Dealing with word frequency lists

(un)reliable frequency counts

Serge Sharoff

Centre for Translation Studies
University of Leeds

16 April 2018

Outline

Frequency of linguistic data

First attempts Friedrich Wilhelm Kaeding (1843-1928)
Andrey Markov (1856-1922)
George Kingsley Zipf (1902–1950)

Frequency of linguistic data

First attempts Friedrich Wilhelm Kaeding (1843-1928)
Andrey Markov (1856-1922)
George Kingsley Zipf (1902–1950)

English lists General Service List (West, 1953)
Brown Corpus (Kučera, Francis, 1967)
BNC list (Kilgarriff, 1997)

Frequency of linguistic data

First attempts Friedrich Wilhelm Kaeding (1843-1928)
Andrey Markov (1856-1922)
George Kingsley Zipf (1902–1950)

English lists General Service List (West, 1953)
Brown Corpus (Kučera, Francis, 1967)
BNC list (Kilgarriff, 1997)

Statistics in linguistics Harald Baayen: frequency distributions
Stefan Gries: frequency dispersions Adam Kilgarriff: frequency lists

Applications of frequency research

- How many words are there in a language?

Applications of frequency research

- How many words are there in a language?

? BNC: 665,238; ukWac: 10,937,392 lemmas

Applications of frequency research

- How many words are there in a language?
- ? BNC: 665,238; ukWac: 10,937,392 lemmas
- ? *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch*; *environmentally-friendly*

Applications of frequency research

- How many words are there in a language?
- ? BNC: 665,238; ukWac: 10,937,392 lemmas
- ? *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch*; *environmentally-friendly*
- Detection of core lexicon
  Corpus as random Library metaphor (Stefan Evert)

Applications of frequency research

- How many words are there in a language?
- ? BNC: 665,238; ukWac: 10,937,392 lemmas
- ? *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch*; *environmentally-friendly*
- Detection of core lexicon
  Corpus as random Library metaphor (Stefan Evert)
- Language Technology issues:
  Estimation of priors for probabilities
  More likely words for Machine Translation

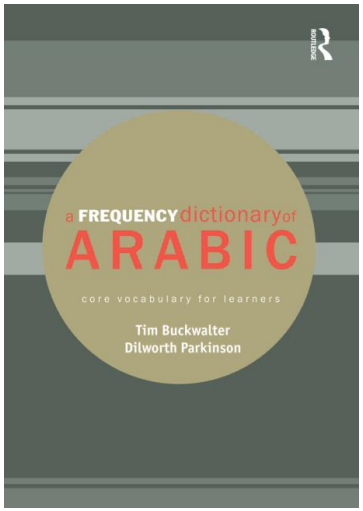
Applications of frequency research

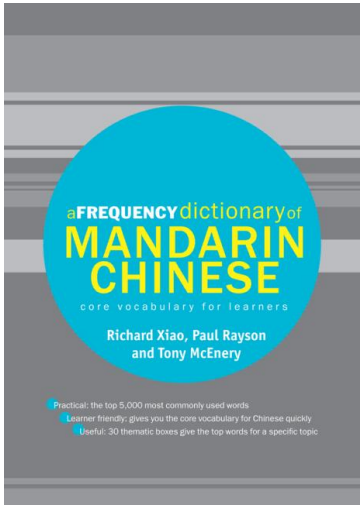
- How many words are there in a language?
- ? BNC: 665,238; ukWac: 10,937,392 lemmas
- ? *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch*; *environmentally-friendly*
- Detection of core lexicon
  Corpus as random Library metaphor (Stefan Evert)
- Language Technology issues:
  Estimation of priors for probabilities
  More likely words for Machine Translation
- Lexicography: dictionary coverage

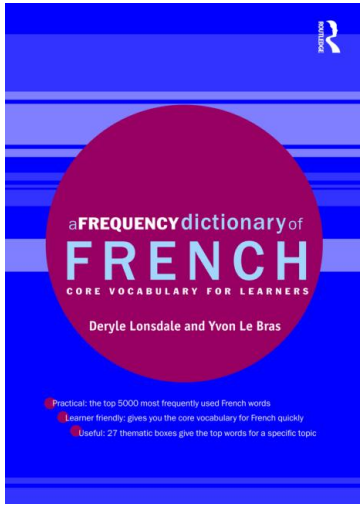
Applications of frequency research

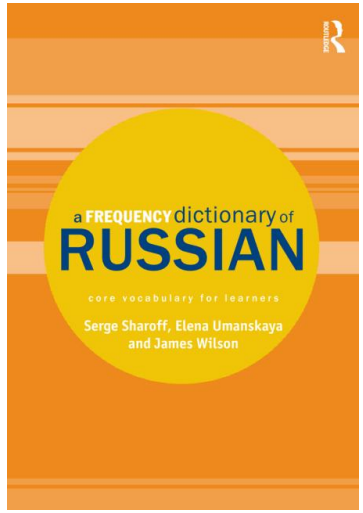
- How many words are there in a language?
- ? BNC: 665,238; ukWac: 10,937,392 lemmas
- ? *Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch*; *environmentally-friendly*
- Detection of core lexicon
  Corpus as random Library metaphor (Stefan Evert)
- Language Technology issues:
  Estimation of priors for probabilities
  More likely words for Machine Translation
- Lexicography: dictionary coverage
- Language learning:
  which words to teach at what stage

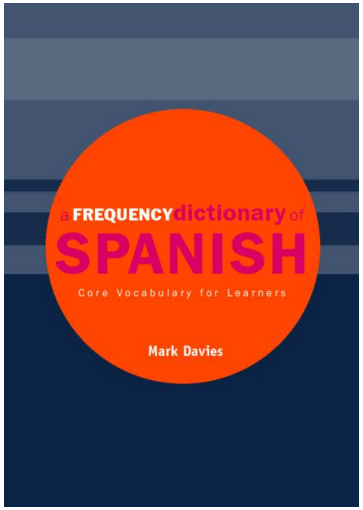
Frequency dictionaries from Routledge











Outline

Problems with frequencies

- Frequency estimation using the BNC:
  is *gastric.J* more common than *toothbrush.N*?

Problems with frequencies

- Frequency estimation using the BNC:
  is *gastric.J* more common than *toothbrush.N*?
- Raw counts and $\pm^{.95}$-confidence intervals:
  $gastric.J = 2057^{\pm 89}$
  $toothbrush.N = 183^{\pm 27}$

Problems with frequencies

- Frequency estimation using the BNC:
  is *gastric.J* more common than *toothbrush.N*?
- Raw counts and $\pm^{.95}$-confidence intervals:
  $gastric.J = 2057^{\pm 89}$
  $toothbrush.N = 183^{\pm 27}$
- Baayen: overdispersed words:
  $Range(gastric.J) = 65$; $Range(toothbrush.N) = 132$
  Out of 4,096 documents in the BNC

Problems with frequencies

- Frequency estimation using the BNC:
  is *gastric.J* more common than *toothbrush.N*?
- Raw counts and $\pm^{.95}$-confidence intervals:
  $gastric.J = 2057^{\pm 89}$
  $toothbrush.N = 183^{\pm 27}$
- Baayen: overdispersed words:
  $Range(gastric.J) = 65$; $Range(toothbrush.N) = 132$
  Out of 4,096 documents in the BNC
- What is the core lexicon for Language X?

Problems with frequencies

- Frequency estimation using the BNC:
  is *gastric.J* more common than *toothbrush.N*?
- Raw counts and $\pm^{.95}$-confidence intervals:
  $gastric.J = 2057^{\pm 89}$
  $toothbrush.N = 183^{\pm 27}$
- Baayen: overdispersed words:
  $Range(gastric.J) = 65$; $Range(toothbrush.N) = 132$
  Out of 4,096 documents in the BNC
- What is the core lexicon for Language X?

→ What are **robust** word frequencies in Corpus Y?
What are possible **pitfalls** of using Corpus Y?

Whelks and bananas (Kilgarriff et al, 2014)

- Frequency bursts (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly planner, evil, cage,* ***pylorus***$_{5955}$*, disguise, sunlight, repay*

Whelks and bananas (Kilgarriff et al, 2014)

- Frequency bursts (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly planner, evil, cage,* ***pylorus***$_{5955}$*, disguise, sunlight, repay*

  Journal of Gastroentorology and Hepatology: 713 kW in the BNC

Whelks and bananas (Kilgarriff et al, 2014)

- Frequency bursts (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly*
  planner, evil, cage, ***pylorus***$_{5955}$*, disguise, sunlight, repay*

  Journal of Gastroentorology and Hepatology: 713 kW in the BNC
- Frequency drops (*banana-toothbrush* problem)
  anchor, instrumental, sodium, ***banana***$_{6965}$*, tilt, hunter, armour*
  leer, enthrall, sheaf, ***toothbrush***$_{19676}$*, dungeon, stocky, lawsuit*

Whelks and bananas (Kilgarriff et al, 2014)

- Frequency bursts (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly*
  planner, evil, cage, ***pylorus***$_{5955}$*, disguise, sunlight, repay*

  Journal of Gastroentorology and Hepatology: 713 kW in the BNC
- Frequency drops (*banana-toothbrush* problem)
  anchor, instrumental, sodium, ***banana***$_{6965}$*, tilt, hunter, armour*
  leer, enthrall, sheaf, ***toothbrush***$_{19676}$*, dungeon, stocky, lawsuit*
- Other senses and constructions:
  With ***banana skins*** *like VAT on fuel some ministers may. . .*
  USA turning into a ***banana republic*** *only without the bananas.*
  turning crime levels in Germany into ***banana republic*** *proportions.*

Frequencies in the BNC

	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
Count	603	2053	2057	2065	38	2042	183
Range	338	1038	65	701	32	1100	123

Frequencies in the BNC

	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
Count	603	2053	2057	2065	38	2042	183
Range	338	1038	65	701	32	1100	123
Frequency distributions in ipm per text:							
Mean	$3.6^{\pm 0.6}$	$21.3^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17.4^{\pm 3.8}$	$0.2^{\pm 0.08}$	$18.0^{\pm 2.0}$	$1.8^{\pm 0.7}$

Frequencies in the BNC

	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
Count	603	2053	2057	2065	38	2042	183
Range	338	1038	65	701	32	1100	123
Frequency distributions in ipm per text:							
Mean	$3.6^{\pm 0.6}$	$21.3^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17.4^{\pm 3.8}$	$0.2^{\pm 0.08}$	$18.0^{\pm 2.0}$	$1.8^{\pm 0.7}$
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.000
1st Q	0.00	0.00	0.00	0.00	0.00	0.00	0.000
Med.	0.00	0.00	0.00	0.00	0.00	0.00	0.000
3rd Q	0.00	12.50	0.00	0.00	0.00	18.73	0.000
Max	479.16	3897.79	2957.45	4143.39	79.37	2028.40	1046.025

Outline

Word frequencies and dispersions

- Juilland's D: $D = 1 - \frac{\sigma}{\mu\sqrt{\|n_i\|-1}}$

  (Juilland's, 1964; Leech, et al 2001; Sharoff, et al, 2013)

Word frequencies and dispersions

- Juilland's D: $D = 1 - \frac{\sigma}{\mu\sqrt{\|n_i\|-1}}$

  (Juilland's, 1964; Leech, et al 2001; Sharoff, et al, 2013)
- Gries' deviation of proportions (2008)

$$DP = 1 - \frac{\sum |\frac{c_i}{C} - \frac{n_i}{N}|}{2}$$

Word frequencies and dispersions

- Juilland's D: $D = 1 - \frac{\sigma}{\mu\sqrt{\|n_i\|-1}}$

  (Juilland's, 1964; Leech, et al 2001; Sharoff, et al, 2013)
- Gries' deviation of proportions (2008)

$$DP = 1 - \frac{\sum |\frac{c_i}{C} - \frac{n_i}{N}|}{2}$$

- Katz (1996) on burstines of a term:

$p_r = \frac{\|c_i=r\|}{\|n_i\|}$	probability of exactly r instances per text
$\alpha = 1 - p_0$	proportion of texts containing the term
$\gamma = 1 - \frac{p_1}{1-p_0}$	proportion of topical texts
$B = \frac{\sum rp_r}{\sum p_r}(r \geq 2)$	topical burstiness parameter

Problems of existing dispersion measures

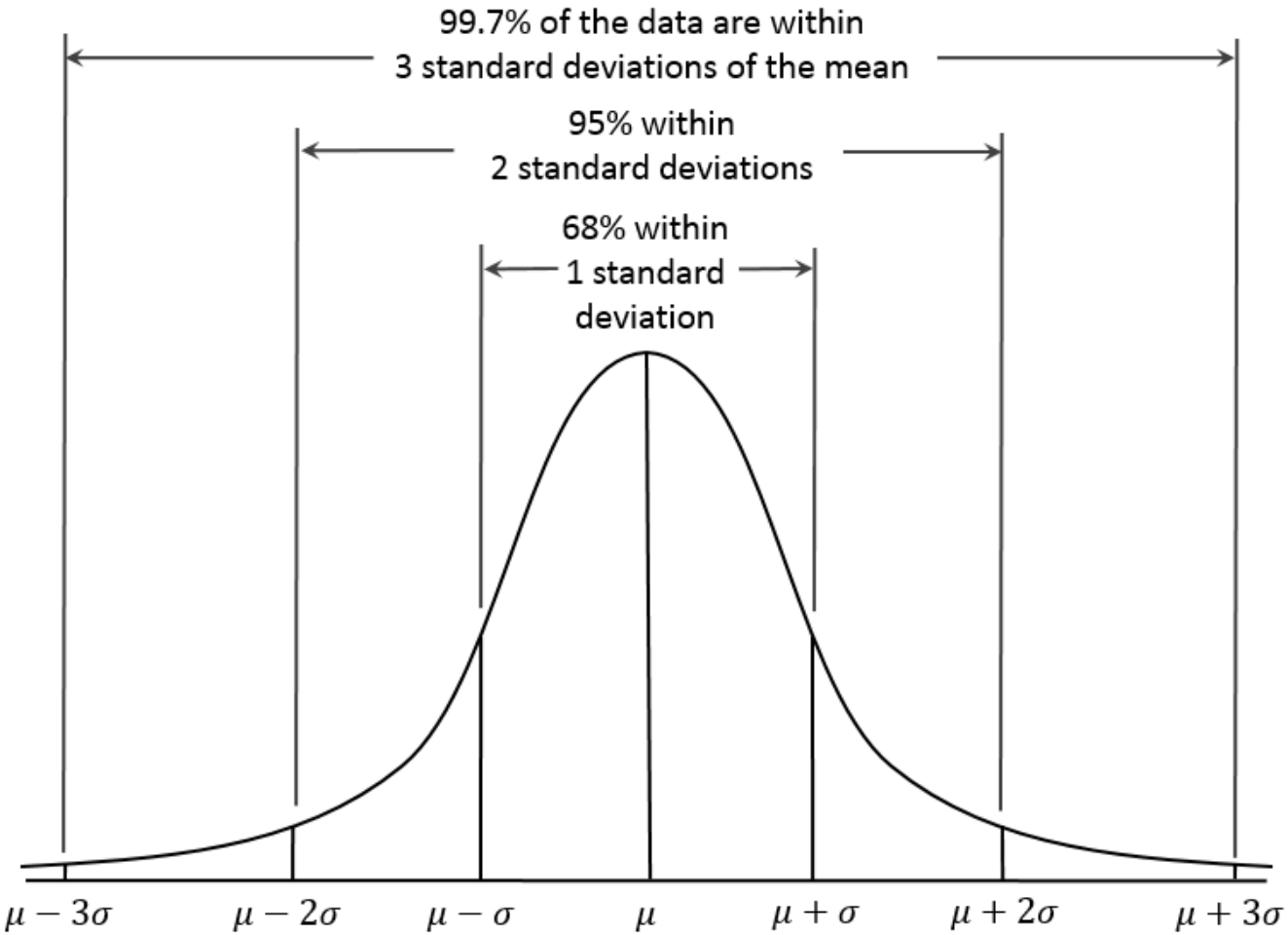
	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
St Dev (σ)	18.05	108.57	83.57	123.02	3.13	64.01	23.56
Juilland's D	0.91	0.89	0.60	0.87	0.72	0.93	0.78
Gries' DP_{norm}	0.37	0.38	0.03	0.27	0.01	0.41	0.06
Katz' α	0.08	0.26	0.02	0.17	0.008	0.27	0.03
Katz' γ	0.36	0.39	0.37	0.49	0.13	0.41	0.28
Katz' B	3.12	3.48	84.00	4.98	2.50	3.11	2.71

Problems of existing dispersion measures

	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
St Dev (σ)	18.05	108.57	83.57	123.02	3.13	64.01	23.56
Juilland's D	0.91	0.89	0.60	0.87	0.72	0.93	0.78
Gries' DP_{norm}	0.37	0.38	0.03	0.27	0.01	0.41	0.06
Katz' α	0.08	0.26	0.02	0.17	0.008	0.27	0.03
Katz' γ	0.36	0.39	0.37	0.49	0.13	0.41	0.28
Katz' B	3.12	3.48	84.00	4.98	2.50	3.11	2.71

Word	α	γ	B
not	0.98	0.98	114.31
be	1.00	1.00	885.61
have	0.99	0.99	293.40
pylorus	0.002	0.78	160.86
do	0.98	0.98	136.25

Two sigma rule for confidence intervals

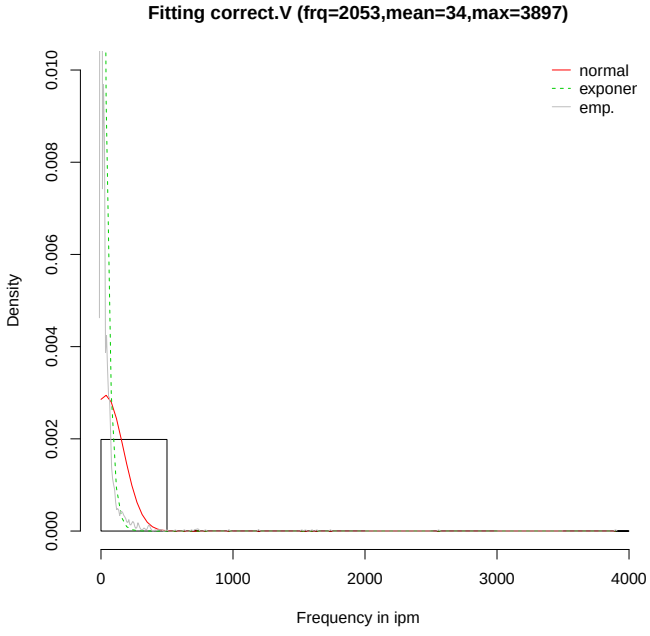


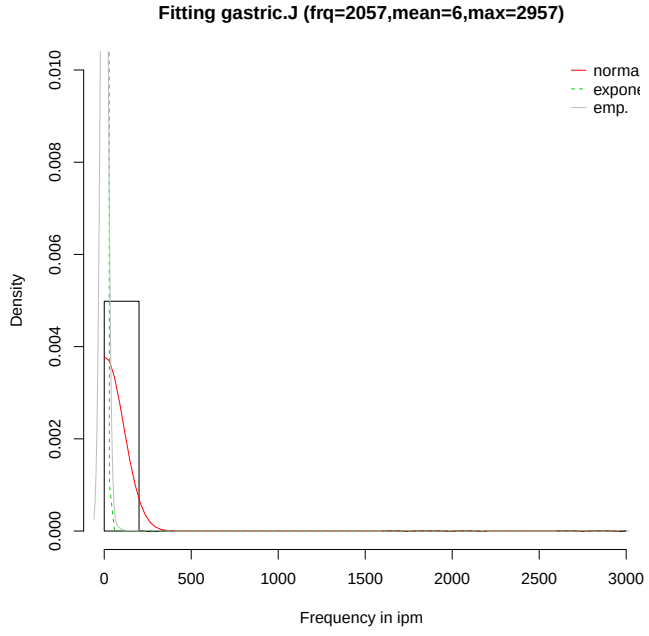
95% within $\mu^{\pm 2\sigma}$; $MSE = \frac{2\sigma}{\sqrt{\|n_i\|}}$; $F(gastric.J) = 3.6^{\pm 2.6}$

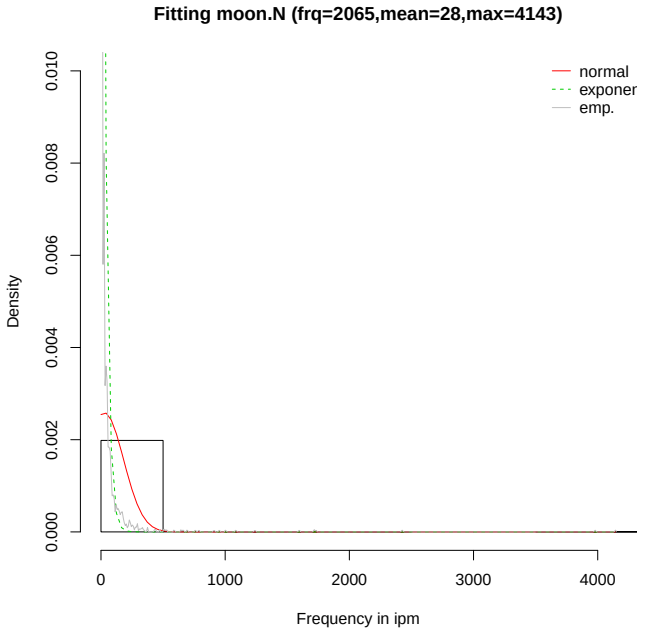
Statistical assumptions

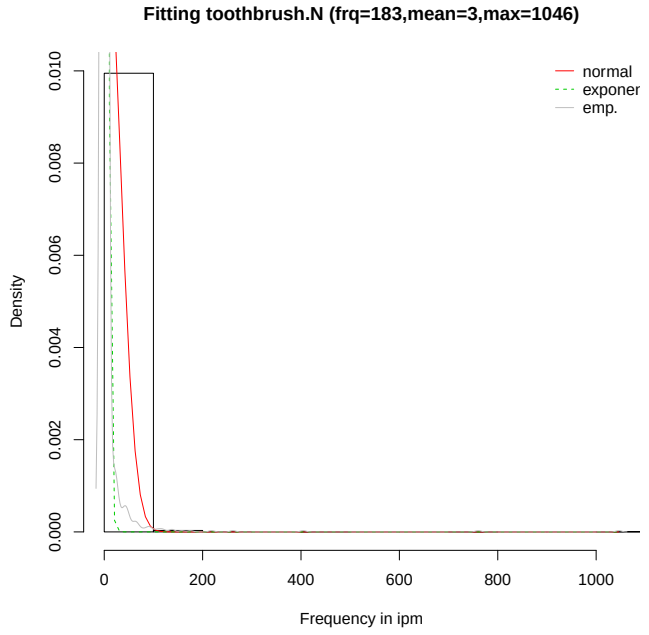
- Independence of observations
  One occurrence of a word is independent from another occurrence
- Normal (Gaussian) distribution
  Frequencies vary following a bell shape
- Homoscedasticity, i.e., equal variance of data:
  Word frequencies in documents vary in similar ways
- Linearity (for linear models)

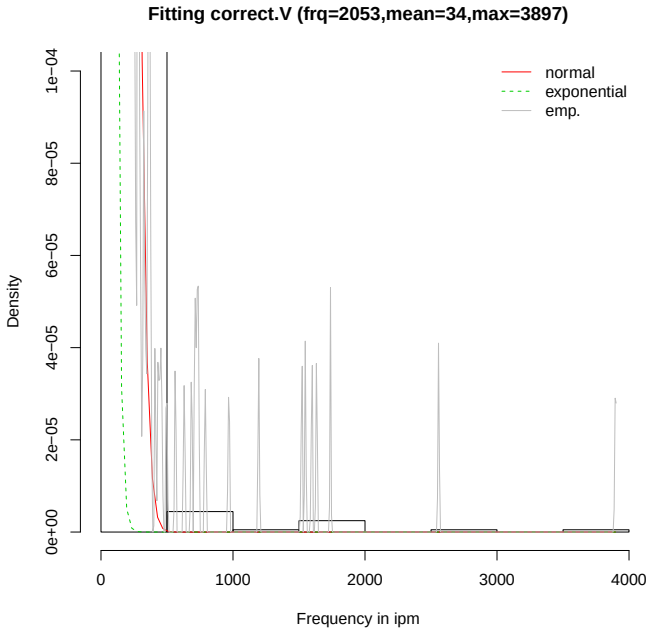
e.g. For confidence intervals or for ANOVA,
deviations from the mean are independently, identically, and normally distributed

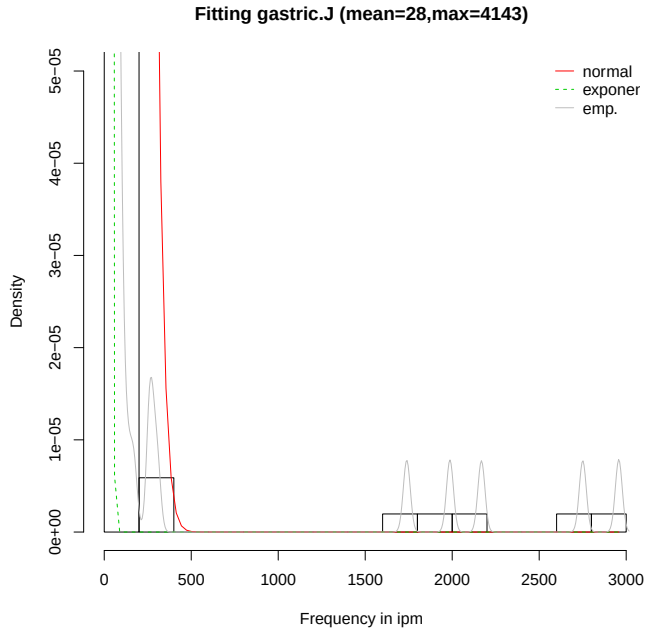

Fitting correct.V (frq=2053,mean=34,max=3897)
normal
exponer
emp.
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000
4000
Frequency in ipm

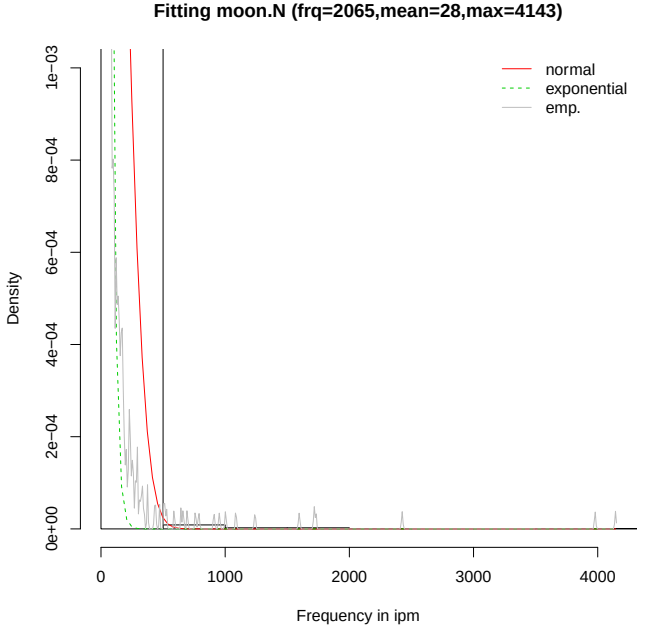

Fitting gastric.J (frq=2057,mean=6,max=2957)
normal
exponer
emp.
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
500
1000
1500
2000
2500
3000
Frequency in ipm

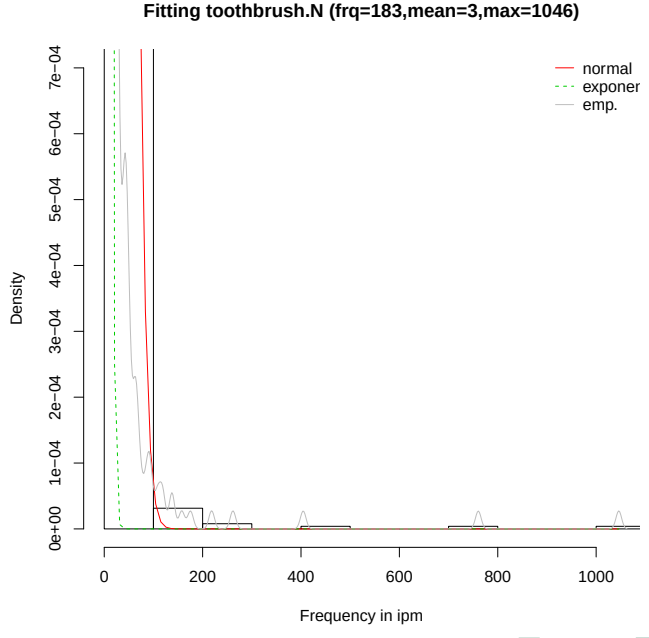

Fitting moon.N (frq=2065,mean=28,max=4143)
normal
exponer
emp.
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
1000
2000
3000
4000
Frequency in ipm


Fitting toothbrush.N (frq=183,mean=3,max=1046)
normal
exponer
emp.
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
200
400
600
800
1000
Frequency in ipm

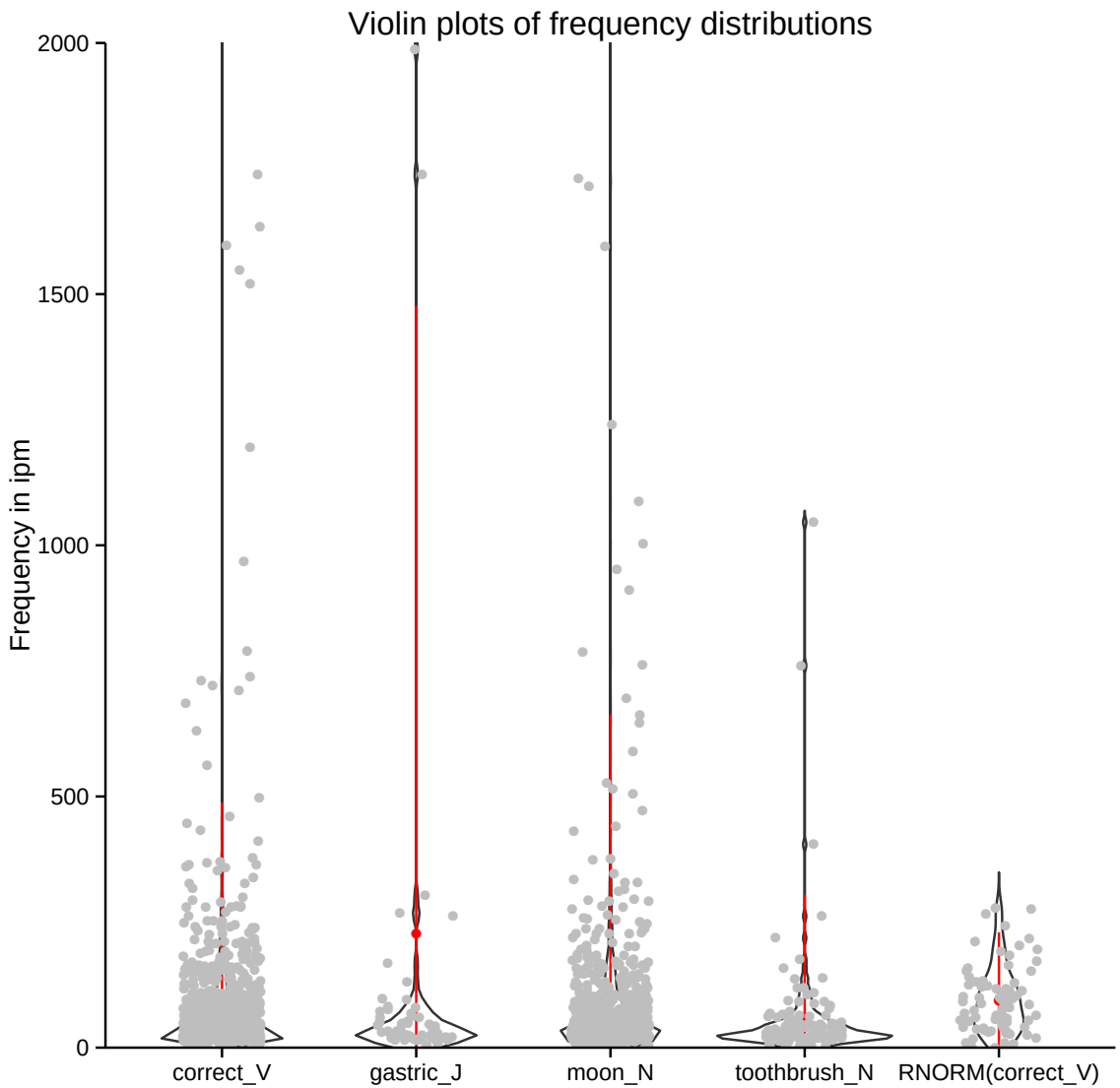

Fitting correct.V (frq=2053,mean=34,max=3897)
normal
exponential
emp.
Density
0e+00
2e−05
4e−05
6e−05
8e−05
1e−04
0
1000
2000
3000
4000
Frequency in ipm


Fitting gastric.J (mean=28,max=4143)
normal
exponer
emp.
Density
0e+00
1e−05
2e−05
3e−05
4e−05
5e−05
0
500
1000
1500
2000
2500
3000
Frequency in ipm


Fitting moon.N (frq=2065,mean=28,max=4143)
normal
exponential
emp.
Density
0e+00
2e−04
4e−04
6e−04
8e−04
1e−03
0
1000
2000
3000
4000
Frequency in ipm


Fitting toothbrush.N (frq=183,mean=3,max=1046)
normal
exponer
emp.
Density
0e+00
1e−04
2e−04
3e−04
4e−04
5e−04
6e−04
7e−04
0
200
400
600
800
1000
Frequency in ipm

Violin plots



Outline

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

E Huber's M-estimator $\frac{|x_i - \mu_k|}{MAD} \leq 1.28 \Rightarrow$ good

$v = (1, 2, 3, 4, |5|, 10, 100, 1000, 10000); h(v) = 8.61$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

E Huber's M-estimator $\frac{|x_i - \mu_k|}{MAD} \leq 1.28 \Rightarrow$ good

$v = (1, 2, 3, 4, |5|, 10, 100, 1000, 10000); h(v) = 8.61$

S Improvement over MAD $S_n = c \times M_i\,(M_j(\,|x_i - x_j|\,))$

$\sigma(v) = 3302.621; MAD(v) = 5.930; S_n = 5.395$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

E Huber's M-estimator $\frac{|x_i - \mu_k|}{MAD} \leq 1.28 \Rightarrow$ good
$v = (1, 2, 3, 4, |5|, 10, 100, 1000, 10000); h(v) = 8.61$

S Improvement over MAD $S_n = c \times M_i\,(M_j(\,|x_i - x_j|\,))$
$\sigma(v) = 3302.621; MAD(v) = 5.930; S_n = 5.395$

CI Robust confidence intervals via bootstrap

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

E Huber's M-estimator $\frac{|x_i - \mu_k|}{MAD} \leq 1.28 \Rightarrow$ good

$v = (1, 2, 3, 4, |5|, 10, 100, 1000, 10000); h(v) = 8.61$

S Improvement over MAD $S_n = c \times M_i\,(M_j(\,|x_i - x_j|\,))$

$\sigma(v) = 3302.621; MAD(v) = 5.930; S_n = 5.395$

CI Robust confidence intervals via bootstrap

- Huge number of zeros for language: $M, MAD, S_n = 0$

Robust measures of Estimation and Scatter

- Mean (μ) and sd (σ) are not robust (unlike Median): frequency bursts move them out of bounds

  $v = (1, 2, 3, 4, |5|, 10, 11, 12, 13); \mu(v) = 6.778, M(v) = 5$

  $(1, 2, 3, 4, |5|, 10, 100, 1000, 10000); \mu(v) = 1236, M(v) = 5$

S Median Absolute Deviation $MAD = b \times M\,|x_i - M(x)|$

E Huber's M-estimator $\frac{|x_i - \mu_k|}{MAD} \leq 1.28 \Rightarrow$ good

$v = (1, 2, 3, 4, |5|, 10, 100, 1000, 10000); h(v) = 8.61$

S Improvement over MAD $S_n = c \times M_i\,(M_j(\,|x_i - x_j|\,))$

$\sigma(v) = 3302.621; MAD(v) = 5.930; S_n = 5.395$

CI Robust confidence intervals via bootstrap

- Huge number of zeros for language: $M, MAD, S_n = 0$

$\rightarrow$ Robust estimates on non-zero docs

Contrasting non-zero docs

Word	Total	Docs	Min.	1stQu.	Median	Mean	3rdQu.	Max.
correct.V	2053	1038	3.87	22.60	41.70	83.30	72.80	3900.00
gastric	2057	**65**	5.24	21.35	30.23	226.00	69.12	2957.00

Contrasting non-zero docs

Word	Total	Docs	Min.	1stQu.	Median	Mean	3rdQu.	Max.
correct.V	2053	1038	3.87	22.60	41.70	83.30	72.80	3900.00
gastric	2057	**65**	5.24	21.35	30.23	226.00	69.12	2957.00
moon.N	2065	701	2.72	22.20	42.30	101.00	81.40	4140.00
toothbrush	183	**123**	2.72	20.59	28.51	59.94	50.24	1046.00

Contrasting non-zero docs

Word	Total	Docs	Min.	1stQu.	Median	Mean	3rdQu.	Max.
correct.V	2053	1038	3.87	22.60	41.70	83.30	72.80	3900.00
gastric	2057	**65**	5.24	21.35	30.23	226.00	69.12	2957.00
moon.N	2065	701	2.72	22.20	42.30	101.00	81.40	4140.00
toothbrush	183	**123**	2.72	20.59	28.51	59.94	50.24	1046.00

Outlying observations to be Winsorised by $E[f] + 2S[f]$

Contrasting non-zero docs

Word	Total	Docs	Min.	1stQu.	Median	Mean	3rdQu.	Max.
correct.V	2053	1038	3.87	22.60	41.70	83.30	72.80	3900.00
gastric	2057	**65**	5.24	21.35	30.23	226.00	69.12	2957.00
moon.N	2065	701	2.72	22.20	42.30	101.00	81.40	4140.00
toothbrush	183	**123**	2.72	20.59	28.51	59.94	50.24	1046.00

Outlying observations to be Winsorised by $E[f] + 2S[f]$

Word	$\mu + 2\sigma$	Out	*huber*	*MAD*	S_n	$h + 2M$	Out	$h + 2S_n$	Out
correct.V	487.80	18	48.69	30.43	28.68	109.56	158	106.05	163
gastric	1476.98	5	40.30	21.71	18.41	83.73	12	77.11	**14**

Contrasting non-zero docs

Word	Total	Docs	Min.	1stQu.	Median	Mean	3rdQu.	Max.
correct.V	2053	1038	3.87	22.60	41.70	83.30	72.80	3900.00
gastric	2057	**65**	5.24	21.35	30.23	226.00	69.12	2957.00
moon.N	2065	701	2.72	22.20	42.30	101.00	81.40	4140.00
toothbrush	183	**123**	2.72	20.59	28.51	59.94	50.24	1046.00

Outlying observations to be Winsorised by $E[f] + 2S[f]$

Word	$\mu + 2\sigma$	Out	*huber*	*MAD*	S_n	$h + 2M$	Out	$h + 2S_n$	Out
correct.V	487.80	18	48.69	30.43	28.68	109.56	158	106.05	163
gastric	1476.98	5	40.30	21.71	18.41	83.73	12	77.11	**14**
moon.N	663.69	14	51.05	32.53	31.03	116.10	124	113.12	127
toothbrush	304.36	3	34.87	18.91	21.47	72.68	20	77.82	**19**

BNC	Frq capped	Frq orig	LL.score
patient.N	5433	22100	7014
cell.N	4105	12868	2790
award.N	3876	12132	2624
teacher.N	7696	19089	2532
ref.N	422	3994	2498
player.N	4923	13432	2227
module.N	1066	5375	2154
speaker.N	2872	9250	2100
pupil.N	3673	10321	1820
social.J	19344	36261	1816
user.N	4551	11879	1784
language.N	10068	20780	1566
gastric.J	170	2057	1441
studio.N	2747	7838	1426
king.N	2299	6820	1338
study.N	16305	29816	1318
share.N	8735	17821	1285
student.N	11037	21492	1281
company.N	33005	54372	1245

ukWac	Frq capped	Frq orig	LL.score
insurance.N	175528	308124	27209
loan.N	117165	178319	8425
puzzle.N	28831	57625	7660
HMS.N	15844	36643	6929
wedding.N	88168	135645	6783
RAF.N	30458	56383	5968
course.N	978425	1174112	5819
campus.N	51436	84551	5768
God.N	456901	573930	5750
mortgage.N	66821	103805	5454
dog.N	119777	169642	5183
child.N	1107349	1309272	4811
Select.N	45875	70812	3602
pension.N	107261	146971	3516
credit.N	212619	272102	3413
Sale.N	22842	39722	3389
Estate.N	46976	71189	3287
Mulder.N	5133	12765	2793
nigritude.N	188	1936	1541

titleid	left	match	right
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. Google Tests Expanded
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. Dynamic Pages Dynamic
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. Search Engine Keywords
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. SEOs Relationship
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. The Myth of Search
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. How Search Engines
>>	and felicitations for the ...	nigritude	ultramarine is rising again. After
>>	waiting (bis repetita), our	nigritude	ultramarine page is coming from
>>	software bottom of the google	nigritude	best anti virus software ultramarine
>>	middle of it. let 's see if our	nigritude	ultramarine team anti virus software
>>	waiting, our anti virus software uk	nigritude	ultramarine page is now indexed
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. Search Engine Marketing
>>	webpage for a non-sensical phrase: '	nigritude	ultramarine '. On May 7th, the
>>	On May 7th, the day the terms '	nigritude	ultramarine ' was announced, typing
>>	their page to #1 for the phrase '	nigritude	ultramarine '. The contest is

Google nigritude ultramarine

Web Images Shopping Videos Maps More Search tools

About 7,680 results (0.33 seconds)

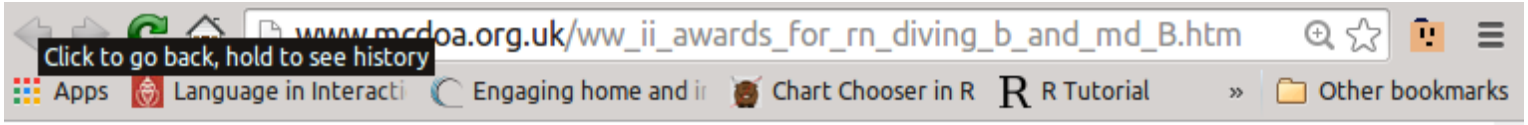
SEO contest - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/SEO_contest
In the English-language world, the **nigritude ultramarine** competition created by DarkBlue.com and run by SearchGuild is widely acclaimed as the mother of all ...

Nigritude Ultramarine - Anil Dash
dashes.com/anil/2004/06/**nigritude**-ultra.html
Nigritude Ultramarine. June 4, 2004. Update: The contest is over, and this entry did pretty well but didn't win the initial prize. So the best purpose this page can ...

Nigritude Ultramarine SEO Competition SEO Blog ... - Sim64
www.sim64.co.uk/uk/**nigritude-ultramarine**/
Nigritude Ultramarine was an early SEO competition organised by SEO forum SearchGuild and sponsored by the Australian affiliate network Dark Blue (hence ...

Images for nigritude ultramarine Report images





A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

NAME	RANK/RATE	UNIT	DATE GAZETTED	AWARD	REMARKS
		B			
BABINGTON John Herbert	Sub Lt (Sp) RNVR (Later Lt Cdr)	HMS President (London) HMS Volcano (Ravenglass)	27 Dec 40 8 Jun 43	GC OBE	BD and RMS - See below.
	GC awarded for great gallantry and undaunted devotion to duty. Sub-Lieutenant Babington had experimented with the dismantling of all types of bombs and had worked on the first suspended parachute magnetic mines. He volunteered to deal with a very dangerous bomb in Chatham dockyard in the autumn of 1940 where an anti-withdrawal device was suspected - an RAF officer had shortly before lost his life in trying to deal with a bomb of this description. Sub-Lieutenant Babington was lowered into a 16 ft pit where he tied a line to the head of the fuse but the line broke and he again went into the pit. He had to make three attempts to remove the fuse before he finally directed the lifting of the bomb which was then taken away. At that time the Bomb Disposal Authorities had very little knowledge of the mechanism of these mines and much was learnt from this incident. Appointed OBE for coolness and courage in operations involving great risk to himself.				

Up Canasius, The Paragon, 02 September 2003, 11:53:04

Canasius enters.
Canasius says, "Im sorry"
-> Kiania looks up
Canasius says, "can I sit?"
You nod
Canasius falls unconscious...
Canasius has regained consciousness.
You say, "Why Sorry?"
Canasius says, "your crying"
Canasius says, "he upset you"

Back to gastric-banana problems

- Frequency spikes (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly planner, evil, cage,* ***pylorus***$_{5955}$*, disguise, sunlight, repay*

Back to gastric-banana problems

- Frequency spikes (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly*
  planner, evil, cage, ***pylorus***$_{5955}$*, disguise, sunlight, repay*

$\rightarrow$ *vindictive, bitumen, cleave,* ***gastric***$_{16727}$*, minke, railwayman*
verger, rigorist, ***pylorus***$_{37868}$*, moonbeam, correlative, gallivant*

Back to gastric-banana problems

- Frequency spikes (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly*
  planner, evil, cage, ***pylorus***$_{5955}$*, disguise, sunlight, repay*

→ *vindictive, bitumen, cleave,* ***gastric***$_{16727}$*, minke, railwayman*
verger, rigorist, ***pylorus***$_{37868}$*, moonbeam, correlative, gallivant*

- Frequency drops (*banana-toothbrush* problem)
  anchor, instrumental, sodium, ***banana***$_{6965}$*, tilt, hunter, armour*
  leer, enthrall, sheaf, ***toothbrush***$_{19676}$*, dungeon, stocky, lawsuit*

Back to gastric-banana problems

- Frequency spikes (*whelk-gastric* problem);
  moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly planner, evil, cage,* ***pylorus***$_{5955}$*, disguise, sunlight, repay*

→ *vindictive, bitumen, cleave,* ***gastric***$_{16727}$*, minke, railwayman verger, rigorist,* ***pylorus***$_{37868}$*, moonbeam, correlative, gallivant*

- Frequency drops (*banana-toothbrush* problem)
  anchor, instrumental, sodium, ***banana***$_{6965}$*, tilt, hunter, armour leer, enthrall, sheaf,* ***toothbrush***$_{19676}$*, dungeon, stocky, lawsuit*

→ *overview, floating, group,* ***banana***$_{5148}$*, wounded, catch, philosopher balancing, unhappily, suicidal,* ***toothbrush***$_{12017}$*, cuisine, retaliate, take-off*

Comparing BNC and ukWac

BNC	anxiously	correct.V	gastric.J	moon.N	moon.V	thoroughly	toothbrush
Raw rank	9157	3769	3763	3746	47820	3783	19676
New rank	7341	2641	16727	3275	27278	2647	12017
Raw IPM	$3.6^{\pm 0.6}$	$21^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17^{\pm 3.8}$	$0.20^{\pm 0.08}$	$18^{\pm 2.0}$	$1.8^{\pm 0.7}$
Estimate	$2.8^{\pm 0.3}$	$13^{\pm 0.9}$	$0.7^{\pm 0.2}$	$10^{\pm 0.8}$	$0.17^{\pm 0.06}$	$13^{\pm 0.8}$	$1.2^{\pm 0.2}$

Comparing BNC and ukWac

BNC	anxiously	correct.V	gastric.J	moon.N	moon.V	thoroughly	toothbrush
Raw rank	9157	3769	3763	3746	47820	3783	19676
New rank	7341	2641	16727	3275	27278	2647	12017
Raw IPM	$3.6^{\pm 0.6}$	$21^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17^{\pm 3.8}$	$0.20^{\pm 0.08}$	$18^{\pm 2.0}$	$1.8^{\pm 0.7}$
Estimate	$2.8^{\pm 0.3}$	$13^{\pm 0.9}$	$0.7^{\pm 0.2}$	$10^{\pm 0.8}$	$0.17^{\pm 0.06}$	$13^{\pm 0.8}$	$1.2^{\pm 0.2}$
ukWac							
Raw IPM	$0.6^{\pm 0.05}$	$16^{\pm 0.3}$	$1.7^{\pm 0.15}$	$13^{\pm 0.4}$	$0.08^{\pm 0.02}$	$20^{\pm 0.3}$	$1.3^{\pm 0.13}$
Estimate	$0.5^{\pm 0.03}$	$12^{\pm 0.2}$	$1.1^{\pm 0.07}$	$9^{\pm 0.2}$	$0.05^{\pm 0.01}$	$16^{\pm 0.2}$	$0.8^{\pm 0.05}$

Comparing BNC and ukWac

BNC	anxiously	correct.V	gastric.J	moon.N	moon.V	thoroughly	toothbrush
Raw rank	9157	3769	3763	3746	47820	3783	19676
New rank	7341	2641	16727	3275	27278	2647	12017
Raw IPM	$3.6^{\pm 0.6}$	$21^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17^{\pm 3.8}$	$0.20^{\pm 0.08}$	$18^{\pm 2.0}$	$1.8^{\pm 0.7}$
Estimate	$2.8^{\pm 0.3}$	$13^{\pm 0.9}$	$0.7^{\pm 0.2}$	$10^{\pm 0.8}$	$0.17^{\pm 0.06}$	$13^{\pm 0.8}$	$1.2^{\pm 0.2}$
ukWac							
Raw IPM	$0.6^{\pm 0.05}$	$16^{\pm 0.3}$	$1.7^{\pm 0.15}$	$13^{\pm 0.4}$	$0.08^{\pm 0.02}$	$20^{\pm 0.3}$	$1.3^{\pm 0.13}$
Estimate	$0.5^{\pm 0.03}$	$12^{\pm 0.2}$	$1.1^{\pm 0.07}$	$9^{\pm 0.2}$	$0.05^{\pm 0.01}$	$16^{\pm 0.2}$	$0.8^{\pm 0.05}$
Wikipedia							
Raw IPM	$0.19^{\pm 0.03}$	$7.8^{\pm 0.2}$	$1.4^{\pm 0.15}$	$14^{\pm 0.4}$	$0.07^{\pm 0.02}$	$3.5^{\pm 0.1}$	$0.36^{\pm 0.06}$
Estimate	$0.16^{\pm 0.09}$	$6.1^{\pm 0.2}$	$1.0^{\pm 0.08}$	$10^{\pm 0.2}$	$0.05^{\pm 0.01}$	$2.9^{\pm 0.1}$	$0.25^{\pm 0.03}$
Giga-EN							
Raw IPM	$1.3^{\pm 0.07}$	$61^{\pm 1.6}$	$0.3^{\pm 0.04}$	$13^{\pm 0.3}$	$0.09^{\pm 0.02}$	$6.4^{\pm 0.2}$	$0.7^{\pm 0.06}$
Estimate	$1.1^{\pm 0.05}$	$22^{\pm 0.3}$	$0.2^{\pm 0.02}$	$11^{\pm 0.2}$	$0.08^{\pm 0.01}$	$5.5^{\pm 0.1}$	$0.5^{\pm 0.03}$

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*
- Robust frequency estimates
  Possibly with their 95% confidence intervals

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*
- Robust frequency estimates
  Possibly with their 95% confidence intervals
- Adam Kilgarriff's metaphor: garden vs jungle:

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*
- Robust frequency estimates
  Possibly with their 95% confidence intervals
- Adam Kilgarriff's metaphor: garden vs jungle:

  BNC is a walled garden, but this does not guarantee clean data

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*
- Robust frequency estimates
  Possibly with their 95% confidence intervals
- Adam Kilgarriff's metaphor: garden vs jungle:

  BNC is a walled garden, but this does not guarantee clean data

  Jungle of the Web suffers from clean data issues anyway

Take-home message

- Don't use raw counts: *gastric* $\neq$ *correct*
- Robust frequency estimates
  Possibly with their 95% confidence intervals
- Adam Kilgarriff's metaphor: garden vs jungle:

  BNC is a walled garden, but this does not guarantee clean data

  Jungle of the Web suffers from clean data issues anyway
- Lexical cohesion: some words are inherently topical
  moon.N, gastric vs *moon.V, thoroughly*