Statistics of Illumination

Allan J. Rossman
Department of Statistics
Cal Poly – San Luis Obispo

arossman@calpoly.edu
http://statweb.calpoly.edu/arossman/

presented for NCTM Conference, Indianapolis
October 2014

"Most people use statistics as a drunk uses a lamppost – more for support than for illumination."

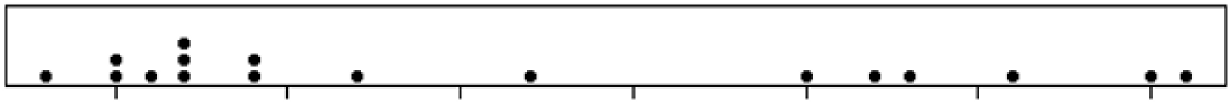
Contrary to this popular saying, statistics is a very important scientific discipline, providing useful tools for shedding light on important issues.

I present 10 examples in which very simple statistical tools provide illumination, where careful thinking about data helps to avoid mistaken conclusions.

Example 1: J_____ T____________ (mystery)
Example 2: Geyser Eruptions
Example 3: Cancer Pamphlets
Example 4: Draft Lottery
Example 5: L_____ E____________ (mystery)
Example 6: Speaking and Intelligence
Example 7: Home Court Disadvantage?
Example 8: Murderous Nurse?
Example 9: Sex Discrimination?
Example 10: AIDS Testing

Example 1: J_____ T____________ (mystery)

Consider the following dotplot:

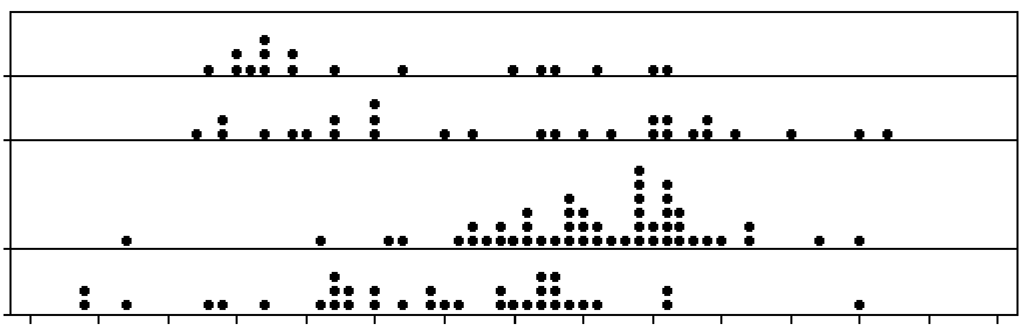


a) Describe what this graph reveals.

b) As more information is provided to you, insert an axis label and scale on the graph above.

c) After you know the context for these data, describe what this graph reveals about the context. Also suggest explanations for the unusual features in the graph based on the context.

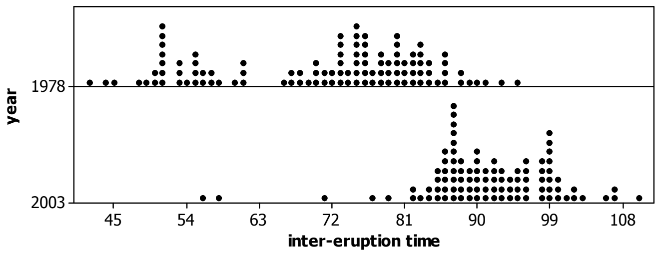
Now consider the following dotplots:

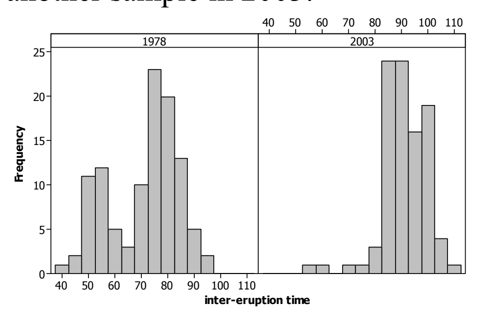


d) As more information is provided, insert axis labels and scales. Then with the benefit of knowing the context, describe what the graphs reveal.

Example 2: Geyser Eruptions

The following graphs display times between eruptions (in minutes) of Old Faithful geyser in Yellowstone National Park, for one sample in 1978 and another sample in 2003:





a) Has the distribution of inter-eruption times changed noticeably between 1978 and 2003? Describe primary differences that you observe between the two distributions. (Comment on center, variability, shape, and unusual observations.)

b) In which year did visitors tend to have a shorter wait for the next eruption? Explain how you decide.

c) Which year showed more predictability (less variability) in how long a visitor would have to wait for the next eruption? Explain how you decide.

d) Based on these inter-eruption times, in which year (1978 or 2003) would you have preferred to be a tourist at Old Faithful? Justify your choice based on the data and your answers above.

Example 3: Cancer Pamphlets

Researchers in Philadelphia investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels of 63 cancer patients and also the readability levels of 30 cancer pamphlets (based on such factors as the lengths of sentences and number of polysyllabic words). These numbers correspond to grade levels, but patient reading levels of under grade 3 and above grade 12 are not determined exactly.

The following tables indicate the number of patients at each reading level and the number of pamphlets at each readability level:

Patients' reading levels	< 3	3	4	5	6	7	8	9	10	11	12	> 12	Total
Count (number of patients)	6	4	4	3	3	2	6	5	4	7	2	17	63

Pamphlets' readability levels	6	7	8	9	10	11	12	13	14	15	16	Total
Count (number of pamphlets)	3	3	8	4	1	1	4	2	1	2	1	30

a) Explain why the form of the data do not allow one to calculate the *mean* reading skill level of a patient.

b) Determine the *median* reading level of a patient and the median readability level of a pamphlet.

Patient:

Pamphlet:

c) How do these medians compare? Are they fairly close?

d) Does the closeness of these medians indicate that the pamphlets are well matched to the patients' reading levels? Explain.

e) What proportion of the patients do not have the reading skill level necessary to read even the simplest pamphlet in the study?

f) Do you want to re-think your answer to d) in light of your answer to e)?

Example 4: Draft Lottery

The following data are the draft numbers (1-366) assigned to birthdates in the 1970 draft lottery. Men born on the date assigned a draft number of 1 were the first to be drafted, followed by those born on the date assigned draft number 2, and so on.

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

a) What draft number was assigned to *your* birthday? Is this draft number in the top third, middle third, or last third of the draft order? Is your draft number 213?

b) Use technology to examine a scatterplot of draft number vs. birthdate number (i.e., let January 1 be 1, January 31 be 31, February 1 be 32, and so on through December 31 as 366). Does the scatterplot reveal any association between draft number and birthdate?

The following table arranges the draft numbers for each month *in order*:

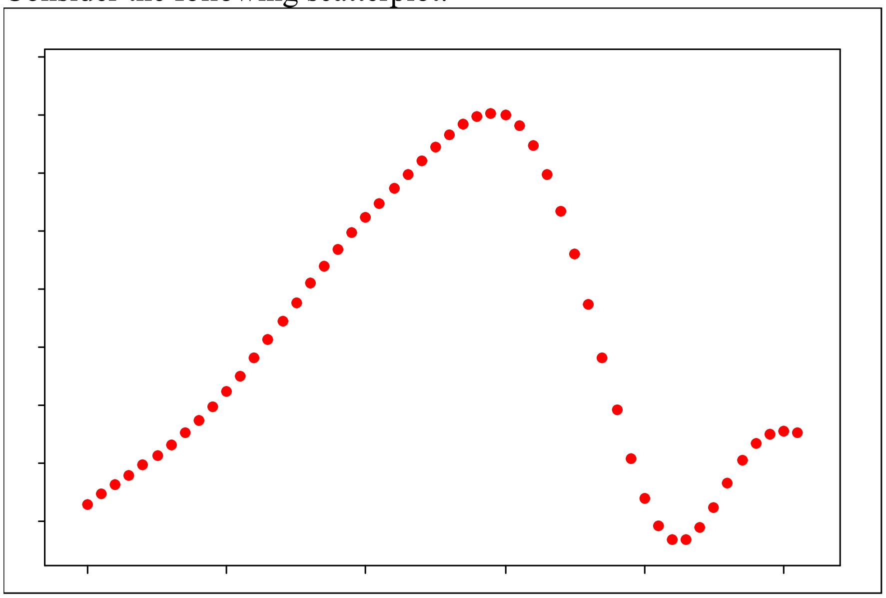
rank	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	17	4	29	2	31	20	13	11	1	5	9	3
2	52	25	30	14	35	22	15	21	6	7	19	10
3	58	57	33	32	37	28	23	36	8	24	34	12
4	59	68	108	62	40	60	27	44	18	38	46	16
5	77	86	122	74	55	64	42	45	49	72	47	26
6	92	89	136	81	65	69	50	48	63	79	51	39
7	101	91	139	83	75	73	67	54	71	87	66	41
8	118	144	166	90	103	85	88	61	82	94	76	43
9	121	150	169	124	112	104	93	102	113	117	80	53
10	140	152	170	147	130	109	98	106	119	125	97	56
11	159	179	200	148	133	110	115	111	149	138	99	70
12	164	181	213	191	155	134	120	114	151	171	107	78
13	186	189	217	208	178	137	153	116	158	176	126	84
14	194	205	223	218	183	180	172	141	160	192	127	95
15	199	210	239	219	197	206	187	142	161	196	131	96
16	211	212	256	231	226	209	188	145	175	201	132	100
17	215	214	258	252	250	222	190	154	177	202	143	105
18	221	216	259	253	276	228	193	167	184	220	146	123
19	224	236	265	260	278	247	227	168	195	229	156	128
20	235	285	267	262	295	249	248	198	204	234	174	129
21	238	290	268	269	296	272	270	245	207	237	182	135
22	251	292	275	271	298	274	277	261	225	241	185	157
23	280	297	293	273	308	301	279	286	232	243	203	162
24	305	299	300	312	313	335	284	291	233	244	230	163
25	306	302	317	316	319	341	287	307	242	254	266	165
26	318	338	323	336	321	353	289	311	246	264	281	173
27	325	347	332	340	326	356	303	324	255	283	282	240
28	329	363	334	345	330	358	322	333	257	288	309	304
29	337	365	343	346	357	360	327	339	263	294	310	314
30	349		354	351	361	366	331	344	315	342	348	320
31	355		362		364		350	352		359		328

c) Use this information to calculate the median draft number for *your* birth month. Is this number in the top half or bottom half of the draft order?

d) Pool the findings of the class and examine the median draft number for each month. Do you notice any tendency in these median draft numbers over time? What does this reveal about whether the draft lottery was conducted fairly?

Example 5: L_____ E____________ (mystery)

Consider the following scatterplot:



a) Describe what this graph reveals.

b) As more information is provided to you, insert axis labels and scales on the graph above.

c) After you know the context for these data, describe what this graph reveals about the context. Also suggest explanations for the unusual features in the graph based on the context.

Example 6: Speaking and Intelligence

The following data are from a study that assessed whether the age (in months) at which a child speaks his/her first word is related to the child's score on the Gesell aptitude test taken later.

a) Before you look closely at the data, do you expect to see a positive association, a negative association, or no association between age of first word and aptitude score? Explain.

Child ID#	Age (months)	Gesell score	Child ID#	Age (months)	Gesell score	Child ID#	Age (months)	Gesell score
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

b) Use technology to produce a scatterplot to see if age of first speaking is of use in predicting Gesell score. Comment on whether there seems to be a relationship/association between these two variables.

c) Based on the scatterplot, make an educated guess for the value of the correlation coefficient between these variables.

d) Use technology to superimpose the least squares line for predicting aptitude score from age of first speaking on the scatterplot. Also determine the value of the correlation coefficient r and the equation of the least squares line. Record these and also the value of r^2. Comment on whether this analysis suggests a relationship between age of first speaking and aptitude score.

e) Do any of the children appear to be outliers in the *age* variable? If so, which child? How long did it take him/her to speak?

f) Remove this unusual child from the analysis. Then look at a scatterplot and report the least squares equation and value of r^2. How have these changed?

g) Now remove the child who took the next longest time to speak, look at a scatterplot, and report the least squares equation and value of r^2. How have these changed?

h) Explain why the results of your analyses change so much depending on whether the unusual children are included. Also summarize what you have learned concerning the relationship between age of first speaking and aptitude.

Example 7: Home Court Disadvantage?

The 2008-09 Oklahoma City Thunder, a National Basketball Association team in its second year after moving from Seattle, found that their win-loss record was actually worse for home games with a sell-out crowd (3 wins and 15 losses) than for home games without have a sell-out crowd (12 wins and 11 losses). (These data were noted in the April 20, 2009 issue of *Sports Illustrated* in the Go Figure column.)

a) Identify the observational units and variables in this study.

Observational units:

Explanatory variable:

Response variable:

b) Organize the given data into a 2×2 table of counts:

	Sell-out crowd	Smaller crowd	Total
Win			
Loss			
Total			

c) Calculate the proportion of wins for each group. When did the team have a higher winning percentage: in front of a sell-out crowd or not?

Sell-out crowd: Smaller crowd:

d) Would you conclude that playing in front of a sell-out crowd causes the team to play worse, or can you think of an alternative explanation for the relationship you've found?

e) Suggest a confounding variable in this study that plausibly explains the observed relationship. Also explain how this confounding variable could account for the relationship.

Example 8: Murderous Nurse?

For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine. Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU (Cobb and Gelbach, 2005). The resulting data are organized in the following 2×2 table:

	Gilbert working on shift	Gilbert not working on shift
Death occurred on shift	40	34
Death did not occur on shift	217	1350

a) Identify the observational units and variables in this study.

Observational units:

Explanatory variable:

Response variable:

b) Notice that the number of shifts with a death was 40 when Gilbert worked and 34 when Gilbert did not work. These numbers seem to be pretty close. Explain what's inappropriate about making this comparison. Also propose how to make a more meaningful comparison.

c) Calculate the proportion of Gilbert shifts in which a death occurred. Then do the same for the non-Gilbert shifts. Do these proportions appear to differ substantially?

d) Calculate the ratio of the proportions in part c), and interpret what this ratio says.

e) Is it reasonable to conclude from these data that Gilbert is responsible for the deaths? Explain.

Example 9: Sex Discrimination?

The University of California at Berkeley was charged with having discriminated against women in their graduate admissions process for the fall quarter of 1973. The two-way table below shows the number of men accepted and denied and the number of women accepted and denied for two of the university's graduate programs (Bickel and O'Connell, 1975).

	Men	Women
Accepted	533	113
Denied	665	336
Total	1198	449

a) Calculate the proportion of men applicants who were accepted and the proportion of women applicants who were accepted. Is there evidence that men were accepted at a much higher rate (proportion) than women? Explain.

Men: Women:

The table below identifies the number of acceptances and denials for both men and women applicants, broken down into the two graduate programs identified as A and F. (Notice that the column totals of the two programs match the counts in the two-way table above.)

	Men		Women	
	Accepted	Denied	Accepted	Denied
Program A	511	314	89	19
Program F	22	351	24	317
Total	533	665	113	336

b) *Within each program*, calculate the proportion of men who were accepted and the proportion of women who were accepted. Did men have the higher rate of acceptance in both programs? Does this seem consistent with your results in (a)? Explain.

Program A Men: Women:

Program F Men: Women:

c) There appears to be a "paradox" in your answers to (a) and (b) – describe it.

d) Using the data provided in the table, explain how the paradox happened in this study and what this means about the issue of whether the university was guilty of sex discrimination.

Example 10: AIDS Testing

The ELISA test for AIDS is used in the screening of blood donations. As with most medical diagnostic tests, the ELISA test is not infallible. If a person actually carries the AIDS virus, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry the AIDS virus, ELISA gives a negative result 92.6% of the time (the *specificity* of the test). Recent estimates are that 0.5% of the American public carries the AIDS virus (the *base rate* with the disease).

a) Suppose that someone tells you that they have tested positive. Given this information, how likely do you think it is that the person actually carries the AIDS virus?

Imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. You will fill in a two-way table as you derive *Bayes' Theorem* to address the question above.

	Positive test	Negative test	Total
Carries AIDS virus	(c)	(c)	(b)
Does not carry AIDS	(d)	(d)	(b)
Total	(e)	(e)	1,000,000

b) Assuming that 0.5% of the population of 1,000,000 people carries AIDS, how many such carriers are there in the population? How many non-carriers are there? (Record these in the table.)

c) Consider for now just the carriers. If 97.7% of them test positive, how many test positive? How many carriers does that leave who test negative? (Record these in the table.)

d) Now consider only the non-carriers. If 92.6% of them test negative, how many test negative? How many non-carriers does that leave who test positive? (Record these in the table.)

e) Determine the total number of positive test results and the total number of negative test results. (Record these in the table.)

f) Of those who test positive, what proportion actually carry the disease? How does this compare to your prediction in a)? Explain why this probability is smaller than most people expect.

g) Of those who test negative, what proportion are actually free of the disease?