

Simulation Study of Horvitz- Thompson estimator

by Julie Yee

September 2009

Horvitz-Thompson estimator

$$\hat{N} = \sum_{i=1}^x \frac{1}{\beta_i}$$

Where x is the number of fatalities found and β_i is the probability of finding each respective fatality.

Sometimes termed “adjusted fatalities”

Variations:

- A: estimates fatalities at sampled turbines
- B: extrapolates to all turbines, using # turbines as extrapolation factor
- C: extrapolates to all turbines, using MW capacity as extrapolation factor

Two Versions:

- 1: assumes β is known
- 2: uses an estimate, i.e. $\hat{\beta}$

Horvitz-Thompson estimators

Estimated fatalities
at sampled turbines

$$\bullet A1 = \sum_{i=1}^x \frac{1}{\beta_i}$$

$$\bullet A2 = \sum_{i=1}^x \frac{1}{\hat{\beta}_i}$$

...and extrapolated to all turbines

$$\bullet B1 = \left(\sum_{i=1}^x \frac{1}{\beta_i} \right) \times \frac{\text{total turbines}}{\text{sampled turbines}}$$

$$\bullet B2 = \left(\sum_{i=1}^x \frac{1}{\hat{\beta}_i} \right) \times \frac{\text{total turbines}}{\text{sampled turbines}}$$

$$\bullet C1 = \left(\sum_{i=1}^x \frac{1}{\beta_i} \right) \times \frac{\text{total MW}}{\text{sampled MW}}$$

$$\bullet C2 = \left(\sum_{i=1}^x \frac{1}{\hat{\beta}_i} \right) \times \frac{\text{total MW}}{\text{sampled MW}}$$

Distribution of kW capacity

size category	kW capacity	total # turbines	total kW	total turbines
Very small	40	150	34,200	750
Very small	65	600		
Small	100	3000	351,000	3400
Small	120	300		
Small	150	100		
Medium	250	20	45,000	100
Medium	330	40		
Medium	400	40		

Stratified sampling

size category	Total MW	total # turbines	Sampled turbines	Sampled MW
Very small	34.2	750	750	34.2
Small	351	3400	1650	Variable
Medium	45	100	100	45
Total	430.2	4250	2500	Variable

Each estimator is applied separately to each strata (i.e. size category)
Then estimated fatalities are summed across strata

Definitions of β and $\hat{\beta}$

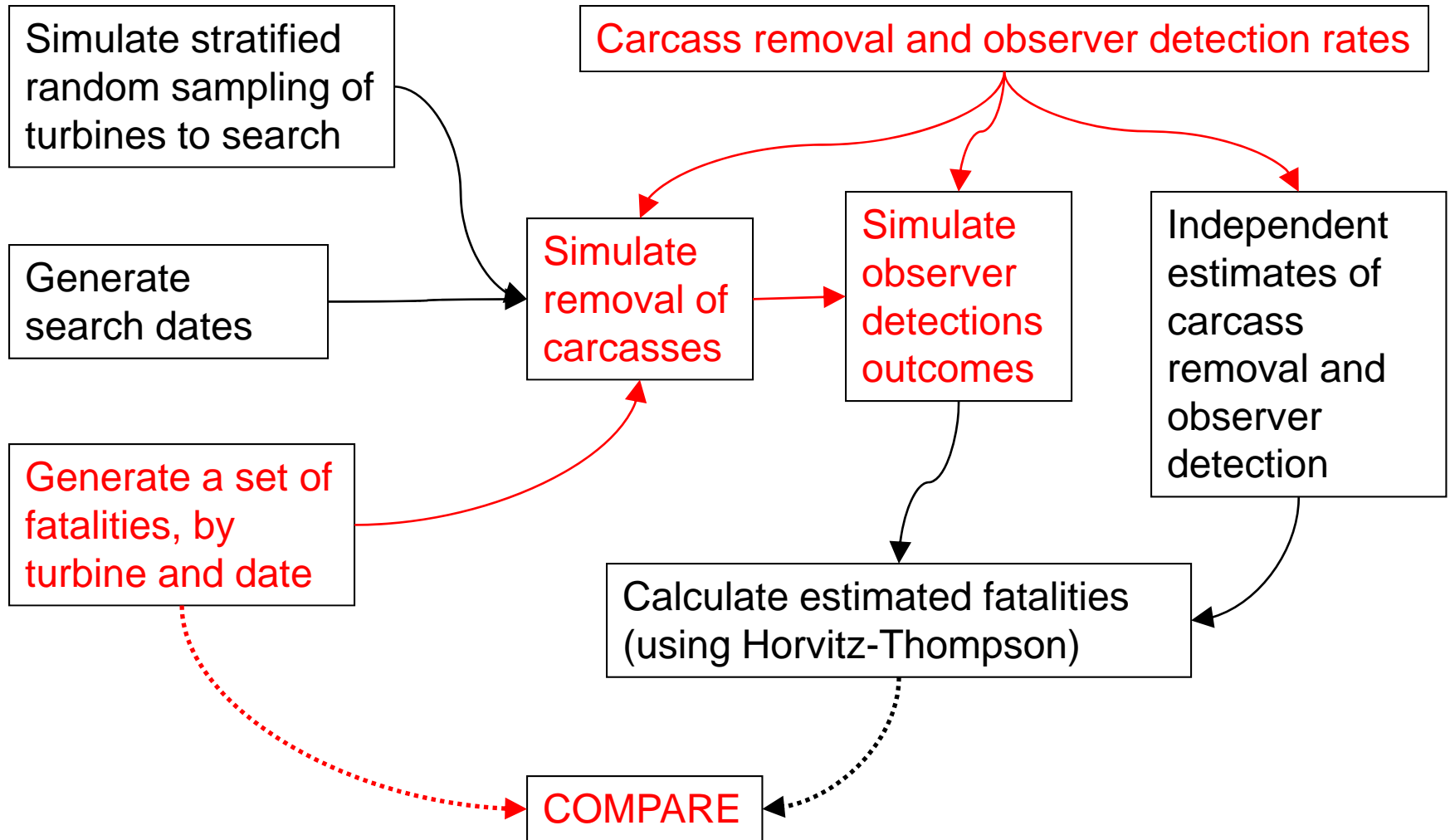
$$\begin{aligned}\beta &= \text{Prob \{fatality becomes found\}} \\ &= \text{Prob \{fatality is still present on survey date and observer detects it\}} \\ &= \text{Prob \{fatality is still present on survey date\}} \\ &\quad \times \text{Prob\{observer detects fatality, conditional on its presence\}} \\ &= R_i \times p\end{aligned}$$

Where i is the number of days between date of fatality and search date

$$\hat{\beta} = \hat{R}_c \times \hat{p} = \left(\frac{1}{L} \sum_{i=1}^L \hat{R}_i \right) \times \hat{p}$$

Where L is the length of search interval in which fatality was found, and \hat{R}_i and \hat{p} are unbiased estimators of R_i and p

Simulation flowchart

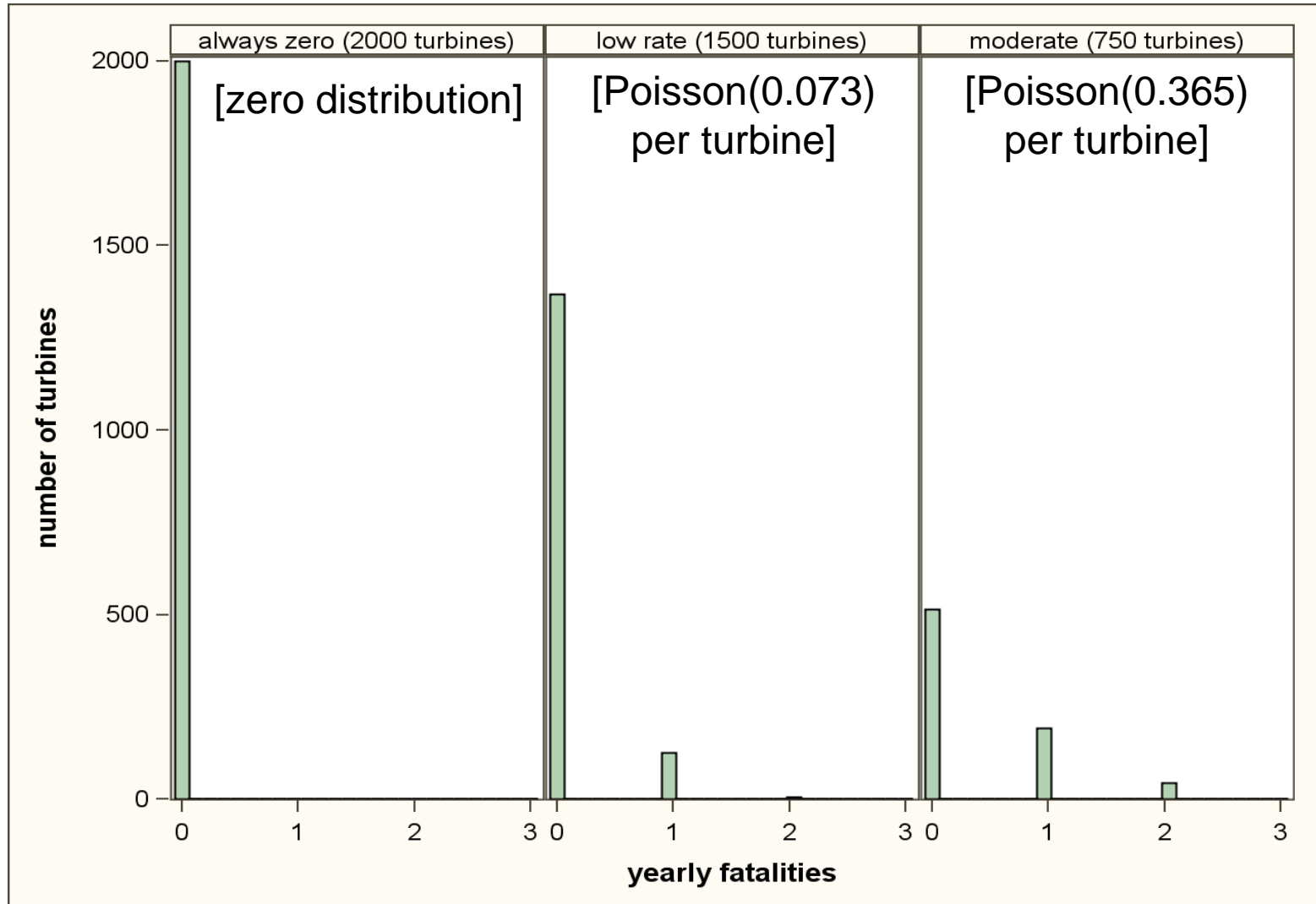


Fatalities and search dates

- “Period 1”
 - Fatality rate 383.25 per year
 - Search dates between 30 and 90 days
- “Period 2”
 - 50% reduction in fatality rate (191.625/yr)
 - Search dates between 25 and 50 days

Fatalities distributed heterogeneously

zero-inflated Poisson mixture (ex: "Period 1")



Simulated fatalities

The screenshot shows the SAS software interface with a data table open in browse mode. The table has the following columns: **turbnum**, **size_cat**, **size_kw**, **surveyed**, **day**, and **fallen**. The data consists of 19 rows, each representing a simulated fatality event. The 'size_cat' column is consistently 'SM' and 'size_kw' is consistently '100'. The 'surveyed' column is consistently 'Y'. The 'day' column ranges from 1 to 19, and the 'fallen' column is consistently '0'.

	turbnum	size_cat	size_kw	surveyed	day	fallen
1	1	SM	100	Y	1	0
2	1	SM	100	Y	2	0
3	1	SM	100	Y	3	0
4	1	SM	100	Y	4	0
5	1	SM	100	Y	5	0
6	1	SM	100	Y	6	0
7	1	SM	100	Y	7	0
8	1	SM	100	Y	8	0
9	1	SM	100	Y	9	0
10	1	SM	100	Y	10	0
11	1	SM	100	Y	11	0
12	1	SM	100	Y	12	0
13	1	SM	100	Y	13	0
14	1	SM	100	Y	14	0
15	1	SM	100	Y	15	0
16	1	SM	100	Y	16	0
17	1	SM	100	Y	17	0
18	1	SM	100	Y	18	0
19	1	SM	100	Y	19	0

NOTE: Table has been opened in browse mode.

Definitions of β and $\hat{\beta}$

Simulate “found” fatalities using two separate scenarios:

- $R_i = 1 - 0.2062888 \times \log(i)$ and $p = 1$ (perfect detection)
- $R_i = 1 - 0.2062888 \times \log(i)$ and $p = 0.75$ (imperfect detection)

Simulate estimators $\hat{\beta} = \hat{R}_c \times \hat{p}$

- \hat{R}_c = unbiased estimator of R_c with 0.25 CV (coef. of variation)
 - \hat{p} = unbiased est. of p with 0.05 CV (coef. of variation)
- (use $\hat{p} = 1$ when detection is perfect)

Simulating finds: example 1

The screenshot displays the SAS software interface. The main window shows a data table with the following columns: sampturbnum, day, fallen, beg, end, intlenth, days, Ri, condRi, prevmiss, remain, and found. The data is organized into 19 rows. The Explorer window on the left shows the contents of the 'Work' directory, with 'Fatalities4b' selected. The taskbar at the bottom shows several open windows, including 'Output - (Untitled)', 'Log - (Untitled)', 'simulate2009091...', 'macros simulate.sas', and 'VIEWTABLE: Wo...'. The status bar at the bottom indicates the current directory is 'E:\Reviews\Altamont\simulate'.

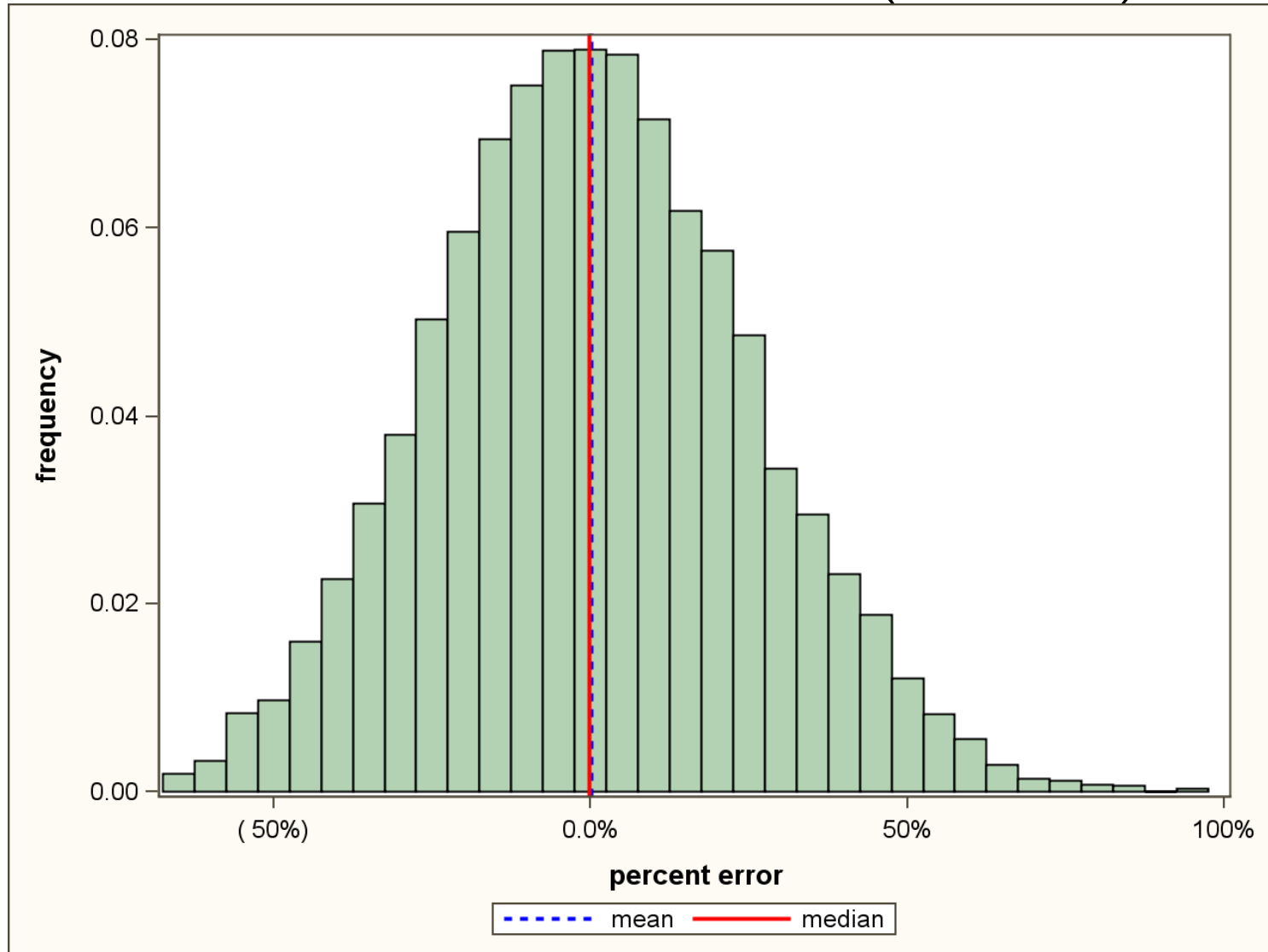
	sampturbnum	day	fallen	beg	end	intlenth	days	Ri	condRi	prevmiss	remain	found
1	1255	89	1	63	103	40	15	0.44136	.	.	1	0
2	1255	89	1	103	139	36	51	0.18891	0.42801	1	0	0
3	1255	89	1	139	169	30	81	0.09347	0.49481	.	.	.
4	1255	89	1	169	210	41	122	0.00898	0.09611	.	.	.
5	1255	89	1	210	257	47	169	0	0	.	.	.
6	1255	89	1	257	299	42	211	0
7	1255	89	1	299	344	45	256	0
8	1255	89	1	344	365	45	277	0
9	1264	338	1	333	365	39	28	0.3126	.	.	0	0
10	1276	90	1	68	113	45	24	0.3444	.	.	0	0
11	1276	90	1	113	158	45	69	0.12655	0.36745	.	.	.
12	1276	90	1	158	203	45	114	0.02298	0.18155	.	.	.
13	1276	90	1	203	250	47	161	0	0	.	.	.
14	1276	90	1	250	292	42	203	0
15	1276	90	1	292	338	46	249	0
16	1276	90	1	338	365	46	276	0
17	1356	224	1	190	235	45	12	0.48739	.	.	1	1
18	1356	224	1	235	262	27	39	0.24425	0.50113	.	.	.
19	1356	224	1	262	310	48	87	0.07873	0.32235	.	.	.

Simulating finds: example 2

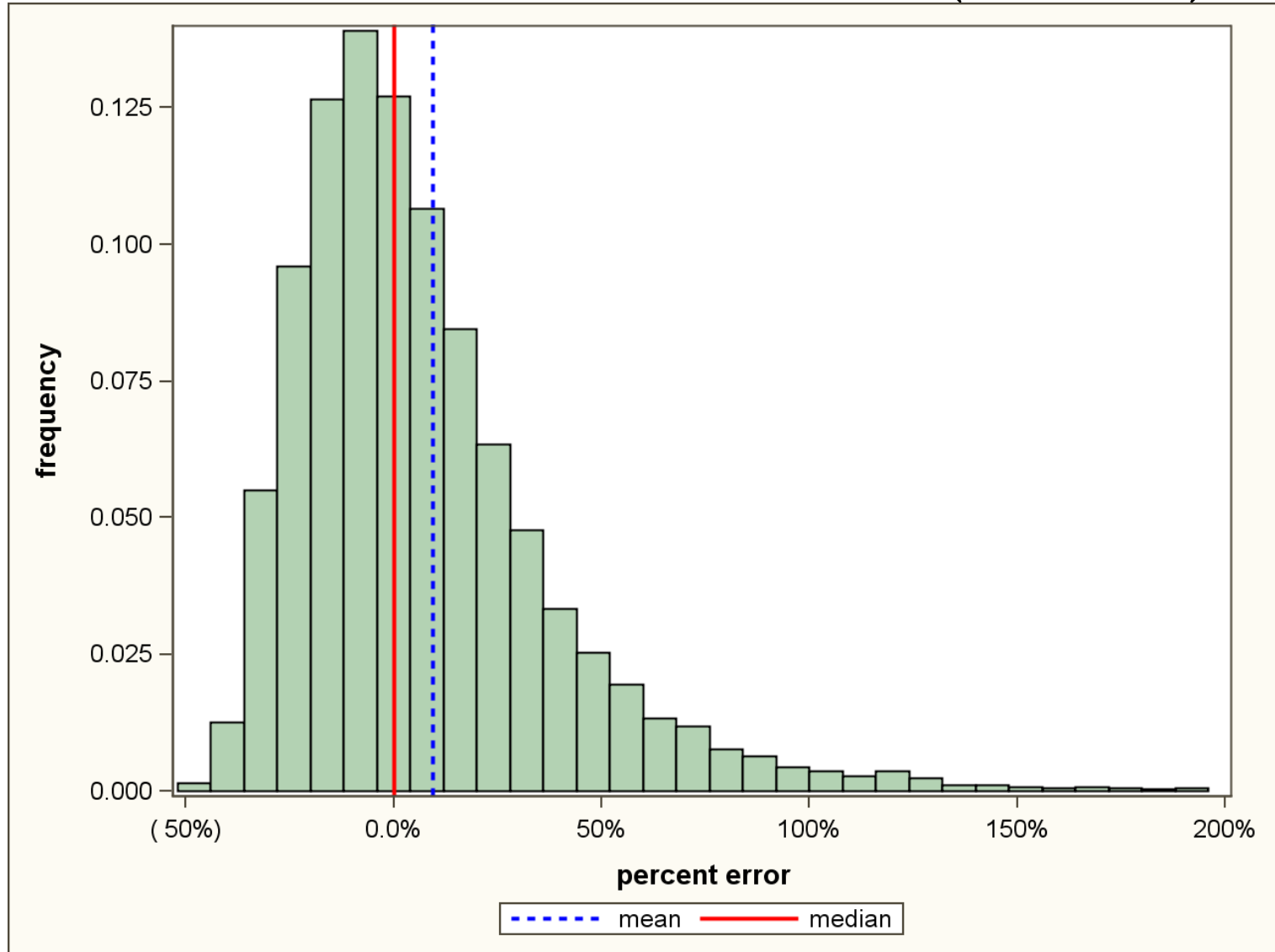
The screenshot displays the SAS software interface. The main window shows a data table with the following columns: sampturbnum, day, fallen, beg, end, intlength, days, Ri, condRi, prevmiss, remain, and found. The data is organized into 19 rows, with the first 15 rows having a sampturbnum of 1622 and the last 4 rows having a sampturbnum of 1629. The 'fallen' column is consistently 1. The 'found' column shows the results of the simulation, with values 0 or 1.

	sampturbnum	day	fallen	beg	end	intlength	days	Ri	condRi	prevmiss	remain	found
66	1622	140	1	110	148	38	9	0.54674	.	.	1	0
67	1622	140	1	148	192	44	53	0.18097	0.33101	1	1	0
68	1622	140	1	192	218	26	79	0.09863	0.54501	1	1	1
69	1622	140	1	218	257	39	118	0.01586	0.16081	.	.	.
70	1622	140	1	257	306	49	167	0	0	.	.	.
71	1622	140	1	306	354	48	215	0
72	1622	140	1	354	365	48	226	0
73	1629	72	1	45	87	42	16	0.42805	.	.	0	0
74	1629	72	1	87	113	26	42	0.22896	0.5349	.	.	.
75	1629	72	1	113	152	39	81	0.09347	0.40826	.	.	.
76	1629	72	1	152	197	45	126	0.00233	0.02492	.	.	.
77	1629	72	1	197	225	28	154	0	0	.	.	.
78	1629	72	1	225	258	33	187	0
79	1629	72	1	258	294	36	223	0
80	1629	72	1	294	330	36	259	0
81	1629	72	1	330	365	36	294	0
82	1639	197	1	186	229	43	33	0.27871	.	.	1	0
83	1639	197	1	229	272	43	76	0.10662	0.38254	1	0	0
84	1639	197	1	272	304	32	108	0.03413	0.3201	.	.	.

Percent error of $(\hat{R}_c \times \hat{p})$



Percent error of $1/(\hat{R}_c \times \hat{p})$



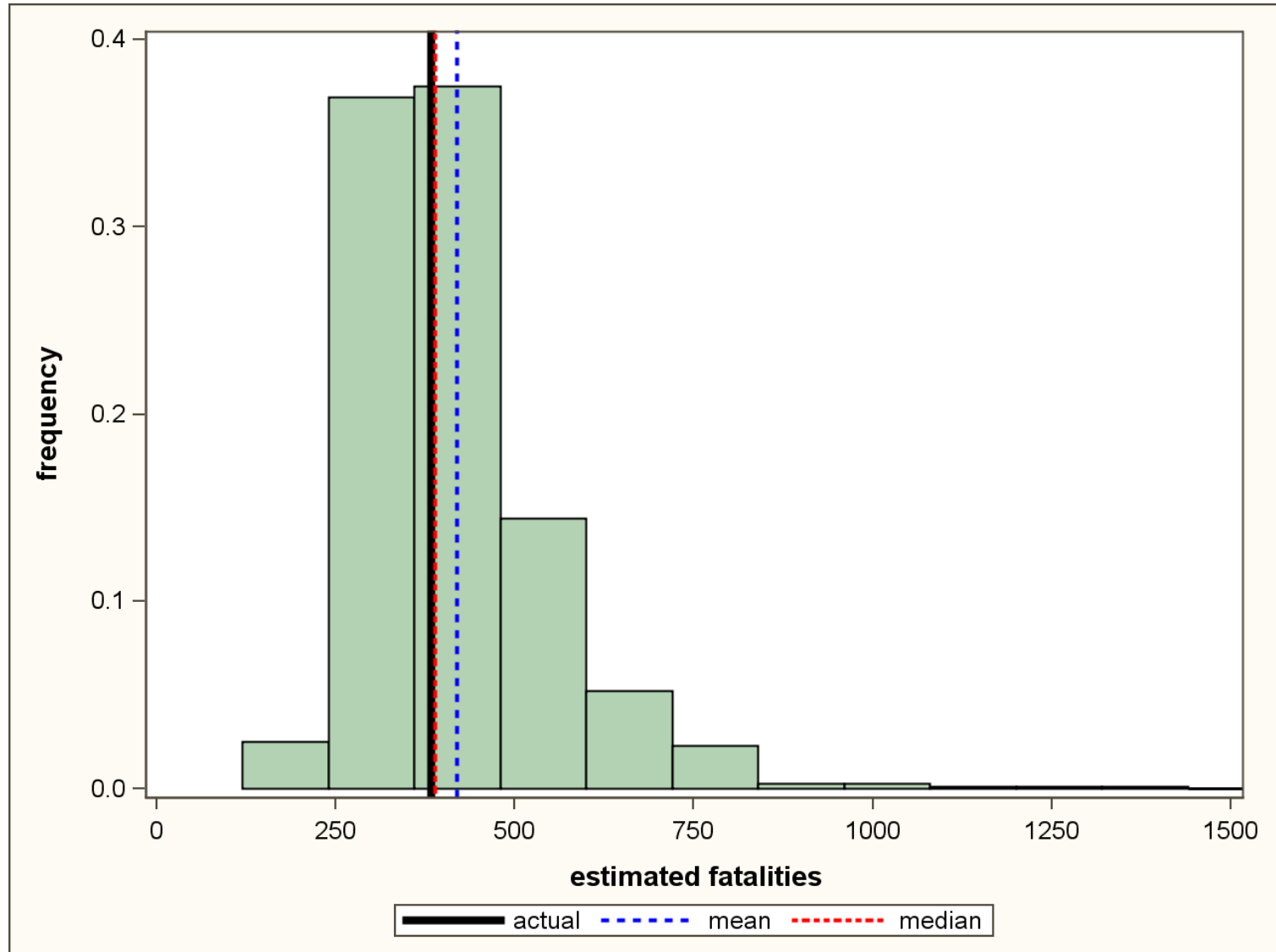
Simulation results, by iteration

example: "Period 1" perfect detection

iter	allfallen	samplefallen	found	estA1	estA2	estB1	estB2	estC1	estC2
1	333	193	72	209.914	219.425	361.470	377.849	361.588	377.972
2	385	222	71	203.795	240.892	350.692	414.528	351.142	415.060
3	382	222	72	208.217	244.737	339.791	399.388	340.270	399.950
4	371	237	65	187.529	216.063	302.220	348.206	302.140	348.113
5	370	217	82	228.850	197.065	376.173	323.926	376.523	324.228
6	342	200	67	200.815	184.748	328.663	302.369	329.114	302.783
7	374	228	74	224.571	272.867	368.550	447.808	369.106	448.484
8	382	229	78	224.963	220.657	373.426	366.279	373.542	366.392
9	361	204	81	239.283	265.181	390.466	432.726	389.125	431.239
10	390	216	62	180.731	183.935	313.476	319.033	313.640	319.200
11	377	254	84	245.209	259.395	415.071	439.085	414.662	438.652
12	386	206	74	210.000	305.056	381.349	553.967	381.228	553.792
13	361	209	65	192.767	259.538	316.541	426.186	316.935	426.716
14	383	219	68	199.421	196.000	344.062	338.160	344.373	338.466
15	370	240	87	258.085	301.000	456.928	532.907	457.378	533.432
16	370	213	71	219.774	363.575	385.505	637.748	386.849	639.972
17	363	217	66	189.758	288.001	307.480	466.672	307.962	467.404
18	411	244	81	243.676	241.966	415.095	412.182	414.662	411.752

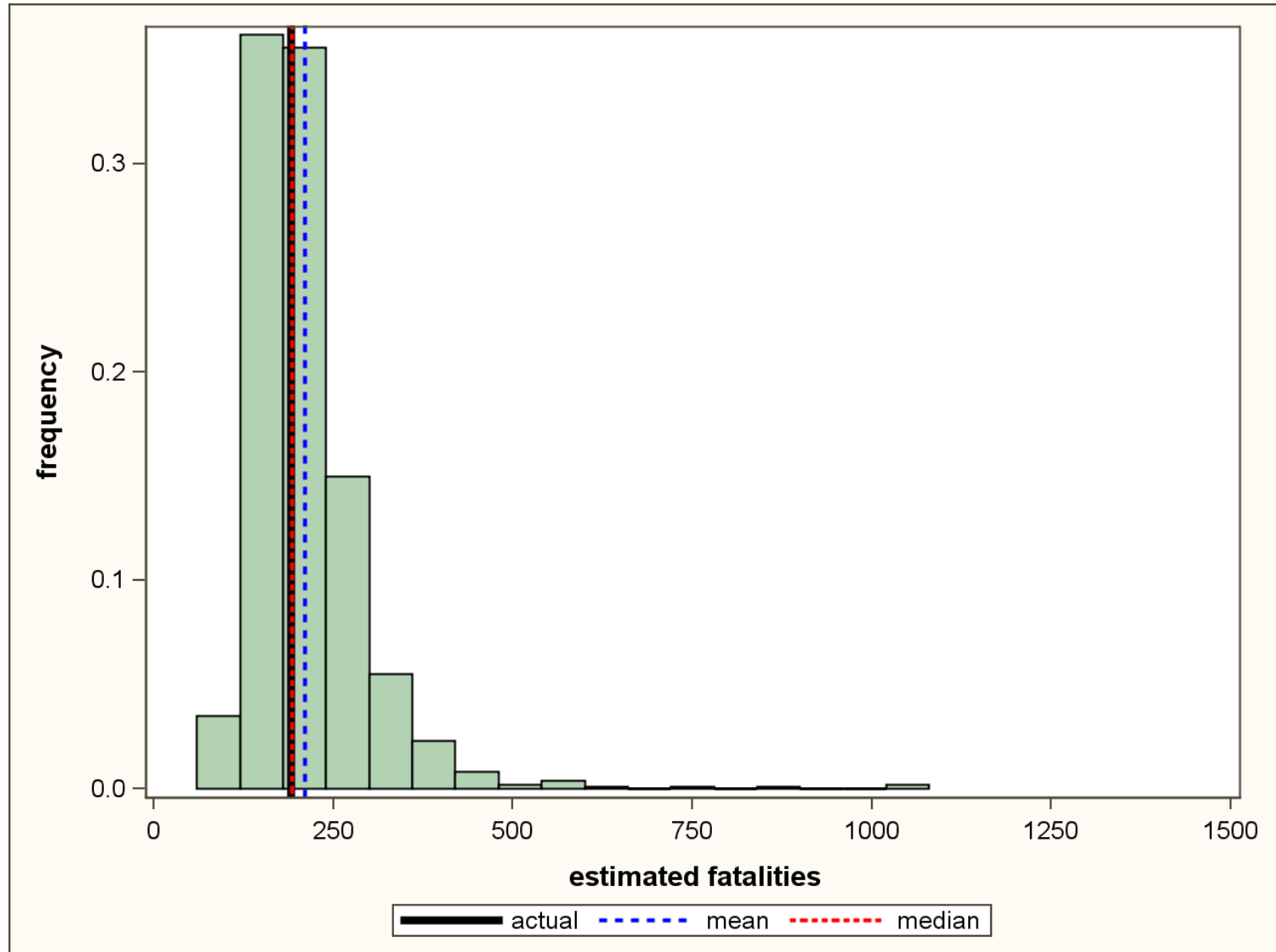
Histogram of fatality estimates

("Period 1" B2 or C2 estimator, perfect detection)



Histogram of fatality estimates

("Period 2" B2 or C2 estimator, perfect detection)



Accuracy and Precision

(perfect detection)

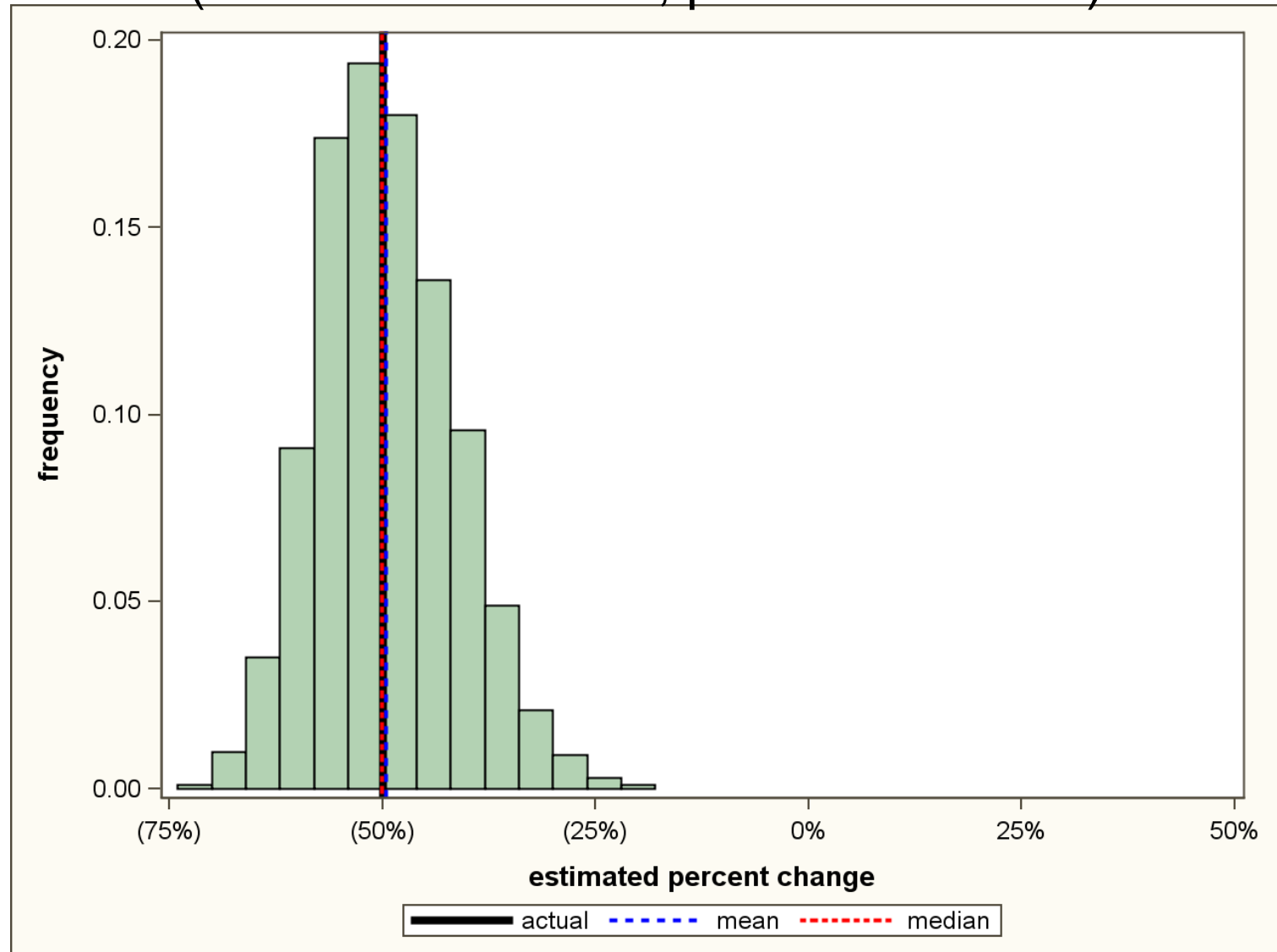
	“Period 1”			“Period 2”		
	bias	stderr	odds	bias	stderr	odds
A1	-0.5	20.6	50:50	0.2	12.1	51:49
A2	20.7	92.5	53:47	11.2	48.6	50:50
B1	-0.3	41.1	48:52	0.2	25.2	49:51
B2	35.8	159.3	52:48	18.9	84.6	50:50
C1	-0.3	41.1	48:52	0.2	25.2	48:52
C2	35.8	159.3	52:48	18.9	84.6	50:50

bias = average difference between H-T estimator and fatalities

stderr = standard error of H-T estimator

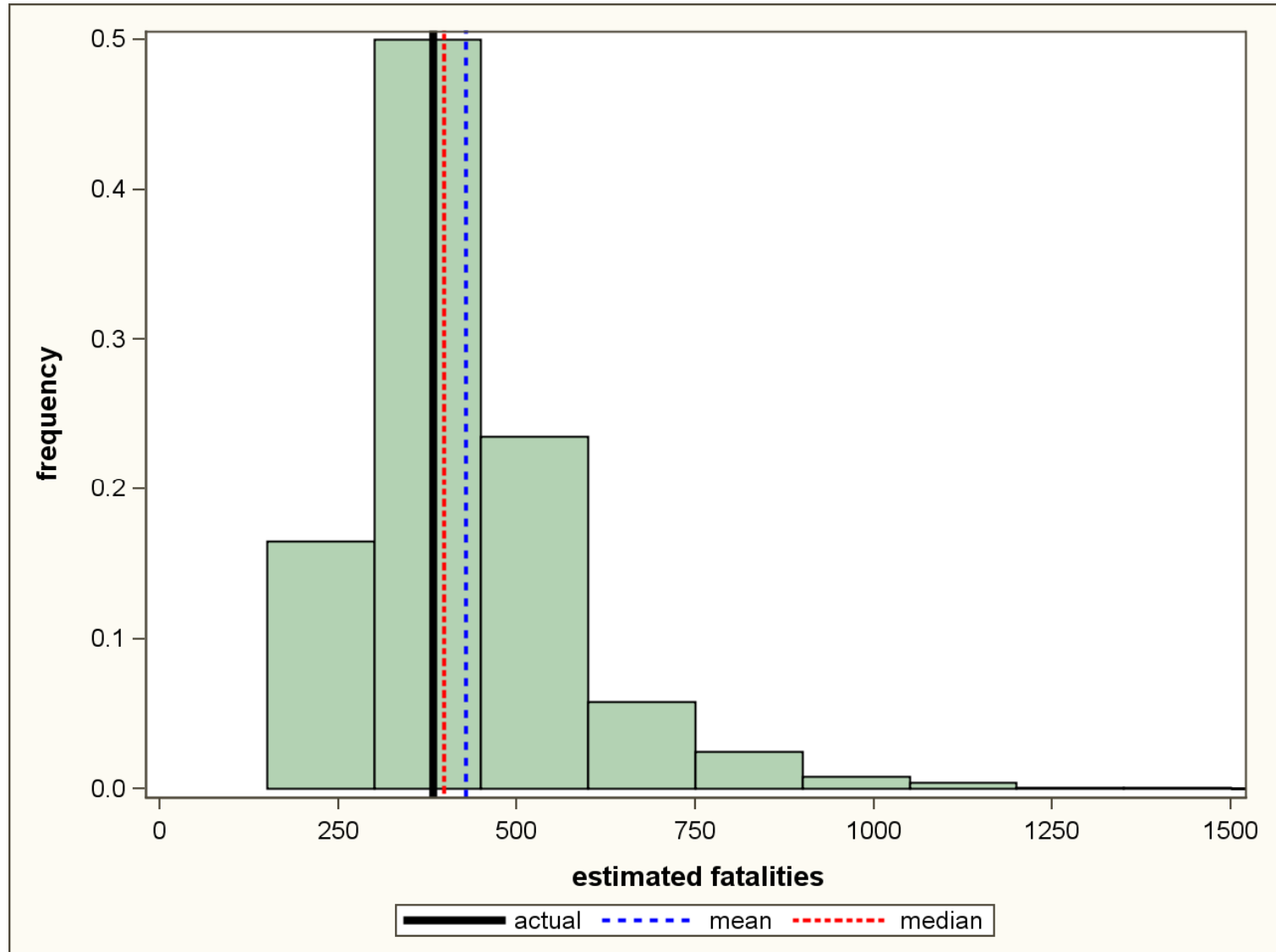
odds = odds H-T overestimating fatalities

% Change from "Period 1" to "Period 2" (B2 or C2 estimator, perfect detection)



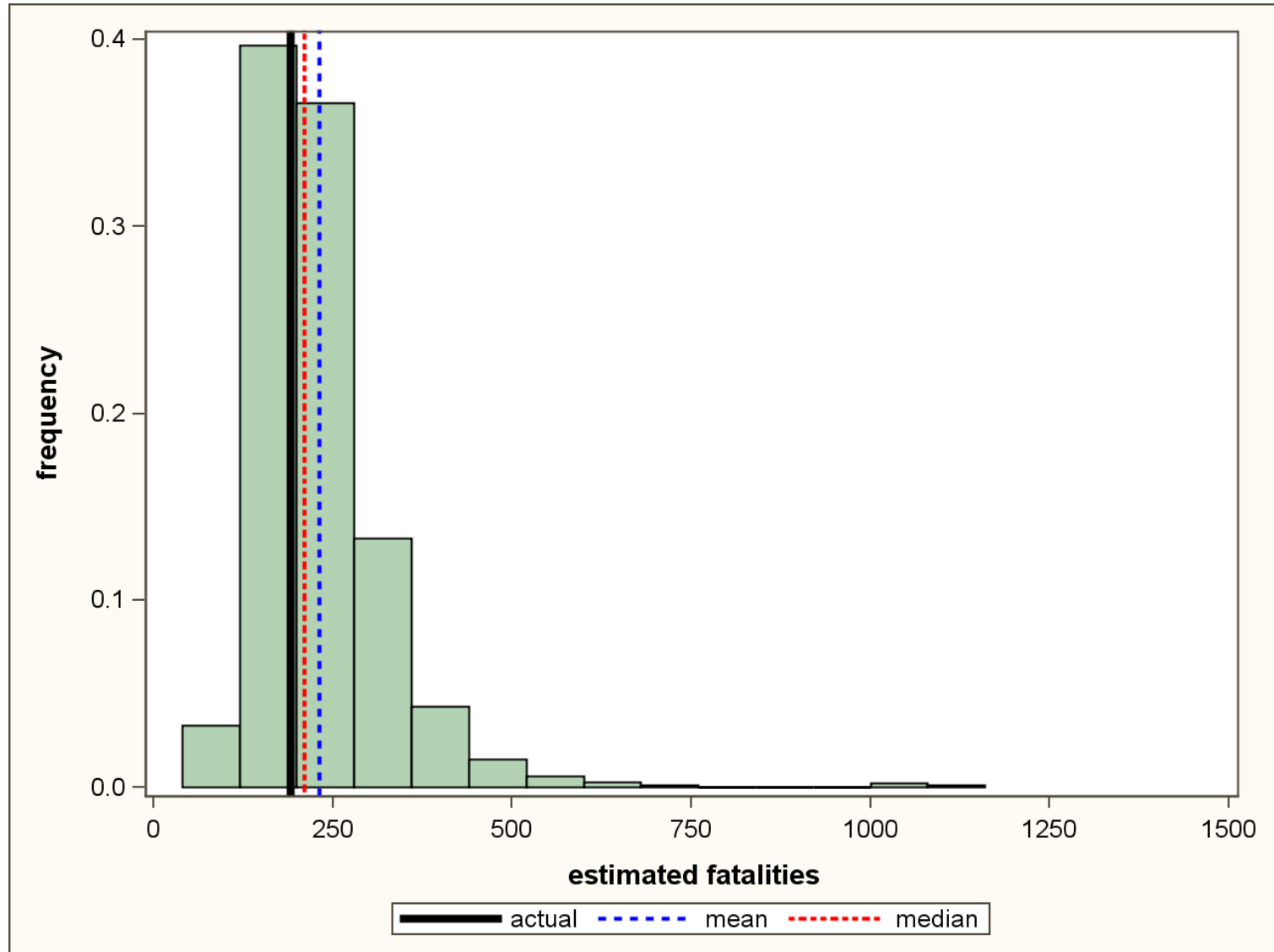
Histogram of fatality estimates

("Period 1" B2 or C2 estimator, 0.75 detection)



Histogram of fatality estimates

("Period 2" B2 or C2 estimator, 0.75 detection)

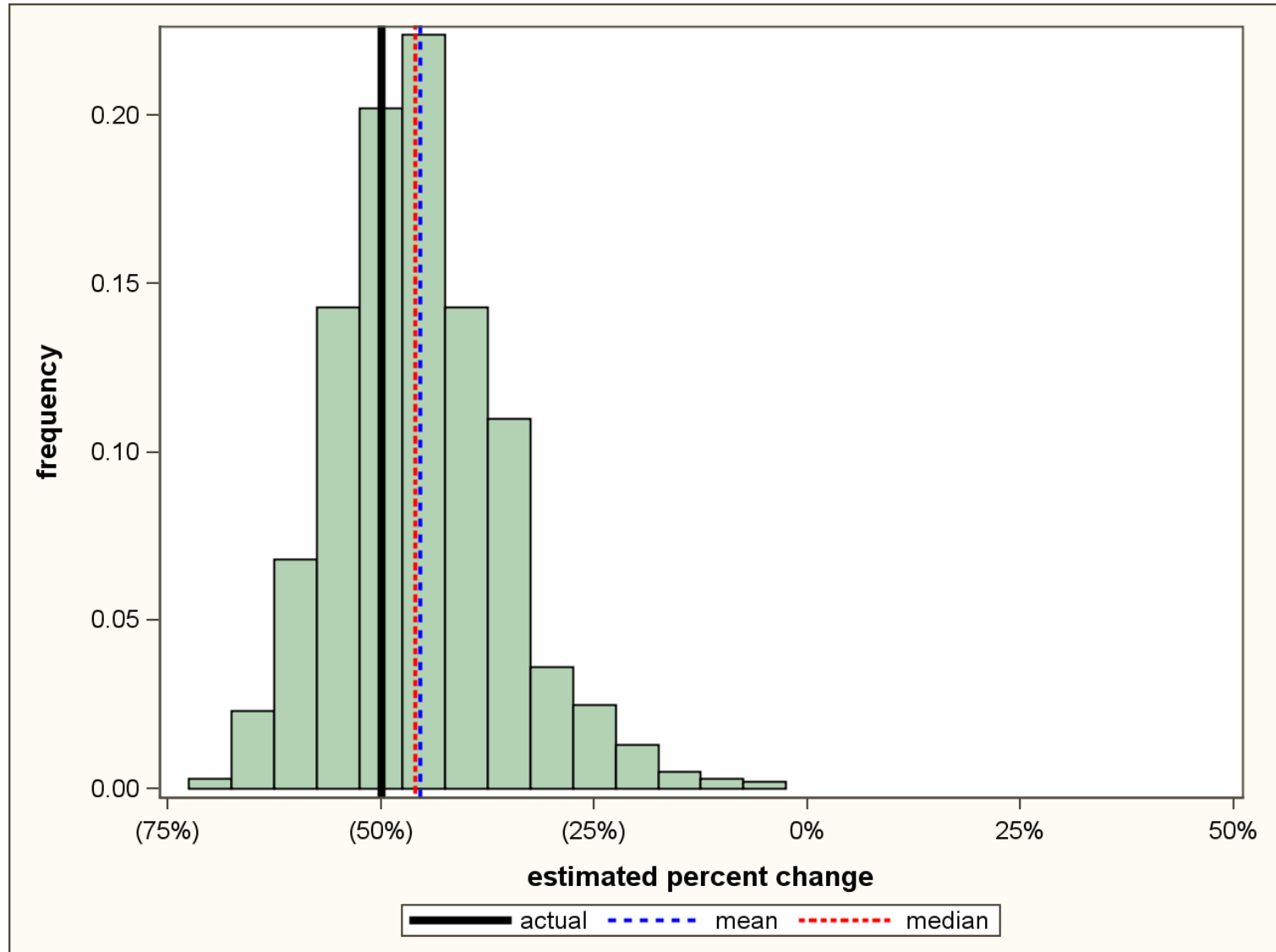


Accuracy and Precision

(0.75 detection)

	"Period 1"			"Period 2"		
	bias	stderr	odds	bias	stderr	odds
A1	3.2	30.2	54:46	10.9	19.3	75:25
A2	26.3	102.0	55:45	23.6	58.7	65:35
B1	6.8	56.0	54:46	18.6	36.5	72:28
B2	46.2	177.2	54:46	40.1	101.5	65:35
C1	6.9	56.0	54:46	18.6	36.6	72:28
C2	46.2	177.2	54:46	40.1	101.5	65:35

% Change from "Period 1" to "Period 2" when observer detection = 0.75



% Change from “Period 1” to “Period 2”

Simulation summary

	average estimate	bias	stderr	odds of overestimating reduction
perfect detection	-50%	0	8%	51:49
0.75 detection	-45%	5%	11%	33:67

Simulation Results

Fatality estimates had

- no bias when sighting rate (β) precisely known
- positive bias (but similar odds of over or underestimation) when β estimated with error.
- extra positive bias when detection was imperfect.

Percent change estimates had

- no bias (and similar odds of over or underestimation) when detection was perfect, even if β estimated with error
- negatively biased reductions when detection imperfect