

BIOL 300: Biostatistics

Course web address:

<http://www.zoology.ubc.ca/~whitlock/bio300/>

Canvas: <https://canvas.ubc.ca/courses/39708>

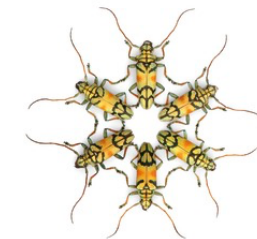
Office hours: Mon. 1:30-2:30
and after class most days

Prof. Michael Whitlock
Department of Zoology
Office: 216 Biodiversity
whitlock@zoology.ubc.ca

Dr. Matt Pennell
Dept. of Zoology
pennell@zoology.ubc.ca

Textbook

Whitlock and Schluter,
The Analysis of Biological Data



The Analysis of Biological Data
WHITLOCK · SCHLUTER
THIRD EDITION

Lab manual

Available at course web site

R

Statistical software for PCs and
Macs

Available for free download:

[http://www.zoology.ubc.ca/~whitlock/
bio300/labs/downloadingR.html](http://www.zoology.ubc.ca/~whitlock/bio300/labs/downloadingR.html)

Evaluation

Final 25%
Mid-term 30%
Project final 25%
Assignments (homework) 10%
Lab assignments 10%

Lab

Begins **second** week of term
(September 14-18)

Biosci 2004

Virtual labs

Midterm

Mid-October:

Class vote: Oct. 22 or 29

Assignments

Available on canvas

Due on Fridays at noon

Lab assignments

Due in canvas the following week

Statistics course pairings

Credit given for only one of BIOL 300, FRST 231, STAT 200, PSYC 218 or 366.

These are paired with BIOL 300, but *do not count* as biology courses.

Introduction to statistics

Statistics are "a quantitative technology for empirical science; it is a logic and methodology for the measurement of *uncertainty* and for an examination of that uncertainty."

The key word here is "uncertainty." Statistics become necessary when observations are variable.

Statistics is also about
good scientific practice

Goals of statistics

- Estimate the values of important parameters
- Test hypotheses about those parameters

Feline High-Rise Syndrome (FHRS)

The injuries associated with a cat falling out of a window.



High falls show *lower* injury rates

“The diagnosis of high-rise syndrome is not difficult. Typically, the cat is found outdoors, several stories below, and a nearby window or patio door is open.”

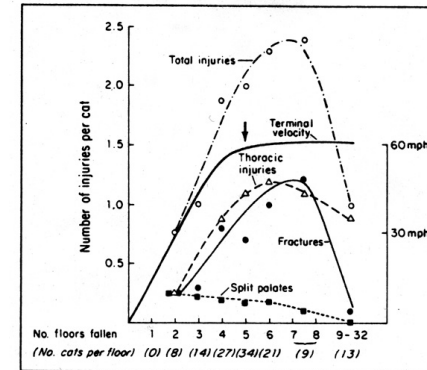


Figure 2—Relationship of injuries to distance fallen and velocity in 132 cats with high-rise syndrome: ↓ points to terminal velocity (—); total number of injuries/cat (○, - - - -); number of thoracic injuries (pulmonary contusions + pneumothorax)/cat (△, - - - -); number of fractures/cat (●, —); number of split palates/cat (■, - - - -).

Whitney and Mehloff, *Journal of the American Veterinary Medicine Association*, 1987

Why?



1. Cats have high surface-to-volume ratios
2. Cats have excellent vestibular systems
3. Cats reach terminal velocity quickly, relax, and therefore absorb impact better
4. Cats land on their limbs and absorb shock through soft tissue

Jared Diamond, *Nature* 1988

Or not...



A sample of convenience is a collection of individuals that happen to be available at the time.

A newer study reports more injuries with longer falls

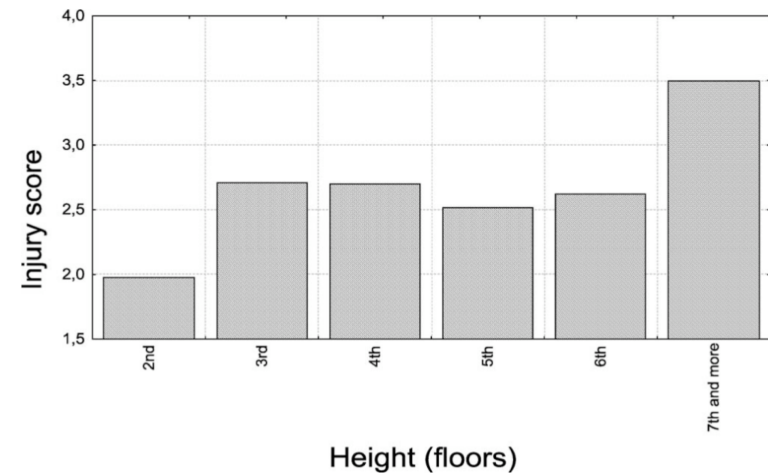


Figure 5 Graph showing the relationship between injury score and height of fall.

Vnuk et al. 2004. Feline high-rise syndrome: 119 cases (1998-2001). *J. Fel. Med. Surg.* 6:305-312.

Variable

A **variable** is a characteristic measured on individuals drawn from a population under study.

Data are measurements of one or more variables made on a collection of individuals.

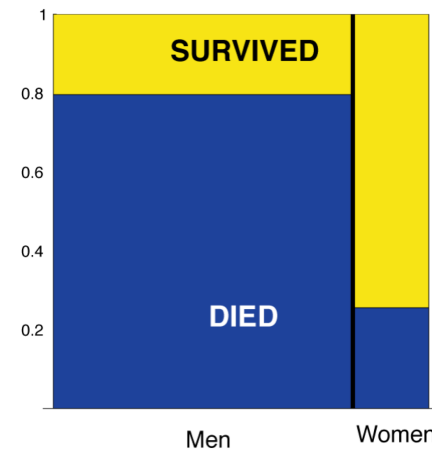
Read: Chapters 1 & 2

Explanatory and response variables

We try to predict or explain a **response variable** from an **explanatory variable**.

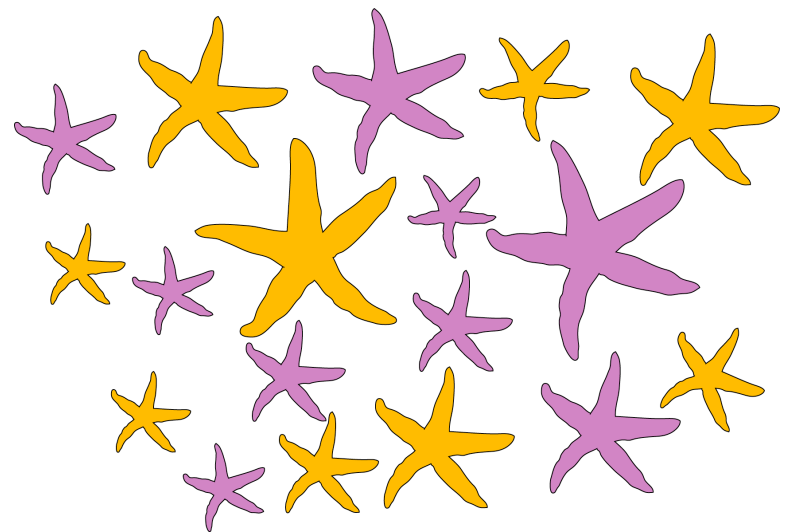
Older terminology:
dependent variable and *independent variable*

Mortality on the *Titanic*, as predicted by gender



Populations and samples

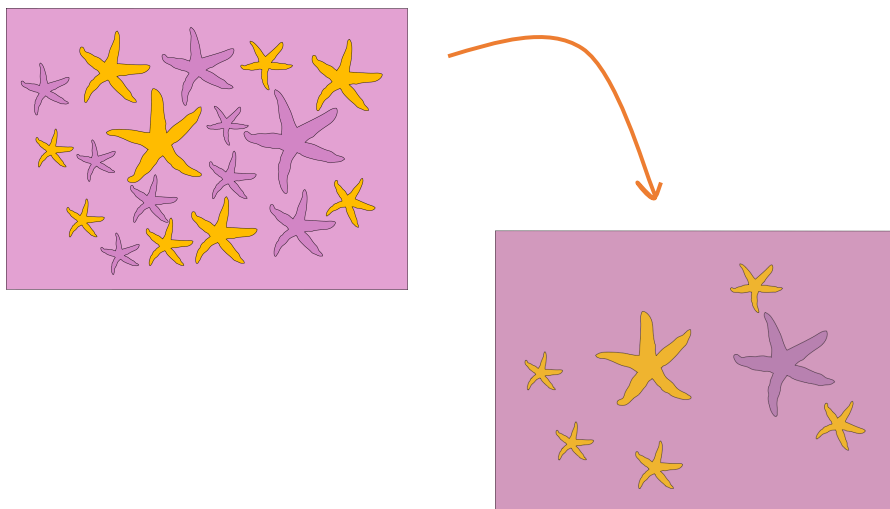
A population of starfish





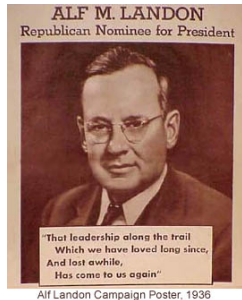
Populations \leftrightarrow Parameters;
Samples \leftrightarrow Estimates

A biased sample



Bias is a systematic discrepancy between estimates and the true population characteristic.

The 1936 US presidential election



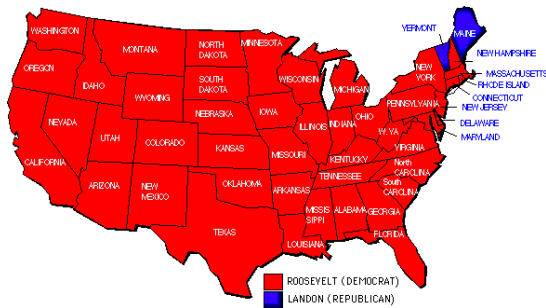
Alf Landon
Republican

vs.



Franklin Roosevelt
Democrat

1936 election results



Roosevelt won with 62% of the vote

1936 *Literary Digest* Poll

2.4 million respondents

Based on questionnaires mailed to 10 million people, chosen from telephone books and club lists

Predicted Landon wins: Landon 57%
over Roosevelt 43%

What went wrong?

Subjects given the questionnaire were chosen from telephone books and clubs, biasing the respondents to be those with greater wealth

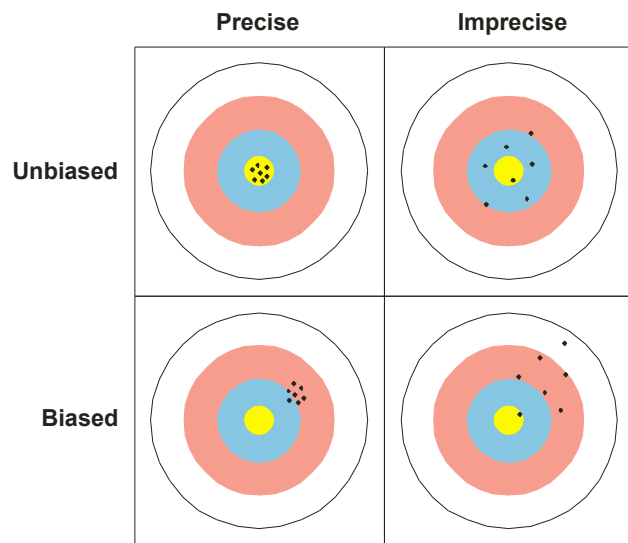
Voting and party preference is correlated with personal wealth

Volunteer bias

Volunteers for a study are likely to be different, on average, from the population.

For example:

- Volunteers for sex studies are more likely to be open about sex
- Volunteers for medical studies may be sicker than the general population



Each point represents an estimate of a parameter.

Goals of estimation

- Accuracy (on average gets the correct answer)
- Precision (gives a similar answer repeatedly)

Properties of a good sample

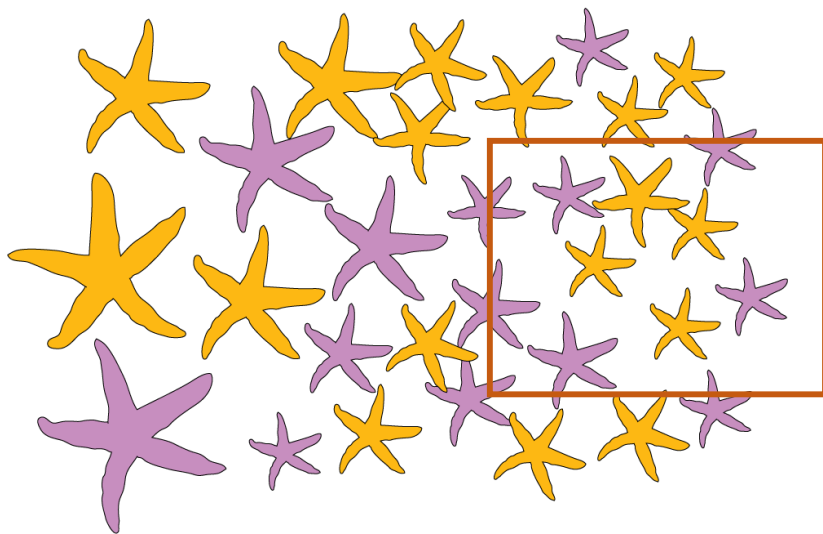
- Independent selection of individuals
- Random selection of individuals
- Sufficiently large

Independent sampling

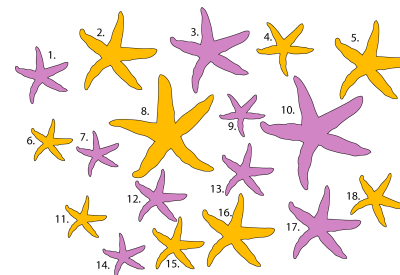
The chance of an individual being included in the sample does not depend on who else is sampled.

In a *random sample*, each member of a population has an equal and independent chance of being selected.

Non-independent sample



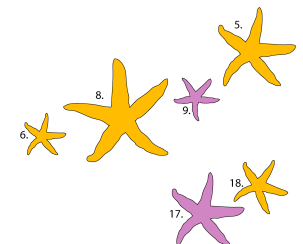
One procedure for random sampling



Number each individual

18, 6, 8, 5, 9, 17

Choose random numbers



Sample those individuals with matching numbers

Population parameters are *constants* whereas estimates are *random variables*, changing from one random sample to the next from the same population.

Larger samples on average will have smaller sampling error.

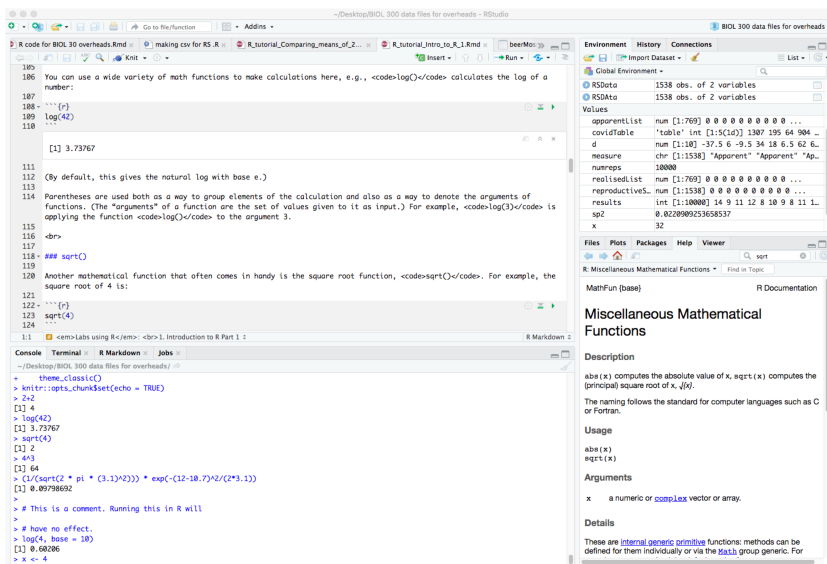
Sampling error

The difference between the estimate and average value of the estimate

Introduction to R



R is a free, open-source statistical language, widely used by scientists worldwide.

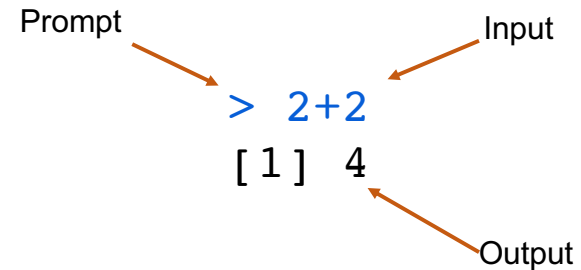


```

105 You can use a wide variety of math functions to make calculations here, e.g., <code>log()</code> calculates the log of a
106 number:
107
108 > log(42)
109 [1] 3.73767
110
111
112 (By default, this gives the natural log with base e.)
113
114 Parentheses are used both as a way to group elements of the calculation and also as a way to denote the arguments of
115 functions. (The "arguments" of a function are the set of values given to it as input.) For example, <code>log(3)</code> is
116 applying the function <code>log()</code> to the argument 3.
117
118
119 <br>
120 Another mathematical function that often comes in handy is the square root function, <code>sqrt()</code>. For example, the
121 square root of 4 is:
122
123 > sqrt(4)
124 [1] 2
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

A simple command in R



Basic arithmetic

```

> log(42)
[1] 3.73767
> sqrt(4)
[1] 2
> 4^3
[1] 64

```

Functions can have multiple inputs

```

> log(4, base = 10)
[1] 0.60206

```

You can define variables

```
> x <- 4
> x + 3
[1] 7
```

Calculations on data

```
> mean(titanicData$age, na.rm=TRUE)
[1] 31.19418
```

Data frames

```
> titanicData <- read.csv("DataForLabs/titanic.csv")
> head(titanicData)
```

	passenger_class	name	age	embarked
1	1st	Allen, Miss Elisabeth Walton	29.0000	Southampton
2	1st	Allison, Miss Helen Loraine	2.0000	Southampton
3	1st	Allison, Mr Hudson Joshua Creighton	30.0000	Southampton
4	1st	Allison, Mrs Hudson J.C. (Bessie Waldo Daniels)	25.0000	Southampton
5	1st	Allison, Master Hudson Trevor	0.9167	Southampton
6	1st	Anderson, Mr Harry	47.0000	Southampton

Each row is an individual;
each column is a variable describing each individual.

Calculations on data

```
> t.test(titanicData$age ~ titanicData$survive)
```

Welch Two Sample t-test

```
data: titanicData$age by titanicData$survive
t = 1.9947, df = 570.96, p-value = 0.04655
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 0.03644633 4.71184176
sample estimates:
mean in group no mean in group yes
    32.24811      29.87396
```