# Non-convergent extremes, coupon collecting and computer-based tests

Charles M. Goldie

Mathematics Department
University of Sussex

17th October 2011

Joint work with Rosie Cornish (Univ. of Bristol) & Carol L. Robinson (Loughborough University).

# Contents I

# 1. Extremes I

$Y,\ Y_1,\ Y_2$, i.i.d. $\tilde{F}$.

$$¿\,\exists\ a_n > 0,\ \ b_n \text{ so that } \frac{\max(Y_1, \dots, Y_n) - b_n}{a_n} \Longrightarrow \text{non-degenerate limit?}$$

Necessary for weak convergence (convergence in law) that

$$\frac{F(x)}{F(x-)} \to 1 \text{ as } x \to \infty.$$

So if $Y$ discrete, with probabilities geometrically decaying:

$$\frac{P(Y = k)}{P(Y = k+1)} \to c > 1,$$

weak convergence of $\max(Y_1, \dots, Y_n)$, however centred & normed, can't occur [Anderson, 1970].

# 2. Coupon collecting I

There are $a$ types of coupon. Each cereal packet has one.
$Y :=$ # packets needed to get at least 1 coupon of each type.

$$Y = X_1 + X_2 + \cdots + X_a,$$

$$X_1,\ X_2 \text{ independent},\ X_k \sim \text{Geom}_1\left(\frac{a-k+1}{a}\right),$$

where the $\text{Geom}_1$ law has probabilities $p(1-p)^{k-1}$ at $k = 1, 2, \ldots$.

# 2. Coupon collecting I

## Law of $Y$

Let the coupon types be $1, \ldots, a$. Let

$$A_i := \{\text{type } i \text{ doesn't occur in the first } y \text{ cereal packets bought.}\}$$

So $\{Y > y\} = A_1 \cup A_2 \cup \cdots \cup A_a$, hence

$$\begin{aligned}
&P(Y > y) \\
&= \sum_1^a P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \\
&\qquad \cdots + (-)^{a+1} P(A_1 \cap \cdots \cap A_a) \\
&= \sum_1^a \left(1 - \frac{1}{a}\right)^y - \sum_{i<j}\left(1 - \frac{2}{a}\right)^y + \sum_{i<j<k}\left(1 - \frac{3}{a}\right)^y - \cdots + (-)^{a+1}\left(1 - \frac{a}{a}\right)^y \\
&= \sum_{k=1}^a (-)^{k+1} \binom{a}{k} \left(1 - \frac{k}{a}\right)^y,
\end{aligned}$$

## 2. Coupon collecting II

This formula,

$$P(Y > y) = \sum_{k=1}^{a} (-)^{k+1} \binom{a}{k} \left(1 - \frac{k}{a}\right)^{y},$$

is a classical one for the probability that not all cells are occupied when $y$ balls are distributed at random among $a$ cells.

For large $y$ the $1^{\text{st}}$ term dominates, i.e.

$$P(Y > y) \sim a\left(1 - \frac{1}{a}\right)^{y} \text{ as } y \to \infty \ (y \in \mathbb{N},\ a \text{ fixed}).$$

# 3. Computer-based tests I

Each student takes a test of $q$ questions.

For each question there is a bank of $a$ alternatives.

The computer generates a test by selecting, for each of the $q$ questions, one of the $a$ alternatives for that question.

Let $N_q := \#$ tests one needs to generate to see all $aq$ alternatives in the $q$ question banks at least once.

I fix $a$, for instance $a := 10$, and consider how $N_q$ behaves for various $q$.

The case $q = 1$, i.e. a 1-question test, is coupon-collecting.

Coupon-collecting asymptotics are for $Y = N_1$ as $a \to \infty$, but I'm interested in $N_q$ as $q$ grows, for fixed $a$.

The case considered is coupon collecting when $q$ brands of cereal bought simultaneously, each brand having a different set of $a$ coupons to collect. Therefore

$$N_q = \max(Y_1, \ldots, Y_q)$$

where the $Y_i$ are independent with the coupon-collecting distribution.

# 4. $N_q$ for specific values of $a$ & $q$ I

$$
\begin{aligned}
EN_q &= \sum_{n=0}^{\infty} P(N_q > n) \\
&= \sum_{n=0}^{\infty} \left(1 - \prod_{i=1}^{q} P(Y_i \le n)\right) \\
&= \sum_{n=0}^{\infty} \left(1 - \big(1 - P(Y > n)\big)^q\right).
\end{aligned}
$$

# 4. $N_q$ for specific values of $a$ & $q$ II

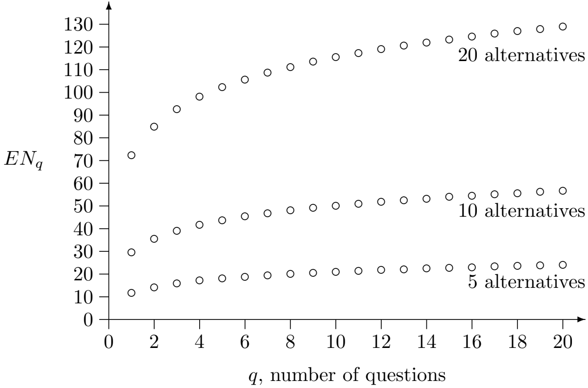


Figure 1: $EN_q$, the expected number of tests that need to be generated in order for all questions to have appeared at least once, for tests with up to 20 questions and 5, 10, and 20 alternatives for each question.

# 4. $N_q$ for specific values of $a$ & $q$ III

Note that in a 20-question test with 5 alternatives for each question, there are $5^{20} = 95\,367\,431\,640\,625$ different possible tests and a total bank of 100 questions; however, on average all questions will have appeared at least once by the time only 24 tests have been generated.

# 4. $N_q$ for specific values of $a$ & $q$ I

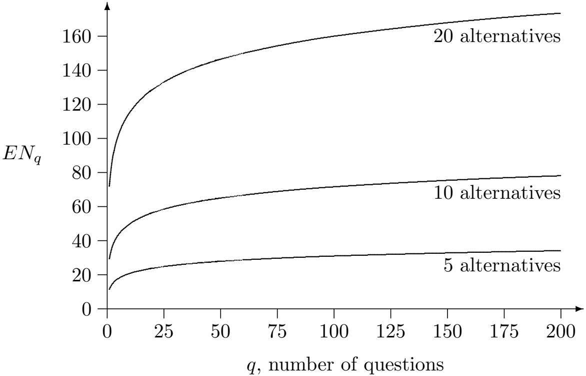


Figure 2: $EN_q$, the expected number of tests that need to be generated in order for all questions to have appeared at least once, for tests with up to 200 questions and 5, 10, and 20 alternatives for each question.

## 5. $N_q$ for $q$ large I

Set $\alpha := \ln \frac{a}{a-1} > 0$, then I showed

$$P(Y > y) \sim ae^{-\alpha y} \text{ as } y \to \infty, \;\; y \in \mathbb{N}.$$

Ignoring the restriction to $\mathbb{N}$,

$$\begin{aligned} P\left(N_q - \frac{\ln(aq)}{\alpha} \le x\right) &= \left(P\left(Y \le \frac{\ln(aq)}{\alpha} + x\right)\right)^q \\ &= \left(1 - ae^{-\ln(aq) - \alpha x}(1 + o(1))\right)^q \\ &= \left(1 - \frac{e^{-\alpha x}(1 + o(1))}{q}\right)^q \to e^{-e^{-\alpha x}} = \Lambda(\alpha x), \end{aligned}$$

where $\Lambda(x) := e^{-e^{-x}}$ is the Gumbel distribution function.

**Theorem 1.**
*With* $b_q := \frac{1}{\alpha}\ln(aq)$,

$$\liminf_{q\to\infty} P(N_q - b_q \le x) = \Lambda\big(\alpha(x-1)\big);$$

$$\limsup_{q\to\infty} P(N_q - b_q \le x) = \Lambda(\alpha x).$$

# 5. $N_q$ for $q$ large II

Thus $N_q - b_q$ is, asymptotically, in distribution between $\frac{Z}{\alpha}$ and $\frac{Z}{\alpha} + 1$ where $Z$ Gumbel, and the bounds are sharp.

Let $\lfloor x \rfloor$ denote the integer part, $\{x\} := x - \lfloor x \rfloor$ the fractional part, of $x$.

**Theorem 2 (extending [Anderson, 1980, Ferguson, 1993]).**

$$P(N_q - b_q = n + 1 - \{b_q\}) = P\left(\frac{Z}{\alpha} \leq n + 1 - \{b_q\}\right) - P\left(\frac{Z}{\alpha} \leq n - \{b_q\}\right) + o_n(1),$$

*where* $\sum_{n \in \mathbb{Z}} o_n(1) \to 0$ *as* $q \to \infty$.

# 6. $L_p$-boundedness I

**Theorem 3.**
*$N_q - b_q$ is $L_p$-bounded for all $p$, i.e.* $\sup_{q\in\mathbb{N}} E\big(|N_q - b_q|^p\big) < \infty$ *for all* $p \geq 1$.

Proof.
Fix $n \in \mathbb{N}$; set $R_q := N_q - b_q$. I prove $\sup_q E(R_q^{2n}) < \infty$, which suffices. Now

$$E(R_q^{2n}) = -2n\int_{-\infty}^{0} x^{2n-1}P(R_q \leq x)\,dx + 2n\int_{0}^{\infty} x^{2n-1}P(R_q > x)\,dx =: A+B.$$

For $B$, show

$$P(Y > x + b_q) \leq \frac{2}{q}e^{\alpha-\alpha x} \quad \forall x \geq 0,\; q \geq q_0;$$

$$\therefore\; P(R_q > x) \leq 1 - \left(1 - \frac{2}{q}e^{\alpha-\alpha x}\right)^q \leq 4e^{\alpha-\alpha x} \quad \forall\, x \geq 0,\; q \geq q_1;$$

$$\therefore\; B \leq 8n\int_0^\infty x^{2n-1}e^{\alpha-\alpha x}\,dx < \infty.$$

For $A$, adapt a split-and-bound technique from [Resnick, 1987]. □

# 7. Mean growth I

$L_p$-boundedness implies asymptotic bounds on moments. Recall

$$\frac{Z}{\alpha} \le N_q - b_q \le \frac{Z}{\alpha} + 1 \quad \text{asymptotically},$$

and $EZ = \gamma \doteq 0{\cdot}5772$.

**Theorem 4.**

$$\frac{\gamma}{\alpha} \le \lim_{q\to\infty} \sup/\inf (EN_q - b_q) \le \frac{\gamma}{\alpha} + 1.$$

# 7. Mean growth I

| $q$ | 1 | 10 | 100 | 1000 |
|---|---|---|---|---|
| $EN_q$ | 29·29 | 49·90 | 71·57 | 93·40 |
| $b_q + \gamma/\alpha$ | 27·33 | 49·19 | 71·04 | 92·90 |
| excess | 1·956855 | 0·715025 | 0·527514 | 0·503224 |

| $q$ | 10 000 | $10^5$ | $10^6$ | $10^7$ |
|---|---|---|---|---|
| $EN_q$ | 115·25 | 137·10 | 158·96 | 180·81 |
| $b_q + \gamma/\alpha$ | 114·75 | 136·60 | 158·46 | 180·31 |
| excess | 0·500358 | 0·500039 | 0·500004 | 0·500000 |

Table 1: For $a = 10$, values of $EN_q$, its approximant $b_q + \gamma/\alpha$, and the excess $EN_q - (b_q + \gamma/\alpha)$.

**Conjecture.**

*As* $q \to \infty$, $EN_q - b_q - \frac{\gamma}{\alpha} \to$ *limit, maybe* 0·5.

# 8. Variance stability I

**Lemma.**

$$
\begin{aligned}
&E\Big(\Big(1+\frac{Z}{\alpha}\Big)^2\mathbf{1}_{1+\alpha^{-1}Z\le 0}+\Big(\frac{Z}{\alpha}\Big)^2\mathbf{1}_{Z>0}\Big)\\
&\quad\le \lim_{q\to\infty}\sup/\inf E\big((N_q-b_q)^2\big)\\
&\qquad\le E\Big(\Big(\frac{Z}{\alpha}\Big)^2\mathbf{1}_{Z\le 0}+\Big(1+\frac{Z}{\alpha}\Big)^2\mathbf{1}_{1+\alpha^{-1}Z>0}\Big)
\end{aligned}
$$

## 8. Variance stability I

Note var $Z = \pi^2/6$, so without discreteness we'd get var $N_q \to \frac{\pi^2}{6\alpha^2}$.

**Theorem 5.**

$$\limsup_{q\to\infty}\left|\operatorname{var} N_q - \frac{\pi^2}{6\alpha^2}\right| \le \theta(\alpha) + 1 - \frac{1}{e} + \frac{2(\gamma + E_1(1))}{\alpha},$$

*where*

$$\theta(\alpha) = E\left(\left(1+\frac{Z}{\alpha}\right)^2 \mathbf{1}_{0<1+\alpha^{-1}Z\le 1}\right) \in (0,1),$$

$$E_1(1) = \int_1^\infty \frac{e^{-t}}{t}\,dt \doteq 0{\cdot}2194.$$

| $a$ | 2 | 3 | 4 | 5 | 10 | 20 |
|---|---|---|---|---|---|---|
| sd($N_q$) | 1·873 | 3·176 | 4·468 | 5·755 | 12·176 | 25·006 |
| $\pi/(\alpha\sqrt{6})$ | 1·850 | 3·163 | 4·458 | 5·748 | 12·173 | 25·004 |
| Min s.d. | 0·641 | 2·323 | 3·697 | 5·024 | 11·507 | 24·362 |
| Max s.d. | 2·537 | 3·823 | 5·107 | 6·390 | 12·804 | 25·630 |

Table 2: Asymptotic standard deviation of $N_q$, its approximant, and bounds.

# 8. Variance stability II

**Conjecture.**
*As* $q \to \infty$, $\operatorname{var} N_q \to$ *limit.*

# References I

[Anderson, 1970] Anderson, C. W. 1970. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probab.*, **7**, 99–113.

[Anderson, 1980] Anderson, C. W. 1980. Local limit theory for the maxima of discrete random variables. *Math. Proc. Cambridge Philos. Soc.* **88**, 161–165.

[Goldie, Cornish & Robinson, 2010] C. M. Goldie, R. Cornish and C. L. Robinson, 'Applying coupon-collecting theory to computer-aided assessments'. In: Bingham, N. H., and Goldie, C. M. (eds), *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, pp. 299–318. London Math. Soc. Lecture Note Series **378**. Cambridge: Cambridge Univ. Press.

[Ferguson, 1993] T. S. Ferguson, 'On the asymptotic distribution of max and mex'. *Statistische Hefte* **34**, 97–111.

[Resnick, 1987] Resnick, S. I. 1987. *Extreme Values, Regular Variation, and Point Processes*. New York: Springer-Verlag.