Chapter 7. Inference concerning the mean (A)

March 17, 2011

Outline

Point Estimation

Interval Estimation

Tests of Hypothesis

Hypothesis Concerning One Mean

The Relation between Tests and Confidence Intervals, Power & Sample Size

Notations

- $\bar{X}$ and $\bar{x}$ (also call $\bar{X}$ an estimator of μ, and $\bar{x}$ an estimate of μ)

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} \qquad \text{a r.v. before observation}$$
$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \qquad \text{a value after observation}$$

- S and s (also call S an estimator of σ, and s an estimate of σ)

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad \text{a r.v. before observation}$$
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \text{a value after observation}$$

- S^2 and s^2 ....

  See the item above

- Z, z and z_α (Z: statistic; z statistic value; z_α critical value)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad \text{a r.v. before observation}$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \text{a value after observation}$$

$$z_\alpha \qquad \text{a critical value with } P(Z > z_\alpha) = \alpha.$$

- t, t^* and t_α (t: statistic; t^* statistic value; t_α critical value)

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \qquad \text{a r.v. before observation}$$

$$t^* = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \text{a value after observation}$$

$$t_\alpha \qquad \text{a critical value with } P(t > t_\alpha) = \alpha.$$

- F, F^* and F_α, χ^2, χ^* and χ^2_α.

1 Point Estimation

Estimator

- Definition: An estimator is a rule, usually expressed as a formula, that tells us how to calculate an estimate based on information in the sample.
- Estimators are used in two different ways:
  - Point estimation: Based on sample data, a single number is calculated to estimate the population parameter.
    * The rule or formula that describes this calculation is called the point estimator.

- * the resulting number is called a point estimate.

- Interval estimation: Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie.

- Example:
  - In order to estimate the average waiting time, μ, for a bus of a student attending ST2334, the lecturer asked 4 students their waiting time for a bus, the results are

$$x_1 = 6, x_2 = 1, x_3 = 4, x_4 = 9$$

  - Use $\bar{x} = 5$ to estimate μ.
  - $\bar{X}$ is the estimator, the computed value 5 is the estimate

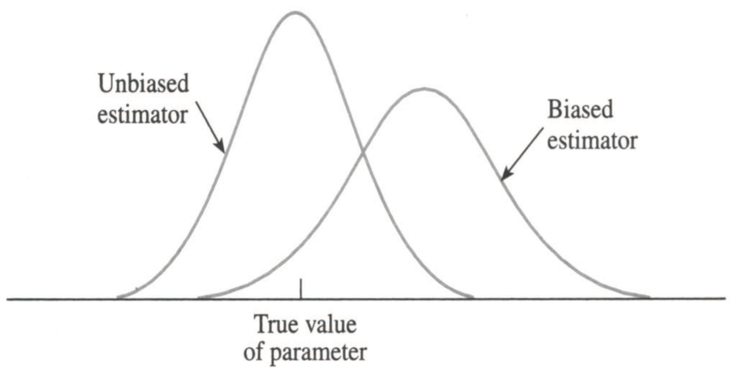
2 Point Estimation of a Mean

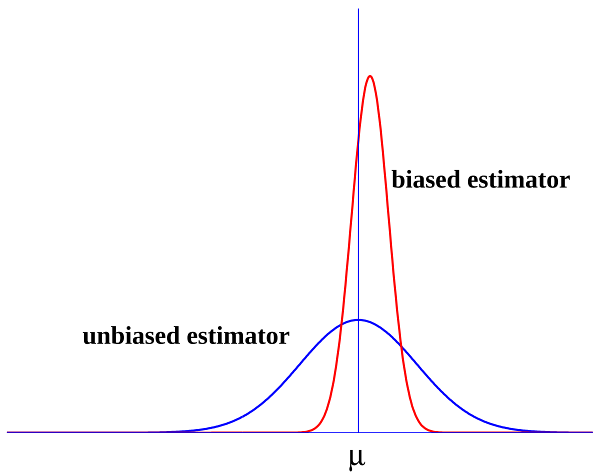
- Parameter of interest: μ, the population mean.
- Data: A random sample $X_1, X_2, ..., X_n$.
- Estimator: $\bar{X}$.
- The population standard error (SE) of $\bar{X}$ is $\sigma/\sqrt{n}$.
- Use S to estimate σ, then the estimator of the standard error of $\bar{X}$ is $S/\sqrt{n}$.

Unbiased Estimator

- Note that an estimator is a random variable. E.g.: $\bar{X}$.
- Bear in mind that $\bar{X}$ is to estimate μ.
- Refer to Theorem 6.1, $E(\bar{X}) = \mu$.
- More generally,
  - θ: the parameter of interest. for example p, μ, σ
  - $\hat{\theta}$ an estimator of θ. It is a random variable based upon the sample.
  - If $E(\hat{\theta}) = \theta$, $\hat{\theta}$ is called unbiased estimator of θ.
  - Figure Illustration.

The following figure shows distributions for biased and unbiased estimators.





- Unbiased estimator may not be better than biased estimator

Examples of Unbiased Estimators

- If a bus arrives at the bus stop every θ (unknown) mins. The lecturer wants to estimate θ, so, this morning, he randomly selected 4 students and asked their waiting time for a bus. X_1, X_2, X_3, X_4 are obtained.
  - Is $\bar{X}$ an unbiased estimator of θ? NO!

$$E\bar{X} = E(X) = \frac{\theta - 0}{2} = \theta/2 \neq \theta$$

  - Find an unbiased estimator of θ. 2 $\bar{X}$ is one, since

$$E(2\bar{X}) = 2E(\bar{X}) = 2\theta/2 = \theta$$

  - Is $2\bar{X}$ the only unbiased estimator of θ? NO! e.g: $2X_1$.

- Unbiased Estimator: Yes & No

$X_1, X_2, ..., X_n$is a random sample from the same population with mean μ and variance σ^2.

  - Is $\bar{X}$ an unbiased estimator of μ ? Yes
  - Is X_1 an unbiased estimator of μ ? Yes
  - Is S^2 an unbiased estimator of σ^2? Yes
  - Is S an unbiased estimator of σ? No
  - Is $\sum_{i=1}^{n}(X_i - \bar{X})^2/n$ an unbiased estimator of σ? No
  - If μ is a known value, $\sum_{i=1}^{n}(X_i - \mu)^2/n$ an unbiased estimator of σ^2?

Comparing Unbiased Estimators

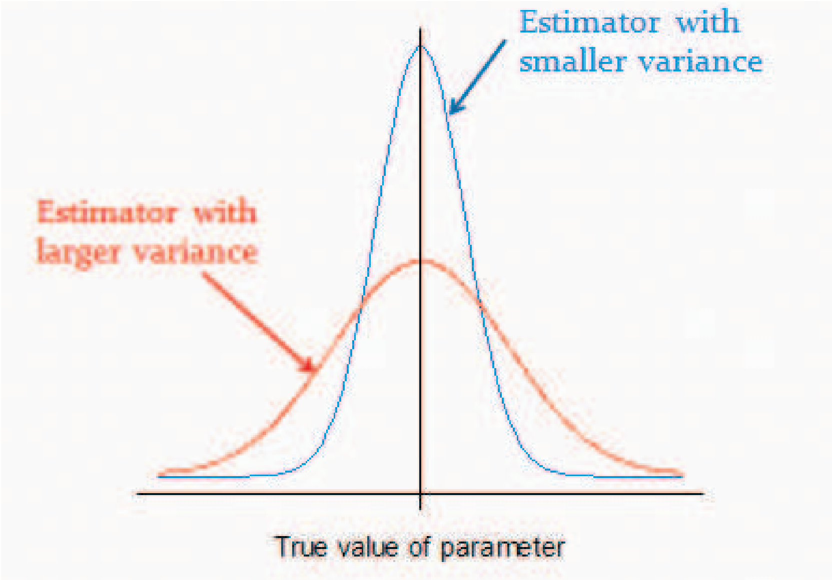
Usually we can find more than one possible unbiased estimators of a parameterθ.

So which one is better?

- More efficient unbiased estimator is better
- $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of θ.
- If $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$, and ...
- "$<$" is true for at least one value of θ
- $\hat{\theta}_1$ is said to be more efficient (better) estimator than $\hat{\theta}_2$.

Figure Illustration

The following figure compares estimator variability



3 Maximum Error of Estimate

- Usually $\bar{X} \neq \mu, \bar{X} - \mu$ measures the difference between the estimator and the true value of the parameter. If the population is normal or n is large, refer to Theorem 6.1 and 6.3, then

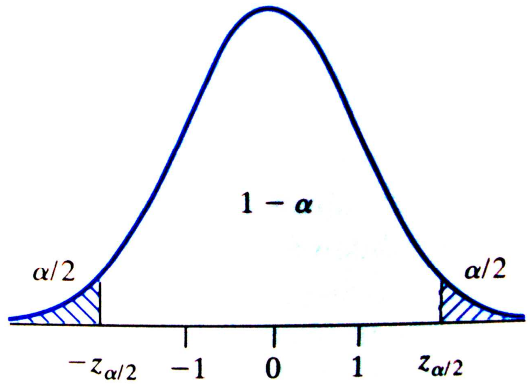
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- either has standard normal or approximately follows a standard normal distribution.

  – With probability equal to $1 - \alpha$,

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

– See the following figure



– That is, with probability equal to $1-\alpha$,

$$\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \le z_{\alpha/2}$$

- Or with probability equal to $1-\alpha$, $|\bar{X}-\mu|$ is less than

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

which is called Maximum Error of Estimate.

- **Meaning of E**: with probability $1-\alpha$, the error $\bar{X}-\mu$ is at most E.

- E increases with probability $1-\alpha$ and decreases with sample size n

Example: Maximum Error of Estimate

- An investigator is interested in the possibility of merging the capabilities of television and the Internet. He proposes to conduct a random sample of n $=$ 50 internet users to poll about the time they spend watching television per week.

- If, based on historical experience, the investigator can assume that $\sigma = 3.5$ hours.

- He proposed to use the mean time of the sample to estimate the population mean time Internet users spend watching television (unbiased estimator of μ).
- What can he assert with probability 0.99 about the maximum error of estimation?
- Now that $n = 50(\geq 30)$ is a large value, $\sigma = 3.5$ and checking from Table 3, $z_{0.01/2} = 2.575$. Use the formula for E:

$$E$$

- Conclusion: the investigator can assert with probability 0.99 that his error will be at most 1.275 hours.

Determination of Sample Size

- What is the minimum n such that with probability (or with confidence, if data available) $1 - \alpha$, the error is at most E. Recall that

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Solve for n, we have

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Example: Determination of Sample Size

- Example: refer to the television example, what is the sample size n required such that the investigator can assert with 99% probability (or confidence when data are collected) that his estimation error is at most 0.5 hour?

- Now that $z_{0.01/2} = 2.575,\ E = 0.5, \sigma = 3.5$:

$$n = \left(\frac{2.575 \cdot 3.5}{0.5}\right)^2$$

Other cases

σ	n	Population	Statistic	E	with E and α given sample size n needed
known	S. or L.	Normal	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$
unknown	S.	Normal	$t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{\alpha/2}s}{E}\right)^2$
known	L.	any	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$
unknown	L.	any	$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}s}{E}\right)^2$

when $n \geq 30$, it is large (L.), otherwise, it is small (S.)

Example: σ Unknown, Data Normal, ~~n Small~~

- Refer to the television example. Still, n = 50 internet users to poll about the time they spend watching per week.

- But now, if σ is unknown.

- After collecting the data, the sample standard deviation can be computed, $s = 3.6$.

- $n = 50, s = 3.6, z_{0.01/2} = 2.575,$

$$E = z_{\alpha/2}\frac{s}{\sqrt{n}} =$$

4 Interval Estimation

- Usually, $\bar{X} \neq \mu$, the error $\bar{X} - \mu$ can be quantified by E.

- **Confidence Interval (C.I.)** approach estimates μ by an interval estimator.

- Interval Estimator: a rule for calculating two numbers, say a and b, that create an interval [a, b] that you are fairly certain that it contains the parameter of interest. a and b are usually functions of $X_1, X_2, ..., X_n$.

- The concept of "fairly certain" can be quantified using a statistical concept called the **degree of confidence** and designated by $(1 - \alpha)$.

$$P(a < \mu < b) = 1 - \alpha$$

i.e. the probability for the random interval [a,b] to cover μ is $(1-\alpha)$.

- Definition: [a,b] is called the confidence interval; the probability that a confidence interval will contain the estimated parameter is called the **degree of confidence** (or **confidence level**).

- We discuss several cases in constructing the C.I.

Case I: σ Known, Data Normal

- Assumption: $X_1, X_2, ..., X_n$is a random sample from a normal population and σ is known, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution.

- With probability equal to $1 - \alpha$,

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

i.e.

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

- That is:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

- When sampled data is available, $\bar{x}$ can be computed, then C.I. with **confidence 100(1-α)%** is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Or written as

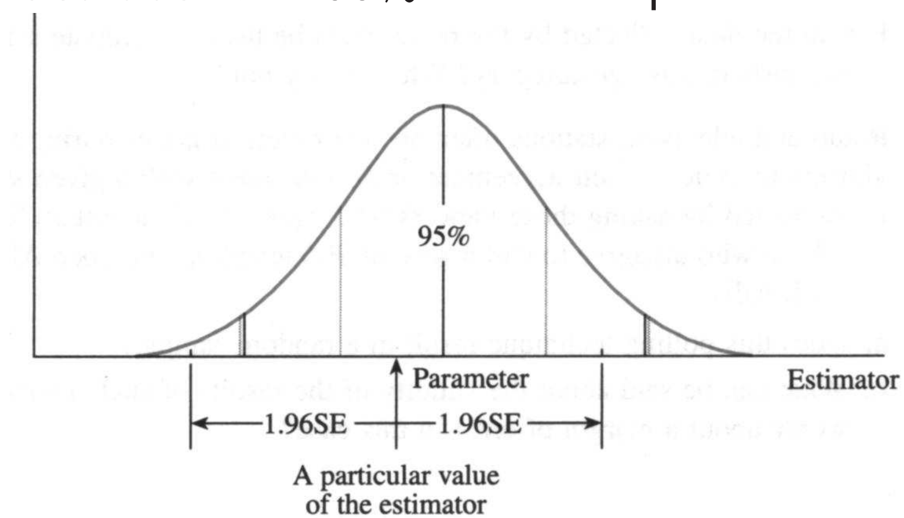
$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm E$$

• i.e. a C.I. for μ with degree of confidence 1-α or 100(1-α)% is

$$[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

where $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ are called lower confidence limit (LCL) and upper confidence limit (UCL)

A Figure to Illustrate C.I. 95% C.I. for a particular value of $\bar{x}$,



Example A computer company samples demand during lead time over 25 time periods:

235 374 309 499 253 421 361 514 462 369 394 439 348 344 330 261 374 302 466 535 386 316 296 332 334

It is known that the standard deviation of demand over lead time is 75 computers. Assuming the demands follows normal. We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels

- Thus, the parameter to be estimated is the population mean: μ
- (because the population is normal and σ is known) So our confidence interval estimator will be: $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

$\bar{x}$	370.16
$z_{\alpha/2}$	1.96
σ	75
n	25

Calculated from the data…

$1-\alpha = .95, \ \therefore \alpha/2 = .025$

$so \ \ z_{\alpha/2} = z_{.025} = 1.96$

Given

- $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 24.9$
- The lower and upper confidence limits are 340.76 and 399.56, or 95% confidence interval is [340.76, 399.56]

Television Example Revisit

- Assume that in the television example, the data (i.e. sampled n=50 internet users time of watching television per week) are normal, $\sigma = 3.5, \bar{X} = 11.5$ hours.
- Recall that we have computed E for this case: E=1.275.
- then, the 100(1-0.01)% (=99%) C.I. for μ is given by

$$\bar{x} \pm E =$$

Case II: σ Known, n Large

- Assumption: $X_1, X_2, ..., X_n$is a random sample from any population, σ is known, n is large. Then by the CLT, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$

- Use Theorem 6.3, similar arguments to Case I, we obtain 100(1-α)% C.I. for μ

- Or equivalently,

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm E$$

which shares the same form as Case I, but now, please be reminded that the interval is not exact, it is approximately true, since C.L.T. is applied.

Case III: σ Unknown, Data Normal, n Small

- Assumption: σ is unknown, $X_1, X_2, ..., X_n$is a random sample, normally distributed, n is small (optional). Use Theorem 6.4, we have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Similar arguments to Case I, we obtain 100(1-α)% C.I. for μ

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- Or equivalently $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$

- Now that σ is replaced with S, and the degrees of freedom for t-distribution is (n-1).

Case IV: σ Unknown, n Large

- Assumption: σ is unknown, $X_1, X_2, ..., X_n$is a random sample, normally distributed, n is large.

- Use Theorem 6.3,

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \approx N(0, 1)$$

- similar arguments to Case I, we obtain 100(1-σ)% C.I. for μ

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

- Or equivalently

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm E$$

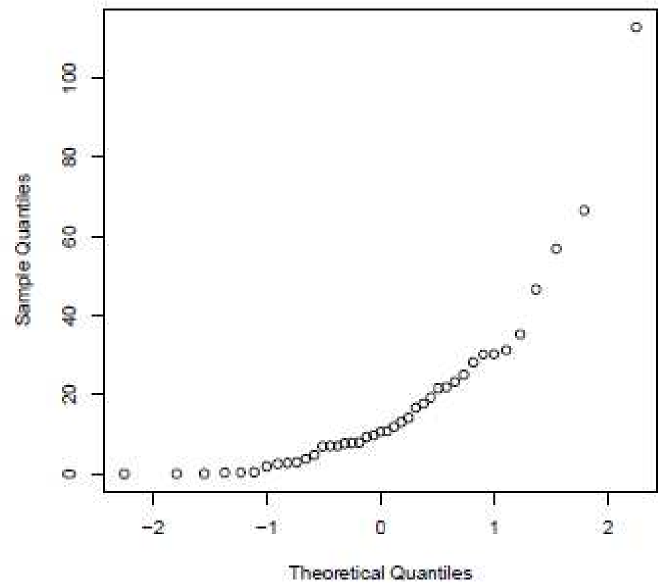
Example: Which Case to Use

– The following data set collects n=41 randomly sampled waiting times of students from ST2334 to receive reply for their email from a survey in the day time.

2.50	23.28	19.34	4.74	7.03	21.85	2.72	10.64
17.73	21.55	9.71	30.24	0.37	31.26	35.24	13.10
7.81	16.69	66.54	1.88	14.14	46.59	28.17	7.92
0.06	9.32	0.03	10.75	6.97	56.86	2.89	112.77
7.67	30.16	0.33	0.44	3.77	25.07	7.05	11.93
0.08							

– Question: construct 98% C.I. for the mean waiting time of students.

- Which Case to Use:
  - In this example, first, we should be clear that σ is unknown. Therefore there are two possible cases for us to choose from: Case III or Case IV.
  - Checking the assumptions carefully, since n=41 is quite a large number, we know that both cases can be used providing that the normal assumption is correct.
  - Normal assumption can be checked by using normal scores plot
  - Normal Scores plot is provided below: normality is not satisfied! Therefore, only Case IV can be used.



– Now that $\bar{x} = 17.736, s = 21.727, z_{0.01} = 2.33$, the 98% C.I. is given by

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \bar{x} \pm E =$$

Interpreting C.I.

- Use Case III as an example. Similar arguments apply for other cases.
- Interpretation I: the interval $[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}]$ covers the true mean μ with probability 1-α.
- Pay attention to the statement in interpretation I! The statement is made before observations are made. i.e. $\bar{X}$ and S are both random variables!

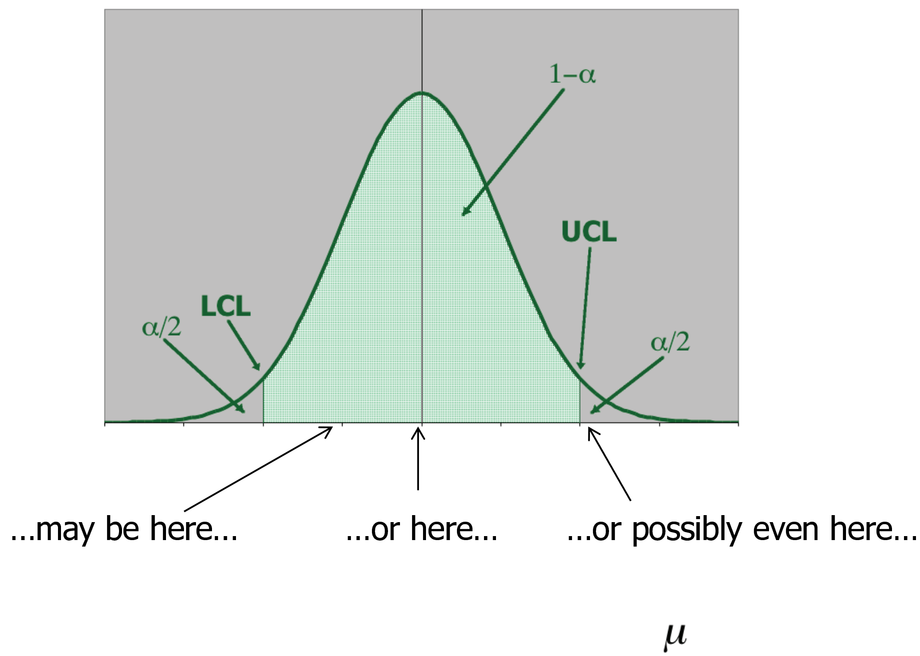
- Interpretation II: If we repeatedly sample n observations from the population, for each sampling, a confidence can be constructed, then, continue this procedure vast number of times,

$$\frac{\#\text{ of intervals that cover } \mu}{\#\text{ of intervals calculated}} \to 1 - \alpha$$

- Interpretation III: when the observations are available, we have 100(1-α)% confidence (note that it is not well-defined!) that μ will be located within the interval

$$[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}]$$

- the actual location of the population mean μ. The population mean is a fixed but unknown quantity. It's incorrect to interpret the confidence interval estimate as a probability statement about μ. The interval acts as the lower and upper limits of the interval estimate of the population mean.



Four commonly used confidence levels

Confidence Level

cut & keep handy!

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

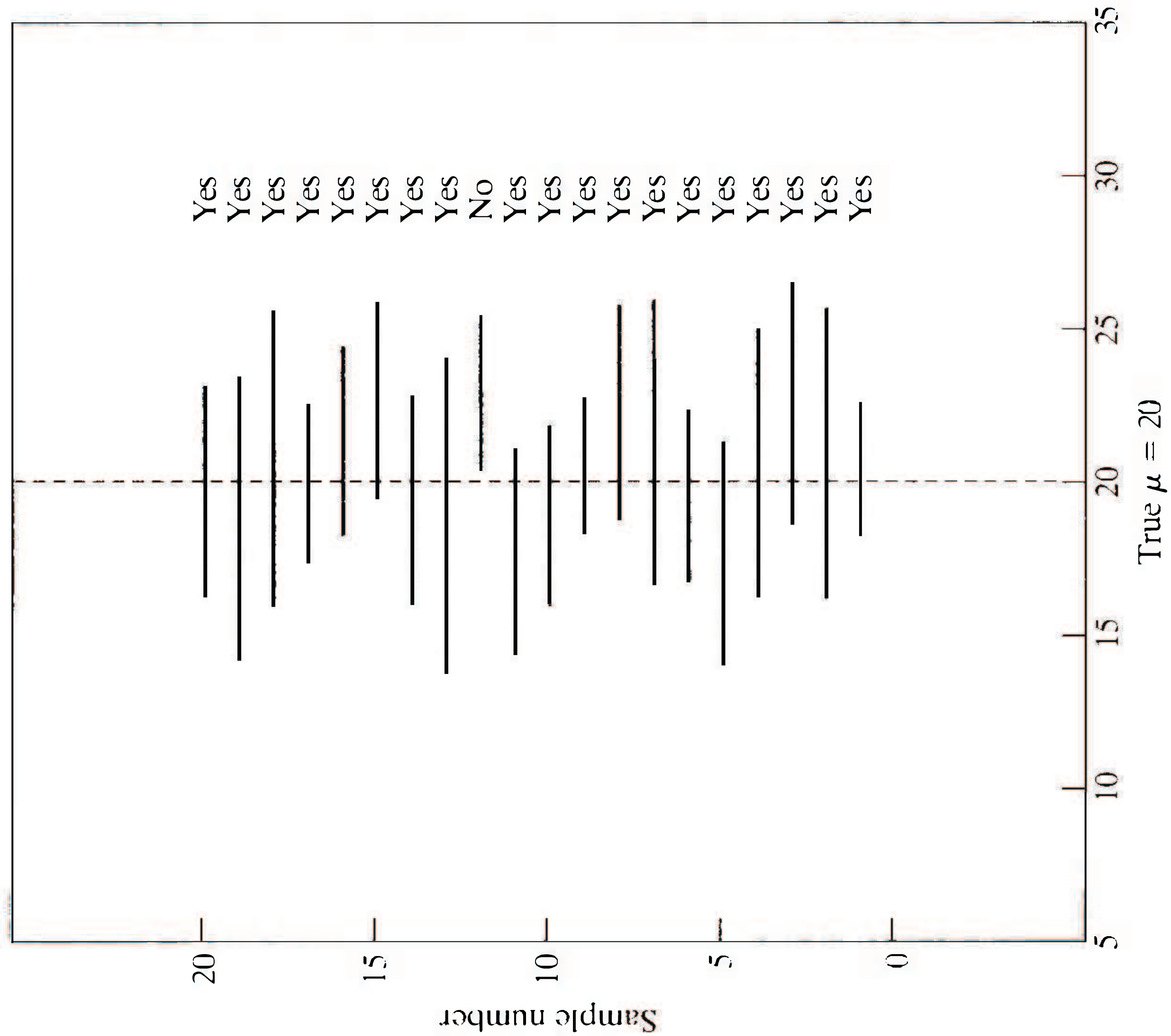
and C.I. based on t-distribution.

Example: Interpretation II

- We conduct a sample of size n $= 10$ from N($\mu = 20$, $\sigma^2 = 25$). Pretending that μ and σ are both unknown, we then construct 95% C.I. for μ based on Case III.

- The above sampling and C.I. construction procedure is repeated 20 times. That is, 20 C.I.s are constructed.

- Some contain the true $\mu = 0$, while the others don't. The results are displayed in the Figure of next slide


Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
No
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
Yes
True μ = 20
Sample number
5
10
15
20
25
30
35
0
5
10
15
20