal{U}_m - \mathcal{V}_n)(y)|, \tag{9}$$

and so the distribution thus will again not depend on F, proving Theorem 5.

Proof of the Glivenko–Cantelli theorem, Theorem 4: By Proposition 1, and since $\mathcal{U} \circ F \equiv F$, it suffices to prove this for the $U[0, 1]$ distribution $\mathcal{U}$. Given $\varepsilon > 0$, take a positive integer k such that $1/k < \varepsilon/2$. For

each $j = 0, 1, ..., k$, $\mathcal{U}_n(j/k) \to j/k$ as $n \to \infty$ with probability 1 by the ordinary strong law of large numbers. Take $n_0 = n_0(\omega)$ such that for all $n \geq n_0$ and all $j = 0, 1, ..., k$, $|\mathcal{U}_n(j/k) - j/k| < \varepsilon/2$. For t outside $[0, 1]$ we have $\mathcal{U}_n(t) \equiv \mathcal{U}(t) = 0$ or 1. For each $t \in [0, 1]$ there is at least one $j = 1, ..., k$ such that $(j-1)/k \leq t \leq j/k$. Then for $n \geq n_0$,

$$(j-1)/k - \varepsilon/2 < \mathcal{U}_n((j-1)/k) \leq \mathcal{U}_n(t) \leq \mathcal{U}_n(j/k) < j/k + \varepsilon/2.$$

It follows that $|\mathcal{U}_n(t) - t| < \varepsilon$, and since t was arbitrary, the theorem follows. □

5. Central limit theorems and the asymptotic distribution

Now we'll consider the limiting behavior of $\alpha_n := n^{1/2}(F_n - F)$ as $n \to \infty$. For any fixed t, the central limit theorem in its most classical form, for binomial distributions, says that $\alpha_n(t)$ converges in distribution to $N(0, F(t)(1-F(t)))$, in other words a normal (Gaussian) law, with mean 0 and variance $F(t)(1 - F(t))$.

In what follows, "RAP" will mean the book *Real Analysis and Probability* (Dudley, 2002).

For any finite set T of values of t, the multidimensional central limit theorem (RAP, Theorem 9.5.6) tells us that $\alpha_n(t)$ for t in T converges in distribution as $n \to \infty$ to a normal law $N(0, C_F)$ with mean 0 and covariance $C_F(s, t) = F(s)(1 - F(t))$ for $s \leq t$.

Although the multidimensional central limit theorem was proved by mathematicians only about 1930, Karl Pearson, in his publication of the χ^2 test of goodness of fit in 1900, implicitly assumed a central limit theorem for multinomial distributions.

For any set T and any probability space Ω on which a probability P is defined, a real-valued *stochastic process* $\{x_t(\omega),\ t \in T,\ \omega \in \Omega\}$ is a function of $t \in T$ and $\omega \in \Omega$ such that for each $t \in T$, $x_t(\cdot)$ is a real-valued random variable defined on Ω. Empirical distribution functions $F_n(t)$ for any n and empirical processes $\sqrt{n}(F_n - F)(t)$ are examples of stochastic processes.

The *Brownian bridge* (RAP, Section 12.1) is a stochastic process $B_t(\omega)$ defined for $0 \leq t \leq 1$ and ω in some probability space Ω, such that for any finite set $S \subset [0, 1]$, B_t for t in S have distribution $N(0, C)$, where $C = C_{\mathcal{U}}$ for the uniform distribution function $\mathcal{U}(t) = t$, $0 \leq t \leq 1$. So by Proposition 1 the empirical process α_n converges in distribution to the Brownian bridge composed with F, namely $t \mapsto B_{F(t)}$, at least when restricted to finite sets.

It was then natural to ask whether this convergence extends to infinite sets or the whole interval or line. Kolmogorov (1933) showed that

when F is continuous, the supremum $\sup_t \alpha_n(t)$ and the supremum of absolute value, $\sup_t |\alpha_n(t)|$, converge in distribution to the laws of the same functionals of B_F. Then, these functionals of B_F have the same distributions as for the Brownian bridge itself, since F takes $\mathbb{R}$ onto an interval including $(0,1)$ and which may or may not contain 0 or 1; this makes no difference to the suprema since $B_0 \equiv B_1 \equiv 0$. Also, $B_t \to 0$ almost surely as $t \downarrow 0$ or $t \uparrow 1$ by sample continuity; the suprema can be restricted to a countable dense set such as the rational numbers in $(0,1)$ and are thus well-defined random variables (in measure-theoretic terms, they are measurable).

To work with the Brownian bridge process it will help to relate it to the well-known Brownian motion process x_t, defined for $t \geq 0$, also called the Wiener process. This process is such that for any any finite set $T \subset [0, +\infty)$, the joint distribution of $\{x_t\}_{t \in F}$ is $N(0, C)$ where $C(s,t) = \min(s,t)$. This process has independent increments, namely, for any $0 = t_0 < t_1 < \cdots < t_k$, the increments $x_{t_j} - x_{t_{j-1}}$ for $j = 1, ..., k$ are jointly independent, with $x_t - x_s$ having distribution $N(0, t-s)$ for $0 \leq s < t$. Recall that for jointly Gaussian (normal) random variables, joint independence, pairwise independence, and having covariances equal to 0 are equivalent. Having independent increments with the given distributions clearly implies that $E(x_s x_t) = \min(s,t)$ and so is equivalent to the definition of Brownian motion with that covariance.

Brownian motion can be taken to be sample continuous, i.e. such that $t \mapsto x_t(\omega)$ is continuous in t for all (or almost all) ω. This theorem, proved by Norbert Wiener in the 1920's (while teaching 20 hours a week here at MIT), is Theorem 12.1.5 in RAP; a proof (due to Paul Lévy) will be indicated here. If Z has $N(0,1)$ distribution then for any $c > 0$, $\Pr(Z \geq c) \leq \exp(-c^2/2)$ (RAP, Lemma 12.1.6(b)). Thus if X has $N(0, \sigma^2)$ distribution for some $\sigma > 0$ then $\Pr(X \geq c) = \Pr(X/\sigma > c/\sigma) \leq \exp(-c^2/(2\sigma^2))$. It follows that for any $n = 1, 2, ...$ and any $j = 1, 2, ...,$

$$\Pr\left(\left| x_{j/2^n} - x_{(j-1)/2^n} \right| \geq \frac{1}{n^2} \right) \leq 2 \exp\left(-2^n/(2n^4) \right).$$

It follows that for any integer $K > 0$, the probability of any of the above events occurring for $j = 1, ..., 2^n K$ is at most $2^{n+1} K \exp(-2^n/(2n^4))$, which approaches 0 very fast as $n \to \infty$, because of the dominant factor -2^n in the exponent. Also, the series $\sum_n 1/n^2$ converges. It follows by the Borel–Cantelli Lemma (RAP, Theorem 8.3.4) that with probability 1, for all $t \in [0, K]$, for a sequence of dyadic rationals $t_n \to t$ given by the binary expansion of t, x_{t_n} will converge to some limit X_t, which

equals x_t almost surely. Specifically, for $t < K$, let $t_n = (j-1)/2^n$ for the unique $j \le 2^n K$ such that $(j-1)/2^n \le t < t/2^n$. Then $t_{n+1} = t_n = 2j/2^{n+1}$ or $t_{n+1} = (2j-1)/2^{n+1}$, so that t_{n+1} and t_n are either equal or are adjacent dyadic rationals with denominator 2^{n+1}, and the above bounds apply to the differences $x_{t_{n+1}} - x_{t_n}$.

The process X_t is sample-continuous and is itself a Brownian motion, as desired. From here on, a "Brownian motion" will always mean a sample-continuous one.

One can represent B_t for $0 \le t \le 1$ as $X_t - tX_1$ or as X_t conditional on $X_1 = 0$ or more rigorously, as the limit of X_t given $|X_1| < \varepsilon$ as ε decreases to 0.

Under H_0, $\sqrt{n}D_n$ converges in distribution to $\sup_{-\infty<x<\infty} |B_{F(x)}|$ (some details are given in Subsection 5.1). If F is continuous, this supremum equals $\sup_{0\le t\le 1} |B_t|$ (it doesn't matter whether F takes the values 0 or 1, because $B_t = 0$ with probability 1 for $t = 0$ or 1.) The distribution of this supremum is known: for each $M > 0$, $\Pr(\sup_{0\le t\le 1} B_t \ge M) = \exp(-2M^2)$ (RAP Prop. 12.3.3) where $\exp(x) \equiv e^x$ and for the absolute value we're actually interested in,

$$\Pr\left(\sup_{0\le t\le 1} |B_t| \ge M\right) = 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2M^2) \tag{10}$$

(RAP, Prop. 12.3.4), which is the same as the asymptotic distributions given in (1) and (3).

5.1. **The Komlós–Major–Tusnády statement with Bretagnolle and Massart's constants.** To show how the uniform empirical process $\sqrt{n}(\mathcal{U}_n - \mathcal{U})$ converges to the Brownian bridge process B_t with respect to uniform convergence for $0 \le t \le 1$, a rate of convergence has been found. It's formulated as follows.

Komlós, Major, and Tusnády (1975), to be called KMT hereafter, stated a rate of convergence, namely that on some probability space there exist X_i i.i.d. $U[0,1]$ and Brownian bridges B_n such that

$$P\left(\sup_{0\le t\le 1} |(\alpha_n - B_n)(t)| > \frac{x + c\log n}{\sqrt{n}}\right) \;<\; Ke^{-\lambda x} \tag{11}$$

for all n and x, where c, K, and λ are positive absolute constants. Assuming this and taking x also to be of the form $C\log n$, we see that the probability will go to 0 like any desired power of n, so that α_n and B_n will be uniformly close of order $(\log n)/\sqrt{n}$. This is rather remarkable, as in the one-dimensional central limit theorem, the best possible rate of convergence in general (given non-zero skewness) is of order $1/\sqrt{n}$, and here we have an infinite-dimensional central limit

theorem with convergence slower only by the very moderate factor $\log n$.

KMT formulated a construction giving a joint distribution of α_n and B_n, and this construction has been accepted by later workers. But KMT gave hardly any proof for (11). After partial proofs by others, Bretagnolle and Massart (1989) gave a proof of the inequality (11) with specific constants, as follows.

Theorem 8 (Bretagnolle and Massart)**.** *The approximation (11) of the $U[0,1]$ empirical process by the Brownian bridge holds with $c = 12$, $K = 2$ and $\lambda = 1/6$ for $n \geq 2$.*

A proof, with more details than in Bretagnolle and Massart's (1989) paper, is given in Dudley (2014), §1.4.

The Komlós–Major–Tusnády–Bretagnolle–Massart theorem is very important theoretically, as it establishes how uniform empirical processes become close to Brownian bridges. But let's compare what we might learn from this theorem in practice about the distribution of the one-sample Kolmogorov statistic, as compared to what we get from the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (2). From the latter, we get that we can reject H_0 in the one-sample case at the 0.05 level if $\sup_x \sqrt{n}|(F_n - F)(x)| = \sqrt{n}D_n = M$ where $2\exp(-2M^2) \leq 0.05$. This is equivalent to $M \geq \sqrt{(\log 40)/2} \doteq 1.358102 \doteq 1.3581$. But now supposing n is large so that the asymptotic distribution holds to a sufficient approximation, and we could apply the distribution given by (10). We can compute the series up through k terms by the R function supabsbt(x,k). R gives
supabsbt(1.3581,k) = 0.05000041
for $k = 1$ (just giving the Massart bound), whereas for $k = 2, 3$ or larger,
supabsbt(1.3581,k) = 0.04999963,
which is the same to 7 decimal places or rather to 5 significant digits. Thus, there is only a tiny, almost imperceptible improvement in adding the further terms beyond the first term of (10) for the critical value 1.3581 of M. It seems that n would have to be extremely large to apply the KMT–Bretagnolle–Massart theorem to take advantage of this small possible improvement, given that $(\log n)/\sqrt{n}$ does not approach 0 very fast.

Neither of the methods just compared, based on proved theorems, shows how for finite n, $\sqrt{n}D_n$ is actually smaller in distribution than the limit distribution. For example, for $n = 40$, $\alpha = 0.05$, the Hollander and Wolfe table gives the critical value for $\sqrt{n}D_n$ as $0.210 \cdot \sqrt{40} \doteq 1.33 < 1.3581$.

5.2. **The two-sample case.** The two-sample statistic $D_{m,n}$ defined as $\sup_x |(F_m - G_n)(x)|$ is normalized by multiplying it times $\sqrt{mn/(m+n)}$, as has been mentioned. Here is an explanation of that. Recall that Y is a Bernoulli(p) random variable if $\Pr(Y = 1) = p = 1 - \Pr(Y = 0)$. Such a variable has mean p and variance $p(1-p)$. If $Y_1, ..., Y_n$ are i.i.d. Bernoulli(p), then their sample mean $\overline{Y}$ has mean p and variance $p(1-p)/n$. If X is a random variable with distribution function F, then for each x, $1_{X \le x}$ is Bernoulli(p) with $p = F(x)$. The empirical distribution function $F_n(x)$ is the sample mean of n such variables, so it has mean $F(x)$ and variance $F(x)(1 - F(x))/n$. It follows that $E((F_n - F)(x)^2) = F(x)(1 - F(x))/n$.

We can evaluate, under the hypothesis $G = F$,

$$E((F_m - G_n)(x)^2) = E([(F_m - F)(x) - (G_n - F)(x)]^2) =$$

$$E((F_m - F)(x)^2) + E((G_n - G)(x)^2) = \left(\frac{1}{m} + \frac{1}{n}\right) F(x)(1 - F(x),$$

where the second equality holds because $(F_m - F)(x)$ and $(G_n - G)(x)$ are independent with mean 0, so the expectation of their product is 0. So to get a multiple of $F_m - G_n$ whose mean-square at each x doesn't depend on m or n, we multiply it by $1/\sqrt{m^{-1} + n^{-1}}$, which equals $\sqrt{mn/(m+n)}$.

It follows from (9) that if the two-sample null hypothesis holds, so that F_m and G_n are independent empirical distribution functions from the same F, and if F is continuous, then $\sqrt{\frac{mn}{m+n}} \sup_x |(F_m - G_n)(x)|$ has the same distribution as

$$\sqrt{\frac{mn}{m+n}} \sup_{0 \le t \le 1} |(\mathcal{U}_m - \mathcal{V}_n)(t)|$$

where $\mathcal{U}_m$ and $\mathcal{V}_n$ are independent empirical distribution functions from $U[0,1]$. It will be shown here that as m and n both go to $+\infty$, the displayed random variable converges in distribution to the same limit distribution as in the one-sample case, namely that of the supremum of the absolute value of the Brownian bridge, given by (10). We can write $\mathcal{U}_m - \mathcal{V}_n$ as $(\mathcal{U}_m - \mathcal{U}) - (\mathcal{V}_n - \mathcal{U})$ and then

$$\sqrt{\frac{mn}{m+n}}(\mathcal{U}_m - \mathcal{V}_n) = \sqrt{\frac{n}{m+n}}\alpha_m + \sqrt{\frac{m}{m+n}}\beta_n \tag{12}$$

where $\alpha_m = \sqrt{m}(\mathcal{U}_m - \mathcal{U})$ and $\beta_n = \sqrt{n}(\mathcal{V}_n - \mathcal{U})$ are two independent empirical processes for $U[0,1]$. By the KMT–Bretagnolle–Massart Theorem 8, we can approximate α_m by a Brownian bridge $B_{\alpha,m}$ and β_n by another $B_{\beta,n}$ where we can take these Brownian bridges to be independent because the empirical processes are. If X, Y are

two i.i.d. normal random variables with mean 0 and a, b are two constants with $a^2 + b^2 = 1$, then $aX + bY$ has the same distribution as X or Y. The same holds if X, Y are i.i.d. normal vectors or in fact Gaussian processes with mean 0. For each m and n, taking $a = \sqrt{n/(m+n)}$ and $b = \sqrt{m/(m+n)}$ we see that $a^2 + b^2 = 1$ and so $aB_{\alpha,m} + bB_{\beta,n}$ is also a Brownian bridge, which approximates the two-sample empirical process (12). This shows why the distribution of $\sqrt{mn/(m+n)}\sup_x |(F_m - G_n)(x)|$ under the null hypothesis $F = G$ continuous has a distribution converging to (10). If $F = G$ is not necessarily continuous then we can apply the upper bound in (8) to get (4).

NOTES

The table in Hollander and Wolfe (1999) for the Kolmogorov (one-sample) statistic is based on Table 1 of the paper by Miller (1956). There are various differences. (1) Miller's table has $n = 1, ..., 100$, and H&W's only $1, ..., 40$; (2) The numbers in the body of Miller's table are given to 5 decimal places, rounded to 3 by H&W; (3) Miller does not actually claim to give 0.8 quantiles of D_n, but rather 0.9 quantiles of $\sup_x(F_n - F)(x)$ without absolute values; anyhow, these would seem not very interesting. (4) Also, Miller says that the quantiles are exactly computed for $n \le 20$ (to 7 places, then rounded to 5) and approximated for $n > 20$. He gives a further discussion of approximations.

Some Russian probabilists, including A. N. Kolmogorov, published much of their work in German, in German books or journals, in the late 1920's and early 1930's. Kolmogorov had also on occasion published papers in French. In 1933, whether this was related to earth-shaking events in Germany that year I don't know, both Glivenko and Kolmogorov decided to publish papers in the Italian actuarial journal, edited by F. P. Cantelli. Glivenko's paper proved the "Glivenko–Cantelli" theorem in case F is continuous, and Cantelli (1933), having early knowledge of Glivenko's work, extended the theorem to general F. At the time, I believe there were no specialized probability journals in the world. There were statistics journals in England (*Biometrika*; the British Royal Statistical Society was founded in 1834, and its *Journal* started publishing in 1838) and since 1930 in North America, *Annals of Mathematical Statistics*, but apparently, there were none per se on the continent of Europe.

My expanded proof of the Bretagnolle–Massart (1989) theorem appeared in the lecture notes Dudley (2000) and is now incorporated as §1.4 in Dudley (2014).

REFERENCES

Bretagnolle, Jean, and Massart, Pascal (1989). Hungarian constructions from the nonasymptotic viewpoint. *Ann. Probab.* **17**, 239–256.

*Cantelli, F. P. (1933). Sulla determinazione empirica della leggi di probabilità. *Giorn. Ist. Ital. Attuari* **4**, 421-424.

Dudley, R. M. (2000). *Notes on Empirical Processes. MaPhySto Lecture Notes* **4**, Aarhus, Denmark.

Dudley, R. M. (2002). *Real Analysis and Probability.* Second ed., Cambridge University Press.

Dudley, R. M. (2014). *Uniform Central Limit Theorems*, 2d ed., Cambridge Univ. Press. Draft Chap. 1: math.mit.edu/∼rmd/998/ch1.pdf.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27**, 642-669.

*Glivenko, V. I. (1933). Sulla determinazione empirica della leggi di probabilità. *Giorn. Ist. Ital. Attuari* **4**, 92-99.

Hollander, M., and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, 2d ed., Wiley, New York.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari* **4**, 83-91.

Komlós, J., Major, P., and Tusnády, G. (1975). An approximation of partial sums of independent RV'-s and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111–131.

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Ann. Probability* **18**, 1269-1283.

Miller, L. H. (1956). Table of percentage points of Kolmogorov statistics. *J. Amer. Statist. Assoc.* **51**, 111-121.

Wei, Fan, and Dudley, R. M. (2011). Dvoretzky–Kiefer–Wolfowitz inequalities for the two-sample case (Preprint). Available in MIT DSpace, also at http://arxiv.org/abs/1107.5356v2 [math.ST].

Wei, Fan, and Dudley, R. M. (2012). Two-sample Dvoretzky–Kiefer–Wolfowitz inequalities. *Statist. Probab. Letters* **82**, 636–644.

* I have not seen these references in the original. I learned of them from secondary sources.