The Central limit theorem

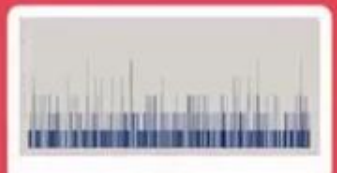
#UnmatchedPracticalApplication

Dataset

985, 978, 435, 389, 79, 926, 299, 538, 571, 828, 681,
302, 13, 518, 873, 256, 899, 864, 314, 470, 547, 440,
699, 867, 860, 202, 155, 792, 64, 406, 906, 859, 584,
375, 996, 466, 401, 428, 714, 453, 194, 487, 993, 34,
829, 317, 865, 296, 197, 895, 208, 613, 98, 487, 963, 81,
808, 182, 5, 869, 291, 549, 489, 49, 941, 473, 116, 705,
340, 209, 547, 156, 735, 573, 234, 259, 704, 711, 892,
509, 680, 280, 819, 385, 618, 666, 599, 389, 229, 862,
288, 971, 656, 18, 774, 226, 990, 786, 828, 605

any distribution



Sampling

985, 978, 435, 389, 7[illegible]26, 299, 538, 571, 828, 681,
302, 13, 518, 873, 256[illegible]9, 864, 314, 470, 547, 440,
699, 867, 860, 20[illegible]35, 792, 64, 406, 906, 859, 584,
375, 996, 466, 401, 428, 714, 453, 194, 487, 993, 34,
829, 317, 865, 296, 197, 895, 208, 613, 98, 487, 963, 81,
808, 182, 5, 869, 291, 549, 489, 49, 941, 473, 116, 705,
340, 209, 547, 156, 735, 573, 234, 259, 704, 711, 892,
509, 680, 280, 819, 385, 618, 666, 599, 389, 229, 862,
288, 971, 656, 18, 774, 226, 990, 786, 828, 605

Mean of sample k

$\bar{x}_1 = 555.2$

$\bar{x}_2 = 439.5$

$\bar{x}_3 = 625.3$

$\bar{x}_k = 567.5$

any distribution




The Central limit theorem

No matter the distribution

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k$

$$N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

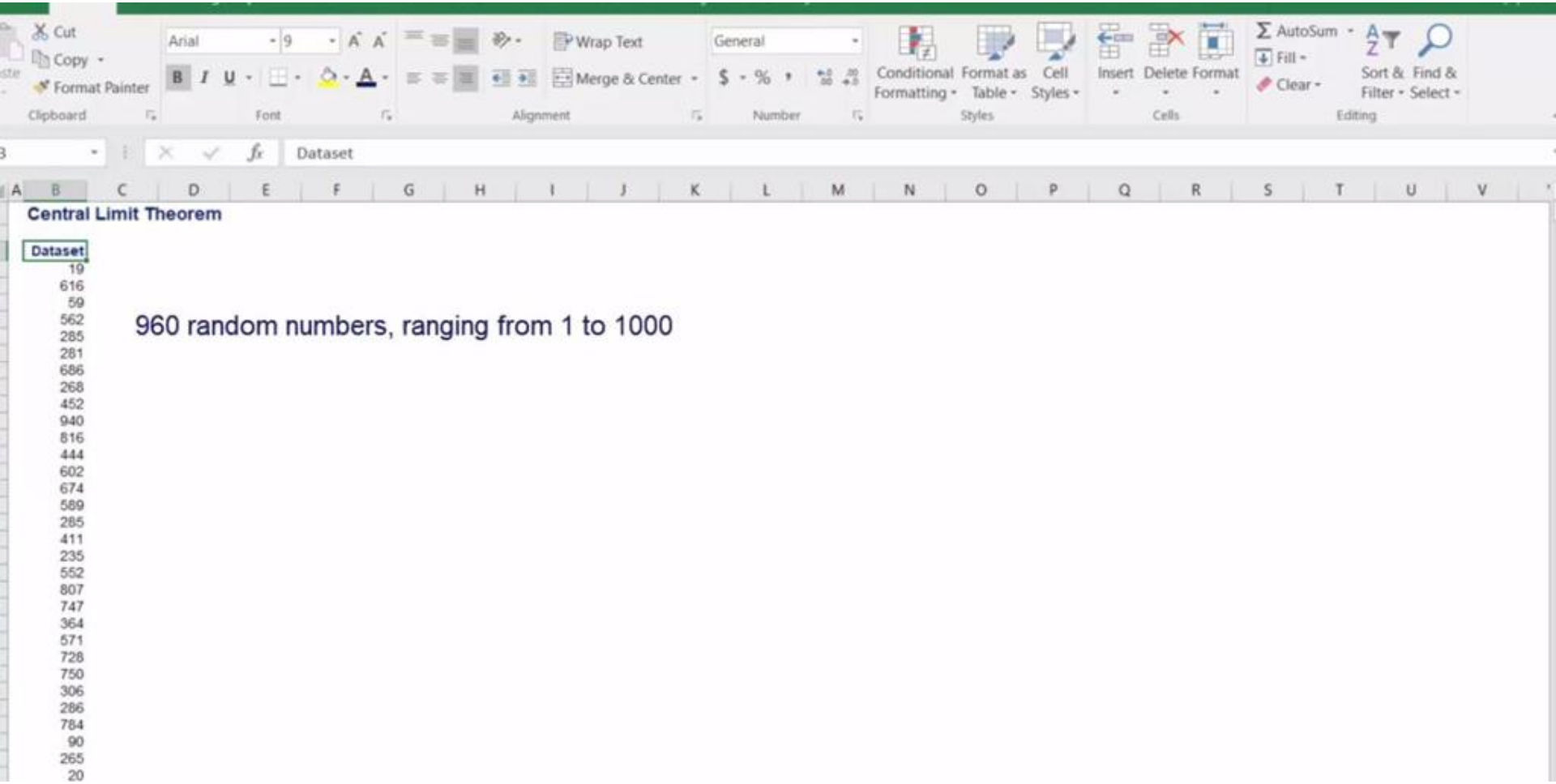
The more samples you extract

$k \longrightarrow \infty$

The bigger the samples

$n \longrightarrow \infty$

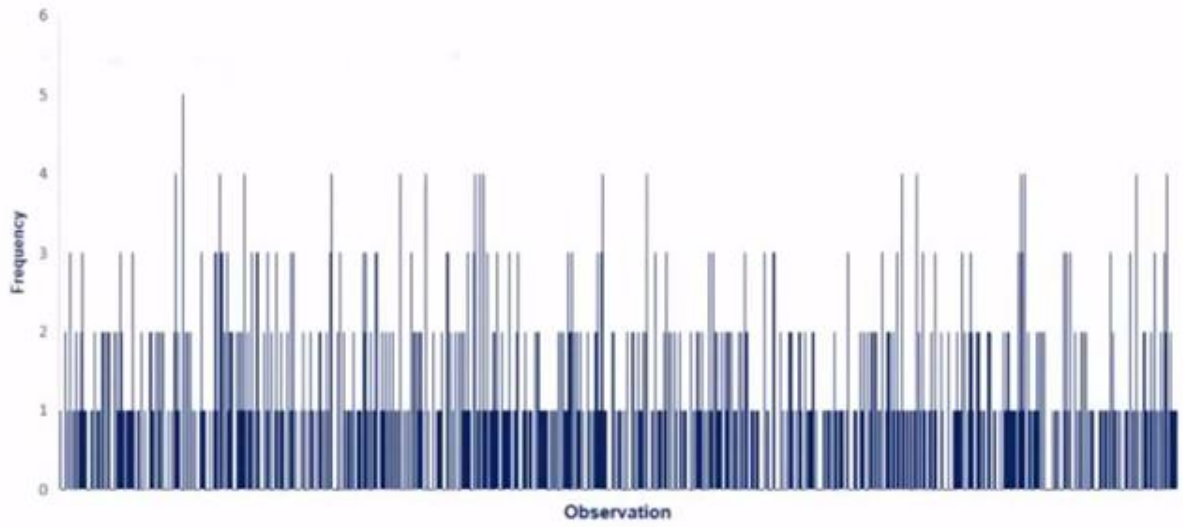
#UnmatchedPracticalApplication

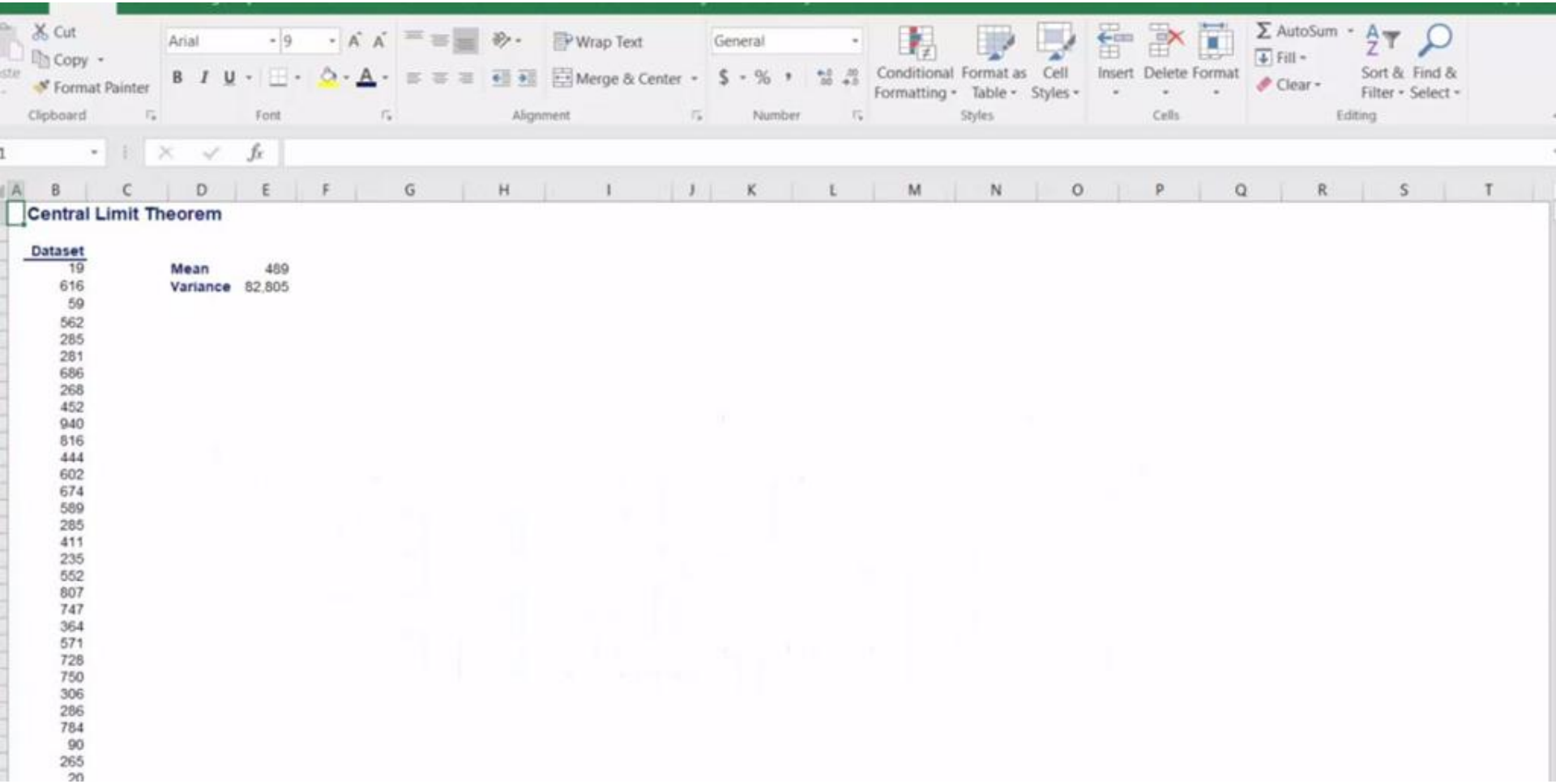

Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
Dataset
Central Limit Theorem
Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
20
960 random numbers, ranging from 1 to 1000

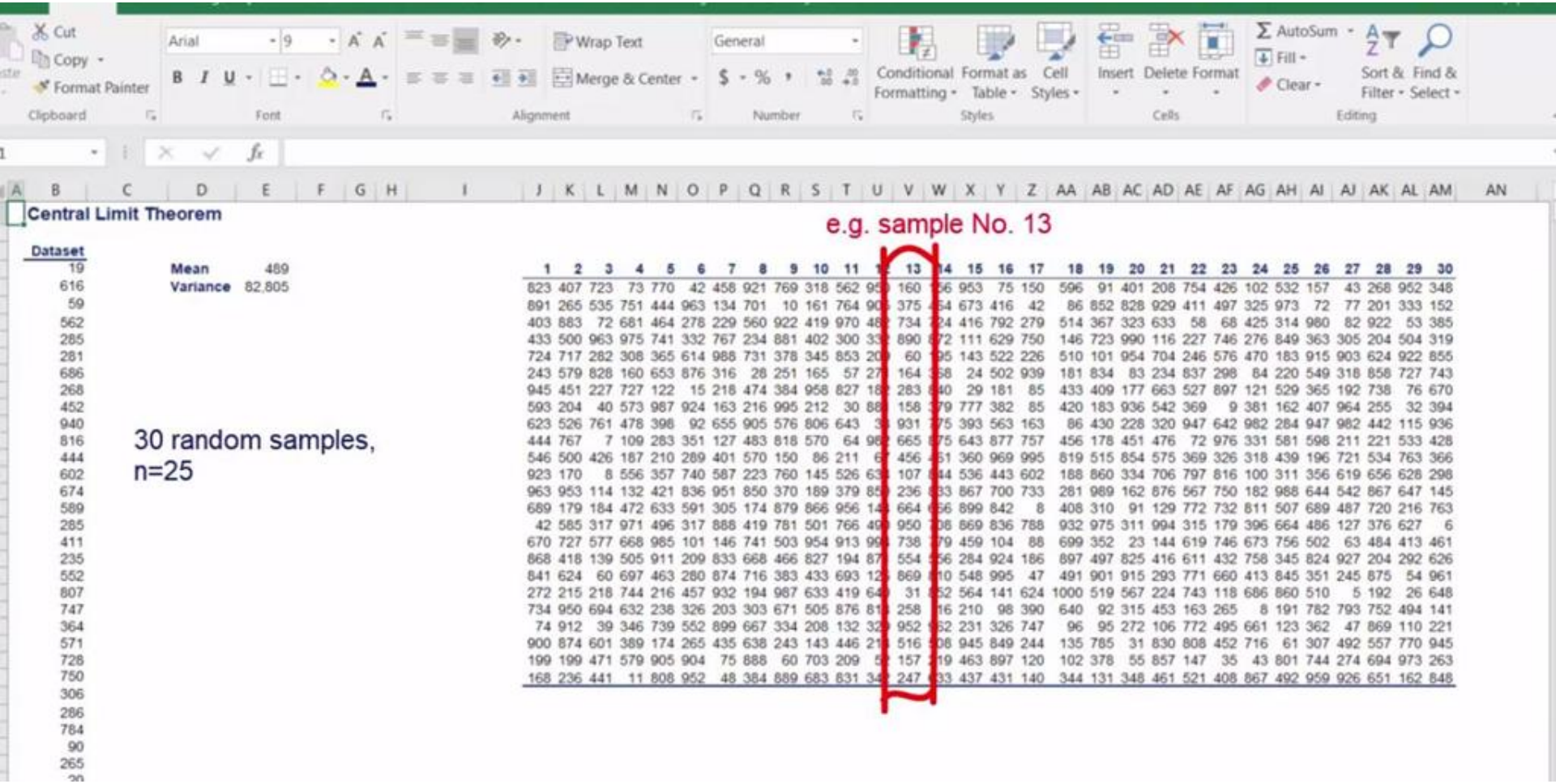
Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing

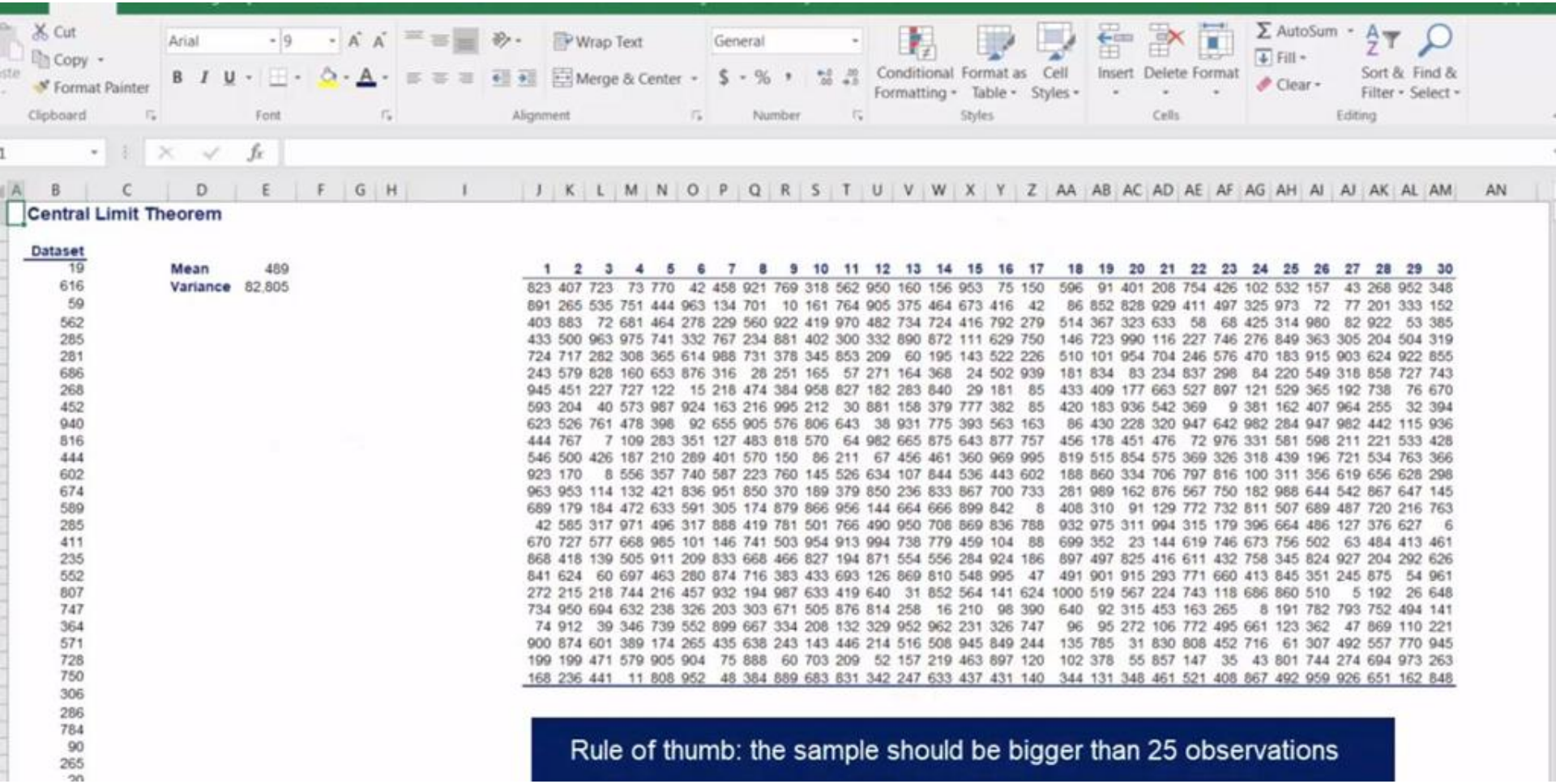
Central Limit Theorem

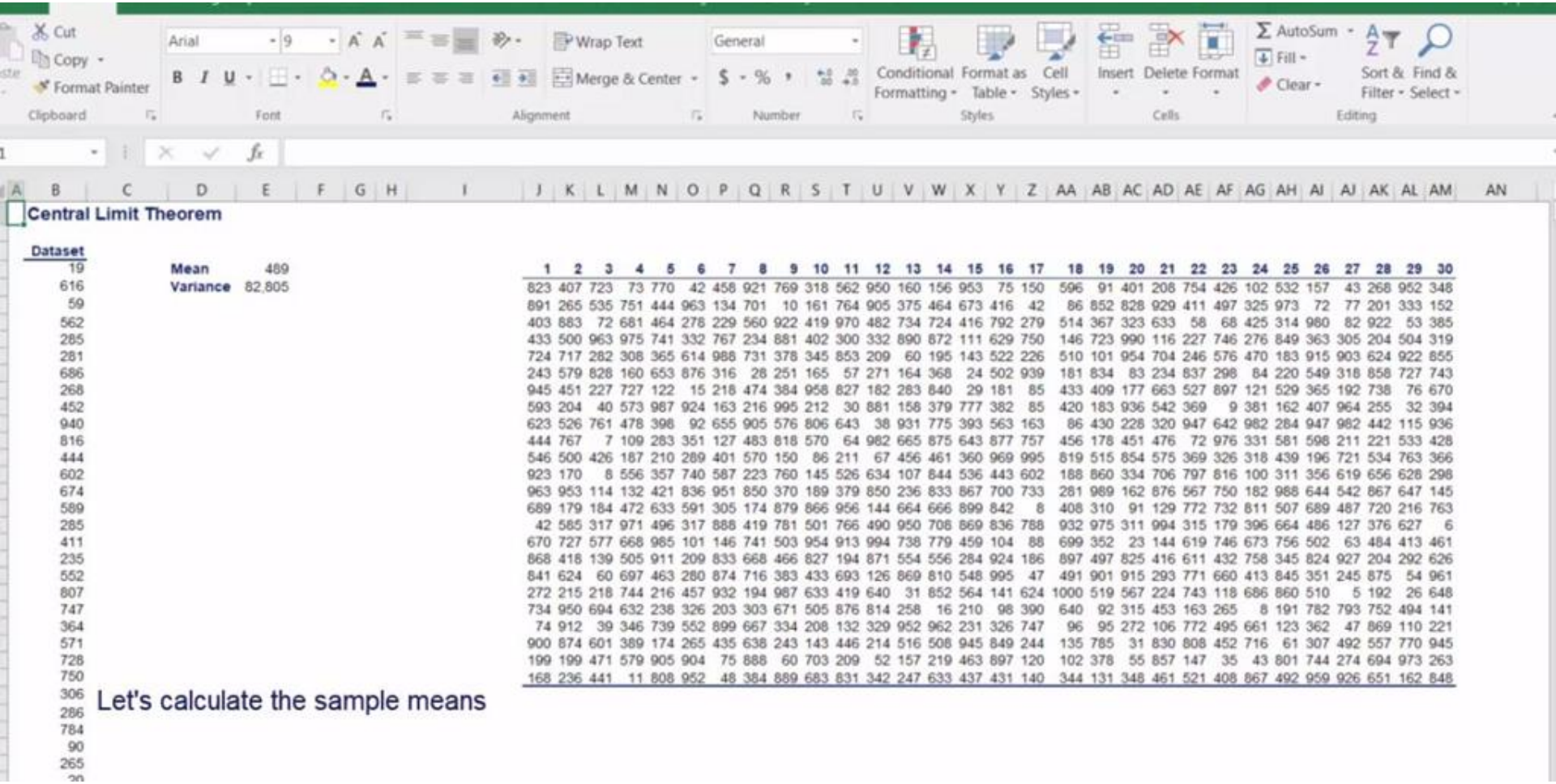
Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
20




Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
Central Limit Theorem
Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
20
Mean 489
Variance 82,805


Cut
Copy
Format Painter
Arial
9
Wrap Text
General
Merge & Center
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
Central Limit Theorem
e.g. sample No. 13
Dataset
Mean 489
Variance 82.805
30 random samples,
n=25


Cut
Copy
Format Painter
Arial
9
Wrap Text
General
Merge & Center
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
Central Limit Theorem
Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
Mean 489
Variance 82.805
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
823 407 723 73 770 42 458 921 769 318 562 950 160 156 953 75 150 596 91 401 208 754 426 102 532 157 43 268 952 348
891 265 535 751 444 963 134 701 10 161 764 905 375 464 673 416 42 86 852 828 929 411 497 325 973 72 77 201 333 152
403 883 72 681 464 278 229 560 922 419 970 482 734 724 416 792 279 514 367 323 633 58 68 425 314 980 82 922 53 385
433 500 963 975 741 332 767 234 881 402 300 332 890 872 111 629 750 146 723 990 116 227 746 276 849 363 305 204 504 319
724 717 282 308 365 614 988 731 378 345 853 209 60 195 143 522 226 510 101 954 704 246 576 470 183 915 903 624 922 855
243 579 828 160 653 876 316 28 251 165 57 271 164 368 24 502 939 181 834 83 234 837 298 84 220 549 318 858 727 743
945 451 227 727 122 15 218 474 384 958 827 182 283 840 29 181 85 433 409 177 663 527 897 121 529 365 192 738 76 670
593 204 40 573 987 924 163 216 995 212 30 881 158 379 777 382 85 420 183 936 542 369 9 381 162 407 964 255 32 394
623 526 761 478 398 92 655 905 576 806 643 38 931 775 393 563 163 86 430 228 320 947 642 982 284 947 982 442 115 936
444 767 7 109 283 351 127 483 818 570 64 982 665 875 643 877 757 456 178 451 476 72 976 331 581 598 211 221 533 428
546 500 426 187 210 289 401 570 150 86 211 67 456 461 360 969 995 819 515 854 575 369 326 318 439 196 721 534 763 366
923 170 8 556 357 740 587 223 760 145 526 634 107 844 536 443 602 188 860 334 706 797 816 100 311 356 619 656 628 298
963 953 114 132 421 836 951 850 370 189 379 850 236 833 867 700 733 281 989 162 876 567 750 182 988 644 542 867 647 145
689 179 184 472 633 591 305 174 879 866 956 144 664 666 899 842 8 408 310 91 129 772 732 811 507 689 487 720 216 763
42 585 317 971 496 317 888 419 781 501 766 490 950 708 869 836 788 932 975 311 994 315 179 396 664 486 127 376 627 6
670 727 577 668 985 101 146 741 503 954 913 994 738 779 459 104 88 699 352 23 144 619 746 673 756 502 63 484 413 461
868 418 139 505 911 209 833 668 466 827 194 871 554 556 284 924 186 897 497 825 416 611 432 758 345 824 927 204 292 626
841 624 60 697 463 280 874 716 383 433 693 126 869 810 548 995 47 491 901 915 293 771 660 413 845 351 245 875 54 961
272 215 218 744 216 457 932 194 987 633 419 640 31 852 564 141 624 1000 519 567 224 743 118 686 860 510 5 192 26 648
734 950 694 632 238 326 203 303 671 505 876 814 258 16 210 98 390 640 92 315 453 163 265 8 191 782 793 752 494 141
74 912 39 346 739 552 899 667 334 208 132 329 952 962 231 326 747 96 95 272 106 772 495 661 123 362 47 869 110 221
900 874 601 389 174 265 435 638 243 143 446 214 516 508 945 849 244 135 785 31 830 808 452 716 61 307 492 557 770 945
199 199 471 579 905 904 75 888 60 703 209 52 157 219 463 897 120 102 378 55 857 147 35 43 801 744 274 694 973 263
168 236 441 11 808 952 48 384 889 683 831 342 247 633 437 431 140 344 131 348 461 521 408 867 492 959 926 651 162 848
Rule of thumb: the sample should be bigger than 25 observations



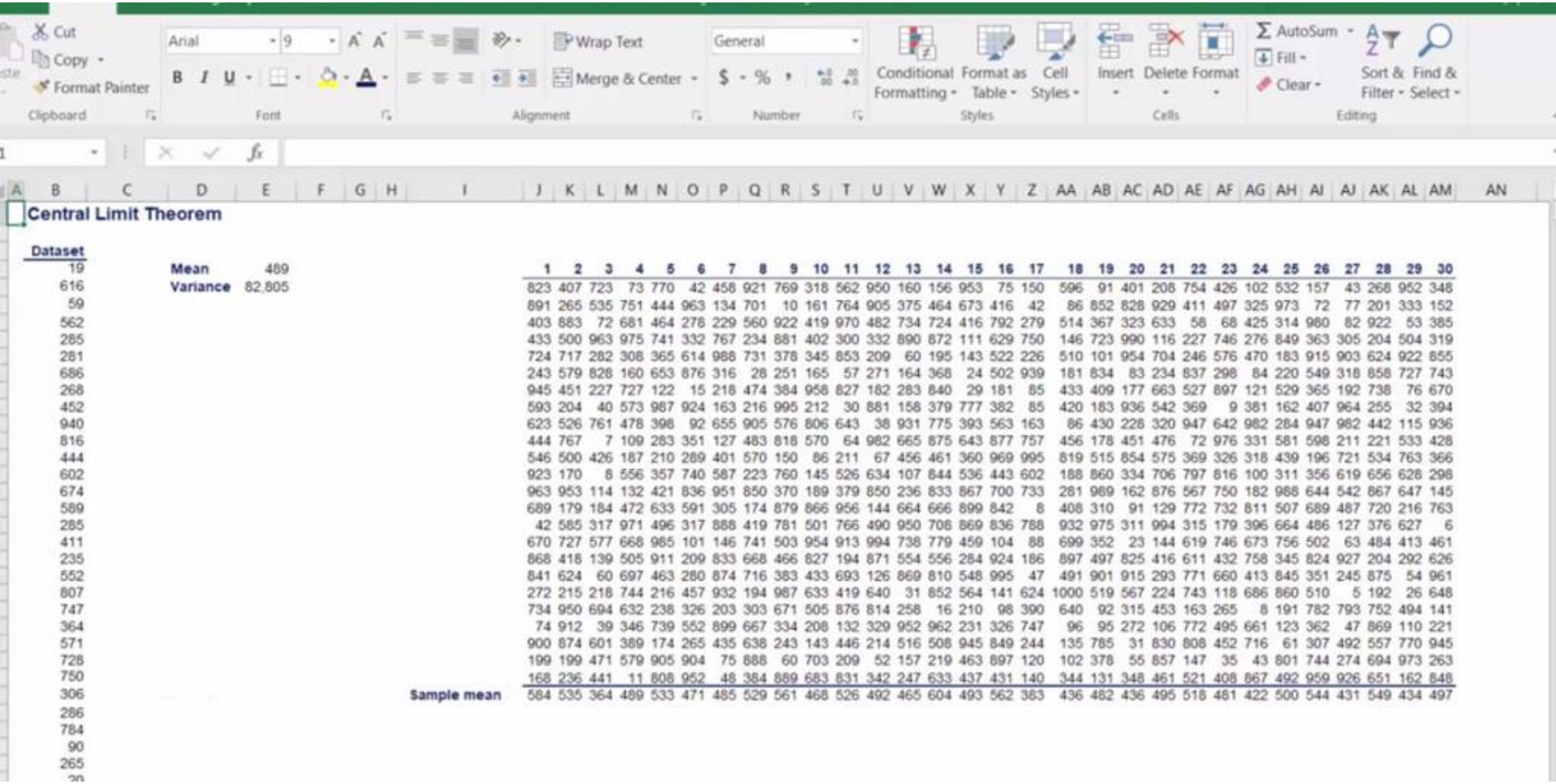
Central Limit Theorem

Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265

Mean	489
Variance	82,805

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
823	407	723	73	770	42	458	921	769	318	562	950	160	156	953	75	150	596	91	401	208	754	426	102	532	157	43	268	952	348
891	265	535	751	444	963	134	701	10	161	764	905	375	464	673	416	42	86	852	828	929	411	497	325	973	72	77	201	333	152
403	883	72	681	464	278	229	560	922	419	970	482	734	724	416	792	279	514	367	323	633	58	68	425	314	980	82	922	53	385
433	500	963	975	741	332	767	234	881	402	300	332	890	872	111	629	750	146	723	990	116	227	746	276	849	363	305	204	504	319
724	717	282	308	365	614	988	731	378	345	853	209	60	195	143	522	226	510	101	954	704	246	576	470	183	915	903	624	922	855
243	579	828	160	653	876	316	28	251	165	57	271	164	368	24	502	939	181	834	83	234	837	298	84	220	549	318	858	727	743
945	451	227	727	122	15	218	474	384	958	827	182	283	840	29	181	85	433	409	177	663	527	897	121	529	365	192	738	76	670
593	204	40	573	987	924	163	216	995	212	30	881	158	379	777	382	85	420	183	936	542	369	9	381	162	407	964	255	32	394
623	526	761	478	398	92	655	905	576	806	643	38	931	775	393	563	163	86	430	228	320	947	642	982	284	947	982	442	115	936
444	767	7	109	283	351	127	483	818	570	64	982	665	875	643	877	757	456	178	451	476	72	976	331	581	598	211	221	533	428
546	500	426	187	210	289	401	570	150	86	211	67	456	461	360	969	995	819	515	854	575	369	326	318	439	196	721	534	763	366
923	170	8	556	357	740	587	223	760	145	526	634	107	844	536	443	602	188	860	334	706	797	816	100	311	356	619	656	628	298
963	953	114	132	421	836	951	850	370	189	379	850	236	833	867	700	733	281	989	162	876	567	750	182	988	644	542	867	647	145
689	179	184	472	633	591	305	174	879	866	956	144	664	666	899	842	8	408	310	91	129	772	732	811	507	689	487	720	216	763
42	585	317	971	496	317	888	419	781	501	766	490	950	708	869	836	788	932	975	311	994	315	179	396	664	486	127	376	627	6
670	727	577	668	985	101	146	741	503	954	913	994	738	779	459	104	88	699	352	23	144	619	746	673	756	502	63	484	413	461
868	418	139	505	911	209	833	668	466	827	194	871	554	556	284	924	186	897	497	825	416	611	432	758	345	824	927	204	292	626
841	624	60	697	463	280	874	716	383	433	693	126	869	810	548	995	47	491	901	915	293	771	660	413	845	351	245	875	54	961
272	215	218	744	216	457	932	194	987	633	419	640	31	852	564	141	624	1000	519	567	224	743	118	686	860	510	5	192	26	648
734	950	694	632	238	326	203	303	671	505	876	814	258	16	210	98	390	640	92	315	453	163	265	8	191	782	793	752	494	141
74	912	39	346	739	552	899	667	334	208	132	329	952	962	231	326	747	96	95	272	106	772	495	661	123	362	47	869	110	221
900	874	601	389	174	265	435	638	243	143	446	214	516	508	945	849	244	135	785	31	830	808	452	716	61	307	492	557	770	945
199	199	471	579	905	904	75	888	60	703	209	52	157	219	463	897	120	102	378	55	857	147	35	43	801	744	274	694	973	263
168	236	441	11	808	952	48	384	889	683	831	342	247	633	437	431	140	344	131	348	461	521	408	867	492	959	926	651	162	848

Let's calculate the sample means



Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing

Central Limit Theorem

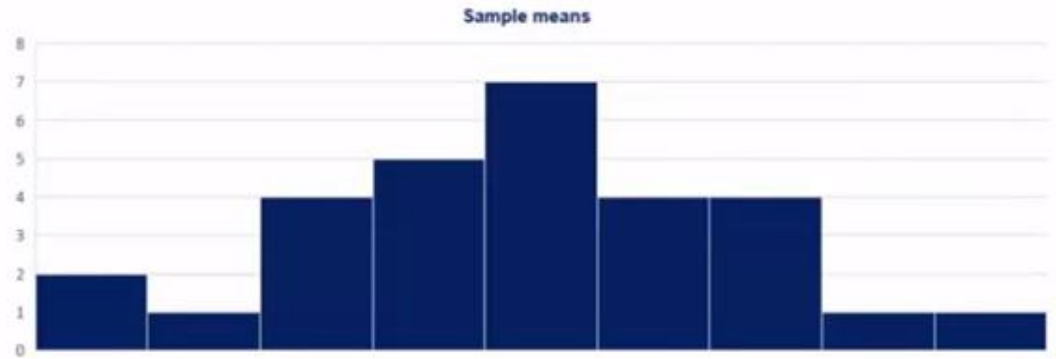
Dataset		
19	**Mean**	489
616	**Variance**	82,805
59		
562		
285		
281		
686		
268		
452		
940		
816		
444		
602		
674		
589		
285		
411		
235		
552		
807		
747		
364		
571		
728		
750		
306		
286		
784		
90		
265		

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	823	407	723	73	770	42	458	921	769	318	562	950	160	156	953	75	150	596	91	401	208	754	426	102	532	157	43	268	952	348
	891	265	535	751	444	963	134	701	10	161	764	905	375	464	673	416	42	86	852	828	929	411	497	325	973	72	77	201	333	152
	403	883	72	681	464	278	229	560	922	419	970	482	734	724	416	792	279	514	367	323	633	58	68	425	314	980	82	922	53	385
	433	500	963	975	741	332	767	234	881	402	300	332	890	872	111	629	750	146	723	990	116	227	746	276	849	363	305	204	504	319
	724	717	282	308	365	614	988	731	378	345	853	209	60	195	143	522	226	510	101	954	704	246	576	470	183	915	903	624	922	855
	243	579	828	160	653	876	316	28	251	165	57	271	164	368	24	502	939	181	834	83	234	837	298	84	220	549	318	858	727	743
	945	451	227	727	122	15	218	474	384	958	827	182	283	840	29	181	85	433	409	177	663	527	897	121	529	365	192	738	76	670
	593	204	40	573	987	924	163	216	995	212	30	881	158	379	777	382	85	420	183	936	542	369	9	381	162	407	964	255	32	394
	623	526	761	478	398	92	655	905	576	806	643	38	931	775	393	563	163	86	430	228	320	947	642	982	284	947	982	442	115	936
	444	767	7	109	283	351	127	483	818	570	64	982	665	875	643	877	757	456	178	451	476	72	976	331	581	598	211	221	533	428
	546	500	426	187	210	289	401	570	150	86	211	67	456	461	360	969	995	819	515	854	575	369	326	318	439	196	721	534	763	366
	923	170	8	556	357	740	587	223	760	145	526	634	107	844	536	443	602	188	860	334	706	797	816	100	311	356	619	656	628	298
	963	953	114	132	421	836	951	850	370	189	379	850	236	833	867	700	733	281	989	162	876	567	750	182	988	644	542	867	647	145
	689	179	184	472	633	591	305	174	879	866	956	144	664	666	899	842	8	408	310	91	129	772	732	811	507	689	487	720	216	763
	42	585	317	971	496	317	888	419	781	501	766	490	950	708	869	836	788	932	975	311	994	315	179	396	664	486	127	376	627	6
	670	727	577	668	985	101	146	741	503	954	913	994	738	779	459	104	88	699	352	23	144	619	746	673	756	502	63	484	413	461
	868	418	139	505	911	209	833	668	466	827	194	871	554	556	284	924	186	897	497	825	416	611	432	758	345	824	927	204	292	626
	841	624	60	697	463	280	874	716	383	433	693	126	869	810	548	995	47	491	901	915	293	771	660	413	845	351	245	875	54	961
	272	215	218	744	216	457	932	194	987	633	419	640	31	852	564	141	624	1000	519	567	224	743	118	686	860	510	5	192	26	648
	734	950	694	632	238	326	203	303	671	505	876	814	258	16	210	98	390	640	92	315	453	163	265	8	191	782	793	752	494	141
	74	912	39	346	739	552	899	667	334	208	132	329	952	962	231	326	747	96	95	272	106	772	495	661	123	362	47	869	110	221
	900	874	601	389	174	265	435	638	243	143	446	214	516	508	945	849	244	135	785	31	830	808	452	716	61	307	492	557	770	945
	199	199	471	579	905	904	75	888	60	703	209	52	157	219	463	897	120	102	378	55	857	147	35	43	801	744	274	694	973	263
	168	236	441	11	808	952	48	384	889	683	831	342	247	633	437	431	140	344	131	348	461	521	408	867	492	959	926	651	162	848
Sample mean	584	535	364	489	533	471	485	529	561	468	526	492	465	604	493	562	383	436	482	436	495	518	481	422	500	544	431	549	434	497

Central Limit Theorem

$$N \sim (\mu, \sigma^2)$$

Dataset			Sample means
19	Mean	489	564
616	Variance	82,805	535
59			364
562			489
285			533
281			471
686			485
268			561
452			468
940			526
816			492
444			465
602			604
674			493
589			562
285			383
411			436
235			482
552			436
807			495
747			518
364			481
571			422
728			500
750			544
306			431
286			549
784			434
90			497
265			
20			



Central Limit Theorem

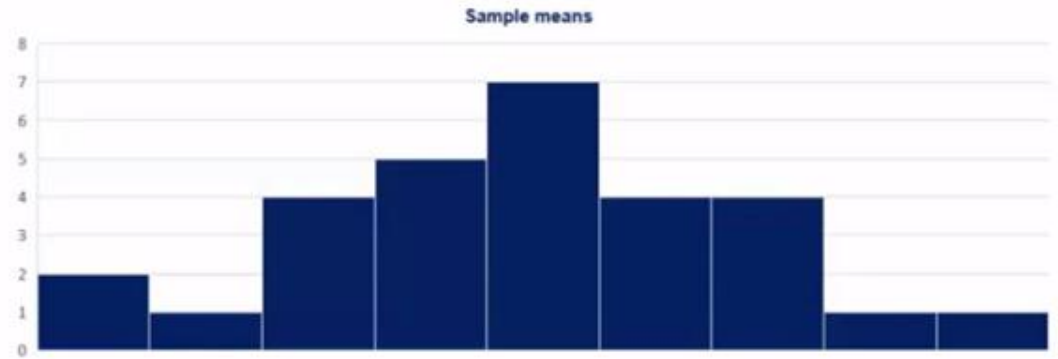
Observed: $N \sim (492, 3171)$

Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
20

Mean	489
Variance	82,805

Sample means
564
535
364
489
533
471
485
561
468
526
492
465
604
493
562
383
436
462
436
495
518
481
422
500
544
431
549
434
497

Mean	492
Variance	3,171



Cut
Copy
Format Painter
Clipboard
Arial
9
Font
Wrap Text
Merge & Center
Alignment
Number
Number
Conditional Formatting
Format as Table
Cell Styles
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing

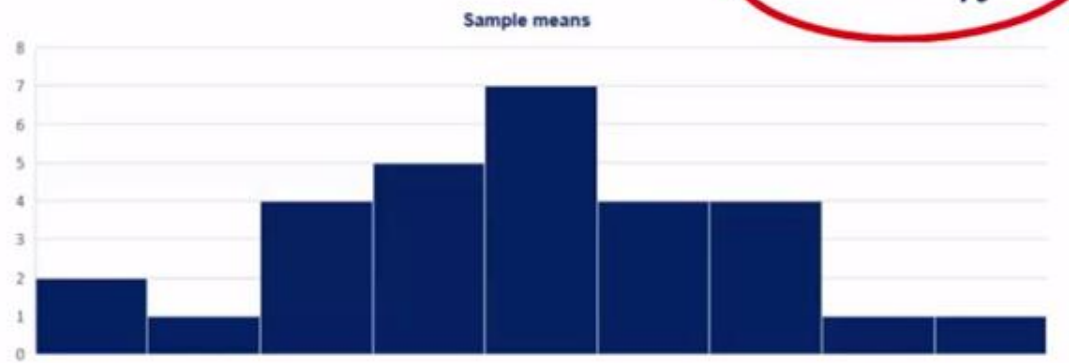
B3
823

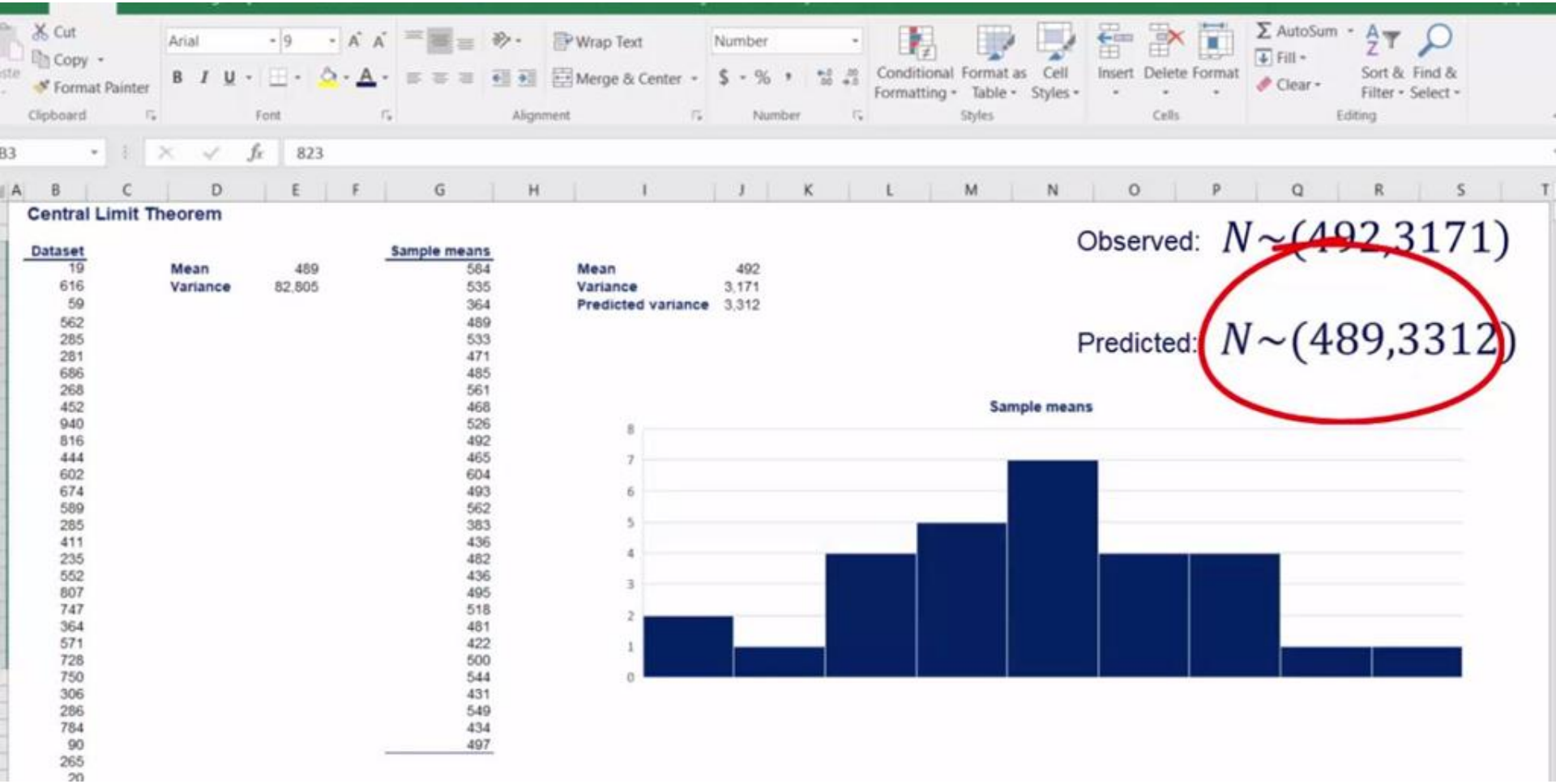
Central Limit Theorem

Dataset			Sample means		
19	Mean	489	564	Mean	492
616	Variance	82,805	535	Variance	3,171
59			364		
562			489		
285			533		
281			471		
686			485		
268			561		
452			468		
940			526		
816			492		
444			465		
602			604		
674			493		
589			562		
285			383		
411			436		
235			482		
552			436		
807			495		
747			518		
364			481		
571			422		
728			500		
750			544		
306			431		
286			549		
784			434		
90			497		
265					
20					

Observed: $N \sim (492, 3171)$

Predicted: $N \sim (\mu_o, \frac{\sigma_o^2}{n})$




Cut
Copy
Format Painter
Arial
9
Wrap Text
Merge & Center
Number
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
B3
823
Central Limit Theorem
Dataset
19
616
59
562
285
281
686
268
452
940
816
444
602
674
589
285
411
235
552
807
747
364
571
728
750
306
286
784
90
265
Mean 489
Variance 82,805
Sample means
564
535
364
489
533
471
485
561
468
526
492
465
604
493
562
383
436
482
436
495
518
481
422
500
544
431
549
434
497
Mean 492
Variance 3,171
Predicted variance 3,312
Observed: $N \sim (492, 3171)$
Predicted: $N \sim (489, 3312)$
Sample means
0
1
2
3
4
5
6
7
8

Standard error

Def. The standard deviation of the distribution formed by the sample means

How to find the standard error?

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k \qquad N \sim \left(\mu, \frac{\sigma^2}{n}\right)$$

Variance

$$\text{Standard deviation} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Meaning of the standard error

Like any standard deviation, it shows variability

Meaning of the standard error

Variability.. of the sample means

Meaning of the standard error

Used in most statistical tests

Because it shows how well you approximated the true mean

#VeryImportant

Note

$$\frac{\sigma}{\sqrt{n}}$$ ↓ n ↑

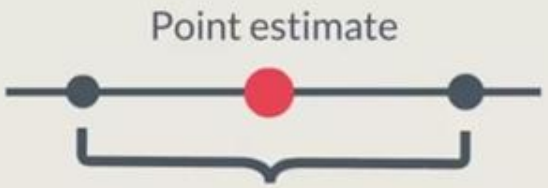
Standard error decreases when sample size increases

Bigger sample - better approximation

Estimators and estimates

single number

Point estimates



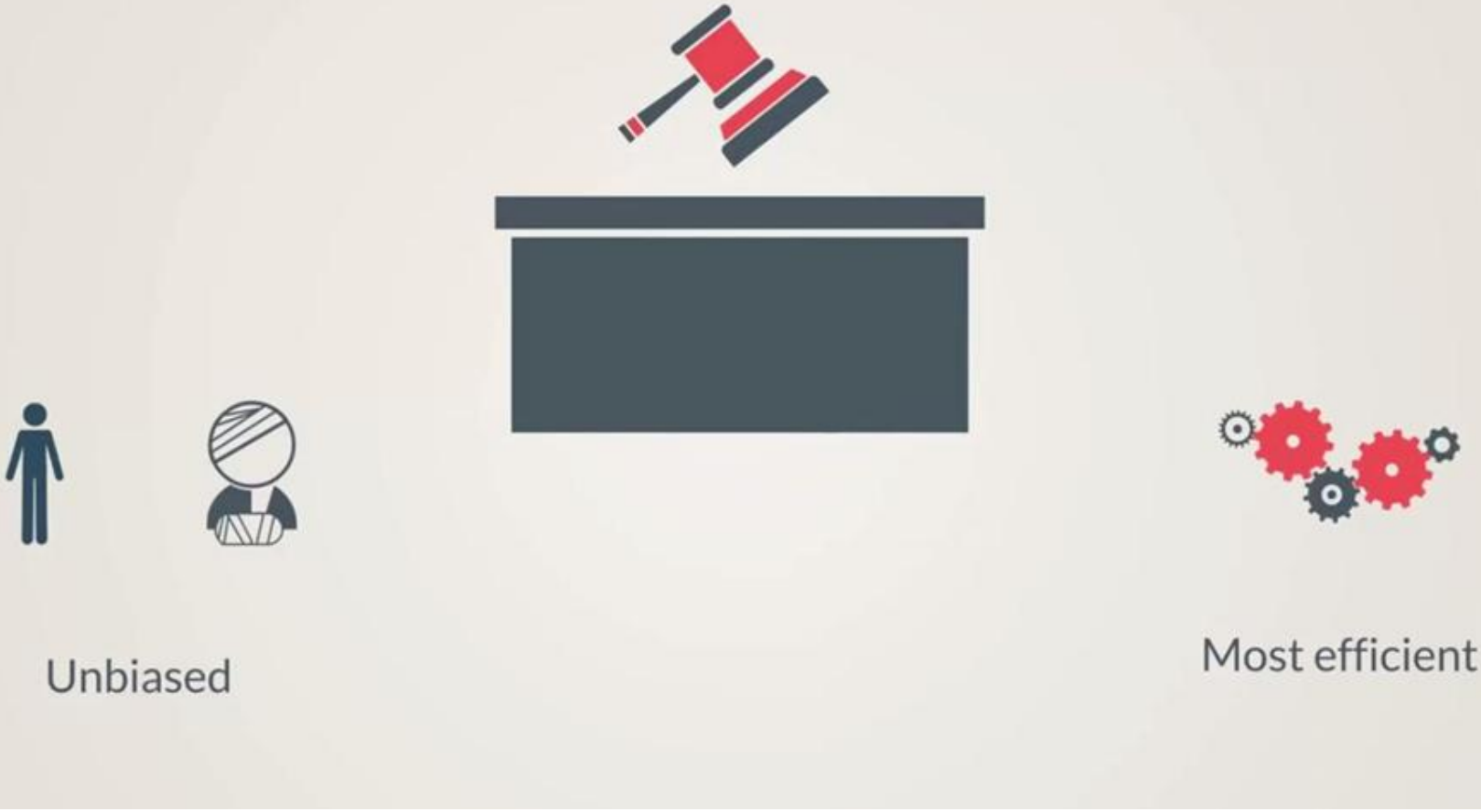
Confidence interval estimates

Point estimators and estimates

Estimator /how to estimate/		Parameter /what to estimate/		Estimate /concrete result/
$\bar{x}$	of	μ	→	52.22
s^2	of	σ^2	→	1724.93

Bias

Efficiency


Unbiased
Most efficient

Bias

An unbiased estimator has an expected value equal to the population parameter

e.g. $\bar{x}$ has an expected value of μ

Bias

$\bar{x}$ + 1 ft estimates μ

Bias

$\bar{x}$ estimates μ with no bias

$\bar{x} + 1\text{ ft}$ estimates μ with a bias of 1 ft

Efficiency

The most efficient estimator is the unbiased estimator with the smallest variance

London

You visit 5% of the restaurants

Average price

22.50£

You visit 5% of the restaurants

Average price

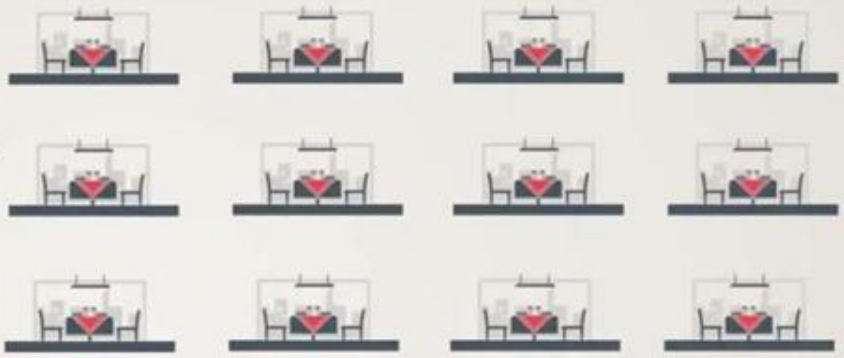
22.50£

20 £ 25 £

Confidence interval: [20£, 25£]

You are 95% confident that the population parameter lies between 20£ and 25£

You go through the entire population =>
all restaurants in London



100% confidence!

You visit 5% of the restaurants

Confidence interval: [20£, 25£]

5% chance that the population parameter is outside the range

$$1-\alpha$$

$$0 \leq \alpha \leq 1$$

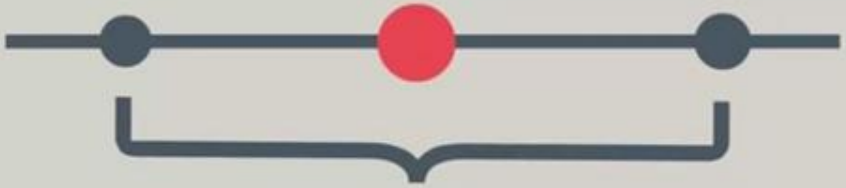
Confidence level

α	Confidence level
5%	95%
1%	99%

$$\left[\text{Point estimate} - \text{reliability factor} * \text{standard error}, \text{Point estimate} + \text{reliability factor} * \text{standard error}\right]$$

$$\left[\bar{x} - \text{reliability factor} * \text{standard error}, \bar{x} + \text{reliability factor} * \text{standard error} \right]$$

Point estimate



$$\left[\bar{x} - \text{reliability factor} * \frac{\sigma}{\sqrt{n}}, \bar{x} + \text{reliability factor} * \frac{\sigma}{\sqrt{n}} \right]$$

Confidence intervals

A confidence interval is the range within which you expect the population parameter to be.

Confidence intervals

Population variance known

Population variance unknown

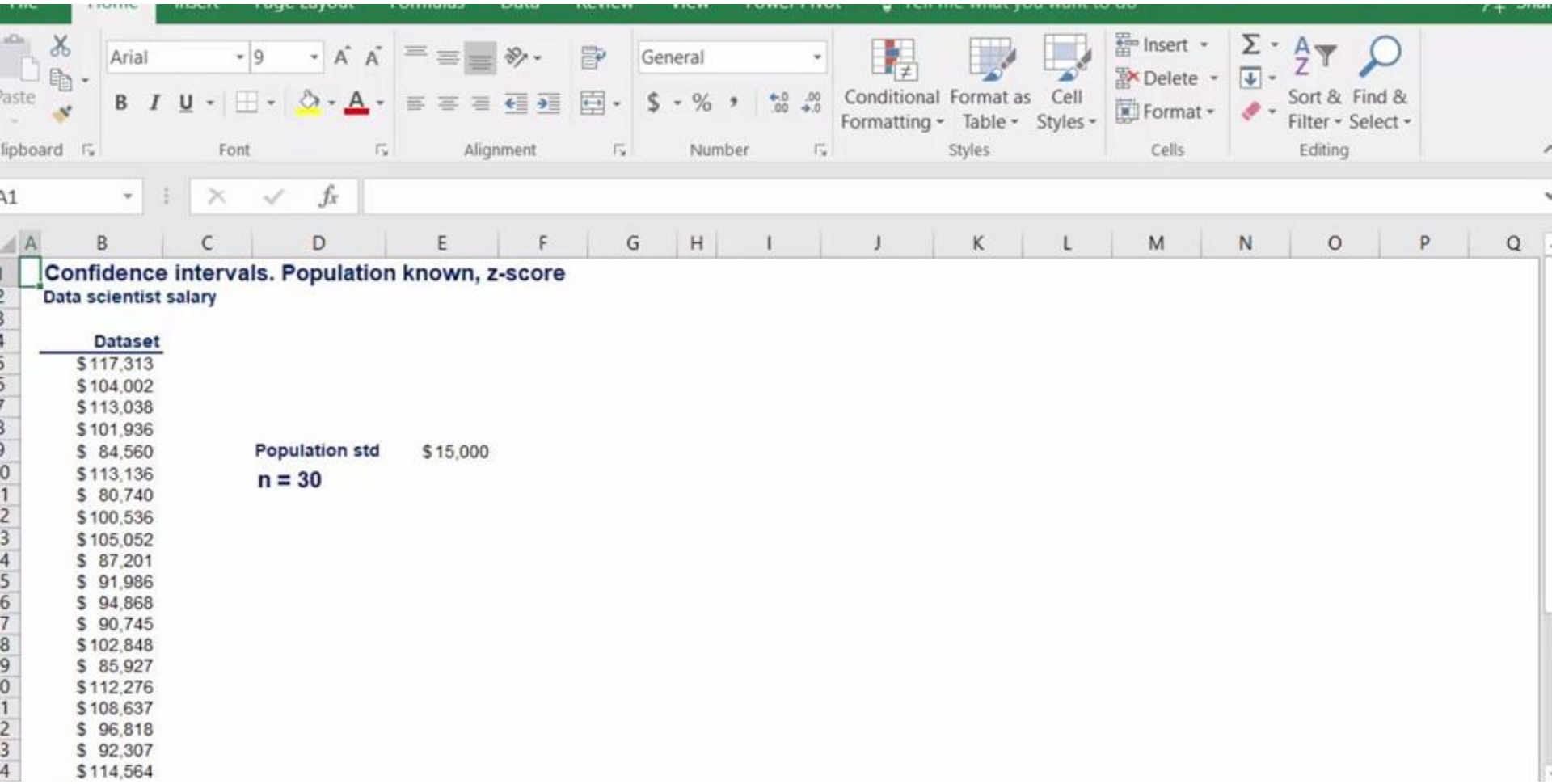
Confidence intervals

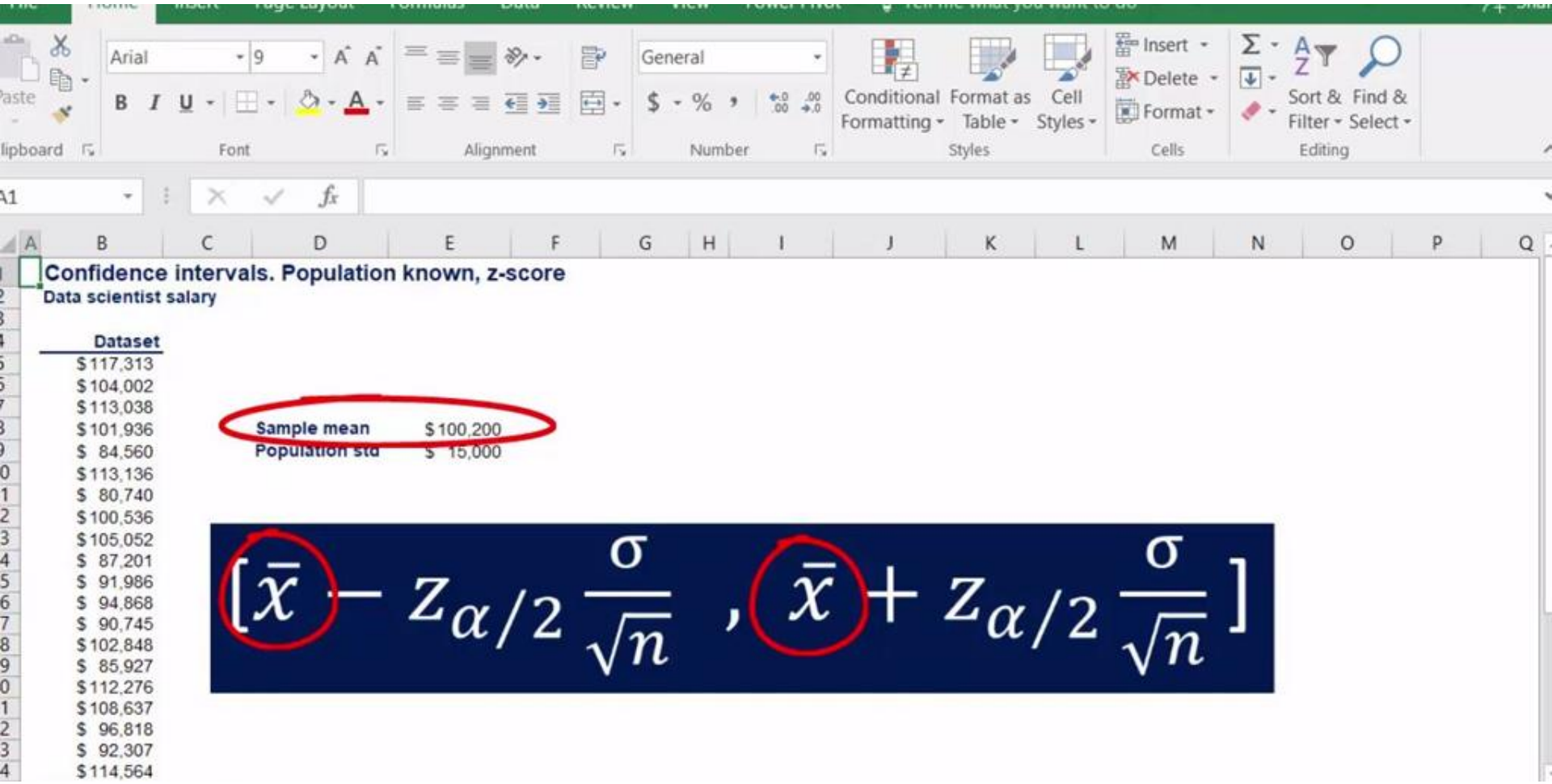
$$N \sim (\mu, \sigma^2)$$

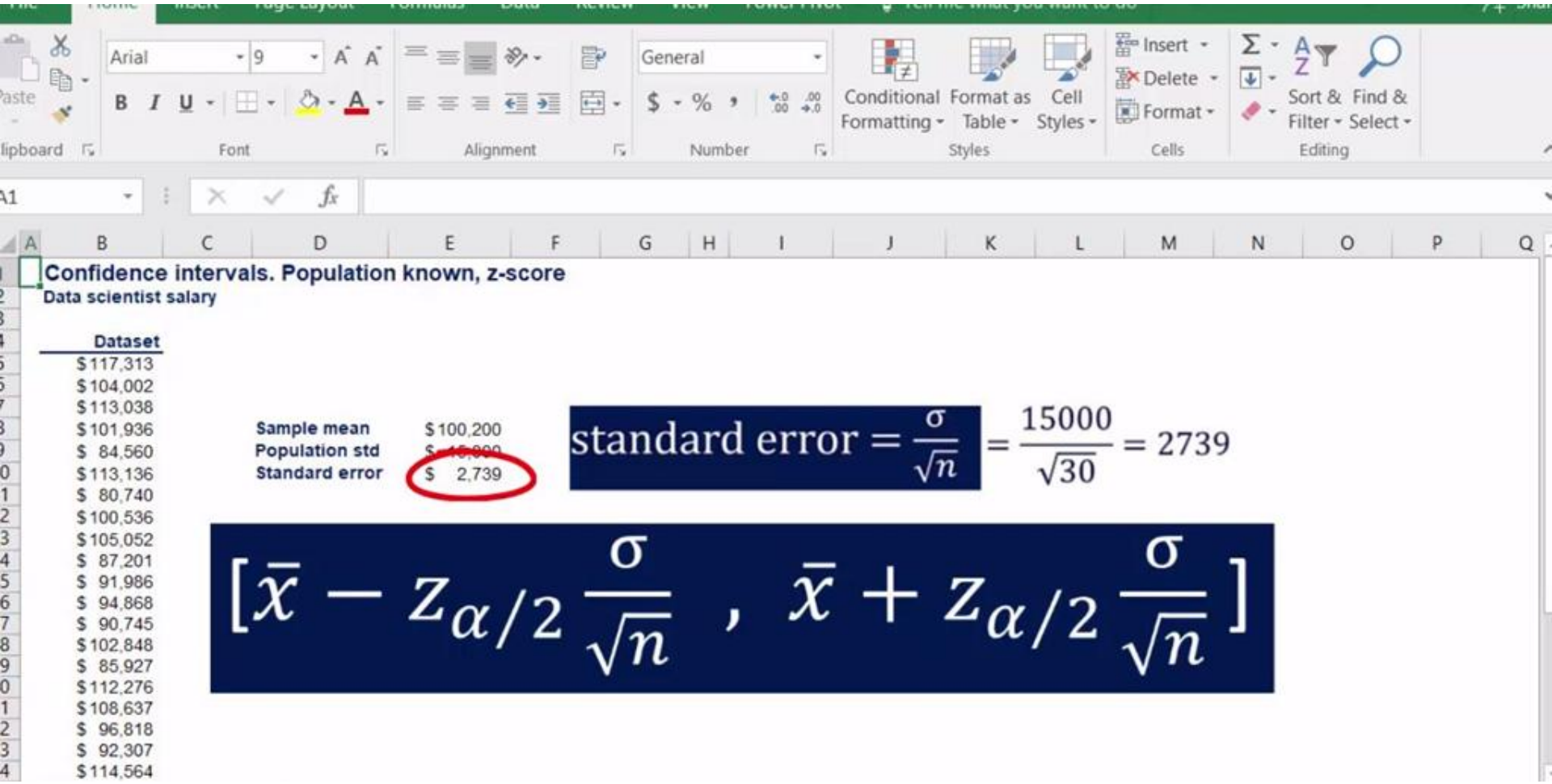
CLT

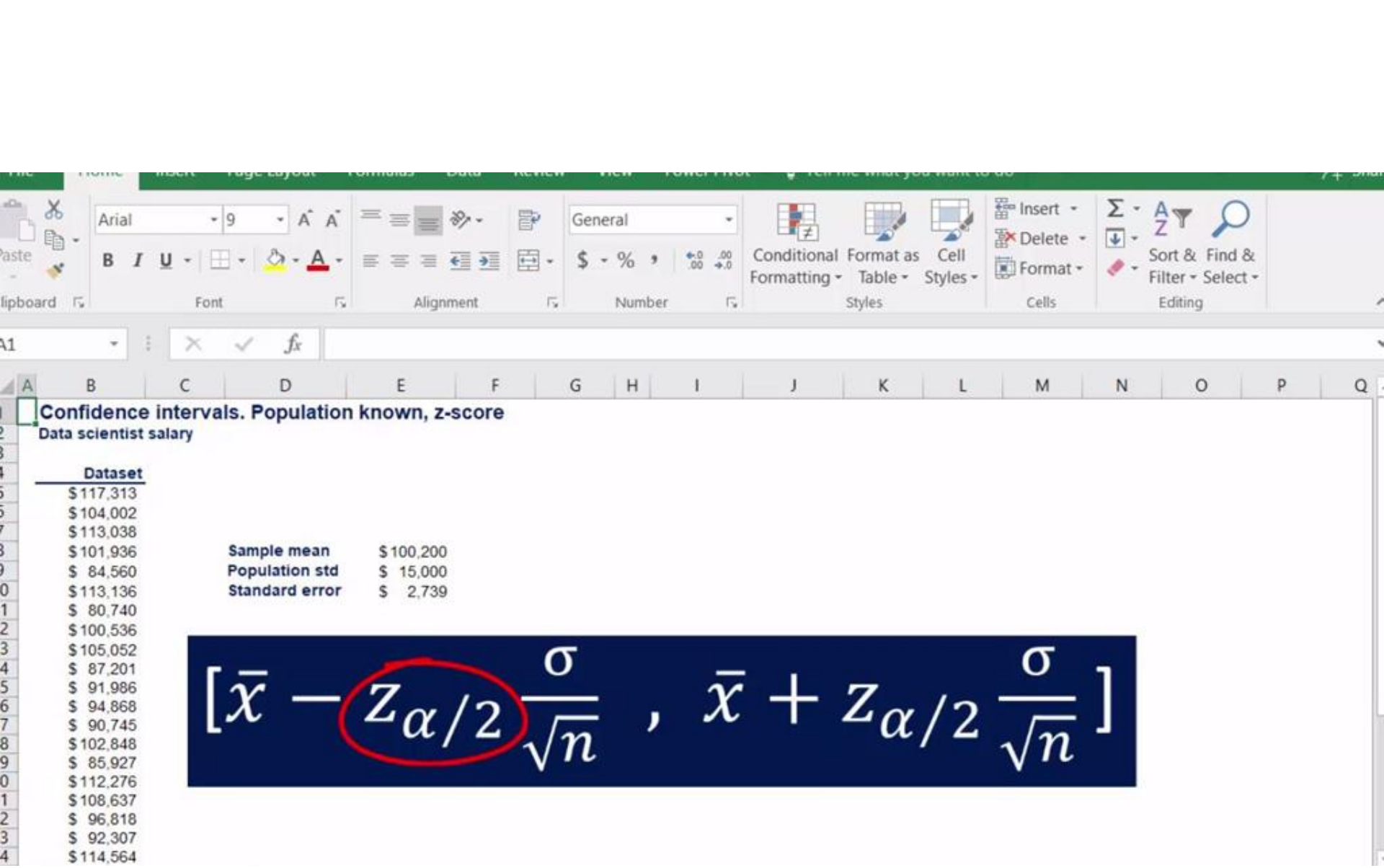
Population variance known

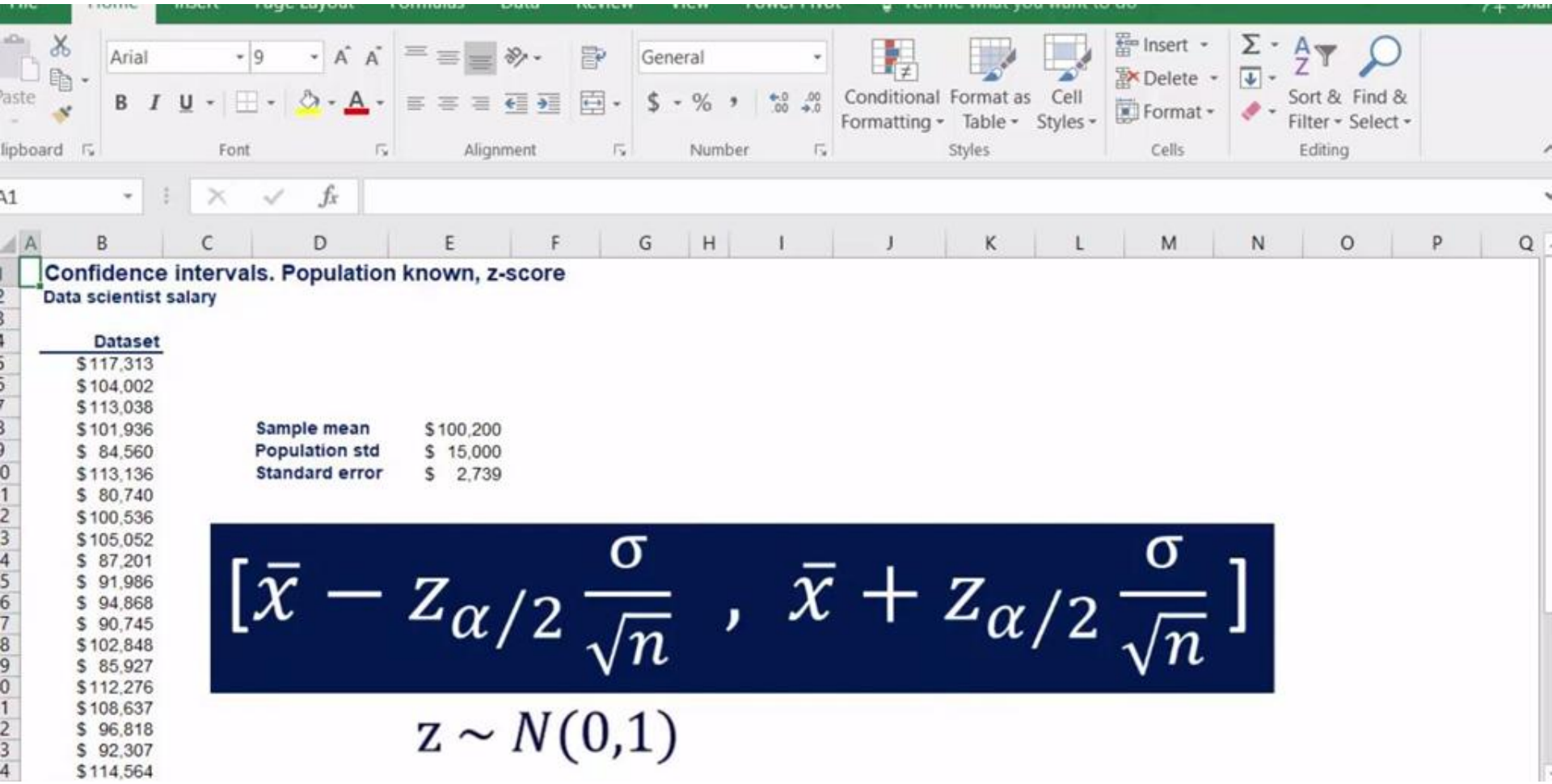
Example

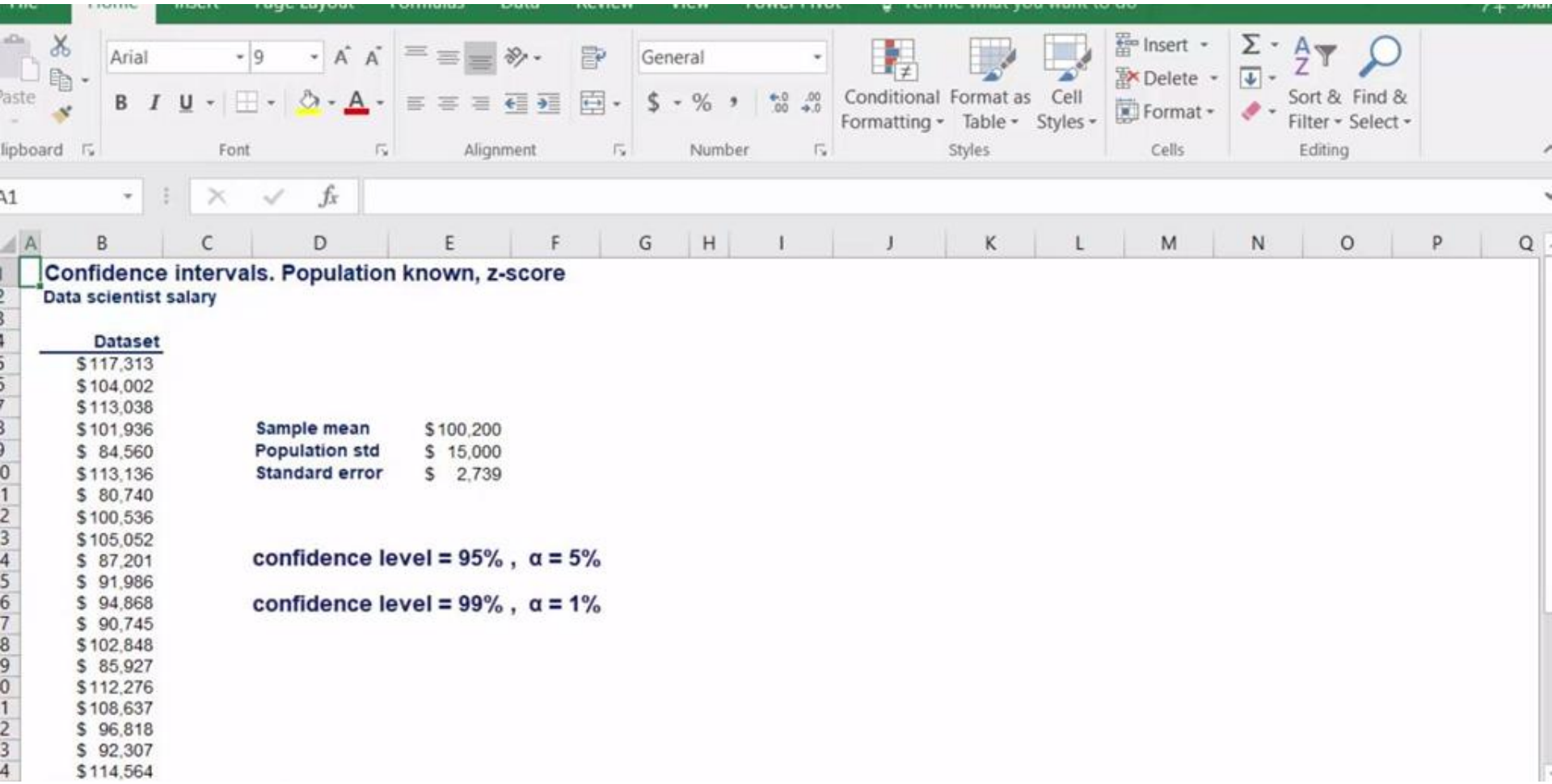

Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Population std
$15,000
n = 30

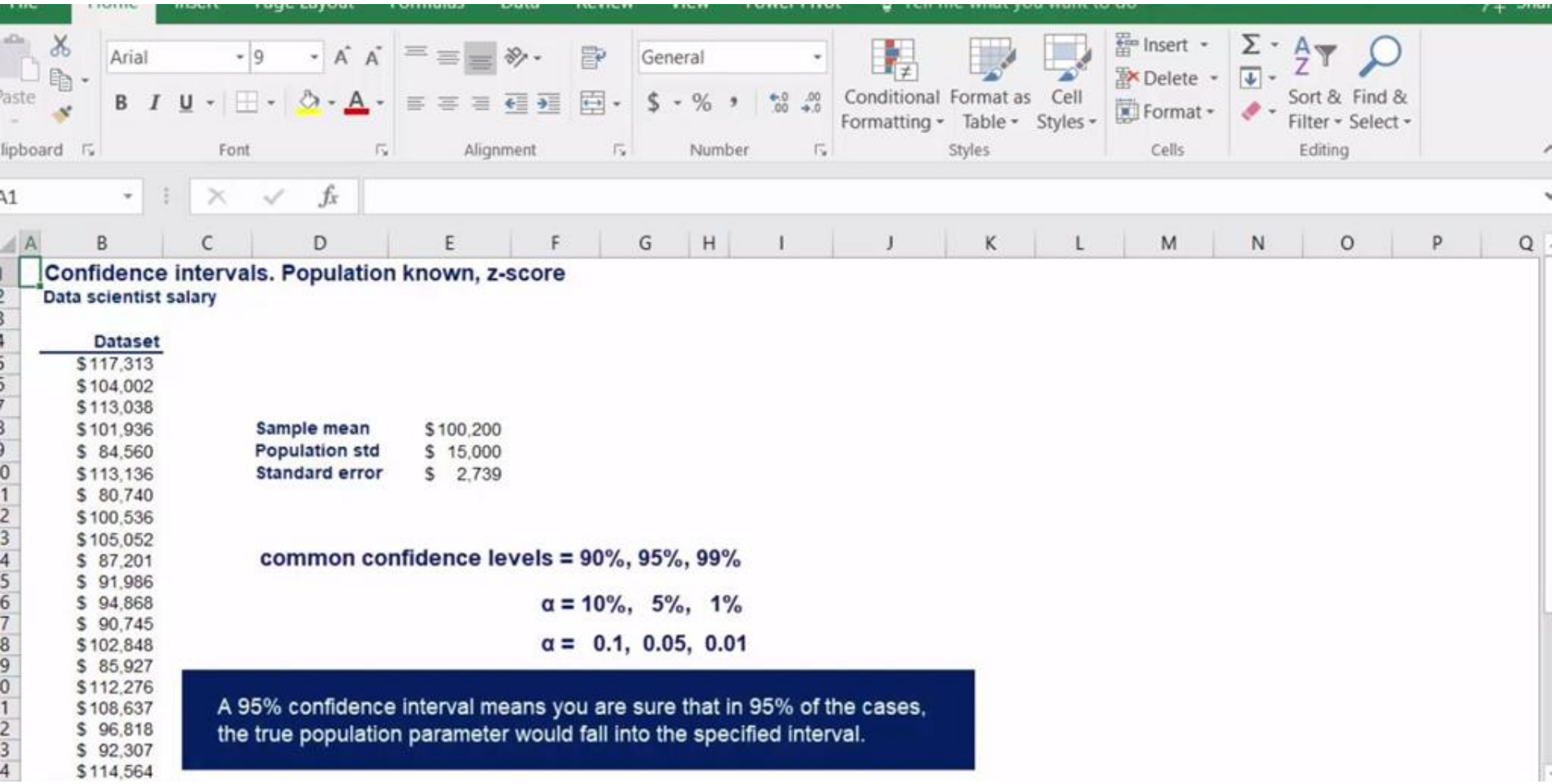

Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std
Standard error $ 2,739
standard error = $\frac{\sigma}{\sqrt{n}} = \frac{15000}{\sqrt{30}} = 2739$
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$
$$z \sim N(0,1)$$


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
confidence level = 95% , α = 5%
confidence level = 99% , α = 1%


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
common confidence levels = 90%, 95%, 99%
α = 10%, 5%, 1%
α = 0.1, 0.05, 0.01
A 95% confidence interval means you are sure that in 95% of the cases, the true population parameter would fall into the specified interval.

Standard normal distribution

z-table

The table summarizes the standard normal distribution critical values and the corresponding (1-α)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

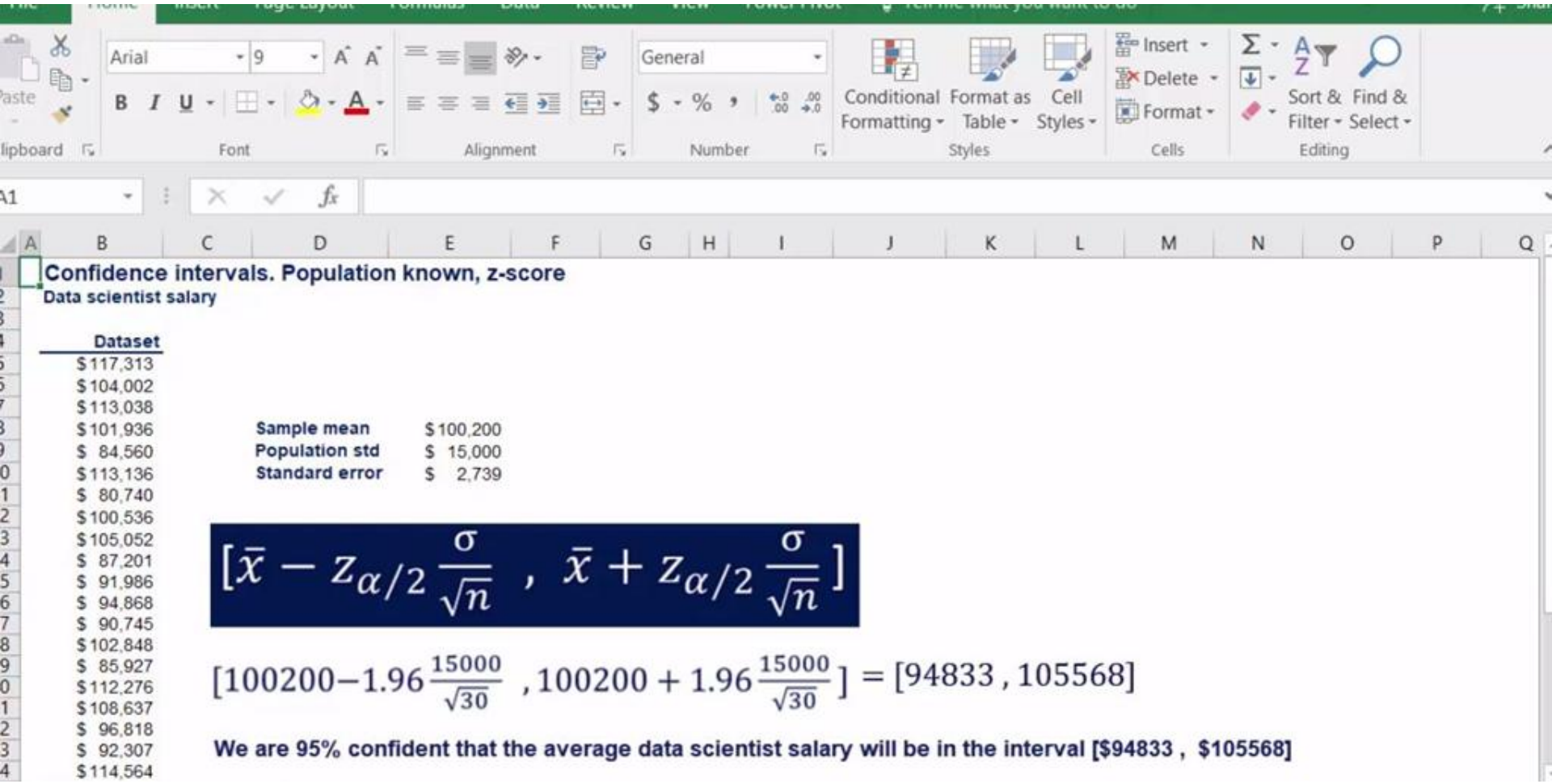
Confidence interval: 95%

α = 0.05

$z_{0.025}$

$1 - 0.025 = 0.975$

$z_{0.025} = 1.9 + 0.06 = 1.96$



Confidence intervals. Population known, z-score

Data scientist salary

Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564

Sample mean	$100,200
Population std	$ 15,000
Standard error	$ 2,739

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$\left[100200 - 1.96\frac{15000}{\sqrt{30}} \;, 100200 + 1.96\frac{15000}{\sqrt{30}}\right] = [94833\,,\,105568]$$

We are 95% confident that the average data scientist salary will be in the interval [$94833 , $105568]

Standard normal distribution

z-table

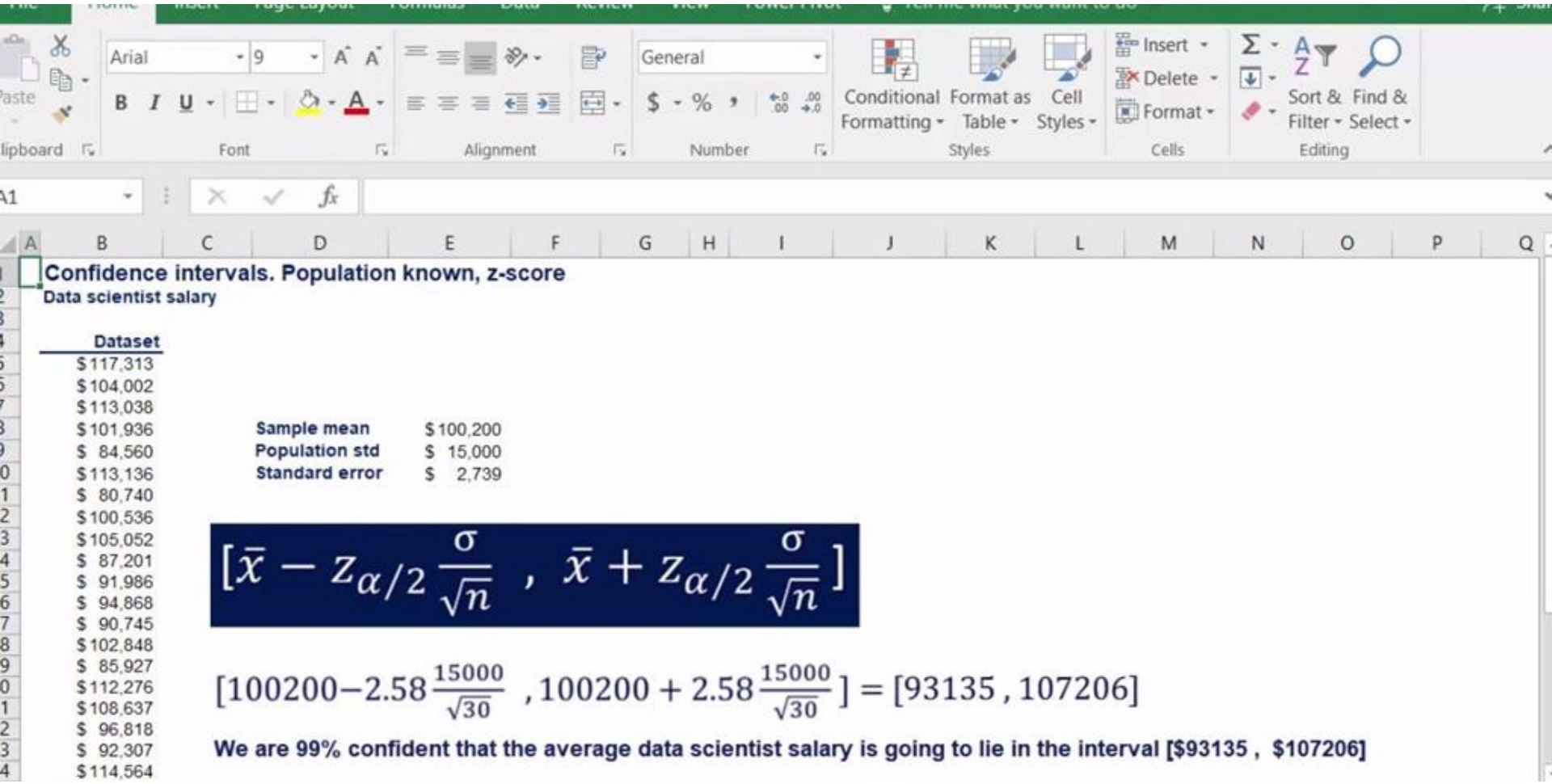
The table summarizes the standard normal distribution critical values and the corresponding (1-α)

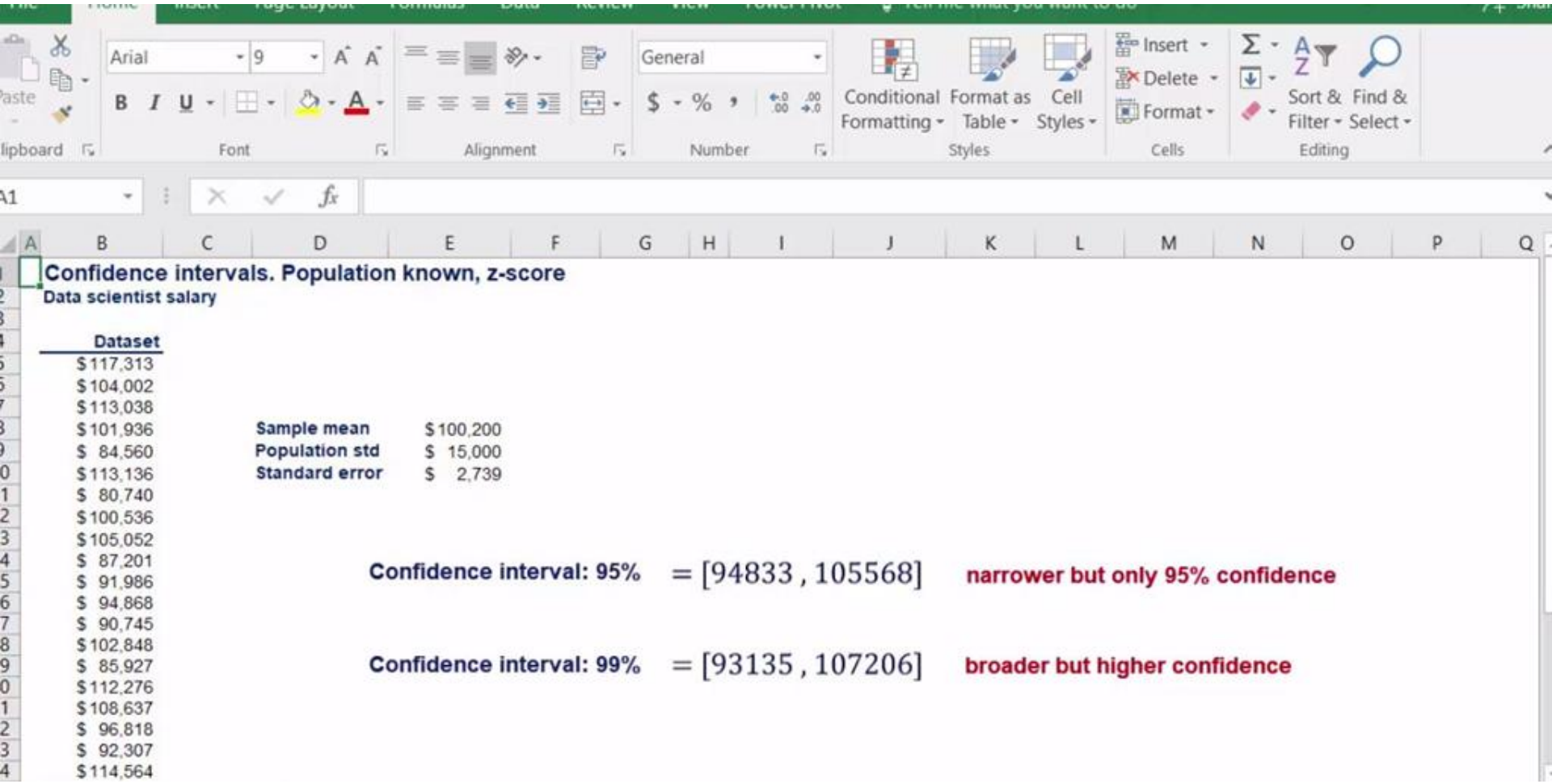
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

Confidence interval: 99%

α = 0.01

1-0.005=0.995


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$
$[100200-2.58\frac{15000}{\sqrt{30}} , 100200 + 2.58\frac{15000}{\sqrt{30}}] = [93135 , 107206]$
We are 99% confident that the average data scientist salary is going to lie in the interval [$93135 , $107206]


Arial
9
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Font
Alignment
Number
Styles
Cells
Editing
A1
Confidence intervals. Population known, z-score
Data scientist salary
Dataset
$117,313
$104,002
$113,038
$101,936
$ 84,560
$113,136
$ 80,740
$100,536
$105,052
$ 87,201
$ 91,986
$ 94,868
$ 90,745
$102,848
$ 85,927
$112,276
$108,637
$ 96,818
$ 92,307
$114,564
Sample mean $100,200
Population std $ 15,000
Standard error $ 2,739
Confidence interval: 95% = [94833 , 105568]
narrower but only 95% confidence
Confidence interval: 99% = [93135 , 107206]
broader but higher confidence

What is a Hypothesis?

- A hypothesis is an assumption which we make about a population parameter.
- The hypothesis which we wish to test is called the null hypothesis because it implies that there is no difference between the true value and the hypothesized value.

Cont..

- A thesis is something that has been proven to be true. However, a hypothesis is something that has not yet been proven to be true.
- Hypothesis testing is the process of determining whether or not a given hypothesis is true.
- Hypothesis testing along with estimation forms the foundation of inferential statistics.

The Null Hypothesis

- The first step in a hypothesis testing is to formalize it by specifying the null hypothesis.
- A Null hypothesis is an assertion about the value of a population parameter. It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.
- According to R.A. Fisher, "Hypothesis is tested for possible rejection under the assumption that it is true called null hypothesis. It is denoted by H0.

Cont..

- The alternate hypothesis is the negation of the null hypothesis. H1 symbol is used to denote alternate hypothesis.
- Because the null and alternate hypothesis assert exactly opposite statements, only one of them can be true. Rejecting one is equivalent to accepting the other.

Cont..

- BUT the question is how does one formulate hypothesis?. There is no hard-and-fast rule.
- Very often the phenomenon under study will suggest the nature of the null and alternate hypothesis.
- Theoretical expectation or prior empirical work or both can be relied upon to formulate hypotheses.

2. Evidence Collection

- After the null and alternate hypotheses are spelled out, the next step is to gather evidence.
- The best evidence is that where you are 100% confident. That rarely happens.
- In all cases, evidence is gathered from a random sample of the population.
- An important limitation of making inferences from sample data is that we cannot be 100% confident about it.

Type I and Type II error

- In our professional or personal lives, we often have to make an accept or reject type decision based on incomplete data.
- a) A quality control inspector has to accept or reject a batch of parts supplied by a vendor usually based on test results of a random sample.
- b) A recruiter has to accept or reject a job applicant usually based on evidence gathered from resume and interview.

Cont..

- As long as such decisions are made based on evidence that does not provide 100% confidence , there will be chances for error.
- No error is committed when a good prospect is accepted or a bad one is rejected. But there is small chance that a bad prospect is accepted or good one is rejected.
- A researcher has to minimize the chances of such errors.

The p-Value

- Suppose the null and alternate hypotheses are:
- H0: $\mu = 1000$
- H1: $\mu < 1000$
- A random sample of size 30 gives a sample mean of 995.
- Because the sample mean is less than 1000, the evidence goes against the null.
- Can we reject null based on this evidence?

Cont…

- Immediately we realize the dilemma:
- A) If we reject, there is some chance that we might be committing a type 1 error, and
- B) If we accept it, there is some chance that we might be committing a type II error.
- A natural question is to ask for the credibility of the null (H0)in light of unfavourable evidence.

Cont..

- We have to ask:

When actual μ=1000, and with sample size 30, what is the probability of getting a sample mean that is less than or equal to 995.

Suppose the answer is 20%. That is there is 20% chance for a sample of 30 to give a mean less than or equal to 995 when the actual μ=1000. Statistician call this 20% the p-value.

Cont…

- The p-value is a kind of ‘credibility rating’ of null (H0) in light of evidence.
- A credibility rating of 20% means that there is 20% probability that H0 is true, despite the evidence.
- Conversely, we can be roughly 80% confident that H0 is false in light of the evidence.

Cont..

- The implication is that if we reject H0, then there is 80% chance that we are doing the right thing and 20% chance that we are committing a type I error.

The significance Level

- The most common policy in statistical hypothesis is to establish a statistical significance denoted by α.
- When this policy is followed, one can be sure that the maximum probability of type 1 error is α.
- The standard values of α are 10%, 5%, and 1%.

Optimal α and Types of Error

- Relative costs of two types of error.
- In such cases where type II error is more costly, we keep a large value for α, namely, 10%.
- In such cases where type I error is more costly, we keep a small value for α, namely, 1%.

β and power

- The symbol used for the probability of type II error.
- β depends on the actual value of the parameter being tested, the sample size and α.
- If the actual value is 993 rather than 994, H0 would be 'even more wrong'. This should make it easier to detect that it is wrong.
- If the sample size increases, the evidence become more reliable and the possibility of error including β will decrease.

Sample Size and Errors

- What if both types of error are costly and we want to have low α and β.
- The only way to do this is to make our evidence more reliable, which can be done only by increasing the sample size.
- When the costs of both types of error are high the best policy is to have a large sample and a low α, such as 1%.

Two –tails Test

- Let us consider the following null hypothesis and alternate hypothesis:

$$H_0 : \mu = 0.20$$

$$H_1 : \mu \neq 0.20$$

- Thus, in the null hypothesis the value of μ is 0.20 which is a single hypothesis. However, the alternate hypothesis is a composite hypothesis.

- This implies that value of μ is either greater or less than 0.20. Thus, we are interested in both tails of the distribution. Hence, it is called two-tail test.

- Such two-tail alternate hypothesis is very often formulated when the researcher do not have a strong a priori theoretical idea to frame alternate hypothesis.

One –Tail Test

- One-tail test is resorted when we have strong a priori theoretical basis to suggest that the alternate hypothesis is unidirectional.

- For example, consider the following null and alternate hypotheses:

$$H_0 : \mu \leq 0.20$$

$$H_1 : \mu \succ 0.20$$

- In the above, alternate hypothesis tells us that the value of μ is necessarily greater than 0.20. Thus, in this case, the researcher is only interested in the right tail of the distribution.

- Since only one tail is the relevant while conducting this hypothesis testing, it is termed as one-tail test.

The Procedure of Hypothesis Testing

A sample random normal variable produces a sample mean of $\bar{x}$ =0.5022. Let's say the true population μ is 0.75. Now, we have to conduct hypothesis test to find whether there exist enough statistical evidence to claim that the true μ is 0.75.

The Procedure of Hypothesis Testing

A sample random normal variable produces a sample mean of $\bar{x}$ =0.5022. Let's say the true population μ is 0.75. Now, we have to conduct hypothesis test to find whether there exist enough statistical evidence to claim that the true μ is 0.75.

The procedure of hypothesis testing is as follows:

a) Frame the null and alternate hypotheses as:

$$H_0 : \mu = 0.75$$

$$H_1 : \mu \neq 0.75$$

The Procedure of Hypothesis Testing

A sample random normal variable produces a sample mean of $\bar{x}$ =0.5022. Let's say the true population μ is 0.75. Now, we have to conduct hypothesis test to find whether there exist enough statistical evidence to claim that the true μ is 0.75.

The procedure of hypothesis testing is as follows:

a) Frame the null and alternate hypotheses as:

$$H_0 : \mu = 0.75$$

$$H_1 : \mu \neq 0.75$$

b) Choose an appropriate significance level i.e. α . In general, decisions in social sciences are made at 10 percent, 5 percent and 1 percent level of significance. Lets take α =0.05 i.e. 5% to test this hypothesis.

The Procedure of Hypothesis Testing

A sample random normal variable produces a sample mean of $\bar{x}$ =0.5022. Let's say the true population μ is 0.75. Now, we have to conduct hypothesis test to find whether there exist enough statistical evidence to claim that the true μ is 0.75.

The procedure of hypothesis testing is as follows:

a) Frame the null and alternate hypotheses as:

$$H_0 : \mu = 0.75$$

$$H_1 : \mu \neq 0.75$$

b) Choose an appropriate significance level i.e. α . In general, decisions in social sciences are made at 10 percent, 5 percent and 1 percent level of significance. Lets take α =0.05 i.e. 5% to test this hypothesis.

c) Choose an appropriate test such as Z – Test or t –Test.

d) Compute the value of the Z test or t-test

The Procedure of Hypothesis Testing

A sample random normal variable produces a sample mean of $\bar{x}$ =0.5022. Let's say the true population μ is 0.75. Now, we have to conduct hypothesis test to find whether there exist enough statistical evidence to claim that the true μ is 0.75.

The procedure of hypothesis testing is as follows:

a) Frame the null and alternate hypotheses as:

$$H_0 : \mu = 0.75$$

$$H_1 : \mu \neq 0.75$$

b) Choose an appropriate significance level i.e. α . In general, decisions in social sciences are made at 10 percent, 5 percent and 1 percent level of significance. Lets take α =0.05 i.e. 5% to test this hypothesis.

c) Choose an appropriate test such as Z – Test or t –Test.

d) Compute the value of the Z test or t-test

e) Compare the calculated t value with table value of t i.e. critical t value

f) Take the decision. The rule is when the calculated t value is greater than critical t value, reject the null hypothesis.

Hypothesis Testing of a Population Mean: Large Sample

In case of single population, when the sample size is large ($n \geq 30$) and drawn from a normal population, hypothesis test about a single population mean is done by Z-test. Suppose a recent RBI study says that average real growth rate of the Indian economy will be 5.5 percent per annum. You believe that the figure is somewhat underestimated and decide to test this claim. For this purpose, a random sample of 36 economists taken in the survey, resulting in a sample growth rate of 6.1 percent per annum and standard deviation of 1.68. Test the hypothesis that the average growth rate is 5.5 percent per annum at 5 percent level of significance.

Cont..

The steps involved in the procedure of are the following:

a) Frame the null and alternate hypothesizes

Null hypothesis $H_0: \mu = 5.5$

Alternate hypothesis $H_1: \mu \neq 5.5$

b) Choose an appropriate significance level i.e. α . For α =0.05 i.e. 5% , the table value of Z = 1.96.

c) Choose an appropriate test. As we know that the sample size is more than 30. So in this case Z-test is the appropriate test.

d) Compute the value of the Z-test as follows:

$$z = \frac{sample \quad mean \quad - (population \quad mean \;)}{S\,\text{tan}\;dard \quad Error \quad of \quad mean} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{6.1 - 5.5}{\frac{1.68}{\sqrt{36}}} = 2.14$$

e) Compare the calculated Z value i.e. 2.14 with table value of Z i.e. critical Z value. Critical Z value is 1.96.

f) Take the decision. The rule is when the calculated Z value is greater than critical Z value, reject the null hypothesis. In our case, calculated Z value is 2.14 while the critical Z value is 1.96, so reject the null hypothesis.

Hypothesis Testing of Population Mean: Small Sample

The Centre for Science and Environment recently in a study stated that all brands of edible oil in India is not good for health due to high levels of trans fats in oils. The standard of 2 percent level of trans fats is only set by Denmark which is considered to be safe for our health. To test this, a random sample of 12 brands of refined oil were taken, resulting in mean level of 3.68 trans fats and standard deviation of 1.1 trans fats. Conduct a hypothesis test to conclude that refined oil in India is not safe for health at 1 percent level of significance.

a) Frame the null and alternate hypothesizes

Null hypothesis $H_0: \mu = 2$

Alternate hypothesis $H_1: \mu > 2$

b) Choose an appropriate significance level i.e. α . For α =0.01 i.e. 1% , the table value of t = 2.718.

c) Choose an appropriate test. As we know that the sample size is less than 30. So in this case t-test is the appropriate test.

d) Compute the value of the t-test as follows:

$$z = \frac{sample \quad mean \quad - (population \quad mean)}{S\,tan \quad dard \quad Error \quad of \quad mean} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{3.68 - 2.00}{\frac{1.1}{\sqrt{12}}} = 5.41$$

e) Compare the calculated t value i.e. 5.41 with table value of t i.e. critical t value. Critical Z value is 2.718

f) Take the decision. The rule is when the calculated t value is greater than critical t value, reject the null hypothesis. In our case, calculated t value is 5.41 while the critical t value is 2.718, so reject the null hypothesis.

We can infer with 99 percent confidence that the average level of trans fats in Indian edible oils is greater than 2 percent. Hence, they are not safe for eating.

Hypothesis Test of a Proportion

Sometimes, we are concerned about market share, customer mark up etc which are expressed in percentages. For example, one may be interested in testing what proportions of population prefer voting versus those who are not voting.

A recent survey of 300 people in Gurgaon suggests that 24 percent people are using credit card of some bank. The Oriental Bank of Commerce management wants to find out whether the true percentage of credit card holders is less than 20 before targeting the customer with new package. Test the hypothesis at 5 percent significance level. This is clearly a one-tailed test.

The steps involved for the hypothesis test of proportion are as follows:

a) Frame the null and alternate hypothesizes

Null hypothesis $H_0: \mu \geq 0.20$

Alternate hypothesis $H_1: \mu < 0.20$

b) Choose an appropriate significance level i.e. α . For α =0.05 i.e. 5% , the table value of Z for 0.05 in case of one tail is 1.65.

c) Choose an appropriate test. In this case Z-test is the appropriate test.

d) Compute the value of the Z-test as follows:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z = \frac{0.24 - 0.20}{\sqrt{\frac{0.20\,(1-0.20)}{300}}} = 1.73$$

e) Compare the calculated Z value i.e. 1.73 with table value of Z i.e. critical Z value. Critical Z value is 1.65

f) Take the decision. The rule is when the calculated Z value is greater than critical Z value, reject the null hypothesis. In our case, calculated Z value is 1.73 while the critical Z value is 1.65, so reject the null hypothesis.

The bank can infer with 95 percent confidence that the less than 20 percent people are using credit card and can target people for credit card with new package.

Hypothesis Test of Two Population Mean Assuming Equal variance: Independent Samples

In many situations, you may be interested in comparing means of two different populations. For example, a business analyst may compare stock market mean returns in 2000 with those of 2015 to find out whether any change in mean return occurred over time, the high income customers spends more on junk food than low-income customer, smokers are more prone to lung cancer than non-smokers, mean salary of males are higher than mean salary of females.

When samples are drawn randomly from different population they are termed as independent samples because the units or people sampled under each group are in no way linked to units of other group.

Cont..

When comparing means for two independent samples, the hypotheses may take the following form:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

When variance of both the populations are same, a pooled variance is estimated using both the samples variances as follows:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The standard error of the above statistic is given as follows:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Finally, to test hypothesis about two population means, the appropriate test is given as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$\bar{x}_1 - \bar{x}_2$ = Difference in two samples mean

$\mu_1 - \mu_2$ = Difference in two population mean

S^2 = pooled Variance

n_1 = Sample size 1

n_2 = sample size 2

Cont..

Let us consider Business Statistics subject is taken by two faculty members in two different courses. The mean mark of 15 randomly selected students is 68 and variance is 25 in course Section A. The mean of 18 randomly selected students is 76 with variance of 16 in section B. Test the hypothesis that mean marks in both the courses are equal assuming equal variance at 5 percent significance level.

a) In this case our null and alternate hypothesis are:

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

b) Choose a level of significance. The critical value for 33 degrees of freedom at 0.05 significance level is 1.69.

c) Choose an appropriate test. In this case t test is the appropriate test.

d) Compute the value of the t test as follows

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t = \frac{(8) - (0)}{\sqrt{20.06\left(\frac{1}{15} + \frac{1}{18}\right)}} = 5.19$$

e) Take the decision. Since the computed value is greater than critical t value, we will reject the null hypothesis in favour of alternate hypothesis implying the mean marks of two courses are not equal. In other words, the difference in marks of two courses are statistically significant.

	A	B	C	D	E	F	G	H	I	J	K	L
2	Sample 1	Sample 2			t-Test: Two-Sample Assuming Equal Variances							
3	12	20										
4	8	17				*Sample 1*	*Sample 2*					
5	14	17			Mean	12.64285714	14.375					
6	8	13			Variance	15.01648352	17.05					
7	9	17			Observations	14	16					
8	10	18			Pooled Variance	16.10586735						
9	11	15			Hypothesized Mean Difference	0						
10	15	14			df	28						
11	18	10			t Stat	-1.179383453						
12	13	16			P(T<=t) one-tail	0.12408495						
13	16	20			t Critical one-tail	1.701130908						
14	20	18			P(T<=t) two-tail	0.2481699						
15	8	9			t Critical two-tail	2.048407115						
16	15	8										
17		16										
18		8										
19												
20												

Hypothesis Test of Population Mean: Two Independent Samples ($\sigma_1^2 \neq \sigma_2^2$)

When the population variances are not equal, one cannot use the pooled variance estimate as discussed in earlier section. In this case, population variance is estimated by its sample variance. It is important to note that the sampling distribution of the following resulting statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

do not follow either normal distribution or t-distribution. However, in practical it is approximated by t-distribution with degrees of freedom given by the following expression:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

which is often rounded to the nearest integer.

Let us consider the following statistics relating to random samples from two normally distributed population,

$\bar{x}_1 = 210 \qquad s_1^2 = 49 \qquad n_1 = 18$

$\bar{x}_2 = 198 \qquad s_2^2 = 16 \qquad n_2 = 4$

Test the hypothesis that the population mean is not equal at 1 percent significance level.

a) In this case our null and alternate hypothesis are:

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

b) Choose a level of significance. The degrees of freedom is calculated as follows:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 - 1} + \frac{(s_2^2 / n_2)^2}{n_2 - 1}}$$

$$df = \frac{\left(\frac{49}{18} + \frac{16}{44}\right)^2}{\frac{(49 / 18)^2}{17} + \frac{(16 / 44)^2}{43}} = 21.60 \ (\textit{rounded to } 22)$$

For 22 degrees of freedom, the critical t value at 1 percent significance level is 2.508.

c) Choose an appropriate test. In this case t test is the appropriate test.

d) Compute the value of the t test as follows

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$t = \frac{(12) - (0)}{\sqrt{\left(\frac{49}{18} + \frac{16}{44}\right)}} = 6.85$$

e) Take the decision. Since the computed value is greater than critical t value, we will reject the null hypothesis in favour of alternate hypothesis implying the mean of two populations are not equal. In other words, the differences in mean values of two populations are statistically significant.

Analysis of Variance (ANOVA)

Introduction

Analysis of variance is an important statistical technique used to test the hypothesis that the means of two or more populations are equal.

In case of more than two means, one can also use t-test for comparing means but the chances of type I error increases.

In order to avoid this situation, in case of more than two population means, the appropriate test statistic for testing equality of more than two means is **analysis of variance**.

R. Fisher, the father of statistics, developed a technique called 'experimental design' to establish cause and effect relationship between variables. In fact, ANOVA is an important part of a large 'experimental design' setup.

In ANOVA, we have a dependent variable which is quantitative in nature and one or more independent variables which are categorical in nature.

The independent variables which are categorical variables are also called **factors**. Combination of factors or categories is called **treatment**.

When there is a single independent variable or a single factor, it is called **one-way ANOVA.** If there are two or more factors it is termed as **n-way ANOVA.**

One-Way ANOVA

In one-way ANOVA, we have one dependent variable and one categorical independent variable.

The idea is to find how much variation in dependent variable is explained by categorical independent variable and how much variation is not accounted by this independent variable.

In fact, we will try to decompose total variation in dependent variable (Y) into variation explained by categorical independent variable (X) and variation not explained by X, that is, error. SS_Y is the total variation in Y.

SS_X is the variation in Y that is due to the variations in the means of groups of X. SS_{Error} is the variation in Y that is linked with variation within each category of X.

The total variation in dependent(Y), denoted by SS_Y, is decomposed into:

$$SS_Y = SS_X + SS_{Error}$$

where

$$SS_Y = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

$$SS_X = \sum_{j=1}^{c} n(\bar{Y}_j - \bar{Y})^2$$

$$SS_{Error} = \sum_{j}^{c} \sum_{i}^{n} (Y_{ij} - \bar{Y}_j)^2$$

Y_i = individual observation

$\bar{Y}_j$ = average for category j

$\bar{Y}$ = Grand mean

Y_{ij} = ith observation in jth category

In analysis of variance, the aim is to test the null hypothesis that the means of two or more population are equal. In other words, our null and and the alternate hypotheses are:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$$
$$H_1 : at\ least\ one\ mean\ is\ different$$

The above hypothesis is tested by the F statistic with (c-1) and (N-c) degrees of freedom in the numerator and denominator respectively. The F statistic is given by the following formula:

$$F = \frac{\frac{SS_X}{(c-1)}}{\frac{SS_{Error}}{(N-c)}}$$

The rule is when the calculated value of F is greater than crtical F value reject the null hypothesis.

The ICICI Bank has three branches in New Delhi, and the management wants to find out whether there is any difference in the average business (Rs. Crores) of the three branches. The following table gives data relating to 8 randomly selected months' business at each branch.

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18
24	27	25
20	32	30
16	34	28
22	31	27
26	29	22

Test the hypothesis that the average businesses of three branches are equal at 5 percent significance level.

Solution

In this case our null and alternate hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : at\ least\ one\ mean\ is\ different$

The above hypothesis will be tested by the F statistic. First we will compute category mean and Grand mean and then various sums of squares will be computed.

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : at\ least\ one\ mean\ is\ different$

The above hypothesis will be tested by the F statistic. First we will compute category mean and Grand mean and then various sums of squares will be computed.

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18
24	27	25
20	32	30
16	34	28
22	31	27
26	29	22
$\bar{X}_1 = 20.25$	$\bar{X}_2 = 29.37$	$\bar{X}_3 = 25$
Grand Mean	$\frac{20.25 + 29.37 + 25}{3}$	

The above hypothesis will be tested by the F statistic. First we will compute category mean and Grand mean and then various sums of squares will be computed.

Branch 1	Branch 2	Branch 3
15	30	26
18	28	24
21	24	18
24	27	25
20	32	30
16	34	28
22	31	27
26	29	22
$\bar{X}_1 = 20.25$	$\bar{X}_2 = 29.37$	$\bar{X}_3 = 25$
Grand Mean	$\frac{20.25 + 29.37 + 25}{3}$ =24.87	

Now

$SS_Y = (15-24.87)^2 + (18-24.87)^2 + (21-24.87)^2 + (24-24.87)^2 + (20-24.87)^2 + (16-24.87)^2 + (22-24.87)^2 + (26-24.87)^2 + (30-24.87)^2 + (28-24.87)^2 + (24-24.87)^2 + (27-24.87)^2 + (32-24.87)^2 + (34-24.87)^2 + (31-24.87)^2 + (29-24.87)^2 + (26-24.87)^2 + (24-24.87)^2 + (18-24.87)^2 + (25-24.87)^2 + (30-24.87)^2 + (28-24.87)^2 + (27-24.87)^2 + (22-24.87)^2$

$= 600.26$

$SS_X = 8(20.25-24.87)^2 + 8(29.37-24.87)^2 + 8(25-24.87)^2$

$= 332.89$

$SS_{Error} = (15-20.25)^2 + (18-20.25)^2 + (21-20.25)^2 + (24-20.25)^2 + (20-20.25)^2 + (16-20.25)^2 + (22-20.25)^2 + (26-20.25)^2 + (30-29.37)^2 + (28-29.37)^2 + (24-29.37)^2 + (27-29.37)^2 + (32-29.37)^2 + (34-29.37)^2 + (31-29.37)^2 + (29-29.37)^2 + (26-25)^2 + (24-25)^2 + (18-25)^2 + (25-25)^2 + (30-25)^2 + (28-25)^2 + (27-25)^2 + (22-25)^2$

$= 267.37$

$= 600.26$

$SS_X = 8(20.25-24.87)^2 + 8(29.37-24.87)^2 + 8(25-24.87)^2$

$= 332.89$

$SS_{Error} = (15-20.25)^2 + (18-20.25)^2 + (21-20.25)^2 + (24-20.25)^2 + (20-20.25)^2 + (16-20.25)^2 + (22-20.25)^2 + (26-20.25)^2 + (30-29.37)^2 + (28-29.37)^2 + (24-29.37)^2 + (27-29.37)^2 + (32-29.37)^2 + (34-29.37)^2 + (31-29.37)^2 + (29-29.37)^2 + (26-25)^2 + (24-25)^2 + (18-25)^2 + (25-25)^2 + (30-25)^2 + (28-25)^2 + (27-25)^2 + (22-25)^2$

$= 267.37$

It can be verified that

$SS_Y = SS_X + SS_{Error}$

$600.26 = 332.89 + 267.37$

The above null hypothesis can now be tested as follows:

$$F = \frac{\frac{SS_X}{(c-1)}}{\frac{SS_{Error}}{(N-c)}} = \frac{\frac{332.89}{(3-1)}}{\frac{267.37}{(24-3)}} = 13.07$$

Thus, the calculated F value is 13.07. Now we have to compare this calculated value with critical F value. The critical F value for 2 degrees of freedom in numerator and 21 degrees of freedom in denominator is 3.47 for $\alpha = 0.05$. Since the calculated value F is greater than the critical F value, we will reject the null hypothesis. This implies that average businesses of three branches are not equal.

	A	B	C
2	15	30	26
3	18	28	24
4	21	24	18
5	24	27	25
6	20	32	30
7	16	34	28
8	22	31	27
9	26	29	22

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Branch 1	8	162	**20.25**	14.5
Branch 2	8	235	**29.375**	9.696428571
Branch 3	8	200	**25**	14

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	333.25	2	166.625	**13.08695652**	0.000204	3.466800112
Within Groups	267.375	21	12.73214286			
Total	600.625	23				

N-Way ANOVA

We can extend the concept of one way ANOVA to study the effect of more than one factor.

For example, how do age of the readers (less than 20, 20-50, more than 50) and educational levels (higher secondary, under graduate, post-graduate, M.phil/Ph.D) affect the circulation of a particular newspaper?

Similarly, if we want to study the effect of students' familiarity with a university (very high, high, medium, low, very low) and image of the university (highly positive, positive, neutral, negative, highly negative) on the preference for the university, n-way anova can be used to determine such effects.

This also helps the researchers to find the interactions between the factors. Let us consider two factors, namely, X_1 and X_2 with categories c_1 and c_2. In this case, total variation is decomposed into:

$$SS_{total} = SS_{X_1} + SS_{X_2} + SS_{X_1X_2} + SS_{error}$$

where

SS_{total} = Total variation

$$SS_{total} = SS_{X_1} + SS_{X_2} + SS_{X_1X_2} + SS_{error}$$

where

SS_{total} = Total variation

SS_{X1} = variance explained by X_1

SS_{X2} = variance explained by X_2

SS_{X1X2} = variance jointly explained by X_1 and X_2

One can test the significance of the overall effect by F test given below:

$$F = \frac{(SS_{X_1} + SS_{X_2} + SS_{X_1X_2}) / df_1}{SS_{error} / df_2}$$

where

df_1 = degrees of freedom in the numerator = c_1c_2-1
df_2 = degrees of freedom in the denominator = N-c_1c_2

f

X_1 is tested as follows:

$$F = \frac{\frac{(SS_{X_1})}{(c_1 - 1)}}{\frac{SS_{error}}{(N - c_1 c_2)}}$$

Similarly, you can test the significance of X_2 as follows:

$$F = \frac{\frac{(SS_{X_2})}{(c_2 - 1)}}{\frac{SS_{error}}{(N - c_1 c_2)}}$$

If you are interested in finding the significance of interaction effect, you can test the null of no interaction effect by the following statistic

$$F = \frac{\frac{(SS_{X_1X_2})}{(c_1 - 1)(c_2 - 1)}}{\frac{SS_{error}}{(N - c_1c_2)}}$$

The decision rule is same. When the calculated F value is greater than critical F value reject the null hypothesis.

Thapar University is an ISO certified university which conducts a student survey every year to assess the satisfaction levels of its students.

The students were asked to rate the university on a sacle of 1 to 7 (7 representing excellently) on various attributes of quality.

One of the questions of the survey was overall, how well do you think that the Thapar University has prepared you for a bright career in the corporate sector.

The following data give responses of the students to this question. The students were divided by the regional centers and type of courses offered.

	Centres			
		New Delhi	Hyderabad	Mumbai
Course Type	MBA(Finance)	4 2 6 5	6 2 3 5	3 5 2 6
	MBA(Marketing)	4 4 5 6	6 4 5 6	2 3 3 2

The following data give responses of the students to this question. The students were divided by the regional centers and type of courses offered.

	Centres			
Course Type		New Delhi	Hyderabad	Mumbai
	MBA(Finance)	4 2 6 5	6 2 3 5	3 5 2 6
	MBA(Marketing)	4 4 5 6	6 4 5 6	2 3 3 2

Determine whether there are significant differences in the responses using two-way ANOVA to this question at 5 percent significance level.

Solution

We will test the following hypotheses concerning to two-way ANOVA.
Our null and alternate hypotheses for row effects are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 : at\ least\ one\ mean\ is\ different$

Our null and alternate hypotheses for column effects are:

$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : at\ least\ one\ mean\ is\ different$

For interaction effects:

H_0: There is no interaction effect

	New Delhi	Hyderabad	Mumbai
MBA(Finance)	4	6	3
	2	2	5
	6	3	2
	5	5	6
MBA(Marketing)	4	6	2
	4	4	3
	5	5	3
	6	6	2

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
MBA(Finance)	3	13	4.333333	2.333333
	3	9	3	3
	3	11	3.666667	4.333333
	3	16	5.333333	0.333333
MBA(Marketing)	3	12	4	4
	3	11	3.666667	0.333333
	3	13	4.333333	1.333333
	3	14	4.666667	5.333333
New Delhi	8	36	4.5	1.714286
Hyderabad	8	37	4.625	2.267857
Mumbai	8	26	3.25	2.214286

5	5	3
6	6	2

MBA(Marketing)	3	12	4	4	
	3	11	3.666667	0.333333	
	3	13	4.333333	1.333333	
	3	14	4.666667	5.333333	
New Delhi	8	36	4.5	1.714286	
Hyderabad	8	37	4.625	2.267857	
Mumbai	8	26	3.25	2.214286	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	10.625	7	1.517857	0.648855	0.710146	2.764199
Columns	9.25	2	4.625	**1.977099**	0.175281	3.738892
Error	32.75	14	2.339286			
Total	52.625	23				

Thapar University is an ISO certified university which conducts a student survey every year to assess the satisfaction levels of its students.

The students were asked to rate the university on a scale of 1 to 7 (7 representing excellently) on various attributes of quality.

One of the questions of the survey was overall, how well do you think that the Thapar University has prepared you for a bright career in the corporate sector.

The following data give responses of the students to this question. The students were divided by the regional centers and type of courses offered.

		Centres		
		New Delhi	Hyderabad	Mumbai
Course Type	MBA(Finance)	4 2 6 5	6 2 3 5	3 5 2 6
	MBA(Marketing)	4 4 5 6	6 4 5 6	2 3 3 2

Determine whether there are significant differences in the responses using two-way ANOVA to this question at 5 percent significance level.

We will test the following hypotheses concerning to two-way ANOVA.

Our null and alternate hypotheses for row effects are:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_1 : at\ least\ one\ mean\ is\ different$

Our null and alternate hypotheses for column effects are:

$H_0 : \mu_1 = \mu_2 = \mu_3$
$H_1 : at\ least\ one\ mean\ is\ different$

For interaction effects:
H_0: There is no interaction effect
H_1: There is interaction effect

O7

	A	B	C	D
1		New Delhi	Hyderabad	Mumbai
2	MBA(Finance)	4	6	3
3		2	2	5
4		6	3	2
5		5	5	6
6	MBA(Marketing)	4	6	2
7		4	4	3
8		5	5	3
9		6	6	2

Anova: Two-Factor With Replication

SUMMARY	New Delhi	Hyderabad	Mumbai	Total
MBA(Finance)				
Count	4	4	4	12
Sum	17	16	16	49
Average	4.25	4	4	4.083333333
Variance	2.916666667	3.333333333	3.333333333	2.628787879
MBA(Marketing)				
Count	4	4	4	12
Sum	19	21	10	50
Average	4.75	5.25	2.5	4.166666667
Variance	0.916666667	0.916666667	0.333333333	2.151515152
Total				
Count	8	8	8	

Sheet1 Sheet2 Sheet3

O7

	A	B	C	D
3		2	2	5
4		6	3	2
5		5	5	6
6	**MBA(Marketing)**	4	6	2
7		4	4	3
8		5	5	3
9		6	6	2

	G	H	I	J	K
3	SUMMARY	New Delhi	Hyderabad	Mumbai	Total
4	*MBA(Finance)*				
5	Count	4	4	4	12
6	Sum	17	16	16	49
7	Average	4.25	4	4	4.083333333
8	Variance	2.916666667	3.333333333	3.333333333	2.628787879
9					
10	*MBA(Marketing)*				
11	Count	4	4	4	12
12	Sum	19	21	10	50
13	Average	4.75	5.25	2.5	4.166666667
14	Variance	0.916666667	0.916666667	0.333333333	2.151515152
15					
16	*Total*				
17	Count	8	8	8	
18	Sum	36	37	26	
19	Average	4.5	4.625	3.25	
20	Variance	1.714285714	2.267857143	2.214285714	

Sheet1 Sheet2 Sheet3

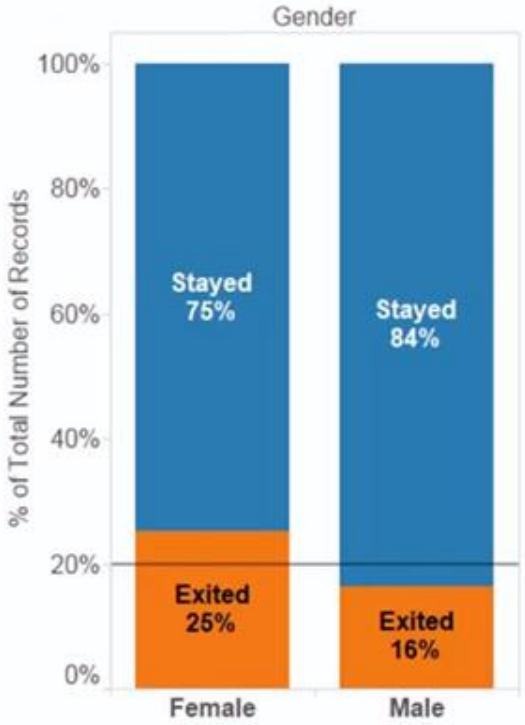
M27 3.55455714571373

				Total				
				Count	8	8	8	
				Sum	36	37	26	
				Average	4.5	4.625	3.25	
				Variance	1.714285714	2.267857143	2.214285714	
				ANOVA				
				Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
				Sample	0.041666667	1	0.041666667	**0.021276596**	0.885649	4.413873
				Columns	9.25	2	4.625	2.361702128	0.122798	3.554557
				Interaction	8.083333333	2	4.041666667	2.063829787	0.155967	3.554557
				Within	35.25	18	1.958333333			
				Total	52.625	23				

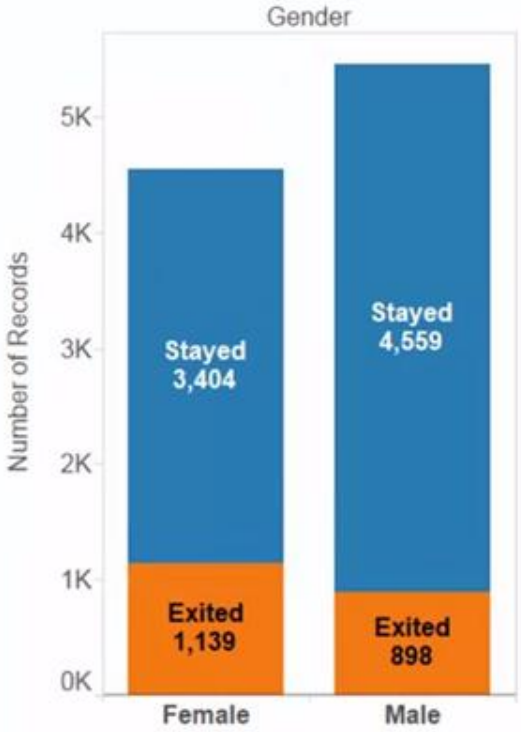
Sheet1 Sheet2 Sheet3

Chi-Squared

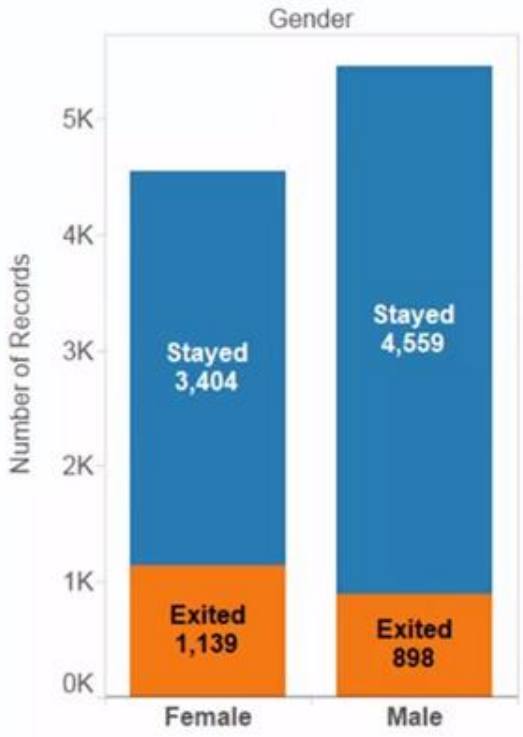
Chi-Squared


Gender
100%
80%
60%
40%
20%
0%
% of Total Number of Records
Stayed
75%
Exited
25%
Stayed
84%
Exited
16%
Female
Male

Chi-Squared


Gender
Number of Records
5K
4K
3K
2K
1K
0K
Stayed
3,404
Exited
1,139
Stayed
4,559
Exited
898
Female
Male

Chi-Squared



Observed:

	Stayed	Exited
Male	4,559	898
Female	3,404	1,139

Expected:

	Stayed	Exited
Male	4,366	1,091
Female	3,634	909

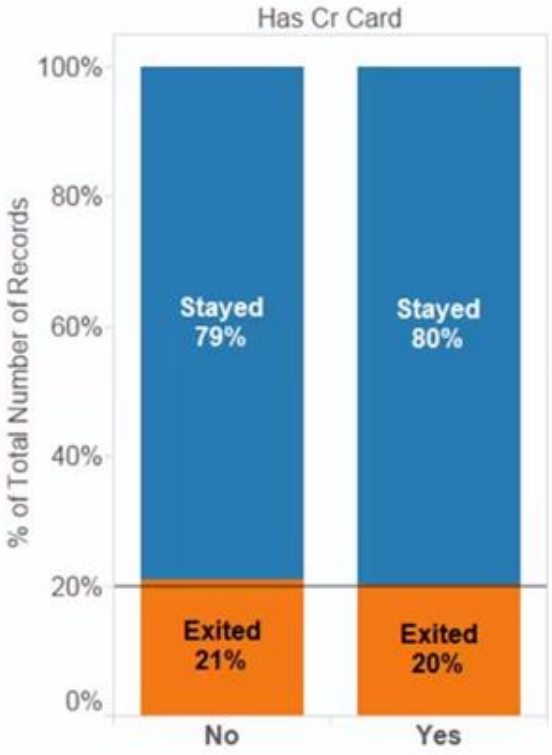
20% x Total Males

20% x Total Females

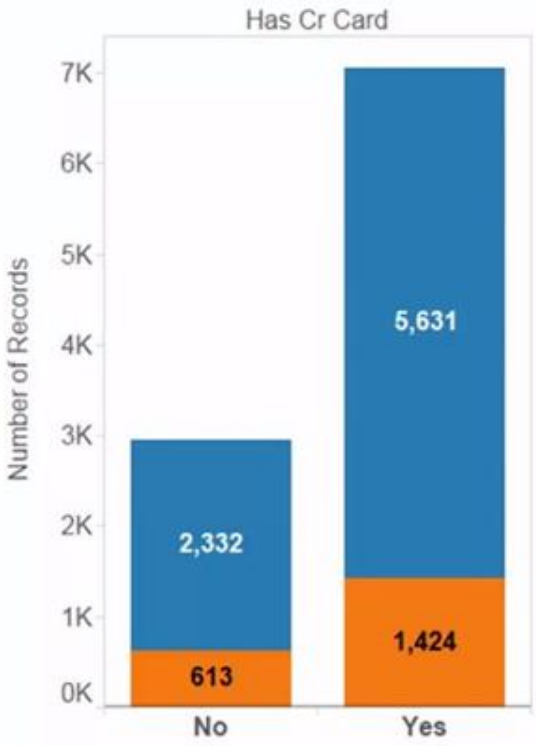
Chi-Squared

Chi-Squared is a test designed to test the **probability of independence**

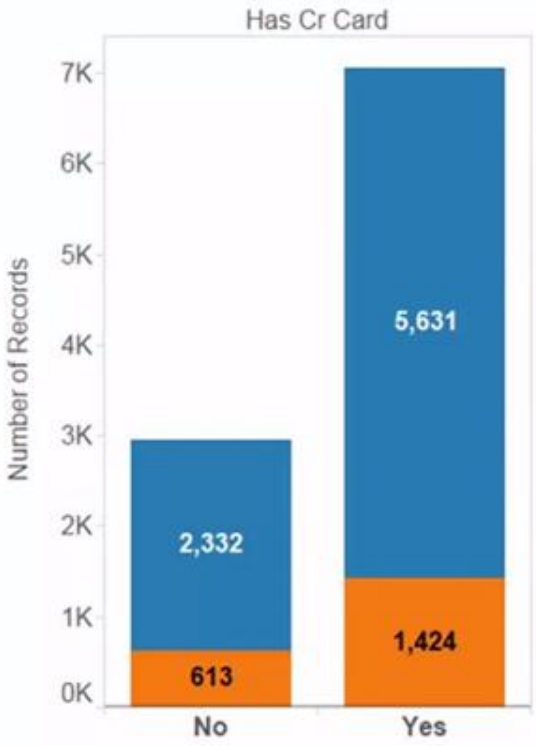
Chi-Squared


Has Cr Card
100%
80%
60%
40%
20%
0%
% of Total Number of Records
Stayed
79%
Stayed
80%
Exited
21%
Exited
20%
No
Yes

Chi-Squared


Has Cr Card
Number of Records
7K
6K
5K
4K
3K
2K
1K
0K
2,332
613
5,631
1,424
No
Yes

Chi-Squared



Observed:

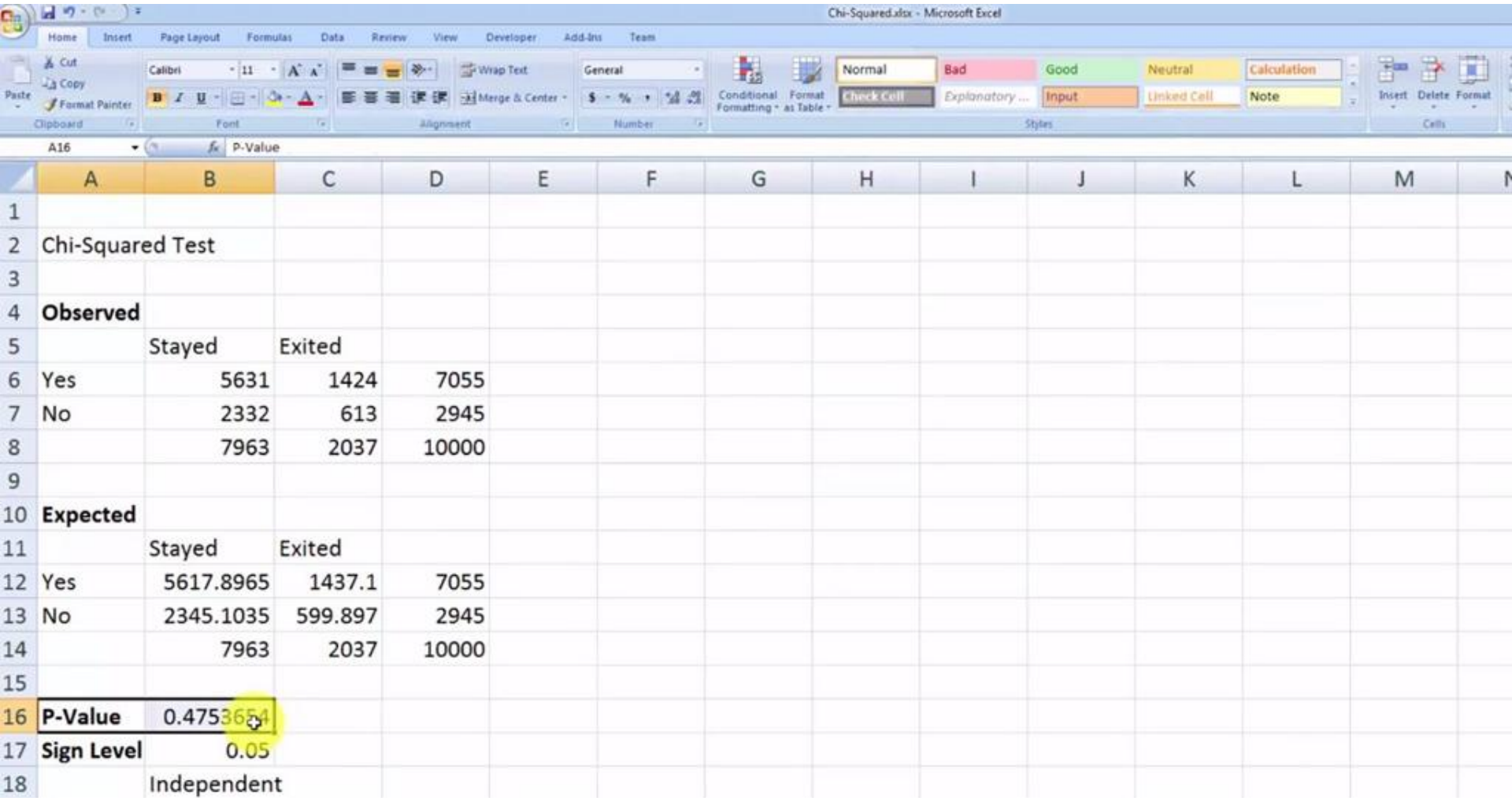
	Stayed	Exited
Yes	5,631	1,424
No	2,332	613

Expected:

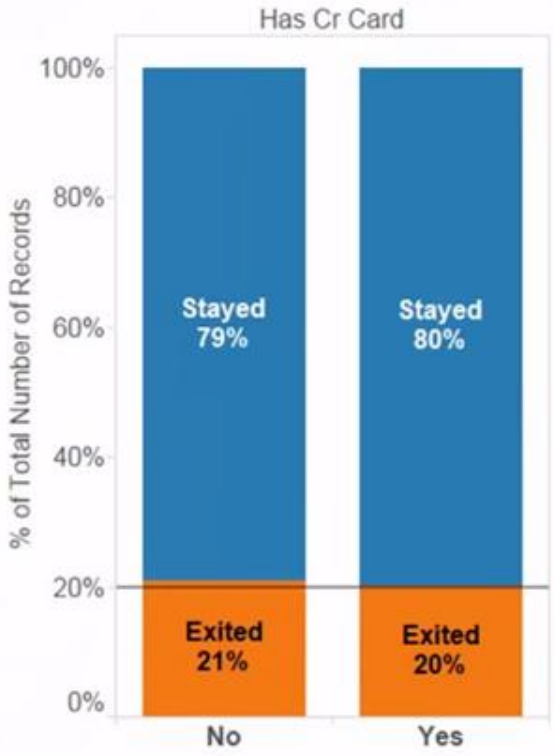
	Stayed	Exited
Yes	5,644	1,411
No	2,356	589

20% x Total Yes

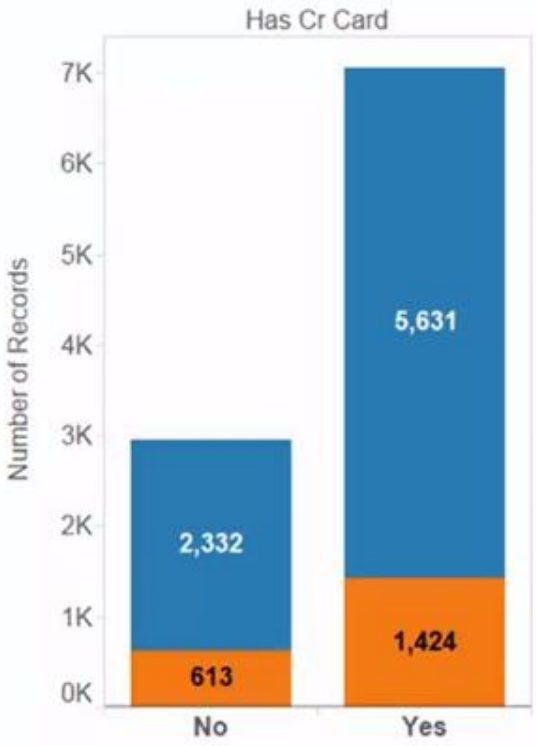
20% x Total No


Chi-Squared.xlsx - Microsoft Excel
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Add-Ins
Team
Cut
Copy
Format Painter
Paste
Clipboard
Calibri
11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Normal
Bad
Good
Neutral
Calculation
Check Cell
Explanatory ...
Input
Linked Cell
Note
Styles
Insert
Delete
Format
Cells
A16
P-Value
Chi-Squared Test
Observed
Stayed
Exited
Yes 5631 1424 7055
No 2332 613 2945
7963 2037 10000
Expected
Stayed
Exited
Yes 5617.8965 1437.1 7055
No 2345.1035 599.897 2945
7963 2037 10000
P-Value 0.4753654
Sign Level 0.05
Independent

Chi-Squared


Has Cr Card
100%
80%
60%
40%
20%
0%
% of Total Number of Records
Stayed
79%
Stayed
80%
Exited
21%
Exited
20%
No
Yes

Chi-Squared



Observed:

	Stayed	Exited
Yes	5,631	1,424
No	2,332	613

Expected:

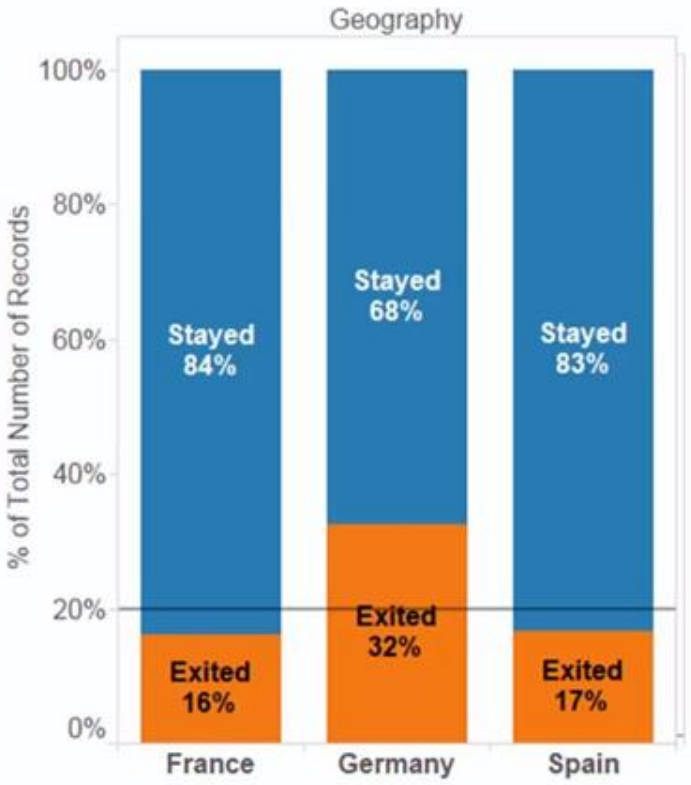
	Stayed	Exited
Yes	5,644	1,411
No	2,356	589

Chi-Squared

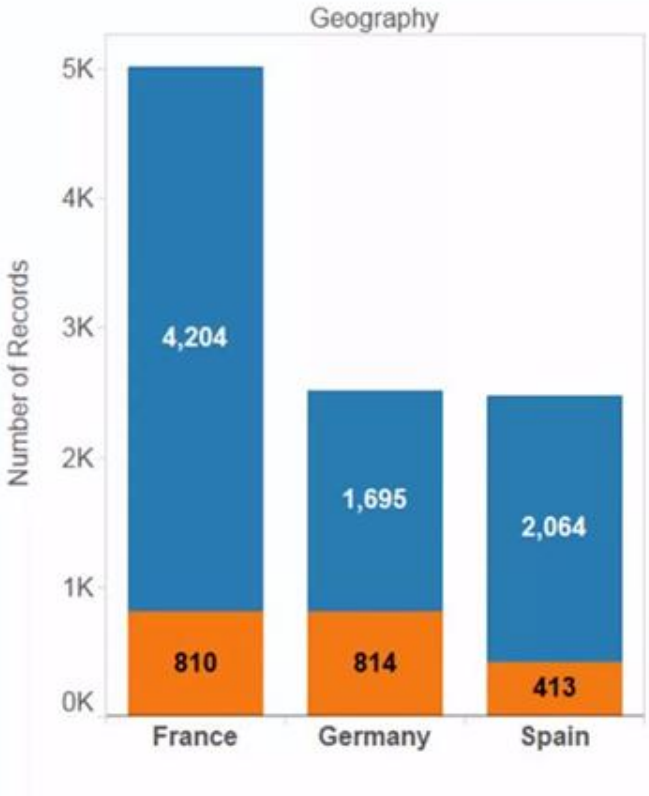
Rules

1. Probability of independence
2. NOT the relationship between variables
3. Cannot use %, need absolute values
4. Categories must be Mutually Exclusive
5. Never exclude one of the outcomes
6. Minimum 5 observations in each cell

Chi-Squared


Geography
100%
80%
60%
40%
20%
0%
% of Total Number of Records
Stayed 84%
Exited 16%
Stayed 68%
Exited 32%
Stayed 83%
Exited 17%
France
Germany
Spain

Chi-Squared



Observed:

	Stayed	Exited
France	4,204	810
Germany	1,695	814
Spain	2,064	413

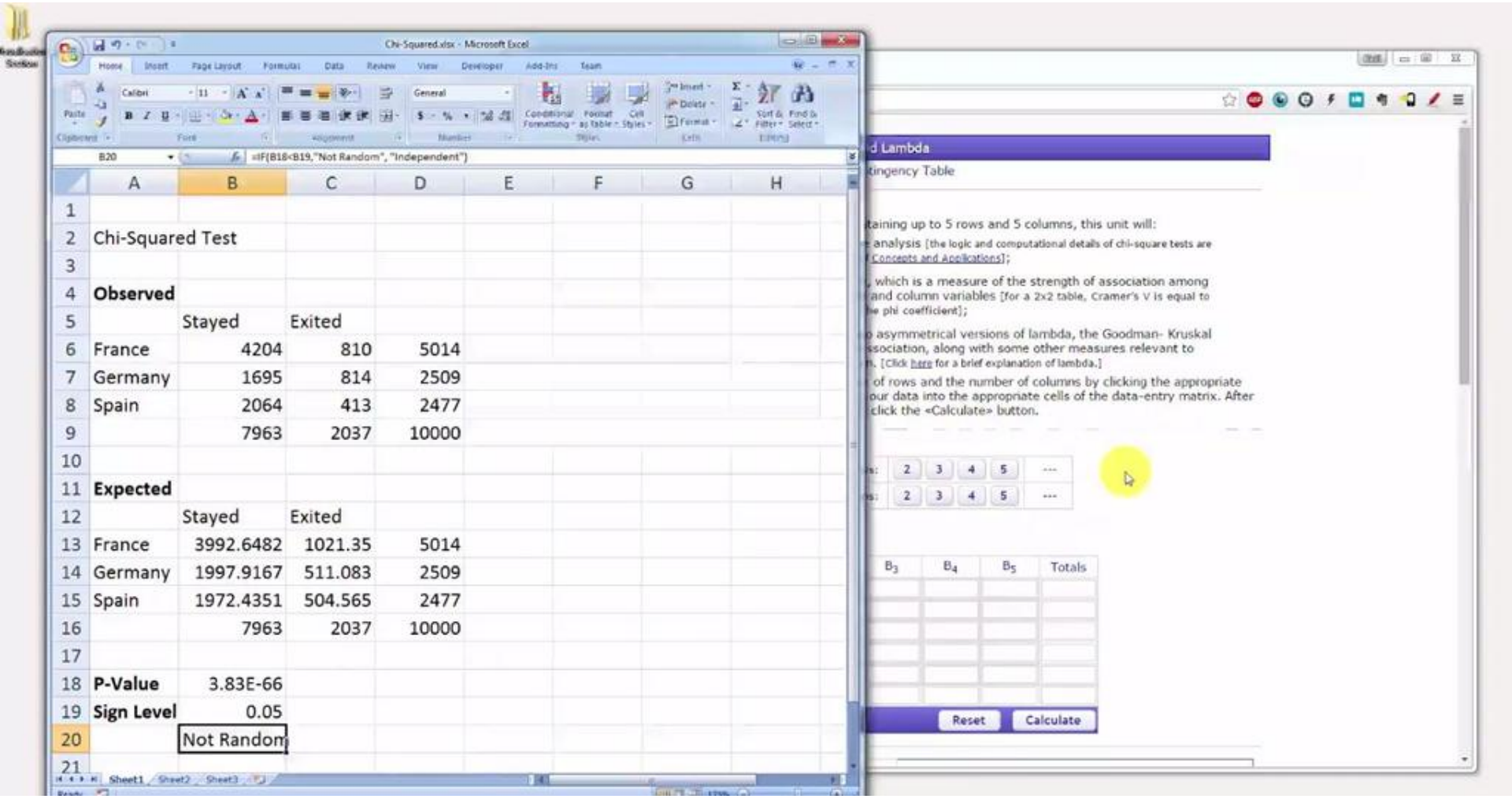
Expected:

	Stayed	Exited
France	4,011	1,003
Germany	2,007	502
Spain	1,982	495

20% x Total France

20% x Total Germany

20% x Total Spain


Chi-Squared.xlsx - Microsoft Excel
=IF(B18<B19,"Not Random","Independent")
Chi-Squared Test
Observed
Stayed Exited
France 4204 810 5014
Germany 1695 814 2509
Spain 2064 413 2477
7963 2037 10000
Expected
Stayed Exited
France 3992.6482 1021.35 5014
Germany 1997.9167 511.083 2509
Spain 1972.4351 504.565 2477
7963 2037 10000
P-Value 3.83E-66
Sign Level 0.05
Not Random
Sheet1 Sheet2 Sheet3
d Lambda
tingency Table
taining up to 5 rows and 5 columns, this unit will:
analysis [the logic and computational details of chi-square tests are
Concepts and Applications];
which is a measure of the strength of association among
and column variables [for a 2x2 table, Cramer's V is equal to
e phi coefficient];
asymmetrical versions of lambda, the Goodman- Kruskal
ssociation, along with some other measures relevant to
. [Click here for a brief explanation of lambda.]
of rows and the number of columns by clicking the appropriate
our data into the appropriate cells of the data-entry matrix. After
click the «Calculate» button.
2 3 4 5 ...
2 3 4 5 ...
B3 B4 B5 Totals
Reset Calculate