Comments on the Iraqi Data,

Background: Jodie Tillman of the Valley News recently asked me to comment on the following data concerning the wars in Iraq and Afghanistan:

State	Deployment to Iraq and Afghanistan per state	Casualties in Iraq and Afghanistan	Casualty Rate
District of Columbia	594	4	0.67%
Vermont	1,613	9	0.56%
Massachusetts	7,146	27	0.38%
Arizona	9,515	33	0.35%
….	….	…	….
Nevada	5,466	5	0.09%
Florida	62,527	54	0.09%
Hawaii	2,355	2	0.08%
Alaska	8,746	0	0.00%

The whole data set can be found at the end of this document. This data was collect by Jodie Tillman and corresponds to the dates between March 1, 2003 and October 31, 2004. The only thing I know about this data set is °ßAfter much back-and-forth between the Pentagon, I finally ended up with some figures that are accurate,°® Jodie Tillman January 7th 2005. I interpreted the situation as follows: It had been noticed that Vermont°¶s casualty rate was higher than most. For example, from a na_ve point of view this data suggest that service members from Vermont were perhaps subjected to as much as 6 times the risk of casualty as service members from say Florida. This observation led people to the following question:

Question: From a statistical point of view, were Vermont service members truly subjected to a higher risk than other service members? In other words, can Vermont°¶s high casualty rate as seen in the above table be reasonably attributed to chance variation alone? In still other words, is there any statistical evidence suggesting that Vermont service members were subjected to some bias that resulted in placing them at greater risk?

To statistically approach such a question we usually articulate a hypothesis.

The Hypothesis: The risk faced by a service member is independent of the state from which the service member comes from.

Then we test The Hypothesis. The test I used was based on this following data alone:

	Deployment to Iraq and Afghanistan per state*	Casualties in Iraq and Afghanistan*	Casualty Rate
Total	644660	1174	0.0018
Vermont	1,613	9	0.0056

Such a hypothesis test is based on estimating how likely it is that Vermont would have the above casualty rate given a reasonable mathematical articulation of The Hypothesis. The exact test of The Hypothesis is described below, under The Vermont Hypothesis Test. Using this test I was unable to reject The Hypothesis and would be forced to conclude the following answer to the above question:

Answer: It is reasonable to attribute Vermont°¶s casualty rate to chance alone.

Conclusions About The Whole Data Set: There is not enough information in the whole data set to decide whether The Hypothesis is compatible with this data in general. This is because we must include the possibility of multiple casualties in a good model. For example, if one Vermonter is a casualty in a given incident then further Vermont casualties are likely if the Vermonter's are stationed to together in the same unit (since attacks and accidents can result in multiple casualties). Due in part to the large number of national guard and reserve service members deployed in these wars a statistical look at this data would not be complete without acknowledging this.

Technical Comments:

Comment: To really model this situation we need at two parameters. The Casualty Rate

P = Pr(A deployed service member is a casualty)

and parameter capturing the possibility of multiple casualties, as discussed the section Conclusions About The Whole Data Set. For example, we could make a reasonable model using the following parameter:

E = E(Casualties from given state resulting from an incident given this number > 0).

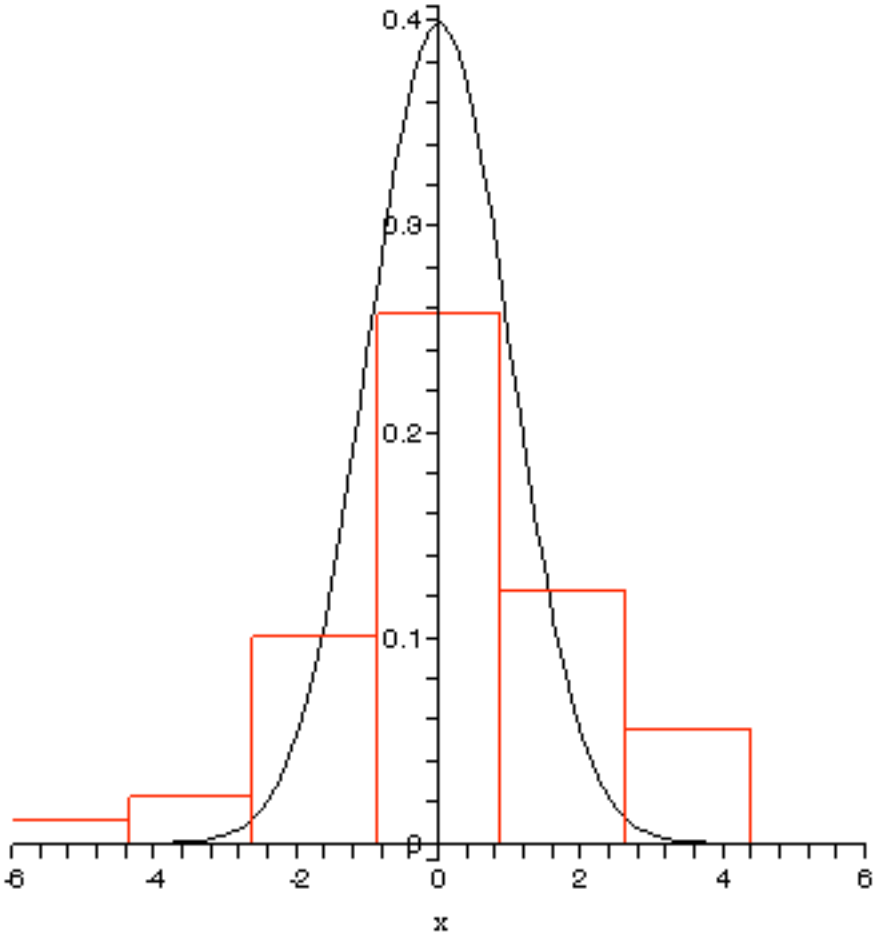
P can be estimated from the data via

P = (Total Casualties)/(Total Deployment) = 1,174/644,660 = 0.0018,

E could be estimated from the data requested in the section The Numbers That Would Be Useful. Assuming E=1 makes the mathematics easier but using E=1 essentially views incidents resulting in a casualties as totally unrelated events with regard to the states and may exaggerate how unlikely the given data is (potentially resulting in an unwarranted rejection of The Hypothesis).

The Vermont Hypothesis: I will assume E=1 and use the above P. This allows me to estimate how likely it is to find 9 **or more** casualties in a sample of 1,613 using the binomial probability distribution, and this is found to be 0.003. Hence, given a collection of 50 states the chance that one or more of these states experiences a casualty rate this high or higher has a probability of (1-(1-0.003)^50) = 0.14. In conclusion, the Vermont data is statistically consistent with the fact that **someone noticed it** and The Hypothesis even in the case where E is assumed to be 1. (Comments: 0.14 is an estimate of the test's P-value. The test was chosen to be one sided since we are much more likely to notice an abnormally high casualty rate than an abnormally low casualty rate. We cannot use the original 0.003 as our P-value since there are 50 states and **t**he event was noticed before the data was analyzed. Hence we must ask what is the chance that someone notices an event this dramatic in 50 trials. Perhaps 50 is a bit of an under estimate, since any city (county, town, etc..) that notices a high casualty rate might also be inclined to question The Hypothesis. But this will only increase our 0.14, so will not change our conclusions.)

Testing The Whole Data Set: Assuming E=1, a statistician could test the whole data set by computing the number of standard deviations from the expect rate of the rates we found using the binomial distribution's standard deviation. I computed these standard deviations and they can be found in the data set at the end of this document under the column standard deviations from the mean. These numbers may seem rather uninteresting to most people, but to anyone with a grasp on statistics they are extremely surprising! The exclamation point is there in part due to the –5.62, which to a statistician is incredibly dramatic evidence in favor of rejecting this mathematical articulation of The Hypothesis with E=1. The chance that we would see something as dramatic as –5.6 in a data set of 50 states is less than 1 in ten million! Under The Hypothesis and E=1 the computed standard deviations from the mean should behave approximately like a *standard normal* (the black curve in the below graph). However the red *histogram* of the actual data is clearly not behaving this way.



There are two possibilities. The Hypothesis is actually false or our E=1 model of the hypothesis is too naïve. It is worth exploring a reasonable model where we can vary E in order to explore how sensitive our standard deviation statistic is to our choice of E. The casualties from the k_{th} state could reasonably be modeled by

$$C_k = \sum_{i=1}^{Bin(P/E,n_i)} Geo(1/E)$$

where Geo(1/E) is the geometric random variable with rate 1/E, and $Bin(P/E,n_k)$ is the binomial with rate P/E, and n_k is the number deployed from the k_{th} state.
In this model, we have that

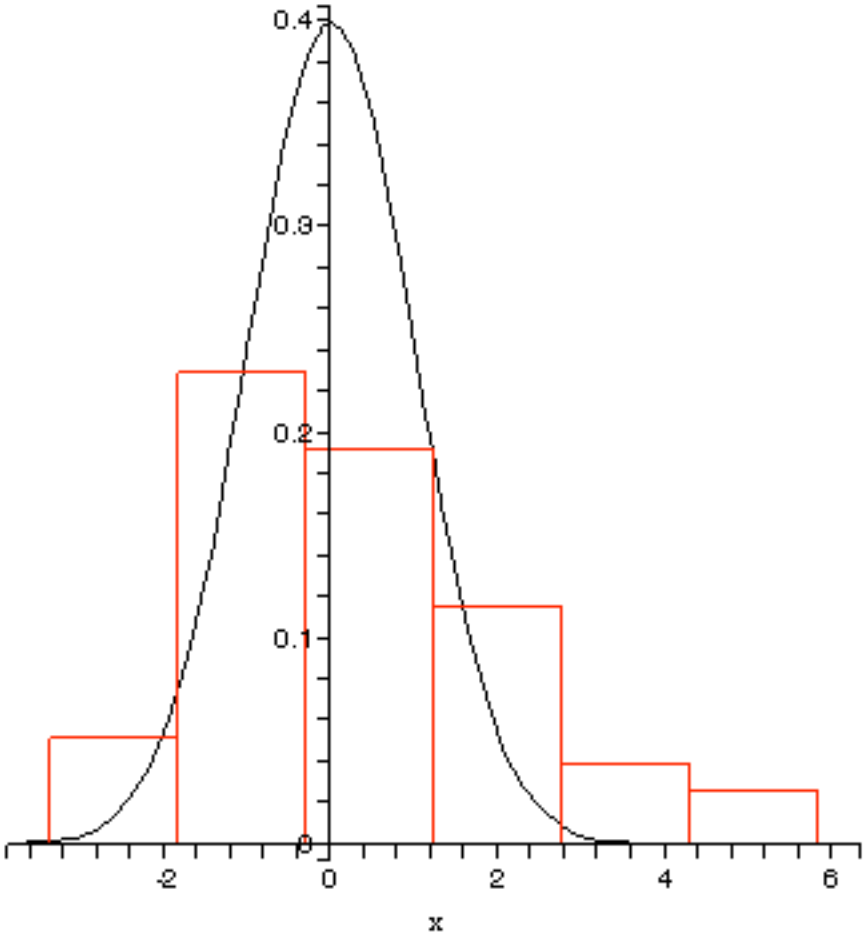
$$E(C_k) = n_k P$$
$$Var(C_k) = n_k P(2E - 1 - P)$$

so

$$K_k = (C_k - n_k P) / \sqrt{n_k}$$

will have mean zero and variance $V = P(2E - 1 - P)$ for any k.

We can simulate this model for say P=0.0018 and E=2 (which seems like a rather high E), and we find:



Certainly this make our original data look a bit less fishy (at least we see how the seemingly abnormally high variance could be accounted for by a more honest model).

Notice, in our data set we have samples of the K_k, and can estimate V, which we find to be to be 0.0063. From the above equation we see

$$E = \left(\frac{1}{2}\right)\left(\frac{V}{P} + 1 + P\right)$$

so we can estimate E to be 2.25. This seem way too high and we could formulate a hypothesis test if we had the following data:

The Numbers I Need: This number I need to really test The Hypothesis could be (with some effort) estimated from available data (see for example http://albertarose.org/fallen/). One would need to look at all incidents in the war between March 1, 2003 and October 31, 2004 that resulted in casualties, and for each incident compute the number of states involved. For most incidents this is likely to equal the number of casualties. However, for example, if an incident results in two Vermont National Guard casualties, then the number corresponding to this incident would be one since only one state is involved (see the 2nd incident listed in the following example table). A good table would look like this.

Incident	Casualties	Number of States Involved	National Guard Casualties	Reserve Casualties	Active Duty Casualties
1		1	10	0	1
2		2	12	0	0
3		2	20	1	1

The sum of casualty column should equal 1174, the total casualties. From this table I could accurately test The Hypothesis. The exact test of The Hypothesis I would use is it really only requires the first 3 columns (though the other data would be VERY useful in a more comprehensive analysis).

A Proposed Hypothesis Test: Now if we could get the data described in The Numbers I Need section, then I can properly estimate E as the ratio of the sum of column 2 and column 3. Call this estimate E_0. Using the model developed above a fair hypothesis test would be to estimate

Pr(V >.0063 | P=2(0.0018), E=E_0).

I would propose to reject the The Hypothesis if this probability is less than 0.01. The most obvious argument against such a result would be the choice of Geo(1/E) in the model. We would need to look at all the data from the The Numbers I Need section to see if such an assumption is truly reasonable. (As well as a break down of the original table according to national guard, reserve, and active duty.)

Confession: When I initially computed the number of standard deviations from the mean, I was expecting Vermont to be the extreme number in this list of standard deviations. I was very wrong! I was initially tempted to reject The Hypothesis based on this standard deviation evidence. However, I decided to try the E=2 and P=0.0018 simulation above, just make sure that the choice of E was not critical to the analysis. Looking at the simulated results I was forced to confess that E is potentially critical.

The Data Set: I have ordered the data set below according to the standard deviations from the expected casualty rates, a statistically meaningful way to order the casualty rates. We have seen (and the above graph illustrates) that this order cannot be explained by chance alone if E is assumed to be 1.

	Deployment to Iraq and Afghanistan per state*	**Casualties in Iraq and Afghanistan***	Standard Deviations	Casualty Rate
Massachusetts	7,146	27.0000	3.8806	0.003
Arizona	9,515	33.0000	3.7683	0.003
California	54,296	136.0000	3.7365	0.002
Vermont	1,613	9.0000	3.5405	0.005
District of Columbia	594	4.0000	2.8084	0.006
Wisconsin	8,459	24.0000	2.1919	0.002
Nebraska	4,435	14.0000	2.0862	0.003
Pennsylvania	26,416	62.0000	2.0049	0.002
Oregon	8,020	22.0000	1.9367	0.002
Indiana	9,441	25.0000	1.8845	0.002
North Dakota	2,333	8.0000	1.8216	0.003
Delaware	1,737	6.0000	1.5964	0.003
Kentucky	5,783	15.0000	1.3782	0.002
Oklahoma	8,662	21.0000	1.3169	0.002
Illinois	23,929	52.0000	1.2770	0.002
South Carolina	10,674	24.0000	1.0355	0.002
Rhode Island	1,870	5.0000	0.8648	0.002
Maine	3,242	8.0000	0.8634	0.002
New Jersey	12,615	27.0000	0.8409	0.002
Connecticut	4,715	11.0000	0.8244	0.002
Arkansas	7,660	17.0000	0.8174	0.002
Colorado	7,343	16.0000	0.7192	0.002

Ohio	20,731	42.0000	0.6917	0.002
Iowa	7,049	15.0000	0.6042	0.002
Mississippi	7,768	16.0000	0.4933	0.002
Idaho	3,233	7.0000	0.4588	0.002
Kansas	5,866	12.0000	0.4034	0.002
Virginia	15,988	30.0000	0.1640	0.001
New York	27,961	52.0000	0.1515	0.001
North Carolina	16,441	30.0000	0.0108	0.001
Georgia	15,473	28.0000	-0.0336	0.001
Michigan	17,709	32.0000	-0.0441	0.001
Maryland	8,441	15.0000	-0.0950	0.001
Wyoming	2,307	4.0000	-0.0983	0.001
New Mexico	4,264	7.0000	-0.2749	0.001
Alabama	12,403	21.0000	-0.3343	0.001
Utah	4,664	7.0000	-0.5130	0.001
Minnesota	7,143	11.0000	-0.5573	0.001
Missouri	13,167	20.0000	-0.8132	0.001
West Virginia	6,609	9.0000	-0.8758	0.001
New Hampshire	4,164	5.0000	-0.9389	0.001
South Dakota	4,905	6.0000	-0.9821	0.001
Tennessee	18,575	28.0000	-1.0028	0.001
Hawaii	2,355	2.0000	-1.1062	0.000
Louisiana	12,905	18.0000	-1.1359	0.001
Montana	4,058	4.0000	-1.2482	0.001
Nevada	5,466	5.0000	-1.5717	0.000
Washington	20,892	26.0000	-1.9548	0.001
Texas	82,352	102.0000	-3.9209	0.001
Alaska	8,746	0.0000	-3.9946	0.000
Florida	62,527	54.0000	-5.6156	0.000

*Between March, 1, 2003 and October 31, 2004

Source