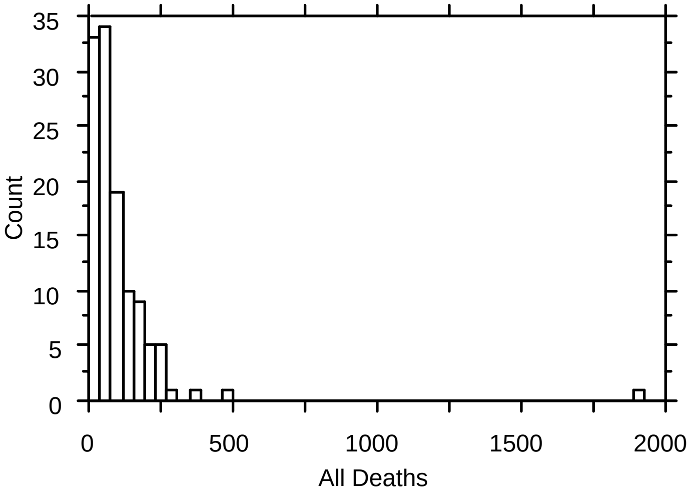
BE.104 Spring
Problem Set 3 Answers
Sherley

See Table III from the Feb. 25, 2005 issue of *Mortality Morbidity Weekly Report* (Vol 54, p. 191).
http://www.cdc.gov/mmwr/mmwr_wk.html

1) Generate a frequency histogram for all deaths (i.e., "All Ages" entries) in the listed cities. What should you do with the “U”’s?

Answer: 0.5pt



Answer: 0.25pt
Nothing. There is no data, which is not the same as a zero. The information is not known.

2) Describe the distribution.

Answer: 0.25pt
Skewed

3) Suggest at least 3 distinct classes of “processes or factors” that might yield this distribution.
Justify your answers.
Will a similar distribution occur next week?

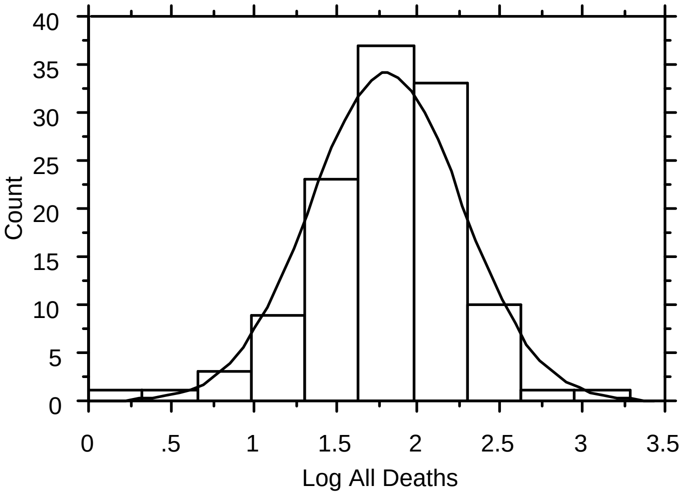
Answers (any 3 are acceptable, including multiple versions of c): 1.5pts (0.5pt for each)

a) Difference in population of cities. Number of deaths will have some relationship to number of individuals. If the distribution of population size among the cities is skewed, then the distribution of deaths might be skewed accordingly
b) Differences in the age distribution of cities. Older individuals have a higher risk of dying. If the median age of the cities is not normally distributed, then the distribution of deaths might be skewed accordingly
c) There may be other factors such as crowding, poor sanitation, or population demographics (which may be related or unrelated to population size) that result in a skewed distribution of deaths.
d) This group of cities is self-selected. This selection bias may lead to a non-normal distribution of deaths.

Answer: 0.25pt
Yes, given that the same cities will be reporting, and the responsible factors are likely to be *chronic*, it is probable that the similar distribution will occur each week (and it does!).

4) Generate a new distribution based on the log(X), where X = the number of deaths for each city. Describe the log-transformed distribution. How would you now describe the original data set? How might this transformation be useful to you if you were comparing death rates among U.S. cities to those among European cities?

Answer:



Answer:
Near normal. **0.5pt**
“Log Normal” **0.25pt**
The mean of the log X distribution is a better measure of the “center” of greatly skewed distributions. The antilog of the mean of the log X distribution is called the geometric mean. For highly skewed distributions, the geometric mean is a better basis for comparison than the arithmetic mean.
Arithmetic mean = 180
Geometric mean = 64
0.5pt

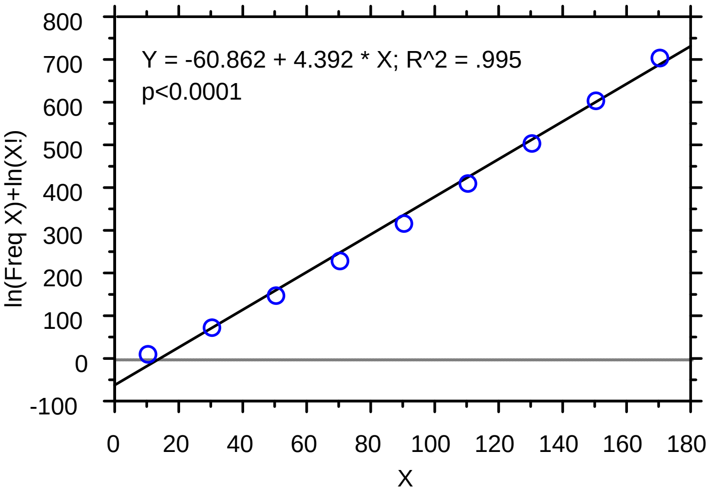
5) Re-evaluate the frequency distribution for all deaths (i.e., "All Ages" entries) in the listed cities for its "Poisson-ness".

Answer: 2pts
Perform Poisson Plot test
Rank the data
Decide on constant size bins
Develop list of bin median (= x) vs number of cities with in the bin (number of cities in bin/119 = frequency of x).

Bin	bin median	number	Frequency (=number/119)
1-20	10	14	0.12
21-40	30	19	0.16
41-60	50	17	0.14
61-80	70	19	0.16
81-100	90	8	0.07
101-120	110	10	0.08
121-140	130	6	0.05
141-160	150	5	0.04

161-180	170	5	0.04
181-200	190	4	0.03
201-220	210	2	0.02
221-240	230	2	0.02
241-260	250	3	0.03
261-280	270	1	0.01
281-300	290	1	0.01
361-380	370	1	0.01
461-480	470	1	0.01
1921-1940	1930	1	0.01

Perform Poisson test plot: X vs ln (Freq X) + ln(X!)



The distribution is modeled well by a Poisson distribution, even though the absolute number of deaths is large in many cases. Thus, deaths are **rare observations** compared to the entire population; a major fraction are **random (unpredictable)** in occurrence; a major fraction do not effect their source populations in any significant manner; and a major fraction are **independent** of other deaths. Thus, a significant fraction of deaths are Poisson distributed.

Insight: this problem can be made more calculator friendly by scaling the X. For example, dividing all deaths by 10 will not change the structure of the distribution. Try this on your own.

6) What level of confidence do you have that NYC's higher number of deaths (1926) is not simply due to chance and error?

Answer: 1pt for either method below
Based on the results of question #5:

Method 1

If you remove 1926 from the distribution, then the mean = 93

For x = 1926, 99% CI for the mean about x $=1926 + 3.32 \pm 2.567\,(1926 + 1.7)^{1/2}$

$=1929 \pm 113$

$= 1816\text{-}2042$, does not include 93

Therefore, greater than 99% confidence; $p < 0.01$

Method 2

Scale the problem by dividing all deaths by 93, the mean when 1926 is removed from the distribution. Now phrase the question as, “What is the probability of observing 21 (1926/93) when 1 (=93/93) was expected? Remember the structure of the distribution is not altered by scaling it.

$P = (e^{-1})(1^{21})/21! = 7 \text{ x } 10^{-21}$

7) Is there a significant relationship between number of deaths and population size for U.S. cities?

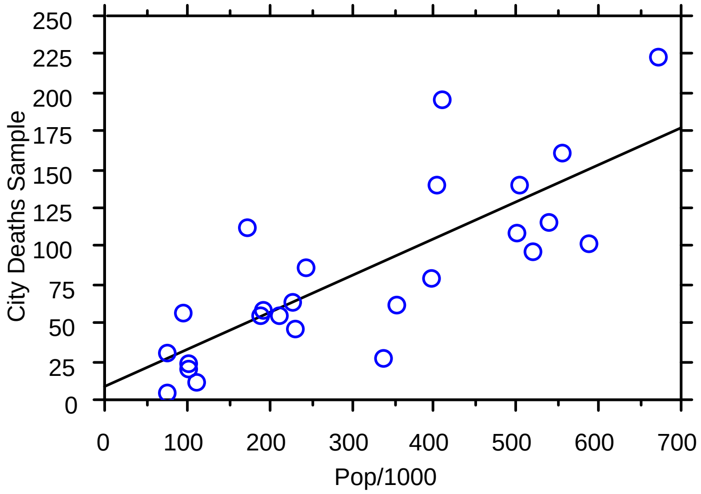
Answer: 2pts

Values for the populations of cities in the table were required to complete this problem. A few minutes of creative web-searching with terms like 122 U.S. cities, CDC, populations, U.S. census, should have yielded websites with the needed information. Here are two examples:

http://www.census.gov/population/estimates/metro-city/SC100K-T1.txt
http://www.census.gov/population/estimates/metro-city/SC10K-T3.txt

No need to use all of the data. Devise a random sampling method:
E.g., Use every 5th entry in Table III (as below); or
Put the city names in a “hat” and draw them out one-by-one; or
Use a computer program to randomly draw numbers from 1-122, as pre-assigned to cities list

Perform a linear regression analysis and determine R^2 and p



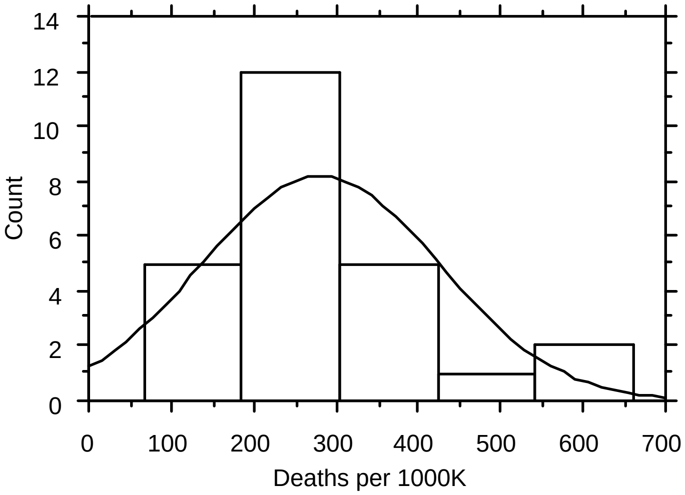
Y = 8.248 + .242 * X; R^2 = .624; p < 0.0001

Yes, the relationship is highly significant (p < 0.0001)

Insight: This analysis indicates that about 62% of the variance in deaths among cities can be accounted for by the fact that they differ in population size. The rest must be due to other factors.

8) Is there anything going on in NYC regarding the risk of dying?

Answer: 1pt
Insight: Risk assessment requires an "incidence" evaluation. For deaths that is mortality rate.
Use your population data to estimate the distribution mortality rate ("deaths in one week per persons") and evaluate whether NYC's rate (259 per 1000K) is significantly different by either CI interval analyses or standard value analysis. Note that the mortality rate distribution is more normal than the distribution for number of deaths, but it is still somewhat skewed.



95% CI = 222-341 (includes NYC value of 259; there for NYC is not significantly different for mortality rate)

p for standard value analysis = 0.4439

There appears to be nothing going on in NYC except that it has a lot of people.

(T-statistic table can be found on page 419 of Schork and Remington; Table A-5)
You may use whatever software packages you have available to complete the assignment, but know how to perform the required calculations and what any presented parameters MEAN (no pun intended).

Statistical Packages Online

Here are web sites that will allow you to generate histograms and calculate distribution parameters. (Of course, these can be done with Excel or other common statistical packages as well.)

To create frequency plots, etc.:
http://dostat.stat.sc.edu/webstat/3.0.1.0/

To calculate distribution parameters:
http://bardeen.physics.csbsju.edu/stats/

Others sources of information and statistical operations
http://www.graphpad.com/quickcalcs/index.cfm
http://faculty.vassar.edu/lowry/t_ind_stats.html