STAB22

Chong Chen
University of Toronto, Scarborough
Department of Computer and Mathematical Sciences

2019

1 Sampling distribution

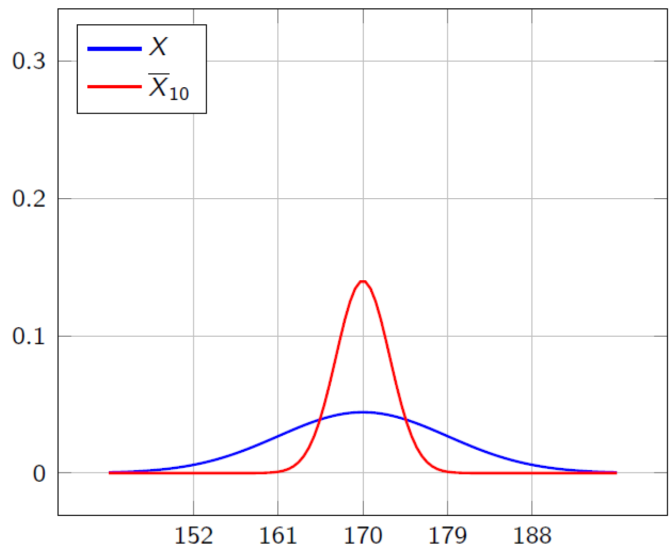
- Statistics(sample count, sample mean $\bar{X}$, sample proportion $\hat{p}$) can change from one sample to another.
- They are random.
- Therefore they have a distribution, called the **sampling distribution**.

1.1 Experiment

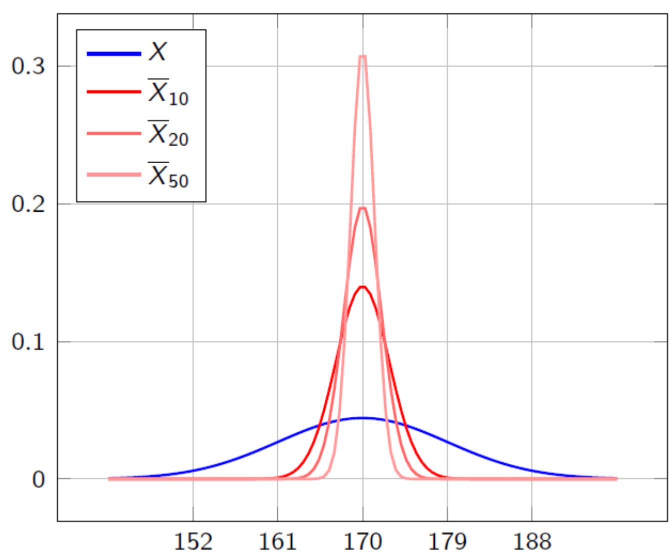
- Generate $n = 10$ values X_1, X_2, ... , X_{10} independently from $N(\mu = 270, \sigma = 9)$
- Compute the sample mean.
- Repeat 10000 times.

The results are shown below:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	$\overline{X}_{10}$
Sample 1	164.4	162.8	172.1	175.6	168.0	174.7	173.1	161.0	156.0	169.1	167.7
Sample 2	171.7	160.5	172.2	178.0	173.2	165.6	173.7	170.2	171.8	166.5	170.3
Sample 3	162.5	160.7	164.2	166.2	171.4	180.2	171.1	166.0	179.4	189.2	171.1
					...						
Sample 10,000	184.4	159.3	152.6	167.3	168.9	180.9	155.7	177.6	170.6	175.0	169.2



X
$\overline{X}_{10}$
0.3
0.2
0.1
0
152
161
170
179
188



X
$\overline{X}_{10}$
$\overline{X}_{20}$
$\overline{X}_{50}$
0.3
0.2
0.1
0
152
161
170
179
188

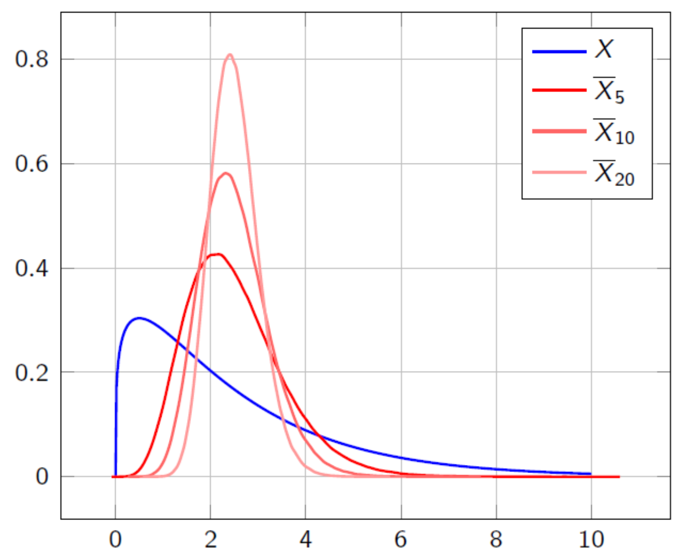
1.2 Summary

If X follows the normal distribution $N(\mu, \sigma)$, then

- $\bar{X}_n$ is normally distributed,
- with mean μ
- with standard deviation(sd) $<$ than σ.
- the sd of $\bar{X}_n$ decreases as n increases.
  That is, $\bar{X}_n$ estimates μ with greater precision when n is larger.

REMARK: **Central limit theorem**

Draw a Simple Random Sample of size n from any population with mean μ and finite standard deviation σ. When n is large enough, the sampling distribution of the sample mean $\bar{X}$ follows the normal distribution $(\mu, \frac{\sigma}{\sqrt{n}})$.



1.3 Normal Approximation for Counts and Proportions

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes. When n is large, the sampling distributions of these statistics are approximately Normal:

X is approximately $N(np.\sqrt{np(1-p)})$,

$\hat{p}$ is approximately $N(p, \sqrt{\frac{p(1-p)}{n}})$

As a rule of thumb, we will use this approximation for values of n and p that satisfy: $np \geq 10$ and $n(1-p) \geq 10$.

1.4 Example

A sample of size $n = 25$ is drawn from a population with mean 40 and sd 10. What is the probability that the sample mean will be between 36 and 44? (Assume that the Central Limit Theorem applies.)

1.5 Example

Suppose that the weights of airline passengers are known to have a distribution with a mean of 75kg and a sd of 10kg. A certain plane has a passenger weight capacity of 7700kg. What is the probability that a flight of 100 passengers will exceed the capacity?

1.6 Sampling Distribution of Sample Proportions

Sampling distribution, as well as Central limit theorem, can also be applied to sample proportions.

Draw an SRS of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample and $\hat{p} = X/n$ be the sample proportion of successes. When n is large, the sampling distributions of these statistics are approximately Normal:

$\hat{p}$ follows the normal distribution $(p, \sqrt{\frac{p(1-p)}{n}})$.

2 Confidence Intervals for Proportions

An interval estimate is an interval of numbers (around the point estimate) which is highly likely to contain the parameter value. Interval estimate is also known as **confidence interval**.

Confidence interval for the population proportion p is given by

$$\hat{p} \pm z^* * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where standard error of the estimate

$$SE(\hat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

also margin of error of the CI is

$$z^* * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the critical value.

For any experiment:

- p = population proportion (true probability of success, unknown)
- n = sample size (number of trials)
- Point estimate: $\hat{p}$
- CLT: Sampling distribution of $\hat{p}$ is approximately normal $(p, \sqrt{\frac{p(1-p)}{n}})$.
- Problem: p is unknown and we cannot compute the true sd of $\hat{p}$
- Solution: substitute $\hat{p}$ to find the standard error

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Remark: 95% confident interval for p:

$$[\hat{p} - 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

3 Exercises

3.1

An epidemiologist intends to study the proportion p of oral contraceptive users in a certain population (unknown). She randomly selects 200 women. In her sample 30 women out of 200 use oral contraceptive. The sample proportion is $\hat{p} = 0.15$. What is a 95% confidence interval for the proportion of oral contraceptive users in the population?

3.2

University students will vote on proposal. You ask random sample of 400 students, out of which 250 are in favor. Build 95% CI for proportion in favor.