Sta 111 - Summer II 2017
Probability and Statistical Inference

6. Variability in estimates

Lu Wang

Duke University, Department of Statistical Science

July 10, 2017

Outline

1. Sample statistics vary from sample to sample

2. CLT describes the shape, center, and spread of sample mean

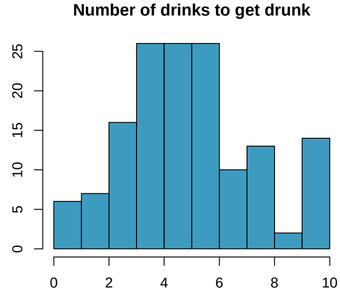
3. Homework 2

Parameter estimation

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

- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect, 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.

The following histogram shows the distribution of number of drinks it takes a group of college students to get drunk. We will assume that this is our population of interest. We would like to estimate the average number of drinks it takes these college students to get drunk.



- If we randomly select observations from this data set, which values are most likely to be selected, which are least likely?
- Suppose that you don't have access to the population data. You might sample 10 observations from the population and use your sample mean as the best guess for the unknown population mean.

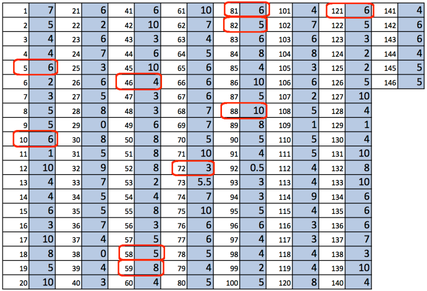
Sample mean

- 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:

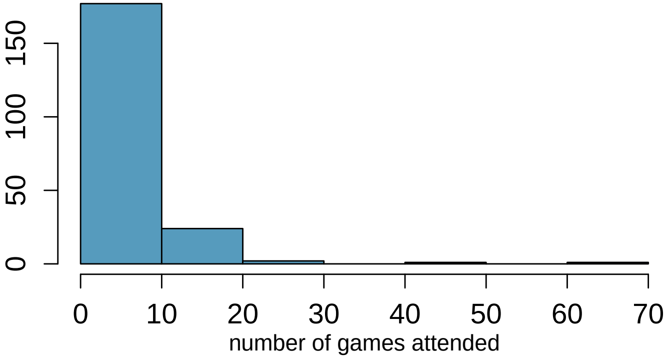
$$(8 + 6 + 10 + 4 + 5 + 3 + 5 + 6 + 6 + 6)/10 = 5.9$$

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

Average number of Duke games attended

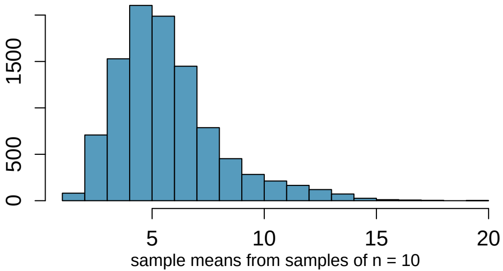
Next let's check how sample means vary from sample to sample.

The histogram below shows population data for the number of Duke basketball games attended:



Average number of Duke games attended (cont.)

- Repeatedly draw 10 observations from the population and compute the sample mean.
- Histogram of sample means from lots of samples with $n = 10$:

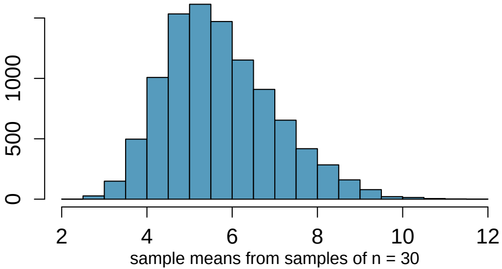


What does each observation in this distribution represent?

Is the variability of the sample mean smaller or larger than the variability of the population distribution?

Average number of Duke games attended (cont.)

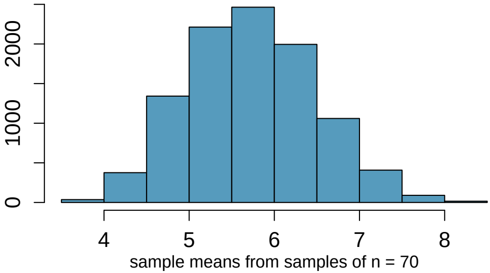
Go to samples with size n = 30. Histogram of sample means:



How does the shape, center, and spread of the distribution of sample means change going from $n = 10$ to $n = 30$?

Average number of Duke games attended (cont.)

Go to samples with size n = 70. Histogram of sample means:



Central Limit Theorem

Under certain conditions, the distribution of the sample mean is well approximated by a normal distribution:

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

We use *SE* to denote the SD of sample statistics. A cheat: If σ is unknown, use *s*.

- So it wasn't a coincidence that the distribution of the sample mean we saw earlier was symmetric.
- Note that $SE = \frac{\sigma}{\sqrt{n}}$ implies that as *n* increases *SE* decreases.
- Intuitively as the sample size increases we would expect samples to yield more accurate sample means, hence the variability among the sample means would be lower.

CLT - conditions

Certain conditions must be met for the CLT to apply:

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 if the population distribution is skewed, the sample size is large.
   - the more skewed the population distribution, the larger sample size we need for the CLT to apply
   - for moderately skewed distributions, $n > 30$ is a widely used rule of thumb

We won't go into proving why $\bar{x}$ has a normal approximation, but a simple proof of why $SE = SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$.

Practice

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. 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,

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

- If you only have access to the CDF of $N(0,1)$, transform $\bar{x}$ to a Z score:

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

Homework 2

Graded questions:

- Ch 4: 4.8, 4.12, 4.24, 4.32

Practice questions:

- Variability in estimates and the Central Limit Theorem: 4.1, 4.3, 4.5, 4.33, 4.35, 4.37, 4.41
- Confidence intervals: 4.9, 4.11, 4.13, 4.15
- Hypothesis tests: 4.17, 4.19, 4.23, 4.25, 4.27