Chapter 35
Data Spread: Standard Deviations, One Sample Z-Test, One Sample Binomial Test

1 General Purpose

With continuous outcome data (Chap. 1), the standard deviation is generally used to estimate the spread in a data sample. The standard deviation is, then, used for multiple purposes like null-hypothesis testing and the computation of confidence intervals. With binary outcome data things are different. Instead of a mean value the number of responders is calculated as a "kind of" mean value. In a data sample with binary outcome (yes-no outcome), the spread is estimated with the equation, $\sqrt{(p(1-p)}$, where p = the proportion of responders, otherwise called the (yes-data fraction versus all data). This chapter assesses how these estimators can be used in practice for testing null-hypotheses and confidence intervals of binary outcome data.

2 Schematic Overview of Type of Data File

Outcome binary
.
.
.
.
.
.
.
.
.

T.J. Cleophas, A.H. Zwinderman, *Clinical Data Analysis on a Pocket Calculator*,
DOI 10.1007/978-3-319-27104-0_35

3 Primary Scientific Question

How do we compute the standard deviation of a binary data set.

4 The Computation of the Proportion of Responders and Its Standard Deviation (SD)

Why is SD of the proportion responders $= \sqrt{(p(1-p)}$? The proportion of responders can be looked at as the "mean" of a yes/no data set.

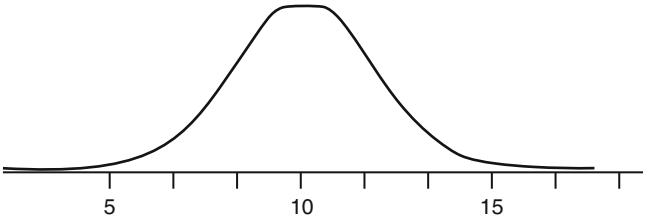
Proportion = mean of yes/no data.
For example, mean of the 6 values [1, 0, 1, 0, 0, 1] is 3/6 yes = 50 % = proportion p.

SD of continuous data $= \sqrt{[\Sigma (a - \bar{a})^2/n}$, where n = sample size.
SD of proportional data $= \sqrt{[p(1-p)]}$, where p = proportion, e.g., 5/15

What does SD of proportion $= \sqrt{[p(1-p)]}$ mean in practice?

We assume for example: on average 10/15 in a random population say yes to the question, are you sometimes sleepy through the day. Then, 10/15 saying yes will be encountered most frequently.

The chance to find answer <10/15 or >10/15 gets gradually smaller.



The above graph has on the x-axis the numbers of patients saying yes, and on the y-axis the chance of finding this number. The chance of 8/15 or less is only 15 %, the chance of 7/15 or less only 2.5 %, the chance of 5/15 or less is only 1 %. With many samples the above graph follows a normal frequency distribution, where the equation $[\sqrt{[p(1-p)]}$ is a very good approximation of its standard deviation. This is how nature works, and it can even be proven to be true with the one sample binomial formula requiring a hypergeometric distribution, but this is beyond the scope of the current work.

5 One Sample Z-Test

Out of a sample of 100 patients only 10 patients were yes-responders. And, so, the proportion of yes responders is 10 % = 0.1. For testing, whether this is significantly different from 0 responders, a standard error of the response is required.

The standard error (SE) can be calculated from the standard deviation according to:

SE = SD/n, where n = sample size.
SE $= \sqrt{[p(1-p)]}/\sqrt{100}$
$= \sqrt{(0.1 \times 0.9)}/10$
$= 0.03$

the z-value is the test statistic and equals [proportion/(its SE)] = 0.1/0.03 = 3.33

The bottom row of the underneath t-table gives p-values from the z-values. With a z-value of 3.33, the p-value, two-tail as usual, should be <0.001. This would mean that the 10 % yes responders is significantly better than a zero response would have been.

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

The t-table has a left-end column giving degrees of freedom (≈ sample sizes), and two top rows with p-values (areas under the curve = p-values), one-tail meaning that only one end of the curve, two-tail meaning that both ends are assessed simultaneously. The t-table is, furthermore, full of t-values, that, with ∞ degrees of freedom, are equal to z-values. The z-values and t-values are to be understood as mean results of studies, but not expressed in mmol/l, kilograms, or proportions of responders, but in so-called SEM-units (Standard error of the mean units), that are obtained by dividing your mean result by its own standard error. For continuous outcome data, with many degrees of freedom (large samples) the curve will be a little bit narrower, and more in agreement with nature. For binary outcome data, nature has determined that the curves will always be as narrow as can be, according to the row at the bottom.

6 Computing the 95 % Confidence Intervals of a Data Set

The example is taken from the Chap. 13. What is the standard error (SE) of a study with events in 10 % of the patients, and a sample size of 100 (n). Ten percent events means a proportion of events of 0.1. The standard deviation (SD) of this proportion is defined by the equation

$$\sqrt{}\,[\text{proportion} \times (1 - \text{proportion})] = \sqrt{}\,(0.1 \times 0.9) = \sqrt{}\,0.09 = 0.3,$$

$$\text{the standard error} = \text{standard deviation} / \sqrt{n},$$
$$= 0.3 / 10 = 0.03,$$

the 95 % confidence interval is given by

$$\text{proportion given} \pm 1.960 \times 0.03 = 0.1 \pm 1.960 \times 0.03,$$
$$= 0.1 \pm 0.06,$$
$$= \text{between } 0.04 \text{ and } 0.16.$$

The 95 % confidence intervals can be used for multiple purposes, for example for noninferiority testing (Chap. 15), and equivalence testing (Chap. 14).

7 Conclusion

With binary outcome data, instead of a mean value the number of responders is calculated as a "kind of" mean value. The spread is estimated with the equation, $\sqrt{}(p(1-p)$, where p = the proportion of responders, otherwise called the (yes-data fraction from all data), in a data sample. This chapter assesses how these estimators can be used in practice for testing null-hypotheses and confidence intervals of binary data.

8 Note

More background, theoretical and mathematical information of standard deviations of binary data is given in Statistics applied to clinical studies 5th edition, Chap. 3, Springer Heidelberg Germany, 2012, from the same authors.