Math/Stat 341: Probability
Least Squares Lecture

Steven J Miller
Williams College

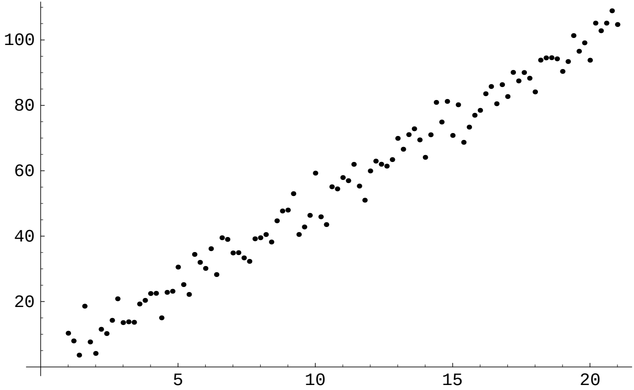
sjm1@williams.edu

http://www.williams.edu/Mathematics/sjmiller/public_html/341

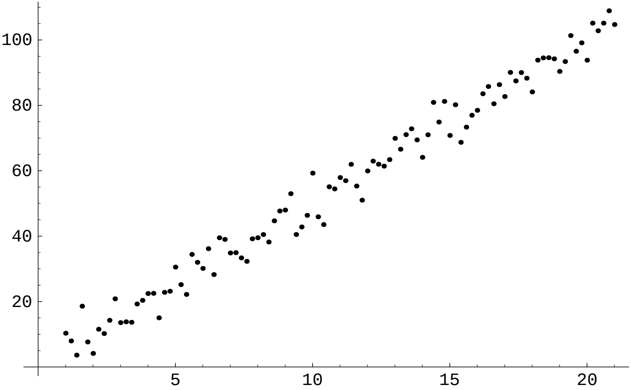
Bronfman 106
Williams College, May 11, 2015

Introduction

Spring Test



Spring Test



Data from $x_n = 5 + .2n$, $y_n = 5x_n$ plus an error randomly drawn from a normal distribution with mean zero and standard deviation 4. Best fit line of $y = 4.99x + .48$; thus $a = 4.99$ and $b = .48$.

Spring Test (continued)

Our value of b is significantly off: $a = 4.99$ and $b = .48$.

Spring Test (continued)

Our value of b is significantly off: $a = 4.99$ and $b = .48$.

Using absolute values for errors gives best fit value of a is 5.03 and the best fit value of b is less than 10^{-10} in absolute value.

Spring Test (continued)

Our value of b is significantly off: $a = 4.99$ and $b = .48$.

Using absolute values for errors gives best fit value of a is 5.03 and the best fit value of b is less than 10^{-10} in absolute value.

The difference between these values and those from the Method of Least Squares is in the best fit value of b (the least important of the two parameters), and is due to the different ways of weighting the errors.

Zipf's Law

City Populations

The twenty-five most populous cities (I believe this is American cities from a few years ago):

8,363,710	1,540,351	912,062	754,885	620,535
3,833,995	1,351,305	808,976	703,073	613,190
2,853,114	1,279,910	807,815	687,456	604,477
2,242,193	1,279,329	798,382	669,651	598,707
1,567,924	948,279	757,688	636,919	598,541

City Populations

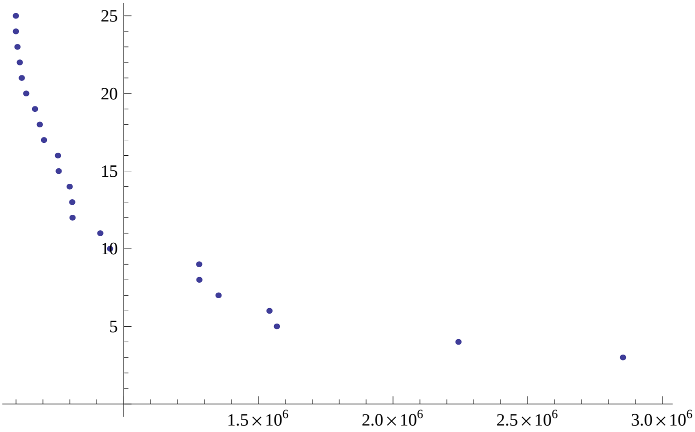


Figure: Plot of rank versus population

City Populations

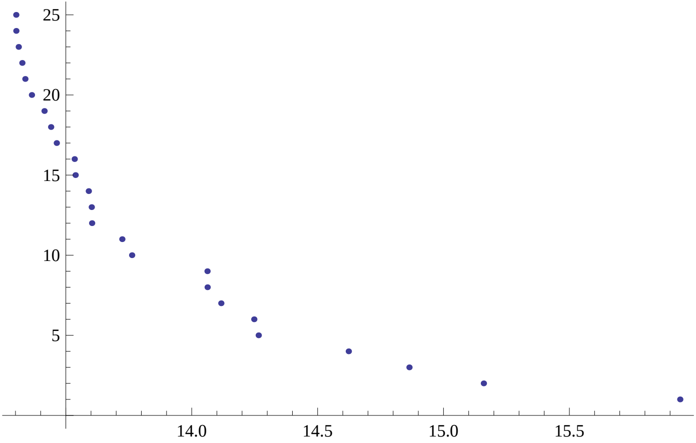


Figure: Plot of rank versus log(population)

City Populations

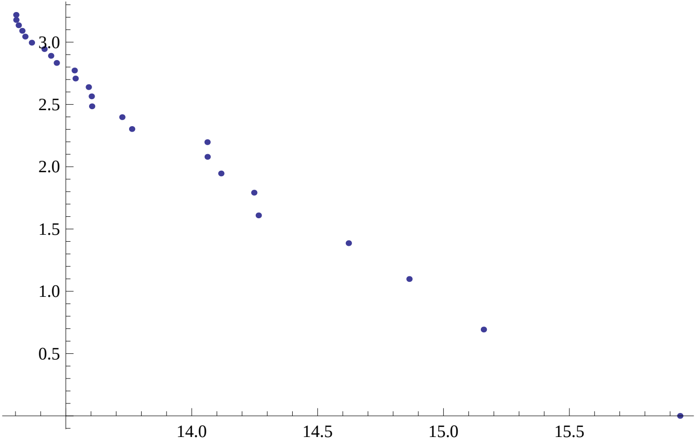


Figure: Plot of log(rank) versus log(population)

City Populations

Plot of 100 most populous cities

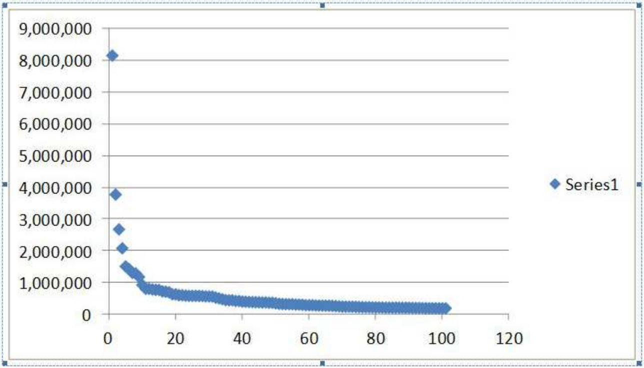


Figure: Plot of rank versus population

City Populations

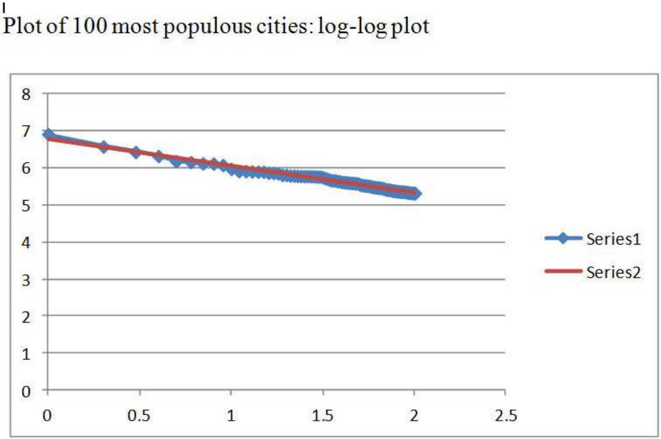


Figure: Plot of log(rank) versus log(population)

Word Counts

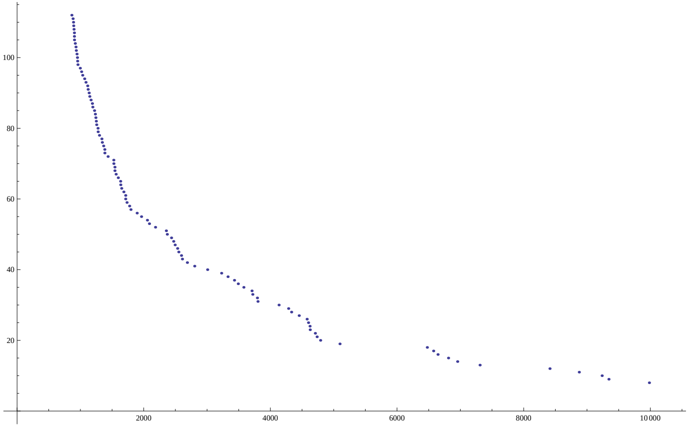


Figure: Plot of rank versus occurrences

Word Counts

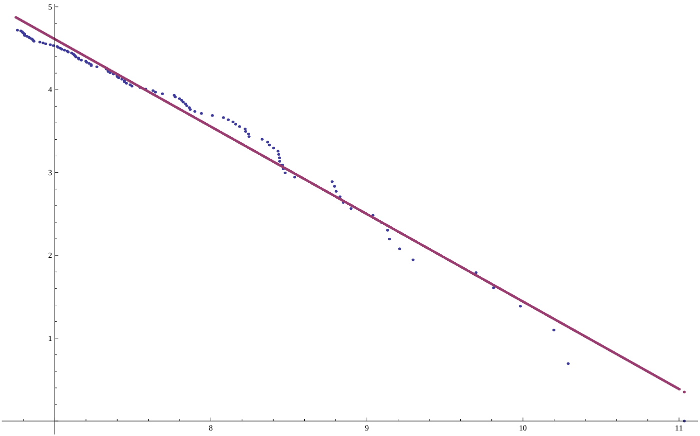


Figure: Plot of log(rank) versus log(occurrences)

Word Counts

The First Hundred				
1. the	21. at	41. there	61. some	81. my
2. of	22. be	42. use	62. her	82. than
3. and	23. this	43. an	63. would	83. first
4. a	24. have	44. each	64. make	84. water
5. to	25. from	45. which	65. like	85. been
6. in	26. or	46. she	66. him	86. call
7. is	27. one	47. do	67. into	87. who
8. you	28. had	48. how	68. time	88. oil
9. that	29. by	49. their	69. has	89. its
10. it	30. word	50. if	70. look	90. now
11. he	31. but	51. will	71. two	91. find
12. was	32. not	52. up	72. more	92. long
13. for	33. what	53. other	73. write	93. down
14. on	34. all	54. about	74. go	94. day
15. are	35. were	55. out	75. see	95. did
16. as	36. we	56. many	76. number	96. get
17. with	37. when	57. then	77. no	97. come
18. his	38. your	58. them	78. way	98. made
19. they	39. can	59. these	79. could	99. may
20. I	40. said	60. so	80. people	100. part

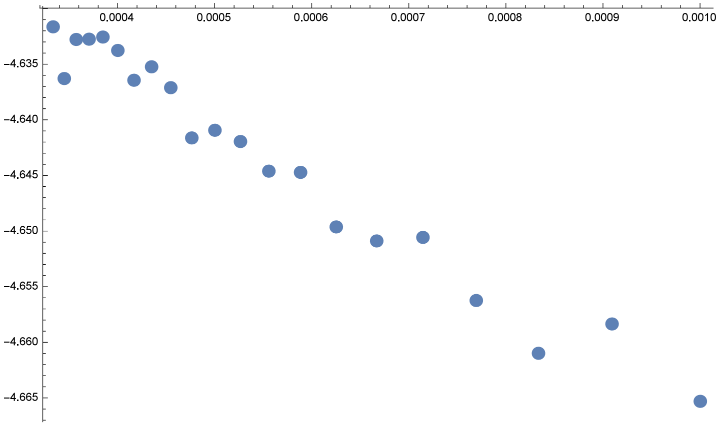
Coin Problem

- Toss n die, each k-sides, numbers $1, 2, \ldots,$ k all equally likely.
- Take largest number as result, call it m.
- Get $1 - \frac{m}{k}$.
- Theory predicts....

Coin Problem

- Toss n die, each k-sides, numbers $1, 2, \ldots, k$ all equally likely.
- Take largest number as result, call it m.
- Get $1 - \frac{m}{k}$.
- Theory predicts.... $e^{-n/k}/(n+1)$.

Numerical Data (semi-log plot, using $1/k$)



Least Squares Analysis

```
In[269]:= Clear[x]
          lm = LinearModelFit[list, x, x]
          Clear[x]
          llm = LinearModelFit[loglist, x, x]
          Clear[x]
          slm = LinearModelFit[semiloglist, x, x]

Out[270]= FittedModel[ 0.00990019 - 0.518556 x ]

Out[272]= FittedModel[ -4.88236 - 0.0314637 x ]

Out[274]= FittedModel[ -4.61468 - 54.2127 x ]

          Answer should be Exp[-n / k] / (n + 1)
          As n = 100 this gives y = Exp[-100 / k] / 101
          We took logs so get log (y) = -100 (1 / k) - log (100)

In[301]:= {-Log[100.], -100}

Out[301]= {-4.60517, -100}
```

Note answer is very good in constant term but slope is off by almost a factor of 2!