

Inference and One-sample Test

Key two concepts

- Hypothesis testing
 - Testing your hypothetical argument based on inferential statistics.
 - Again, parameter describes population, and it is unknown.
 - So, you calculate statistics you are interested in from sample data to estimate parameter.
- Confidence level
 - With how much confidence can you substantiate your hypothesis? How likely is it that your result from sample can be generalized to wider population? This depends on two things: the standard error and the size of sample.

What are your goals here?

- You can estimate the mean of population with the mean from your sample, but with some degree of errors.

Probability sampling

- Probability sampling (vs non-probability sampling)
- We should be able to guarantee probability sampling if you are trying to do your own research, which means the probability of selecting A and the probability of B should be equal. Otherwise, your sample cannot be representative of population:
Biased!

Sampling distribution of a statistics

- What is sampling distribution and why important?
 - 1) The real distribution of a population is about a single sample. But, sampling distribution is about all possible samples with the size of n we can draw from population. (sample 1, sample 2,...)
 - 2) Sampling distribution is not interested in scores of individual case. There are usually n individuals in each sample, though the size of each sample is different in your survey in the class.
 - 3) You are interested in sampling distribution of a statistics as of now, especially sampling distribution of the mean in this chapter.

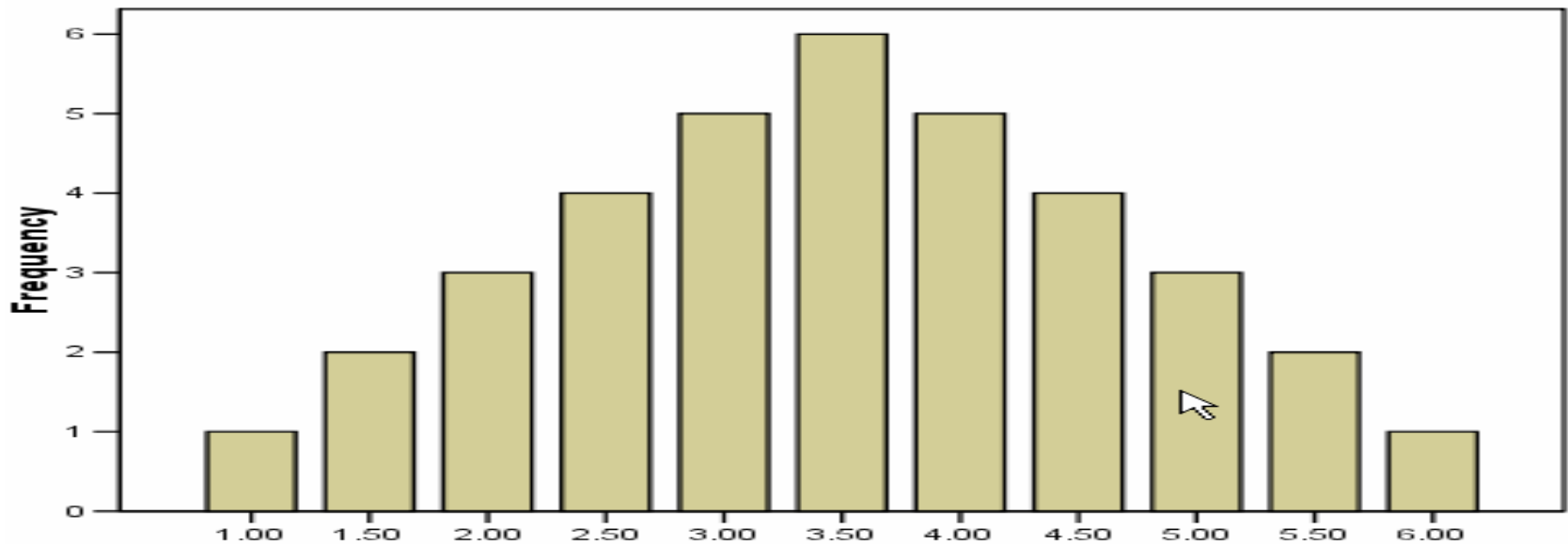
Sampling distribution of the mean

- Recall the survey you already did in the class. Here are six students. Their mid-term scores are (1, 2, 3, 4, 5, 6) Although we know the mean is 3.5, we cannot know the mean of population generally!
- From this population, let's make samples with $n=2$. Replacement is acceptable, which means you might be chosen several times. The number of samples we can draw is $6 \times 6 = 36$. (See the next page)
- (1,1) (1,2) ... (2,1)... (3,1)... (4,1)... (5,1)... (6,1)... (6,6) This is all possible sets of sample, but you might choose (1,2) sometimes, and (3,1) some other time in real situation. In other words, your sample is only one possibility, so you can't avoid some amount of error. (This leads to the concept of "standard error" later on!)
- We are interested in the mean of each sample: $(1+1)/2=1$ for the first sample, $(1+2)/2=1.5$ for the second sample,... (See the next page)

sample	mean	sample	mean	sample	Mean
1,1	1	3,1	2	5,1	3
1,2	1.5	3,2	2.5	5,2	3.5
1,3	2	3,3	3	5,3	4
1,4	2.5	3,4	3.5	5,4	4.5
1,5	3	3,5	4	5,5	5
1,6	3.5	3,6	4.5	5,6	5.5
2,1	1.5	4,1	2.5	6,1	3.5
2,2	2	4,2	3	6,2	4
2,3	2.5	4,3	3.5	6,3	4.5
2,4	3	4,4	4	6,4	5
2,5	3.5	4,5	4.5	6,5	5.5
2,6	4	4,6	5	6,6	6

VAR0001

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	2.8	2.8	2.8
	1.50	2	5.6	5.6	8.3
	2.00	3	8.3	8.3	16.7
	2.50	4	11.1	11.1	27.8
	3.00	5	13.9	13.9	41.7
	3.50	6	16.7	16.7	58.3
	4.00	5	13.9	13.9	72.2
	4.50	4	11.1	11.1	83.3
	5.00	3	8.3	8.3	91.7
	5.50	2	5.6	5.6	97.2
	6.00	1	2.8	2.8	100.0
	Total		36	100.0	100.0



Comparison of population distribution and sampling distribution of the mean

- Surprisingly, the mean of sampling distribution of the mean is also 3.5. In general, both are always the same. ($\mu = \mu_x$)
- This implies that your sample mean could be used to estimate population mean. (Unbiased estimator)

- How about standard deviation of population (1, 2, ... 6), say σ , and standard deviation of sampling distribution's mean (1.0, 1.5, ... 6.0), say σ_x ? In general, when the sample size is n , $\sigma_x = \sigma / \sqrt{n}$. This is called "standard error." (Exactly saying, $n-1$)
- However, we cannot actually know about the standard deviation of population, as the same with the mean of population. This is why we use "s" (standard deviation in your sample) instead of σ_x .
- Finally, standard error of the mean is s / \sqrt{n} .

Standard Error, why important?

- The standard error (of the mean) tells us about how far your sample mean is from the true score of population mean. (The mean vary from sample to sample by SE)
- Notice that standard error decreases, as sample size n increases.
- Also notice that higher variability of population (or your sample), standard error increases.
- Put together, if you have large sample and lower dispersion of population (or your sample), your sample mean is more likely to be around the population mean. In other words, smaller standard error.

Central Limit Theorem and Normal Distribution

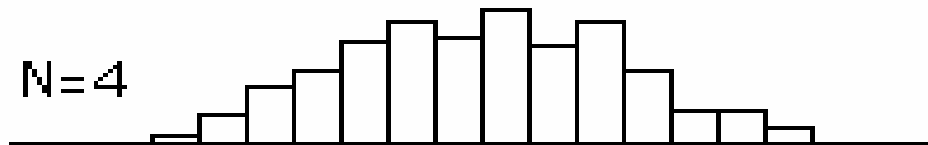
- If your population is normal, then “your sampling distribution of the mean” is also normal. (not just “sampling distribution”!)
- Central limit theorem
- When your population is not normal, as long as your sample size is big enough, then your sampling distribution of the mean is approximately normal distribution.
- Population distribution (μ, σ^2) ; Sampling distribution of the mean $(\mu, \sigma^2/n)$

- How large? Usually, $n > 30$.

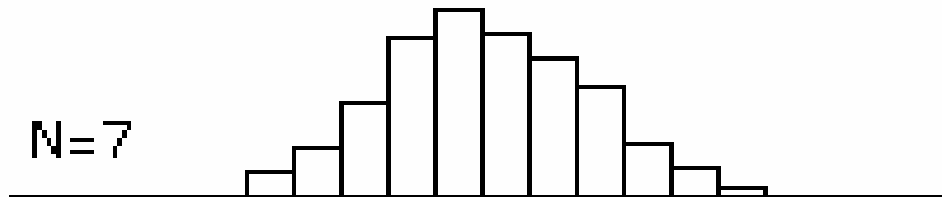
$N = 1$



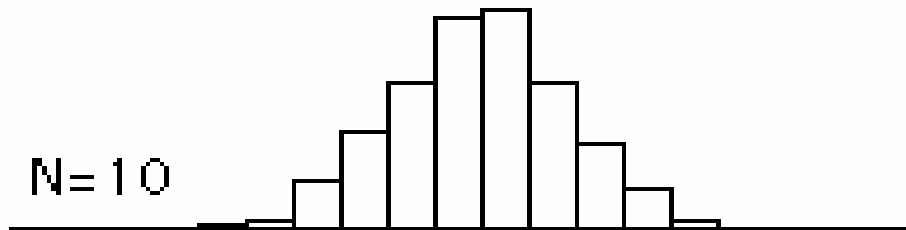
$N = 4$



$N = 7$



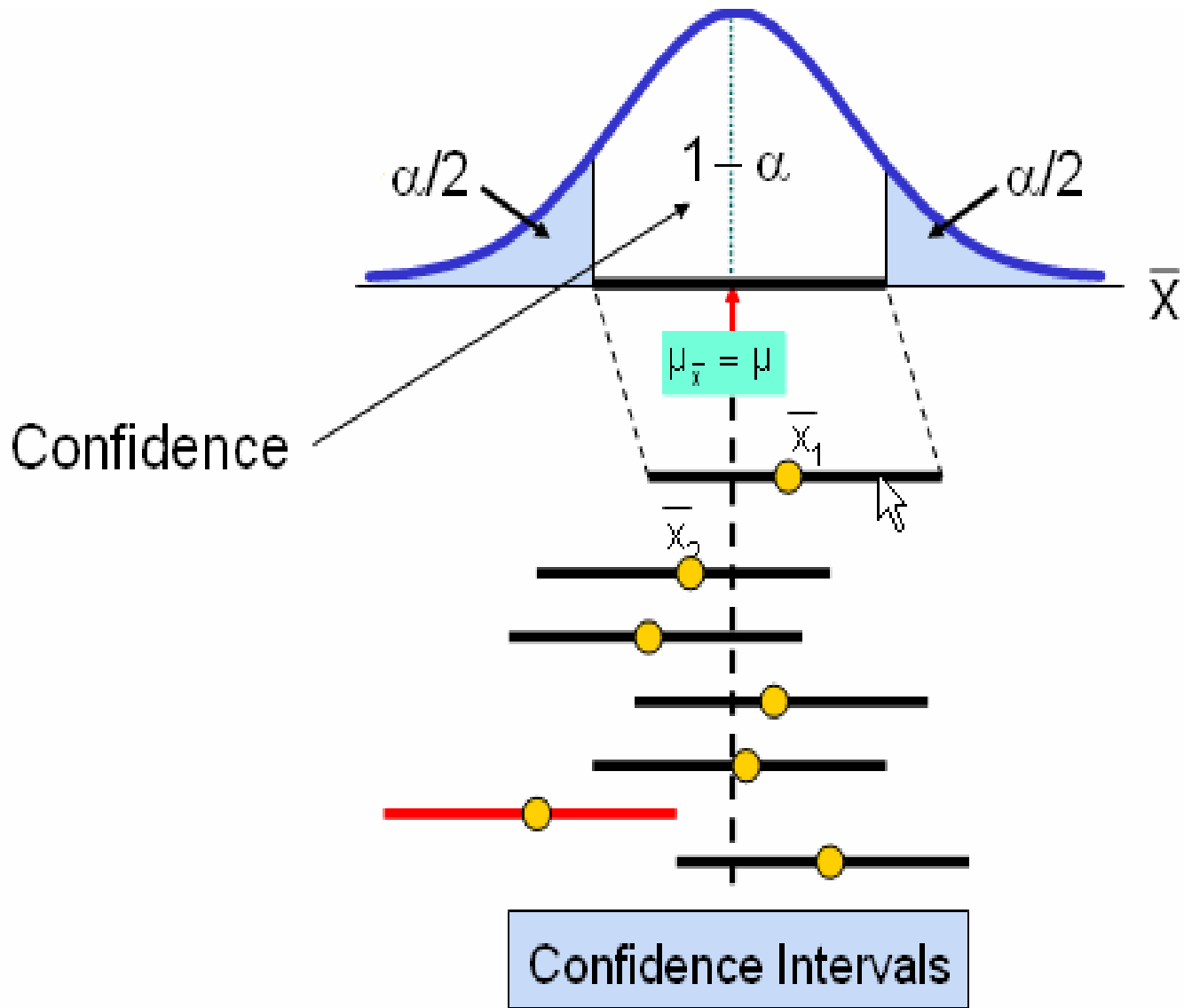
$N = 10$

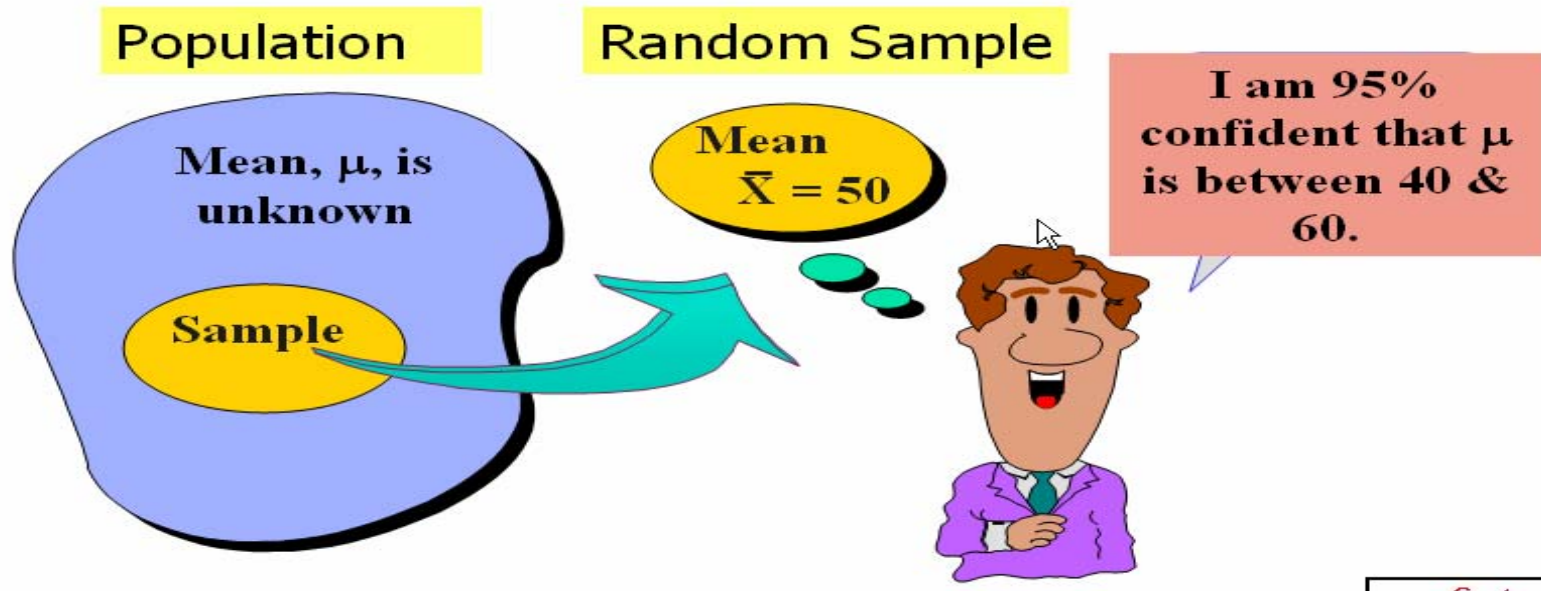


Confidence Interval for Population Mean

- A range of values around a point estimate that makes it possible to state the probability that an interval contains the population parameter between its lower and upper bounds.
- The margin of error can be obtained by replacing σ by s because we don't know about standard deviation of population in most of cases, as I said

before.
$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$





- What does his saying mean in statistical sense? The mean of his one sample is 50. He can use this mean as statistics to estimate population mean based on inferential statistics theory. He is sure about his guess with the confidence level of $95\% = 100(1-\alpha)\%$. $\alpha=0.05$. The normal distribution is symmetric, so z is 1.96 in the z -score table when $\alpha/2=0.025$. His estimation falls between $50-10$ and $50+10$. Therefore, 20 is the 95% confidence interval width, which is $1.96 \times (s/\sqrt{n})=10$. Wow! 5% is the chance that your guess is wrong!