Estimator and Confidence Interval

Point Estimation

A **point estimate** of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ of the statistic $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a point estimate of the population parameter μ. Similarly, $\hat{p} = x/n$ is a point estimate of the true proportion p for a binomial experiment.

- Estimate is not expected to be error free (sampling bias)
- There might be more than one estimate
- We have to select the right one.

Desirable properties of a "good" decision function for choosing estimator

Let $\bar{\Theta}$ be an estimator whose value $\bar{\theta}$ is a point estimate of some unknown population parameter θ. Certainly, we would like the sampling distribution of $\bar{\Theta}$ to have a mean equal to the parameter estimated.

Unbiased Estimator: A statistic $\bar{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if $E(\bar{\Theta}) = \theta$

Show that S^2 is an unbiased estimator of σ^2.

Why S^2 is an unbiased estimator of σ^2

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2$$

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})$$

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.$$

$$E(S^2)=E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i-\mu)^2-nE(\bar{X}-\mu)^2\right]$$

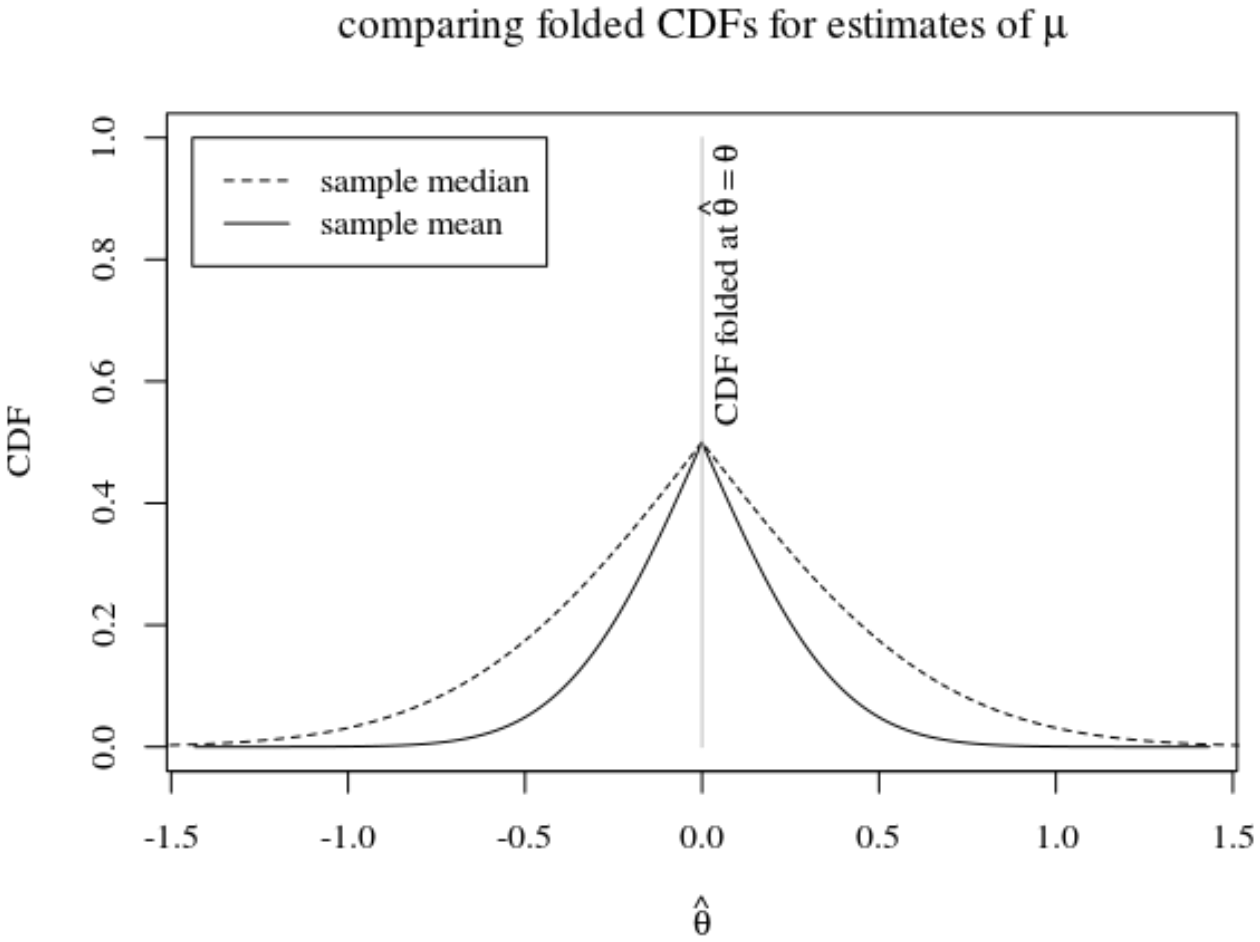
$$=\frac{1}{n-1}\left(\sum_{i=1}^{n}\text{Var}(X_i)-n\,\text{Var}(\bar{X})\right)\quad=\frac{1}{n-1}\left(n\sigma^2-n\frac{\sigma^2}{n}\right)=\sigma^2.$$

Variance of a Point Estimator

- If Θ_1 and Θ_2 are two unbiased estimators of the same population parameter θ, we want to choose the estimator whose sampling distribution has the smaller variance. Hence, if $\sigma^2_{\Theta 1} < \sigma^2_{\Theta 2}$, we say that Θ1 is a **more efficient estimator** of θ than Θ2.

Most Efficient Estimator

If we consider all possible unbiased estimators of some parameter θ, the one with the **smallest variance** is called the *most efficient estimator* of θ.



Interval Estimation

- Even the most efficient unbiased estimator is unlikely to estimate the population parameter exactly.

- There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the parameter.

- Such an interval is called an **interval estimate**.

Interval Estimation

An *interval estimate* of a population parameter θ is an interval of the form

$$\hat{\theta}_L < \theta < \hat{\theta}_U,$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value of the statistic $\hat{\Theta}$ for a particular sample and also on the sampling distribution of $\hat{\Theta}$.

Example

- SAT verbal scores for students in the entering freshman class might produce an interval from 530 to 550, within which we expect to find the true average of all SAT verbal scores for the freshman class.

- The values of the endpoints, 530 and 550, will depend on the computed sample mean $\bar{x}$ and the sampling distribution of $\bar{X}$.

- As the sample size increases, estimate is likely to be closer to the parameter μ.

Let us now look at an example.

The heights of the freshmen at UMD are supposed to follow a *normal* distribution with mean μ and standard deviation $\sigma = 10$ (in cm). A random sample of size $n = 36$ is taken, the sample mean $\overline{x} = 160$.

- Use $\overline{x} = 160$ to estimate the value of μ. This is a point estimation of μ.
- Is μ equal to 160?
- We would like to convert this point estimate into a statement, like "the value of μ is between 150 cm and 170 cm" and attached to the statement a measure of degree of confidence of it being true.

From the distribution of $\overline{X}$,

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,\ 1)$$

About 95% of the values of $\overline{X}$ are expected to fall within $2(\sigma/\sqrt{n})$ of μ, i.e.,

$$\mathrm{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Exchanging the positions of μ and $\overline{X}$,

$$\mathrm{P}\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Our sample gives $\overline{x} = 160$, then the interval is from $\left(\overline{x} - 2\frac{\sigma}{\sqrt{n}}\right)$ to $\left(\overline{x} + 2\frac{\sigma}{\sqrt{n}}\right)$, or,

$$\left(160 - 2\frac{10}{\sqrt{36}},\ 160 + 2\frac{10}{\sqrt{36}}\right)$$

or,

$$(157,\ 163)$$

This is an interval estimate of the unknown parameter of μ.

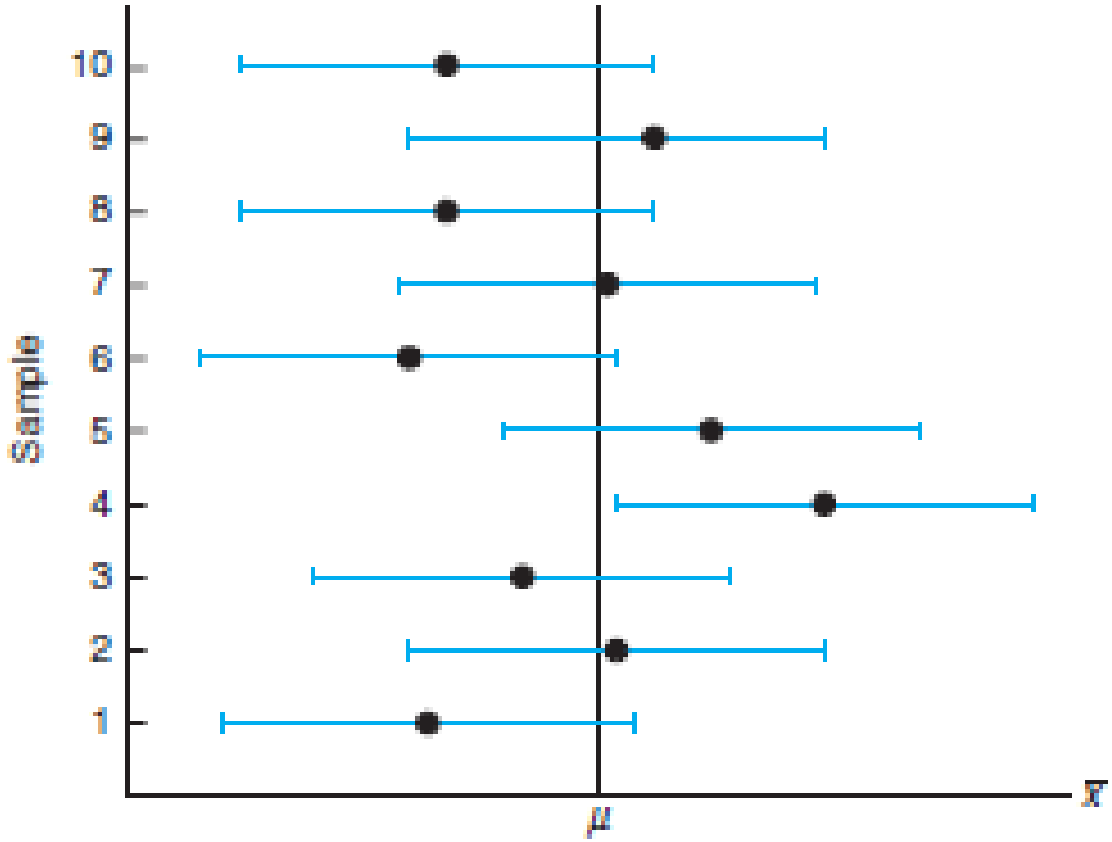
- 0.95, confidence level or confidence coefficient
- $(157,\ 163)$, 95% confidence interval of μ
- 157, lower confidence limit
- 163, upper confidence limit
- $6 = 163 - 157$, interval width

Interpretation of 95% confidence interval

- We are 95% confident that the interval from 157 cm and 163 cm will contain the true value of μ.
- We are 95% confident that the true value of μ lies between 157 cm and 163 cm.
- If we repeat the sampling processes over and over again, then approximately 95% of the similarly constructed intervals are expected to contain the true value of μ.

Caution – *Don't say*

- 95% of all freshmen at UMD are expected to have heights between $157cm$ and $163cm$.
- We are 95% confident that a randomly selected UMD freshman has a height between $157cm$ and $163cm$.

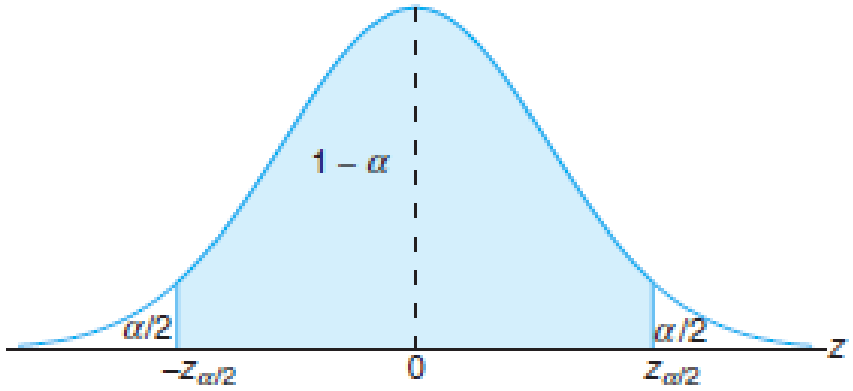

10
9
8
7
6
5
4
3
2
1
Sample
μ
X

Interpretation of Interval Estimates

From the sampling distribution of $\hat{\Theta}$, we shall be able to determine $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1 - \alpha,$$

for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ.


1 – α
α/2
α/2
z
$-z_{\alpha/2}$
0
$z_{\alpha/2}$

The case of known σ

$$\mathrm{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha$$

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

$$\mathrm{P}\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- Example: Average zinc concentration recovered from a sample of zinc measurements in 36 locations of river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence Intervals for the mean zinc concentration in the river. Assume that population standard deviation is 0.3.

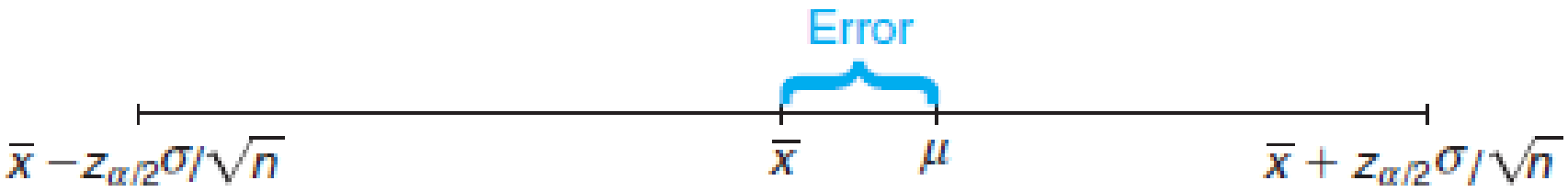
- Point estimate of μ is $\bar{x} = 2.6$.
  - Z value leaving an area of 0.025, is $z_{0.025} = 1.96$
  - Hence 95% confidence Interval is
  - $2.6 - 1.96*(0.3/6) < \mu < 2.6 + 1.96*(0.3/6)$
  - $2.5 < \mu < 2.7$
- Similarly for 99% confidence Interval $z_{0.005}=2.575$
- $2.6 - 2.575*(0.3/6) < \mu < 2.6 + 2.575*(0.3/6)$
- $2.47 < \mu < 2.73$

- High school students who take the SAT mathematics exam second time, generally score higher compared to their first time score. Change in score follows a normal distribution with σ^2=2500. When a random sample is taken of 1000 students, it is found that average gain is 22. Please find out 90% confidence interval for mean score gain μ.

- $\bar{x}$=22, σ^2=2500 α=0.1 n=1000
- P($\bar{x}$ - $Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$ <μ < $\bar{x}$ + $Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$)= 1- α
- $Z_{\alpha/2}$=1.645
- $\frac{\sigma}{\sqrt{n}}=\frac{50}{31.62}$=1.58
- $Z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$=1.645*1.58=2.599
- Lower limit of C.I = 22 – 2.599=19.4
- Upper limit of C.I = 22 + 2.599=24.6

Error in estimating μ by $\bar{x}$.

- If μ is actually the center value of the interval, then $\bar{x}$ estimates μ without error.
- The size of this error will be the absolute value of the difference between μ and $\bar{x}$, and we can be 100(1 – α)% confident that this difference will not exceed $z_{\alpha/2}(\frac{\sigma}{\sqrt{n}})$.



- Frequently, we wish to know how large a sample is necessary to ensure that the error in estimating μ will be less than a specified amount e.
- If we assume e = $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$
- $n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$

- The average zinc concentration recovered from a sample of measurements taken from n different locations in a river is found to be 2.6 grams per milliliter. How large a sample is required (n) if we want to be 95% confident that our estimate of μ is off by less than 0.05?. Assume that the population standard deviation is 0.3 gram per milliliter.

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

- Therefore, we can be 95% confident that a random sample of size 139 will provide an estimate $\bar{x}$ differing from μ by an amount less than 0.05.

One-Sided Confidence Interval on μ, σ Known

If $\overline{x}$ is the mean of a random sample of size n from a population with standard deviation σ, the one-sided $100(1-\alpha)\%$ confidence intervals for μ are given by

upper one-sided C.I.: $$-\infty < \mu < \overline{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

lower one-sided C.I.: $$\overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}} < \mu < \infty.$$

- In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec^2 and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time.

The upper 95% bound is given by

$$\begin{aligned}\bar{x} + z_\alpha \sigma/\sqrt{n} &= 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658 \\ &= 6.858 \text{ seconds.}\end{aligned}$$

Hence, we are 95% confident that the mean reaction time is less than 6.858 seconds.

The Case of σ Unknown

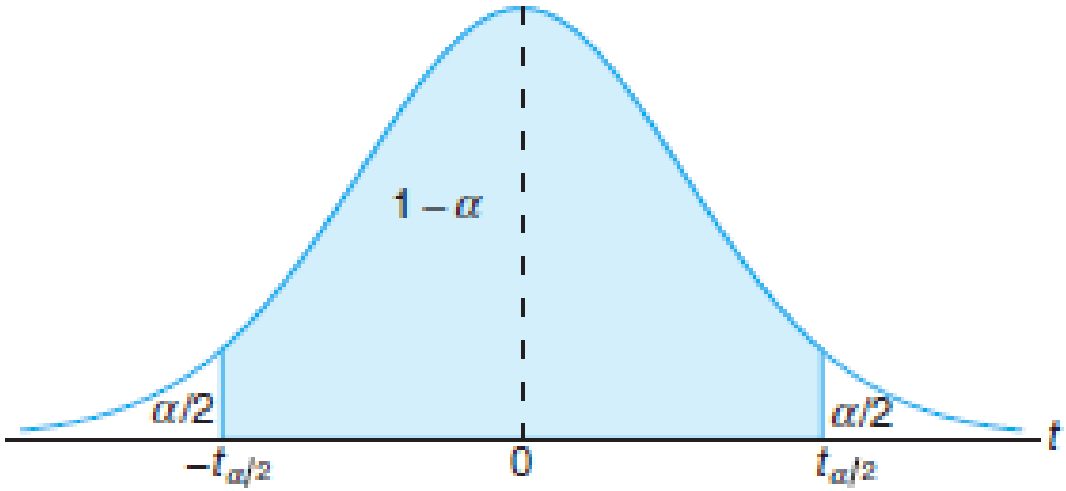
- Frequently, we must attempt to estimate the mean of a population when the variance is unknown.

- if we have a random sample from a *normal distribution*, then the random variable $T=\frac{X-\mu}{S/\sqrt{n}}$ has a Student *t*-distribution with $n-1$ degrees of freedom

Confidence Interval on μ, when σ Unknown

If $\overline{x}$ and s are the mean and the standard deviation of a random sample of size n from a population with unknown standard deviation σ, a $100(1-\alpha)\%$ confidence interval for μ is given by

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


1 − α
α/2
α/2
t
−t_α/2
0
t_α/2

- Example: The contents of 7 similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0 10.2 and 9.6 liters. Find a 95% CI for the mean of all such containers assuming an approximate normal distribution.

	α						
ν	0.40	0.30	0.20	0.15	0.10	0.05	0.025
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228

- Sample mean $\bar{x}$ = 10.0
- Sample standard deviation s = 0.283
- α=0.05 $\alpha/2$=0.025
- $T_{0.025}$=2.447 for v=6 degree of freedom
- $10.0 - 2.447*(0.283/\sqrt{7}) < \mu < 10.0 + 2.447*(0.283/\sqrt{7})$
- $9.74 < \mu < 10.26$

EXAMPLE *How much do users pay for Internet service?* Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000: (Data from the August 2000 supplement to the Current Population Survey, from the Census Bureau Web site, www.census.gov.)

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

Give a 95% confidence interval for the mean monthly cost of Internet access in August 2000.

- $\overline{x}$=20.9 s^2=58.459 s=7.65 n=50
- α=0.05 $\alpha/2$=0.025
- As n>=30, instead of $t_{\alpha/2}$, we may use $Z_{\alpha/2}$.
- $Z_{0.025}$=1.96
- $s/\sqrt{n}$ =1.08
- Lower limit = 20.9 – 1.96*1.08 = 20.9-2.12=18.78
- Upper limit = 20.9 + 1.96*1.08 = 20.9+2.12=23.02