

APPENDIX A. EXAMPLE OF CONSERVATIVE KOLMOGOROV-SMIRNOV AND KUIPER TESTS

Our examples are as follows. Let  $X_i$  be a Bernoulli random variable with parameter  $p = \frac{1}{2}$ . Under  $H_0$ , the empirical cdf of a sample of size  $N$ ,  $F_N$  is  $F_N(x) = [1 - \bar{X}] \mathbf{1}_{[0,1)}(x) + \mathbf{1}_{[1,\infty)}(x)$  while the true cdf is  $F(x) = (1/2) \mathbf{1}_{[0,1)}(x) + \mathbf{1}_{[1,\infty)}(x)$ . The definitions of the Kolmogorov-Smirnov ( $D_N$ ) and Kuiper ( $V_N \equiv D_N^+ + D_N^-$ ) are given by

$$\begin{aligned} D_N &= \sup_x |F_N(x) - F(x)| = \left| \frac{1}{2} - \bar{X} \right| \\ D_N^+ &= \sup_x [F_N(x) - F(x)]_+ = \max\left\{ \frac{1}{2} - \bar{X}, 0 \right\} \\ D_N^- &= \sup_x [F_N(x) - F(x)]_- = \max\left\{ \bar{X} - \frac{1}{2}, 0 \right\} \end{aligned}$$

So that  $D_N = V_N = \left| \frac{1}{2} - \bar{X} \right|$ . By the CLT,  $\sqrt{N}D_N$  and  $\sqrt{N}V_N$  both converge in distribution to a  $N(0, 1/4)$ , giving asymptotic test values for a .99 level test of  $\approx 2.58/2 = 1.29$ . This shows that the respective test levels based on the assumption of a continuous  $F$ , namely 1.628 for  $D_N$  and 2.001 for  $V_N$  are much too large. In particular for  $V_N$  and large  $N$ ,  $\Pr(|\sqrt{N}(\bar{X} - \frac{1}{2})| \leq 2.001) \approx .99994$ . In other words instead of falsely rejecting the null 1% of the time, by using the 2.001 cutoff rule will falsely reject it only .006% of the time which is far too conservative.

APPENDIX B. PROOFS

It will be convenient to partition  $(0, \infty)$  into sets  $\{A_{d,k}\}$  related to First Significant Digits.

**Definition.** For  $k \in \mathbb{R}$  define the  $d^{\text{th}}$  FSD set of order  $k$ ,  $A_{d,k}$  by

$$A_{d,k} \equiv [d \cdot 10^k, (d+1) \cdot 10^k)$$

Clearly for any  $x \in \mathbb{R}_{++}$  the FSD of  $x$  is  $d$  iff  $\exists k \in \mathbb{Z}$  s.t.  $x \in A_{d,k}$ , so that  $x$  has FSD equal to  $d$  iff  $x \in A_d$  where  $A_d \equiv \bigcup_{k \in \mathbb{Z}} A_{d,k}$ . Note that in particular we have that

$$\log_{10} A_{d,k} = [\log_{10} d \cdot 10^k, \log_{10}(d+1) \cdot 10^k) = [k + \log_{10} d, k + \log_{10}(d+1))$$

so that (where  $|\cdot|$  denotes Lebesgue measure when appropriate)  $|\log_{10} A_{d,k}| = \log_{10} 1 + \frac{1}{d}$  for any  $k$ .<sup>1</sup>

B.1. Proofs for the Main Text.

**Lemma.** Suppose  $X$  is a random variable on  $(0, \infty)$  with continuous pdf and let  $Y \sim \log_{10} X$ . If  $Y \in I(\epsilon)$  then  $X$   $\epsilon$ -satisfies Benford's Law.

*Proof.* Let  $f$  denote the pdf of  $Y$ , and by definition of  $A_{k,d}$  and  $A_d$  we have that

$$(B.1) \quad \Pr(X \text{ has FSD} = d) = \Pr(Y \in \log_{10} A_d) = \sum_{k=-\infty}^{\infty} \int_{\log_{10} A_{d,k}} f(y) dy$$

---

<sup>1</sup>Carrying over the results to a general base  $b$  as now appears common in some of the literature presents no overwhelming difficulties. However, as the literature has focused on applications to testing which revolve around base 10 we stick to base 10 avoiding the extra baggage of more general notation.

By assumption  $Y \in I(\epsilon)$  so there exist constants  $\{c_i\}$  such that for each  $d$ ,

$$(B.2) \quad \begin{aligned} \epsilon &\geq \left| \sum_{k=-\infty}^{\infty} \int_{\log_{10} A_{d,k}} f(y) dy - \int_{\log_{10} A_d} \sum c_i \mathbf{1}_{[i,i+1)}(y) dy \right| \\ &= \left| \Pr(X \text{ has FSD} = d) - \sum_{k=-\infty}^{\infty} \int_{\log_{10} A_{d,k}} \sum c_i \mathbf{1}_{[i,i+1)}(y) dy \right| \end{aligned}$$

where the second line follows from Equation (B.1). Using the fact that  $\log_{10} d < 1$  we know that  $\mathbf{1}_{[k+\log_{10} d, k+\log_{10} d+1)}(y) \mathbf{1}_{[i,i+1)}(y) = 0$  unless  $k = i$  so

$$(B.3) \quad \mathbf{1}_{[k+\log_{10} d, k+\log_{10} d+1)}(y) \sum c_i \mathbf{1}_{[i,i+1)}(y) = c_k \mathbf{1}_{\log_{10} A_{d,k}}(y)$$

Using Equation (B.3), we have

$$(B.4) \quad \sum_{k=-\infty}^{\infty} \int_{\log_{10} A_{d,k}} \sum c_i \mathbf{1}_{[i,i+1)}(y) dy = \sum_{k=-\infty}^{\infty} \int_{\log_{10} A_{d,k}} c_k dy = \left[ \sum_{k=-\infty}^{\infty} c_k \right] \log_{10} \left(1 + \frac{1}{d}\right)$$

Pairing Equations (B.4) with Equation (B.2) we have that

$$(B.5) \quad \epsilon \geq \left| \Pr(X \text{ has FSD} = d) - \left[ \sum_{k=-\infty}^{\infty} c_k \right] \log_{10} \left(1 + \frac{1}{d}\right) \right|$$

Finally from Lemma ?? we may assume WLOG that  $c_i = \int_{[i,i+1)} f(x) dx$  so that  $\sum c_k = 1$ , giving the desired inequalities. Examination of the proof shows a connection to the work of Allaart<sup>2</sup>.  $\square$

**Lemma.** *Let  $f$  be an  $L^1$  function. Then*

$$\arg \min_c \sup_{A \text{ measurable}} \left| \int_{[0,1] \cap A} [f(x) - c] dx \right| = \int_{[0,1]} f(x) dx$$

and for  $c^* \equiv \int_{[0,1]} f(y) dy$ , the minimum attained is  $\frac{1}{2} \int_{[0,1]} |f(x) - c^*| dx$ .

*Proof.* We define three set mappings  $A^+(c), A^-(c), A^0(c)$  respectively by

$$A^+(c) \equiv \{x : f(x) - c > 0\} \quad A^-(c) \equiv \{x : f(x) - c < 0\} \quad A^0(c) \equiv \{x : f(x) - c = 0\}$$

and since  $f$  is measurable, each of  $A^+(c), A^-(c), A^0(c)$  is measurable. For any fixed  $c$  we also have

$$\sup_{A \text{ measurable}} \left| \int_{[0,1] \cap A} [f(x) - c] dx \right| = \max \left\{ \int_{[0,1] \cap A^+(c)} [f(x) - c] dx, - \int_{[0,1] \cap A^-(c)} [f(x) - c] dx \right\}$$

Define functions  $B^+(c), B^-(c)$  corresponding to the sets  $A^+(c), A^-(c)$  by

$$B^+(c) \equiv \int_{[0,1] \cap A^+(c)} [f(x) - c] dx \quad B^-(c) \equiv - \int_{[0,1] \cap A^-(c)} [f(x) - c] dx$$

Since  $c' > c$  implies  $A^+(c') \subset A^+(c)$ ,  $[f(x) - c] \mathbf{1}_{A^+(c)}$  is decreasing in  $c$  so that  $B^+(c)$  is decreasing and similarly  $B^-(c)$  is increasing. Since we have

$$(B.6) \quad \sup_{A \text{ measurable}} \left| \int_{[0,1] \cap A} [f(x) - c] dx \right| = \max \{B^+(c), B^-(c)\}$$

<sup>2</sup>Approximately, the stronger characterization of Benford's law made in ? (pg. 290) holds. Consequently, when the hypotheses of the lemma hold, a random variable is "approximately sum invariant" in the language of Allaart. Since we resort to this lemma throughout in the following results, both the stronger characterization and sum invariance "approximately hold" for the transformations of random variables considered.

any  $\tilde{c}$  s.t.  $B^+(\tilde{c}) = B^-(\tilde{c})$  minimizes Equation (B.6). Note that identically we have

$$(B.7) \quad B^+(c) - B^-(c) = \int_{[0,1]} [f(x) - c] dx = \int_{[0,1]} f(x) dx - c$$

so that  $c^* \equiv \int_{[0,1]} f(x) dx$  minimizes Equation (B.6) and  $c^* \in \mathbb{R}$  since  $f \in L^1$ . Furthermore, we claim  $c^*$  is unique. If a minimum is attained at any  $c' \neq c^*$  we must have

$$(B.8) \quad \max\{B^+(c'), B^-(c')\} = B^+(c^*) = B^-(c^*)$$

and Equation (B.7) shows  $B^+(c') \neq B^-(c')$  so  $B^+(c^*) = B^-(c^*) > 0$  and either  $B^+(c') < B^+(c^*)$  or  $B^-(c') < B^-(c^*)$ . The former of these two cases implies  $c' > c^*$  and  $B^-(c') = B^-(c^*)$  while inspection shows this requires  $B^-(c') = B^-(c^*) = 0$ , contradicting  $B^-(c^*) > 0$ . The latter case is similar, showing that  $c^*$  is the unique minimizer. The second claim follows from Equation (B.6) and  $B^+(c^*) = \frac{1}{2}[B^+(c^*) + B^-(c^*)]$ .  $\square$

**Lemma.**  $Y \in I(\epsilon)$  iff  $aY + b \in I(\epsilon)$  for all  $a, b \in \mathbb{Z}$  with  $a \neq 0$ .

*Proof.* One direction is obvious by taking  $a = 1, b = 0$ . Considering the other direction, fix  $Y \in I(\epsilon)$  and by assumption there exist positive constants  $c_i$  s.t. for every measurable set  $A$ ,

$$(B.9) \quad \left| \int_A f(y) dy - \int_A \sum c_i \mathbf{1}_{[i, i+1)}(y) dy \right| \leq \epsilon$$

and for any strictly monotone transformation  $T$  of  $Y$  with differentiable inverse we have  $\int_A f(y) dy = \int_{TA} f \circ T^{-1}(y) \cdot (T^{-1})'(y) dy$  where  $g(y) \equiv f \circ T^{-1}(y) \cdot (T^{-1})'(y)$  is the pdf of  $T(Y)$ . Referring to Equation (B.9), we also have

$$\int_A \sum c_i \mathbf{1}_{[i, i+1)}(y) dy = \int_{TA} \sum c_i \mathbf{1}_{[T(i), T(i+1))}(y) \cdot (T^{-1})'(y) dy$$

Assuming  $T$  is measurable, since Equation (B.9) holds for any  $A$ , in particular  $T^{-1}(A)$ , we have for any measurable  $A$  that

$$(B.10) \quad \left| \int_A g(y) dy - \int_A \sum c_i \mathbf{1}_{[T(i), T(i+1))}(y) \cdot (T^{-1})'(y) dy \right| \leq \epsilon$$

Considering  $T(x) \equiv ax + b$  for  $a, b \in \mathbb{Z}$  and appealing to Equation (B.10), we have for every  $A$  that

$$\left| \int_A g(y) dy - \int_A \sum ac_i \mathbf{1}_{[ai+b, a(i+1)+b)}(y) dy \right| \leq \epsilon$$

Defining  $d_j \equiv \sum ac_i \mathbf{1}_{[ai+b, a(i+1)+b)}(j)$ , from the last equation we have

$$\left| \int_A g(y) dy - \int_A \sum d_j \mathbf{1}_{[j, j+1)}(y) dy \right| \leq \epsilon$$

so that  $T(Y) \in I(\epsilon)$  as claimed.  $\square$

**Theorem** (Mean-Scale Approximation). *For any random variable  $Y$  with continuous pdf, and  $\epsilon > 0$  there exists a  $s \in \mathbb{R}$  s.t.  $\sigma \leq s$  implies*

$$Y/\sigma \in I(\epsilon)$$

*Additionally  $\forall \mu \in \mathbb{R}$ ,*

$$2(Y - \mu)/\sigma \in I(\epsilon)$$

*Proof.* **I first show**  $\sigma Y \in I(\epsilon)$ . Fix  $\epsilon > 0$  and denote the pdf of  $Y$  as  $f$ . Since  $\lim_{n \rightarrow \infty} \int_{[-n,n]} f(y) dy = 1$  there exists an  $N$  s.t.  $\int_{[-(N-2), N-2]} f(y) dy > 1 - \frac{\epsilon}{3}$ . Since  $f$  is uniformly continuous on  $[-N, N]$  compact,  $\exists \delta \in (0, 1)$  s.t.

$$(B.11) \quad \sup_{y \in \overline{B(x, \delta)}} f(y) - \inf_{y \in \overline{B(x, \delta)}} f(y) < \frac{\epsilon}{6N} \quad \forall x \in [-N, N]$$

where  $\overline{B(x, \delta)}$  denotes the closure of an open ball of radius  $\delta$  around  $x$ . Now fix  $s \in \mathbb{R}$  s.t.  $0 < 1/s < \delta$  and let  $r \geq s$ . Using Equation (B.11) we have

$$(B.12) \quad \sum_{i \in \mathbb{Z}, |\frac{i}{r}| < N} \left[ \sup_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) - \inf_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) \right] \frac{1}{r} < \frac{\epsilon}{6N} \frac{2rN}{r} = \frac{\epsilon}{3}$$

Now for each  $r$  define constants

$$c_{i,r} \equiv \left[ \sup_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) + \inf_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) \right] / 2 \quad d_{i,r} \equiv \left[ \sup_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) - \inf_{y \in [\frac{i}{r}, \frac{i+1}{r}]} f(y) \right] / 2$$

for  $|\frac{i}{r}| < N$  and  $c_{i,r}, d_{i,r} \equiv 0$  for  $|\frac{i}{r}| \geq N$ . for which we have  $c_{i,r} - d_{i,r} \leq f(x) \leq c_{i,r} + d_{i,r}$  for all  $x \in [\frac{i}{r}, \frac{i+1}{r}]$ . This implies that

$$|f(x) - c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x)| \leq d_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x)$$

and therefore for all measurable sets  $A$  we have

$$\begin{aligned} \left| \int_{A \cap [-N, N]} f(x) - \sum_{|\frac{i}{r}| < N} c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x) dx \right| &\leq \sum_{|\frac{i}{r}| < N} \int_{A \cap [-N, N]} |f(x) - c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x)| dx \leq \\ &\sum_{|\frac{i}{r}| < N} \int_{[-N, N]} d_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x) dx \leq \sum_{|\frac{i}{r}| < N} d_{i,r} \frac{1}{r} < \frac{\epsilon}{3} \end{aligned}$$

where the last line follows from Equation (B.12). Since

$$\left| \int_{A \cap [-N, N]^c} f(x) - \sum_{|\frac{i}{r}| \geq N} c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x) dx \right| = \int_{A \cap [-N, N]^c} f(x) dx < \frac{\epsilon}{3}$$

Putting everything together we have

$$\left| \int_A f(x) - \sum c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x) dx \right| < \epsilon$$

Letting  $h(x) \equiv f(\frac{x}{r}) \frac{1}{r}$  denote the pdf of  $rY$  we also have

$$\begin{aligned} \left| \int_A f(x) - \sum c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]}(x) dx \right| &= \left| \int_{rA} f\left(\frac{x}{r}\right) \frac{1}{r} - \sum c_{i,r} \mathbf{1}_{[\frac{i}{r}, \frac{i+1}{r}]} \left(\frac{x}{r}\right) \frac{1}{r} dx \right| \\ &= \left| \int_{rA} f\left(\frac{x}{r}\right) \frac{1}{r} - \sum \frac{c_{i,r}}{r} \mathbf{1}_{[i, i+1]}(x) dx \right| \end{aligned}$$

so that  $rY \in I(\epsilon)$  for constants  $\{\frac{c_{i,r}}{r}\}$ .

**I now show**  $\sigma(Y - \mu) \in I(\epsilon)$ . Now fix  $\mu \in \mathbb{R}$  and let  $g(x) = f(x - \mu)$  be the pdf of  $Y - \mu$ . As we want to show that  $\sigma(Y - \mu) = \sigma Y - \sigma \mu \in I(\epsilon) \forall \sigma \geq s$ , from Lemma ?? for any particular  $\sigma$  it is sufficient to consider only  $\sigma \mu \in [0, 1)$  or rather  $\mu \in [0, \frac{1}{\sigma})$ . Let  $N, s$  and  $r$  be as above so we may assume WLOG that  $\mu \in [0, \frac{1}{r})$ . Since  $\mu < 1$  we have  $\int_{[-(N-1), N-1]} g(y) dy > 1 - \frac{\epsilon}{2}$ . Also let  $c_{i,r}$  and

$d_{i,r}$  be as above and consider that

$$\begin{aligned} & \left| \int_{A \cap [-N, N]} f(x - \mu) - \sum c_{i,r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x) dx \right| \leq \\ & \left| \int_{A \cap [-N, N]} f(x - \mu) - \sum c_{i,2r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x - \mu) dx \right| + \\ & \left| \int_{A \cap [-N, N]} \sum c_{i,2r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x - \mu) - \sum c_{i,2r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x) dx \right| \end{aligned}$$

First term is  $< \frac{2}{3}\epsilon$  second term have

$$\begin{aligned} & \left| \int_{A \cap [-N, N]} \sum c_{i,2r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x - \mu) - \sum c_{i,2r} \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x) dx \right| \leq \\ & \int_{[-N, N]} \left| \sum c_{i,2r} [\mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x - \mu) - \mathbf{1}_{[\frac{i}{2r}, \frac{i+1}{2r}]}(x)] \right| dx = \\ & \int_{[-N, N]} \left| \sum c_{i,2r} [\mathbf{1}_{[\frac{i+1}{2r}, \frac{i+1}{2r} + \mu]}(x) - \mathbf{1}_{[\frac{i}{2r}, \frac{i}{2r} + \mu]}(x)] \right| dx = \\ & \int_{[-N, N]} \sum |c_{i-1,2r} - c_{i,2r}| \mathbf{1}_{[\frac{i}{2r}, \frac{i}{2r} + \mu]}(x) dx = \\ & \sum |c_{i-1,2r} - c_{i,2r}| \mu < \sum |c_{i-1,2r} - c_{i,2r}| \frac{1}{r} \end{aligned}$$

For each term  $|c_{i-1,2r} - c_{i,2r}|$  we have

$$\begin{aligned} |c_{i-1,2r} - c_{i,2r}| &= \frac{1}{2} \left| \left[ \sup_{y \in [\frac{i-1}{2r}, \frac{i}{2r}]} f(y) - \inf_{y \in [\frac{i}{2r}, \frac{i+1}{2r}]} f(y) \right] - \left[ \sup_{y \in [\frac{i}{2r}, \frac{i+1}{2r}]} f(y) - \inf_{y \in [\frac{i-1}{2r}, \frac{i}{2r}]} f(y) \right] \right| \\ &\leq \frac{1}{2} \max \left\{ \left| \sup_{y \in [\frac{i-1}{2r}, \frac{i}{2r}]} f(y) - \inf_{y \in [\frac{i}{2r}, \frac{i+1}{2r}]} f(y) \right|, \left| \sup_{y \in [\frac{i}{2r}, \frac{i+1}{2r}]} f(y) - \inf_{y \in [\frac{i-1}{2r}, \frac{i}{2r}]} f(y) \right| \right\} \\ &\leq \frac{1}{2} \left| \sup_{y \in [\frac{i-1}{2r}, \frac{i-1}{2r} + \frac{1}{r}]} f(y) - \inf_{y \in [\frac{i-1}{2r}, \frac{i-1}{2r} + \frac{1}{r}]} f(y) \right| \end{aligned}$$

Since by construction  $\frac{1}{r} < \delta$  from Equation (B.11) we conclude  $|c_{i-1,2r} - c_{i,2r}| \leq \frac{\epsilon}{6N}$  so that

$$\sum |c_{i-1,2r} - c_{i,2r}| \frac{1}{r} \leq \frac{\epsilon}{6N} 2Nr \frac{1}{r} = \frac{\epsilon}{3}$$

□