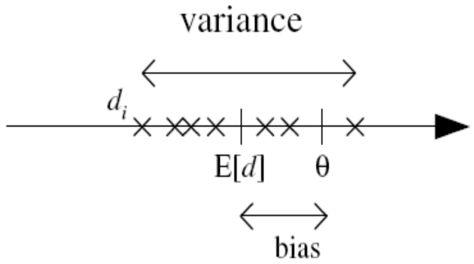
Confidence Intervals

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:

1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 5 & Teaching Material

Introduction



- We have discussed point estimates:
  - $\hat{p}$ as an estimate of a success probability, p (Bernoulli trials)
  - $\overline{X}$ as an estimate of population mean, μ
- These point estimates are almost never exactly equal to the true values they are estimating
  - In order for the point estimate to be useful, it is necessary to describe just how far off from the true value it is likely to be
  - Remember that one way to estimate how far our estimate is from the true value is to report an estimate of the standard deviation, or uncertainty, in the point estimate
- In this chapter, we can obtain more information about the estimation precision by computing a confidence interval when the estimate is normally distributed

Revisit: The Central Limit Theorem

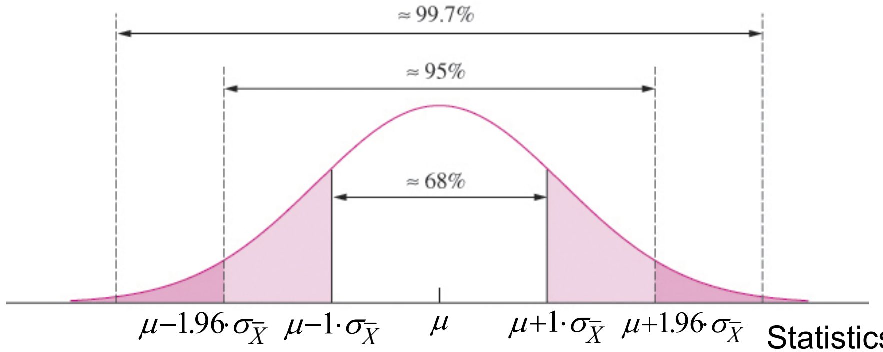
- The Central Limit Theorem
  - Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2 (n is large enough)
  - Let $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ be the sample mean
  - Let $S_n = X_1 + \cdots + X_n$ be the sum of the sample observations. Then if n is sufficiently large,
    - $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ sample mean is approximately normal !
    - And $S_n \sim N(n\mu, n\sigma^2)$ approximately

Example

- Assume that a large number of independent unbiased measurements, all using the same procedure, are made on the diameter of a piston. The sample mean $\overline{X}$ of the measurements is 14.0 cm (coming from a normal population due to the Central Limit Theorem), and the uncertainty in this quantity, which is the standard deviation $\sigma_{\overline{X}}$ of the sample mean $\overline{X}$, is 0.1 cm
- So, we have a high level of confidence that the true diameter is in the interval (13.7, 14.3). This is because it is highly unlikely that the sample mean will differ from the true diameter by more than three standard deviations

$$\underbrace{\text{Measured Value}}_{\text{random variable}} = \underbrace{\text{True Value}}_{\text{constant}} + \underbrace{\text{Bias}}_{\text{constant}} + \underbrace{\text{Random Error}}_{\text{random variable}}$$

$$E[\text{Random Error}] = 0$$



Large-Sample Confidence Interval for a Population Mean

- **Recall the previous example**: Since the population mean will not be exactly equal to the sample mean of 14, it is best to construct a **confidence interval** around 14 that is likely to cover the population mean
  - We can then quantify our level of confidence that the population mean is actually covered by the interval

- To see how to construct a confidence interval, let μ represent the unknown population mean and let σ^2 be the unknown population variance. Let $X_1,\ldots,X_{100}$ be the 100 diameters of the pistons. The observed value of $\overline{X}$ is the mean of a large sample, and the Central Limit Theorem specifies that it comes from a normal distribution with mean μ and whose standard deviation is

  $\sigma_{\overline{X}} = \sigma / \sqrt{100}$

Illustration of Capturing True Mean

- Here is a normal curve, which represents the distribution of $\overline{X}$. The middle 95% of the curve, extending a distance of $1.96\sigma_{\overline{X}}$ on either side of the population mean μ, is indicated. The following illustrates what happens if $\overline{X}$ lies within the middle 95% of the distribution:

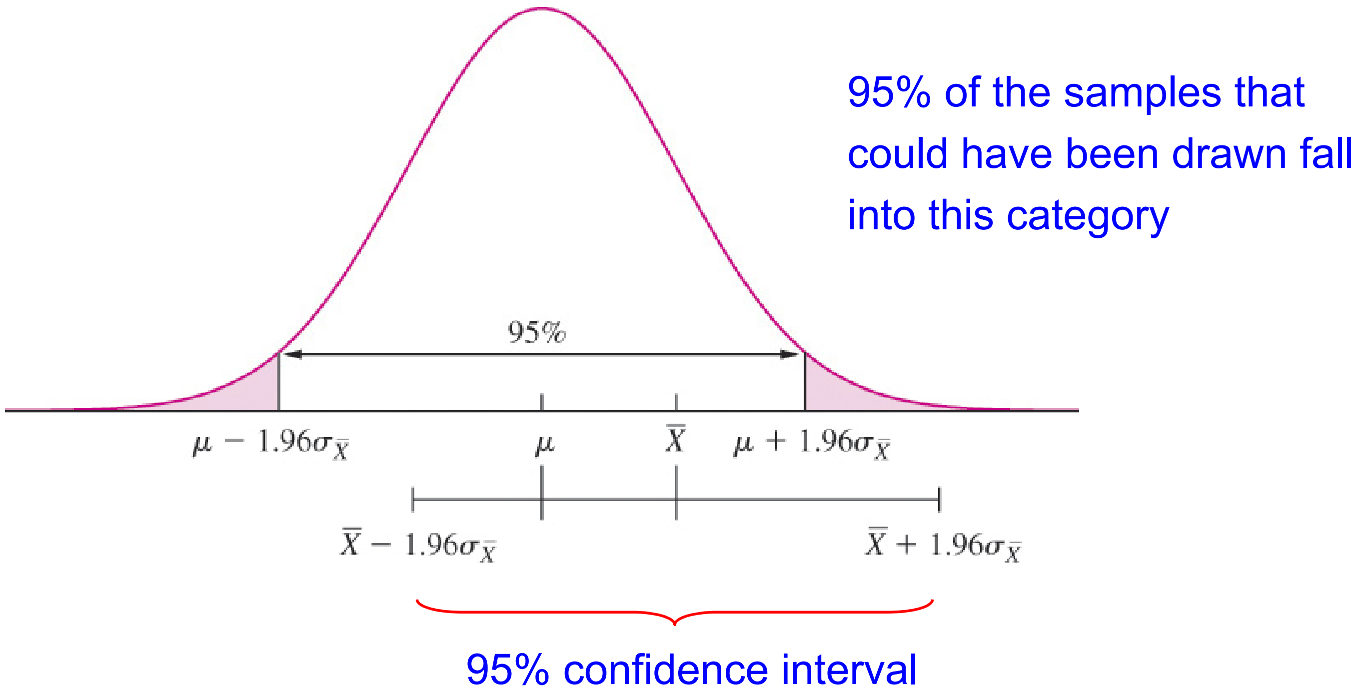
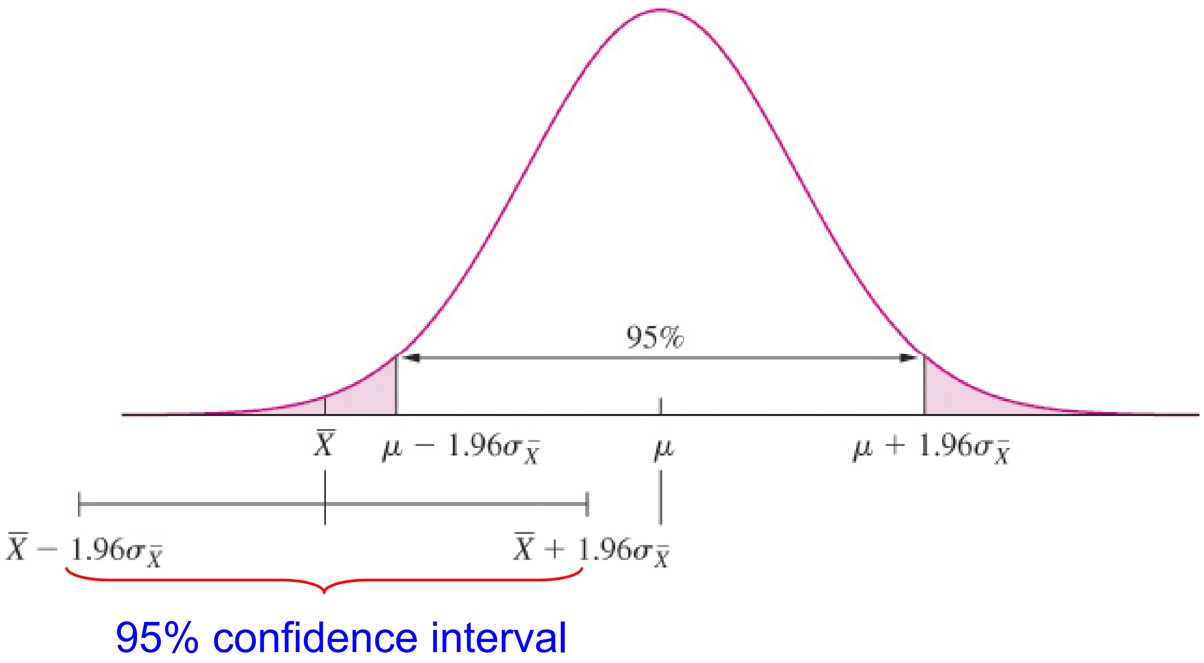


Illustration of Not Capturing True Mean

- If the sample mean lies outside the middle 95% of the curve: Only 5% of all the samples that could have been drawn fall into this category. For those more unusual samples the 95% confidence interval $\overline{X} \pm 1.96\sigma_{\overline{X}}$ fails to cover the true population mean μ



Computing a 95% Confidence Interval

- The 95% confidence interval (CI) is $\overline{X} \pm 1.96\sigma_{\overline{X}}$
- So, a 95% CI for the mean is 14 ± 1.96 (0.1). We can use the sample standard deviation as an estimate for the population standard deviation, since the sample size is large
- We can say that we are 95% confident, or confident at the 95% level, that the population mean diameter for pistons lies, between 13.804 and 14.196
- Warning: The methods described here require that the data be a random sample from a population. When used for other samples, the results may not be meaningful

Question?

- Does this 95% confidence interval actually cover the population mean μ ?
  - It depends on whether this particular sample happened to be one whose mean (i.e. sample mean) came from the middle 95% of the distribution or whether it was a sample whose mean (i.e. sample mean) was unusually large or small, in the outer 5% of the population
  - There is no way to know for sure into which category this particular sample falls
  - In the long run, if we repeated these confidence intervals over and over, then 95% of the samples will have means (i.e. sample mean) in the middle 95% of the population. Then 95% of the confidence intervals will cover the population mean

Extension

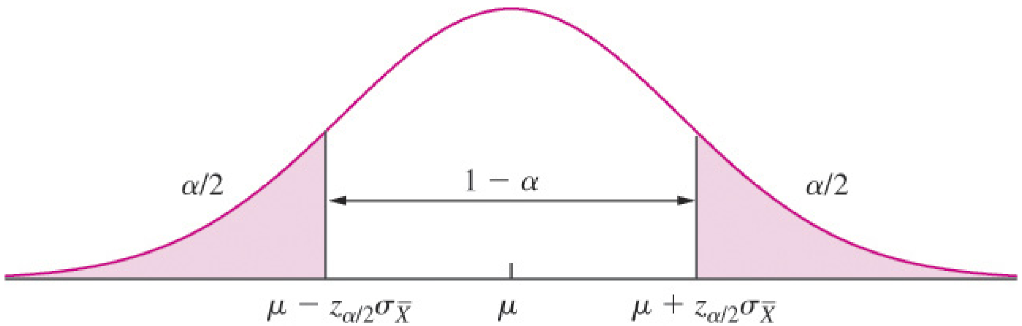
- We are not always interested in computing 95% confidence intervals. Sometimes, we would like to have a different level of confidence
  - We can use this reasoning to compute confidence intervals with various confidence levels
- Suppose we are interested in 68% confidence intervals, then we know that the middle 68% of the normal distribution is in an interval that extends $1.0\sigma_{\overline{X}}$ on either side of the population mean μ
  - It follows that an interval of the same length around $\overline{X}$ specifically, will cover the population mean for 68% of the samples that could possibly be drawn
  - For our example, a 68% CI for the diameter of pistons is $14.0 \pm 1.0(0.1)$, or (13.9, 14.1)

100(1 - α)% CI

- Let $X_1,\ldots,X_n$ be a *large* (n > 30) random sample from a population with mean μ and standard deviation σ, so that is approximately normal. Then a level 100(1 - α)% confidence interval for μ is

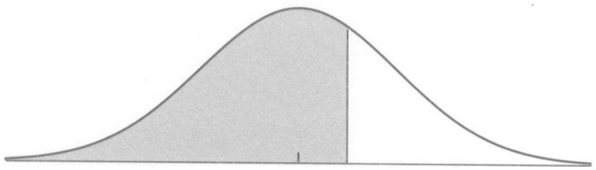
$$\overline{X} \pm z_{\alpha/2}\sigma_{\overline{X}}$$

  - $z_{\alpha/2}$ I s the *z*-score that cuts off an area of $\alpha/2$ in the right-hand tail
  - where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. When the value of σ is unknown, it can be replaced with the sample standard deviation *s*



Z-Table

TABLE A.2 Cumulative normal distribution (continued)



E.g., $\overline{X} \pm z_{\alpha/2}\sigma_{\overline{X}}$ and $\alpha = 0.05$

$\Rightarrow z_{\alpha/2} = 1.96$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Particular CI's

- $\bar{X} \pm \frac{s}{\sqrt{n}}$ is a 68% interval for μ
- $\bar{X} \pm 1.645\frac{s}{\sqrt{n}}$ is a 90% interval for μ
- $\bar{X} \pm 1.96\frac{s}{\sqrt{n}}$ is a 95% interval for μ
- $\bar{X} \pm 2.58\frac{s}{\sqrt{n}}$ is a 99% interval for μ
- $\bar{X} \pm 3\frac{s}{\sqrt{n}}$ is a 99.7% interval for μ

Note that even for large samples, the distribution of $\bar{X}$ is only approximately normal, rather than exactly normal. Therefore, the levels stated for confidence interval are approximate.

Example (CI Given a Level)

- **Example 5.1**: The sample mean and standard deviation for the fill weights of 100 boxes are $\overline{X}$ = 12.05 and *s* = 0.1. Find an 85% confidence interval for the mean fill weight of the boxes.

<u>**Answer**</u>: To find an 85% CI, set 1 - α = .85, to obtain α = 0.15 and $\alpha/2$ = 0.075. We then look in the table for $z_{0.075}$, the *z*-score that cuts off 7.5% of the area in the right-hand tail. We find $z_{0.075}$ = 1.44. We approximate $\sigma_{\overline{X}} \approx s/\sqrt{n} = 0.01.$

So the 85% CI is 12.05 $\pm$ (1.44)(0.01) or (12.0356, 12.0644).

Another Example (The Level of CI)

- **Question**: There is a sample of 50 micro-drills with an average lifetime (expressed as the number of holes drilled before failure) was 12.68 with a standard deviation of 6.83. Suppose an engineer reported a confidence interval of (11.09, 14.27) but neglected to specify the level. What is the level of this confidence interval?

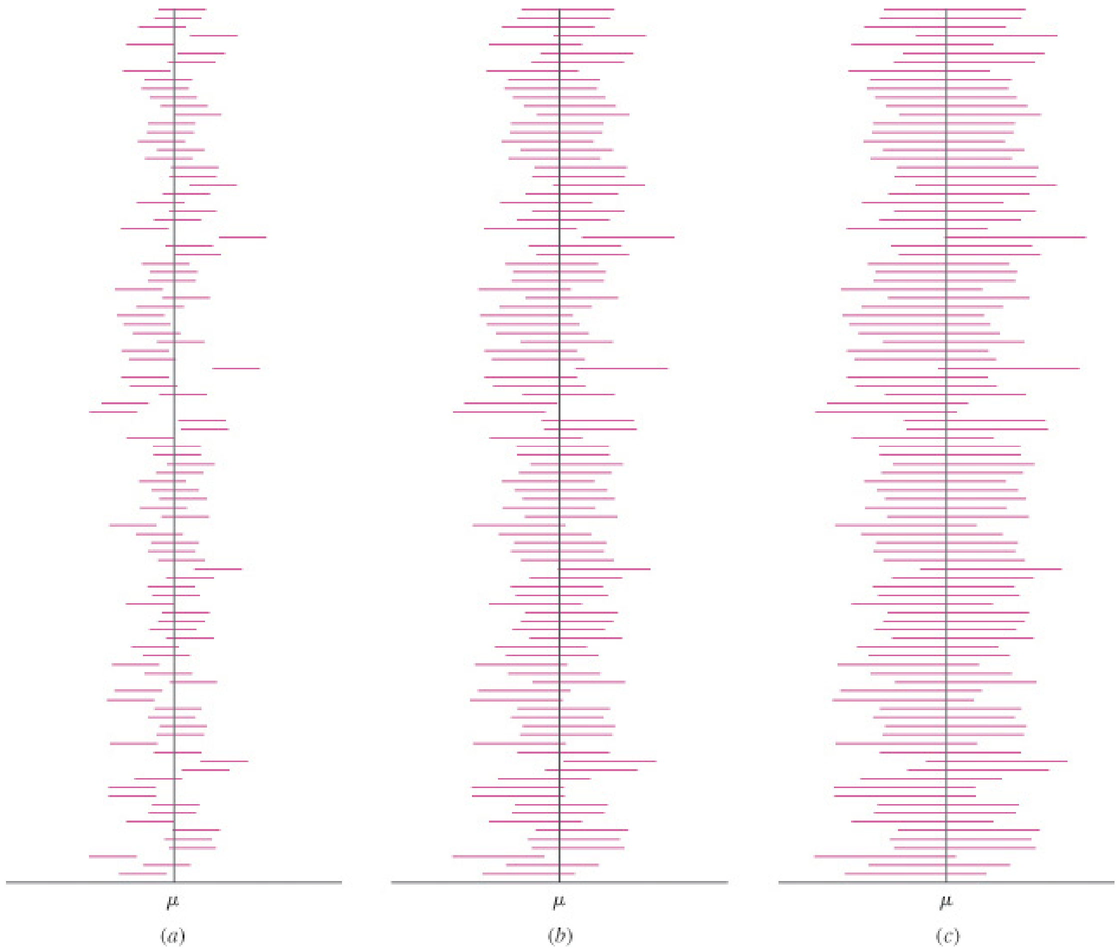
  Answer: The confidence interval has the form $\overline{X} \pm z_{\alpha/2} s / \sqrt{n}$. We will solve for $z_{\alpha/2}$, and then consult the z table to determine the value of α. The upper confidence limit of 14.27 therefore satisfies the equation 14.27 = 12.68 + $z_{\alpha/2}(6.83/\sqrt{50})$. Therefore, $z_{\alpha/2} = 1.646$. From the z table, we determine that $\alpha/2$, the area to the right of 1.646, is approximately 0.05. The level is 100(1 - α)%, or 90%.

More About CI's (1/2)

- The confidence level of an interval measures the reliability of the method used to compute the interval
- A level 100(1 - α)% confidence interval is one computed by a method that in the long run will succeed in in covering the population mean a proportion 1 - α of all the times that it is used
- In practice, there is a decision about what level of confidence to use
- This decision involves a trade-off, because intervals with greater confidence are less precise

More About CI's (2/2)

100 samples



68% confidence intervals 95% confidence intervals 99.7% confidence intervals

Probability vs. Confidence

- In computing CI, such as the one of diameter of pistons: (13.804, 14.196), it is tempting to say that the probability that μ lies in this interval is 95%
- The term probability refers to random events, which can come out differently when experiments are repeated
- 13.804 and 14.196 are fixed not random. The population mean is also fixed. The mean diameter is either in the interval or not
  - There is no randomness involved
- So, we say that we have 95% confidence that the population mean is in this interval
  - It is correct to say that a method for computing a 95% confidence interval has probability 95% of covering the population mean

Determining Sample Size

- Back to the example of diameter of pistons: We had a CI of (13.804, 14.196).
  - This interval specifies the mean to within ±0.196. Now assume that the interval is too wide to be useful

Question: Assume that it is desirable to produce a 95% confidence interval that specifies the mean to within ± 0.1

- To do this, the sample size must be increased. The width of a CI is specified by $\pm z_{\alpha/2}\sigma/\sqrt{n}$. If we know α and σ is specified, then we can find the *n* needed to get the desired width
- For our example, the $z_{\alpha/2}$ = 1.96 and the estimated standard deviation of the population is 1. So, 0.1 =1.96(1)/$\sqrt{n}$, then the *n* accomplishes this is 385 (always round up)

One-Sided Confidence Intervals (1/2)

- We are not always interested in CI's with an upper and lower bound
- For example, we may want a confidence interval on battery life. We are only interested in a lower bound on the battery life. There is not an upper bound on how long a battery can last (confidence interval =(low bound,∞))
- With the same conditions as with the two-sided CI, the level 100(1-α)% lower confidence bound for μ is

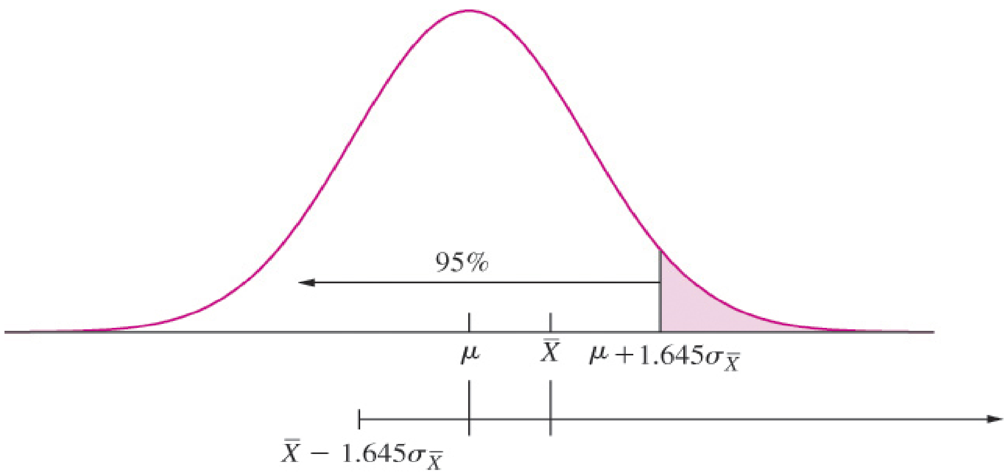
$$\overline{X} - z_{\alpha}\sigma_{\overline{X}}.$$

and the level 100(1-α)% upper confidence bound for μ is

$$\overline{X} + z_{\alpha}\sigma_{\overline{X}}.$$

One-Sided Confidence Intervals (2/2)

- **Example**: One-sided Confidence Interval (for Low Bound)



$$\left(\overline{X} - 1.645\sigma_{\overline{X}},\ \infty\right)$$

Confidence Intervals for Proportions

- The method that we discussed in the last section (Sec. 5.1) was for mean from any population from which a large sample is drawn
- When the population has a Bernoulli distribution Y, this expression takes on a special form (the mean is equal to the success probability)
  - If we denote the success probability as p and the estimate for p as $\hat{p}$ which can be expressed by

$$\hat{p} = \frac{X}{n}$$

n : the sample size

X :number of sample items Y_i that success

$$X = Y_1 + Y_2 + \cdots + Y_n$$

  - A 95% confidence interval (CI) for p is

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Comments

- The limits of the confidence interval contain the unknown population proportion p
  - We have to somehow estimate this (p)
    - E.g., using $\hat{p}$

- Recent research shows that a slight modification of n and an estimate of p improve the interval
  - Define

$$\tilde{n} = n + 4$$

    - And

$$\tilde{p} = \frac{X + 2}{\tilde{n}}$$

CI for p

- Let X be the number of successes in n independent Bernoulli trials with success probability p , so that $X \sim \mathrm{Bin}(n,p)$

- Then a 100(1 - α)% confidence interval for p is

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}.$$

  - If the lower limit is less than 0, replace it with 0.
  - If the upper limit is greater than 1, replace it with 1

Determine the Sample Size to having a specific CI for p

- Sometimes we wish to compute a necessary sample size without having a reliable estimate $\tilde{p}$ available
  - The quantity $\tilde{p}(1-\tilde{p})$ and $\tilde{n}$ determine the width of the confidence interval
  - The quantity $\tilde{p}(1-\tilde{p})$ is greatest (=0.25) when $\tilde{p}=0.5$

- Example 5.14: How large a sample is needed to guarantee that the width of the 95% confidence interval of p will on larger than ±0.08?

$$CI \text{ has width } \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

$$\Rightarrow 1.96\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} \le 0.08$$

$$\Rightarrow 1.96\sqrt{0.5(1-0.5)/(n+4)} \le 0.08 \quad (\text{with a conservative sample size})$$

$$\Rightarrow n \approx 147$$

Small Sample CI for a Population Mean

- The methods that we have discussed for a population mean previously require that the sample size be large
- When the sample size is small, there are no general methods for finding CI's
- If the population is approximately normal, a probability distribution called the Student's *t* distribution can be used to compute confidence intervals for a population mean

$$\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}}=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\neq\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

More on CI's

- What can we do if $\bar{X}$ is the mean of a *small* sample?
- If the sample size is small, *s* may not be close to σ, and $\bar{X}$ may not be approximately normal. If we know nothing about the population from which the small sample was drawn, there are no easy methods for computing CI's
- However, if the population is approximately normal, $\bar{X}$ will be approximately normal even when the sample size *n* is small. It turns out that we can use the quantity $(\bar{X}-\mu)/(s/\sqrt{n})$, but since *s* may not be close to σ, this quantity instead has a Student's t distribution with *n*-1 degrees of freedom, which we denote t_{n-1}

Student's *t* Distribution (1/2)

- Let $X_1, \ldots, X_n$ be a *small* (*n* < 30) random sample from a *normal* population with mean μ. Then the quantity

$$\frac{(\bar{X} - \mu)}{s / \sqrt{n}}.$$

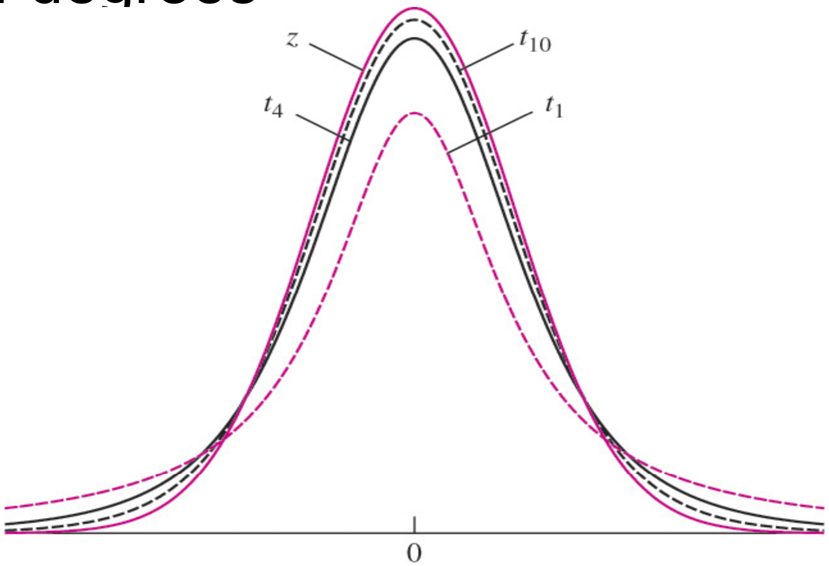
has a Student's *t* distribution with *n* -1 degrees of freedom (denoted by t_{n-1}).

- When *n* is large, the distribution of the above quantity is very close to normal, so the normal curve can be used, rather than the Student's *t*

Cf. http://en.wikipedia.org/wiki/Student's_t-distribution

Student's *t* Distribution (2/2)

- Plots of probability density function of student's t curve for various of degrees

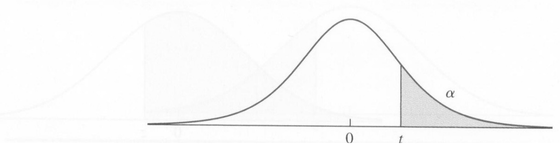


  - The normal curve with mean 0 and variance 1 (*z* curve) is plotted for comparison
  - The *t* curves are more spread out than the normal, but the amount of extra spread out decreases as the number of degrees of freedom increases

More on Student's *t*

- Table A.3 called a ***t* table**, provides probabilities associated with the Student's *t* distribution

TABLE A.3 Upper percentage points for the Student's *t* distribution



ν	α = 0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.682	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Examples

- **Question 1**: A random sample of size 10 is to be drawn from a normal distribution with mean 4. The Student's *t* statistic $t = (\overline{X} - 4)/(s/\sqrt{10})$ is to be computed. What is the probability that *t* > 1.833?
  - **Answer**: This *t* statistic has 10 – 1 = 9 degrees of freedom. From the *t* table, *P*(*t* > 1.833) = 0.05

- **Question 2**: Find the value for the t_{14} distribution whose lower-tail probability is 0.01
  - **Answer**: Look down the column headed with "0.01" to the row corresponding to 14 degrees of freedom. The value for *t* = 2.624. This value cuts off an area, or probability, of 1% in the upper tail. The value whose lower-tail probability is 1% is -2.624

Student's *t* CI

- Let $X_1, \ldots, X_n$ be a *small* random sample from a *normal* population with mean μ. Then a level 100(1 - α)% CI for μ is

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}.$$

Two-sided CI

- To be able to use the Student's *t* distribution for calculation and confidence intervals, you must have a sample that comes from a population that it approximately normal

Other Student's *t* CI's

- Let $X_1,\ldots,X_n$ be a *small* random sample from a *normal* population with mean μ
  - Then a level 100(1 - α)% upper confidence bound for μ i

$$\overline{X} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}.$$

one-sided CI

  - Then a level 100(1 - α)% lower confidence bound for μ is

$$\overline{X} - t_{n-1,\alpha}\frac{s}{\sqrt{n}}.$$

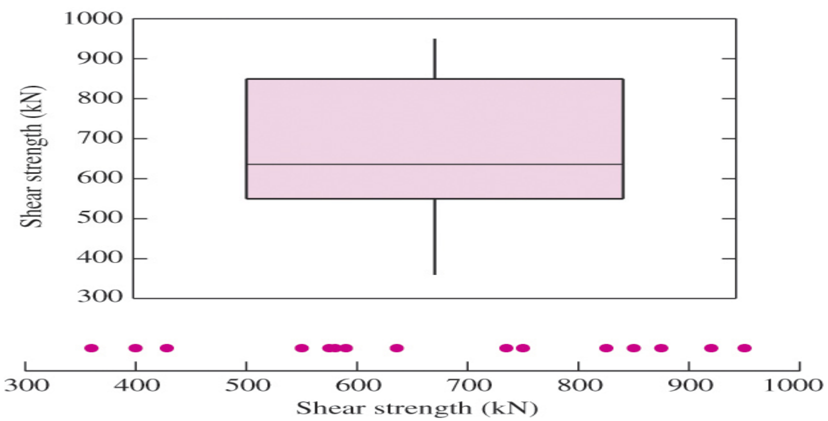
one-sided CI

- Occasionally a small sample may be taken from a normal population whose standard deviation σ is known. In these cases, we do not use the Student's *t* curve, because we are not approximating σ with *s*. The CI to use here, is the one using the *z* table, that we discussed in the first section $\frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$

Determine the Appropriateness of Using *t* Distribution (1/2)

- We have to decide whether a population is approximately normal before using *t* distribution to calculate CI
  - A reasonable way is construct a boxplot or dotplot of the sample
  - If these plots do not reveal a strong asymmetry or any outliers, the it most cast the Student's t distribution will be reliable
- **Example 5.9**: Is it appropriate to use *t* distribution to calculate the CI for a population mean given a a random sample with 15 items shown below

580, 400, 428, 825, 850,
875, 920, 550, 575, 750,
636, 360, 590, 735, 950.

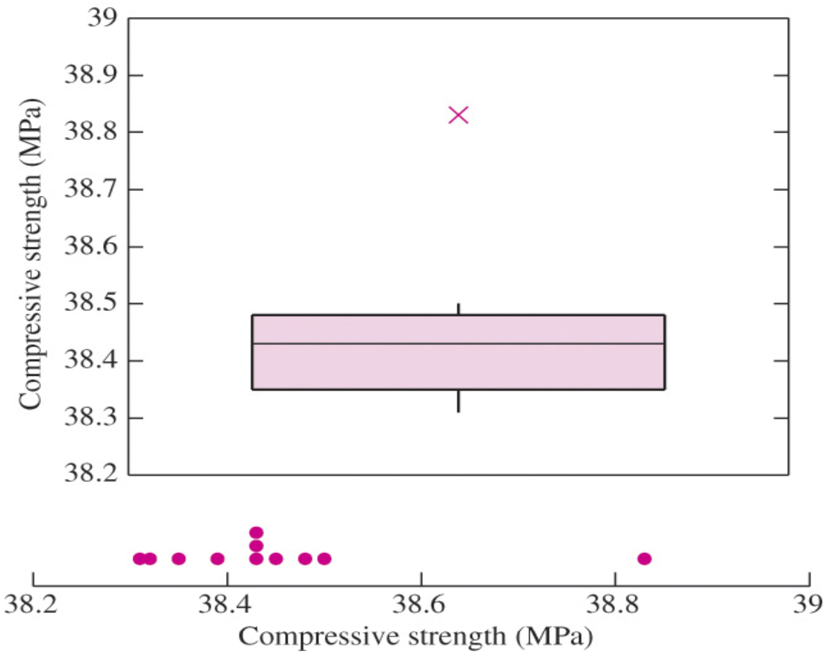


Yes !

Determine the Appropriateness of Using *t* Distribution (2/2)

- Example 5.20: Is it appropriate to use *t* distribution to calculate the CI for a population mean given a a random sample with 11 items shown below

38.43, 38.43, 38.39, 38.83, 38.45, 38.35, 38.43, 38.31, 38.32, 38.38, 38.50.



No !

CI for the Difference in Two Means (1/2)

- We also can estimate the difference between the means μ_X and μ_Y of two populations X and Y
  - We can draw two independent random samples, one from X and the other one from Y, each of which respectively has sample means $\overline{X}$ and $\overline{Y}$
  - Then construct the CI for $\mu_X - \mu_Y$ by determining the distribution of $\overline{X} - \overline{Y}$
- Recall the probability theorem:
  Let X and Y be independent, with $X \sim N\left(\mu_X, \sigma_X^2\right)$ and $Y \sim N\left(\mu_Y, \sigma_Y^2\right)$
  Then

$$X + Y \sim N\left(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$

  - And

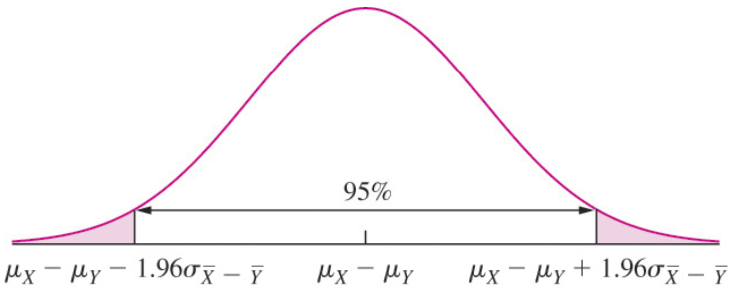
$$X - Y \sim N\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2\right)$$

CI for the Difference in Two Means (2/2)

- Let $X_1,\ldots,X_{n_X}$ be a large random sample of size n_X from a population with mean μ_X and standard deviation σ_X, and let $Y_1,\ldots,Y_{n_Y}$ be a large random sample of size n_Y from a population with mean μ_Y and standard deviation σ_Y . If the two samples are independent, then a level 100(1- α)% CI for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

Two-sided CI



$\alpha = 0.05$

  - When the values of σ_X and σ_Y are unknown, they can be replaced with the sample standard deviations s_X and s_Y

CI for Difference Between Two Proportions (1/3)

- Recall that in a Bernoulli population, the mean is equal to the success probability (population proportion) p
- Let X be the number of successes in n_X independent Bernoulli trials with success probability p_X, and let Y be the number of successes in n_Y independent Bernoulli trials with success probability p_Y, so that $X \sim \text{Bin}(n_X, p_X)$ and $Y \sim \text{Bin}(n_Y, p_Y)$
  - The sample proportions

$$\hat{p}_X = \frac{X}{n_X} \sim N\left(p_X, \frac{p_X(1-p_X)}{n_X}\right)$$

following from the central limit theorem (n_X and n_Y are large)

$$\hat{p}_Y = \frac{Y}{n_Y} \sim N\left(p_Y, \frac{p_Y(1-p_Y)}{n_Y}\right)$$

$$\Rightarrow \hat{p}_X - \hat{p}_Y = \frac{X}{n_X} \sim N\left(p_X - p_Y, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right)$$

CI for Difference Between Two Proportions (2/3)

- The difference satisfies the following inequality for 95% of all possible samples

$$\hat{p}_X - \hat{p}_Y - 1.96\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$$
$$< p_X - p_Y <$$
$$\hat{p}_X - \hat{p}_Y + 1.96\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}$$

Two-sided CI

  – Traditionally in the above inequality, p_X is replaced by $\hat{p}_X$ and p_Y is replaced by $\hat{p}_Y$

CI for Difference Between Two Proportions (3/3)

- Adjustment (In implementation):
  - Define

$$\tilde{n}_X = n_X + 2, \tilde{n}_Y = n_Y + 2, \tilde{p}_X = (X+1)/\tilde{n}_X, \text{and } \tilde{p}_Y = (Y+1)/\tilde{n}_Y$$

  - The 100(1-α)% CI for the difference $p_X - p_Y$ is

$$\tilde{p}_X - \tilde{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_X(1-\tilde{p}_X)}{n_X} + \frac{\tilde{p}_Y(1-\tilde{p}_Y)}{n_Y}}.$$

    - If the lower limit of the confidence interval is less than -1, replace it with -1
    - If the upper limit of the confidence interval is greater than 1, replace it with 1

Small-Sample CI for Difference Between Two Means (1/2)

- Let $X_1,\ldots,X_{n_X}$ be a random sample of size n_X from a normal population with mean μ_X and standard deviation σ_X, and let $Y_1,\ldots,Y_{n_Y}$ be a random sample of size n_Y from a normal population with mean μ_Y and standard deviation σ_Y. Assume that the two samples are independent. If the populations do not necessarily have the same variance, a level 100(1- α)% CI for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm t_{v,\alpha/2}\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}.$$

Two-sided CI

  – The number of degrees of freedom, v (pronounced "nu"), is given by (rounded down to the nearest integer)

$$v = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{\left(s_X^2/n_X\right)^2}{n_X - 1} + \frac{\left(s_Y^2/n_Y\right)^2}{n_Y - 1}}$$

Small-Sample CI for Difference Between Two Means (2/2)

- If we further know the populations X and Y are known to have nearly the same variance σ. Then a 100(1-α)% CI for $\mu_X - \mu_Y$ is

$$\overline{X} - \overline{Y} \pm t_{n_X + n_Y - 2, \alpha/2} s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}.$$

Two-sided CI

  - Degrees of freedom: $n_X + n_Y - 2$

  – The quantity s_p is the pooled variance, used to approximate σ and given by

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}.$$

- *Don't assume the population variance are equal just because the sample variance are close!*

CI for Paired Data (1/3)

- The methods discussed previously for finding CI's on the basis of two samples have required the samples are independent
- However, in some cases, it is better to design an experiment so that each item in one sample is paired with an item in the other (the variability between the cars disappears)
  - **Example**: Tread wear of tires made of two different materials

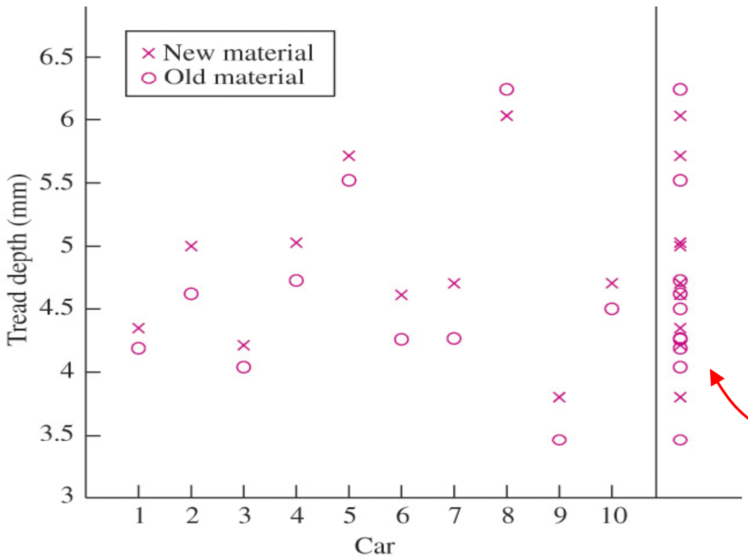


TABLE 5.1 Depths of tread, in mm, for tires made of new and old material

	Car									
	1	2	3	4	5	6	7	8	9	10
New material	4.35	5.00	4.21	5.03	5.71	4.61	4.70	6.03	3.80	4.70
Old material	4.19	4.62	4.04	4.72	5.52	4.26	4.27	6.24	3.46	4.50
Difference	0.16	0.38	0.17	0.31	0.19	0.35	0.43	−0.21	0.34	0.20

include the variability between cars
and the variability in wear between tires

CI for Paired Data (2/3)

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be sample pairs. Let $D_i = X_i - Y_i$. Let μ_X and μ_Y represent the population means for X and Y, respectively. We wish to find a CI for the difference $\mu_X - \mu_Y$. Let μ_D represent the population mean of the differences, then $\mu_D = \mu_X - \mu_Y$. It follows that a CI for μ_D will also be a CI for $\mu_X - \mu_Y$

- Now, the sample $D_1, \ldots, D_n$ is a random sample from a population with mean μ_D , we can use one-sample methods to find CIs for μ_D

CI for Paired Data (3/3)

- Let $D_1,\ldots,D_n$ be a *small* random sample (n < 30) of differences of pairs. If the population of differences is approximately normal, then a level 100(1-α)% CI for μ_D is

$$\overline{D} \pm t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}.$$

- If the sample size is large, a level 100(1-α)% CI for μ_D is

$$\overline{D} \pm z_{\alpha/2}\sigma_{\overline{D}}.$$

  – In practice, $\sigma_{\overline{D}}$ is approximated with $\frac{s_D}{\sqrt{n}}$

Summary

- We learned about large and small CI's for means
- We also looked at CI's for proportions
- We discussed large and small CI's for differences in means
- We explored CI's for differences in proportions