Unit 3: Foundations for inference

1. Variability in estimates and CLT

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

Prof. White

Slides posted at *https://www2.stat.duke.edu/courses/Summer18/sta104.001-1/*

Announcements

- Problem set 2 is due tomorrow (11:55 pm)
- Performance Assessment (PA) 2 is due tomorrow (11:55 pm)
- No class on Monday (Memorial day)

Sample statistics vary from sample to sample

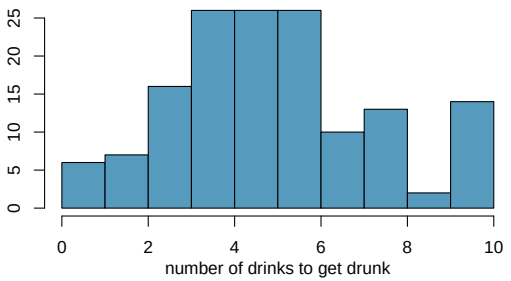
- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample 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 of their ages to be the same, somewhat different, or very different?

We would like to estimate the average number of drinks it takes students to get drunk.

- We will assume that our population is comprised of 146 students.
- Assume also that we don't have the resources to collect data from all 146, so we will take a sample of size $n = 10$.

If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?



- Sample, with replacement, ten student IDs:

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

```
[1]  59 121  88  46  58  72  82  81   5  10
```

- Find the students with these IDs:

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		

- Calculate the sample mean:
$(8 + 6 + 10 + 4 + 5 + 3 + 5 + 6 + 6 + 6)/10 = 5.9$

Activity: Creating a sampling distribution

Repeat this in teams, and report your sample mean.

1. Sample, with replacement, ten student IDs:

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

2. Find the students with these IDs:

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		

3. Calculate the sample mean, round it to 2 decimal places, and submit it using your clicker. Submit once per sample!

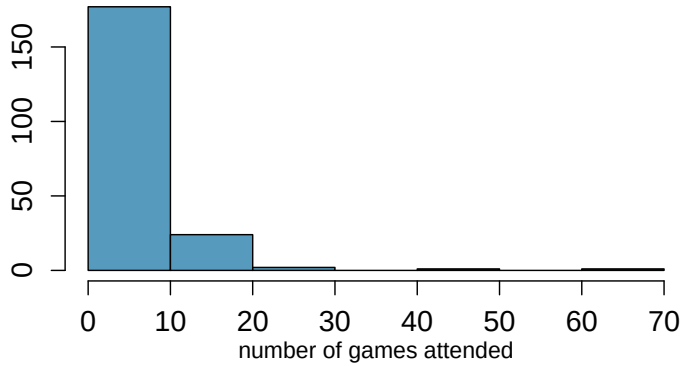
Sampling distribution

What you 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?

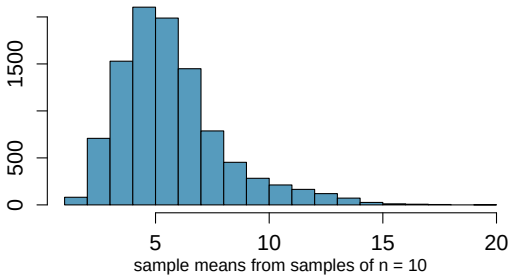
Average number of Duke games attended

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



Average number of Duke games attended (cont.)

Sampling distribution, n = 10:

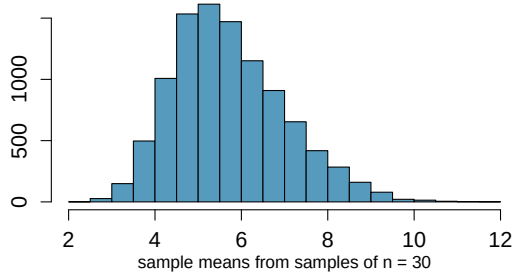


What does each observation in this distribution represent?

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

Average number of Duke games attended (cont.)

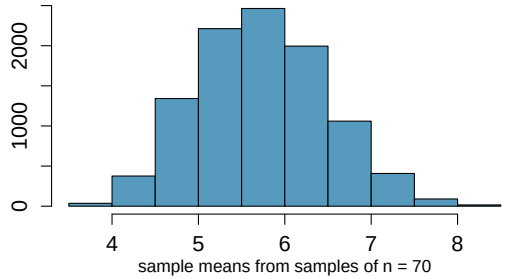
Sampling distribution, n = 30:



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:



Average number of Duke games attended (cont.)

Clicker question

The mean of the sampling distribution is 5.75, and the standard deviation of the sampling distribution (also called the *standard error*) is 0.75. Which of the following is the most reasonable guess for the 95% confidence interval for the true average number of Duke games attended by students?

(a) 5.75 ± 0.75
(b) $5.75 \pm 2 \times 0.75$
(c) $5.75 \pm 3 \times 0.75$
(d) cannot tell from the information given

2. CLT describes the shape, center, and spread of sampling distributions

Under the right conditions, the distribution of the sample means is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

A cheat: If σ is unknown, use s.

- So it wasn't a coincidence that the sampling distributions we saw earlier were symmetric.
- We won't go into the proving why $SE = \frac{\sigma}{\sqrt{n}}$, but note that as n increases SE decreases.
- As the sample size increases we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower.

3. CLT only applies when independence and sample size/skew conditions are met

1. *Independence:* Sampled observations must be independent.

   This is difficult to verify, but is more likely if
   - random sampling/assignment is used, and,
   - if sampling without replacement, $n < 10\%$ of the population.
2. *Sample size/skew:* Either
   - the population distribution is normal or
   - $n > 30$ and the population dist. is not extremely skewed, or
   - $n >> 30$ (approx. gets better as n increases).

   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.

3. CLT only applies when independence and sample size/skew conditions are met

Amongst other things, the central limit theorem is useful for

- constructing confidence intervals and
- conducting hypothesis tests.

Clicker question

Which of the below visualizations is not appropriate for checking the shape of the sample distribution of a numerical variable, and hence the population?

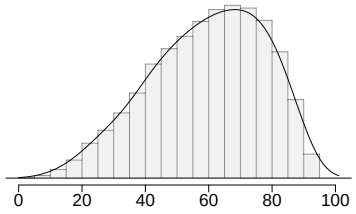
(a) histogram
(b) boxplot
(c) normal probability plot
(d) mosaicplot

Summary of main ideas

1. Sample statistics vary from sample to sample
2. CLT describes the shape, center, and spread of sampling distributions
3. CLT only applies when independence and sample size/skew conditions are met

Clicker question

Four plots: Determine which plot (A, B, or C) is which.
(1) At top: distribution for a population ($\mu = 60, \sigma = 18$),
(2) a single random sample of 500 observations from this population,
(3) a distribution of 500 sample means from random samples with size 18,
(4) a distribution of 500 sample means from random samples with size 81.

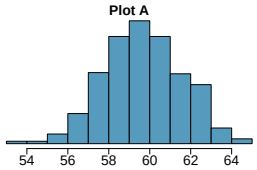


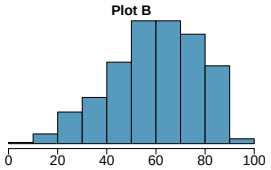
(a) (2) - B; (3) - A; (4) - C

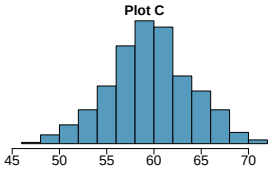
(b) (2) - A; (3) - B; (4) - C

(c) (2) - C; (3) - A; (4) - D

(d) (2) - B; (3) - C; (4) - A







A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Would you expect most houses in Topanga to cost more or less than $1.3 million?
Hint: What is most likely the shape of this distribution?

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly $1.3 million with a standard deviation of $300,000. There were no houses listed below $600,000 but a few houses above $3 million.

Clicker question

Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?

(a) yes

(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

Clicker question

Can we estimate the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

(a) yes

(b) no

A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

In order to calculate $P(\bar{X} > 1.4\ mil)$, we need to first determine the distribution of $\bar{X}$. According to the CLT,

$$\bar{X} \sim N\left(mean = 1.3, SE = \frac{0.3}{\sqrt{60}} = 0.0387\right)$$

$$\begin{aligned} P(\bar{X} > 1.4) &= P\left(Z > \frac{1.4 - 1.3}{0.0387}\right) \\ &= P(Z > 2.58) \\ &= 1 - 0.9951 = 0.0049 \end{aligned}$$