Lecture 10 - Confidence Intervals for Sample Means

Sta102/BME102

Colin Rundel

February 18, 2015

Central Limit Theorem

Let $X_1, X_2, X_3, \ldots, X_n \sim D$ be n iid samples from the distribution D then:

Central limit theorem - sum of iid RVs (S_n)

The distribution of the *sum* of n independent and identically distributed random variables X is approximately normal when n is large.

$$X_1 + X_2 + \cdots + X_n = S_n \sim N\left(\mu = n\,E(X),\, \sigma^2 = n\,Var(X)\right)$$

Central limit theorem - avergae of iid RVs ($\bar{X}$)

The distribution of the *average* of n independent and identically distributed random variables X is approximately normal when n is large.

$$(X_1 + X_2 + \cdots + X_n)/n = \bar{X} \sim N\left(\mu = E(X),\, \sigma^2 = Var(X)/n\right)$$

CLT and Sampling Distribution

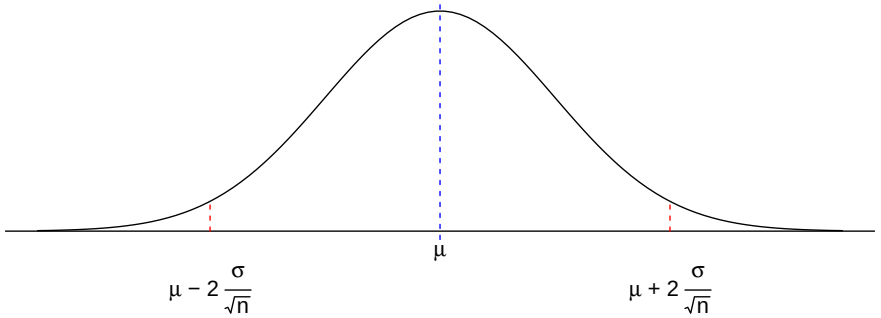
Remember for last time we went through the process of generating a sampling distribution (by generating a bunch of samples and calculating the sample average of each).

The sampling distribution is the distribution of the sample statistic ($\bar{X}$ in this case).

So for some samples (large enough sample size, reasonable population distribution) then the sampling distribution of either the sum or average of the sames is given by the Central Limit Theorem.

Why do we care?

$$\bar{X} \sim N(E(X) = \mu,\ Var(X)/n = \sigma^2/n)$$



As such, we know that 96% of the time our sample $\bar{X}$ will be within $\pm 2\frac{\sigma}{\sqrt{n}}$ of the true mean.

Confidence intervals and the CLT

We have a point estimate $\bar{X}$ for the population mean μ, but we want to design a "net" to have a reasonable chance of capturing μ.

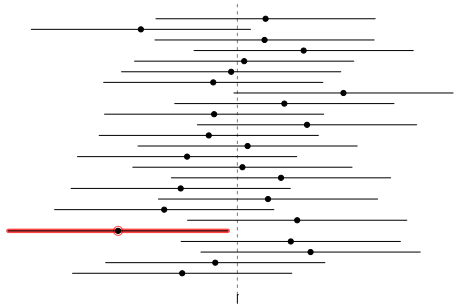
- From the CLT we know that we can think of $\bar{X}$ as a sample from $N(\mu,\ \sigma^2/n)$.
- Therefore, 96% of observed $\bar{X}$'s should be within 2 SEs ($2\ \sigma/\sqrt{n}$) of μ.
- Clearly then for 96% of random samples from the population, μ must then be with in 2 SEs of $\bar{X}$.

Note that we are being very careful about the language here - the 96% here only applies to random samples in the abstract. Once we have actually taken a sample $\bar{X}$ will either be within 2 SEs or outside of 2 SEs

What does 96% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $\textit{point estimate} \pm 2 \times \textit{SE}$.
- Then about 96% of those intervals would contain the true population mean (μ).

- The figure on the left shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



- It **does not** mean there is a 96% probability the CI contains the true value

Changing the confidence level

In general, confidence intervals will always be of the form:

$$\textit{point estimate} \pm \textit{CV} \times \textit{SE}.$$

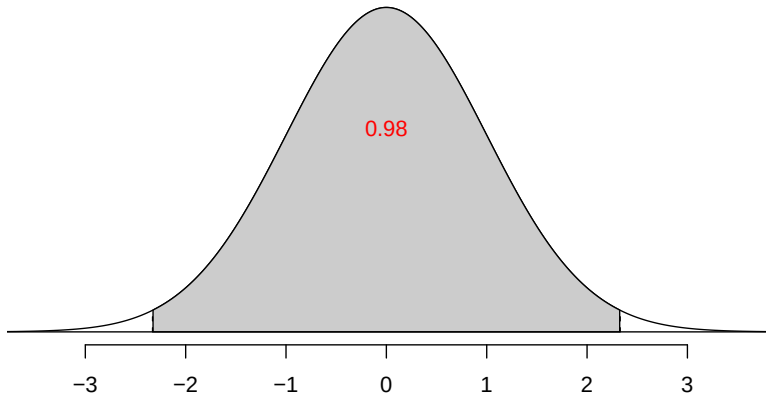
- In order to change the confidence level all we need to do is adjust the critical value in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

If the conditions for the CLT are met then, $CV = Z^\star$ which comes from the normal distribution and $SE = SD(X)/\sqrt{n}$.

- For a 95% confidence interval, $CV = Z^\star = 1.96$.
- Using the Z table it is possible to find the appropriate $Z^\star$ for any desired confidence level.

Example - Calculating $Z^\star$

What is the appropriate value for $Z^\star$ when calculating a 98% confidence interval?



Example - Sample Size

Coca-Cola wants to estimate the per capita number of Coke products consumed each year in the United States, in order to properly forecast market demands they need their margin of error to be 5 items at the 95% confidence level. From previous years they know that $\sigma \approx 30$. How many people should they survey to achieve the desired accuracy? What if the requirement was at the 99% confidence level?

Common Misconceptions

1. The confidence level of a confidence interval is the probability that the interval contains the true population parameter.

   This is incorrect, CIs are part of the frequentist paradigm and as such the population parameter is fixed but unknown. Consequently, the probability any given CI contains the true value must be 0 or 1 (it does or does not).

2. A narrower confidence interval is always better.

   This is incorrect since the width is a function of both the confidence level and the standard error.

3. A wider interval means less confidence.

   This is incorrect since it is possible to make very precise statements with very little confidence.

A small problem

Lets assume we are collecting a large sample (n=200) from a population and measuring some numeric characteristic that has distribution D, where $E(D) = \mu$ and $Var(X) = \sigma^2$ (e.g. blood pressure of high school athletes).

We want to make some inference about the population mean, to do this we can construct a 95% confidence interval based on our observed sample average:

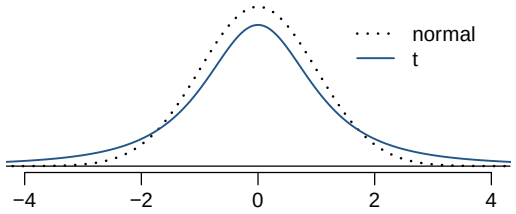
$$CI_{95\%} = \bar{X} \pm Z^\star \, SE = \bar{X} \pm 1.96 \, \frac{\sigma}{\sqrt{n}}$$

Anyone see a problem here?

The t distribution

When working with samples the population standard deviation is almost always unknown, this is addressed by using a new distribution - the *t distribution*.

- We must estimate the standard error using the sample standard deviation, this adds uncertainty to anything else were are doing.
- This distribution also is bell shape, but its tails are *thicker* than the normal distribution.
- Observations are more likely to fall beyond two SDs from the mean than with the normal distribution.
- These thick tails are helpful for resolving our problem with a less reliable estimate of the standard error (using s instead of σ)



History of the t distribution

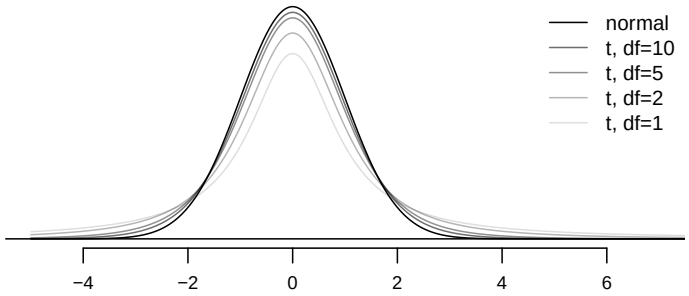
First discovered by William Gosset ...

- Oxford Graduate with a degree in Chemistry and Mathematics
- Hired by the Guinness Brewery in 1899
- Spent 1906 - 1907 studying with Karl Pearson
- Published "The probable error of a mean" in 1908 under the pseudonym "A. Student"
- Much of his work was promoted by R.A. Fisher

Properties of the t distribution

The t distribution ...

- is always centered at zero, like the standard normal (Z) distribution.
- has a single parameter, *degrees of freedom* (*df*), which dictates the thickness of the tails.



- as *df* increases the t distribution converges to the unit normal distribution.

Finding probabilities

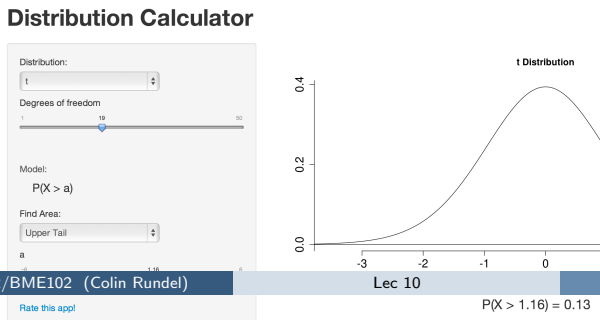
As before the probabilities are calculated as the area under the tail of a distribution, but here we use the fatter tailed t distribution. Here we calculate $P(T > 1.16)$ where $T \sim t_{df=19}$.

- Using R:

```
1-pt(1.16,df=19)

## [1] 0.1302092
```

- Using a web applet (`http://bit.ly/dist_calc`):



Finding Probabilities - t table

Locate the T value on the appropriate *df* row, obtain the probability from the corresponding column heading (one or two tail).

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Finding the p-value (cont.)

$$P(T > 1.16) > 0.10$$

$$P(T < -1.16 \text{ or } T > 1.16) > 0.20$$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	17	1.33	1.74	2.11	2.57	2.90
	18	1.33	1.73	2.10	2.55	2.88
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

CLT vs. t

From the Central Limit Distribution we have,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Since σ is unknown we are modifying the later such that we have

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

Implications of t distribution for Confidence intervals

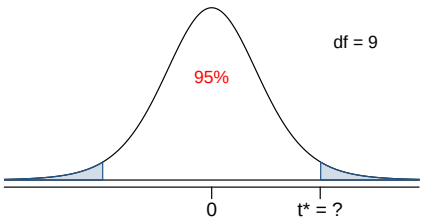
Confidence intervals are always of the form

$$\text{point estimate} \pm CV \times SE$$

If our point estimate is a sample mean and σ is unknown, then our sample mean follows a t distribution (and not a Z distribution), the critical value is then given by $t^\star_{df}$ (as opposed to a $Z^\star$) and the SE is $s/\sqrt{n}$ (and not $\sigma/\sqrt{n}$).

$$\bar{X} \pm t^\star_{df} \times \frac{s}{\sqrt{n}}$$

Finding the critical t ($t^\star$)



$n = 10$, $df = 10 - 1 = 9$

$t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

Constructing a CI

We would like to calculate a 95% confidence interval for the average rental price of an apartment in Durham. We sample craigslist and find

$$\text{Rent} = \{625, 733, 895, 929, 775, 1349, 599, 749, 1020, 799, 705, 665, 1282, 1143, 1209, 500, 1495, 1076, 975, 879\}$$

$$\bar{X} = 920.1 \qquad s = 271 \qquad n = 20 \qquad SE = s/\sqrt{n} = 60.6$$

$$\begin{aligned} CI &= \bar{X} \pm t^\star_{df} \times SE \\ &= 920.1 \pm 2.26 \times 60.6 \\ &= 920.1 \pm 137 \\ &= \textit{(783.1, 1057.1)} \end{aligned}$$

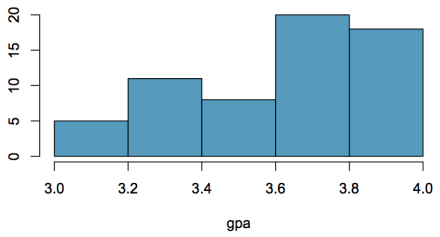
In context - we are 95% confident that the true average rental price in Durham is between (783.1, 1057.1) dollars per month.

Example - NCSSM

The College Board reports that the mean combined SAT score in the United States in 2010 was 1509. Recently 25 randomly selected graduating seniors at the North Carolina School of Science and Mathematics were asked about their combined SAT scores. Among these students the average combined SAT score was 1879 with a standard deviation of 160. Based on this sample can was infer that NCSSM students have a significantly higher combined SAT average than US students as a whole?

Example - Grade Inflation

In 2001 the average GPA of students at Duke University was 3.37. Last semester 63 introductory statistics students reported their GPA on an in class survey. The mean was 3.58, and the standard deviation 0.53. A histogram of the data is shown below.



Assuming that this sample is random and representative of all Duke students, do these data provide convincing evidence that the average GPA of Duke students has changed over the last decade and a half?

Example - Fair Dice

Imagine you are going to roll a die 100 times and record the average value of the rolls, under what circumstances should you conclude that the die is not fair at a 95% confidence level? Hint - be careful with your choice of critical value.

Recap: Inference using CIs for sample means

If σ is unknown, then $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a t distribution with $df = n-1$ when the CLT holds.

Conditions (same as CLT):

- independence of observations (often verified by a random sample, and if sampling without replacement, $n < 10\%$ of population)
- sample size is large (> 30 is usually reasonable) or population not overly skewed or heavy/light tailed

Confidence interval:

$$\bar{X} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}, \text{ where } df = n-1$$

Error Rate:

$$1 - \text{Confidence level}$$