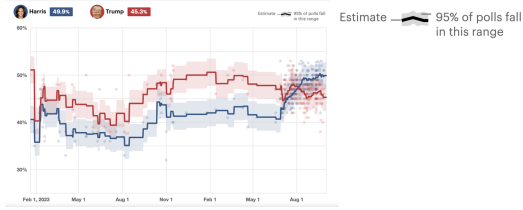


In polling, a 95% confidence interval means that if the same poll were repeated many times (i.e., sampling variation), the true population parameter (e.g., the percentage of people supporting a candidate) would fall within the calculated range in 95% of those polls.

The confidence interval accounts for the margin of error, giving us a range of values that we are fairly confident includes the true result. However, it's important to remember that this does not mean there is a 95% probability that the true value is in that specific range from one poll—it means that, over many polls, 95% of those calculated intervals will capture the true population parameter.



1

Lecture 10: Videos integrating lecture 10 (confidence intervals part 2) and tutorial 5.

- Tutorial 1: Introducing R
- Lecture 2: Key Jargon
- Lecture 3: Displaying data
- Tutorial 2: The R environment
- Lecture 4: Frequency distributions
- Lecture 5: Describing data
- Tutorial 3: Graphs
- Lecture 6: Describing data (part 2)
- Lecture 7: Sampling variation
- Tutorial 4: Describing data
- Lecture 8: Sampling distributions
- Lecture 9: Confidence Intervals
- Tutorial 5: Sampling variation
- Lecture 10: Confidence Intervals par...

part 1

General theory: from a computational approach to develop sampling distributions to number of sample means to a general statistical theory that can be used for "infinite" number of sample means.

part 2

2

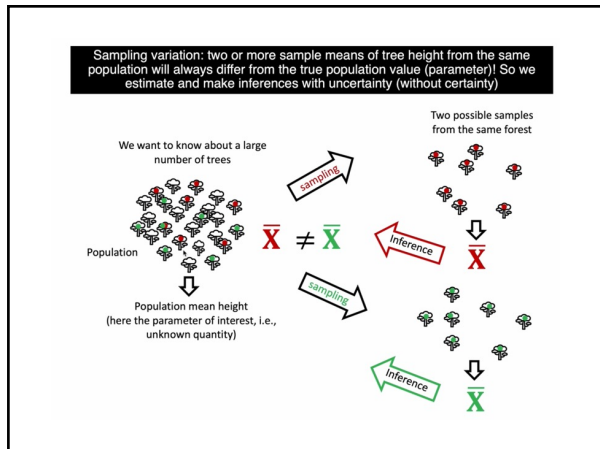
Estimating with uncertainty with some certainty

The statistical road: estimate with uncertainty but know how confident you can be!

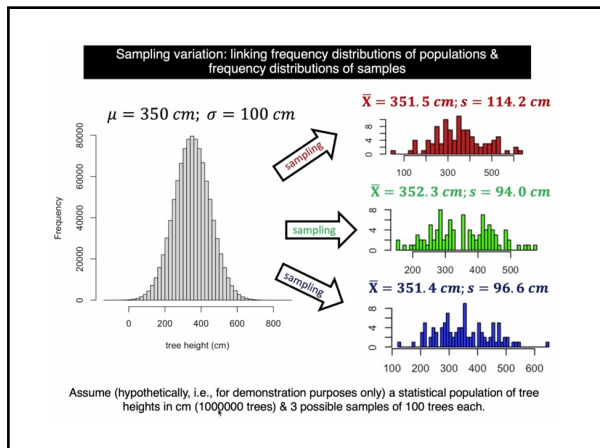
3

Sampling variation and Sampling distributions

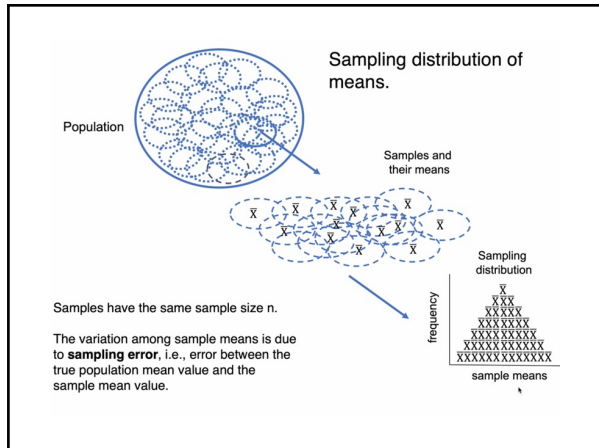
4



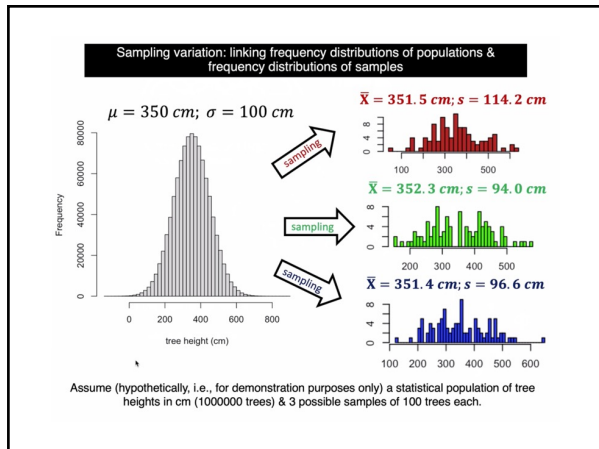
5



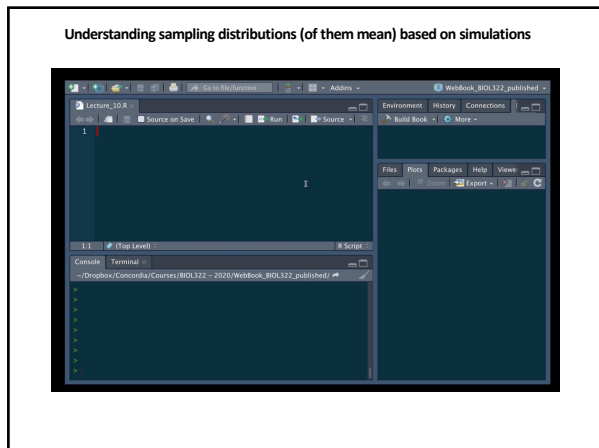
6



7



8



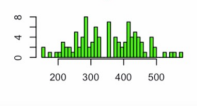
9

**Estimating with uncertainty with certainty
(i.e., with some confidence)**

A confidence interval is a range of values surrounding the sample estimate that is likely to contain the population parameter.

A large confidence interval (e.g., 95% or 99%) provides a most plausible range for a parameter. Values lying within the interval are most plausible, whereas those outside are less plausible, based **ON A SINGLE sample data alone**.

$\bar{x} = 352.3 \text{ cm}; s = 94.0 \text{ cm}$



$$\bar{Y} \pm 2SE_{\bar{Y}} \therefore SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

$$352.3 \pm 2 \times \frac{94.0}{\sqrt{100}}$$

333.5 cm 371.1 cm

10

The properties of sampling distributions: location and spread

The mean of all sample means equal the population mean:

$$\mu = \sum_{i=1}^{\infty} \bar{Y}_i$$

The number of samples is so large that for mathematical purposes it can be considered infinite (∞)

11

The properties of sampling distributions: location and spread

Sampling error - the difference between sample means and the population mean. The estimate of this error is the standard deviation of the sampling distribution, i.e., the average difference between all sample means and the true mean:

The standard deviation of the sampling distribution $\sigma_{\bar{Y}}$ is called standard error (SE) and is exactly:

$$\sigma_{\bar{Y}} = \sqrt{\sum_{i=1}^{\infty} \frac{(\bar{Y}_i - \mu)^2}{\infty}} = SE_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

The number of samples is so large that can be considered infinite (∞)

$\sigma =$ the standard deviation of the population

12

The properties of sampling distributions: location and spread

Given that we almost never know the population standard deviation, we estimate it with the sample value:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{y}}$ = the standard deviation of the sampling distribution of means

σ = the standard deviation of the population

13

The variation among observations within samples (standard deviation) can inform us about how far sample means in general might be from the true population mean (estimate how wrong one could be).

$\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$

$\bar{X} = 352.3 \text{ cm}; s = 94.0 \text{ cm}$

Variation among samples

Variation within samples (among observations) can estimate some certainty (confidence) about uncertainty (variation among sample means)

14

Understanding standard errors via simulations

```

32
33
34
35
36
37 lots.normal.samples.n100 <- replicate(100000, rnorm(n=100, mean=350, sd=100))
38 sample.means.n100 <- apply(X=lots.normal.samples.n100, MARGIN=2, FUN=mean)
39
40 mean(sample.means.n100)
41
42
    
```

40:24 (Top Level) R Script

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

The number of samples is so large that can be considered infinite (∞)

σ = the standard deviation of the population

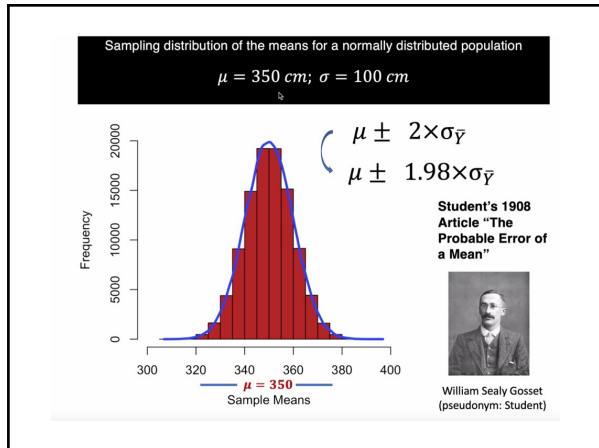
$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

Given that we almost never know the population standard deviation, we estimate it with the sample value:

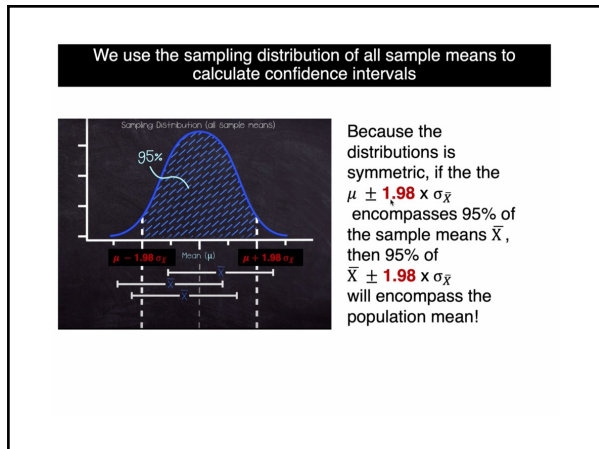
$\sigma_{\bar{y}}$ = the standard deviation of the sampling distribution of means

σ = the standard deviation of the population

15



16



17

Sampling variation: linking frequency distributions of populations & frequency distributions of samples

How many possible samples of 100 trees out of 100000 trees?

1e+15 (zeros)

The **human body** consists of some 37.2 trillion **cells** (3.72e+13 zeros)

18

Sampling variation: linking frequency distributions of populations & frequency distributions of samples

How many possible samples of 100 trees out of 1000000 trees?
1e+15 (zeros)

How many possible samples of 100 trees out of 1000000?
10768272362e+432 (zeros)

The **human body** consists of some
 37.2 trillion **cells**
 (3.72e+13 zeros)

19

Sampling distribution of the means for a normally distributed population
 $\mu = 350 \text{ cm}; \sigma = 100 \text{ cm}$

Frequency
 0 5000 10000 15000 20000

Sample Means
 300 320 340 360 380 400

density
 0.0 0.1 0.2 0.3 0.4

sample mean
 320 340 360 380

William Sealy Gosset
 (pseudonym: Student)

Student's 1908 Article "The Probable Error of a Mean"

20

Sample size increases precision

Frequency distribution of the gene Population

frequency
 0 200 400 600 800 1000 1200 1400

Gene length (number of nucleotides)

Sampling distributions for the sample means of the gene population (varying n)

probability
 0 0.02 0.04 0.06 0.08 0.10

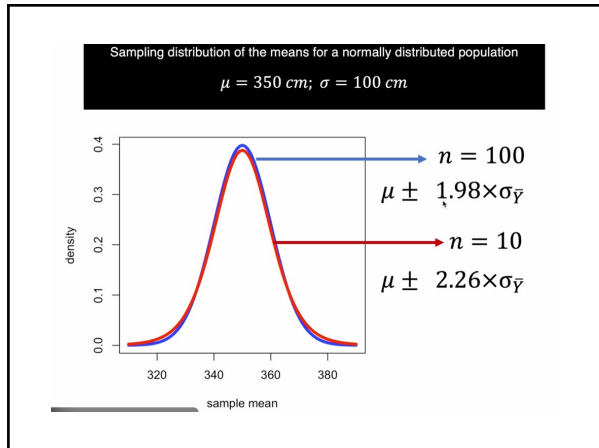
Sample mean length \bar{Y} (nucleotides)

n=20
 n=100
 n=500

precision

Whitlock & Schluter, 2nd edition; 3rd edition has a different set of genes.

21



22

The properties of sampling distributions: location and spread

Given that we almost never know the population standard deviation, we estimate it with the sample value:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

$$SE_{\bar{y}} = \frac{s}{\sqrt{n}}$$

$\sigma_{\bar{y}}$ = the standard deviation of the sampling distribution of means

σ = the standard deviation of the population

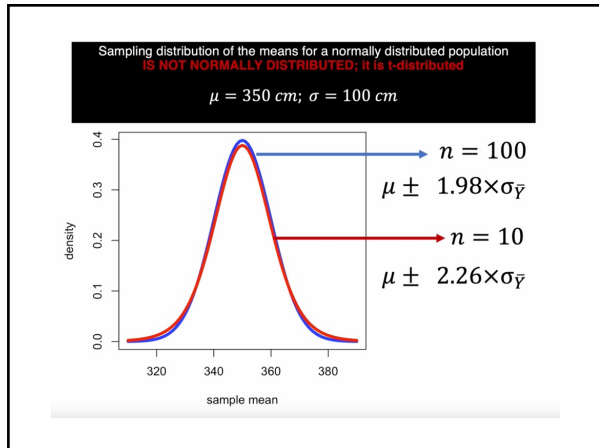
23

Connecting standard errors to confidence intervals
 Understanding confidence intervals via simulations

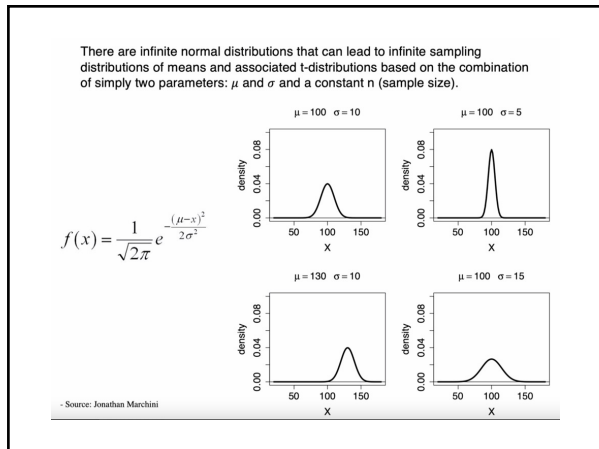
```

116
117
118
119
120
121
122
123
124
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
    
```

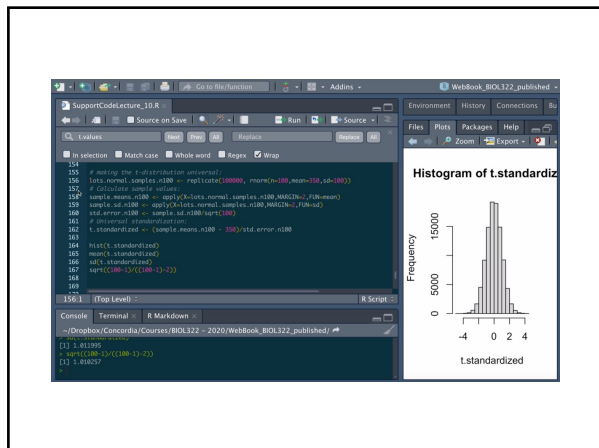
24



25



26



27

By now, you should suspect that one of the inconveniences is that the exact value needed to be multiplied by SE to create 95% confidence intervals changes as a function of sample size.

The sampling distribution of means that varies as a function of the sample size (here $v = \text{degrees of freedom}; v = n - 1$).

This t distribution is a sampling distribution of the the number of sample standard errors away from the mean (now always 0 after the standardization) necessary to produce a confidence interval of the desired coverage (e.g., 95%).

$$t = \frac{\bar{X}_i - \mu}{SE_{\bar{X}_i}} \rightarrow \bar{X}_i \pm t \times SE_{\bar{X}_i}$$

28

29

How to find the appropriate values of t ?

the old days of tables allow to understand the principle – in practice (today) we use software (e.g., R).

Degrees of freedom	Two-sided	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	2.5%	1%
1	1.000	1.378	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6	1273	3183
2	0.818	1.060	1.386	1.886	2.920	4.303	6.965	9.925	14.08	22.32	31.83	63.66	127.3
3	0.766	0.978	1.250	1.699	2.353	3.182	4.541	5.841	7.453	10.21	13.71	19.00	25.01
4	0.741	0.941	1.190	1.601	2.132	2.776	3.747	4.604	5.598	7.173	8.610	10.21	11.91
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	8.151	9.247
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	7.007	7.877
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	6.314	7.007
8	0.706	0.889	1.108	1.397	1.860	2.308	2.896	3.358	3.833	4.501	5.041	5.833	6.314
9	0.701	0.883	1.101	1.389	1.833	2.262	2.821	3.264	3.690	4.297	4.791	5.581	6.055
10	0.697	0.879	1.093	1.372	1.812	2.228	2.750	3.183	3.591	4.144	4.587	5.377	5.851
11	0.693	0.876	1.088	1.363	1.796	2.201	2.718	3.146	3.547	4.095	4.537	5.327	5.801
12	0.689	0.873	1.083	1.356	1.782	2.179	2.681	3.105	3.498	4.044	4.487	5.277	5.751

Assume a sample size of $n = 9$, then the degrees of freedom would be 8 for the t value to calculate the confidence interval for the sample mean.

$$\bar{X}_i \pm t \times SE_{\bar{X}_i} \therefore \bar{X}_i \pm 2.306 \frac{s_i}{\sqrt{9}}$$

30
