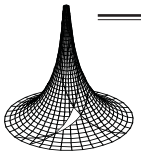


2/2 (2004), 265–273
tmcs@math.klte.hu
http://tmcs.math.klte.hu

Teaching
Mathematics and
Computer Science

Statistical inference in school

Judit Szász-Simon

Abstract. The paper explains a classroom example for convincing students about the utility and applicability of statistical methods in learning getting people's opinions. The emphasis is on convincing instead of proving. The necessary statistical data may be obtained from the Internet as a digital text.

Key words and phrases: short "hands-on" classroom practices in statistics, illustrating opinion polls, convincing instead of proving, sampling, utilizing PC-edited text as a random sample.

ZDM Subject Classification: K70, B50, D40, K90.

1. Introduction

For a statistics teacher the role and importance of sampling is clear. Sampling might be performed for several reasons: first, the total counting could be physically impossible, for example in the case of birds; in other cases it is possible, that during the process of examination, the sample gets destroyed, for example in the case of light bulbs. It is also frequent that the examination of each entity is economically ineffective. TA typical examples for this is are: public opinion polls.

In these cases their case, a random sample is taken to be examined, and the conclusion is applied for all varieties. It is also clear, how and why it functions works. There are two ways to transfer this knowledge to students:

- one may follow the theoretical way of proving by the methods of mathematical statistics

- or by demonstrating the usefulness of sampling by hand-on experiments

The average school-level knowledge is generally not sufficient to follow the first way apply the first method; therefore we should try the second one. This possibility could only be applied only by choosing appropriate situations, experiments, which are not very time-consuming, do not need special skills or experience to perform them, and the "raw material" is always available. Besides, statistical conclusions should be most natural and convincing. This paper presents such an illustrating an example to illustrate this.

2. The statistical task

The skeleton of the sampling problem may be stated on a rather abstract formulation:

> We have a population of size $\boldsymbol{N}$, where $\boldsymbol{M}$ of them is marked and the marked ones can be easily identified. A sample of size $\boldsymbol{n}$ is taken (with equal probability for all members of the population, and there are $\boldsymbol{m}$ marked elements in the sample.

Typical statistical inference problems are that 3 of these 4 parameters are given, the fourth is unknown, and a "good" estimation for this unknown parameter is requested. This skeleton could be dressed on in several ways.

(1) Suppose N is unknown. This is the case, e.g., when we wish want to estimate the number of wild animals. The (nearly) optimal estimation is the integer part of $M \cdot n/m$, $\tilde{N} \approx M \cdot \frac{n}{m}$

(2) Let M be unknown. This is the case in opinion polls. The (nearly) optimal estimation of M is $\tilde{M} \approx N \cdot \frac{m}{n}$

(3) Relatively rare is the case where n (the sample size) is the unknown. The (nearly) optimal estimation is $\tilde{n} \approx N \cdot \frac{m}{M}$

(4) It may be necessary to predict m before sampling. Such problems are important when in planning experiments and preparing guidelines for receiving a shipment of products of many pieces. The (nearly) optimal estimation is $\tilde{m} \approx M \cdot \frac{n}{N}$

For demonstration the purpose of demonstration, we have chosen a type of statistical populations that the readers can generate for himself themselves in suitable the necessary size without huge too much effort. The population includes letters (symbols) from a meaningful written English text written in English (or

in other languages). Such a text may be drawn, for example, from the Internet, as an article in a newspaper, or, from a CD-ROM disc of a novel. Almost all text editor program will tell N (the number of letters) immediately. The teacher (or the students) may choose the marking arbitrarily, say as like all the vowels, or the single letter e. Some special signs (like punctuation marks, numbers, unprintable characters) may be deleted. Then the marked elements are to be counted. So we learn what M exactly is. Then we need a sample of size n or approximately n. Here n could be chosen arbitrarily.

There are several ways for generating random samples. If we are satisfied with "approximately" $\boldsymbol{n}$, then we may define a probability $\boldsymbol{p}$ as $\boldsymbol{p = n/N}$, and choose all characters from the (possibly truncated) text with probability $\boldsymbol{p}$, independently of each other. This choice yields an expected sample size of $\boldsymbol{N \cdot p = n}$. The technical realization on a PC violates the independence: The command RND usually generates pseudo-random numbers, which are not independent (not even random), but although statistical tests show that they are sufficiently random. Once the sampling is performed we may find n and m by direct counting.

Now we know all the four parameters exactly, which allows us the demonstration. Perhaps a better phrase is that this allows the students to have a "real insight" into the nature of statistical inference from a sample. We recall that our goal is to show what kind of and how reliable conclusions can be attained: we cannot expect perfect matching, nevertheless, the "(nearly) optimal estimations" shall be close to the parameter that we wanted to estimate.

3. Working with the PC

As for row material, we have chosen a meaningful English text only consisting of only English letters and spaces, deleting all other characters except the carriage return, which was considered as a special space symbol. After deletion the number of remaining symbols was $\boldsymbol{N = 50\,000}$ character The text consisted of several excerpts from the daily magazine paper *The Times* within the period August–October 2000.

Demo 1

We have considered the vowels as marked elements and found that $\boldsymbol{M = 17\,050}$ (number of vowels). We wanted to take samples of size n, and also the letters to be included in the sample were drawn letter-by-letter. This means that we

wanted to choose all letters independently by according to probability $p = n/N$, yielding in expectation the desired sample size n.

In our own experiments the value of $\boldsymbol{p}$ changed from 0.01 to 0.15 by 0.01, and we repeated the process 10 times with each p. Accordingly, the expected sample sizes are $\mathbf{50.000 \cdot p}$. These are in order: 500, 1000, ..., 7500. Table 1 lists the obtained sample sizes, where the first column shows the value of p, and the next 10 columns show the number of characters in the samples. This is the first occasion for obtaining insight: *The expected sample sizes and the obtained sample sizes can be compared.*

Table 1. 10 sample sizes with various p

p	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0.01	495	472	495	532	462	485	525	498	508	470
0.02	957	1017	1040	1016	990	1007	1021	1019	1000	1042
0.03	1568	1522	1550	1487	1475	1521	1474	1487	1545	1535
0.04	1960	2004	1994	2006	1969	1948	1951	2003	2062	2087
0.05	2376	2451	2516	2437	2582	2527	2537	2572	2486	2490
0.06	3029	2942	3013	3008	2965	2967	3008	3014	2971	2976
0.07	3472	3480	3437	3599	3556	3496	3570	3512	3534	3561
0.08	3993	3927	4076	4098	3995	4008	4035	3838	3957	3926
0.09	4519	4490	4566	4522	4471	4567	4452	4514	4365	4574
0.10	4943	4957	4980	5060	5022	4935	5023	5012	5041	4947
0.11	5489	5556	5620	5562	5418	5497	5519	5508	5589	5511
0.12	5946	6099	5878	5953	6133	5916	6118	6020	5987	5956
0.13	6355	6467	6498	6600	6546	6540	6576	6413	6506	6441
0.14	6978	6945	6853	7076	6924	7059	6901	7012	7080	7116
0.15	7449	7484	7351	7493	7422	7576	7545	7506	7497	7516

We summarize the statistical characteristics of these data in Table 2. These are listed here to illustrate what can be expected from the students work. It can be utilized to demonstrate usual inference principle, like the 3-sigma rule, etc.

We have altogether 150 samples altogether. These correspond to different expected sizes, therefore the expected number of vowels is different in the different rows, but the relative frequencies of the vowels vary around the relative frequency of the vowels in the complete text. Therefore we may investigate their distribution in the 150 samples.

4. Random samples with exactly n elements

We may demand that every subset of n characters should be chosen with the same probability from a population of size N. We may number the elements of the population by 0 through $N-1$, and it suffices to generate a sample of the integers $0, 1, \ldots, N-1$. To this end we can use as many random numbers $U_1, U_2, \ldots, U_K$ as necessary, which are uniformly distributed on the $[0, 1)$ interval. There are two trivial cases:

(1) Case $n = 1$: The integer part of $N \cdot U_1$ gives a random integer between 0 and $N-1$, as desired.

(2) Another trivial case is $n = N$, when all elements are included in the sample.

Table 2. Characteristics of the data in Table 1

$100 \cdot p$	Min	Max	$\bar{x}$	d^2	d	$\bar{x} - 3d$	$\bar{x} + 3d$
1	462	532	494.2	531.5	23.05	425.0	563.4
2	957	1042	1010.9	613.4	24.77	936.6	1085.2
3	1474	1568	1516.4	1138.7	33.74	1415.2	1617.6
4	1948	2087	1998.4	2127.8	46.13	1860.0	2136.8
5	2376	2582	2497.4	4026.3	63.45	2307.0	2687.8
6	2942	3029	2989.3	809.3	28.45	2904.0	3074.6
7	3437	3599	3521.7	2590.9	50.90	3369.0	3674.4
8	3838	4098	3985.3	5942.2	77.09	3754.0	4216.6
9	4365	4574	4504.0	4074.7	63.83	4312.5	4695.5
10	4935	5060	4992.0	2036.7	45.13	4856.6	5127.4
11	5418	5620	5526.9	3273.9	57.22	5355.2	5698.6
12	5878	6133	6000.6	7886.7	88.81	5734.2	6267.0
13	6355	6600	6494.2	5817.7	76.27	6265.4	6723.0
14	6853	7116	6994.4	7737.6	87.96	6730.5	7258.3
15	7351	7576	7483.9	4077.9	63.86	7292.3	7675.5

In the general case we may proceed step-by-step: At first we decide if the first element (i.e. 0) should be included in the sample or not. The decision on a later element (integer) depends on all earlier decisions. Suppose that after the decision

on the K^{th} element (the integer $K-1$), the current, existing sample contains k elements. Then $(n-k)$ elements should be chosen from the remaining $(N-K)$ integers. The decision on the inclusion of the first element in the case of n-out-of-N is a stochastic decision: Let U be a uniformly distributed random variable on $[0,1)$. The first element is included in the sample if $U < n/N$. Proceeding step-by-step, we will reach one of the two trivial cases above. It is a simple exercise in combinatory to show that such a choice will generate a sample with the desired randomness we wanted.

Table 3. Number of vowels in different samples

p	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0.01	203	177	210	216	190	206	210	211	198	209
0.02	405	411	407	399	407	405	409	402	413	424
0.03	638	604	626	563	621	634	610	601	627	635
0.04	798	809	826	802	782	820	777	825	827	881
0.05	972	1014	1025	1007	1059	1087	1001	1090	976	1032
0.06	1239	1159	1218	1223	1237	1214	1280	1237	1221	1185
0.07	1410	1418	1364	1501	1423	1451	1463	1423	1436	1461
0.08	1687	1614	1661	1666	1610	1596	1635	1612	1637	1585
0.09	1844	1874	1838	1814	1805	1828	1886	1823	1778	1902
0.10	2052	1968	2067	2063	1996	1984	2089	2014	2118	2066
0.11	2215	2263	2306	2266	2220	2236	2219	2236	2341	2327
0.12	2448	2520	2389	2386	2539	2437	2530	2437	2483	2460
0.13	2598	2680	2661	2748	2668	2653	2686	2655	2635	2697
0.14	2922	2832	2796	2934	2802	2915	2899	2877	2878	2913
0.15	3062	3016	2934	3084	2979	3063	3122	3016	3072	3051

It may be an educational goal to demonstrate that the two ways of *drawing results similar samples*, so that the extra care effort of generating exactly n element samples does not pay off.

Demo 2

The number of vowels in the complete sample can be easily counted. It is also simple is to find the number of vowels in a given sample. Samples can be generated by each of the two methods explained and compared previously in this paper. Now we provide statistical data on the number of the vowels in different samples generated by the first method (Table 3).

Table 4. Statistical characteristics of data given in Table 3

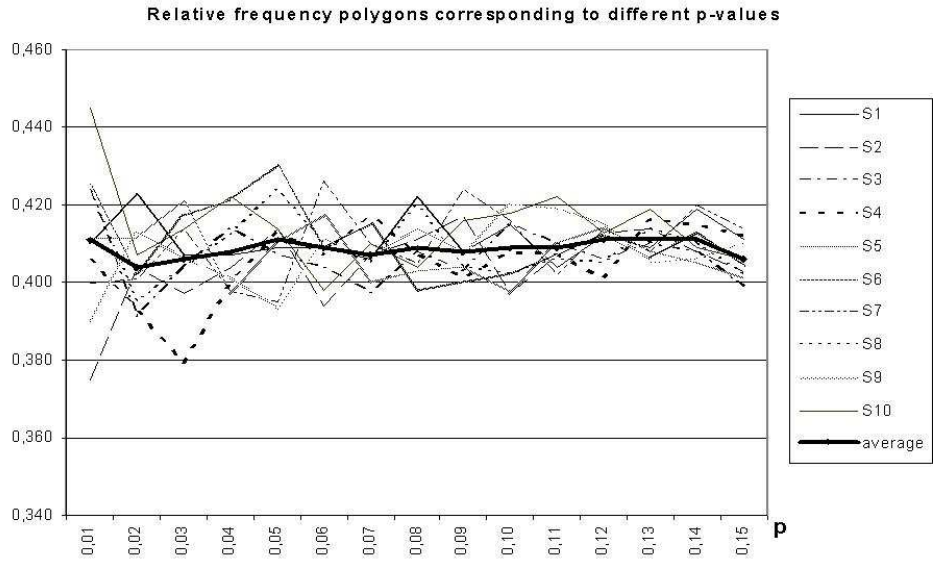
$100 \cdot p$	min	max	$\bar{x}$	d^2	d	$\bar{x} - 3d$	$\bar{x} + 3d$
1	177	216	203.0	138.4	11.77	167.7	238.3
2	399	424	408.2	47.5	6.89	387.5	428.9
3	563	638	615.9	514.3	22.68	547.9	683.9
4	777	881	814.7	863.6	29.39	726.5	902.9
5	972	1090	1026.3	1720.9	41.48	901.8	1150.8
6	1159	1280	1221.3	1057.6	32.52	1123.7	1318.9
7	1364	1501	1435.0	1368.4	36.99	1324.0	1546.0
8	1585	1687	1630.3	1082.2	32.90	1531.6	1729.0
9	1778	1902	1839.2	1478.6	38.45	1723.8	1954.6
10	1968	2118	2041.7	2385.1	48.84	1895.2	2188.2
11	2215	2341	2262.9	2176.1	46.65	2123.0	2402.8
12	2386	2539	2462.9	2989.4	54.68	2298.9	2626.9
13	2598	2748	2668.1	1571.2	39.64	2549.2	2787.0
14	2796	2934	2876.8	2523.3	50.23	2726.1	3027.5
15	2934	3122	3039.9	2991.9	54.70	2875.8	3204.0

It is very instructive to discuss the behavior of the relative frequencies. We summarize our data in Table 5. It is to seen how small the difference among the entries of the table is.

Table 5. Relative frequencies of the individual samples and their averages

P	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	average
0.01	0.410	0.375	0.424	0.406	0.411	0.425	0.400	0.424	0.390	0.445	0.411
0.02	0.423	0.404	0.391	0.393	0.411	0.402	0.401	0.395	0.413	0.407	0.404
0.03	0.407	0.397	0.404	0.379	0.421	0.417	0.414	0.404	0.406	0.414	0.406
0.04	0.407	0.404	0.414	0.400	0.397	0.421	0.398	0.412	0.401	0.422	0.408
0.05	0.409	0.414	0.407	0.413	0.410	0.430	0.395	0.424	0.393	0.414	0.411
0.06	0.409	0.394	0.404	0.407	0.417	0.409	0.426	0.410	0.411	0.398	0.409
0.07	0.406	0.407	0.397	0.417	0.400	0.415	0.410	0.405	0.406	0.410	0.407
0.08	0.422	0.411	0.408	0.407	0.403	0.398	0.405	0.420	0.414	0.404	0.409
0.09	0.408	0.417	0.403	0.401	0.404	0.400	0.424	0.404	0.407	0.416	0.408
0.10	0.415	0.397	0.415	0.408	0.397	0.402	0.416	0.402	0.420	0.418	0.409
0.11	0.404	0.407	0.410	0.407	0.410	0.407	0.402	0.406	0.419	0.422	0.409
0.12	0.412	0.413	0.406	0.401	0.414	0.412	0.414	0.405	0.415	0.413	0.411
0.13	0.409	0.414	0.410	0.416	0.408	0.406	0.408	0.414	0.405	0.419	0.411
0.14	0.419	0.408	0.408	0.415	0.405	0.413	0.420	0.410	0.406	0.409	0.411
0.15	0.411	0.403	0.399	0.412	0.401	0.404	0.414	0.402	0.410	0.406	0.406

It is also instructive to follow the frequency polygons corresponding to the different p-values.



Graph 1. Relative frequency polygons corresponding to different p-values

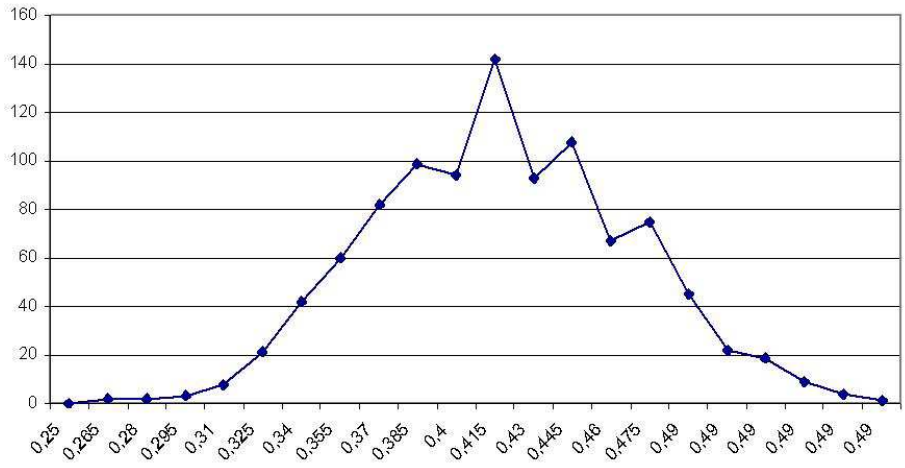
It is the same cost to investigate much larger collection collections of samples. We have generated samples 1000 times with $p = 0.002$. The results are illustrated on Graph 2. (The classes are not detailed here.)

Finally, we would like to add, that the above ideas were presented and experimented in the Fazekas Mihály Secondary School, in Budapest, were where the students enjoyed a heavy serious discussion on the subject. We would be delighted to hear about the findings of any possible experiment.

Summary

This paper discusses the classroom treatment of the following statistical sampling problem:

A sample of size $\boldsymbol{n}$ is taken from a population of size $\boldsymbol{N}$. There are $\boldsymbol{M}$ marked elements in the population, $\boldsymbol{m}$ of them belong to the sample. We attacked the statistical inference problem where 3 of the parameters $\boldsymbol{M}$, $\boldsymbol{m}$, $\boldsymbol{N}$,



Graph 2. Samples with $p = 0.002$. (1000 samples)

$\boldsymbol{n}$ is are given and a – good – estimation for the unknown fourth is requested. Instead of providing giving the a theoretical method of doing this, a convincing experimental classroom demonstration of the solution is provided. The underlying statistical population is a written text, which is easily available for classroom inspection. Some ideas of generating random samples is also explained. Results are tabularized and graphed.

JUDIT SZÁSZ-SIMON
FAZEKAS MIHÁLY SECONDARY SCHOOL
BUDAPEST
HUNGARY

E-mail: simonj@fazekas.hu

(Received February, 2004)