THE PROBABILITIES OF MODERATE DEVIATIONS OF U-STATISTICS AND EXCESSIVE DEVIATIONS OF KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS

Thesis for the Degree of Ph. D.

MICHIGAN STATE UNIVERSITY

GERALD MARLOWE FUNK

1971

This is to certify that the

thesis entitled

THE PROBABILITIES OF MODERATE DEVIATIONS OF
U-STATISTICS AND EXCESSIVE DEVIATIONS
OF KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS

presented by

Gerald Marlowe Funk

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Statistics and Probability

Herman Rubin
Major professor

Date May 21, 1971

O-7639

ABSTRACT

THE PROBABILITIES OF MODERATE DEVIATIONS OF U-STATISTICS AND EXCESSIVE DEVIATIONS OF KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS

By

Gerald Marlowe Funk

Let $X_1, X_2, \ldots$ be independently and identically distributed random variables. The U-statistic, $U^{(n)}$ generated by a function symmetric in its k arguments based on $X_1, \ldots, X_n$ was defined by Hoeffding (Ann. Math. Statis (1948)). Under certain moment conditions Rubin and Sethuraman (Sankhyā Ser. A (1965)) obtained order results for probabilities of moderate deviations of these statistics. These results are extended to obtain expressions of the form

$$P(U^{(n)} - EU^{(n)} > ck\sigma(\log n/n)^{1/2}) \sim (2\pi c^2 \log n)^{-1/2} n^{-1/2 c^2}$$

if $c > 0$ and $\sigma > 0$ is the standard deviation of the limiting distribution of $\sqrt{n}\,(U^{(n)} - EU^{(n)})$. Similar results are obtained for Lehmann-generalized U-statistics, and for some functions of U-statistics.

A related problem, that of obtaining asymptotic expansions of probabilities of excessive deviations of the Kolmogorov-Smirnov and Kuiper statistics is considered. Let F denote the c.d.f. of X_i and let F_n be the sample c.d.f. based on $X_1, \ldots, X_n$. Let $D_n^+ = \sup_x (F(x) - F_n(x))$. Suppose F is continuous and let $n\lambda_n^3 = O(1)$ and $n\lambda_n^2 \to \infty$.

Then

Gerald Marlowe Funk

$$P[D_n^+ > \lambda_n] \sim e^{-2n\lambda_n^2} .$$

This result was obtained by Rubin and Sethuraman (Sankhyā Ser. A (1965)).

In this paper a different proof of this result is presented and the problem is solved for larger deviations, i.e., when the condition $n\lambda_n^3 = O(1)$ is relaxed. Similar results are obtained for the two-sided Kolmogorov-Smirnov statistic and the Kuiper statistic.

THE PROBABILITIES OF MODERATE DEVIATIONS OF U-STATISTICS AND EXCESSIVE DEVIATIONS OF KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS

By

Gerald Marlowe Funk

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

1971

G-70279

ACKNOWLEDGMENTS

My most sincere gratitude is extended to Professor Herman Rubin whose enthusiasm and guidance made this undertaking possible. I benefited immeasurably from our many discussions of large sample theory and asymptotic expansions.

TABLE OF CONTENTS

PROBABILITIES OF MODERATE DEVIATIONS OF U-STATISTICS

1.0 Introduction and Summary

The primary result of this chapter is the development of an asymptotic expansion for probabilities of moderate deviations of U-statistics.

Wassily Hoeffding [11] defined a U-statistic based on n independent random variables, $X_1, X_2, \ldots, X_n$, and a real valued function of $m (\leq n)$ arguments, Ψ, as

$$U'_\Psi(X_1, \ldots, X_n) = \left[m! \binom{n}{m}\right]^{-1} \Sigma \, \Psi(X_{i_1}, \ldots, X_{i_m})$$

where the summation is extended over all permutations $(i_1, \ldots, i_m)$ of m different integers, i_j, such that $1 \leq i_j \leq n$.

Without loss of generality, any U-statistic may be written as

$$U_\varphi(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} \varphi(X_{i_1}, \ldots, X_{i_m}) \quad (1.0.1)$$

where φ is a real valued function symmetric in its m arguments and the summation is extended over all distinct sets of m integers $\{i_1, \ldots, i_m\}$ such that $1 \leq i_j \leq n$ and $i_j < i_{j+1}$ for $1 \leq j \leq m-1$. For example, $U'_\Psi(X_1, \ldots, X_n) = U_\varphi(X_1, \ldots, X_n)$ if

$$\varphi(\alpha_1, \alpha_2, \ldots, \alpha_m) = \frac{1}{n!} \Sigma \, \Psi(\alpha_{i_1}, \ldots, \alpha_{i_m})$$

where the summation is extended over all permutations of $\{1, 2, \ldots, m\}$.

Among other results, Hoeffding found that

$$\lim_{n\to\infty} P[U_\varphi(X_1, \ldots, X_n) > m\sigma x/\sqrt{n}] = 1 - \phi(x)$$

provided $E\varphi(X_1, \ldots, X_m) = 0$, $E\left(\varphi(X_1, \ldots, X_m)\right)^2 < \infty$ and $\sigma^2 = E\left(E[\varphi(X_1, \ldots, X_m)|X_1]\right)^2 > 0$. Here

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x_1^2} dx_1.$$

Clearly there was no loss of generality in assuming the first moment of φ to be zero.

If all moments of φ up to order p exist for some $p > x^2 + 2$ then for fixed $x > 0$

$$\lim_{n\to\infty} \frac{\log P\left[U_\varphi(X_1, \ldots, X_n) > m\sigma x \left(\frac{\log n}{n}\right)^{1/2}\right]}{\log n} = -\frac{x^2}{2}.$$

The latter result was obtained by Herman Rubin and Jayaram Sethuraman [17]. They define deviations of the U-statistic from its mean of the form $c\left(\frac{\log n}{n}\right)^{1/2}$ to be moderate deviations.

Our main result is that, with no additional assumptions,

$$\lim_{n\to\infty} \frac{P\left[U_\varphi(X_1, \ldots, X_n) > m\sigma x\left(\frac{\log n}{n}\right)^{1/2}\right]}{1 - \phi(x\sqrt{\log n})} = 1 \qquad (1.0.2)$$

Since the sample mean is a U-statistic, this may be viewed as a generalization of the following theorem due to Herman Rubin and Jayaram Sethuraman [17].

Theorem 1.0 Let $X_1, X_2, \ldots$ be independently and identically distributed real valued random variables such that $EX_1 = 0$, $EX^2 = \sigma^2$ and $E|X|^p < \infty$ for some $p > x^2 + 2$ and $x > 0$. Then

$$\lim_{n\to\infty} \frac{P\left[\frac{1}{n}\sum_{i=1}^{n} X_i > \sigma x\left(\frac{\log n}{n}\right)^{1/2}\right]}{1 - \phi(x\sqrt{\log n})} = 1 \tag{1.0.3}$$

and

$$\lim_{n\to\infty} \frac{P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > \sigma x\left(\frac{\log n}{n}\right)^{1/2}\right]}{2(1 - \phi(x\sqrt{\log n})} = 1 .$$

The theorem remains true if the constant x is replaced by $x + \varepsilon_n$ where $\varepsilon_n = o\left(\frac{1}{\log n}\right)$.

In section 1.1 the proof of (1.0.2) is presented. In section 1.2 some extensions of this result are discussed. However, the problem of extending these results to deviations larger than moderate deviations remains open. In Chapter 2 moderate and excessive deviation results are presented for the Kolmogorov-Smirnov statistics. This problem is formally introduced in section 2.0.

1.1 U-Statistics

Let $X_1, X_2, \ldots$ be a sequence of independently and identically distributed random variables defined on some probability space $(\Omega, \mathfrak{B}, P)$. Let φ be a Borel measurable, real valued function symmetric in its m arguments. The U-statistic generated by φ and the first n random variables is defined as in (1.0.1):

$$U_\varphi^{(n)} = U_\varphi(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{1 \le i_1 < \ldots < i_m \le n} \varphi(X_{i_1}, \ldots, X_{i_m}) \quad (1.1.1)$$

Given this sequence of U-statistics, $\{U_\varphi^{(n)}\}$, we wish to study the rate at which the corresponding sequence of probabilities of moderate deviations from the mean, $\left\{P\left[U_\varphi^{(n)} - EU_\varphi^{(n)} > c\sqrt{\frac{\log n}{n}}\right]\right\}$, tends to zero as n becomes large. Our approach is to exhibit a sequence of real numbers which is asymptotically equivalent to this sequence of probabilities. (Two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$, are defined to be asymptotically equivalent if $\lim_{n\to\infty} a_n/b_n = 1$ in which case we write $a_n \sim b_n$).

Our main result is

Theorem 1.1

Let $\{U_\varphi^{(n)}\}$ be the sequence of U-statistics defined in (1. 1. 1) and suppose that:

$$E|\varphi(X_1, \ldots, X_m)|^p < \infty \quad \text{for some} \quad p > x^2+2 \quad \text{and} \quad x > 0 \,. \quad (1.1.2)$$

$$E\varphi(X_1, \ldots, X_m) = 0 \,. \quad (1.1.3)$$

$$\sigma^2 = E(E[\varphi(X_1, \ldots, X_m | X_1])^2 > 0 \,. \quad (1.1.4)$$

Then

$$P\left[U_\varphi^{(n)} > m\sigma x\left(\frac{\log n}{n}\right)^{1/2}\right] \sim (2\pi x^2 \log n)^{-1/2}\, n^{-\frac{x^2}{2}} \quad (1.1.5)$$

and

$$P\left[|U_\varphi^{(n)}| > m\sigma x\left(\frac{\log n}{n}\right)^{1/2}\right] \sim 2(2\pi x^2 \log n)^{-1/2}\, n^{-\frac{x^2}{2}} \,. \quad (1.1.6)$$

Proof.

It will be shown that there is a real valued function $\varphi_1(\cdot)$ such that the U-statistics generated by φ_1 and $X_1, X_2, \ldots$ have the following properties:

$$P\left[U_{\varphi_1}^{(n)} > x\sigma\left(\frac{\log n}{n}\right)^{1/2} \pm \frac{\varepsilon_n}{(n \log n)^{1/2}}\right] \sim (2\pi x^2 \log n)^{-1/2} n^{-x^2/2} \tag{1.1.7}$$

$$P\left[|U_{\varphi_1}^{(n)}| > x\sigma\left(\frac{\log n}{n}\right)^{1/2} \pm \frac{\varepsilon_n}{(n \log n)^{1/2}}\right] \sim 2(2\pi x^2 \log n)^{-1/2} n^{-x^2/2} \tag{1.1.8}$$

$$P\left[|U_{\varphi}^{(n)} - m U_{\varphi_1}^{(n)}| > \frac{x m \sigma \varepsilon_n}{(n \log n)^{1/2}}\right] = o\left(\frac{n^{-x^2/2}}{\sqrt{\log n}}\right) \tag{1.1.9}$$

where $\{\varepsilon_n\}$ is a positive sequence such that $\varepsilon_n = o(1)$ and $\frac{1}{\log n} = o(\varepsilon_n)$.

Now for any two real valued random variables Y_1 and Y_2 and constants a and $\delta > 0$

$$P[Y_1 + Y_2 > a] \geq P[Y_1 > a+\delta] - P[Y_2 < -\delta]$$

and

$$P[Y_1 + Y_2 > a] \leq P[Y_1 > a-\delta] + P[Y_2 > \delta] .$$

In particular if $Y_1 + Y_2 = U_\varphi^{(n)}$, $Y_2 = U_\varphi - m U_{\varphi_1}$, $a = m x \sigma\left(\frac{\log n}{n}\right)^{1/2}$ and $\delta = (m x \sigma \varepsilon_n)(n \log n)^{-1/2}$ then these inequalities together with (1.1.7) and (1.1.9) will verify (1.1.5). Similarly, the inequalities

$$P[|Y_1 + Y_2| > a] \geq P[|Y_1| > a + \delta] - P[|Y_2| > \delta]$$

and

$$P[|Y_1 + Y_2| > a] \leq P[|Y_1| > a - \delta] + P[|Y_2| > \delta]$$

together with (1. 1. 8) and (1. 1. 9) will verify (1. 1. 6).

Define

$$\begin{aligned} \varphi_1(x_1) &= E[\varphi(X_1', X_2', \ldots, X_m') \mid X_1' = x_1] \\ &= \int \varphi(x_1, x_2, \ldots, x_m) F(dx_2) \ldots F(dx_m) \qquad \text{a. s.} \end{aligned}$$

where $X_1', \ldots, X_m'$ are independently and identically distributed with the common distribution F being that of the X_i's. Then (1. 1. 3) and (1. 1. 4) insure the real valued random variables $\varphi_1(X_1), \varphi_1(X_2), \ldots$ are independently and identically distributed with expectation $E\varphi_1(X_1) = 0$ and variance $E(\varphi_1(X_1))^2 = \sigma^2 > 0$. Finally, (1. 1. 2) and Jensen's inequality yield

$$E|\varphi_1|^p = E(|E[\varphi|X_1]|^p) \leq E\,E[|\varphi|^p|X_1] = E|\varphi|^p < \infty \text{ for some } p > c^2+2.$$

Thus the conditions of Theorem 1. 1 are satisfied for the sequence $\varphi(X_1), \varphi(X_2), \ldots$. Since

$$U_{\varphi_1}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \varphi_1(X_i)$$

and since

$$\lim_{x \to \infty} \left(\frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \right) \Big/ (1 - \phi(x)) = 1$$

the asymptotic equivalences (1. 1. 7) and (1. 1. 8) have been verified.

To show that (1. 1. 9) is also true, $U_{\varphi}^{(n)}$ is decomposed into k U-statistics as follows:

Define $\varphi_m(x_1, \ldots, x_m) = \varphi(x_1, \ldots, x_m)$ and for $1 \leq r < m$ set

$$\varphi_r(x_1, \ldots, x_r) = E[\varphi(X_1', \ldots, X_m') \mid X_i' = x_i \text{ for } 1 \leq i \leq r] \quad (1.1.10a)$$

and for $1 \leq r \leq m$

$$\begin{aligned}\Psi_r&(x_1, \ldots, x_r)\\ &= \varphi_r(x_1, \ldots, x_r) + (-1) \sum_{1 \leq i_1 < \ldots < i_{r-1} \leq r} \varphi_{r-1}(x_{i_1}, x_{i_2}, \ldots, x_{i_{r-1}})\\ &\quad + (-1)^2 \sum_{1 \leq i_1 < \ldots < i_{r-2} \leq r} \varphi_{r-2}(x_{i_1}, \ldots, x_{i_{r-2}}) + \ldots + (-1)^{r-1} \sum_{1 \leq i \leq r} \varphi_1(x_i) .\end{aligned} \quad (1.1.10b)$$

Then $U_\varphi^{(n)} = \sum_{r=1}^{m} \binom{m}{r} U_{\Psi_r}^{(n)}$. In particular $U_\varphi^{(n)} - m U_{\varphi_1}^{(n)} = \sum_{r=2}^{m} \binom{m}{r} U_{\Psi_r}^{(n)}$.

Note this is the same decomposition as that used in [17]. For $m = 1$, $U_\varphi^{(n)} = U_{\varphi_1}^{(n)}$ and (1.1.9) is trivial. Suppose $m > 1$. Evidently,

$$\begin{aligned}P&\left[\left| U_\varphi^{(n)} - m U_{\varphi_1}^{(n)} \right| > \frac{x m \sigma \varepsilon_n}{(n \log n)^{1/2}} \right]\\ &\leq \sum_{r=2}^{m} P\left[\binom{m}{r} \left| U_{\Psi_r}^{(n)} \right| > \frac{m}{m-1} \frac{x \sigma \varepsilon_n}{(n \log n)^{1/2}} \right]\\ &\leq \left(\frac{m-1}{m x \sigma \varepsilon_n} \right)^\nu (n \log n)^{\nu/2} \sum_{r=2}^{m} \binom{m}{r}^\nu E \left| U_{\Psi_r}^{(n)} \right|^\nu \end{aligned} \quad (1.1.11)$$

for $0 < \nu \leq p$.

The latter inequality was obtained using the Markov inequality [15, p. 158].

It is interesting to note that Ψ_r often has zero (marginal) expection. In fact, if $1 \leq i \leq r$ and the x_j are distinct for $1 \leq j \leq r$ then

$$\int \psi_r(x_1, \ldots, x_r)\, F(dx_i) = 0 . \qquad (1.1.12)$$

To verify this fact it suffices to prove

$$\int \psi_r(x_1, \ldots, x_r)\, F(dx_1) = 0 .$$

Referring to the definition of ψ_r (1.1.10b) it is evident that ψ_r may be written as

$$\psi_r(x_1, \ldots, x_r) = \sum_{j=0}^{r-1} (-1)^i \, T^0_{r-j}(x_1, \ldots, x_r)$$

where $T^0_r(x_1, \ldots, x_r) = \varphi_r(x_1, \ldots, x_r) - \varphi_{r-1}(x_2, \ldots, x_r)$ and for $1 \leq j < r-1$

$$T^0_{r-j}(x_1, \ldots, x_r) = \left(\sum_{1=i_1<\ldots<i_{r-j}\leq r} \varphi_{r-j}(x_{i_1}, \ldots, x_{i_{r-j}}) \right)$$

$$-\left(\sum_{2\leq i_1<\ldots<i_{r-j-1}\leq r} \varphi_{r-j-1}(x_{i_1}, \ldots, x_{i_{r-j-1}}) \right)$$

$$T^0_1(x_1, \ldots, x_r) = \varphi_1(x_1) .$$

Clearly

$$\int \varphi_1(x_1)\, F(dx_1) = 0 ,$$

and

$$\int T^0_{r-j}(x_1, \ldots, x_r)\, F(dx_1) = 0 \quad \text{for} \quad 0 \leq j < r-1 \qquad (1.1.13)$$

since

$$\int \varphi_{r-j}(x_{i_1}, \ldots, x_{i_{r-j}}) F(dx_1) = \begin{cases} \varphi_{r-j}(x_{i_1}, \ldots, x_{i_{r-j}}) & \text{if no } x_{i_j} \text{ is } x_1 \\ \varphi_{r-j-1}(x_{i'_1}, \ldots, x_{i'_{r-j-1}}) & \text{if one } x_{i_j} \text{ is } x_1 \end{cases}$$

where $\{x_{i'_1}, \ldots, x_{i'_{r-j-1}}\} = \{x_{i_1}, \ldots, x_{i_{r-j}}\} - \{x_1\}$. This verifies (1.1.13) and hence (1.1.12).

Returning to line (1.1.11), if ν is an even integer then

$$E|U_{\Psi_r}^{(n)}|^{\nu} = \binom{m}{r}^{-\nu}$$

$$= \binom{n}{r}^{-1} E\Sigma \cdots \Sigma \Psi_r(X_{i_{11}}, \ldots, X_{i_{1r}}) \cdots \Psi_r(X_{i_{\nu 1}}, \ldots, X_{i_{\nu r}}) \tag{1.1.14}$$

Of the $\binom{n}{r}^{\nu}$ terms in this summation of products all those terms in which a particular X_i occurs exactly once have zero expectation. Thus a term with non-zero expectation can have no more than $\frac{r\nu}{2}$ distinct X's.

To see that there is a common upper bound to these expectations, observe that repeated applications of Hölder's inequality [15, p. 156] yields

$$E\left|\prod_{j=1}^{\nu} \Psi_r(X_{i_{j1}}, \ldots, X_{i_{jr}})\right| \leq E|\Psi_r(X_1', \ldots, X_r')|^{\nu}$$

provided the moments on the right hand side exist. Referring back to the definition of Ψ_r in line (1.1.10) it is evident that these moments exist provided $E\varphi_j^{\nu} < \infty$ for $1 \leq j \leq r$. Fortunately this is easily shown since

$$\begin{aligned}
E\varphi_j^{\nu} &= E(E[\varphi(X_1', \ldots, X_m') | X_1', \ldots, X_j']^{\nu}) \\
&\leq E(E[\varphi^{\nu}(X_1', \ldots, X_m') | X_1', \ldots, X_j']) \\
&= E\varphi^{\nu} \\
&\leq [E|\varphi|^p]^{\nu/p} < \infty \quad \text{provided } \nu \text{ is an even integer and } \nu \leq p .
\end{aligned}$$

Here the conditional variant of Jensen's inequality [15, p. 159, 348] was used and assumption (1. 1. 2) assures the finiteness of $E|\varphi|^p$. We conclude that each term in (1. 1. 14) is bounded by $2^{m\nu}(E|\varphi|^p)^{\nu/p}$.

Of course only those terms with $\frac{r\nu}{2}$ or fewer distinct X's have nonzero expectation. There are at most

$$c(n) = \sum_{j=r}^{\frac{r\nu}{2}} c_j \binom{n}{j}$$

such terms where c_j is the number of ways j distinct X's may be arranged in the $r\nu$ positions in a term. Apparently $c(n) = O(n^{\frac{r\nu}{2}})$ so that in (1. 1. 14)

$$E|U_{\Psi_r}|^{\nu} = \binom{n}{r}^{-\nu} O\left(n^{\frac{r\nu}{2}}\right) = O\left(n^{-\frac{r\nu}{2}}\right) \quad \text{for } 2 \leq r \leq m .$$

Finally, returning to (1. 1. 11) it is seen that

$$P\left[|U_{\varphi}^{(n)} - mU_{\varphi_1}^{(n)}| > \frac{x m \sigma \varepsilon_n}{(n \log n)^{1/2}}\right] = O\left(\varepsilon_n^{-\nu}(\log n)^{\frac{\nu}{2}} n^{-\frac{\nu}{2}}\right).$$

Let ν be an even integer such that $c^2 < \nu \leq p$ then (1. 1. 9) is verified and the proof is finished.

1.2 Extensions.

Let F denote a distribution function and Ψ a real valued function integrable with respect to the product measure $\prod_{i=1}^{m} F_i$. Define a parameter $\theta(F)$ as

$$\theta(F) = \int \cdots \int \Psi(x_1, x_2, \ldots, x_m) F(dx_1) F(dx_2) \cdots F(dx_m). \quad (1.2.1)$$

Hoeffding [11] calls $\theta(F)$ a regular functional of F. Given $n (\geq m)$ independent random variables, $X_1, X_2, \ldots, X_n$, a reasonable method of estimating $\theta(F)$ is to replace F by the sample distribution function F_n. This leads to a statistic of the form:

$$\theta(F_n) = \left(\frac{1}{n}\right)^m \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \Psi(X_{i_1}, X_{i_2}, \ldots, X_{i_m}). \quad (1.2.2)$$

This statistic is closely related to the U-statistic discussed in the previous section. (Note that Ψ does not depend on n.)

Define

$$\varphi(x_1, \ldots, x_m) = \Sigma\, \Psi(x_{i_1}, \ldots, x_{i_m}) \quad (1.2.3)$$

where the summation is extended over all permutations of $\{1, 2, \ldots, m\}$. For $1 \leq j < m$ define

$$\Psi_{m-j}(x_1, \ldots, x_{m-j}) = \Sigma\, \Psi(x_{i_1}, \ldots, x_{i_m}) \quad (1.2.4)$$

where the summation is over the distinguishable permutations in which each x_i, $1 \leq i \leq m-j$, occurs at least once. Then

$$n^m \theta(F_n) = \binom{n}{m} U_{\varphi}^{(n)} + \binom{n}{m-1} U_{\Psi_{m-1}}^{(n)} + \ldots + \binom{n}{1} U_{\Psi_1}^{(n)} \quad (1.2.5)$$

Hoeffding showed that if $E[\theta(F_n)]^2 < \infty$ then the asymptotic distribution of $\sqrt{n}\,(\theta(F_n) - \theta(F))$ is that of $\sqrt{n}\,(U_\varphi^{(n)} - \theta(F))$.

The following theorem shows, under additional moment conditions, that a similar result is true for moderate deviations. (The notation developed in (1.2.1) through (1.2.4) is used in the statement of the theorem.)

Theorem 1.2.1

Let $\{X_i\}$ be a sequence of independently and identically distributed random variables with common distribution F and let $\{\theta(F_n)\}$ be the corresponding sequence of regular functionals of the sample distribution function (1.2.2). If

$$\int \cdots \int [\varphi(x_1, \ldots, x_m)]^p F(dx_1) \cdots F(dx_m) < \infty \quad \text{for some} \quad p > x^2 + 2 \tag{1.2.6}$$

and for $1 \leq j \leq m-1$,

$$\int \cdots \int [\Psi_j(x_1, \ldots, x_j)]^q F(dx_1) \cdots F(dx_j) < \infty \quad \text{for some} \quad q > 1,\ q > x^2 \tag{1.2.7}$$

and $\sigma^2 > 0$ where

$$\sigma^2 = \int \left[\int \cdots \int \left(\frac{1}{m!}\, \varphi(x_1, \ldots, x_m) - \theta(F) \right) F(dx_1) \cdots F(dx_{m-1}) \right]^2 F(dx_m) \tag{1.2.8}$$

then for $x > 0$

$$P\left[\theta(F_n) - \theta(F) > m x \sigma \left(\frac{\log n}{n} \right)^{1/2} \right] \sim (2\pi x^2 \log n)^{-1/2}\, n^{-x^2/2} \tag{1.2.9}$$

and

$$P\left[|\theta(F_n) - \theta(F)| > m x \sigma \left(\frac{\log n}{n} \right)^{1/2} \right] \sim 2(2\pi x^2 \log n)^{-1/2}\, n^{-x^2/2}. \tag{1.2.10}$$

Proof.

Using the expansion (1.2.5) $\theta(F_n)$ may be written as

$$\theta(F_n) = \left(\prod_{i=1}^{m-1} (1 - \tfrac{i}{n})\right) U^{(n)}_{\varphi/m!} + \sum_{j=1}^{m-1} \frac{1}{n^j} \left(\prod_{i=1}^{m-j-1} (1 - \tfrac{i}{n})\right) U^{(n)}_{\Psi_{m-j}/(m-j)!}$$

Since any weighted sum of finitely many U-statistics is itself a U-statistic (provided the weights are constants independent of n), $\theta(F_n)$ may be written as

$$\theta(F_n) = \sum_{j=0}^{m-1} \frac{1}{n^j} U^{(n)}_{\varphi'_j} \tag{1.2.11}$$

Clearly $\varphi'_0 = \varphi/m!$ and the φ'_j are weighted sums of φ and the Ψ_j. Evidently,

$$\begin{aligned} P &= P\left[\theta(F_n) - \theta(F) > m x \sigma \left(\frac{\log n}{n}\right)^{1/2}\right] \\ &\leq P\left[U^{(n)}_{\varphi'_0} - \theta(F) > m x \sigma \left(\frac{\log n}{n}\right)^{1/2} - \frac{\varepsilon_n}{(n \log n)^{1/2}}\right] \\ &\quad + \sum_{j=1}^{m-1} P\left[U^{(n)}_{\varphi'_j} > \frac{\varepsilon_n n^{j-1/2}}{(m-1)(\log n)^{1/2}}\right] \end{aligned} \tag{1.2.12}$$

and

$$\begin{aligned} P &\geq P\left[U^{(n)}_{\varphi'_0} - \theta(F) > m x \sigma \left(\frac{\log n}{n}\right)^{1/2} + \frac{\varepsilon_n}{(n \log n)^{1/2}}\right] \\ &\quad - \sum_{j=1}^{m-1} P\left[U^{(n)}_{\varphi'_j} < \frac{-\varepsilon_n n^{j-1/2}}{(m-1)(\log n)^{1/2}}\right] \end{aligned} \tag{1.2.13}$$

where ε_n is the positive sequence defined after line (1.1.9). In view of Theorem 1.1, (1.2.6) and (1.2.8) the first term on the right sides of 1.2.12 and 1.2.13 are asymptotically equivalent to the right hand side of (1.2.9). To see that the remaining terms are asymptotically negligible Markov's inequality is used to obtain

$$P\left[|U^{(n)}_{\varphi_j'}| > \delta_n\right] \le \delta_n^{-\nu} E|U^{(n)}_{\varphi_j'}|^{\nu} \tag{1.2.14}$$

and for $\nu > 1$ Minkowski's inequality is repeatedly used to show

$$(E|U_{\varphi_j'}|^{\nu})^{1/\nu} \le (E|_{\varphi_j'}|^{\nu})^{1/\nu} \le K \max_{1\le i\le m-1}\left\{(E|\varphi|^{\nu})^{1/\nu}, (E|\Psi_i|^{\nu})^{1/\nu}\right\}. \tag{1.2.15}$$

Select ν so that the moments on the right hand side of 1.2.15 exist and $\nu > 1$ and $\nu > x^2$. Set

$$\delta_n = \frac{\varepsilon_n n^{+j-1/2}}{(n \log n)^{1/2}} .$$

Then the right hand side of (1.2.14) is asymptotically negligible compared to $(\log n)^{-1/2} n^{-x^2/2}$ for $1 \le j \le m-1$. and (1.2.9) is verified. A similar argument verifies (1.2.10).

The latter part of this proof is a special case of

<u>Lemma 1.2.1</u>

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of real random variables such that

$$P\left[X_n > x\left(\frac{\log n}{n}\right)^{1/2}\right] \sim 1 - \phi(x\sqrt{\log n}) \quad \text{for } x \in (0, x_0).$$

Then

$$P\left[X_n + \frac{Y_n}{n^c} > x\left(\frac{\log n}{n}\right)^{1/2}\right] \sim 1 - \phi(x \log n)$$

if $c > 1/2$ and, for some fixed M, $E|Y_n|^{\nu} \le M < \infty$ for some $\nu > \frac{x^2}{2c-1}$ and all n .

Proof.

Use inequalities analogous to (1.2.12) and (1.2.13) and apply Markov's inequality.

We give one further example of how Theorem 1.1 may be extended. Let $\{X_n^{(j)}\}$, $1 \leq j \leq r$, denote r independent sequences of independently and identically distributed random variables and let F_j denote the common distribution of the random variables in the j^{th} sequence. Let T^0 be a Borel measurable, real valued function of $m = \sum_{j=1}^{r} m_j$ arguments. Let $T^0(\cdots, x_i, \cdots, x_k, \cdots) = T^0(\cdots, x_k, \cdots, x_i, \cdots)$ if $m_j < i < k \leq m_{j+1}$ for some j, $0 \leq j \leq r-1$. Here $m_0 = 0$. Thus T^0 is symmetric in its first m_1 arguments, next m_2 arguments, etc. Define a Lehmann-generalized U-statistic generated by T^0 and $X_1^{(j)}, X_2^{(j)}, \ldots, X_{cj}^{(j)}$, $1 \leq j \leq r$ as:

$$U'^{(n)} = \left(\prod_{j=1}^{r} \binom{c_j n}{m_j} \right)^{-1} \Sigma\, T^0(X_{1_{k_1}}^{(1)}, \ldots, X_{1_{k_{m_1}}}^{(1)}, \ldots, X_{r_{k_1}}^{(r)}, \ldots, X_{r_{k_{m_r}}}^{(r)}) \tag{1.2.15}$$

where the summation is extended over all subscripts for which

$$1 \leq j_{k_1} < \ldots < j_{k_{m_j}} \leq c_j n \; ; \; 1 \leq j \leq r .$$

Here the c_j's (≥ 1) are integral valued and represent the proportion of $X^{(j)}$'s in the "sample."

The sequence of statistics $U'^{(n)}$ may be approximated by the sequence of U-statistics

$$U_\varphi^{(n)} = \binom{n}{m}^{-1} \sum_{1 < i_1 < \ldots < i_m \le n} \varphi(Z_{i_1}, \ldots, Z_{i_m}) \tag{1.2.16}$$

where

$$Z_{i+1} = (X^{(1)}_{c_1 i+1}, \ldots, X^{(1)}_{c_1(i+1)}, \ldots, X^{(r)}_{c_r i+1}, \ldots, X^{(r)}_{c_r(i+1)}) \tag{1.2.17}$$

for $i = 0, 1, 2, \ldots,$; and

$$\varphi(Z_1, \ldots, Z_m) = \left(m! \prod_{j=1}^{r} \frac{c_j}{m_j!}\right)^{-1} \Sigma\, \Sigma T^0(\ldots) \tag{1.2.18}$$

where the summation is over the $\binom{m}{m_1, \ldots, m_r}$ permutations of the subscripts of Z such that $i_{m_{j(j)}+1} < \ldots < i_{m_{(j+1)}}$ for $0 \le j \le r-1$ and over all permuations of the $X^{(j)}$ within each Z.

Theorem 1.2.2 Let $\{U'^{(n)}\}$ be the sequence of statistics defined in (1.2.15). If the sequence of U-statistics, $\{U_\varphi^{(n)}\}$, satisfies the conditions of Theorem 1.1 and if

$$\int \cdots \int [T^0(x_1^{(1)}, \ldots, x_{m_1}^{(1)}, \ldots, x_1^{(r)}, \ldots, x_{m_r}^{(r)})]^\nu \prod_{j=1}^{r} \prod_{i=1}^{m_j} F_j(d x_i^{(j)}) < \infty \tag{1.2.19}$$

for some $\nu > x^2$, $x > 0$, then

$$P\left[U'^{(n)} > m\sigma x\left(\frac{\log n}{n}\right)^{1/2}\right] \sim P\left[U_\varphi^{(n)} > m\sigma x\left(\frac{\log n}{n}\right)^{1/2}\right].$$

Proof.

Note that $\prod_{j=1}^{r}\binom{c_j n}{m_j} U'^{(n)}$ consists of $M_n = \prod_{j=1}^{r}\binom{c_j n}{m_j}$ summands and may be written as

$$M_n U'^{(n)} = \binom{n}{m}\left(m! \prod_{j=1}^{r} \frac{c_j}{m_j!}\right) U_\varphi^{(n)} + M_n T_n$$

where T_n represents the

$$M_n - \binom{n}{m}\left(m! \prod_{j=1}^{r} \frac{c_j}{m_j!}\right) = N_n$$

summands of $U'^{(n)}$ not included in $(N_n - M_n) U_\varphi^{(n)}$. Thus $U^{(n)}$ may be written as

$$U'^{(n)} = U_\varphi^{(n)} + \left(T_n - \frac{N_n}{M_n} U^{(n)}\right).$$

Condition (1.1.2) assures us that $E|U^{(n)}|^\nu < M < \infty$ for some $\nu > x^2+2$ and condition (1.2.19) guarantees $E|T_n|^\nu \leq |\frac{N_n}{M_n}|^\nu E|T^0|^\nu$ and $E|T^0|^\nu < \infty$. Since $\frac{N_n}{M_n} = O(\frac{1}{n})$, Lemma 1.2.1 applies and the theorem is proved.

The theorem remains true if the independence conditions are replaced with the assumption that the Z's are independent (and condition (1.2.19) replaced by the obvious moment conditions) and if $\sum_{j=1}^{r} c_j$ is replaced by $\sum_{j=1}^{r} c_j + o(1)$.

The final theorem in this section deals with functions of U-statistics. Let $\{X_n\}$ denote a sequence of independently and identically distributed random variables and let $U_1^{(n)}, \ldots, U_k^{(n)}$ be k U-statistics generated by $X_1, \ldots, X_n$ and the kernals $\varphi^{(1)}, \varphi^{(2)}, \ldots, \varphi^{(k)}$ respectively. For each U-statistic, assume that conditions (1.1.3) and (1.1.4) of Theorem 1.1 hold and that all moments of $\varphi^{(j)}$ exist and are finite. Define $\alpha_{ij} = E\varphi_1^{(i)} \varphi_1^{(j)}$; $1 \leq i \leq k$, $1 \leq j \leq k$ where $\varphi_1^{(j)}$ is defined as in (1.1.10a). If the determinant $|\alpha_{ij}|$ is positive then

Theorem 7.1 of [11] asserts that the asymptotic joint distribution of the $\sqrt{n}\, U_j^{(n)}$'s is non-singular. Thus each nondegenerate linear combination $\sum_{j=1}^{k} a_j U_j^{(n)}$ is a U-statistic satisfying the conditions of Theorem 1.1. In view of Theorem 8 of [17] we have the following

Theorem 1.2.3 Let $f(x_1, \ldots, x_k)$ be a function of k arguments and

$$f(x_1, \ldots, x_k) = f(0, \ldots, 0) + \sum_{i=1}^{k} b_i x_i + o\left(\frac{||x||}{\log ||x||}\right)$$

where $||x||^2 = \sum_{i=1}^{k} x_i^2$. Then, for the above-mentioned U-statistics, if $x > 0$, and $\sigma_b > 0$,

$$P\left[f(U_1^{(n)}, \ldots, U_k^{(n)}) - f(0, \ldots, 0) > x\,\sigma_b\left(\frac{\log n}{n}\right)^{1/2}\right] \sim (2\pi x^2 \log n)^{-1/2}\, n^{-x^2/2}$$

where

$$\sigma_b^2 = \sum_{j=1}^{k} \sum_{i=1}^{k} b_i m_i b_j m_j \alpha_{ij}$$

and m_j is the number of arguments of $\varphi^{(j)}$ $1 \leq j \leq n$.

Similar results are valid for $\theta(F_n)$ and for Lehmann-generalized U-statistics.

PROBABILITIES OF EXCESSIVE DEVIATIONS OF THE KOLMOGOROV-SMIRNOV AND KUIPER STATISTICS

2.0 Introduction and Summary

Let $X_1, X_2, \ldots, X_n$ be a sequence of one-dimensional, independent identically distributed random variables each having the continuous cumulative distribution function F.

The empirical cumulative distribution function, F_n, associated with the sequence $X_1, \ldots, X_n$ is defined by the relation

$$F_n(x) = \sum_{x_i \leq x} \frac{1}{n} .$$

The Kolmogorov-Smirnov statistics are defined as

$$D_n^+ = \sup_{-\infty < x < +\infty} (F_n(x) - F(x))$$

$$D_n^- = \sup_{-\infty < x < +\infty} (F(x) - F_n(x))$$

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)| .$$

The Kuiper statistic is defined as

$$V_n = \sup_{-\infty < x < \infty} (F_n(x) - F(x)) - \inf_{-\infty < x < \infty} (F_n(x) - F(x))$$

$$= D_n^+ + D_n^- .$$

The limit distributions of these statistics are well known. N. V. Smirnov and A. N. Kolmogorov [10] have proved that, for any constant c greater than 0 ,

$$\lim_{n\to\infty} P(D_n^+ > cn^{-1/2}) = e^{-2c^2}$$

$$\lim_{n\to\infty} P(D_n^- > cn^{-1/2}) = e^{-2c^2}$$

$$\lim_{n\to\infty} P(D_n > cn^{-1/2}) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2c^2} ,$$

and Nicolass Kuiper [14, p. 43] has shown that

$$\lim_{n\to\infty} P(V_n > cn^{-1/2}) = 2 \sum_{j=1}^{\infty} (4j^2c^2 - 1) e^{-2j^2c^2} .$$

Second order terms of these asymptotic expansions are also known. [10, 14].

Following the terminology of Herman Rubin and Jayaram Sethuraman [17], deviations of these statistics from zero of the form $cn^{-1/2}$ for constant c will be called ordinary deviations. Any deviation of the form $c_n n^{-1/2}$, for $\{c_n\}$ a sequence unbounded above, will be called excessive. If $\{c_n\}$ is also an increasing sequence it is apparent that

$$\lim_{n\to\infty} P(D_n^+ > c_n n^{-1/2}) = 0 .$$

Of interest is the rate at which probabilities of excessive deviations tend to zero as the sample size becomes large. Much work has been done in this area.

Herman Rubin and Jayaram Sethuraman [17] made use of a theorem by N. V. Smirnov [10, p. 154] to obtain the following excessive deviation result. Let $c_n \to \infty$ and $c_n^3 n^{-1/2} = O(1)$, then

$$P[D_n^+ > c_n n^{-1/2}] \sim e^{-2c_n^2}$$
$$P[D_n > c_n n^{-1/2}] \sim 2e^{-2c_n^2}.$$

In section 2.3 it is shown that asymptotic expansions similar to these remain valid if the condition $c_n n^{-1/6} = O(1)$ is replaced by $c_n = o(n^{1/2})$.

In section 2.4 similar results are obtained for the Kuiper statistic; e.g., for $c_n > c\sqrt{\log n}$ and $c_n n^{-1/6} = O(1)$, $c > 1/2$

$$P[V_n > c_n n^{-1/2}] \sim 8c_n^2 e^{-2c_n^2}.$$

However, for larger deviations the results are fundamentally different, that is, they are not a "natural" extension of the ordinary deviation results. A constant deviation will be called large. Jayaram Sethuraman [18] proved that for any constant c between zero and one

$$\frac{1}{n} \log P[D_n > c] \sim \log \beta(c)$$

$$\beta(c) = \sup_{0<x<1-c} (x/(x+c))^{x+c} ((1-x)/(1-x-c))^{1-x-c}$$

In fact Sethuraman obtained results similar to this for k-dimensional random variables and variables defined on separable complete metric spaces.

In section 2.5 it is shown that

$$P[D_n^+ > c] \sim \alpha(c)(\beta(c))^n ;$$

where $\alpha(c)$ does not depend on n. Asymptotic expansions are also given for D_n and D_n^-.

Order results for large deviations of the Kuiper statistic have also been worked out. Innis G. Abrahamson [1] has proved that

$$\frac{1}{n} \log P[V_n \geq c] \sim \log \beta(c).$$

In section 2.6 it is proved that $P[V_n \geq c]$ is asymptotically equal to $\alpha_1(c) n(\beta(c))^n$.

Section 2.1 contains a brief review of some well known facts about F_n, D_n and V_n and the relationship between the order statistics of uniform random variables and Poisson processes.

Section 2.2 contains some theorems of a technical nature which are used in the later sections.

2.1 The Sample Distribution Function

In the following paragraphs some well known properties of the sample distribution function are reviewed. No new results are presented in this section.

2.1.1. Let $X_1, X_2, \ldots, X_n$ be n independent random variables with common continuous cumulative distribution function F. The statistics D_n^+, D_n^-, D_n and V_n are independent of F.

For example, set $Y_i = F(X_i)$ then the Y_i are independent uniform random variables on the interval $[0, 1]$. Since

$$P\left[\sum_{X_i \le x} \frac{1}{n} = \sum_{Y_i \le F(x)} \frac{1}{n}\right] = 1 ,$$

the statistics D_n, D_n^+, D_n^- and V_n generated by $X_1, \ldots, X_n$ are, with probability one, the same as those generated by $Y_1, \ldots, Y_n$. In the following sections it will be assumed that the X_i themselves are uniform random variables.

2.1.2 The equation $F_n(x) - x = D_n^+$ can only be valid if x is one of the observed values of the (uniform) random variables $X_1, \ldots, X_n$. From continuity arguments, it is evident that $F_n(x) - x$ takes, with probability one, its maximal value at a unique point, X_{max}. The random variable X_{max} is uniform on the unit interval. In the following sections it will be assumed that the maximum deviation of $F_n(x) - x$ is unique. (Similar remarks apply to the infimum of $F_n(x-) - x$. [13, 2]).

2.1.3 The probability distribution of the order statistics, $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, of n independent uniform random variables on the interval (s, t) and that of the jump points, $T_1 \le T_2 \le \ldots \le T_n$ of a Poisson process, $X(v)$, is the same given that $X(s) = m$ and $X(t) = m + n$ [6 p. 400 and 12, p. 239]. By a Poisson process with parameter λ is meant a separable, real process $X(t)$ with stationary independent integral valued increments and

$$P[X(t) - X(s) = m] = \frac{\lambda^m (t-s)^m e^{-\lambda(t-s)}}{m!}$$

for $\lambda > 0$; $t > s \ge 0$, $X(0) = 0$, $m = 0, 1, 2 \ldots$

2.2 Some Asymptotic Expansions

The idea of obtaining asymptotic results for Kolmogorov-Smirnov statistics by a consideration of $n F_n(t)$ as a Poisson process, $X(nt)$, is not new (see D. A. Darling [20] and R. Pyke [16]).

This idea is employed here to obtain asymptotic expansions for the conditional probabilities of the events

$$\{F_n(t) - t \leq \frac{m}{n} - x \quad \text{for all } t \text{ such that } \quad 0 < t < x\}$$

and

$$\{F_n(t) - t \leq \frac{m}{n} - x \quad \text{for all } t \text{ such that } \quad x < t < 1\}$$

given that the m^{th} order statistic occurs at x.

For the case $x < t < 1$ use is made of the following theorem ([12], p. 247-48).

<u>Theorem 2.2.1</u>

Let $T_1, \ldots, T_m$ be the order statistics based on m independent uniform random variables on the interval $[s, t]$ and let

$$F_{n,m}(y) = \sum_{T_i \leq y} \frac{1}{n} .$$

Then

$$\begin{aligned} P[\sup_{s \leq x \leq t} \left(F_{n,m}(y) - (y-s)\right) \leq 0] &= P[T_k \geq \frac{k}{n} + s \,;\, k = 1, 2, \ldots, m] \\ &= 1 - \frac{m}{n(t-s)} \quad \text{if } \quad 0 < m < n(t-s) \\ &= 0 \quad \text{if } \quad m > n(t-s) . \end{aligned}$$

In particular,

$$P\Big[\sup_{x<t<1}\Big(F_n(t) - t\Big) \leq \frac{m}{n} - x \,\Big|\, X_{(m)} = x\Big] = 1 - \frac{n-m}{n(1-x)} \qquad (2.2.1)$$

For the case $0 < t < x$ the following theorem, due to Lajos Takács ([19], p. 56). is used.

Theorem 2.2.2

If $\{X(t)\,;\, 0 \leq t < \infty\}$ is a separable stochastic process with stationary independent increments for which almost all sample functions are nondecreasing step functions vanishing at $t = 0$ then for $x \geq 0$

$$P\Big[\sup_{0 \leq t < \infty} (t - X(t)) \leq x\Big] = 1 - e^{-wx},$$

where

$$E[e^{-zX(t)}] = e^{-t\,\phi(z)} \quad \text{for} \quad t \geq 0,\ \operatorname{Re} z \geq 0,$$

and w is the largest real root of the equation $\phi(z) = z$.

This theorem was proved for Poisson processes by R. Pyke [16]. If $X(t)$ is Poisson with parameter λ then

$$\phi(z) = \lambda(1 - e^{-z}) \qquad (2.2.2)$$

and $w = \lambda - \lambda^*$ provided $\lambda > 1$, $\lambda^* < 1$ and $\lambda e^{-\lambda} = \lambda^* e^{-\lambda^*}$. Evidently as λ tends to one, w tends to zero.

Lemma 2.2.1

If $\lambda_n = 1 + a_n$; $a_n > 0$, $a_n = o(1)$ and w_n is the largest real root of the equation $w = \lambda_n(1 - e^{-w})$ then $w_n \sim 2a_n$.

Proof.

Noting the remarks following Theorem 2.2.2, $w_n = \lambda_n - \lambda_n^*$, where $\lambda_n e^{-\lambda_n} = \lambda_n^* e^{-\lambda_n^*}$ and since the function $g(\lambda) = \lambda e^{-\lambda}$ is continuous and for $\lambda > 0$ has a unique maximum at $\lambda = 1$ it is evident that $\lambda_n \downarrow 1$ implies $\lambda_n^* \uparrow 1$. An application of the mean value theorem yields:

$$e^{-a} = 1 - a + \frac{a^2}{2} - O(a^3)\,;\ a > 0\,.$$

Thus

$$(1+a)e^{-a} = 1 - a^2/2 + O(a^3)\,,\quad a > 0$$

$$(1-b)e^{b} = 1 - b^2/2 - O(b^3)\,,\quad b > 0\,.$$

Let $\lambda_n = 1 + a_n$; $\lambda_n^* = 1 - b_n$ so that $\lambda_n e^{-\lambda_n} = \lambda_n^* e^{-\lambda_n^*}$, that is $(1+a_n)e^{-a_n} = (1-b_n)e^{b_n}$ which means that $b_n^2 + O(b_n^3) = a_n^2 - O(a_n^3)$ and since $a_n = o(1)$, $b_n = o(1)$ we have $b_n^2(1+o(1)) = a_n^2(1-o(1))$. Thus $b_n \sim a_n$ and $w_n = a_n + b_n \sim 2a_n$.

In fact $w_n = 2a_n(1 + O(a_n))$.

Given that the m^{th} order statistic of n independent uniform random variables on the unit interval occurs at x, the first $m-1$ order statistics, $X_{(1)}, \ldots, X_{(m-1)}$, have the same distribution as the order statistics of $m-1$ independent uniform random variables on the interval $[0, x]$. Also $U_{(i)} = n(x - X_{(m-i)})$; $i = 1, 2, \ldots, m-1$ would be the order statistics of $m-1$ uniform random variables on the interval $[0, nx]$. Evidently

$$P[\sup_{0<t<x}\left(F_n(t) - t\right) \leq \frac{m}{n} - x \,|\, X_{(m)} = x]$$

$$= P[\frac{k}{n} - X_{(k)} \leq \frac{m}{n} - x\,;\ k = 1, 2, \ldots, m-1 \,|\, X_{(m)} = x]$$

$$= P[-(m-1)\,F^u_{m-1}(t) + t \leq 1\,;\ 0 \leq t \leq nx \,|\, X(0) = 0,\ X(nx) = m-1]$$

$$= L_n(x, m) \quad , \text{ say.} \tag{2.2.3}$$

Here $F^u_{m-1}(t)$ is the sample distribution function based on the $\{U_i\}$ and $X(u)$ is a Poisson process with jump points $\{U_i\}$ on $(0,nx)$ [See 2.1.3].

The next theorem shows that, asymptotically, the condition $X(nx) = m - 1$ in (2.2.3) may be dropped for suitable choice of m and λ.

<u>Theorem 2.2.3</u>

Let $\{X(u), u \geq 0\}$ be a Poisson process and, for each $n > 0$, let $\{X_n(u), u \geq 0\}$ be a Poisson process with parameter $\lambda_n = 1 + \frac{c\,d_n}{t}$ where $0 < c < \infty$, $0 < t < 1$ and d_n is a positive sequence such that $d_n = o(1)$ and $n d_n^2 > a \log n$ for some $a > 0$. Let $k_n = [n(t + c\,d_n)]$. Define

$$L_n(t, k_n) = P[u - X(u) \leq 1\,;\ 0 < u < nt \,|\, X(nt) = k_n - 1].$$

Then for any t_0 and t_1 such that $0 < t_0 < t < t_1 < 1$ and sequence c_n such that $c_n > c$ and $c_n d_n = o(1)$,

$$L_n(t, k_n) \sim P[u - X_n(u) \leq 1]$$

and

$$L_n(t, k_n) = \frac{2c\,d_n}{t}\,[1 + a(n, t, c)] \tag{2.2.4}$$

where

$$\lim_{n\to\infty} \sup_{A} a(n,t,c) = 0 \tag{2.2.5}$$

and the supremum is taken over all t and c such that $0 < t_0 < t < t_1 < 1$ and $1 \leq c \leq c_n$.

Proof.

Suppose $\{X(u) ; u > 0\}$ has parameter λ. The probability of the sample functions for which $u - X(u) \leq 1$ for all $u > 0$ is given in Theorem 2.2.2. This probability may be expressed as a sum by conditioning on the events $\{X(nt) = k - 1\}$ for $k \geq nt$.

For reasons of notational simplicity, define

$$Q(\lambda) = P[u - X(u) \leq 1 ; 0 < u < \infty] , \tag{2.2.6}$$

$$Q_n(\lambda, k) = P[u - X(u) \leq 1 ; u > nt | X(nt) = k - 1] \tag{2.2.7}$$

and

$$P_n(\lambda, k) = P[X(nt) = k - 1] . \tag{2.2.8}$$

Since $X(\cdot)$ has independent increments, $Q(\lambda)$ may be written as

$$Q(\lambda) = \sum_{k \geq nt} L_n(t,k) P_n(\lambda,k) Q_n(\lambda,k) . \tag{2.2.9}$$

and since $X(\cdot)$ has stationary increments,

$$Q_n(\lambda, k) = P[u - X(u) \leq k - nt ; u \geq 0 | X(0) = 0]$$

In view of Theorem 2.2.2,

$$Q(\lambda) = 1 - e^{-w} \tag{2.2.10}$$

and

$$Q_n(\lambda, k) = 1 - e^{-(k - nt)w} \quad \text{if} \quad k \geq nt . \qquad (2.2.11)$$

Apparently $Q_n(\lambda, k)$ and $L_n(t, k)$ are nondecreasing functions of k and we are led to the following inequalities:

$$Q(\lambda) \geq \left(\sum_{j \geq k} P_n(\lambda, j) \right) L_n(t, k) Q_n(\lambda, k) \quad \text{if} \quad k \geq nt \qquad (2.2.12)$$

and

$$Q(\lambda) \leq L_n(t, k) + \left(\sum_{j \geq k} P_n(\lambda, j) \right) L_n(t, k') + \sum_{j \geq k'} P_n(\lambda, j)$$

$$\text{if} \quad k' > k > nt . \quad (2.2.13)$$

If k is large, if $\sum_{j > k} P_n(\lambda, j)$ is near one and if $\sum_{j > k'} P_n(\lambda, j)$ is near zero then $L_n(t, k)$ could be approximated by $Q(\lambda)$. Of course $L_n(t, k)$ is independent of λ so the problem reduces to that of selecting λ so as to make the bounds in (2.2.12) and (2.2.13) reasonably tight when $k = k_n$. An intuitively appealing idea is to "aim" the sample functions of the process $\{X(u)\}$ at the point (nt, k_n), i.e., to select λ so that $EX(nt) \doteq k_n$.

Proceeding with this idea, let

$$nt\lambda'_n = [n(t + cd_n)] + [n\varepsilon_n] \qquad (2.2.14)$$

where $\{\varepsilon_n\}$ is a positive sequence such that $\varepsilon_n = o(d_n)$, $(n\varepsilon_n^2)^{-1} = o(1)$ and $n\varepsilon_n^3 = o(1)$. Now if $\{X_i\}$ is a sequence of independently distributed Poisson random variables with parameter 1 then

$$P_n[\lambda'_n, k] = P\left[\sum_{i=1}^{nt\lambda'_n} X_i = k - 1 \right].$$

Cramér's theorem for probabilities of excessive deviations of a sum of independent random variables from its mean [8, p. 517] applies so that

$$\sum_{j<k_n} P_n[\lambda_n', j] \sim \frac{1}{\sqrt{2\pi n \varepsilon_n^2}} \, e^{-\frac{n\varepsilon_n^2}{2}} . \tag{2.2.15}$$

Returning to (2.2.12) with (2.2.10), (2.2.11) and (2.2.15) it is now evident that

$$L_n(t, k_n) \leq \left(1 - e^{-w}\right)\left(1 - e^{-(k_n - nt)w}\right)^{-1}\left(1 + \frac{2}{\sqrt{2\pi n \varepsilon_n^2}} \, e^{-\frac{n\varepsilon_n^2}{2}}\right) \tag{2.2.16}$$

for all $n > N_\varepsilon$ where N_ε is a constant determined by $\{\varepsilon_n\}$ alone, and not on c or t (since $\lambda_n' > n(t_0+d_n)$ for all t and c satisfying the restrictions after (2.2.5)).

Making use of Lemma 2.2.1 it is seen that

$$\left(1 - e^{-w}\right) \leq \left(\frac{2cd_n}{t} + \frac{\varepsilon_n}{t_0}\right)\left(1 + a\, c_n d_n\right)$$

for some constant a and that

$$\left(1 - e^{-(k_n - nt)w}\right)^{-1} \leq \left(1 + 2e^{-\frac{nd_n^2}{t_1}}\right)$$

for n sufficiently large. Thus

$$L_n(t, k_n) \leq \frac{2cd_n}{t}\left(1 + a'(n, t, c)\right) \tag{2.2.17}$$

where $\limsup\limits_{n\to\infty} \underset{A}{} a'(n, t, c) = 0$.

In order to obtain a lower bound for $L_n(t, k_n)$ return to (2.2.13) with $n t \lambda''_n = [n(t + c d_n)] - [n \varepsilon_n]$ and $k'_n = k_n + (n \log n)^{1/2}$ Then, using Lemma 2.2.1, line (2.2.17) and Cramér's Theorem it is seen that

$$Q(\lambda''_n) \geq \left(\frac{2 c d_n}{t} + \frac{\varepsilon_n}{t_1}\right)\left(1 - a c_n d_n\right),$$

$$L_n(t, k'_n) \leq \frac{2}{t_0}\left(c d_n + \frac{\log n}{n}\right)^{1/2},$$

$$\sum_{j \geq k_n} P_n(\lambda''_n, j) \leq e^{-\frac{n \varepsilon_n^2}{2}},$$

and

$$\sum_{j \geq k'_n} P_n(\lambda''_n, j) \leq (n \log n)^{-1/2}$$

for n sufficiently large and some constant a . Thus

$$L_n(t, k_n) \geq \frac{2 c d_n}{t}\left(1 + a''(n, t, c)\right) \tag{2.2.18}$$

where $\limsup_{n \to \infty} \sup_A a''(n, t, c) = 0$. The Theorem is proved.

The use of Cramér's theorem to obtain asymptotic expansions for $\sum_{k \geq k_n} P_n(\lambda'_n, k_n)$ and $\sum_{k \geq k'_n} P_n(\lambda''_n, k)$ was not essential. An alternative approach is to note that if $p(\lambda, k) = \lambda^k e^{-\lambda}/k!$ then repeated integration by parts verifies [7, p. 163]

$$\sum_{k=0} p(\lambda, k) = \frac{1}{n!}\int_{\lambda}^{\infty} e^{-x} x^n \, dx .$$

Of course $n!$ can be evaluated using Stirling's formula and asymptotic expansions for the above integral, an incomplete gamma function, have been worked out by W. Fulks [9].

The following lemma presents an asymptotic expansion for binomial probabilities [7, p. 169].

Lemma 2.2.2

Let $\{d_n\}$ be a positive sequence such that $d_n = o(1)$ and $n d_n^2 > a_0 \log n$ for some $a > 0$. For each $t \in (0,1)$ let $\{c_n(t)\}$ be a sequence such that $c_n(t) \geq 1$, $m = m_n(c,t) = n(t + c_n(t)d_n)$ is integral valued and for some $\varepsilon > 0$, $c_n(t)d_n < 1 - t - \varepsilon$. Define

$$P_n(t,m) = \binom{n}{m} t^m (1-t)^{n-m} \tag{2.2.19}$$

and

$$f_n(t,m) = (2\pi n(t + c_n(t)\, d_n)(1 - t - c_n(t)\, d_n)^{-1/2} .$$

$$\left(1 + \frac{c_n(t)\, d_n}{t}\right)^{-m} \left(1 - \frac{c_n(t)\, d_n}{1-t}\right)^{-(n-m)} \tag{2.2.20}$$

Then

$$P_n(t,m) = f_n(t,m)\left(1 \pm O(\frac{1}{nt})\right) \tag{2.2.21}$$

Proof.

A refined version of Stirlings formula [7, p. 52] states that

$$(2\pi)^{1/2} n^{n+1/2} e^{-n} e^{\frac{1}{12n+1}} \leq n! \leq (2\pi)^{1/2} n^{n+1/2} e^{-n} e^{\frac{1}{12n}}$$

Thus

$$f_n(t, m)\exp\left(\frac{1}{12n+1} - \frac{1}{12m} - \frac{1}{12(n-m)}\right)$$

$$\leq P_n(t, m)$$

$$\leq f_n(t, m)\exp\left(\frac{1}{12n} - \frac{1}{12m+1} - \frac{1}{12(n-m)+1}\right)$$

and the result (2.2.21) follows after slight manipulation.

Lemma 2.2.3.

Let K have a binomial distribution with parameters n and t. Let $m = n(t + c_n)$, $0 < m < n$ $0 < t < 1$, $c_n > 0$. Then

$$P\left[K \geq m\right] \leq \left[\left(1 - \frac{c_n}{1-t}\right)^{(1-t-c_n)} \left(1 + \frac{c_n}{t}\right)^{(t+c_n)}\right]^{-n}.$$

Proof [3].

For $x > 1$, $K \geq 0$, x^K is an increasing function so that

$$P[K \geq m] \leq \min_{s>1} s^{-m}(ts + 1-t)^n$$

where $E s^K = (ts + 1-t)^n$. Differentiate the right hand side with respect to s to find the minimizing value is $s_0 = \left(\frac{m}{n-m}\right)\left(\frac{1-t}{t}\right) > 1$.

2.3. Probabilities of Excessive Deviations of the Kolmogorov-Smirnov Statistics

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent uniform random variables on the unit interval and let $F_n(x)$ denote the sample distribution function. Let $p(t, \frac{m}{n} - t)$ denote the probability "density" that the (unique) maximum of $F_n(s)-s$, for $s\varepsilon[0,1]$, is attained at t and is equal to $\frac{m}{n} - t$. Thus

$$\int_0^1 \sum_{m=1}^{n} p(t, \frac{m}{n} - t)dt = 1 . \tag{2.3.1}$$

If $p_{n,m}(t)$ denotes the probability density of the m-th order statistic, $X_{(m)}$, then

$$p(t, \frac{m}{n} - t) = p_{n,m}(t)P[F_n(s)-s \leq \frac{m}{n} - t;\ 0\leq s\leq 1 | X_{(m)}=t] \tag{2.3.2}$$

and we have the following.

<u>Lemma 2.3.1.</u>

For $p(t, \frac{m}{n} - t)$ defined as above and $1\leq m\leq n$

$$p(t, \frac{m}{n} - t) = m\binom{n}{m}t^{m-1}(1-t)^{n-m}(1-\frac{n-m}{n(1-t)})L_n(t,m) . \tag{2.3.3}$$

Proof:

The probability density of $X_{(m)}$ is [12]

$$p_{n,m}(t) = m\binom{n}{m}t^{m-1}(1-t)^{n-m} \tag{2.3.4}$$

In view of theorem 2.2.1 ,

$$P[F_n(s)-s \leq \frac{m}{n} - t; t \leq s \leq 1 | X_{(m)}=t] = 1 - \frac{n-m}{n(1-t)} \tag{2.3.5}$$

Since $L_n(t,m)$ is defined in (2.2.3) as

$$P[u-X_{(u)} \leq 1; 0<u<nt | X_{(nt)}=m-1] = P[F_n(s)-s \leq \frac{m}{n} - t; 0<s<t | X_{(m)}=t] \tag{2.3.6}$$

and since the conditional probabilities in (2.3.5) and (2.3.6) are of conditionally independent events, the lemma follows.

Evidently,

$$P[D_n^+ \geq d_n] = \int_0^1 \sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t)\,dt. \tag{2.3.7}$$

An asymptotic expansion for the integral on the right (in 2.3.7) is obtained (for suitable sequences $\{d_n\}$) by first finding an asymptotic expansion for $p(t, \frac{m}{n} - t)$ and then one for $\sum\limits_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t)$ for fixed t and finally carrying out the required integration using the method of Laplace.

Let $\{d_n\}$ and $\{c_n(t)\}$ be sequences as defined (and restricted) in Lemma 2.2.2 and $m=m_n(c,t)=n(t+c_n(t)d_n)$. The following Lemma presents an asympotic expansion for $p(t, \frac{m}{n} - t)$.

<u>Lemma 2.3.2</u>

Let $p(t, \frac{m}{n} - t)$ be the "density" defined prior to (2.3.1).

Let $m=m_n(c,t)$ and $c=c_n(t)$.

Define

$$d_n^2 \, g_n(c,t) = -\log(1+\frac{cd_n}{t})^{n(t+cd_n)}(1 - \frac{cd_n}{1-t})^{n(1-t-cd_n)} \tag{2.3.8}$$

and

$$b_n(c,t) = \frac{2c^2}{t(1-t)} [2\pi(t+cd_n)(1-t-cd_n)]^{-1/2} . \tag{2.3.9}$$

Then for any t_0 and t_1 such that $0 < t_0 < t < t_1 < 1$ and sequence c_n such that $c_n > c$ and $c_n d_n = o(1)$

$$p(t, \frac{m}{n} - t) = \sqrt{n}\, d_n^2 b_n(c,t) e^{nd_n^2 g_n(c,t)} (1+a(n,c,t)) \tag{2.3.10}$$

where $\lim_{n\to\infty} \sup_A a(n,t,c)=0$ and the supremum is over all t ; $0 < t_0 < t < t_1 < 1$ and over all c ; $1 \le c \le c_n$.

Proof:

The proof is achieved by using expressions (2.2.4) and (2.2.19-2.2.21) in expression (2.3.3).

Next on the agenda is the task of evaluating

$$\sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t)$$

for fixed t and sequence $\{d_n\}$.

For each $k=0,1,2,\ldots$ define the sequence $\{c_n^k(t)\}$ so that

$$\begin{aligned} n(t+c_n^k(t)d_n) &= [n(t+d_n)]+k \quad \text{if} \quad [n(t+d_n)]<n(t+d_n) \\ &= [n(t+d_n)]+k-1 \quad \text{if} \quad [n(t+d_n)]=n(t+d_n). \end{aligned}$$

Let

$$m_n^*(k,t) = n(t+c_n^k(t)d_n). \tag{2.3.11}$$

Then

$$\sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t) = \sum_{k \geq 1} p(t, \frac{m_n^*(k,t)}{n} - t).$$

It is interesting to observe that, when nd_n^2 is large, the dominant term of (2.3.10) is a decreasing function of c .

In particular, if a_n is such that $a>1$ and

$$n(t+a_n d_n) = [n(t+ad_n)] ,$$

the upper and lower Darboux sums of

$$nd_n \int_{c_n'(t)}^{a_n} b_n(c,t) e^{nd_n^2 g_n(c,t)} dc$$

obtained by breaking $[c'(t), a_n]$ into intervals of width $1/nd_n$ are

$$\sum_{k=1}^{n(a_n-1)d_n-1} b_n(c_m^k(t),t) e^{nd_n^2 g_n(c_n^k(t),t)} \tag{2.3.13}$$

and

$$\sum_{k=2}^{n(a_n-1)d_n} b_n(c_n^k(t),t) e^{nd_n^2 g_n(c_n^k(t),t)} \tag{2.3.14}$$

respectively.

This leads to

Lemma 2.3.3.

Let $p(t, \frac{m}{n} - t)$ be the "density" defined prior to (2.3.1) and $\{d_n\}$ be defined as in Lemma 2.2.2. then

$$\sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t) = \alpha_n(t) e^{nd_n^2 g_n(1,t)} (1+a(n,t)) \tag{2.3.15}$$

where

$$\alpha_n(t) = 2(nd_n^2)^{1/2} (2\pi t(1-t))^{-1/2} , \tag{2.3.16}$$

$g_n(1,t)$ is defined in (2.3.8)

and

$$\lim_{n\to\infty} \sup_{t_0<t<t_1} a(n,t) = 0 .$$

Proof.

For some $a>1$ let $k_n=[(a-1)nd_n]$ and write the left hand side of (2.3.15) as

$$\sum_{k=1}^{k_n} p(t, \frac{m_n^*(k,t)}{n} - t) + \sum_{k\geq k_n+1} p(t, \frac{m_n^*(k,t)}{n} - t) \qquad (2.3.17)$$

$$= Q_1+Q_2 \text{ say .}$$

It will be shown that Q_1 is asymptotically equivalent to the right hand side of 2.3.15 and that Q_2 is asymptotically negligible when "a" is a sufficiently large real number.

Making use of (2.3.10) and (2.3.12) through (2.3.14), Q_1 may be expressed as

$$(\sqrt{n}d_n^2)^{-1} Q_1 = (nd_n \int_1^a b_n(c,t)e^{nd_n^2 g_n(c,t)} dc)(1+a^*(n,t)) \qquad (2.3.18)$$

$$+ \theta_n(b_n(C_n^0(t),t)e^{nd_n^2 g_n(C_n^0(t),t)}$$

where

$$|\theta_n| \leq 2 \text{ and } \lim_{n\to\infty} \sup_{t_0<t<t_1} a^*(n,t) = 0$$

For large n , each integral exhibited in (2.3.10) is determined by its behavior near one since $g_n(c,t)$ is a decreasing function of $c\geq1$. To see this, recall the definition of $g_n(c,t)$ in (2.3.8):

$$d_n^2 g_n(c,t) = -((t+cd_n)\log(1+\frac{cd_n}{t})+(1-t-cd_n)\log(1-\frac{cd_n}{1-t})). \quad (2.3.19)$$

Differentiate with respect to c to obtain

$$\frac{\partial}{\partial c}(d_n^2 g_n(c,t)) = -d_n(\log(1+\frac{cd_n}{t})-\log(1-\frac{cd_n}{1-t})) \quad (2.3.20)$$

$$= -d_n(\log \frac{1-t}{t} \cdot \frac{t+cd_n}{1-t-cd_n}) \ .$$

Clearly $\frac{\partial}{\partial c} d_n^2 g_n(c,t) < 0$ since $1-t/t$ is a decreasing f unction of t .

Differentiate again with respect to c to obtain

$$\frac{\partial^2}{\partial c^2} d_n^2 g_n(c,t) = \frac{-d_n^2}{(t+cd_n)(1-t-cd_n)} < 0 \ . \quad (2.3.21)$$

The mean value theorem is used to write

$$nd_n^2 g_n(c,t) = nd_n^2 g_n(1,t)+(c-1)nd_n^2 (\frac{\partial}{\partial c} g_n(\zeta,t))$$

for some $1 \leq \zeta \leq c$.

Since for large n, $b_n(c,t)$ is an increasing function of c it is now evident that for large n

$$nd_n\int_1^a b_n(c,t)e^{nd_n^2 g_n(c,t)} dc \leq \frac{nd_n b_n(a,t)e^{nd_n^2 g_n(1,t)}}{nd_n^2 g_n'(1,t)} \int_0^{nd_n^2(a-1)g_n'(a,t)} e^u du \ .$$

$$(2.3.22)$$

and

$$\int_1^a b_n(c,t)e^{nd_n^2 g_n(c,t)} dc \geq \frac{nd_n b_n(1,t)e^{nd_n^2 g_n(1,t)}}{nd_n^2 g_n'(a,t)} \int_0^{nd_n^2(a-1)g_n'(a,t)} e^u du \ . \quad (2.3.23)$$

(Here $g_n'(a,t) = \frac{\partial}{\partial c} g_n(c,t))$.

In particular if we let $a_n = 1+\delta_n$ where $\delta_n > 0$, $\delta_n = o(1)$ and $\frac{1}{nd_n^2} = o(\delta_n)$

then

$$nd_n \int_1^{a_n} b_n(c,t)e^{nd_n^2 g_n(c,t)}\,dc = \frac{b_n(1,t)e^{nd_n^2 g_n(1,t)}}{d_n|g_n'(1,t)|}(1+o(1)). \quad (2.3.24)$$

To verify (2.3.24) note that

$$nd_n^2\delta_n g'(a_n,t) \le nd_n^2\delta_n g'(1,t)$$

and

$$nd_n^2\delta_n g'(1,t) = -nd_n\delta_n(\log(1+\frac{d_n}{t})-\log(1-\frac{d_n}{1-t})) \qquad (2.3.25)$$

$$\le \frac{-nd_n^2\delta_n}{2t(1-t)}$$

provided

$$d_n < \min\left\{\frac{t_0(1-t_0)}{2}, \frac{t_1(1-t_1)}{2}\right\}.$$

The latter inequality is obtained by noting that

$$f(x) = \log(\frac{1-t}{t}) - \log\frac{1-t-x}{t+x} - \frac{x}{2t(1-t)}$$

is such that $f(0) = 0$ and f is increasing if x is small.

Since $\lim_{n\to\infty} nd_n^2\delta_n = \infty$ we conclude that the integrals on the right hand side of (2.3.22) and (2.3.23) converge (uniformly for $t_0<t<t_1$) to -1 as n becomes large. Of course b_n and g_n' are continuous and (2.3.24) follows.

Using (2.3.22) and an inequality analogous to 2.3.26 we have

$$\int_{a_n}^{a} b_n(c,t)e^{nd_n^2 g_n(c,t)}dc \Big/ \int_{1}^{a_n} b_n(c,t)e^{nd_n^2 g_n(c,t)}dc$$

$$\leq \frac{b_n(a,t)}{b_n(1,t)} e^{nd_n^2(a-1-2\delta_n)g_n'(a_n,t)} \tag{2.3.26}$$

In view of (2.3.24), (2.3.25) and (2.3.26),

$$nd_n\int_1^a b_n(c,t)e^{nd_n^2 g_n(c,t)}dc = \frac{b_n(1,t)e^{nd_n^2 g_n(1,t)}}{d_n|g_n'(1,t)|}[1+a(n,t)] \tag{2.3.27}$$

for some $a(n,t)$ such that

$$\lim_{n\to\infty}\ \sup_{t_0<t<t_1} a(n,t) = 0 .$$

Note that

$$\frac{\sqrt{n}\ d_n^2 b_n(1,t)}{d_n|g_n'(1,t)|} \sim \alpha_n(t) .$$

To complete the proof we need to show that Q_2 is small.

Returning to (2.3.3) it is seen that

$$p(t, \tfrac{m}{n} - t) \leq \frac{m}{t} \binom{n}{m} t^m (1-t)^{n-m}$$

and, using (2.3.8) (2.3.11), (2.3.17) and Lemma 2.2.3, that

$$Q_2 = \sum_{m\geq n(t+ad_n)+1} p(t, \tfrac{m}{n} - t) < \frac{n}{t} e^{nd_n^2 g_n(a,t)} .$$

Then, by (2.3.20), (2.3.21), the remark following it and 2.3.25, we have

$$Q_2 \le \frac{n}{t} e^{nd_n^2 g_n(1,t)} \left(\frac{1-t}{t} \cdot \frac{t+d_n}{1-t-d_n}\right)^{-(a-1)nd_n}$$

$$\le \frac{n}{t} \left(e^{\frac{-(a-1)nd_n^2}{2t(1-t)}}\right) e^{nd_n^2 g_n(1,t)} \tag{2.3.28}$$

Since $nd_n^2 > a_0 \log n$ for some $a_0 > 0$, the constant a in (2.3.27) may be selected so large as to make Q_2 asympotically negligible compared to Q_1 for all $t_0<t<t_1$ (e.g. $(a-1) \ge t(1-t)/a_0$) .

The equivalences (2.3.17), (2.3.18), (2.3.27) and inequality (2.3.28) verify (2.3.15).

Recalling that the point t at which $F_n(s)-s$; $s\varepsilon(0,1)$ attains its maximum has a uniform distribution over the unit interval, line (2,3.15) in Lemma 2.3.3 may be interpreted as an expression for the conditional probability that

$$\sup_{0<s<1} \left(F_n(s)-s\right) \ge d_n$$

given that the supremum is attained at the point t , $t_0<t<t_1$. The following lemma presents an expression for the unconditional probability that

$$\sup_{0<s<1} \left(F_n(s)-s\right) \ge d_n \ .$$

Lemma 2.3.4.

Let $\{d_n\}$ be a positive sequence as defined in Lemma 2.2.2, and $p(t, \frac{m}{n} - t)$ be the "density" defined prior to (2.3.1) then

$$\int_0^1 \sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t)dt \sim e^{nd_n^2 g_n(1,t_n^*)} \tag{2.3.29}$$

where $g_n(1,t)$ is defined in (2.3.8) and t_n^* is such that

$$\sup_{0<t<1} g_n(1,t) = g_n(1,t_n^*) \tag{2.3.30}$$

Proof.

The method of Laplace ([4], [20]) will be employed to show for $0<t_0<1/2<t_1<1$, that

$$\int_{t_o}^{t_1} \alpha_n(t)e^{nd_n^2 g_n(1,t)} dt \sim e^{nd_n^2 g_n(1,t_n^*)} \tag{2.3.31}$$

and

$$\int_0^{t_0} + \int_{t_1}^1 \sum_{m \geq n(t+d_n)} p(t, \frac{m}{n} - t)dt = o(e^{nd_n^2 g_n(1,t_n^*)}) \tag{2.3.32}$$

To verify (2.3.31), $g_n(1,t)$ is twice differentiated with respect to t :

$$d_n^2 f_n(t) = d_n^2 g_n(1,t) = (t+d_n)\log \frac{t}{t+d_n} + (1-t-d_n)\log(\frac{1-t}{1-t-d_n}) \tag{2.3.33}$$

$$d_n^2 f_n'(t) = \frac{\partial}{\partial t} d_n^2 g_n(1,t) = \frac{d_n}{t(1-t)} - \log(\frac{1-t}{t} \cdot \frac{t+d_n}{1-t-d_n}) \tag{2.3.34}$$

$$d_n^2 f_n''(t) = \frac{\partial^2}{\partial t^2} d_n^2 g_n(1,t) = -d_n^2 \left[\frac{1}{t^2(t+d_n)} + \frac{1}{(1-t)^2(1-t-d_n)}\right] < 0 \tag{2.3.35}$$

By using Taylor expansions of the log-terms in the expression

$$d_n^2 f_n'(t) = \frac{d_n}{t(1-t)} - \log(1+\frac{d_n}{t}) + \log(1-\frac{d_n}{1-t})$$

it is seen that t_n^*, the (unique) maximum of $g_n(1,t)$ is in the interval $(\frac{1}{2}(1-d_n), 1/2)$ for all n sufficiently large. Using the mean value theorem, $g_n(1,t)$ may be expressed as

$$g_n(1,t) = g_n(1,t_n^*) + \frac{1}{2}(t-t_n^*)^2 f_n''(\zeta_n) \qquad (2.3.36)$$

for some

$$\zeta_n \in \{[t_n^*, t_n] \cup [t, t_n^*]\}$$

In evaluating the integral in (2.3.31) the following notation is used:

$$h_1(t) = t^{-2}(t+d_n)^{-1} ,$$

$$h_2(t) = (1-t)^{-2}(1-t-d_n)^{-1} ,$$

$$t_{1,n} = t_n^* + \varepsilon_n ,$$

$$t_{0,n} = t_n^* - \varepsilon_n ,$$

where

$$\varepsilon_n = o(1) \quad \text{and} \quad \varepsilon_n^{-2} = o(nd_n^2) ,$$

$$u_n = (nd_n^2(h_1(t_{1,n})+h_2(t_{0,n})))^{1/2} ,$$

$$v_n = (nd_n^2(h_1(t_{0,n})+h_2(t_{1,n})))^{1/2} ,$$

and for some $\varepsilon>0$, $t_0<1/2-\varepsilon$ and $t_1>1/2+\varepsilon$.

In view of (2.33) through (2.3.36)

$$\int_{t_{0,n}}^{t_{1,n}} \alpha_n(t) e^{nd_n^2 f_n(t)} dt \le \frac{\alpha_n(t_n') e^{nd_n^2 f_n(t_n^*)}}{u_n} \int_{u_n(t_{0,n}-t_n^*)}^{u_n(t_{1,n}-t_n^*)} e^{\frac{-x^2}{2}} dx$$

$$\int_{t_{0,n}}^{t_{1,n}} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt \geq \frac{\alpha_n(t_n')e^{nd_n^2 f_n(t_n^*)}}{v_n} \int_{v_n(t_{0,n}-t_n^*)}^{v_n(t_{1,n}-t_n^*)} e^{\frac{-x^2}{2}}\,dx$$

for some $t_n' \in [t_{0,n}, t_{1,n}]$ and $t_n'' \in [t_{0,n}, t_{1,n}]$.

Because $t_n^* \sim 1/2$ it is seen that

$$\frac{\alpha_n(t_n')}{u_n} \sim \frac{\alpha_n(t_n'')}{v_n} \sim (2\pi)^{-1/2}$$

so that

$$\int_{t_{0,n}}^{t_{1,n}} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt \sim e^{nd_n^2 g_n(1,t_n^*)} . \tag{2.3.37}$$

Also, referring to the definition of $\alpha_n(t)$, (2.3.16), it is evident that

$$\int_{t_0}^{t_{0,n}} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt \leq 2(2\pi t_0(1-t_0))^{-1/2}(nd_n^2)^{1/2}\int_{t_0}^{t_{0,n}} e^{nd_n^2 f_n(t)}\,dt. \tag{2.3.38}$$

Here, $f_n(t)$ may be expanded as

$$f_n(t) = f_n(t_{0,n}) + f_n'(\zeta)(t-t_{0,n}) \text{ for } t_0<t\leq \zeta \leq t_{0,n} .$$

Of course $t-t_{0,n} < 0$ and f_n' is a decreasing function so that

$$f_n(t) \leq f_n(t_{0,n}) + (t-t_{0,n})f_n'(t_{0,n})$$

$$\leq f_n(t_{0,n}) + 15\,\varepsilon_n(t-t_{0,n}) \tag{2.3.39}$$

for n sufficiently large.

The latter inequality is obtained by expanding $f_n'(t_0,n)$ as a Taylor series about t_n^* with the assumption that $d_n = o(1)$ and $t_n = o(1)$.

Since $nd_n^2\,\varepsilon_n$ tends to infinity for large n

$$\int_{t_0}^{t_{0,n}} e^{15nd_n^2\varepsilon_n(t-t_{0,n})}\,dt = \frac{1}{15nd_n^2\varepsilon_n}\int_{15nd_n^2\varepsilon_n(t_0-t_{0,n})}^{0} e^{-u}\,du \sim (15nd_n^2\varepsilon_n)^{-1}. \tag{2.3.40}$$

Finally $f_n(t_{0,n}) < f_n(t_n^*)$ so that lines (2.3.38) through (2.3.40) guarantee

$$\int_{t_0}^{t_{0,n}} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt = o(e^{nd_n^2 f_n(t_n^*)}) \tag{2.3.41}$$

provided $nd_n^2\varepsilon_n^2$ tends to infinity for large n .

A similar argument shows that

$$\int_{t_{1,n}}^{t_1} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt \leq 2(2\pi t_1(1-t_1))^{-1/2} nd_n^2 \int_{t_{1,n}}^{t_1} e^{nd_n^2 f_n(t)}\,dt\ ,$$

and

$$f_n(t) \leq f_n(t_{1,n}) - 15\varepsilon_n(t-t_{0,n})$$

so that

$$\int_{t_{1,n}}^{t_1} \alpha_n(t)e^{nd_n^2 f_n(t)}\,dt = o(e^{nd_n^2 f_n(t_n^*)})\ . \tag{2.3.42}$$

the asymptotic expansion (2.3.37), (2.3.41) and (2.3.42) verify (2.3.31). To verify (2.3.32) recall Lemma 2.2.3 and Lemma 2.3.1 to see that

$$p(t, \tfrac{m}{n} - t) \leq n\binom{n-1}{m-1}t^{m-1}(1-t)^{n-m}$$

for $0 \leq m \leq n$

$$\sum_{m \geq n(t+d_n)} p(t, \tfrac{m}{n} - t)$$

$$\leq n\left(\frac{1-t}{1-t-d_n}\right)^{n(1-t-d_n)} \left(\frac{t}{t+d_n}\right)^{n(t+d_n)} \left(1-\frac{1}{n}\right)^{n-1} \left(1-\frac{1}{n(1-t-d_n)}\right)^{-n(1-t-d_n)} \left(\frac{1-t-d_n-\frac{1}{n}}{1-t}\right).$$

Note that $(1 - \frac{1}{x})^x < e^{-1}$ and $(1 - \frac{1}{x})^{-x} < (1 - \frac{1}{x})^{-1}e$ for $x > 0$ so that

$$\sum_{m \geq n(t+d_n)} p(t, \tfrac{m}{n} - t) \leq n\left(1-\frac{1}{n}\right)^{-1}\left(\frac{1-t-d_n}{1-t}\right)e^{nd_n^2 f_n(t)}$$

where $f_n(t)$ is defined in line (2.3.33).

Thus there is a constant $c > 0$ such that

$$\int_0^{t_0} \sum_{m \geq n(t+d_n)} p(t, \tfrac{m}{n} - t) \leq cn\int_0^{t_0} e^{nd_n^2 f_n(t)}\,dt\ . \tag{2.3.43}$$

As in (2.3.39), $f_n(t)$ may be bounded above by

$$f_n(t) \leq f_n(t_0) + (t-t_0)f_n'(t_0)$$

for $0 < t \leq t_0$. Thus the right hand side of (2.3.43) is bounded above by

$$\frac{cn}{nd_n^2 f_n'(t_0)} e^{nd_n^2 f_n(t_0)} \tag{2.3.44}$$

which is asymptotically negligible when compared to

$$e^{nd_n^2 f_n(t_n^*)} .$$

This may be easily demonstrated by means of the following crude inequalities. For $t < t_n^*$

$$f_n'(t) = (t-t_n^*)f_n''(\zeta_n) \geq (\tfrac{1}{2} - t)$$

and

$$f_n(t) = f_n(2t)-tf_n'(2t) + \frac{t^2}{2}f_n''(\zeta)$$

$$\leq f_n(2t) - \frac{1}{16t}$$

provided $2t < t_n^*$.

Thus the expression in (2.3.44) is less than

$$\frac{ce^{-\frac{nd_n^2}{16t_0}}}{(\frac{1}{2} - t_0)d_n^2} e^{nd_n^2 f_n(2t_0)} \tag{2.3.45}$$

provided $2t_0 < t_n^*$.

Then since $f_n(2t_0) < f_n(t_n^*)$ and $nd_n^2 > a_0 \log n$ selecting $t_0 < a_0/16$ also will assure that expression (2.3.45) and hence expression (2.3.43) is asympotically negligible compared to (2.3.37).

A similar argument is used to show

$$\int_{t_1}^{1} \sum_{m \geq n(t+d_n)} p(t, \tfrac{m}{n} - t)dt = o(e^{nd_n^2 f_n(t_n^*)}). \tag{2.3.46}$$

Recalling Lemma 2.3.1 and Lemma 2.2.3 it is seen that

$$p(t, \tfrac{m}{n} - t) \le \tfrac{m}{t}\tbinom{n}{m} t^m (1-t)^{n-m}$$

and

$$\sum_{m \ge n(t+d_n)} p(t, \tfrac{m}{n} - t) \le \frac{n(t+d_n)}{t} e^{nd_n^2 f_n(t)} \quad \text{if } t < 1-d_n$$

and is zero if $t > 1-d_n$.

Thus

$$\int_{t_1}^{1} \sum_{m \ge n(t+d_n)} p(t, \tfrac{m}{n} - t)dt \le \quad cn\ e^{nd_n^2 f_n(t_1)}$$

and (2.3.46) is verified.

The results of this section are summarized as:

<u>Theorem 2.3.1</u>

If $\{X_n\}$ is a sequence of independent, identically distributed random variables with continuous distribution functions, F , and $D_n^+ = \sup_x F_n(x) - F(x)$ is the Kolmogorov-Smirnov statistic generated by $X_1,\ldots,X_n$ then for any real sequence $\{d_n\}$ such that $d_n > 0$, $d_n = o(1)$ and $nd_n^2 > a_0 \log n$ for some $a_0 > 0$

$$P[D_n^+ \ge d_n] \sim e^{nd_n^2 f_n(t_n^*)}$$

where

$$-d_n^2 f_n(t) = (t+d_n)\log(1+ \frac{d_n}{t})+(1-t-d_n)\log(1- \frac{d_n}{1-t})$$

and t_n^* is the unique value at which $f_n(t)$ attains its maximum for $0 < t < 1$.

Corollary 2.3.1

For the sequences $\{X_n\}$ and $\{d_n\}$ defined in the previous theorem

$$P[D_n^- > d_n] \sim e^{nd_n^2 f_n(t_n^*)}$$

$$P[D_n > d_n] \sim 2e^{nd_n^2 f_n(t_n^*)}$$

2.4. Probabilities of Excessive Deviations of the Kuiper Statistic

For a sequence of random variables, each with distribution function F, the Kuiper statistic is defined in terms of the sample distribution function, F_n, as

$$V_n = \sup_X (F_n(x) - F(x)) - \inf_x (F_n(x) - F(x)).$$

This statistic was originally suggested to test hypotheses about distributions on a circle. It has the property that V_n is the same no matter where on the circle the count to determine $F_n(x)$ begins. To make this statement more precise, define addition on the unit interval by

$$\begin{aligned} s \dotplus t &= s + t \quad \text{if} \quad s + t \leq 1 \\ & s + t - 1 \quad \text{if} \quad s + t > 1 \end{aligned}$$

and define the "interval" $[s, s \dotplus t]$ by $\{x : s \leq x \leq s + t\}$ if $s + t < 1$ and $\{x : s \leq x \leq 1 \text{ or } 0 \leq x \leq s + t - 1\}$ if $s + t > 1$.

Let $X_1, X_2, \ldots, X_n$ be independent uniform random variables on the unit circle. Define $X_n^s(t)$ as the number of X_i in the interval $(s, s \dotplus t]$ and $n F_n^s(t) = X_n^s(t)$. Then

$$\begin{aligned} V_n &= \sup_{0 \leq x \leq 1} (F_n(x) - x) - \inf_{0 \leq x \leq 1} (F_n(x) - x) \\ &= \sup_{0 \leq x \leq 1} (F_n^s(x) - x) - \inf_{0 \leq x \leq 1} (F_n^s(x) - x) \end{aligned}$$

Let $D_n(v) = F_n(v) - v$. If the infimum of $D_n(v)$ occurs at s and the maximum at $s \dotplus t$ and the m^{th} order statistic in the interval $(s, s \dotplus 1)$

occurs at $s+t$ then $V_n = \frac{m}{n} - t$. Let $p_n(t, \frac{m}{n} - t)$ be the joint "density" of the above event.

Then

$$\int_0^1 \int_0^1 \sum_{m=1}^{n-1} p_n(t, \frac{m}{n} - t)\, ds \;\; dt = 1 .$$

Using the notation developed in (2.6.3), $p_n(t, \frac{m}{n} - t)$ may be written as

$$L(t, n, m) \; P(t, n, m) \; R(t, n, m).$$

Let A be the event that an order statistic occurs at s and the m^{th} order statistic in the interval $(s, s+1)$ occurs at $s+t$. The joint density associated with this event is

$$P(t, n, m) = \frac{n!}{(m-1)!\,(n-m-1)!} \, t^{m-1} (1-t)^{n-m-1}$$

for $0 < s < 1$, $0 < t < 1$, and $0 < m < n$. Then $p_n(t, \frac{m}{n} - t)$ may be written as

$$p_n(t, \frac{m}{n} - t) = L(t, n, m) \; P(t, n, m) \; R(t, n, m) .$$

Here, $L(t, n, m)$ and $R(t, n, m)$ are defined by

$$L(t, n, m) = P[F_n(s) - s \leq F_n(u) - u \leq F_n(s+t) - s+t \, ; \, u \in (s, s+t) \,|\, A]$$

$$R(t, n, m) = P[F_n(s) - s \leq F_n(u) - u \leq F_n(s+t) - s+t \, ; \, u \in (s+t, s+1) \,|\, A]$$

The next three lemmas present asymptotic expansions for $P(t, n, m)$, $L(t, n, m)$ and $R(t, n, m)$. Since the proofs are similar to those of

Section 2.3, most of the details are omitted.

Lemma 2.4.1

Let $\{d_n\}$ be a sequence such that $d_n = o(1)$ and $nd_n^2 > c\log n$ for some $c > 0$. Then if $m = [n(t+d_n)]$

$$P(t, n, m) \sim [2\pi t(1-t)]^{-1/2} n^{3/2} e^{nd_n^2 g_n(t)} .$$

Here, $g_n(t)$ is defined by

$$-d_n^2 g_n(t) = (t+d_n)\log(1+\frac{d_n}{t}) + (1-t-d_n)\log(1-\frac{d_n}{1-t}) .$$

The proof involves using Stirling's formula in a manner similar to that used in Lemma 2.3.2.

Lemma 2.4.2

If $\{d_n\}$ is a sequence of real numbers such that $d_n = o(1)$ and $nd_n^2 > c\log n$ for some $c^2 \geq t/2$ then for $m = [n(t+d_n)]$, $0 < t < 1$

$$L(t, n, m) \sim (\frac{2d_n}{t})^2 .$$

Proof.

The proof consists of breaking $L(t, n, m)$ into parts by conditioning on $F_n(\frac{s+t}{2})$. First note that $P[X_n^s(t/2) = j|A]$ follows the binomial distribution with parameters $1/2$ and $m-1$;

$$P[X_n^s(t/2) = j|A] = \binom{m-1}{j}(1/2)^{m-1} \quad \text{for} \quad 0 \leq j \leq m-1 .$$

Let

$$L_1(t,n,m,j) = P[D_n(s) \leq D_n(v) \leq D_n(s+t)\ ;\ v \in [s, s+t/2] \mid X_n^s(t/2) = j, A]$$

$$L_2(t,n,m,j) = P[D_n(s) \leq D_n(v) \leq D_n(s+t)\ ;\ v \in [s+t/2, s+t] \mid X_n^s(t/2) = j, A].$$

Then

$$L(t,n,m) = (1/2)^{m-1} \sum_{j=0}^{m-1} \binom{m-1}{j} L_1(t,n,m,j)\, L_2(t,n,m,j)\ .$$

As in sections 2.2 and 2.3 the relationship between uniform order statistics and a Poisson process, $X(\cdot)$, can be exploited.

$$P[D_n(s) \leq D_n(v)\ ;\ v \in [s, s+t/2] \mid X_n^s(t/2) = j, A]$$

$$= P[u - X(u) \leq 1\ ;\ 0 < u < nt/2 \mid X(nt/2) = j]$$

$$= L_n(t/2, j+1).$$

If $j = n/2(t + d_n + o(d_n))$ then Theorem 2.2.3 yields

$$L_n(t/2, j+1) \sim 2\, d_n/t\, .$$

Clearly,

$$L_1(t,n,m,j) = P[D_n(s) \leq D_n(v)\ ;\ v \in [s,\ s+t/2] \mid X_n^s(t/2) = j, A]$$

$$- P[D_n(s) \leq D_n(v) \text{ and } D_n(v') > D_n(s+t)\ ;\ \text{for all } v \in [s, s+t/2]$$

$$\text{and some } v' \in [s, s+t/2] \mid X_n^s(t/2) = j, A].$$

It will be shown that for $j = n/2(t + d_n + o(d_n))$ that

$$P[D_n(v') > D_n(s+t) \text{ for some } v' \in [s, s+t] \mid X_n^s(t/2) = j, A] = o(d_n)$$

so that

$$L_1(t,n,m,j) \sim 2d_n/t.$$

To evaluate $P[D_n(v) > D_n(s+t)$ for some $v \in (s, s+t/2) | X_n^s(t/2) = j, A]$ observe that, conditionally, there are $j-1$ order statistics in the interval $(s, s+t/2)$ so the conditional probability density that the k^{th} order statistic occurs at $s+v$ is the probability density that the k^{th} of $j-1$ uniform order statistics on the interval $(0,t/2)$ occurs at v;

$$(t/2)^{-1} k\binom{j-1}{k}\left(\frac{v}{t/2}\right)^{k-1}\left(\frac{t/2-v}{t/2}\right)^{j-1-k} \quad 0 \le v \le t/2 .$$

The density with respect to $u = (t/2)^{-1}v$ is

$$k\binom{j-1}{k}(u)^{k-1}(1-u)^{j-1-k} \quad 0 \le u \le 1.$$

The conditional probability that $D_n(x)$ does not exceed $k/n - v$ to the "right" of v is (see Theorem 2.2.1)

$$\begin{aligned} P[F_{n,j-k}(x) - x \le 0,\ 0 < x < t/2] &= 1 - \frac{j-k}{\frac{nt}{2}(1-u)} \quad \text{if } 0 < j-k < n(t/2-v) \\ &= 0 \quad \text{if } j-k \ge n(t/2-v). \end{aligned}$$

The conditional probability that $D_n(x)$ does not exceed $k/n - v$ to the "left" of v is

$$P[F_n(x) - x \le k/n - v \quad 0 < x < v | X_{(k)} = v] = L_n(v,k).$$

Thus the probability density with respect to u and k that the supremum of $D_n(x)$ for $0 < x < t/2$ occurs at v and is equal to $k/n - v$, given $X_n^s(t/2) = j$ and A is

$$k\binom{j}{k}(u)^{k-1}(1-u)^{j-k}\left(1-\frac{j-k}{\frac{nt}{2}(1-u)}\right)L_n(v,k).$$

An asymptotic expansion for this density can be obtained by methods similar to those used to obtain the expansion in Lemma (2.3.2). Let $\varepsilon_n = o(d_n)$ and $j = n/2(t+d_n+\varepsilon_n)$. We require $k/n - v = d_n + t$ so that $k = n(v + c_e d_n)$ for $c_e \geq 1$. Let

$$k_n = j\left[u + \frac{(2c_e - u)d_n - u\varepsilon_n}{(t+d_n+\varepsilon_n)}\right].$$

Since $k_n = ju + c\theta j d_n/t$ for some θ, $1 \leq \theta \leq 2$; $c \geq 1$, we may obtain, using Lemma 2.3.4 that

$$\begin{aligned} &P[D_n(v') > D_n(s+t) \text{ for some } v' \in [s, s+t] \mid X_n^s(t/2) = j, A] \\ &= O\left(e^{j(d_n^2/t_n^2)g_n(1,u_n^*)}\right) \\ &= O\left(e^{-2jd_n^2/t^2}\right) \\ &= O\left(n^{-c^2/t}\right). \end{aligned}$$

Similarly, $L_2(t,n,m,j) \sim 2d_n/t$. This may be verified by observing that

$$\begin{aligned} &P[D_n(v) \leq D_n(s+t)\,;\, v \in [s+t/2, s\ \ t] \mid X_n^s(t/2) = j, A] \\ &= P[u - X(u) \leq 1\quad 0 < u < nt/2 \mid X(nt/2) = m-j] \\ &= L_n(t/2\,,\, m-j+1) \end{aligned}$$

and then proceding with the proof that $L_1(t,n,m,j) \sim 2d_n/t$.

Apparently

$$P[D_n(s) \le D_n(v) \le D_n(s+t)\ v \in [s, s+t] \,|\, X_n^s(t/2)^2 j, A] = 0$$

unless $nt/2 < j < n(\frac{t}{2} + d_n)$. In this case $|\varepsilon_n| < d_n$. For $\varepsilon_n' = \theta d_n$, $|\theta| < 1$ we have

$$L_1(t,n,m,j)\, L_2(t,n,m,j) = O\left[\left(\frac{2(1+\theta)\, d_n}{t}\right)\left(\frac{2(1-\theta)\, d_n}{t}\right)\right]$$

Recall that

$$L(t,n,m) = \left(\frac{1}{2}\right)^{m-1} \sum_{j=0}^{m-1} \binom{m-1}{j} L_1(t,n,m,j)\, L_2(t,n,m,j).$$

Break this summation into three parts:

(a) $j = \frac{n}{2}(t + d_n + o(d_n))$

(b) $j = \frac{n}{2}(t + (1+\theta)\, d_n)$ $\quad |\theta| < 1$

(c) other j

For part (a) the sum is asymptotically $\left(\frac{2\,d_n}{t}\right)^2 \left(1 - o(1)\right)$ and parts (b) and (c) are negligible compared to this. |

Corollary 2.4.1

If $\{d_n\}$ is a sequence of real numbers such that $d_n = o(1)$ and $n d_n^2 > c \log n$ for some $c > 0$ then for $m = [n(t+d_n]$, $0 < t < 1$ $L(t,n,m) = O\{(\frac{d_n}{t})\}$.

Lemma 2.4.3

If $\{d_n\}$ is a sequence of real numbers such that $d_n = o(1)$

and $nd_n^2 > c \log n$ for some $c^2 \geq \frac{(1-t)}{2}$ then for $m = [n(t+d_n)]$, $0 < t < 1$

$$R(t,n,m) \sim \left(\frac{d_n}{1-t}\right)^2 .$$

The proof is similar to that of the previous lemma. Note that $P[X_n^{s+t}(\frac{1-t}{2}) = j \mid A]$ follows a binomial distribution with parameters $1/2$ and $n-m-1$;

$$P[X_n^{s+t}(\tfrac{1-t}{2}) = j \mid A] = \binom{m-n-1}{j}\left(\frac{1}{2}\right)^{n-m-1}, \quad 0 \leq j \leq n-m-1 .$$

Let

$$R_1(t,n,m,j)$$
$$= P[D_n(s) \leq D_n(v) \leq D_n(s+t)\,;\, v \in [s+t, s+t+\tfrac{1-t}{2}] \mid A \text{ and } X_n^{s+t}(\tfrac{1-t}{2}) = j]$$

$$R_2(t,n,m,j)$$
$$= P[D_n(s) \leq D_n(v) \leq D_n(s+t)\,;\, v \in [s+t+\tfrac{1-t}{2}, s+1 \mid A \text{ and } X_n^{s+t}(\tfrac{1-t}{2}) = j]$$

Then

$$R(t,n,m) = \sum_{j=0}^{n-m-1} \binom{n-m-1}{j}\left(\frac{1}{2}\right)^{n-m-1} R_1(t,n,m,j)\, R_2)t,n,m,j) .$$

It will now be shown that for $j = \frac{n}{2}(1-t-d_n + o(d_n))$

$$R_1(t,n,m,j) \sim \left(\frac{d_n}{1-t}\right)$$

Note that

$R_1(t, n, m, j)$

$$= P[D_n(v) \leq D_n(s+t)\,;\, v \in [s+t, s+t+\tfrac{1-t}{2}] \,|\, A \text{ and } X^{s+t}(\tfrac{1-t}{2}) = j]$$

$$- P[D_n(v) \leq D_n(s+t)\,;\, v \in [s+t, s+t+\tfrac{1-t}{2}]$$

and

$$D_n(v') < D_n(s) \quad \text{for some} \quad v' \in [s+t, s+t+\tfrac{1-t}{2}] \,|\, A \text{ and } X^{s+t}(\tfrac{1-t}{2}) = j]$$

and for any Poisson process $X(u)$ and $j = \frac{n}{2}(1-t - d_n + \varepsilon_n)$

$$P[D_n(v) \leq D_n(s+t)\,;\, v \in [s+t, s+t+\tfrac{1-t}{2}] \,|\, A \text{ and } X^{s+t}(\tfrac{1-t}{2}) = j]$$

$$= P[X(u) - u \leq 0 \quad 0 < u < \tfrac{n}{2}(1-t) \,|\, X(\tfrac{n}{2}(1-t)) = j]$$

$$= 1 - \frac{2j}{n(1-t)} \quad \text{for} \quad 0 < j < n(\tfrac{1-t}{2})$$

$$0 \quad \text{if} \quad j > n(\tfrac{1-t}{2}) \;.$$

Also, using the notation of Theorem 2.2.1,

$$P[D_n(v') < D_n(s) \quad \text{for some} \quad v' \in [s+t, s+t+\tfrac{1-t}{2} \,|\, A \text{ and } X^{s+t}(\tfrac{1-t}{2}) = j]$$

$$= P\left[\sup_{0 < u < \frac{1-t}{2}} \left(F_{n,j}(u) - u\right) > \frac{d_n + \varepsilon_n}{2}\right]$$

$$= O\left(e^{-2j(d_n+\varepsilon_n)^2/(1-t)^2}\right)$$

where $T_1, T_2, \ldots, T_j$ are the order statistics based on j uniform random variables on the interval $[0, \frac{1-t}{2}]$. The "O term" was obtained in the same way as the corresponding term in Lemma 2.4.2.

Similarly,

$$R_2(t, n, m, j) = O\left(1 - \frac{2(n-m-j-1)}{n(1-t)}\right) \quad \text{for} \quad j = \frac{n}{2}(1-t-d_n+\varepsilon_n)$$

and

$$R_2(t, n, m, j) \sim \frac{d_n}{1-t} \quad \text{for} \quad \varepsilon_n = o(d_n)$$

and we can deduce

$$R(t, n, m) \sim \left(\frac{d_n}{1-t}\right)^2 .$$

These lemmas are used in the proof of the following Theorem

Theorem 2.4.1

If $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random variables with continuous distribution functions and V_n is the Kuiper statistic (as defined in Section 2.0) generated by $X_1, \ldots, X_n$, and if $\{d_n\}$ is a real sequence such that $d_n > 0$, $d_n = o(1)$ and $n d_n^2 > 1/2 \log n$ then

$$P[V_n > d_n] \sim 8 n d_n^2 e^{n d_n^2 g_n(t_n^*)}$$

where $d_n^2 g_n(t) = (t+d_n)\log\left(\frac{t}{t+d_n}\right) + (1-t-d_n)\log\left(\frac{1-t}{1-t-d_n}\right)$ and t_n^* is the value at which $g(t)$ attains its maximum $0 < t < 1$.

Proof.

The proof is similar to that of Lemma 2.3.4.

By Lemma 2.4.1, 2.4.2, 2.4.3 it is clear that for $m_c = n(t + c\, d_n)$

$$p_n(t, \frac{m_c}{n} - t) \sim (2\pi)^{-1/2}(t(1-t))^{-5/2}(c\,d_n)^4\, n^{3/2}\, e^{n d_n^2 g_n(c,t)},$$

here $d_n^2 g_n(c, t)$ is determined by replacing d_n by $c\,d_n$ in the definition of $d_n^2 g_n(t)$ and, repeating the reasoning of Lemma 2.3.3,

$$\sum_{k=0}^{n-m} p_n(t, \frac{m+k}{n} - t)$$

$$\sim 4n\,d_n \int_1^{\frac{1-t}{d_n}} (2\pi)^{-1/2}(t(1-t)^{-5/2}(c\,d_n)^4\, n^{3/2}\, e^{n d_n^2 g_n(c,t)}\, dc\,.$$

$$\sim \frac{(2\pi)^{-1/2} 4\left(\frac{n d_n^2}{(t)(1-t)}\right)^{5/2}}{n d_n\left[\log(1 + \frac{d_n}{t}) - \log(1 - \frac{d_n}{1-t})\right]}\; e^{n d_n^2 g_n(t)}\,.$$

It remains to integrate this asymptotic density with respect to s and t. Integration with respect to s leaves the expression unchanged. Integration with respect to t is carried out using the method of Laplace (see the proof of Lemma 2.3.4) to obtain

$$\int\int \sum_h p_n(t, \frac{m+h}{n} - t)\, ds\, dt$$

$$\sim \frac{(2\pi)^{-1/2}\left(\frac{n d_n^2}{t^*(1-t^*)}\right)^{5/2}}{n d_n\left(d_n/(t^*)(1-t^*)\right)} \left(\frac{2\pi}{n d_n^2\left[\frac{1}{t^{*2}(t+d_n)} + \frac{1}{(1-t^*)^2(1-t^*-d_n)}\right]}\right)^{1/2} e^{n d_n^2 g_n(t_n^*)}$$

$$\sim 8\, n d_n^2\, e^{n d_n^2 g_n(t_n^*)}\,.$$

2.5. Probabilities of Large Deviations of Kolmogorov-Smirnov Statistics

Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed, continuous random variables with common cumulative distribution function F. Then $Y_1 = F(X_1)$, $Y_2 = F(X_2)$, $\ldots$ is a sequence of independent, uniform random variables on the interval $(0,1)$. Define $D_n^+ = \sup_{0<t<1}\left(F_n(t) - t\right)$ where $F_n(t)$ is the empirical distribution function of $Y_1, Y_2, \ldots, Y_n$. Of course D_n^+ is equivalent to the usual one-sided Kolmogorov-Smirnov statistic, $\sup_{-\infty<X<\infty}\left(G_n(X) - F(X)\right)$, where $G_n(\ )$ is the empirical distribution function of $X_1, X_2, \ldots, X_n$.

Of interest is the rate at which the probability that D_n^+ is greater than a constant tends to zero as n becomes large.

<u>Theorem 2.5.1</u>

Let D_n^+ be defined as above and let c be a real number, $0 < c < 1$. Then

$$P[D_n^+ > c] \sim \alpha(c)[\beta(c)]^n , \tag{2.5.1}$$

where

$$\beta(c) = \left[\left(\frac{t^*}{t^*+c}\right)^{t^*+c}\left(\frac{1-t^*}{1-t^*-c}\right)^{1-t^*-c}\right], \tag{2.5.2}$$

and

$$\alpha(c) = \frac{(1-e^{-w^*})}{c}\left[\frac{1-t^*-c}{t^{*2}(t^*+c)^2} + \frac{1}{(1-t^*)^2(t^*+c)}\right]^{-1/2} . \tag{2.5.3}$$

Also, t^* is a root of the equation

$$\log\left(\frac{t}{t+c} \cdot \frac{1-t-c}{1-t}\right) + \frac{c}{t(1-t)} = 0 ; \; 0 < t < 1-c ,$$

and w^* is the largest real root of the equation

$$\left(1 + \frac{c}{t^*}\right)\left(1 - e^{-w}\right) = w .$$

Proof.

Without loss of generality, it is assumed that $\{X_n\}$ is a sequence of independent, uniform random variables on the unit interval. The structure of the proof is the same as for the excessive deviation case. In particular, Lemma 2.3.1 is taken as the starting point of the proof. The first task, then, is to obtain an asymptotic expansion for $p(t, \frac{m}{n} - t)$ as n becomes large for integer valued m,

$$m = n(t + c + \theta/n) \tag{2.5.4}$$

where $k \leq \theta < k+1$, c is a fixed positive number, and $t+c$ is bounded away from 0 and 1; $0 < x_0 < t+c < x_1$.

In evaluating $L_n(t, m)$ the inequalities (2.2.12) and (2.2.13) are again employed. Let

$$\lambda = (1 + \frac{c}{t}) \tag{2.5.5}$$

$$\lambda_n^+ = \left(1 + \frac{c + \varepsilon_n + \theta/n}{t}\right) \tag{2.5.6}$$

$$\lambda_n^- = \left(1 + \frac{c - \varepsilon_n + \theta/n}{t}\right) \tag{2.5.7}$$

where $\varepsilon_n > 0$ and $n\varepsilon_n^2 = a \log n$ if $\frac{a \log n}{n} < c^2$ and 0 elsewhere. Let K be a Poisson random variable with parameter μ. Then

$$P[K \geq m] \leq \min_{s \geq 1} S^{-m} E S^K = \left(\frac{\mu}{m}\right)^m e^{m-\mu} \tag{2.5.8}$$

In particular, if $\mu = n t \lambda_n^-$ and m is defined as above

$$\sum_{j \geq m} P_n(\lambda_n^-, j) \leq \left(\frac{nt\lambda_n^-}{m}\right) e^{m - nt\lambda_n^-} \leq e^{-\frac{n\varepsilon_n}{2(t+c+\theta/n)}} \leq n^{-a} \tag{2.5.9}$$

$$\sum_{j \geq m+2n\varepsilon_n} P_n(\lambda_n^+, j) \leq e^{-\frac{n\varepsilon_n^2}{2(t+c+2\varepsilon_n+\theta/n}} \leq n^{-a} \tag{2.5.10}$$

and

$$\sum_{j \leq m} P_n(\lambda_n^+, j) \leq \min_{s \geq 1} S^m E S^{-K} \leq e^{-\frac{n\varepsilon_n^2}{2(t+c+\theta/n)} + \frac{n\varepsilon_n^3}{3(t+c+\theta/n)^2}} \leq n^{-a}. \tag{2.5.11}$$

Let w_+ denote the root associated with λ_n^+ as defined in Theorem 2.2.2. Similarly let w_- denote the root associated with λ_n^-. Then, returning to the inequality (2.2.12) it is seen that

$$L_n(t, m) \leq (1 - e^{-w_+})(1 + 2n^{-a}) \tag{2.5.12}$$

for all n sufficiently large, and using inequality (2.2.13) it is seen that

$$L_n(t,m) \geq (1 - e^{-w})(1 - 2n^{-a}) \tag{2.5.13}$$

for all n suficiently large.

Finally, let w denote the root associated with λ. Since $w = \lambda - \lambda^*$ where $\lambda^* < 1$ is the solution to $\lambda e^{-\lambda} = \lambda^* e^{-\lambda^*}$, the following inequalities obtain

$$w < w_+ < w + \frac{2(\varepsilon_n + \theta/n)}{t} \; ; \quad w - \frac{2(\varepsilon_n + \theta/n)}{t} < w_- < w . \qquad \begin{matrix}(2.5.14)\\(2.5.15)\end{matrix}$$

Thus, selecting $a > 1/2$, it is concluded that

$$L_n(t,m) = \left(1 - e^{-w}\right)\left(1 + \bar{o}(n^{-1/2})\right) \tag{2.5.16}$$

An asymptotic expansion for the expression

$$P_n(t,m) = m\binom{n}{m} t^{m-1}(1-t)^{n-m}$$

may be found in a mannersimilar to that used in Lemma (2.2.2). It is found that

$$P_n(t,m) = \left(\frac{(t+c)n}{2\pi t^2(1-t-c)}\right)^{1/2}\left(\beta(t,c)\right)^n\left(a(t,c)\right)^{\theta}\left(1 + O(\frac{k^2}{n})\right) \tag{2.5.17}$$

where

$$\beta(t,c) = \left(\frac{t}{t+c}\right)^{t+c}\left(\frac{1-t}{1-t-c}\right)^{1-t-c}$$

and

$$a(t,c) = \left(\frac{t}{1-t} \cdot \frac{1-t-c}{t+c}\right)$$

Thus

$$p(t, \frac{m}{n} - t) = \left(\frac{c(1-e^{-w})}{t(1-t)}\right)\left(\frac{n(t+c)}{2\pi(1-t-c)}\right)^{1/2}\left(a(t,c)\right)^{\theta}\left(\beta(t,c)\right)^{n}\left(1+a_n(t,k)\right) \tag{2.5.18}$$

and $\sup a_n(t,k) = \bar{o}(n^{-1/2})$ for $x_0 < t+c < x_1$ and $0 \le k \le k_n = o(n^{1/2})$
The probability density that the maximum deviation of D_n^+ from 0 occurs at t and is greater than or equal to $n(t+c)$ is

$$q_n(t,c) = \sum_{k \ge 0} p(t, \frac{m_k}{n} - t) \tag{2.5.19}$$

where $m_k = n(t + c + \frac{k+\theta_n}{n})$ is integral valued and $0 \le \theta_n < 1$. To see that

$$q_n(t,c) = \frac{p(t, \frac{m_0}{n} - t)}{1 - a(t,c)} \tag{2.5.20}$$

write

$$q_n(t,c) = \sum_{k=0}^{k'} p(t, \frac{m_k}{n} - t) + \sum_{k \ge k'+1} p(t, \frac{m_k}{n} - t) .$$

The first sum on the right tends to the required limit for large k';

$$\sum_{k=0}^{k'} p(t, \frac{m_k}{n} - t) = p(t, \frac{m_0}{n} - t)\left(\frac{1 - a^{k'+1}(t,c)}{1 - a(t,c)}\right)\left(1 + \bar{o}(n^{-1/2})\right)$$

if $k' = o(n^{1/2})$.

The second sum on the right is bounded above by the sum of the probability densities of the order statistics themselves. Thus, employing Lemma 2.2.3.

$$\sum_{k \geq k'+1} p(t, \frac{m_k}{n} - t) \leq \sum_{t \geq k'+1} k\binom{n}{k} t^{k-1}(1-t)^{n-k}$$

$$\leq n\left[\left(1 - \frac{c+k'}{1-t}\right)^{1-t-c-k'}\left(1 + \frac{c+k'}{t}\right)^{t+c+k'}\right]^{-n}$$

$$\leq n\left(\frac{t}{1-t} \cdot \frac{1-t-k'-c}{c+t+k'}\right)^{nk'}\left(\beta(t, c)\right)^{n}$$

$$= o\left(\beta(t, c)\right)^{n} \quad \text{for some} \quad k' = O\left(\left(\frac{\log n}{n}\right)^{-1}\right) \tag{2.5.21}$$

Thus $q_n(t, c)$ may be written as

$$q_n(t, c) = f_n(t, c)\left(a(t, c)\right)^{\theta_n}(1 + a_n) \tag{2.5.22}$$

where $|a_n| = o(1)$ independent of s and c for s bounded away from zero and $s + c$ bounded away from 1.

The probability that D_n^+ is at least c is

$$\int_0^{1-c} q_n(s, c)\, ds \tag{2.5.23}$$

and for fixed t such that $n(t+c)$ is an integer

$$\int_{(t-\frac{1}{n})}^{t} q_n(s, c)\, ds = \int_{t-\frac{1}{n}}^{t} f_n(s, c)\,[a(s, c)]^{\theta}(1 + o(1))\, ds$$

$$= (1 + o(1)) \int_{t-\frac{1}{n}}^{t} f_n(s, c)\,[a(s, c)]^{n(t-s)}\, ds. \tag{2.5.24}$$

The integral on the right may be expressed as

$$\int_{t-\frac{1}{n}}^{t} f_n(s,c)\,[a(s,c)]^{n(t-s)}ds = \frac{1}{n}\int_0^1 f_n(t-\tfrac{u}{n}),c)\,(a(t-\tfrac{u}{n},c))^u du$$

$$\sim \frac{1}{n} f_n(t,c)\int_0^1 [e^{\frac{-uc}{t(1-t)}}][a(t,c)]^{-u}[a(t-\tfrac{u}{n},c)]^u\,du$$

$$\sim \frac{1}{n} f_n(t,c)\int_0^1 e^{-\frac{uc}{t(1-t)}}\,du$$

$$= \frac{t(1-t)}{cn} f_n(t,c)\left[1-e^{-\frac{c}{t(1-t)}}\right]. \qquad (2.5.25)$$

This may be seen by letting $s = t - \frac{u}{n}$; $0 \le u < 1$ then

$$\int_{t-\frac{1}{n}}^{t} f_n(s,c)\,[a(s,c)]^{n(t-s)}ds = \frac{1}{n}\int_0^1 f_n(t-\tfrac{u}{n},c)\,[a(t-\tfrac{u}{n},c]^u\,du$$

and observing that $a(t-\frac{u}{n},c) \sim a(t,c)$ and that

$$f_n(t-\tfrac{u}{n},c) \sim f_n(t,c)\left(\frac{t-\frac{u}{n}}{t}\right)^{n(t+c)}\left(\frac{1-t+\frac{u}{n}}{1-t}\right)^{n(1-t-c)}\left(\frac{t+c}{t}\cdot\frac{1-t}{1-t-c}\right)^u$$

$$\sim f_n(t,c)\,e^{-\frac{cu}{t(1-t)}}\,[a(t,c)]^{-u}.$$

This final relationship is used to define $g_n(s,c)$, a smooth approximation to $q_n(s,c)$. Define $b(s,c) = \log a(s,c) + \frac{c}{s(1-s)}$. Define

$$g_n(s,c) = f_n(s,c)\left[\frac{b(s,c)}{1-e^{-b(s,c)}}\right]\left[\frac{s(1-s)}{c}\;1-e^{-\frac{c}{s(1-s)}}\right] \qquad (2.5.26)$$

if $b(s,c) \neq 0$ and

$$g_n(s,c) = f_n(s,c)\left(\frac{s(1-s)}{c}\right)\left(1 - e^{-\frac{c}{s(1-s)}}\right)$$

if $b(s,c) = 0$. Then (by 2.6.24, 2.6.25)

$$\int_{t-\frac{1}{n}}^{t} g_n(s,c)\,ds \sim \int_{t-\frac{1}{n}}^{t} q_n(s,c)\,ds\,.$$

And from the above discussion it is evident that for t bounded away from 0, and $t+c$ bounded away from 1, that

$$\int_{t-\frac{1}{n}}^{t} q_n(s,c)\,ds = \left(\int_{t-\frac{1}{n}}^{t} g_n(s,c)\,ds\right)\left(1 + o_n(t,c)\right)$$

where $|o_n(t,c)| < a_n$ and a_n is a sequence convergent to 0 independent of t and c.

Thus for $0 < t_0 < t_1 < 1-c$

$$\int_{t_0}^{t_1} q_n(s,c)\,ds \sim \int_{t_0}^{t_1} g_n(s,c)\,ds\,.$$

The left hand term is the probability that the maximum deviation of $F_n(s) - s$ is at least c and is obtained at some s, $t_0 < s < t_1$.

The right hand term may be integrated using the

<u>Method of Laplace [4, 20]</u>:

Let $\phi(x)$ and $h(x)$ be two real continuous functions defined on (a,b) such that

(i) $\phi(x)\,e^{nh(x)}$ is absolutely integrable for every positive value of $n > n_0$.

(ii) $h(x)$ has a single maximum in the interval, an interior point of (a, b).

(iii) $h''(x)$ is continuous, $h'(y) = 0$ and $h''(y) < 0$. Then as $n \to \infty$

$$\int_a^b \phi(x)\,e^{nh(x)}\,dx \sim \phi(y)\left[\frac{-2\pi}{nh''(y)}\right]^{1/2} e^{nh(y)}$$

In our case

$$h(s) = \log\left(\frac{t}{t+c}\right)^{t+c}\left(\frac{1-t}{1-t-c}\right)^{1-t-c}$$

$$\phi(s) = \left[\frac{s+c}{2\pi(1-s-c)}\right]^{1/2}\frac{(1-e^{-w})}{1-a}\left[\frac{b}{1-e^{-b}}\right]\left[1-e^{-\frac{c}{s(1-s)}}\right]$$

$a = a(s, c)$, $b = b(s, c)$, $w = w_{\lambda_s}$ as defined above, and

$$\int_{t_0}^{t_1} g_n(s, c)\,ds = n^{1/2}\int_{t_0}^{t_1}\phi(s)\,e^{nh(s)}ds = \phi(y)\left(\frac{-2\pi}{h''(y)}\right)^{1/2} e^{nh(y)}.$$

Since

$$h(s) = \log\left(\frac{t}{t+c}\right)^{t+c}\left(\frac{1-t}{1-t-c}\right)^{1-t-c}$$

$$h'(s) = \left(\log\frac{s}{s+c}\cdot\frac{1-s-c}{1-c}\right) + \frac{c}{s(1-s)} = \log a + \frac{c}{s(1-s)} = b$$

$$h''(s) = -c^2\left[\frac{1}{s^2(s+c)} + \frac{1}{(1-s)^2(1-s-c)}\right]$$

it is evident that $h'(s)$ has at most one root in the interval $[0, 1-c]$ and it must be in the interior of the interval since

$$\lim_{s\to 0} h'(s) = \infty \qquad \lim_{s\to 1-c} h'(s) = -\infty$$

Call this root t^* and select t_0 and t_1, so that $t_0 < t^* < t_1$.

Evidently conditions (i), (ii), (iii) of the method of Laplace are satisfied.

Thus

$$\int_{t_0}^{t_1} q_n(s,c)\,ds \sim \int_{t_0}^{t_1} g_n(s,c)\,ds$$

$$\sim \frac{(1-e^{-w^*})}{c}\left(\frac{1-t^*-c}{t^{*2}(t^*+c)^2} + \frac{1}{(1-t^*)^2(t^*+c)}\right)^{-1/2} e^{-nh(t^*)}$$

$$= \alpha(c)\,[\beta(c)]^n$$

($\alpha(c)$ and $\beta(c)$ are defined in lines (2) and (3) of the Theorem).

To complete the proof it should be noted that

$$P[D_n^+ > c] = \int_0^{t_0} q_n(t,c)\,dt + \int_{t_0}^{t_1} q_n(t,c)\,dt + \int_{t_1}^{1} q_n(t,c)\,dt$$

$$\sim \alpha(c)\,[\beta(c)]^n + \int_0^{t_0} q_n(t,c)\,dt + \int_{t_1}^{1} q_n(t,c)\,dt.$$

It remains to be shown that the two integrals on the right are small compared with $\alpha(c)\,[\beta(c)]^n$.

The integral

$$\int_0^{t_0} q_n(t, c)\, dc$$

denotes the probability of the event that the maximum deviation of $F_n(s) - s$ is at least c and is obtained for some s in the interval $(0, t_0)$. The event described is realized only if at least nc of the order statistics, Y_k, are in the interval $(0, t_0)$. The probability of the latter event is $P[K \geq nc]$ where K is a binomial random variable with parameters n and t_0. Thus

$$\int_0^{t_0} q_n(t, c)\, dt \leq P[K \geq nc] \leq \min_{s>1} s^{-nc} Es^{+K}$$

$$= \min_{s>1} s^{-nc} (ts + 1 - t)^n$$

$$= [\beta(t_0, c)]^n \quad \text{say}$$

Clearly $\beta(t_0, c) < \beta(c)$ if $0 < t_0 < 1/2$ and $t_0 < (1-c)^{\frac{1-c}{c}} c[\beta(c)]^{1/c}$. For any such t_0,

$$\int_0^{t_0} q_n(t, c)\, dt = o(\alpha(c)[\beta(c)]^n) .$$

A similar argument demonstrates that

$$\int_{t_1}^{1} q_n(t,c)\,dt \le P[K > n(t_1+c)] \le [\beta(t_1, t_1+c)]^n$$

(here K is a binomial random variable with parameters n and t_1) so for any t_1 ; $t^* < t_1 < 1-c$

$$\int_{t_1}^{1} q_n(t,c)\,dt = o(\alpha(c)\,[\beta(c)]^n)\,.$$

This concludes the proof.

The following two theorems are immediate.

Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with common, continuous distribution function F. Define the sample distribution function as

$$F_n(x) = \sum_{X_i \le x} \frac{1}{n}\,.$$

Define the Kolmogorov-Smirnov statistics:

$$D_n^- = \sup_{-\infty < x < \infty} \left(F(x) - F_n(x)\right)$$

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|\,.$$

Define functions of c, $0 < c < 1$ as follows:

t^* ; the root of the equation

$$\left(\log\left(\frac{s}{s+c} \cdot \frac{1-s-c}{1-s}\right)\right) + \left(\frac{c}{s(1-s)}\right) = 0\,;\; 0 < s < 1-c$$

w^* ; the largest real root of the equation

$$(1+\frac{c}{t^*})(1-e^{-w})-w=0$$

$$\beta\ ;\ \beta=\exp\{(t^*+c)(\log\frac{t^*}{t^*+c})+(1-t^*-c)(\log\frac{1-t^*}{1-t^*-c})\}=\frac{1-t^*}{1-t^*-c}\exp-\frac{c(c+t^*)}{t^*(1-t^*)}$$

$$\alpha\ ;\ \alpha=\frac{w^*t^*}{c}\left[\frac{1-t^*-c}{t^{*2}}+\frac{t^*+c}{(1-t^*)^2}\right]^{-1/2}.$$

Theorem 2.5.2

Let D_n^- be defined as above, then

$$P[D_n^- > c] \sim \alpha\cdot\beta^n.$$

Theorem 2.5.3

Let D_n be defined as above, then

$$P[D_n > c] \sim 2\alpha\cdot\beta^n.$$

To verify theorem 2.5.2 assume, without loss of generality, that $X_1, X_2, \ldots$ are uniform random variables on the unit interval and define $U_i = 1 - X_i$, $i = 1, 2, 3, \ldots$. Let

$$F_n^x(x)=\sum_{X_i\le x}\frac{1}{n}\quad\text{and}\quad F_n^u(x)=\sum_{U_i<x}\frac{1}{n}$$

then

$$x-F_n^x(x)=F_n^u(1-x)-(1-x).$$

Thus $P[D_n^- > c] = P[\sup_u \left(F_n^u(u) - u\right) > c]$ and the probability on the right is determined in Theorem 2. 5. 1.

Theorem 2. 5. 3. follows from the observation that

$$P[D_n > c] = P[D_n^+ > c] + P[D_n^- > c] - P[D_n^- > c, D_n^+ > c]$$

and the probability on the right is small compared to the others.

2.6. Probabilities of Large Deviations of the Kuiper Statistic.

Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with common continuous distribution function F. As before, the sample distribution function generated by the first n random variables is defined as

$$F_n(x) = \sum_{x_i \leq x} \frac{1}{n} ; \; i = 1, 2, \ldots, n$$

and the Kuiper Statistic is defined as

$$V_n = D_n^+ + D_n^- = \sup_{-\infty < x < \infty} (F_n(x) - F(x)) - \inf_{-\infty < x < \infty} (F_n(x) - F(x)) .$$

I. G. Abrahamson [1] proved that the distributions of the Kuiper statistic and the Kolmogorov-Smirnov statistic are of the same exponential order in the tails,

$$\lim_{n \to \infty} \frac{\log P[D_n > c]}{\log P[V_n > c]} = 1 ; \; 0 < c < 1 .$$

The following theorem, together with Theorem 2.4.3 proves that the probabilities of large deviations of V_n and D_n are not asymptotically equivalent, in fact,

$$\lim_{n \to \infty} \frac{P[D_n > c]}{P[V_n > c]} = 0 ; \; 0 < c < 1 .$$

<u>Theorem 2.6.1</u>

Let V_n be defined as above and let c be a real number, $0 < c < 1$, then

$$P[V_n > c] \sim \alpha_1(c)\, n[\beta(c)]^n . \tag{2.6.1}$$

The definitions of $\beta(c)$, w^* and t^* are given in Theorem 2.5.1 and

$$\alpha_1(c) = \frac{t^2(1-c-t)^{1/2} w^2}{(1-t)^2 (t+c)^{3/2}} \left[\frac{1}{t^2(t+c)} + \frac{1}{(1-t)^2(1-t-c)}\right]^{-1/2} \tag{2.6.2}$$

$$w = w^*, \; t = t^*,$$

Proof.

The method of proof is similar to that of Theorem 2.5.1. Assume, without loss of generality, that $X_1, X_2, \ldots, X_n$ are n independent uniform random variables on the unit interval. The notation $s \dotplus t$ is used to describe addition on the unit circle. That is, for $0 \le s < 1$ and $0 \le t < 1$ define

$$s \dotplus t = \begin{cases} s+t & \text{if } 0 \le s+t \le 1 \\ |1-s-t| & \text{if } 1 < s+t \le 2 \end{cases}$$

and $(s, s\dotplus t)$ denotes the interval $(s, s+t)$ if $0 < s+t \le 1$ and the union of intervals $(s, 1)$ and $[0, |1-s-t|)$ if $1 < s+t \le 2$.

If the infimum of $F_n(v) - v$ occurs at s and the maximum at $s \dotplus t$ and the m^{th} order statistic on the interval $(s, s\dotplus 1)$ occurs at $s \dotplus t$ then $V_n = \frac{m}{n} - t$. For large deviations, an asymptotic expression for the joint probability density of the above event for $m \ge n(c+t)$ is of interest.

To begin with, the probability density that an X_i occurs at s; $0 < s < 1$ is nds.

The conditional probability density that the m^{th} order statistic in the interval $(s, s+1)$ occurs at $s+t$, $0 < t < 1$, given that an X_i occurs at s is the same as the probability density of the m^{th} of $n-1$ order statistics of uniform random variables on the unit interval,

$$\frac{(n-1)!}{(m-1)!\,(n-m-1)!}\; t^{m-1}(1-t)^{n-m-1} .$$

Thus the joint density of an order statistic at s and the m^{th} order statistic in the interval $(s, s+1)$ at $s+t$ is

$$P(t, n, m) = \frac{n!}{(m-1)!\,(n-m-1)!}\; t^{m-1}(1-t)^{n-m-1} .$$

Finally, the joint probability density that the infimum of $F_n(v) - v$ is at s and the maximum is at $s+1$ and that $V_n = \frac{m}{n} - t$ may be written as

$$L(t, n, m)\,P(t, n, m)\,R(t, n, m), \tag{2.6.3}$$

$L(t, n, m) = P[F_n(s)-s \leq F_n(u) - u \leq F_n(s+t) - s+t\,;\; u \in (s, s+t)\,|\,A]$,

$R(t, n, m) = P[F_n(s)-s \leq F_n(u)-u \leq F_n(s+t) - s+t\,;\; u \in (s+t, s+1)\,|\,A]$,

A is the event which has probability density $P(t, n, m)$.

It will be shown that for $m = n(t+c)$ and for large n

$$P(t, n, m) \sim \left(\frac{(t+c)\,(1-t-c)}{2\pi t^2 (1-t)^2}\right)^{1/2} n^{3/2}\left[\left(\frac{t}{t+c}\right)^{t+c}\left(\frac{1-t}{1-t-c}\right)^{1-t-c}\right]^n, \tag{2.6.4}$$

$$R(t, n, m) \sim \left(\frac{c}{1-t}\right)^2,$$

and $\quad L(t, n, m) \sim \left(1 - e^{-w}\right)^2 = \left(\frac{wt}{t+c}\right)^2.$

Thus the joint density is

$$p_n(t, c) = L(t, n, m)\, P(t, n, m)\, R(t, n, m)$$

$$= \frac{(w^2 c^2 t)(1-t-c)^{1/2} n^{3/2}}{(2\pi)^{1/2}(1-t)^3 (t+c)^{3/2}} \; (\beta(t, c))^n. \qquad (2.6.5)$$

As was shown in Theorem 2.5.1

$$\beta^n(t, c + \frac{k}{n}) \sim a^k(t, c)\beta^n(t, c)$$

$$a(t, c) = \left(\frac{t}{1-t}\right)\left(\frac{1-t-c}{t+c}\right) .$$

The joint probability density that the infimum of $F_n(v) - v$ is at s and the maximum is at $s+t$ and that $V_n \geq \frac{m}{n} - t$, $m = n(t+c)$, is

$$q_n(t, c) \sim \frac{1}{1 - a(t, c)} \; p_n(t, c).$$

Apparently $q_n(t, c)$ is independent of s so that

$$\int_0^1 q_n(t, c)ds = q_n(t, c)$$

as in Theorem 2.5.1 we have

$$\int_{t-\frac{1}{n}}^{t} q_n(u, c)du \sim \int_{t-\frac{1}{n}}^{t} g_n(u, c)du$$

for

$$g_n(u, c) = f_n(u, c)\left[\frac{b(u, c)}{1 - e^{b(u, c)}}\right]\left[\frac{u(1-u)}{c}\right]\left[1 - e^{-\frac{c}{u(1-u)}}\right],$$

$f_n(u, c)$ denotes the right hand side of (2. 6. 5) divided by $1 - a(t, c)$, and $b(u, c)$ is defined as in (2. 5.26).

Finally the probability that V_n is greater than c is

$$P[V_n > c] = \int_0^1 q_n(t, c)dt \sim \int_{t_0}^{t_1} g_n(t, c)dt$$

for any t_0 and t_1 such that $0 < t_0 < t^* < t_1 < 1-c$. The last integral may be integrated using the method of Laplace thus obtaining $P[V_n > c] \sim \alpha_1(c)n[\beta(c)]^n$. It remains to verify the asymptotic expansions listed under (2. 6. 4).

The asymptotic expansion of $P(t, n, m)$ is found by using Sterling's approximation of $k!$ as described following line (2. 2. 21).

Next, for any Poisson stochastic process $\{X(v), v \geq 0\}$, define $L'(a, b, k) = P[a \leq v - X(v) \leq 1 ; 0 < v < b \,|\, X(0) = 0, X(b) = k]$. For $L(t, n, m)$ as defined in (2. 6. 3), identifying the $m - 1$ order statistics in the interval $(s, s+1)$ with the corresponding order statistics of $m - 1$ independent random variables on the interval (o, nt) and these, in turn, with the $m - 1$ jump points of a Poisson process in the interval $(0, nt)$ we observe that

$$\begin{aligned} L(t, n, m) &= L'(nt - m, nt, m - 1) \\ &= \sum_j L'(nt - m, \tfrac{nt}{2}, [\tfrac{m-1}{2}] + j) \\ &\quad \cdot P[X(\tfrac{nt}{2}) = [\tfrac{m-1}{2}] + j \,|\, X(nt) = n - 1] \\ &\quad \cdot L'(nt - m, \tfrac{nt}{2}, m - 1 - [\tfrac{m-1}{2}] - j] \end{aligned}$$

The summation is over those integers j for which

$$0 \leq [\frac{m-1}{2}] + j \leq n-1 .$$

The conditional probability of $X(\frac{nt}{2})$ is binomial with parameters 1/2 and m-1. Using an excessive deviation result for binomial random variables it is seen that, for $m = n(c+t)+k$, $k = 0(\sqrt{n})$ and $|j| \leq a_n$, a_n increasing faster than $\sqrt{n}$,

$$\sum_{|j| \leq a_n} P[X(\frac{nt}{2}) = [\frac{m-1}{2}] + j \,|\, X(nt) = m-1] \sim 1 . \qquad (2.6.7)$$

For $L_n(t, k)$ defined as in (2.5.10), it is now apparent that

$$L'(nt-m , \frac{nt}{2} , [\frac{m-1}{2}] + j) \sim L_n(\frac{t}{2} , [\frac{m-1}{2}] + j) \qquad (2.6.8)$$

because

$$P[nt-m < v - X(v) ; 0 < v < \frac{nt}{2} \,|\, X(o) = 0, X(\frac{nt}{2}) = [\frac{m-1}{2}] + j] \sim 1 (2.6.9)$$

and, for m, k and j defined as above and $k_n = [\frac{m-1}{2}] + j$, relation 2.5.16 applies so that

$$L_n(\frac{t}{2} , k_n) \sim 1 - e^{-w} > 0 \qquad (2.6.10)$$

and w is the largest real root of the equation $w = (1 + \frac{c}{t})(1 - e^{-w})$.

Applying (2.6.7) through (2.6.10) to (2.6.6) it is seen that $L(t, n, m) \sim (1 - e^{-w})^2$.

The asymptotic expansion of $R(t, n, m)$ is obtained in a similar manner and we have

$$R(t, n, m) \sim \left(\frac{c}{1-t}\right)^2$$

This completes the proof.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Abrahamson, Innis G. (1967). Exact Bahadur efficiences for the Kolmogorov-Smirnov and Kuiper one-and two-sample statistics. Ann. Math. Stat. 38, pp. 1475-1483.

2. Birnbaum, Z. W. and Ronald Pyke. (1958). On some distributions related to the statistic D_n^+. Ann. Math. Stat. 29, pp. 179-187.

3. Blackwell, David and J. L. Hodges, Jr. (1959). The probability in extreme tail of a convolution. Ann. Math. Stat. 30, pp. 1113-1120.

4. Copson, E. T. (1965). Asymptotic Expansions. Cambridge University Press, Cambridge.

5. Darling, D. A. (1960). On the theorems of Kolmogorov-Smirnov. Theory of Probability and its Applications, 5, pp. 356-60.

6. Doob, J. L. (1953). Stochastic Processes. Wiley, New York.

7. Feller, William. (1957). An Introduction to Probability Theory and Its Applications. Vol. 1, 2nd edition. Wiley, New York.

8. Feller, William (1966). An Introduction to Probability Theory and Its Applications. Vol. 2, Wiley and Sons, New York.

9. Fulks, W. (1951). A Generalization of Laplace's Method. Proceedings of the American Mathematical Society, Vol. 2, p. 613-622.

10. Gnedenko, B. V., V. S. Koroluk and A. V. Skorokhod. (1960). Asymptotic expansions in probability theory. Proc. Fourth Berkeley Symp. Math. Stat. Prob. Vol. 2, pp. 153-170. Univ. of California, Berkeley.

11. Hoeffding, W. (1948). On a class of statistics with asymptotically normal distribution. Ann. Math. Stat. Vol. 19, pp. 293-323.

12. Karlin, Samuel. (1966). A First Course in Stochastic Processes. Academic Press, New York.

13. Kuiper, Nicolaas H. (1960). On the random cumulative frequency function. Proc. of the Royal Neth. Academy of Sciences Series A. Vol. 63, pp. 32-37.

14. Kuiper, Nichlass H. (1960). Tests concerning random points on a circle. Proc. of the Royal Neth. Academy of Sciences Series A. Vol. 63, pp. 38-47.

15. Loéve, Michel. (1963). Probability Theory (3rd edition). Van Nostrand, Princeton.

16. Pyke, Ronald, (1959). The supremum and infimum of the Poisson process. Ann. Math. Stat. Vol. 30, pp. 568-576.

17. Rubin, Herman and J. Sethuraman. (1965). Probabilities of moderate deviations. Sankhyã. Series A. Vol. 27, pp. 325-344.

18. Sethuraman, J. (1964). On the probability of large deviations of sample means. Ann. Math. Stat. Vol. 35, pp. 1304-1316.

19. Takács, Lajos. (1967). Combinatorial Methods in the Theory of Stochastic Processes. Wiley and Sons, New York.

20. Widder, D. V. (1941). The Laplace Transform. Princeton University Press, Princeton.