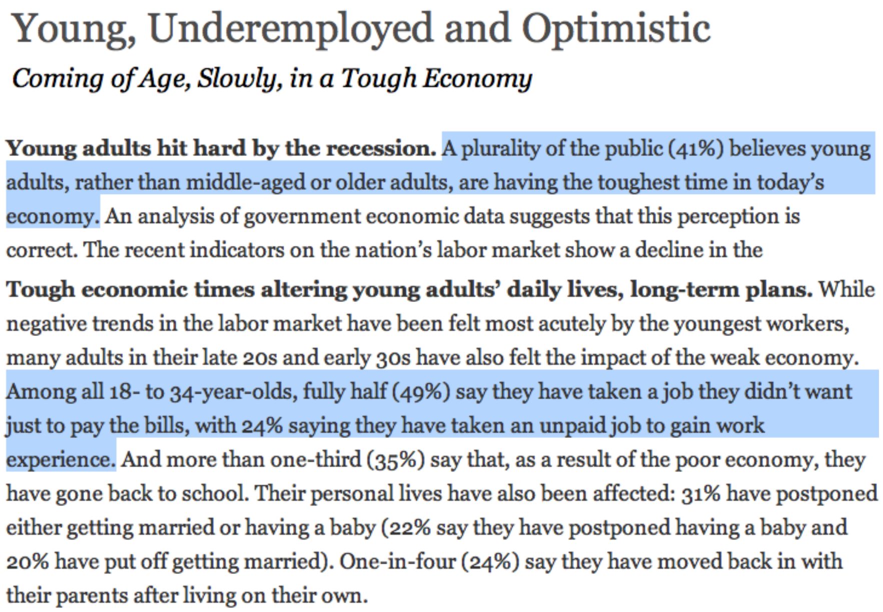
Lecture 9 - Sampling Distributions and the CLT

Sta102/BME102

Colin Rundel

September 22, 2014



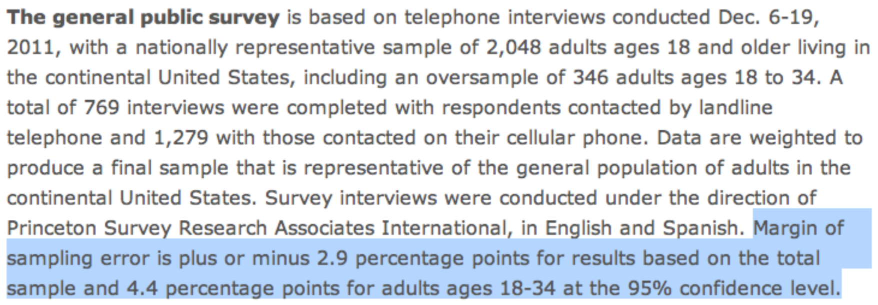
Young, Underemployed and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy. An analysis of government economic data suggests that this perception is correct. The recent indicators on the nation's labor market show a decline in the

Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience. And more than one-third (35%) say that, as a result of the poor economy, they have gone back to school. Their personal lives have also been affected: 31% have postponed either getting married or having a baby (22% say they have postponed having a baby and 20% have put off getting married). One-in-four (24%) say they have moved back in with their parents after living on their own.

Margin of error



The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States, including an oversample of 346 adults ages 18 to 34. A total of 769 interviews were completed with respondents contacted by landline telephone and 1,279 with those contacted on their cellular phone. Data are weighted to produce a final sample that is representative of the general population of adults in the continental United States. Survey interviews were conducted under the direction of Princeton Survey Research Associates International, in English and Spanish. Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.

Mean

- *Sample mean* ($\bar{x}$) -

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Population mean* (μ) -

$$\mu = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The sample mean is a *sample statistics*, or a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population) it is usually a good guess.

Variance

- *Sample Variance* (s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- *Population Variance* (σ^2) -

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Similarly, the sample variance is a *sample statistics*, or a *point estimate* of the population variance.

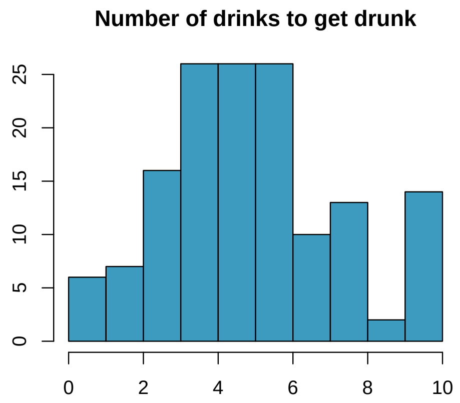
Parameter estimation

- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the *margin of error* associated with our point estimate.
- But before we get to quantifying the variability among samples, let's try to understand how and why point estimates vary from sample to sample.

> Suppose we randomly sample 1,000 adults from each state in the US. Would you expect the sample means to be the same, somewhat different, or very different?

Estimate the avg. # of drinks it takes to get drunk

We would like to estimate the average (self reported) number of drinks it takes a person get drunk, we assume that we have the population data:



Estimate the avg. # of drinks it takes to get drunk (cont.)

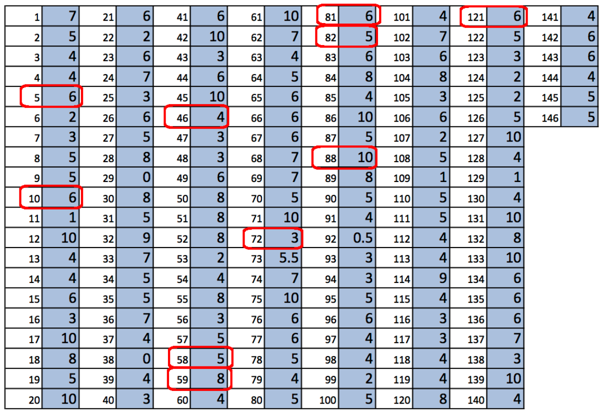
- Sample, with replacement, ten respondents and record the number of drinks it takes them to get drunk.
  - Use RStudio to generate 10 random numbers between 1 and 146

    ```
    sample(1:146, size = 10, replace = TRUE)
    ```

  - If you don't have a computer, ask a neighbor to generate a sample for you.
- Find the sample mean, round it to 1 decimal place, and report it using your clicker.

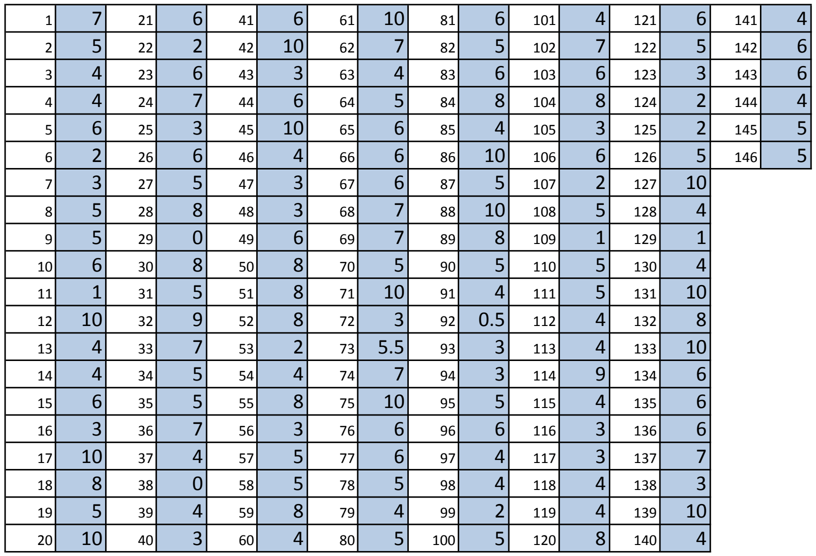
Estimate the avg. # of drinks it takes to get drunk (cont.)

```
sample(1:146, size = 10, replace = TRUE)
## [1] 59 121 88 46 58 72 82 81 5 10
```



1	7	21	6	41	6	61	10	81	6	101	4	121	6	141	4
2	5	22	2	42	10	62	7	82	5	102	7	122	5	142	6
3	4	23	6	43	3	63	4	83	6	103	6	123	3	143	6
4	4	24	7	44	6	64	5	84	8	104	8	124	2	144	4
5	6	25	3	45	10	65	6	85	4	105	3	125	2	145	5
6	2	26	6	46	4	66	6	86	10	106	6	126	5	146	5
7	3	27	5	47	3	67	6	87	5	107	2	127	10		
8	5	28	8	48	3	68	7	88	10	108	5	128	4		
9	5	29	0	49	6	69	7	89	8	109	1	129	1		
10	6	30	8	50	8	70	5	90	5	110	5	130	4		
11	1	31	5	51	8	71	10	91	4	111	5	131	10		
12	10	32	9	52	8	72	3	92	0.5	112	4	132	8		
13	4	33	7	53	2	73	5.5	93	3	113	4	133	10		
14	4	34	5	54	4	74	7	94	3	114	9	134	6		
15	6	35	5	55	8	75	10	95	5	115	4	135	6		
16	3	36	7	56	3	76	6	96	6	116	3	136	6		
17	10	37	4	57	5	77	6	97	4	117	3	137	7		
18	8	38	0	58	5	78	5	98	4	118	4	138	3		
19	5	39	4	59	8	79	4	99	2	119	4	139	10		
20	10	40	3	60	4	80	5	100	5	120	8	140	4		

$$(8 + 6 + 10 + 4 + 5 + 3 + 5 + 6 + 6 + 6)/10 = 5.9$$



1	7	21	6	41	6	61	10	81	6	101	4	121	6	141	4
2	5	22	2	42	10	62	7	82	5	102	7	122	5	142	6
3	4	23	6	43	3	63	4	83	6	103	6	123	3	143	6
4	4	24	7	44	6	64	5	84	8	104	8	124	2	144	4
5	6	25	3	45	10	65	6	85	4	105	3	125	2	145	5
6	2	26	6	46	4	66	6	86	10	106	6	126	5	146	5
7	3	27	5	47	3	67	6	87	5	107	2	127	10		
8	5	28	8	48	3	68	7	88	10	108	5	128	4		
9	5	29	0	49	6	69	7	89	8	109	1	129	1		
10	6	30	8	50	8	70	5	90	5	110	5	130	4		
11	1	31	5	51	8	71	10	91	4	111	5	131	10		
12	10	32	9	52	8	72	3	92	0.5	112	4	132	8		
13	4	33	7	53	2	73	5.5	93	3	113	4	133	10		
14	4	34	5	54	4	74	7	94	3	114	9	134	6		
15	6	35	5	55	8	75	10	95	5	115	4	135	6		
16	3	36	7	56	3	76	6	96	6	116	3	136	6		
17	10	37	4	57	5	77	6	97	4	117	3	137	7		
18	8	38	0	58	5	78	5	98	4	118	4	138	3		
19	5	39	4	59	8	79	4	99	2	119	4	139	10		
20	10	40	3	60	4	80	5	100	5	120	8	140	4		

`http://bit.ly/Sta102_CLT`

Sampling distribution

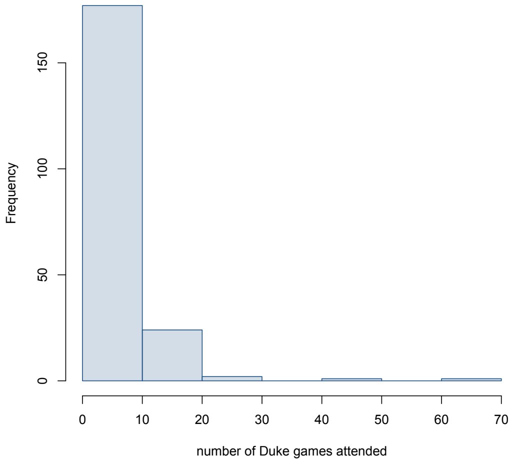
What we just constructed is called a *sampling distribution*.

What is the shape and center of this distribution.

Based on this distribution what do you think is the true population average?

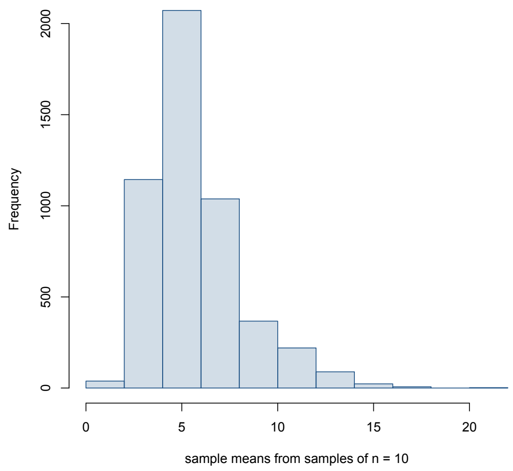
Average number of Duke games attended

Next let's look at the population data for the number of basketball games attended by a class of Duke students:



Average number of Duke games attended (cont.)

Sampling distribution, n = 10:

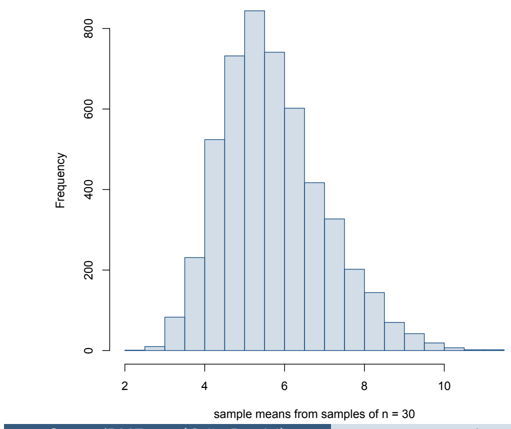


What does each observation in this distribution represent?

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Average number of Duke games attended (cont.)

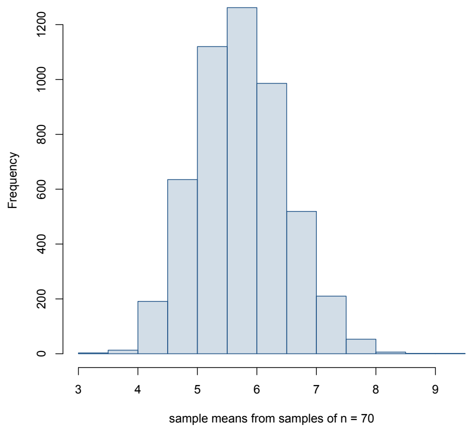
Sampling distribution, n = 30:



How did the shape, center, and spread of the sampling distribution change going from $n = 10$ to $n = 30$?

Average number of Duke games attended (cont.)

Sampling distribution, n = 70:



Sums of iid Random Variables

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} D$ where D is some probability distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

We define $S_n = X_1 + X_2 + \cdots + X_n$

Average of iid Random Variables

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} D$ where D is some probability distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

We define $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n = S_n/n$ then

Central Limit Theorem

Central limit theorem - S_n

The distribution of the *sum* of n independent and identically distributed random variables is well approximated by a normal model:

$$S_n \sim N\left(\mu = n\, E(X_i),\; \sigma^2 = n\; Var(X_i)\right)$$

when n is large.

Central limit theorem - $\bar{x}$

The distribution of the *average* of n independent and identically distributed random variables is therefore well approximated by a normal model:

$$\bar{x} \sim N\left(\mu = E(X_i),\; \sigma^2 = Var(X_i)/n\right)$$

when n is large.

CLT - Conditions

Certain conditions must be met for the CLT to apply:

1. *Independence:* Sampled observations must be independent and identically distributed.

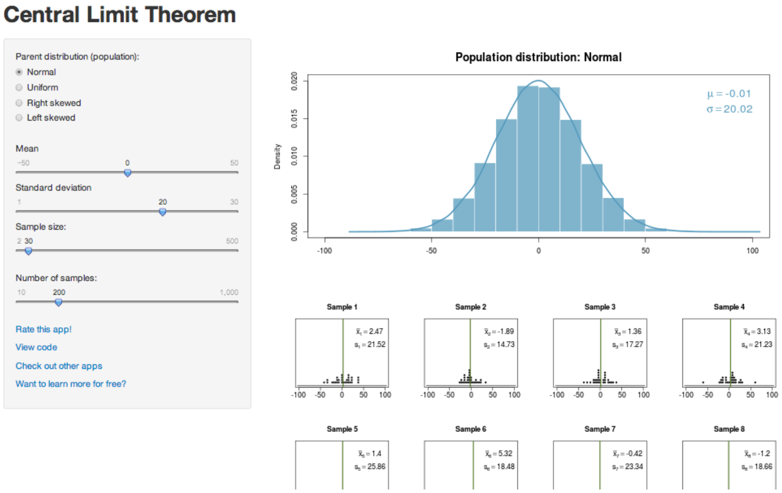
   This is difficult to verify, but is usually reasonable if

   - random sampling/assignment is used, and
   - $n < 10\%$ of the population.

2. *Sample size/skew:* the population distribution must be nearly normal *or* $n > 30$ and the population distribution is not extremely skewed.

   This is also difficult to verify for the population, but we can check it using the sample data, and assume that the sample mirrors the population.

CLT - Simulation



Central Limit Theorem

Parent distribution (population):

● Normal
○ Uniform
○ Right skewed
○ Left skewed

Mean: 0 (−50 to 50)

Standard deviation: 20 (1 to 30)

Sample size: 30 (2 to 500)

Number of samples: 200 (10 to 1,000)

Rate this app!

View code

Check out other apps

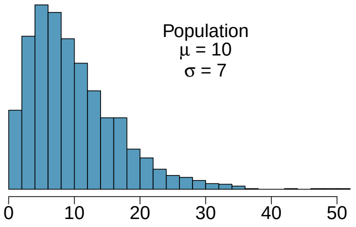
Want to learn more for free?

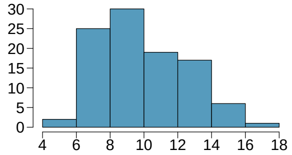
`http://bit.ly/clt_mean`

Review

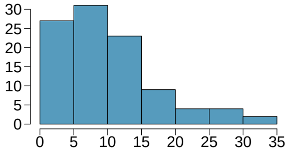
To the right is a plot of a population distribution. Match each of the following descriptions to one of the three plots below.

1. a single random sample of 100 observations from this population
2. a distribution of 100 sample means from random samples with size 7
3. a distribution of 100 sample means from random samples with size 49

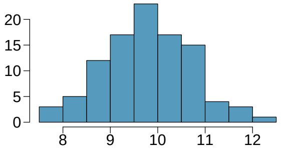




Plot A



Plot B



Plot C

Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a point estimate to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we are more likely to miss. If we toss a net in that area, we have a better chance of catching the fish.

- If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

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 $N(\mu,\ \sigma/\sqrt{n})$.

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 of μ.

Example - Relationships

A sample of 50 Duke students were asked how many long term exclusive relationships they have had. The sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

The 96% confidence interval is defined as

$$\textit{point estimate} \pm 2 \times SE$$

$$\bar{x} = 3.2 \qquad s = 1.74 \qquad SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

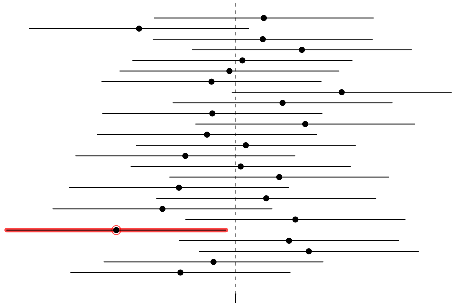
$$\begin{aligned} \bar{x} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (3.15, 3.25) \end{aligned}$$

We are 96% confident that Duke students on average have been in between 3.15 and 3.25 exclusive relationships

What does 96% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* ± 2 × *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

A more accurate interval

Confidence interval, a general formula

$$point\ estimate \pm Z^\star \times SE$$

Conditions when the point estimate = $\bar{x}$:

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - $n <$ 10% of population
2. *Sample size / skew*: $n \geq 30$ and distribution not extremely skewed

Note: We'll talk about what happens when $n < 30$ after the midterm.

Changing the confidence level

$$point\ estimate \pm Z^\star \times SE$$

- In order to change the confidence level all we need to do is adjust $Z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $Z^\star = 1.96$.
- Using the Z table it is possible to find the appropriate $Z^\star$ for any confidence level.

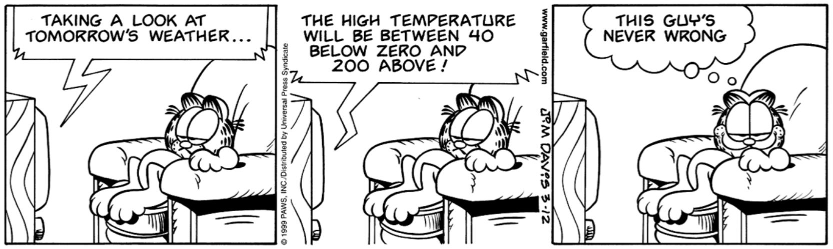
Example - Calculating $Z^\star$

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

Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JIM DAVIS 3-12

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.