Unit 3: Foundations for inference
1. Variability in estimates and CLT

Sta 104 - Summer 2015

Duke University, Department of Statistical Science

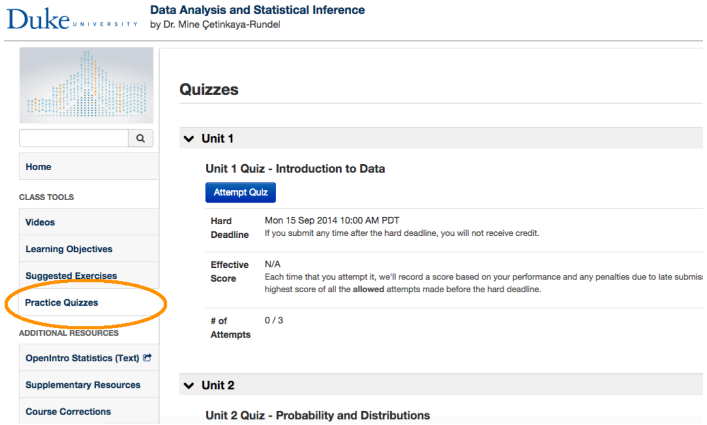
May 26, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bit.ly/sta104su15

Announcements

- Review session 1-2pm today
- Review materials posted on course website + review quizzes on Coursera (not graded)



Data Analysis and Statistical Inference
by Dr. Mine Çetinkaya-Rundel

Quizzes

Unit 1

Unit 1 Quiz - Introduction to Data

Hard Deadline: Mon 15 Sep 2014 10:00 AM PDT
If you submit any time after the hard deadline, you will not receive credit.

Effective Score: N/A
Each time that you attempt it, we'll record a score based on your performance and any penalties due to late submis highest score of all the **allowed** attempts made before the hard deadline.

# of Attempts: 0 / 3

Unit 2

Unit 2 Quiz - Probability and Distributions

RA 3

- Individual - 15 mins
- Team - 10 mins

Sample statistics vary from sample to sample

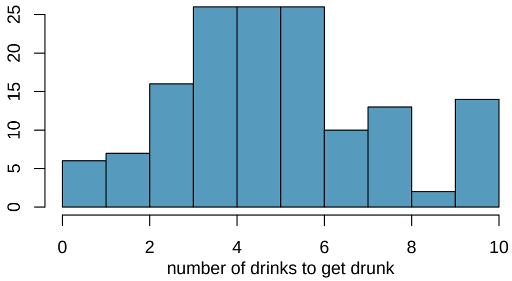
- We are often interested in *population parameters*.
- Since complete populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample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

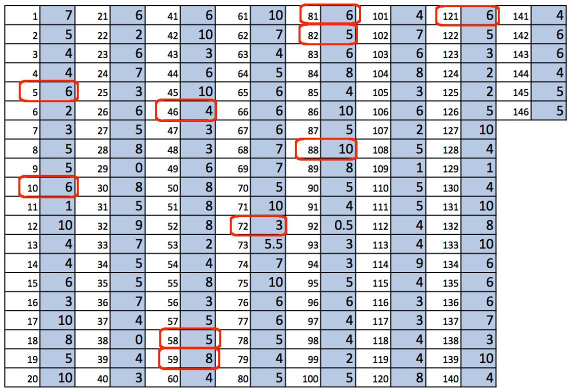


- Sample, with replacement, ten student IDs:

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

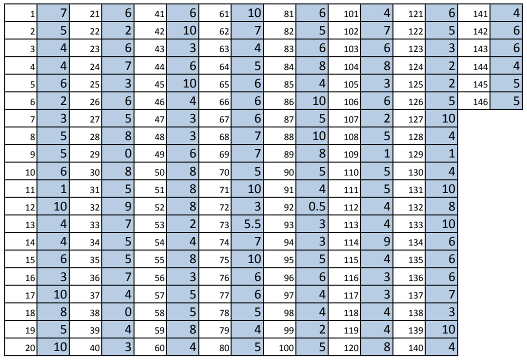
Activity: Creating a sampling distribution

Repeat this, 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 report to me.

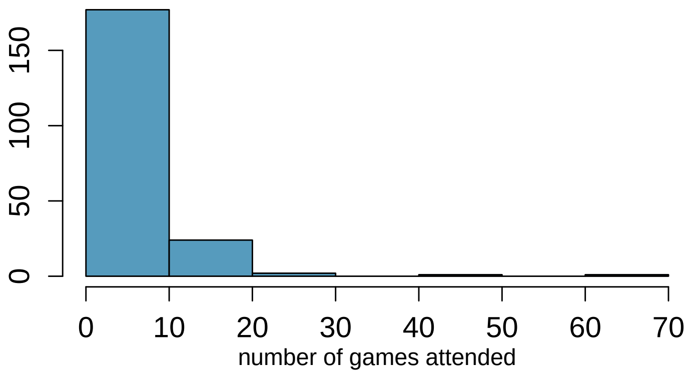
Sampling distribution

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?

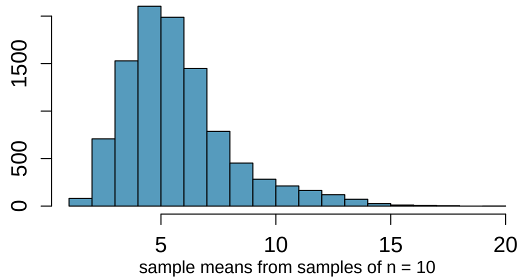
Average number of Duke games attended

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:

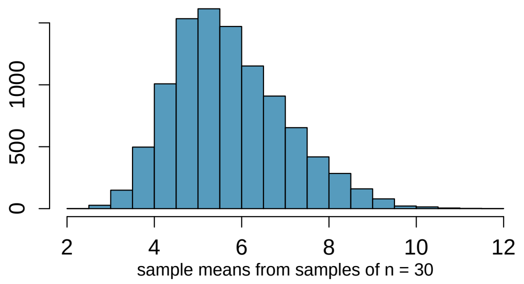


What does each observation in this distribution represent?

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

Average number of Duke games attended (cont.)

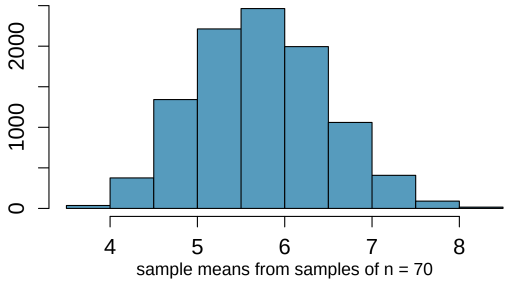
Sampling distribution, n = 30:



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:

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

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 distribution of the sample, and hence the population?

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

4. Statistical inference methods based on the CLT depend on the same conditions as the CLT

Always check these in context of the data and the research question!

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:* Population must be normal or sample size must be large.

5. Use confidence intervals to estimate population parameters

$$CI : point\ estimate \pm margin\ of\ error$$

If the parameter of interest is the population mean, and the point estimate is the sample mean,

$$\bar{x} \pm Z^{\star} \frac{s}{\sqrt{n}}$$

Application exercise: 3.1 Confidence interval for a single mean

See course website for details.

6. Critical value depends on the confidence level

Clicker question

What is the critical value ($Z^\star$) for a confidence interval at the 91% confidence level?

(a) $Z^\star = 1.34$

(b) $Z^\star = 1.65$

(c) $Z^\star = 1.70$

(d) $Z^\star = 1.96$

(e) $Z^\star = 2.33$

Common misconceptions about confidence intervals

1. The confidence level of a confidence interval is the probability that **the specific confidence interval you construct with data from a single sample** contains the true population parameter.
   The confidence level is equal to the proportion of random samples that result in confidence intervals that contain the true population parameter.

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.

7. Use hypothesis tests to make decisions about population parameters

Hypothesis testing framework:

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu =$ *null value*
   - $H_A : \mu <$ or $>$ or $\neq$ *null value*
2. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Sample size / skew: $n \geq 30$ (or larger if sample is skewed), no extreme skew
3. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \text{ where } SE = \frac{s}{\sqrt{n}}$$

4. Make a decision, and interpret it in context of the research question
   - If p-value $< \alpha$, reject H_0, data provide evidence for H_A
   - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A

Application exercise: 3.2 Hypothesis testing for a single mean

See course website for details.

Clicker question

Which of the following is the correct interpretation of the p-value from App Ex 3.2?

(a) The probability that average GPA of Duke students has changed since 2001.

(b) The probability that average GPA of Duke students has not changed since 2001.

(c) The probability that average GPA of Duke students has not changed since 2001, if in fact a random sample of 63 Duke students this year have an average GPA of 3.58 or higher.

(d) The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher, if in fact the average GPA has not changed since 2001.

(e) The probability that a random sample of 63 Duke students have an average GPA of 3.58 or higher or 3.16 or lower, if in fact the average GPA has not changed since 2001.

Common misconceptions about hypothesis testing

1. P-value is the probability that the null hypothesis is true
   A p-value is the probability of getting a sample that results in a test statistic as or more extreme than what you actually observed (in the direction of H_A, if in fact H_0 is correct. It is a conditional probability, conditioned on H_0 being correct.

2. A high p-value confirms the null hypothesis.
   A high p-value means the data do not provide convincing evidence for H_A and hence that H_0 can't be rejected.

3. A low p-value confirms the alternative hypothesis.
   A low p-value means the data provide convincing evidence for H_A, but not necessarily that it is confirmed.

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
4. Statistical inference methods based on the CLT depend on the same conditions as the CLT
5. Use confidence intervals to estimate population parameters
6. Critical value depends on the confidence level
7. Use hypothesis tests to make decisions about population parameters