Exercises of Random Variables

Exercise

- Show that the necessary and suficient condition for a random variable on *N* to have a ***geometric distribution*** is that it should have the property:

$$P(X > n+m / X > m) = P(X > n)$$

  – For each natural number **n** and **m**.

geometric distribution

- Random variable that models the number of trials until a success or failure.
- requirements :
  - number of trials is potentially infinite
  - two outcomes per trial; success and failure
  - outcomes statistically independent
  - trials have the same probability of success

$$P(X = i) = p^{i-1}(1-p) \quad \text{for } i = 1,2,3,\cdots$$

Exercise

- Meaning of:

$$P(X > n+m \,/\, X > m) = P(X > n)$$

- Probability of waiting **n** minuts more given that you have waited **m** is independent of **m**.
  - Applications:
    - Queue at the bus stop (Relate to Poison rv)
    - Queue at a hub or a relay (is the model correct?)
    - Expected survival time
      - Illness, or protocol design.

Like its continuous analogu
(the exponential distribution
the geometric distribution is
memoryless.

Exercise

- Property to be shown:

$$P(X > n+m / X > m) = P(X > n)$$

- Definition: Geometric Random Variable:

$$P(X = i) = p^{i-1}(1-p) \quad \text{for } i = 1,2,3,\cdots$$

- The distribution function is

$$P(X > n) = (1-p)\sum_{i=n+1}^{\infty} p^{i-1} \underset{\text{k=i−(n+1)}}{=\text{c.v.}} (1-p)\sum_{k=0}^{\infty} p^{k+n} = (1-p)\, p^n \sum_{k=0}^{\infty} p^k \underset{\text{Series}}{=\text{Geometric}}$$

$$= (1-p)\, p^n \frac{1}{1-p} = p^n$$

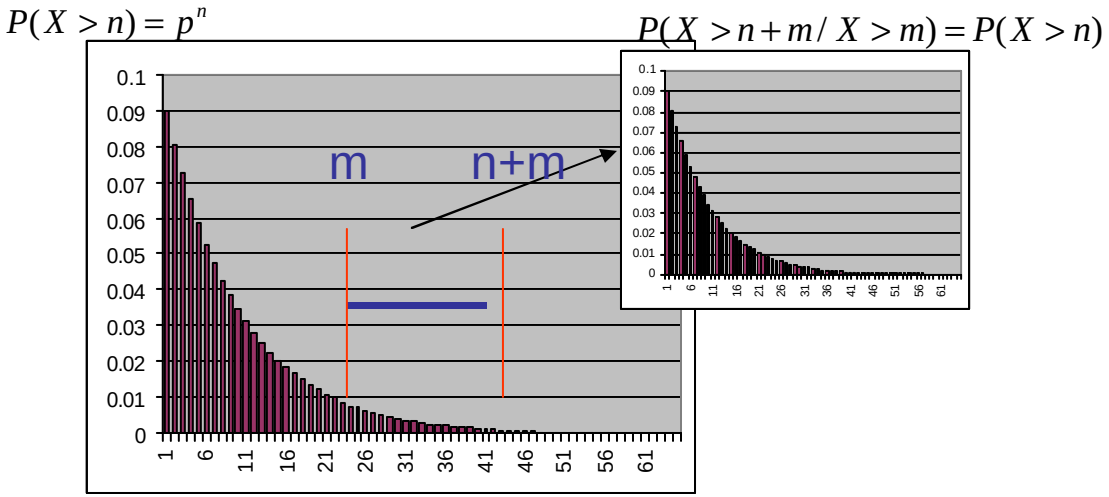
$$P(X > n) = p^n$$

Exercise

- If *A* then *B*:

$$P(X > n) = p^n$$

$$P(X > n+m / X > m) = \frac{P(X > n+m)}{P(X > m)} = \frac{p^{n+m}}{p^m} = p^n = P(X > n)$$



Exercise

- On the other hand If B then A:

$$P(X > n+m / X > m) = \frac{P(X > n+m)}{P(X > m)} = P(X > n)$$

Suppose that $P(X > n) = a_n$ has the property $\frac{a_{n+m}}{a_m} = a_n$

then $a_{n+m} = a_n a_m$ and $a_m = a_1 a_{m-1} = a_1^{\,2} a_{m-2} \cdots = a_1^{\,m}$

then $P(X = m) = P(X > m-1) - P(X > m) = a_1^{\,m-1} - a_1^{\,m} = a_1^{\,m-1}(1-a_1)$

Example of a rv with memory

- A ***Pareto distribution*** when used to model a queue has ***memory***:

$$P(X > n+m / X > m) > P(X > n)$$

  - For each natural number **n** and **m**.
  - Meaning:
    - Probability of waiting n minuts more given that you have **waited m is greater than at the arrival**.
    - Richer get richer: "80-20 rule" which says that 20% of the population owns 80% of the wealth.
    - The more you wait, the more you are expected to wait

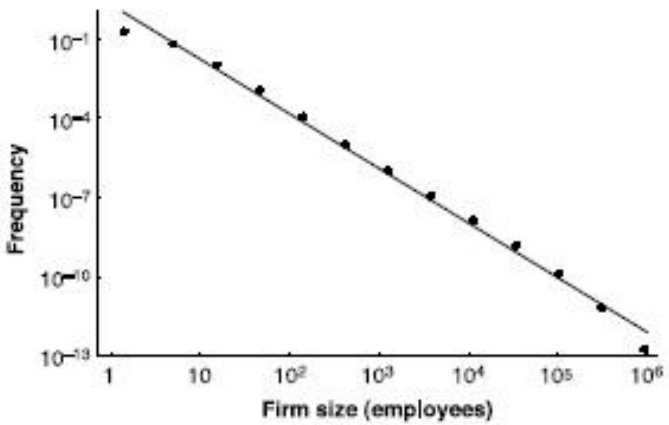
Example of a rv with memory

- Examples of uses of the Pareto Distribution:
  - * Frequencies of words in longer texts (a few words are used often, lots of words are used infrequently)
  - * The sizes of human settlements (few cities, many hamlets/villages)
  - * File size distribution of Internet traffic which uses the TCP protocol (many smaller files, few larger ones)
  - * Clusters of Bose-Einstein condensate near absolute zero
  - * The values of oil reserves in oil fields (a few large fields, many small fields)
  - * The length distribution in jobs assigned supercomputers (a few large ones, many small ones)
  - * The standardized price returns on individual stocks
  - * Sizes of sand particles
  - * Sizes of meteorites
  - * Numbers of species per genus (There is subjectivity involved: The tendency to divide a genus into two or more increases with the number of species in it)
  - * Areas burnt in forest fires

Cities and firms

- Zipf distribution of U.S. firm sizes

$$\Pr[S \geq s_i] = \left(\frac{s_0}{s_i}\right)^{\alpha}, \; s_i \geq s_0, \; \alpha > 0$$



REPORTS

Fig. 1. Histogram of U.S. firm sizes, by employees. Data are for 1997 from the U.S. Census Bureau, tabulated in bins having width increasing in powers of three (*30*). The solid line is the OLS regression line through the data, and it has a slope of 2.059 (SE = 0.054; adjusted $R^2 = 0.992$), meaning that $\alpha = 1.059$; maximum likelihood and nonparametric methods yield similar results. The data are slightly concave to the origin in log-log coordinates, reflecting finite size cutoffs at the limits of very small and very large firms.

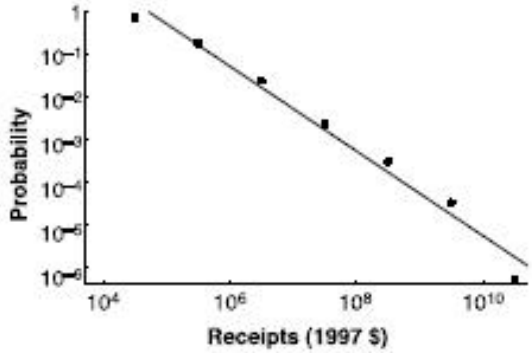
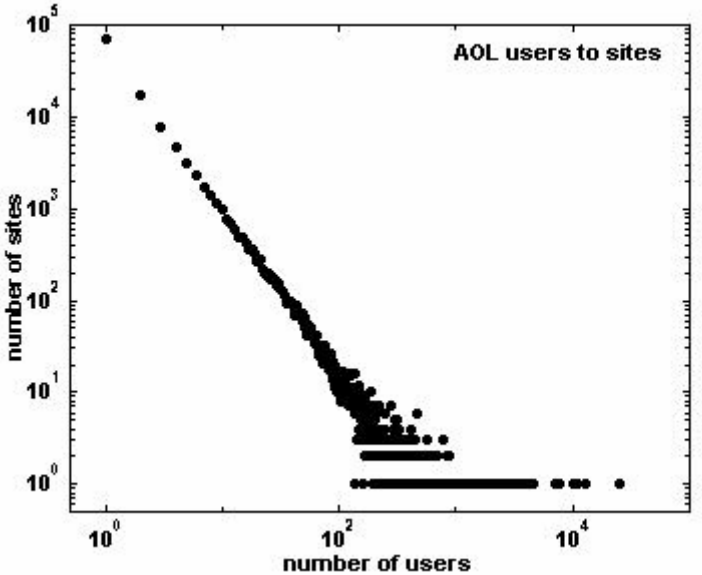


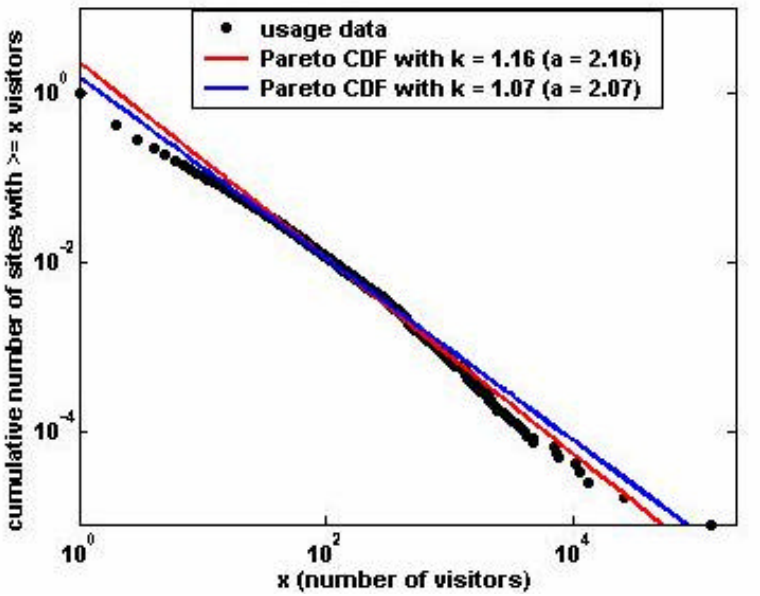
Fig. 2. Tail cumulative distribution function of U.S. firm sizes, by receipts in dollars. Data are for 1997 from the U.S. Census Bureau, tabulated in bins whose width increases in powers of 10. The solid line is the OLS regression line through the data and has slope of 0.994 (SE = 0.064; adjusted $R^2 = 0.976$).

Axtell, R. L. (2001), "Zipf distribution of U.S. firm sizes", Science

Web sites visits

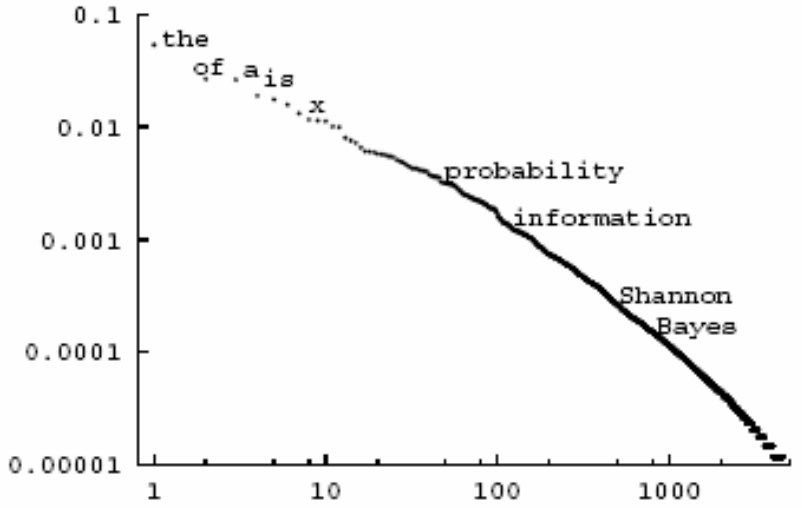
- Distribution of AOL users' visits to various sites on a December day in 1997





Zipf, Power-laws, and Pareto - a ranking tutorial
Lada A. Adamic
Comments from B.A. Huberman

Word frequencies in a text



Speculative Prices

- Mandelbrot's paper on long tail densities

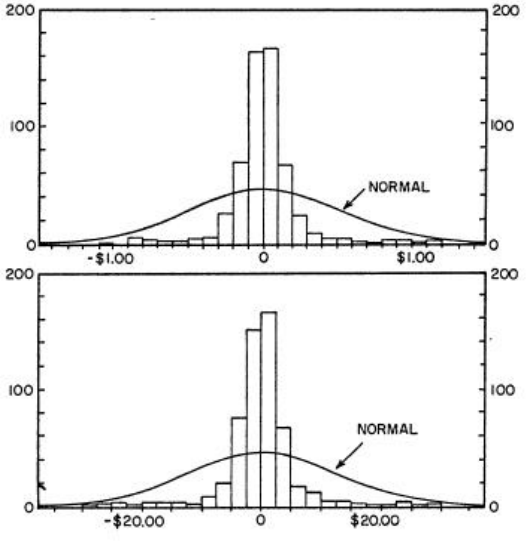


FIG. 1.—Two histograms illustrating departure from normality of the fifth and tenth difference of monthly wool prices, 1890–1937. In each case, the continuous bell-shaped curve represents the Gaussian "interpolate" based upon the sample variance. Source: Gerhard Tintner, *The Variate-Difference Method* (Bloomington, Ind., 1940).

e Variation of Certain Speculative Prices

Benoit Mandelbrot

The Journal of Business, Vol. 36, No. 4. (Oct., 1963), pp. 394-419.

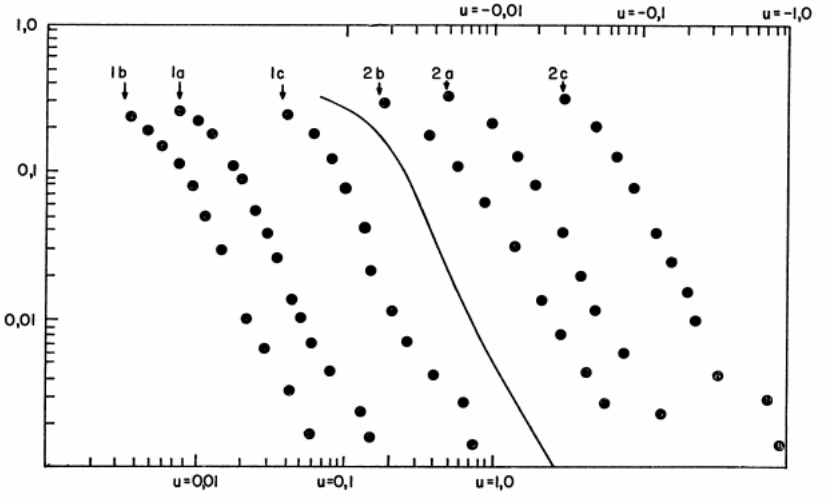


FIG. 5.—Composite of doubly logarithmic graphs of positive and negative tails for three kinds of cotton price relatives, together with cumulated density function of a stable distribution. Horizontal scale u of lines *1a*, *1b*, and *1c* is marked only on lower edge, and horizontal scale u of lines *2a*, *2b*, and *2c* is marked along upper edge. Vertical scale gives the following relative frequencies: (*1a*) $Fr[\log_e Z(t + \text{one day}) - \log_e Z(t) > u]$, (*2a*) $Fr[\log_e Z(t + \text{one day}) - \log_e Z(t) < -u]$, both for the daily closing prices of cotton in New York, 1900–1905 (source: private communication from the United States Department of Agriculture).

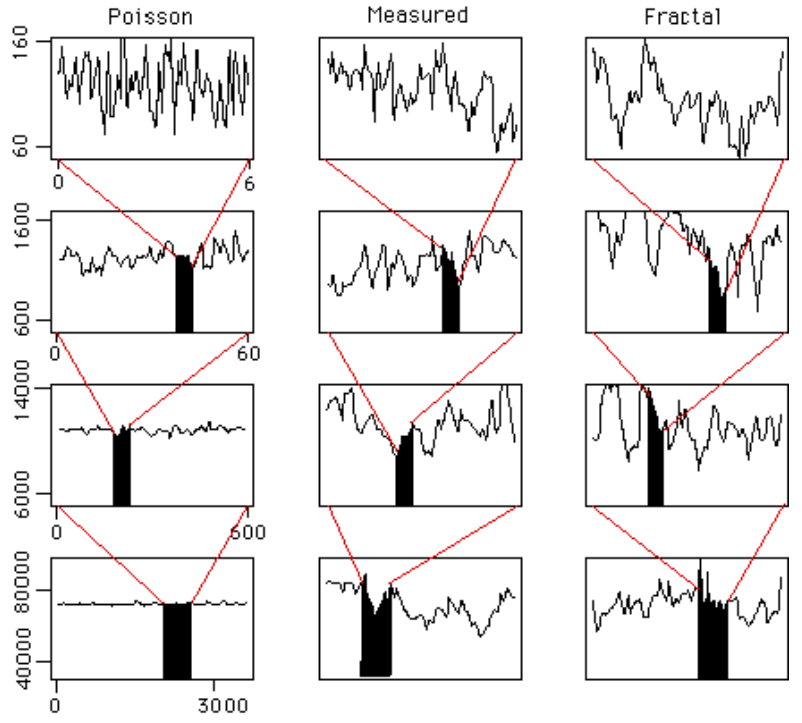
(*1b*) $Fr[\log_e Z(t + \text{one day}) - \log_e Z(t) > u]$, (*2b*) $Fr[\log_e Z(t + \text{one day}) - \log_e Z(t) < -u]$, both for an index of daily closing prices of cotton in the United States, 1944–58 (source: private communication from Hendrik S. Houthakker).

(*1c*) $Fr[\log_e Z(t + \text{one month}) - \log_e Z(t) > u]$, (*2c*) $Fr[\log_e Z(t + \text{one month}) - \log_e Z(t) < -u]$, both for the closing prices of cotton on the 15th of each month in New York, 1880–1940 (source: private communication from the United States Department of Agriculture).

The reader is advised to copy on a transparency the horizontal axis and the theoretical distribution and to move both horizontally until the theoretical curve is superimposed on either of the empirical graphs; the only discrepancy is observed for line *2b;* it is slight and would imply an even greater departure from normality.

Speculative Prices

- Mandelbrot's paper on long tail densities
  - An interesting result



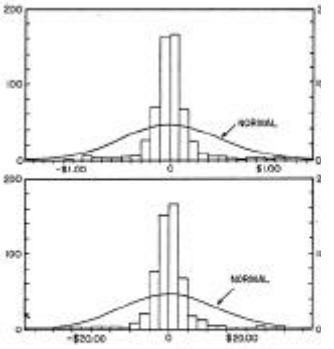


FIG. 1.—Two histograms illustrating departure from normality of the fifth and tenth difference of monthly wool prices, 1890–1937. In each case, the continuous bell-shaped curve represents the Gaussian "interpolate" based upon the sample variance. Source: Gerhard Tintner, *The Variate-Difference Method* (Bloomington, Ind., 1940).

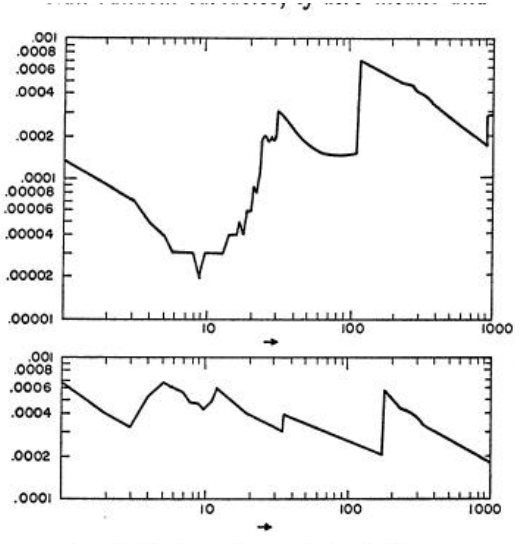


FIG. 2.—Both graphs are relative to the sequential sample second moment of cotton price changes. Horizontal scale represents time in days, with two different origins $T°$: on the upper graph, $T°$ was September 21, 1900; on the lower graph $T°$ was August 1, 1900. Vertical lines represent the value of the function

$$(T-T°)^{-1}\sum_{t=T°}^{t=T}[L(t,1)]^2,$$

where $L(t, 1) = \log_e Z(t+1) - \log_e Z(t)$ and $Z(t)$ is the closing spot price of cotton on day t, as privately reported by the United States Department of Agriculture.

http://classes.yale.edu/fractals/Panorama/ManuFractals/Internet/Internet4.html

Burstiness property

- Burstiness in cities & internet trafic

The image below (composed of several satellite pictures) gives an idea of the degree of economic agglomeration in the world economy.

Analisys of the Pareto distribution

- We will compute the value:

$$P(X > n+m / X > n)$$

- Remember the definition:

$$P(X > m) = \left(\frac{m_0}{m} \right)^a \quad \text{with } a > 0$$

- The conditioned probability is:

$$P(X > n+m / X > m) = \frac{P(X > n+m)}{P(X > m)}$$

Analisys of the Pareto distribution

- We will compute the value:

$$P(X > n+m / X > m) = \frac{P(X > n+m)}{P(X > m)}$$

- The conditioned probability is:

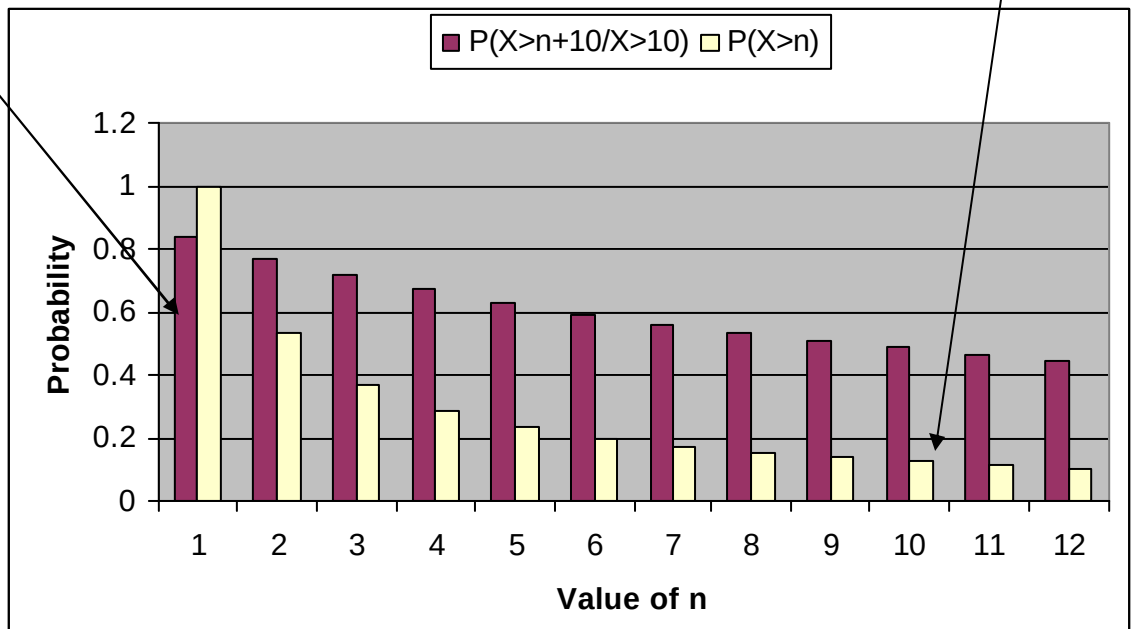
$$P(X > n+m / X > m) = \frac{\left(\frac{n_0+m_0}{n+m}\right)^a}{\left(\frac{m_0}{m}\right)^a} = \left(\frac{m_0}{n_0+m_0}\right)^a \left(\frac{m}{n+m}\right)^a$$

$$\left(\frac{m_0}{n_0+m_0}\right)^a \left(\frac{m}{n+m}\right)^a > P(X > n) = \left(\frac{n_0}{n}\right)^a$$

Analisys of the Pareto distribution

- Simulation: $m_0 = 10 \quad \text{and} \quad n_0 = 1$
  - Message: the longer you wait, the more you will wait

$$(X > n+m / X > m) = \left(\frac{m_0}{n_0 + m_0}\right)^a \left(\frac{m}{n+m}\right)^a \qquad P(X > n) = \left(\frac{n_0}{n}\right)^a$$



Negative Binomial distribution

- Generalization of a Geometric distribution:
- Def. Probability of ***r*** successes in ***n*** Bernouilli trials. Trials independent and identically distributed.

r=1 $$\left\{ \underbrace{HHH\cdots HHHH}_{N-1}\,T \right\} \rightarrow p(1-p)^{N-1} = \binom{N-1}{0} p(1-p)^{N-1}$$

r=2 $$\left\{ \begin{matrix} THH\cdots HHHHT \\ \ddots \\ HHH\cdots HHHTT \end{matrix} \right\} \rightarrow (N-1)\,p^2(1-p)^{N-2} = \binom{N-1}{1} p^2(1-p)^{N-2}$$

General case

r $$\left\{ \begin{matrix} THH\cdots THTHH\;\; T \\ \binom{N-1}{r-1} \;\; \vdots \\ \cdots \qquad\qquad\quad T \end{matrix} \right\} \rightarrow \binom{N-1}{r-1} p^r(1-p)^{N-r}$$

Negative Binomial distribution

- General expression:
  - Probability of ***r*** successes in ***n*** Bernouilli trials. Trials independent and identically distributed.

$$P(X = r) = \binom{N-1}{r-1} p^r (1-p)^{N-r}$$

    - Examples:
      - Disk redundancies
      - Coding theory. Error correction
      - Banach Matches.

Banach's Matches

- **Example**

A pipe-smoking mathematician carries, at all times, 2 matchboxes, 1 in his left-hand pocket and 1 in his right-hand pocket. Each time he needs a match he is equally likely to take it from either pocket. Consider the moment when the mathematician first discovers that one of his matchboxes is empty. If it is assumed that both matchboxes initially contained N matches, what is the probability that there are exactly k matches in he other box, k = 0, 1, ...,N?

See Feller

Banach's Matches

- Note that it is a negative binomial, at least must have N+1 successes in one of the boxes.
- The success number *(N+1)* occurs at the *(N+1)+(N-k)=2N-k* trial.

$$\text{Prob}(k) = 2P(X = (N+1)) = 2\binom{2N-k}{N} p^{N+1}(1-p)^{N-k}$$

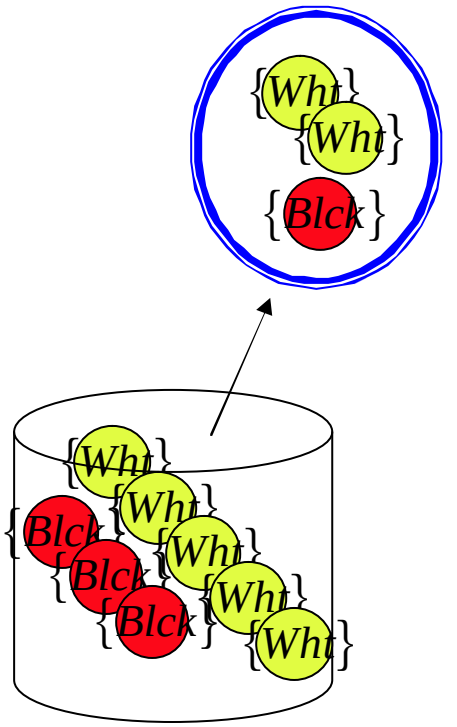
Banach's Matches

- Applications:
  - Allocations of files in a disk system.
  - Heap management.

$$\text{Prob}(k) = 2P(X = (N+1)) = 2\binom{2N-k}{N} p^{N+1}(1-p)^{N-k}$$

Hypergeometric Random

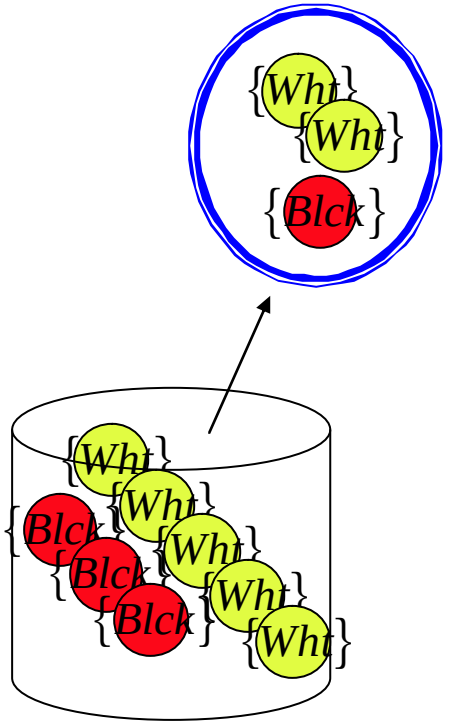
- Models the number of successes **k** in a sequence of **n** draws from a finite population ***without*** replacement.
  - Size of the population: ***m***
  - Observed successes: ***k***
  - Favorable objects: ***r***
  - Number of draws: ***n***



Hypergeometric Random

- Random Variable Y=k
  - Size of the population: ***m***
  - Observed successes: ***k***
  - Favorable objects: ***r***
  - Number of draws: ***n***

$$P(Y=k)=\frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}$$



Application: capture-recapture problem

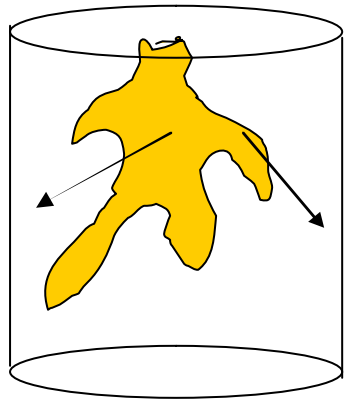
- Lake containing ***m*** fish where m is unknown. We capture ***r*** of the fish, tag them, and return them to the lake.
- Next we capture ***n*** of the fish and observe ***Y***, the number of tagged fish in the sample.

$$\frac{Y}{n} = \frac{r}{m}$$

Size of the population: ***m***
Observed successes: ***k***
Favorable objects: ***r***
Number of draws: ***n***

Application: capture-recapture problem

- Caveat:
  - Diffusion problem



$$P(Y=k)=\frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}$$

$\frac{Y}{n}=\frac{r}{m}$ takes for granted that the observed value is the mean

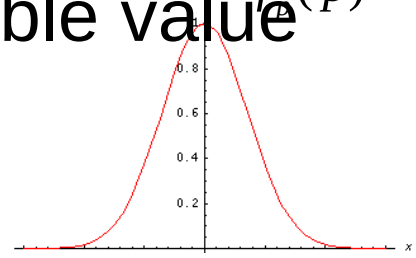
Application: capture-recapture problem

- Caveat:

$\frac{Y}{n} = \frac{r}{m}$ takes for granted that the observed value is the mean

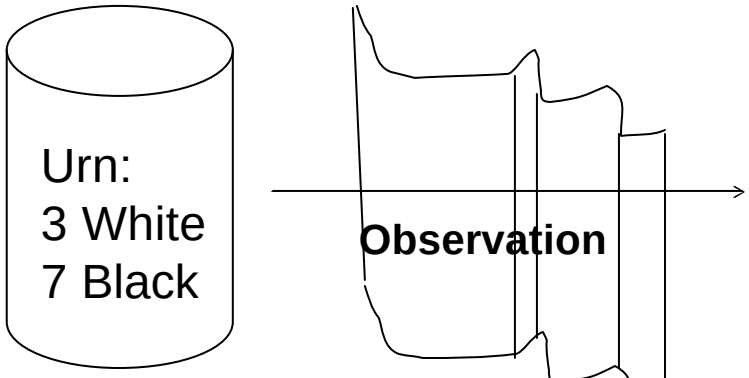
– Variability arround the most probable value

$f_{\hat{p}}(\hat{p})$



p^*

$p^* = P(white\ observation / composition\ of\ the\ urn)$



$P(composition\ of\ the\ urn / white\ observation)$

Example

- A computer cluster of 24 machines, at a given moment has 3 with high load processes. What is the probability of getting ***k*** loaded machines if ***5*** are selected at random?

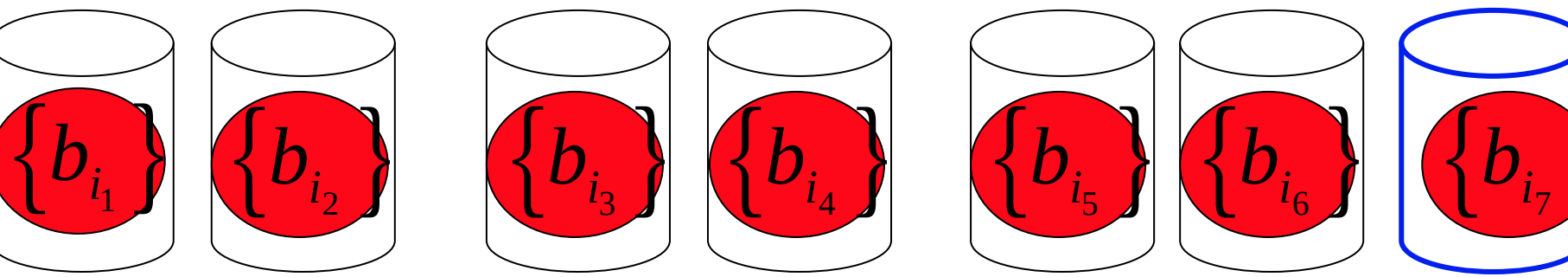
$$P(Y=k)=\frac{\binom{3}{k}\binom{21}{5-k}}{\binom{24}{5}} \qquad P(Y=0)=\frac{\binom{3}{0}\binom{21}{5}}{\binom{24}{5}}=\frac{19*18*17}{24*23*22}=0.478$$

!

Combinatorial Methods.Lotto6/49

- Lotto6/49: 6 numbers+ 1 complementary are selected from 49. A multiple bet means ***selecting r*** from the 49 numbers
  - Probability of guessing k from the winning combination.
  - Probability of guessing k AND the complentary
  - Probability of guessing k AND Not the complentary

$$\{b_1, b_2, b_3, \cdots, b_{48}, b_{49}\} \rightarrow \{i_1, i_2, \cdots, i_7\} \rightarrow \{b_{i_1}, b_{i_2}, \cdots, b_{i_7}\}$$



Example taken from VÉLEZ , HERNÁNDEZ, Cálculo de Probabilidades

Combinatorial Methods.Lotto6/49

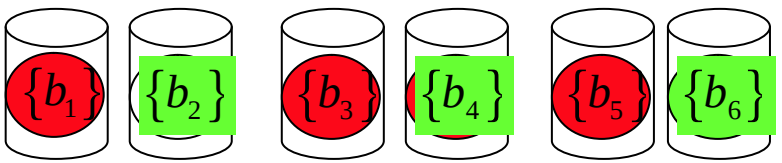
- Number of ways for guessing n results.

Different sets with the winning numbers

$$\binom{r}{k}\binom{49-r}{6-k}$$

Different sets with the non-selected winning numbers.

$$\{i_1, i_2, \cdots, i_k\}$$



- Probability of guessing ***k***

$$\Pr(n) = \frac{\binom{r}{k}\binom{49-r}{6-k}}{\binom{49}{6}}$$

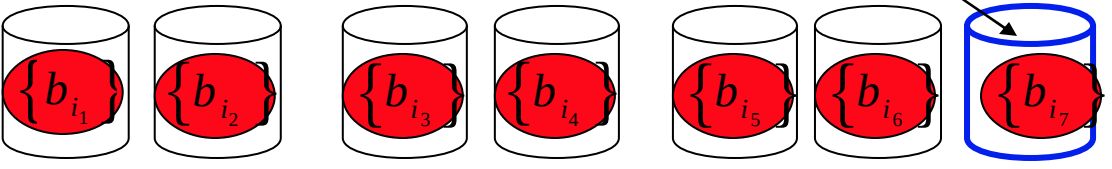
Combinatorial Methods.Lotto6/49

- Probability of guessing k AND the complentary

Different sets with the non-selected winning numbers.

The complementary can be any of the remaining ***r-k***

$$\Pr(n) = \frac{\binom{r}{k}\binom{49-r}{6-k}\binom{r-k}{1}}{\binom{49}{6}43} = \frac{\binom{r}{k}\binom{49-r}{6-k}}{\binom{49}{6}43}(r-k)$$

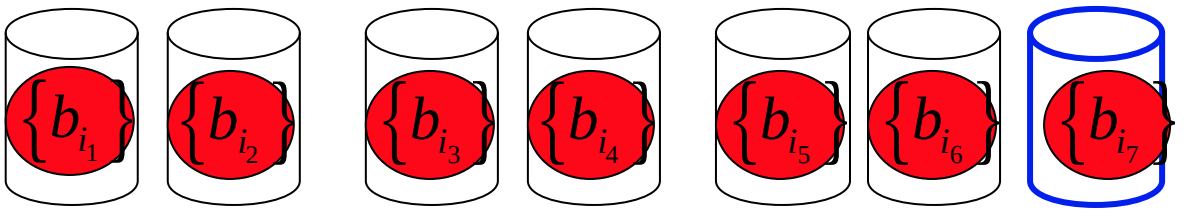


Combinatorial Methods.Lotto6/49

- Probability of guessing k AND NOT the complentary

$$\Pr(n) = \frac{\binom{r}{k}\binom{49-r}{6-k}\binom{49-(r+6-k)}{1}}{\binom{49}{6}43} = \frac{\binom{r}{k}\binom{49-r}{6-k}}{\binom{49}{6}43}\left(43-(r-k)\right)$$

Compementary cannot be
- in the marked ***r***,
- nor in ***(6-k)*** non-marked but winner numbers.



Binomial Random Variables

- Most important discrete probability distribution.
- Model:
  - Two possible outcomes: Success/Failure
  - Probabilities: Success=***p*** / Failure=***1-p***
  - We compound n independent Bernouilli trials.
  - **Define** the random variable:

X=Total number of successes in ***n*** indep. Bernouilli trials

Binomial Random Variables

- Distribution.

X=Total number of successes in ***n*** indep. Bernouilli trials

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad k=0,1,2,3\cdots,n$$

$$THH\cdots HTHHH$$
$$HTH\cdots HHHTH$$
$$\vdots$$
$$\underbrace{HHH\cdots HHTTT}_{k \text{ Successes in } n \text{ trials}}$$

- Model:
  - Two possible outcomes: Success/Failure
  - Probabilities: Success=***p*** / Failure=***1-p***
  - We compound n independent Bernouilli trials.

Binomial Random Variables Example

- Overbooking:
  - An aircraft has a capacity of 150 tickets. The airline management sells 160 tickets in order to protect themselves against no-show passengers.
  - Experience shows that the probability of a passenger being a no-show is of 0.1. The booked passengers act independengly of each other.
  - Given this overbooking strategy, what is the probability that some passengers will be left out?.

Taken from H.Tijms, understanding probability

Binomial Random Variables Example

- Overbooking:
  - The problem can be seen as 160 independent trials of a bernouilli experiment with a success rate of 9/10, where a passenger who shows up for the flight is counted as a success.
  - We define ***X***=number of passengers that show up.
  - X is binomially distributed with parameters n=160, and p=9/10.
  - The probability is P(***X***>150)

$$\underbrace{THHTTHT\cdots THHTTT}_{\text{more than 150 Successes in 160 trials}}$$

$$P(X>150)=\sum_{k=151}^{n}\binom{n}{k}p^{k}(1-p)^{n-k}=0.0359$$

Taken from H.Tijms, understanding probability