Sampling Principles

Ryan Miller

Introduction

- Understanding how data are organized, summarized, and displayed is important, but arguably most important is how they are collected
- All research questions pertain to some target **population**, or comprehensive group of cases
  - In most circumstances it is impossible to collect data on the entire population, instead we must rely on a **sample** or subset of cases
  - A sample must be **representative** of the population in order to produce *reliable conclusions*

Example

- On the next slide I will show you Abraham Lincoln's famous Gettysburg Address
  - I'd like you to choose 5 words as a *representative sample* of all the words in document
  - Next, find the average word length in your sample
  - Report your sample's average word length to the rest of us (I'll display the class results)

The Gettysburg Address

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.

We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living and dead, who struggled here have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here.

It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

Questions

1. What are the observational units and variable in your sample?
2. What are the observational units and variable in the dotplot we constructed?

Answers

1. In your sample, the observational units are individual words, the variable is their length (in letters)
2. Our class dotplot, the observational units are samples of 5 words, the variable is the sample average

It is important to keep these two concepts distinct. The former is a **sample**, or a subset of cases from the population. The second relates to the **sampling distribution**, which displays many *different samples* from the population.

Representative Samples

- For Gettysburg address it is feasible to study the entire population
  - The population's mean word length, in *statistical notation*, is $\mu = 4.295$
- Why did our sample means differ from the true population mean?

Representative Samples

- For Gettysburg address it is feasible to study the entire population
  - The population's mean word length, in *statistical notation*, is $\mu = 4.295$
- Why did our sample means differ from the true population mean?
  - There are few possible explanations, but one important reason why a sampling procedure might not reflect the population is **sampling bias**, or a systematic tendency that makes some cases more likely to end up in a sample than others

Discussion

1. In what ways could you have introduced *sampling bias* when choosing your 5 words?

Discussion

1. In what ways could you have introduced *sampling bias* when choosing your 5 words?
2. Supposed you were blindfolded and chose the words in your sample by throwing darts at a copy of the Gettysburg Address hung upon a wall, would this sampling procedure be un-biased?

Discussion (Some common answers)

1. Choosing more interesting words instead of short words like “a” or “by”, choosing more important words which tend to be longer, choosing words that take up more visual space, etc.
2. No, this would be biased towards words that take up the most space on the page

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?

Simple Random Sampling

- The *ideal sampling procedure* is **simple random sampling**, a protocol where each case in the target population has an identical chance of ending up in the sample
- Do you think it would be easy or hard to collect a simple random sample of Xavier students?
  - It would actually be quite hard since the University is unlikely to give you a list of all enrolled students
  - Often times we attempt to get representative samples in other ways

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are not random, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping people who walked by

Other Sampling Protocols

- A **convenience sample** is exactly what the name suggests, a sample that is easily collected (ie: low monetary or time costs)
  - Convenience samples are not random, but they can be representative if carefully selected
  - You might be able to get a representative sample by standing near the center of campus on a typical day and stopping people who walked by - A **stratified sample** is a more complex scheme where the population is broken into similar subcategories, which are sampled separately (typically simple random sampling)
  - The analysis methods we cover will need extensions in order to apply to this type of data
  - Nevertheless, we should be able to recognize it for precisely that reason

Sampling and Statistical Inference

- A *fundamental goal* of statisticians is to use information from a sample to make *reliable* statements about a population
  - This idea is called **statistical inference**

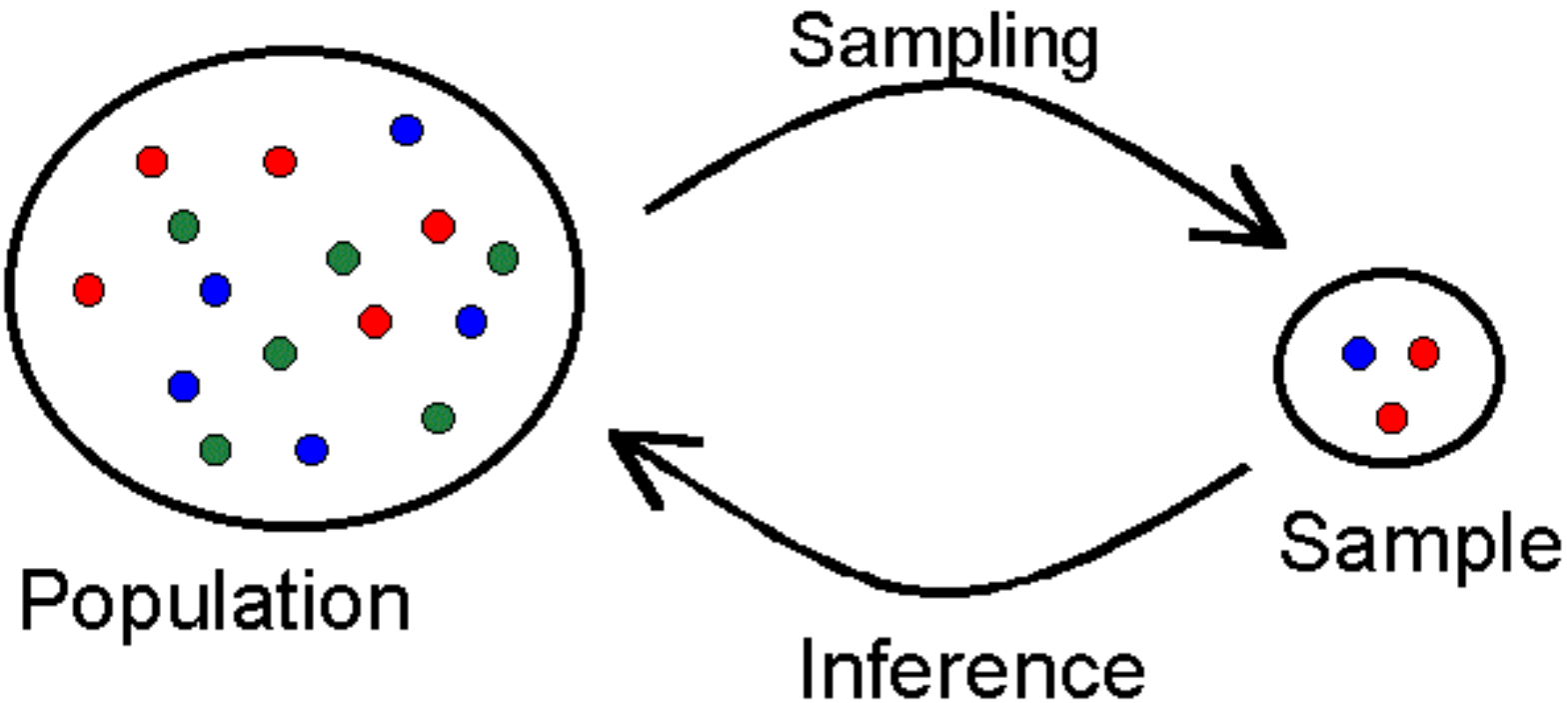


Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

Statistical Inference - Notation

Statisticians use different notation to distinguish *population parameters* (things we want to know) from *estimates* (things derived from a sample). For a few common measures, this notation is summarized below:

	Population Parameter	Estimate (from sample)
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r

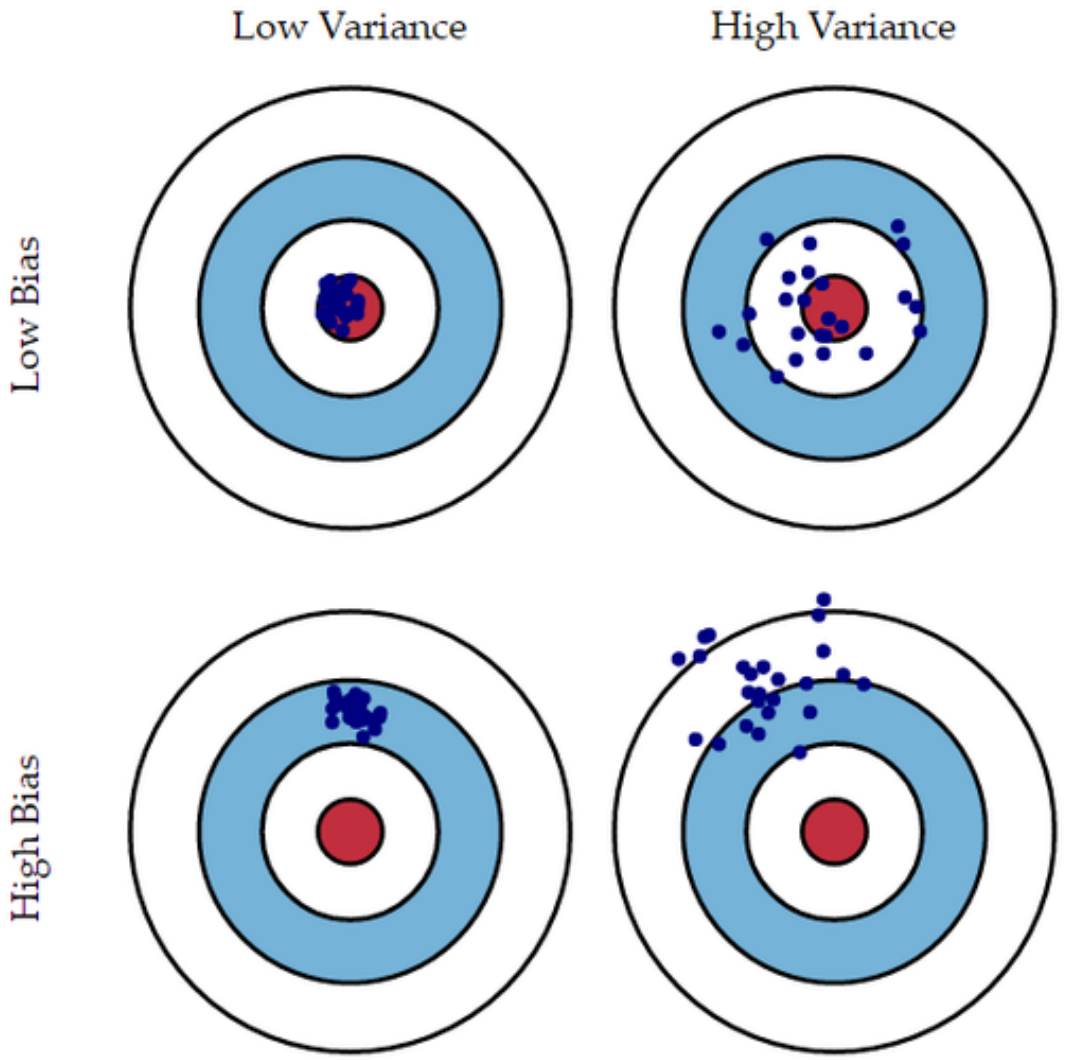
For example, μ is the mean of the target population, while $\bar{x}$ is the mean of the cases that ended up in the sample.

Sampling Variability

- Ideally, we are working with a simple random sample, so the *uncertainty* introduced by the random sampling scheme is the *only explanation* for sample estimates differing from their population parameters
- Put differently, very few samples should be expected to produce estimates that *exactly* equal the population parameter
  - The randomness involved in which cases were selected leads to variability in estimates from different samples

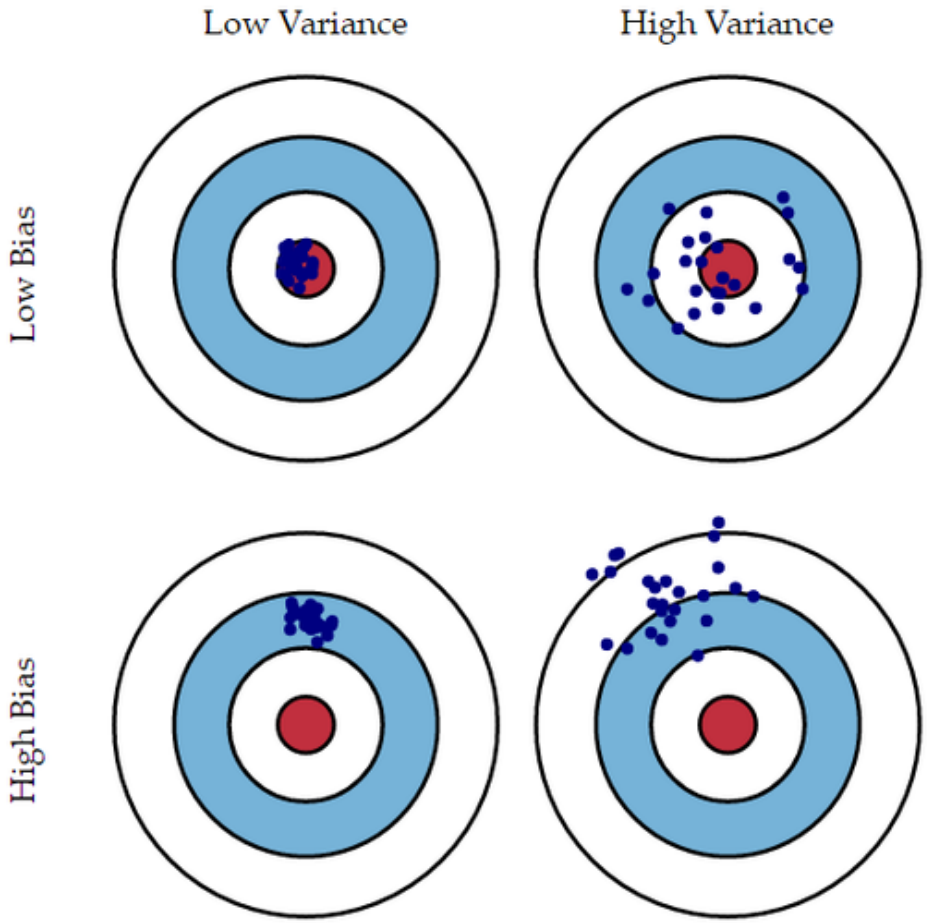
Bias and Variability

Thus, we've now discussed *two reasons* why an estimate might not accurately reflect a population parameter, **bias** and **variability**:



The Role of Sample Size

- The effects of sampling variability are reduced by selecting larger samples
- The effects of sampling bias will remain or might even be exacerbated by selecting larger samples



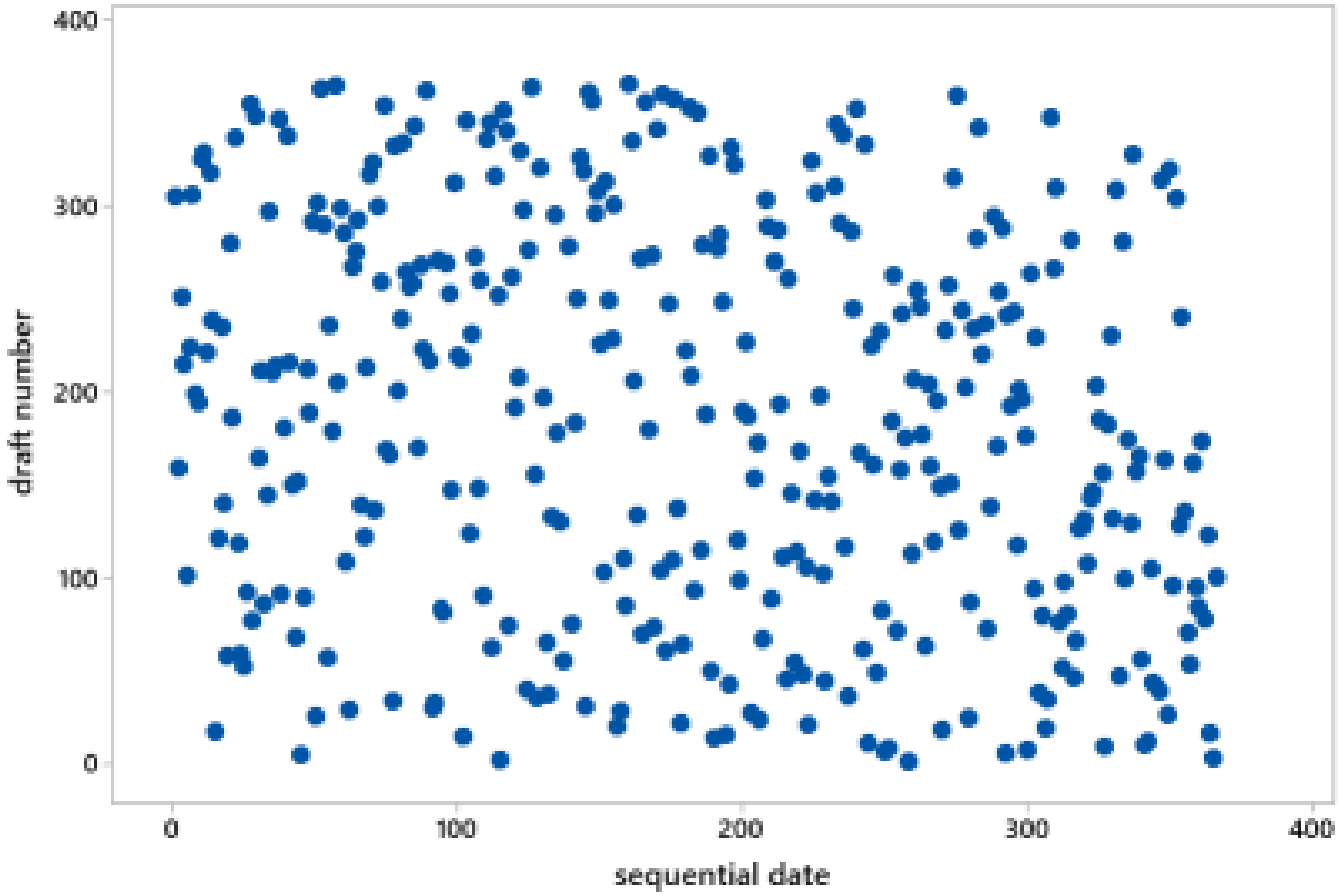
Digression - The Vietnam Draft

- During the Vietnam War, the US government conducted a draft lottery of all eligible adult men serve in the armed forces
  - The lottery was based on birthdays and televised on Dec 1, 1969
  - 366 capsules placed into a bin for a host to draw one-at-time, the results are shown below:

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

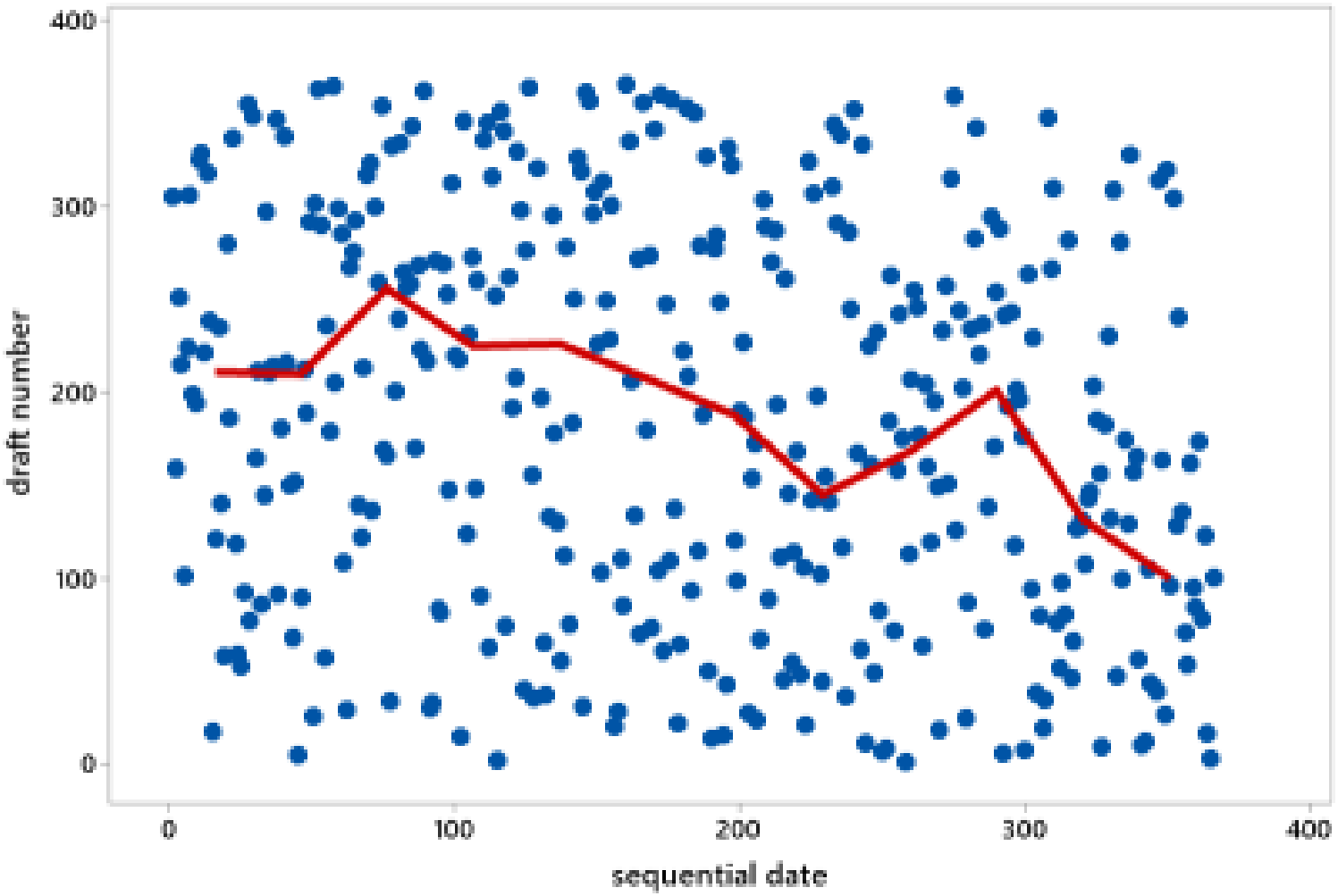
Digression - The Vietnam Draft

- A potentially better display of the draft results is the scatterplot below
  - Does the sampling procedure used in the lottery appear to be fair?



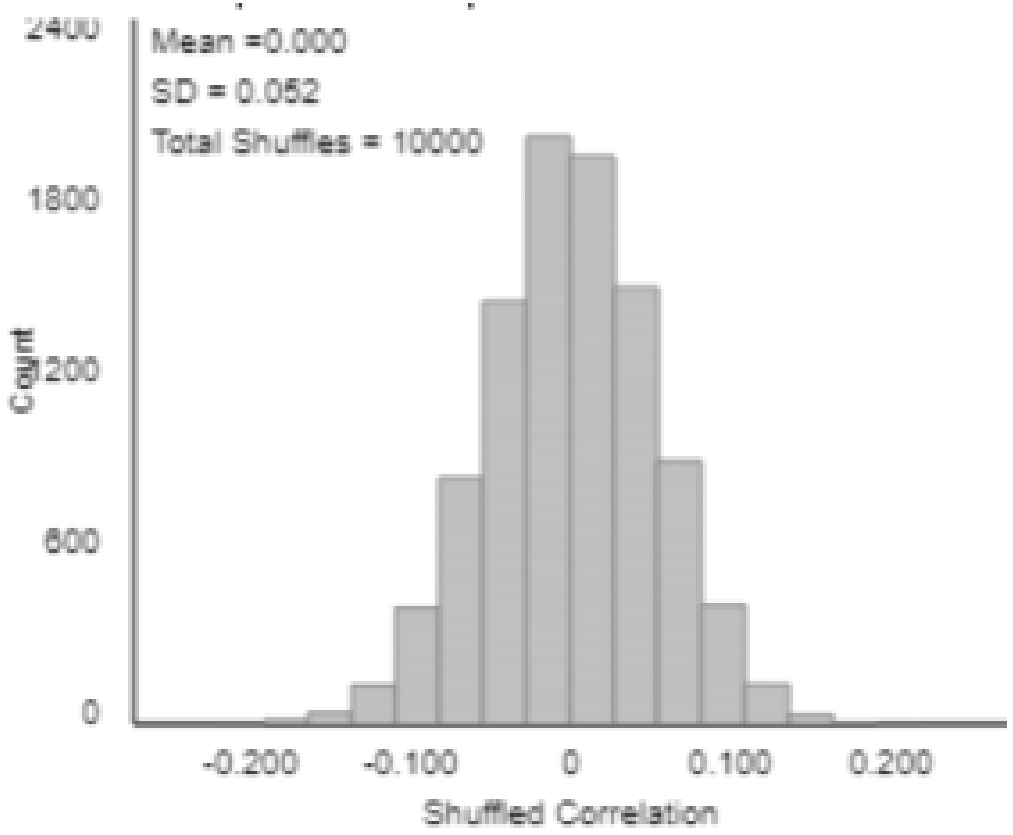
Digression - The Vietnam Draft

Statisticians weren't so sure. . .



Digression - The Vietnam Draft

- The observed correlation coefficient between day and draft number was -0.226
  - The histogram below shows the correlation coefficients from 10,000 simulated random lotteries
  - Do you think the draft was fair?



Digression - The Vietnam Draft

- If the lottery were fair, it would be *extremely unlikely* to see this relationship between later birthdays to have lower drafter numbers
  - The p-value is less than $1/10000$

Digression - The Vietnam Draft

- If the lottery were fair, it would be *extremely unlikely* to see this relationship between later birthdays to have lower drafter numbers
  - The p-value is less than $1/10000$
- Many believe this is an artifact of bias in the sampling procedure
  - The capsules were added to the bin sequentially and likely weren't mixed properly
  - This put the earlier dates closer to the bottom and made them less likely to be drawn early on

Conclusion

- The process by which data are collected is almost always *more important* than what the statistician does to analyze it when it comes to making reliable conclusions
  - It is very difficult for a data analyst to retro-actively address sampling bias in a suitable manner, and any methods for doing so well beyond the scope of this course
- The lesson is to always carefully consider the representativeness of your data before generalizing from them