

**Did you sample the relevant population?**

**Why are handsome men such jerks?**

**Charles S. Barnett**

**Adjunct Mathematics Instructor (retired)**

**Las Positas College**

**Presented at the 43<sup>st</sup> annual fall conference  
California Mathematics Council, Community Colleges  
Monterey, California  
December 11 and 12, 2015**

## Abstract

I will discuss four counterintuitive situations that arise in probabilistic and statistical contexts. One appears in some elementary texts, but I will add a twist to it that leads to a rather startling result. The other cases are less well known; some have important real-world implications. Enliven our discussion by bringing your favorite paradox.

# Attribution

Ellenberg, Jordan, *How Not To Be Wrong* (Penguin Press, New York, 2014)

Shermer, Michael, "Surviving Statistics," *Scientific American*, September 2014

Feller, William, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edition (John Wiley & Sons, NY, 1971)

Karlin, Samuel and Howard M. Taylor, *A First Course in Stochastic Processes*, 2nd edition (Academic Press, NY, 1975)

Various online articles. Phrases such as "probabilistic paradoxes" and "statistical biases" yield a deluge of relevant results. *Caveat emptor*.

Some of what follows is original in the following sense:

*"Originality is the fine art of remembering what you hear but forgetting where you heard it."*  
(Laurence J. Peter, Canadian-American educator)

## Topics to be covered

- The sibling-gender problem and a generalization thereof
- Berkson's Paradox
- The Inspection Paradox
- Survival Bias

# **The Sibling-Gender Problem**

## Assumptions for the sibling-gender probabilistic models

- The standard model

Population of interest consists of two-child families.

Sexes of the first-born and second-born are independent.

At each birth, the probability of a boy equals the probability of a girl equals  $1/2$ .

- The additional-descriptor model

With probability  $p$  ( $0 < p < 1$ ) a boy child possesses an additional attribute. The nature of the attribute is not important, but a popular one is: “The boy was born on a Tuesday and  $p = 1/7$ .” I will analyze the general case in which the attribute remains unspecified and  $p$  remains a variable.

## Two similar-sounding questions (Textbook examples)

**Assumption: Standard sibling-gender model**

Question 1: You randomly select a 2-child *family* and announce that one child is a boy. What is the probability that the other child is a boy?

Answer:  $(1/3)$

Question 2: You randomly select a *child* from the population of offspring of 2-child families; the child is a boy. What is the probability that your selection has a brother?

Answer:  $(1/2)$

The answer to Question 1 surprises many students. It surprised me on first exposure.

*Outcome space for the standard sibling-gender problem*

		(1/2)	(1/2)	
Descriptor of second born	<b>G</b>	BG	GG	(1/2)
		(1/4)	(1/4)	
	<b>B</b>	BB	GB	(1/2)
		(1/4)	(1/4)	
		<b>B</b>	<b>G</b>	
		Descriptor of first born		

*Standard Model*



*A simple but informative illustration of the importance of knowing what population should be sampled to answer the question of interest.*

		(1/2)	(1/2)	
Descriptor of second born	<b>G</b>	BG	GG	(1/2)
		(1/4)	(1/4)	
	<b>B</b>	BB	GB	(1/2)
		(1/4)	(1/4)	
		<b>B</b>	<b>G</b>	
		Descriptor of first born		

### ***Standard Model***

Example 1. Sample a *family* (sample a *cell*) and announce that one component is B. What is the probability that the other component is a B? The event of interest is **E** where **E**={BB}. The conditioning event is **C** where **C**={BG, BB, GB}. Hence,

$$P(\mathbf{E} | \mathbf{C}) = (1/4) / (3/4) = 1/3.$$

Example 2. Sample a *child* [sample a *row* or (exclusive) a *column* – the researcher cannot know whether it is a row or a column] and announce that one component is a B. What is the probability that the other component is a B? Again the event of interest is **E**={BB}. If the sample came from a column, the conditioning event is **C**={BG, BB}. Hence,

$$P(\mathbf{E} | \mathbf{C}) = (1/4) / (2/4) = 1/2.$$

If the sample came from a row, symmetry shows that the same result obtains.

## Sibling-gender problem with a twist

First, recall the standard problem and its solution.

Sample a family from the population of two-child families.

Announce that *one child is a boy*. What is the probability that the other child is a boy? Ans:  $(1/3)$

Now, a common statement of the new problem:

Sample a family from the population of two-child families.

Announce that *one child is a boy that was born on a Tuesday*. What is the probability that the other child is a boy? Ans:  $(13/27)$ ; almost  $(1/2)$  instead of  $(1/3)$ .

I will pose a more general version in which the additional attribute need not be day-of-the-week birthday and the probability need not equal  $(1/7)$ .

Let's call it the Additional-Descriptor Model

## Additional-Descriptor Model

At each birth three possibilities exist:  $B_T$ ,  $B$ ,  $G$ , where

$B_T$  = boy who possesses additional attribute

$B$  = boy who does not possess additional attribute

$G$  = girl

The corresponding probabilities are

$$P(B_T) = (1/2)p, \quad P(B) = (1/2)(1 - p), \quad P(G) = (1/2),$$

where  $0 < p < 1$ .

These assignments lead to the outcome space shown on the next visual.

*Outcome space for the Additional-Descriptor Model*

		$(1/2)(p)$	$(1/2)(1-p)$	$(1/2)(1)$	
Descriptor of second born	<b>G</b>	$B_T G$	$B G$	$G G$	$(1/2)(1)$
		$(1/4)p$	$(1/4)(1-p)$	$(1/4)$	
	<b>B</b>	$B_T B$	$B B$	$G B$	$(1/2)(1-p)$
		$(1/4)p(1-p)$	$(1/4)(1-p)^2$	$(1/4)(1-p)$	
	<b>B<sub>T</sub></b>	$B_T B_T$	$B B_T$	$G B_T$	$(1/2)(p)$
		$(1/4)p^2$	$(1/4)p(1-p)$	$(1/4)p$	
		<b>B<sub>T</sub></b>	<b>B</b>	<b>G</b>	
		Descriptor of first born			

*Additional-Descriptor Model*

## Calculation of probability of two boys, given that one (or more) possesses additional attribute T

Let E represent the event of interest and C represent the conditioning event

Then  $E = \{B_TB, B_TB_T, BB_T, BB\}$  and  $C = \{B_TG, B_TB, B_TB_T, BB_T, GB_T\}$

The next visual shows the conditioning event in tabular form.

Consult that visual, invoke the definition of conditional probability

$$P(E | C) = P(E \cap C) / P(C) ;$$

Do the math and find that

$$P(E | C) = (2 - p) / (4 - p) , \quad 0 < p < 1$$

Observe that  $P(E|C)$  tends to  $(1/2)$  as  $p$  tends to 0 and that  $P(E|C)$  tends to  $(1/3)$  as  $p$  tends to 1.

Sample a family. Announce that one is a  $B_T$ . What is the probability that the other is a boy?  
 Shaded cells show the conditioning event.

		$(1/2)(p)$	$(1/2)(1-p)$	$(1/2)(1)$	
Descriptor of second born	$G$	$B_T G$	$BG$	$GG$	$(1/2)(1)$
		$(1/4)p$	$(1/4)(1-p)$	$(1/4)$	
	$B$	$B_T B$	$BB$	$GB$	$(1/2)(1-p)$
		$(1/4)p(1-p)$	$(1/4)(1-p)^2$	$(1/4)(1-p)$	
	$B_T$	$B_T B_T$	$BB_T$	$GB_T$	$(1/2)(p)$
		$(1/4)p^2$	$(1/4)p(1-p)$	$(1/4)p$	
		$B_T$	$B$	$G$	
		Descriptor of first born			

*Additional-Descriptor Model*

Sample a family. Announce that one is a  $B_T$ . What is the probability that other is a boy?  
 Limit as  $p$  tends to 0.

		$(1/2)(p)$	$(1/2)(1-p)$	$(1/2)(1)$	
<b>G</b>	Descriptor of second born	$B_T G$	BG	GG	$(1/2)(1)$
		$(1/4)p$	$(1/4)(1-p)$	$(1/4)$	
<b>B</b>	Descriptor of second born	$B_T B$	BB	GB	$(1/2)(1-p)$
		$(1/4)p(1-p)$	$(1/4)(1-p)^2$	$(1/4)(1-p)$	
<b><math>B_T</math></b>	Descriptor of second born	$B_T B_T$	$B B_T$	$G B_T$	$(1/2)(p)$
		$(1/4)p^2$	$(1/4)p(1-p)$	$(1/4)p$	
		<b><math>B_T</math></b>	<b>B</b>	<b>G</b>	
		Descriptor of first born			

*Additional-Descriptor Model*

The shaded cells dominate as  $p$  tends to 0, which shows that the probability of two boys tends to  $1/2$ .

*Sample a family. Announce that one is a  $B_T$ . What is the probability that other is a boy?  
Limit as  $p$  tends to 1.*

		$(1/2)(p)$	$(1/2)(1-p)$	$(1/2)(1)$	
<b><i>G</i></b>	$B_T G$	$(1/4)p$	$(1/4)(1-p)$	$(1/4)$	$(1/2)(1)$
	$B_T B$	$(1/4)p(1-p)$	$(1/4)(1-p)^2$	$(1/4)(1-p)$	
<b><i>B</i></b>	$B_T B_T$	$(1/4)p^2$	$(1/4)p(1-p)$	$(1/4)p$	$(1/2)(p)$
	$BB_T$				
		<b><i>B<sub>T</sub></i></b>	<b><i>B</i></b>	<b><i>G</i></b>	

Descriptor of first born

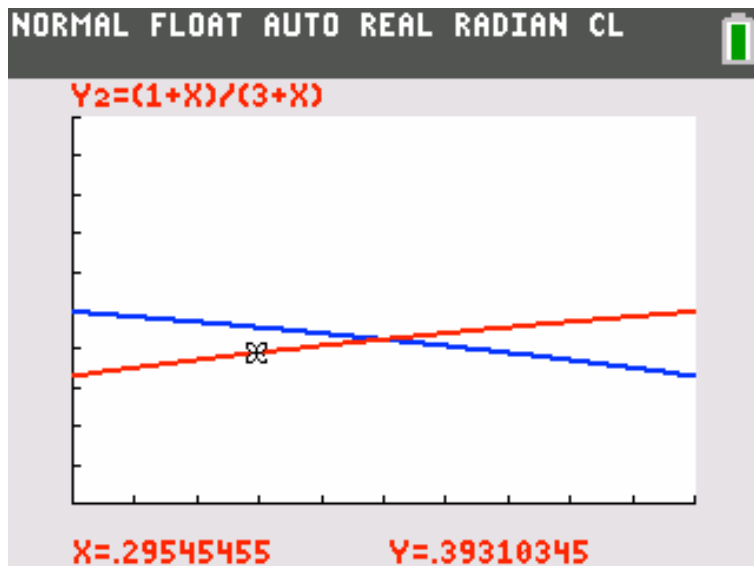
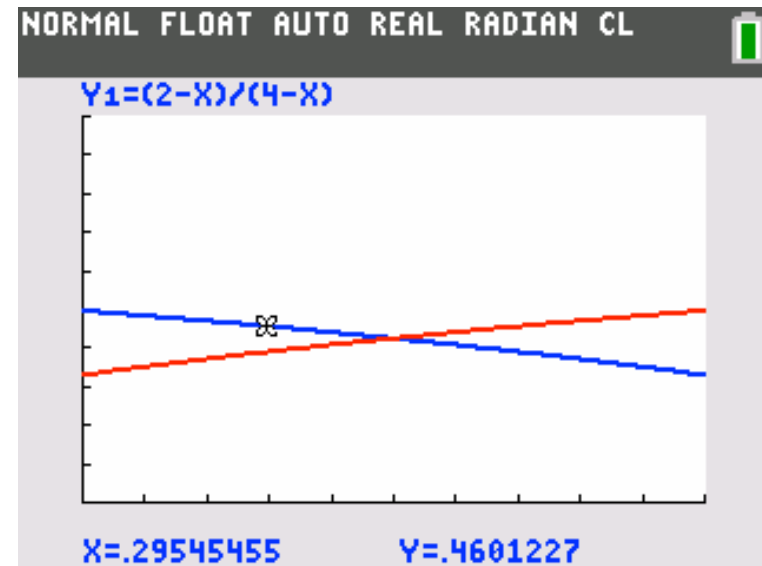
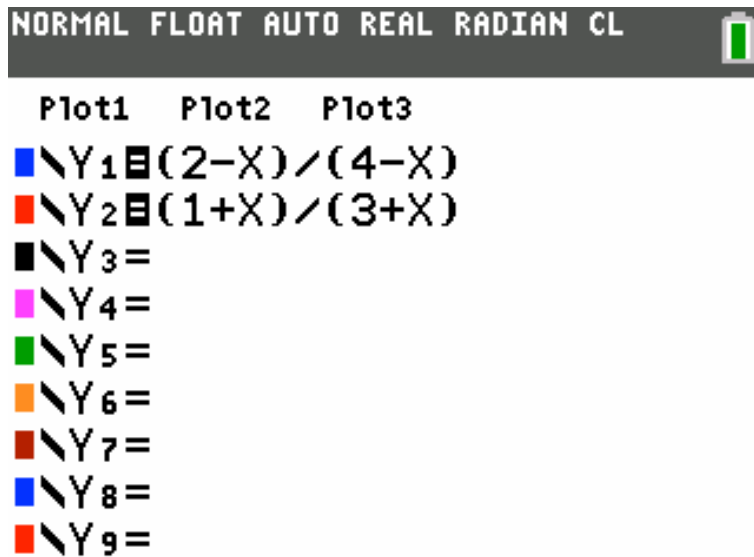
Descriptor of second born

*Additional-Descriptor Model*

*The shaded cells dominate as  $p$  tends to 1 which shows that the probability of two boys tends to  $1/3$ .*



$x$  here is our  $p$  (probability that a boy possesses an additional descriptor)



Select a family; announce that one is a  $B_T$ . What is the probability that the other child is a boy?  $Y_1$  answers that question.

Select a family; announce that one is a  $B$ . What is the probability that the other child is a boy?  $Y_2$  answers that question.

Observe that  $Y_2$  is a reflection of  $Y_1$  about  $x=1/2$

# **Berkson's Paradox**

## **Berkson's Paradox or Fallacy**

- Informal statement
- Mathematical statement and proof
- Some implications

## **Informal statement of Berkson's Paradox**

Two independent events become negatively dependent given that at least one occurs.

# Mathematical statement of Berkson's Paradox

## Hypotheses

1.  $(\Omega, \Sigma, P)$  is a probability space.

$\Omega$  = outcome set;  $\Sigma$  = event set;  $P$  = probability function

2.  $A \in \Sigma$  and  $B \in \Sigma$  are independent events.

3.  $C \equiv A \cup B$ .  $C$  is the conditioning event.

4.  $P_c(E) \equiv P(E|C)$  where  $E \in \Sigma$  is a generic event.

## Conclusion

$P_c(A|B) < P_c(A)$  and  $P_c(B|A) < P_c(B)$ .

The events inhibit each other in the conditioned space but are independent in the parent space.

## Proof of Berkson's Paradox

$$\begin{aligned} P_c(A|B) &= \mathbf{1} \frac{P_c(A \cap B)}{P_c(B)} \\ &= \mathbf{2} \frac{P(A \cap B) / (P(A \cup B))}{P(B) / P(A \cup B)} \\ &= \mathbf{3} \frac{P(A \cap B)}{P(B)} \\ &= \mathbf{4} P(A|B) \\ &= \mathbf{5} P(A) \\ &< \mathbf{6} \frac{P(A)}{P(A \cup B)} \\ &= \mathbf{7} P_c(A) \end{aligned}$$

1 Definition of conditional probability

2 Relation between  $P_c$  and  $P$

3 Clear

4 Definition of conditional probability

5 A and B are independent in the parent space.

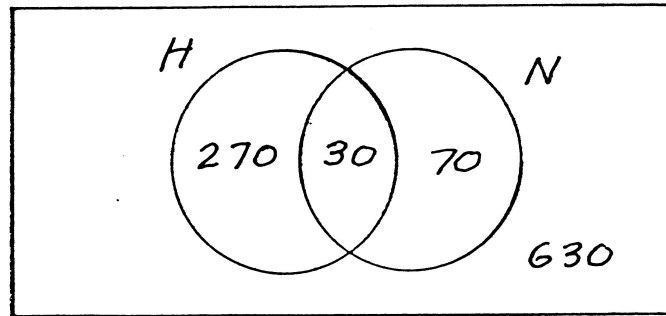
6 Clear

7 Relation between  $P_c$  and  $P$

## An implication of Berkson's Paradox: an answer to the subtitle of the presentation (1)

Situation: Your dating pool contains 1000 men. Of these, 300 are handsome, 100 are nice, 30 are handsome and nice, and the remaining 630 are neither handsome nor nice.

The following Venn diagram illustrates your situation:

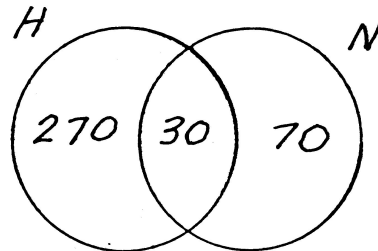


Observe that  $P(H|N) = \frac{n(N \cap H)}{N} = \frac{30}{100} = \frac{300}{1000} = \frac{n(H)}{n(\Omega)} = P(H)$

So H and N are *independent* events in your dating pool.

## An implication of Berkson's Paradox: an answer to the subtitle of the presentation (2)

But you have high standards: you will consider dating only handsome or nice men. So your dating pool shrinks to  $H \cup N \equiv C$  and your Venn diagram changes to



Observe that in this restricted probability space

$$P_c(H|N) = P(H|N) = 0.3 < 0.81 \cong \frac{300}{370} = P_c(H) \quad \text{and}$$

$$P_c(N|H) = P(N|H) = 0.1 < 0.27 \cong \frac{100}{370} = P_c(H).$$

So, in your acceptable dating pool, N inhibits H by almost a factor of 3 and H inhibits N by the same factor.



## **An implication of Berkson's Paradox: an answer to the subtitle of the presentation (3)**

So, what is going on? Is there a *causal* connection between handsomeness and niceness, or a hidden variable that induces a correlation?

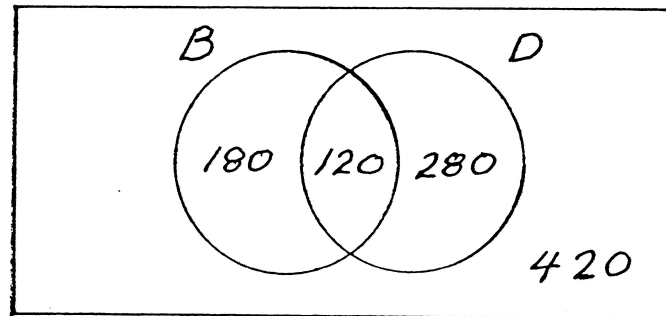
Neither: the correlation is the result of an *effect*: your good taste in men. Your high standards place only handsome or nice men into your acceptable dating pool.

## Another implication of Berkson's Paradox, less entertaining but more important than the previous one (1)

You want to test the conjecture that hypertension is a risk factor for diabetes, so you survey patients in a local hospital. Unexpectedly, you find that high blood pressure is *less* prevalent among the diabetics than the non-diabetics. So, do you advise diabetics to increase their intake of blood-pressure-increasing foods? Before you do, consider the following example:

The population of your town is 1000, and the members of that town like a rich, salty diet. So, 300 have hypertension, 400 have diabetes, 120 have both. All patients wind up in the hospital.

The following Venn diagram illustrates the situation.



## Another implication of Berkson's Paradox

(2)

Observe that

$$P(B) = 0.3 = P(B|D) \text{ and } P(D) = 0.4 = P(D|B)$$

but that if  $C \equiv B \cup D$ , then

$$P_c(B|D) = P(B|D) = 0.3 < 0.52 \cong P_c(B)$$

and

$$P_c(D|B) = P(D|B) = 0.4 < 0.69 \cong P_c(D)$$

Again, the negative correlation is the result of an *effect*. Both B and D put you into the hospital.

# **The Inspection Paradox**

# Inspection Paradox (Length-biased Sampling)

## Example

### Situation:

You own a common-carrier trucking company.

Your fleet of trucks is fairly old. Let's say the average age is several times the average age of a truck battery.

Today you get interested in the question: What is the average length of life of my truck batteries?

So, you monitor the length of life of those in service today to answer your question.

You will be happier than you should be.

# Inspection Paradox

## Informal Statement

### Situation:

A light bulb, electronic component, any gadget is placed into service at time 0. As soon as it fails, it is replaced. This process continues indefinitely.

A component has a random lifetime with mean  $\mu$ .

Pick and fix a time  $t > 0$ . The component in service at time  $t$  tends to have a lifetime greater than a typical component. If  $t$  is large compared to  $\mu$ , the tendency is significant.

## Inspection Paradox: The Poisson Case (1)

Situation:

Random variable  $X$  represents the lifetime of a component.  $X$  is exponentially distributed with parameter  $\lambda$ , i.e.,

$$f_X(x) = \lambda e^{-\lambda x} \text{ if } x \geq 0 \quad (\text{pdf that describes } X)$$

$$E(X) = \frac{1}{\lambda} \quad (\text{expectation of } X)$$

Component 1 lasts  $X_1$  units of time and is replaced by component 2, which lasts  $X_2$  units of time, . . . etc.

$X_1, X_2 \dots X_n \dots$  are independent and are distributed as  $X$ .

Let  $L_t$  represent the lifetime of the component that is in service at time  $t$ . Naively,  $L_t$  is also distributed as  $X$ . But it is not.

## Inspection Paradox: The Poisson Case

(2)

The pdf that describes  $L_t$  is  $f_{L_t}$ , where

$$f_{L_t}(x) = \begin{cases} \lambda^2 x e^{-\lambda x} & \text{if } 0 < x \leq t \\ \lambda(1 + \lambda t)e^{-\lambda x} & \text{if } t < x \end{cases}$$

and

$$E(L_t) = \frac{1}{\lambda} + \frac{1}{\lambda}(1 - e^{-\lambda t}) \rightarrow 2E(X) \text{ as } t \uparrow \infty$$

The next visual displays a graphic example.



*Items in service at time T have longer lives than the standard item*

NORMAL FLOAT AUTO REAL RADIAN CL 

Plot1 Plot2 Plot3

$Y_1 = (0 < X \text{ and } X \leq T)(Xe^{-X}) + (X > T)(1+T)e^{-X}$

$Y_2 = e^{-X}$

$Y_3 =$

$Y_4 =$

$Y_5 =$

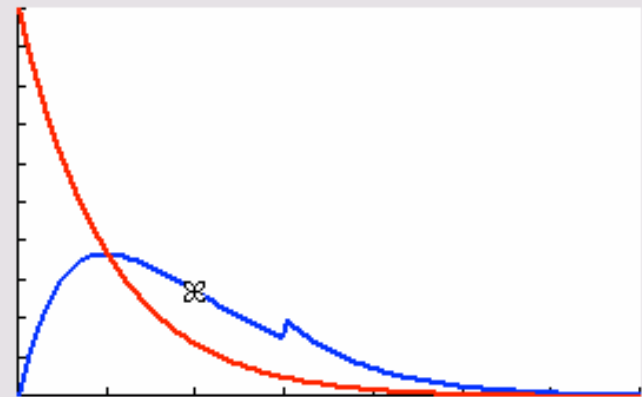
$Y_6 =$

$Y_7 =$

$Y_8 =$

NORMAL FLOAT AUTO REAL RADIAN CL 

$Y_1 = (0 < X \text{ and } X \leq T)(Xe^{-X}) + (X > T)(1+T)e^{-X}$

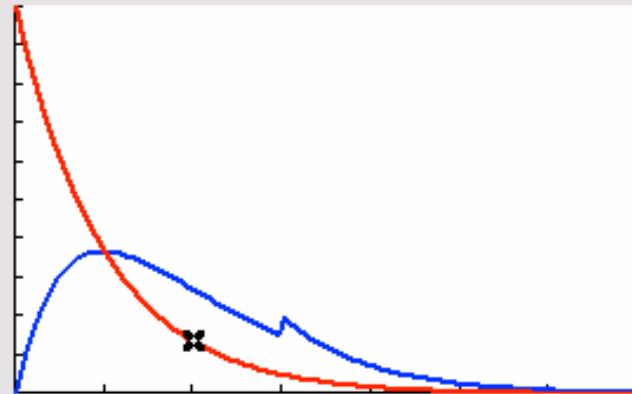


X=2

Y=.27067057

NORMAL FLOAT AUTO REAL RADIAN CL 

$Y_2 = e^{-X}$

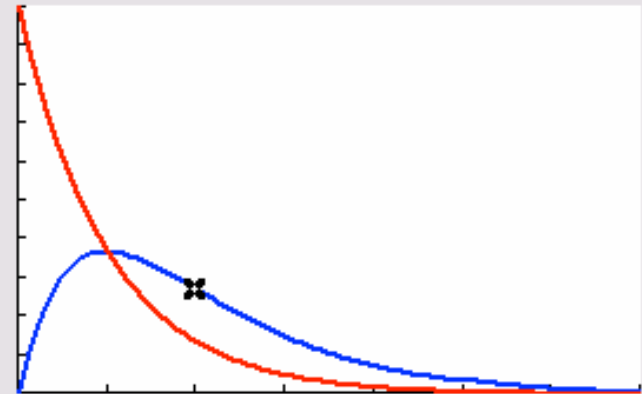


X=2.0151515

Y=.1333002

NORMAL FLOAT AUTO REAL RADIAN CL 

$Y_1 = (0 < X \text{ and } X \leq T)(Xe^{-X}) + (X > T)(1+T)e^{-X}$




X=2

Y=.27067057

*T=3 for graphs 1 and 2. T=5 for graph 3.*

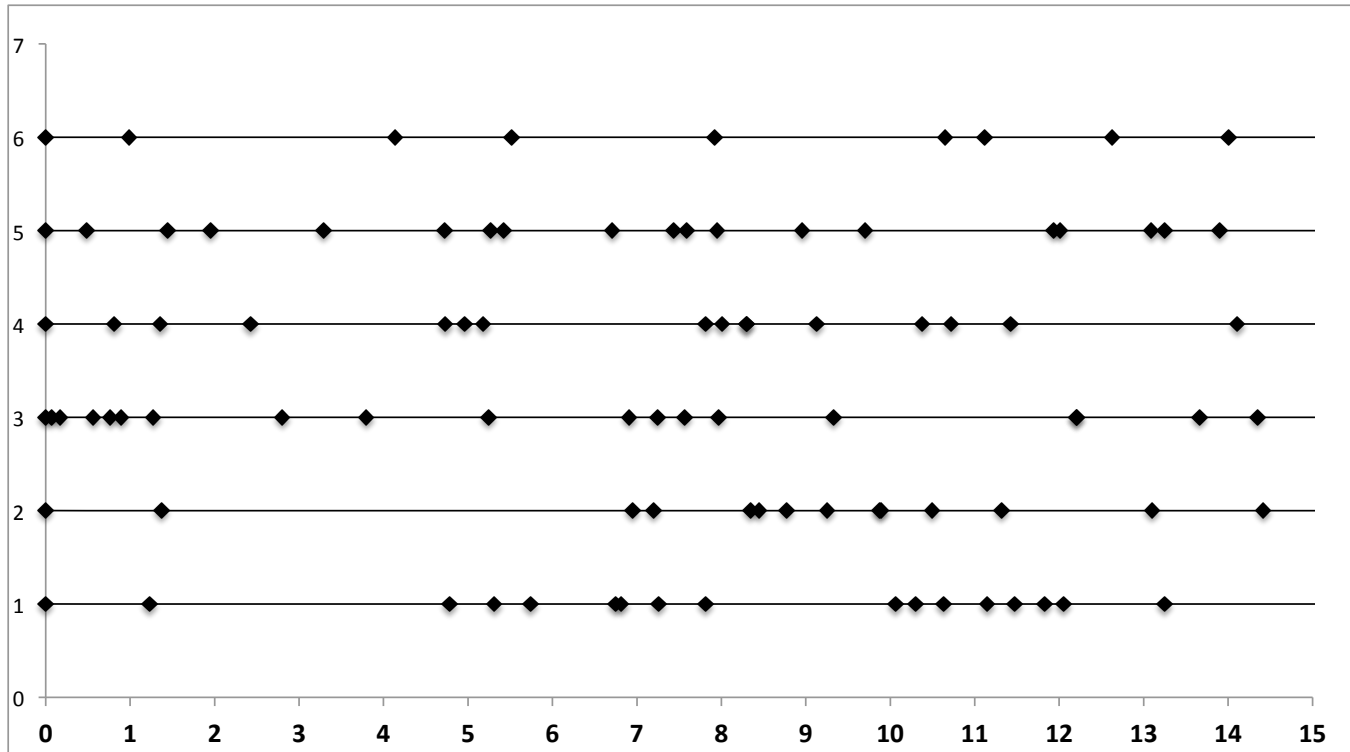
*TI-84 screen to calculate length of segment that straddles location T*

```
NORMAL FLOAT AUTO REAL RADIAN CL   
-ln(rand(20))→L1:cumSum(L1  
 )→L2:seq((L2(I)<T and L2(I  
+1)>T)*(L2(I+1)-L2(I)),I,1  
,19,1)→L3  
{0 0 0 0 0 .3129156362 0 ...
```

Poisson case with  $\lambda=1$  and  $T=5$ .

A run of 20 simulations yielded an average length of 2.4.

*Six realizations of a portion of a Poisson Process*



*Each interarrival is a sample of an exponentially distributed, mean 1 random variable*

# Survival Bias

## Survival Bias

(1)

I begin with what I think is a remarkable story. Epoch: early 1940s. Location: West 118th Street, NYC. Institution: Statistical Research Group (SRG) formed to advise the military on warfare issues. Abraham Wald is a member of this group.

A group of military officers brings to SRG the following data about planes returning from missions over Europe.

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Armor is known to degrade range and performance of aircraft. The officers reasoned that you can get equivalent protection with less armor if you concentrate the armor on the places getting the most hits. But they wanted technical advice on how to redistribute that armor. Instead, Wald says:

*The armor goes where the bullet holes aren't: on the engines.*

## Survival Bias

(2)

Wald asked himself: Where would the missing holes be if the damage were uniformly distributed over the plane?

His answer to himself: The missing holes were on the missing planes. And they are missing because of hits to the engine area, which brought down those planes.

And so the statistical concept of Survival Bias was born



## Abraham Wald's Work on Aircraft Survivability

Marc Mangel, Francisco J. Samaniego

*Journal of the American Statistical Association*, Volume 79, Issue 386 (Jun., 1984), 259-267.

Viewing Wald's work on aircraft survivability in light of the state of the art at the time it was done, it seems to us to be a remarkable piece of work. While the field of statistics has grown considerably since the early 1940's, Wald's work on this problem is difficult to improve upon. Much of the work appears to be ad hoc—there are few allusions to modeling and no reference to classical statistical approaches or results. By the sheer power of his intuition, Wald was led to subtle structural relationships (e.g., Equations (3.3) and (3.24)), and was able to deal with both structural and inferential questions in a definitive way.

[Received May 1981. Revised March 1983.]

# Survival Bias is alive and well today (1)

Articles and books about successful outliers sell. Those about failures do not.

- *Good to Great*, Jim Collins, 2001

An example of predicting those who will do well after finding those who did well.

Screened 11 companies out of 1435 whose stocks beat the market over a 40-year period. Then searched for common characteristics that he believed led to their success.

That is an example of “history”; not useful for predictive purposes.

Incidentally, in the 2001-2012 epoch 6 of the 11 underperformed

- Want to be the next Steve Jobs? Drop out of college, join some acquaintances, start to work in your garage.

Maybe you will succeed; high probability that you will wind up with a garage full of junk.



## Survival Bias is alive and well today

(2)

The usual path to wealth by Silicon Valley startups:

(Get venture capital) → (initial public offering)  
or  
(being acquired)

Last year: 1334 got funded; 81 achieved IPO or acquisition.

And we will never know how many bailed out before getting funding.

(Data from the National Venture Capital Association)

# Closure

## Two famous historical examples of sampling from the wrong population

(1)

- 1936 Presidential Election

Incumbent Franklin Delano Roosevelt (aka FDR) vs Alf Landon

*Literary Digest* magazine polled over 2 million persons via mail and predicted that Landon would win by a large margin. Landon took two states: Maine and Vermont.

George Gallup polled 50,000 and predicted that FDR would win. That prediction put Gallup on the map.

What went wrong for *Literary Digest*? *Literary Digest* polled its readers, registered automobile owners and telephone users. Those populations contained an over-representation of the rich.

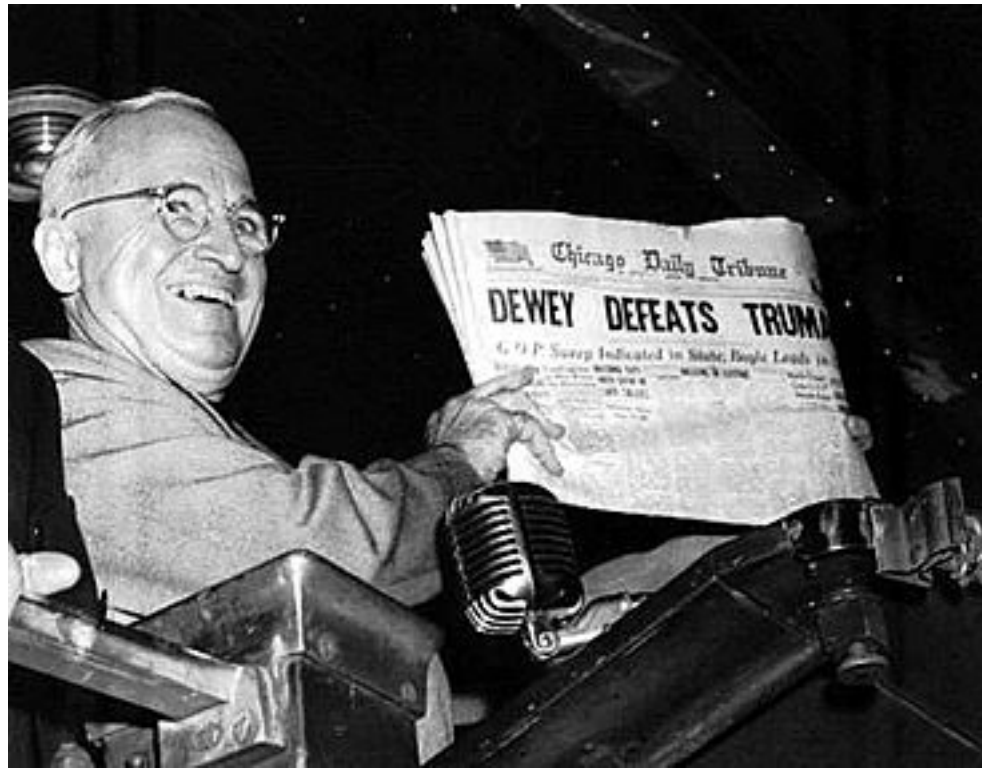
## Two famous historical examples of sampling from the wrong population (2)

- 1948 Presidential Election

Incumbent Harry Truman vs Tom Dewey

*Chicago Tribune* printed DEWEY DEFEATS TRUMAN headline

The editors trusted a phone survey. Telephones were not widespread. (To be fair, almost all polls got it wrong, including Gallup.)



## A Look Back

I have discussed four cases:

- The sibling gender-problem and a generalization thereof
- Berkson's Paradox
- The Inspection Paradox
- Survival Bias

Many additional sampling anomalies, sampling traps and outright fraudulent statistical schemes exist. Some have colorful names: Texas sharpshooter, caveman effect, moving the goal post, Will Rogers phenomenon, cherry picking.

# Opinion

- Acquiring a representative sample in real-world applications is impossible. But decisions have to be made. Awareness of sampling difficulties improves interpretation of and adjustment of imperfect results.
- We probably should not burden elementary statistics students with these observations.

## Motto:

**Tell them the truth and nothing but the truth, but don't tell them the whole truth.**

**Thank you for attending**