Design IV

E232 Fall 07

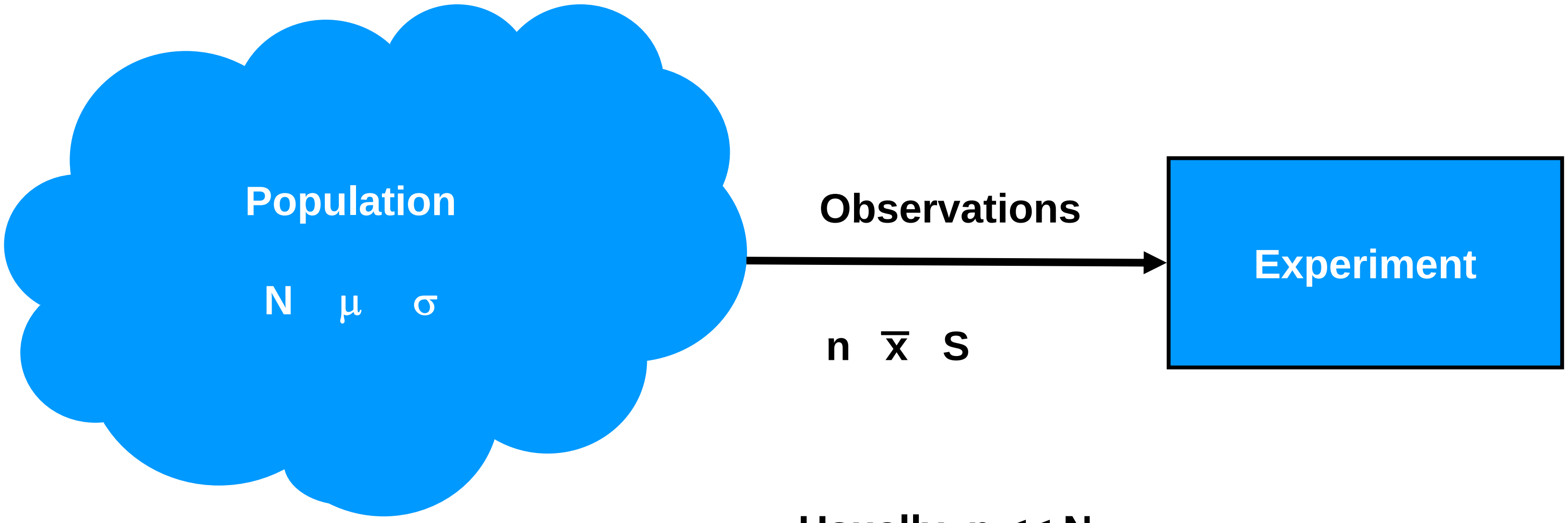
Class 18

Bruce McNair
bmcnair@stevens.edu

Today's topics

- More On Statistical Analysis of Experimental Data
  - Parameter estimation

Parameter Estimation



Population

N μ σ

Observations

n $\bar{x}$ S

Experiment

N potential outcomes

Mean value μ

Standard deviation σ

Usually, n << N
so $\bar{x} \sim \mu$
$S \sim \sigma$

But if $\delta = |\bar{x}-\mu|$,
how large is δ likely to be?

Confidence Interval and Confidence Level

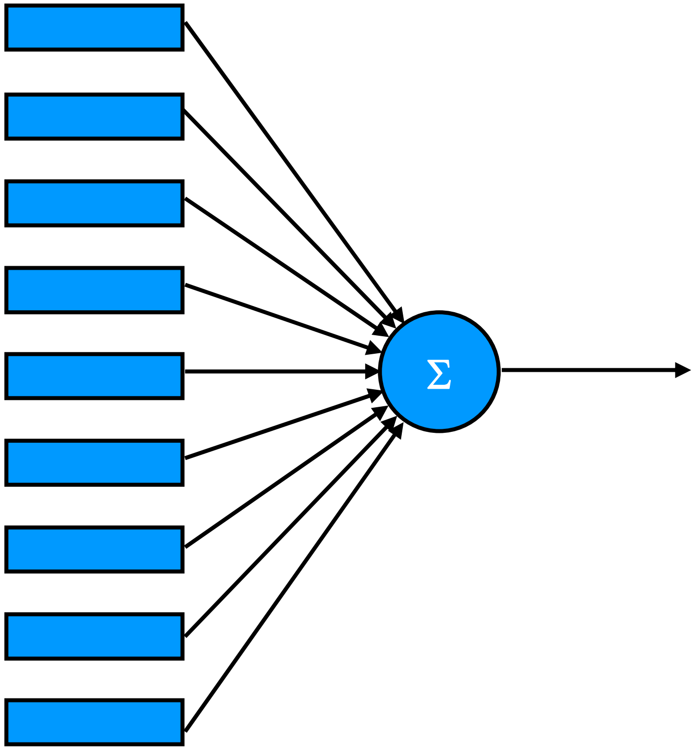
- For a population having a mean μ, the observed mean of n samples measured in one experiment is $\overline{x}$. The confidence interval, i.e., the region within δ of μ is:

$$\overline{x} - \delta \leq \mu \leq \overline{x} + \delta$$

- If α is the probability that the observed mean will not be within δ of μ, the confidence level is:

$$1 - \alpha = P(\overline{x} - \delta \leq \mu \leq \overline{x} + \delta)$$

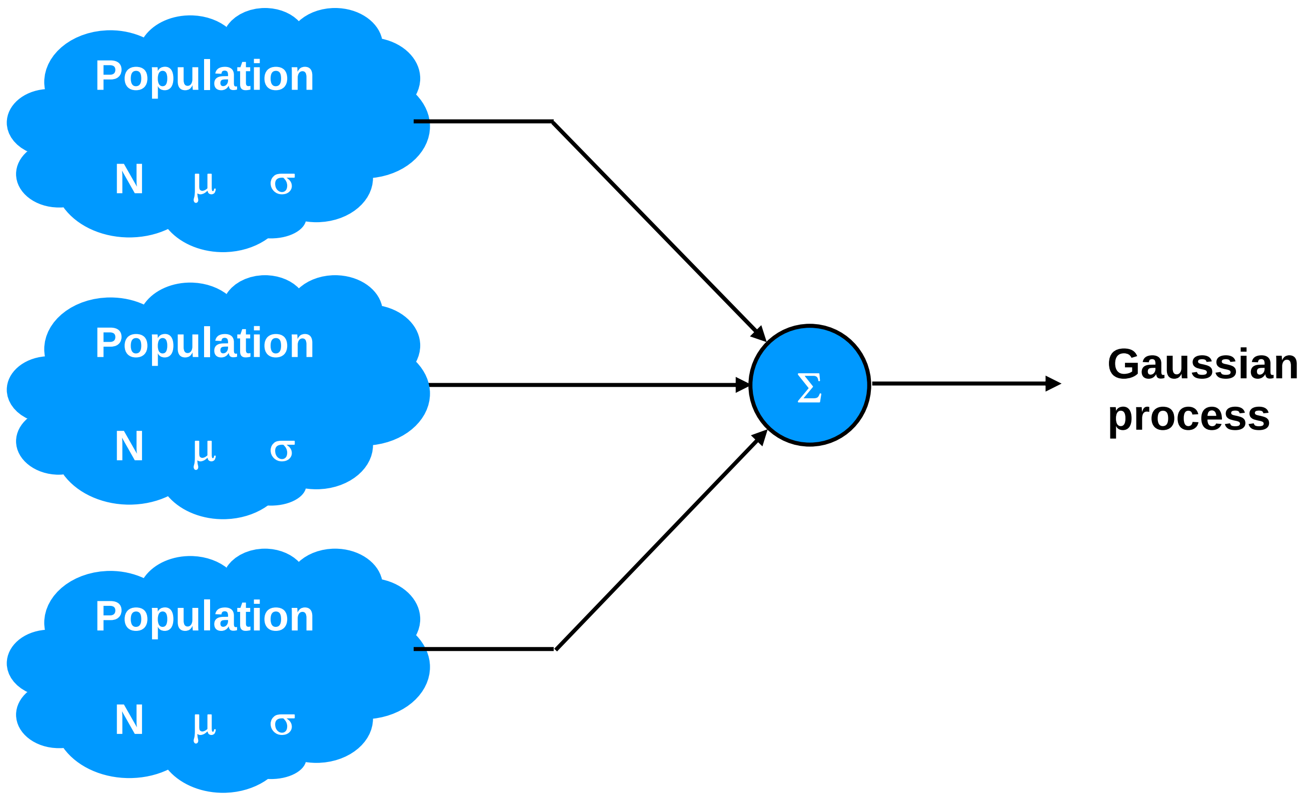
Source of Gaussian Distribution



Gaussian process

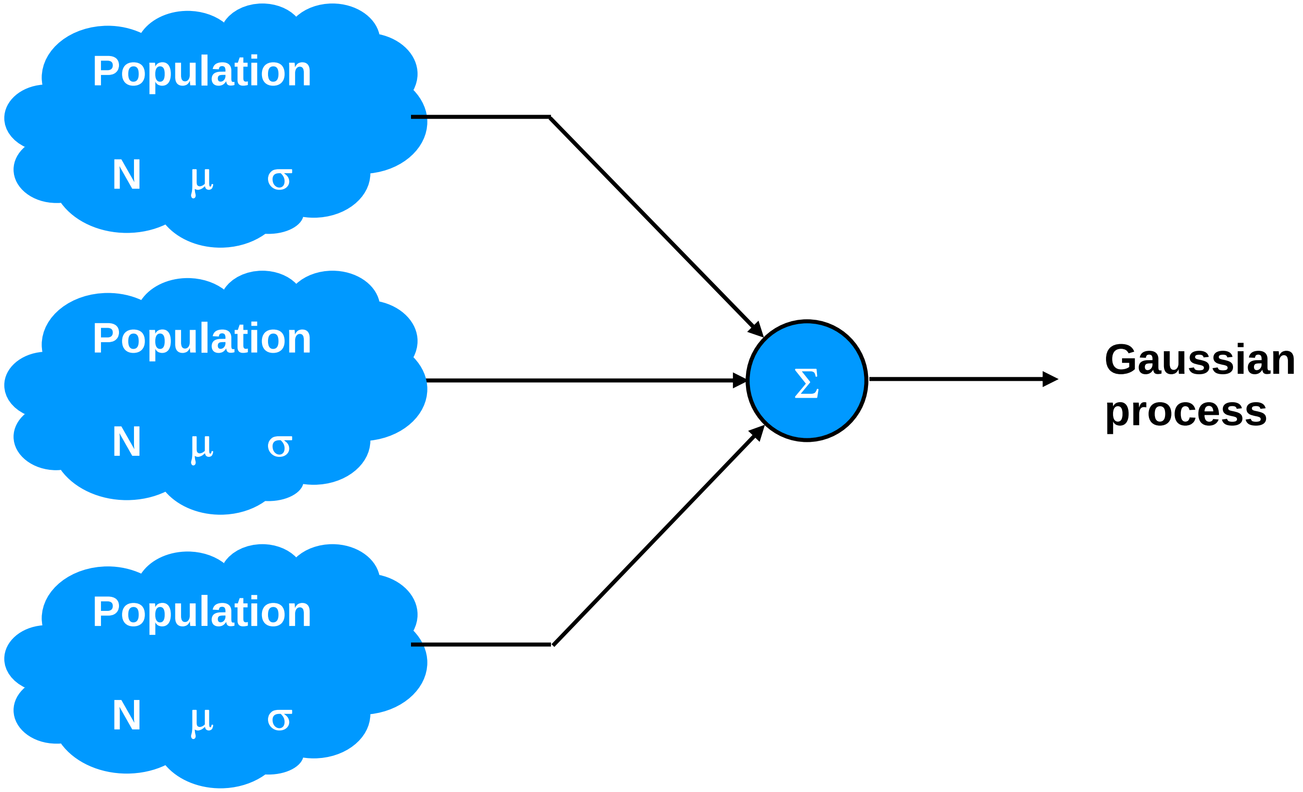
Large number of independent, identically distributed r.v.s

From last time: Source of Gaussian Distribution



Assume experimental samples are picked randomly from population
Assume number of samples is "large",
Assume identically distributed samples

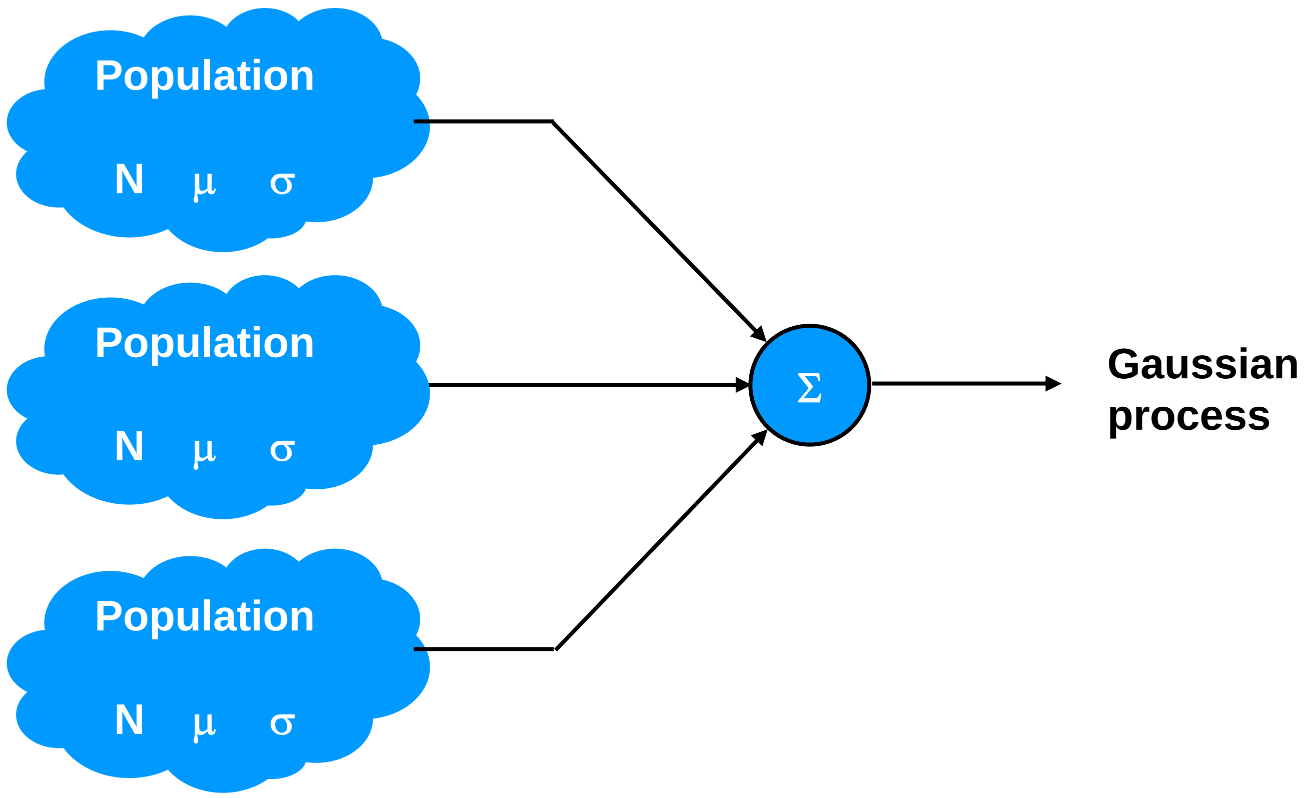
From last time: Source of Gaussian Distribution



Assume experimental samples are picked randomly from population
Assume number of samples is "large",
Assume identically distributed samples

If population is gaussian,
n>0 is gaussian

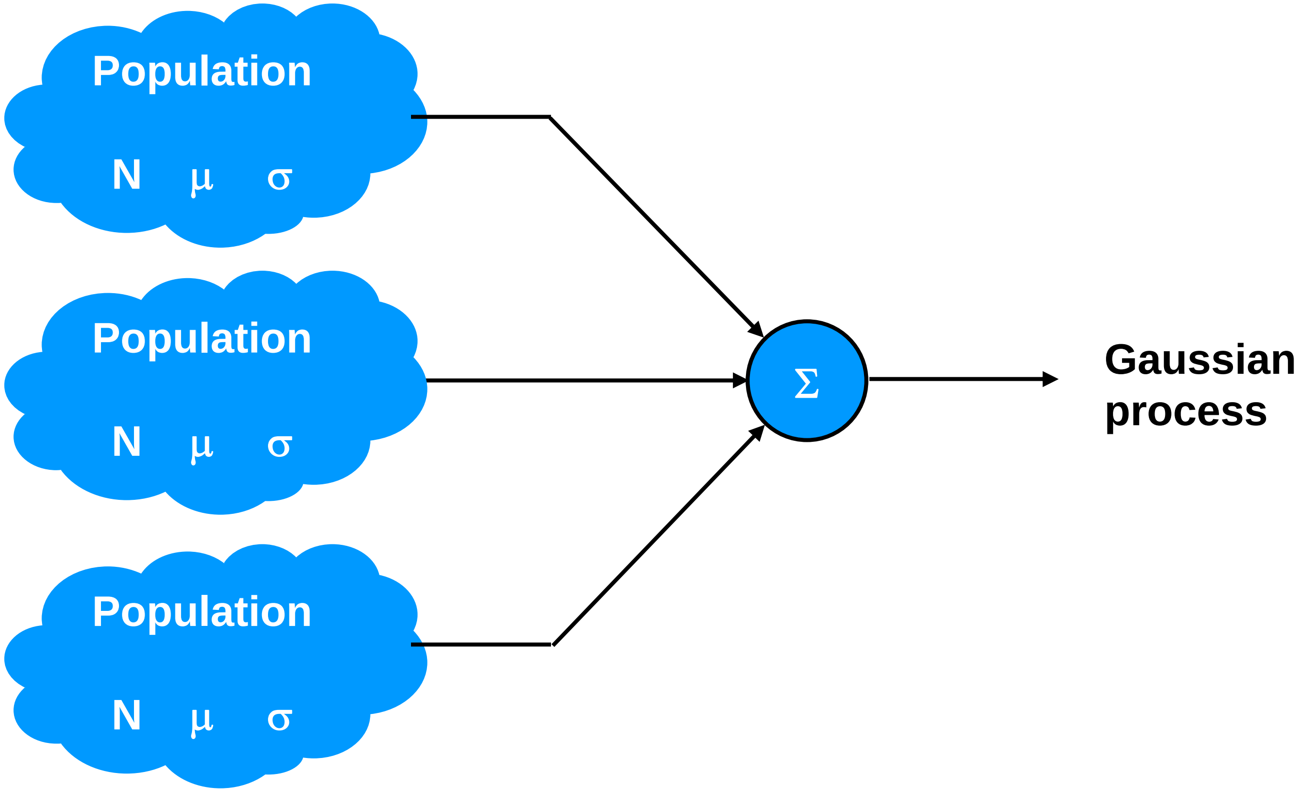
From last time: Source of Gaussian Distribution



Assume experimental samples are picked randomly from population
Assume number of samples is "large",
Assume identically distributed samples

If population is non gaussian, but n>30, result is very close to gaussian

From last time: Source of Gaussian Distribution

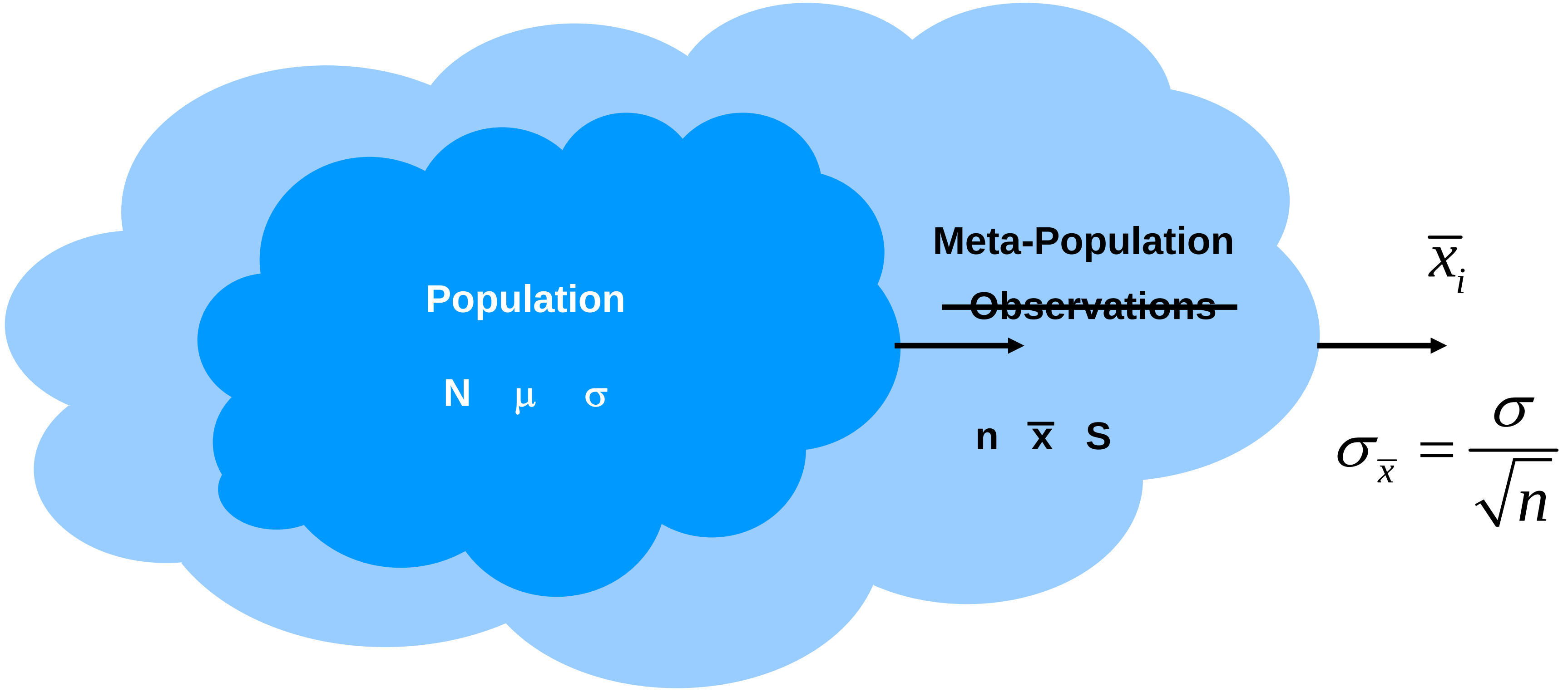


Assume experimental samples are picked randomly from population
Assume number of samples is “large”,
Assume identically distributed samples

If population is non gaussian,
and n<30 result is not gaussian

Estimating Confidence Interval

- Assume a large enough sample size, use Central Limit Theorem



Population

N μ σ

Meta-Population

~~**Observations**~~

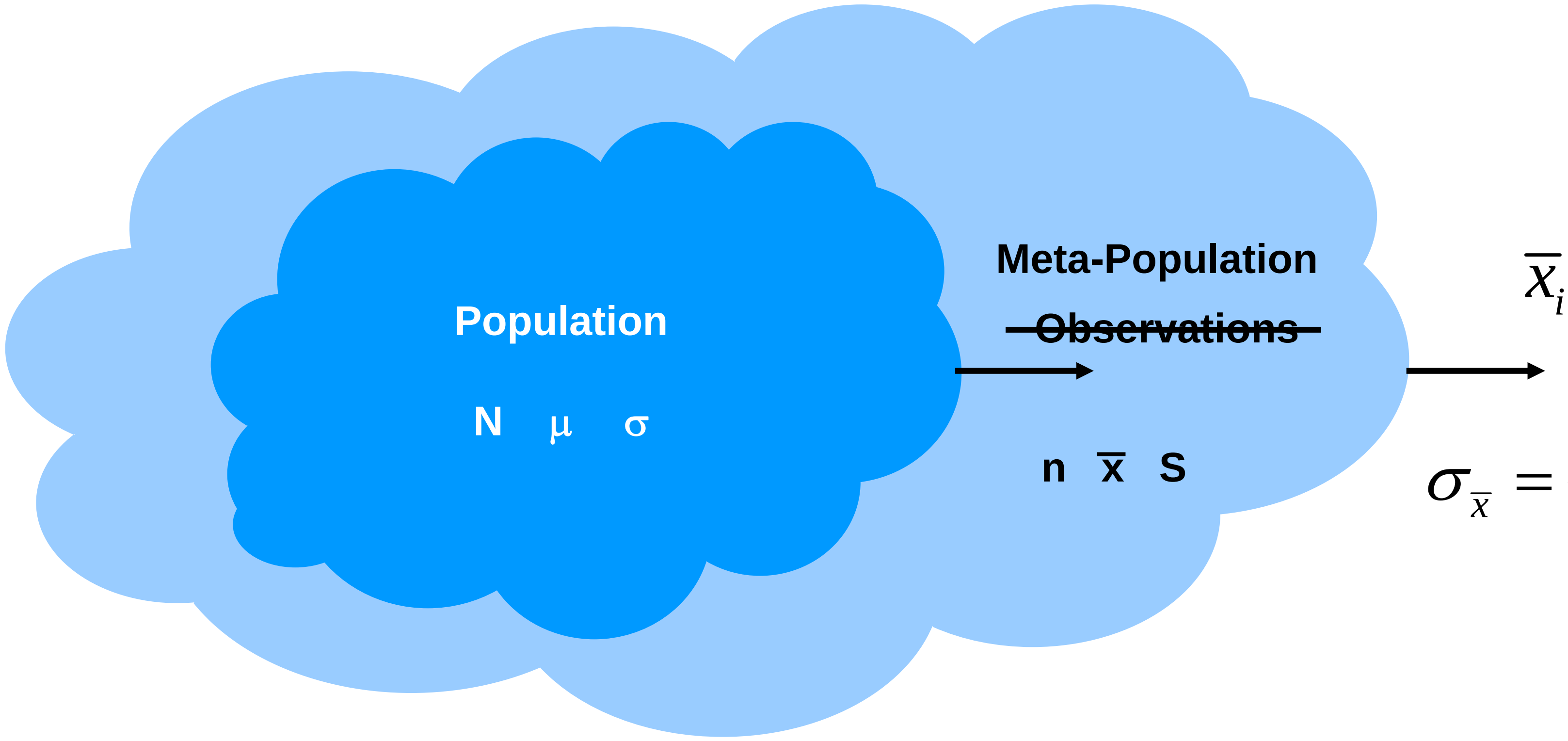
n $\bar{x}$ S

$\bar{x}_i$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Estimating Confidence Interval

- Assume a large enough sample size, use Central Limit Theorem



Population

N μ σ

Meta-Population

~~Observations~~

n $\overline{x}$ S

$\overline{x}_i$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

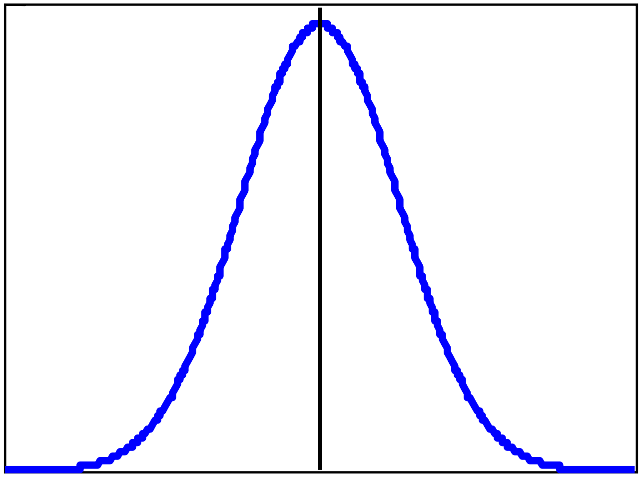
What can be said, statistically, about confidence interval?

Computing Confidence Interval

- Define statistic z:

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$$

z is normally distributed with <u>zero mean</u>



z=0

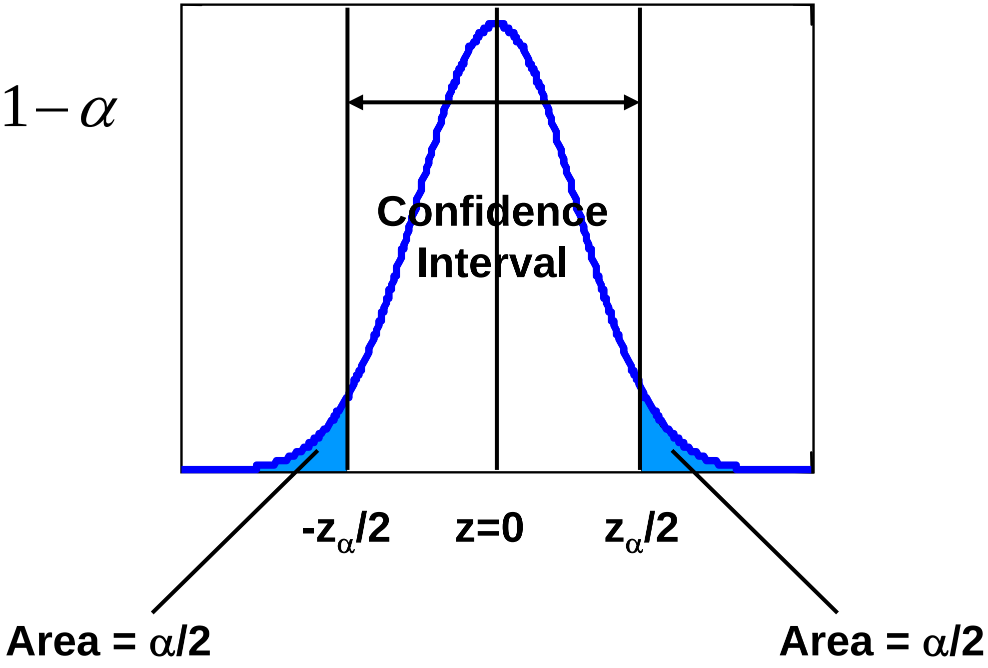
Computing Confidence Interval

- Define statistic z:

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$$

z is normally distributed with <u>zero mean</u>

$$P(-z_{\alpha/2} \le \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}) = 1 - \alpha$$



Computing Confidence Interval

- Define statistic z:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

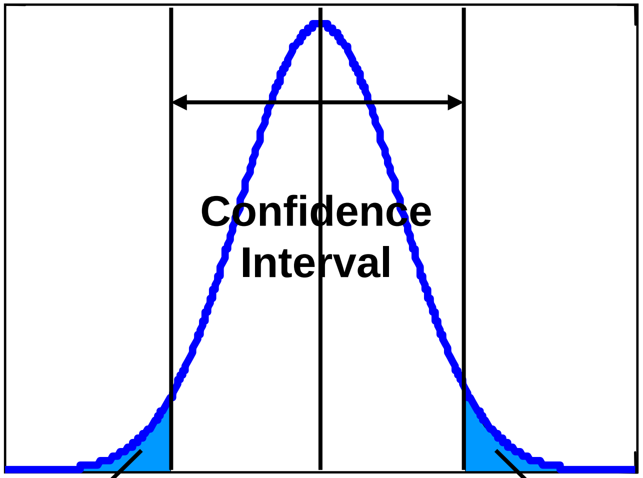
z is normally distributed with <u>zero mean</u>

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

Or,

$$\mu = \bar{x} \pm \left(z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

with confidence 1-α



$-z_{\alpha}/2$ z=0 $z_{\alpha}/2$

Area = α/2

Area = α/2

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^{o}C with a standard deviation of 3 ^{o}C. What is the 98% confidence interval for the average melting temperature?

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^oC with a standard deviation of 3 ^oC. What is the 98% confidence interval for the average melting temperature?
    - 50 samples are enough to assume a Gaussian distribution of results.

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^oC with a standard deviation of 3 ^oC. What is the 98% confidence interval for the average melting temperature?

  - 50 samples are enough to assume a Gaussian distribution of results.
  - 98% confidence implies α = .02

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^{o}C with a standard deviation of 3 ^{o}C. What is the 98% confidence interval for the average melting temperature?

  - 50 samples are enough to assume a Gaussian distribution of results.
  - 98% confidence implies α = .02
  - We need to find $z_{\alpha}/2$ such that

$$\int_{0}^{z_{\alpha}/2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = .5 - \frac{\alpha}{2} = .49$$

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^oC with a standard deviation of 3 ^oC. What is the 98% confidence interval for the average melting temperature?
  - 50 samples are enough to assume a Gaussian distribution of results.
  - 98% confidence implies α = .02
  - We need to find $z_\alpha/2$ such that

$$\int_0^{z_\alpha/2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = .5 - \frac{\alpha}{2} = .49$$

  - Use Table 6.3 with a value .49

$z_\alpha/2$ = 2.326

Confidence Interval Example

- We need to know the melting point of an organic compound being manufactured by a chemical process, but the results depend on the specific composition, which varies randomly. 50 samples are tested, and the average melting point is found to be 80 ^oC with a standard deviation of 3 ^oC. What is the 98% confidence interval for the average melting temperature?

  - 50 samples are enough to assume a Gaussian distribution of results.
  - 98% confidence implies α = .02
  - We need to find $z_\alpha/2$ such that

$$\int_0^{z_\alpha/2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} = .5 - \frac{\alpha}{2} = .49$$

  - Use Table 6.3 with a value .49

    $z_\alpha/2 = 2.326$

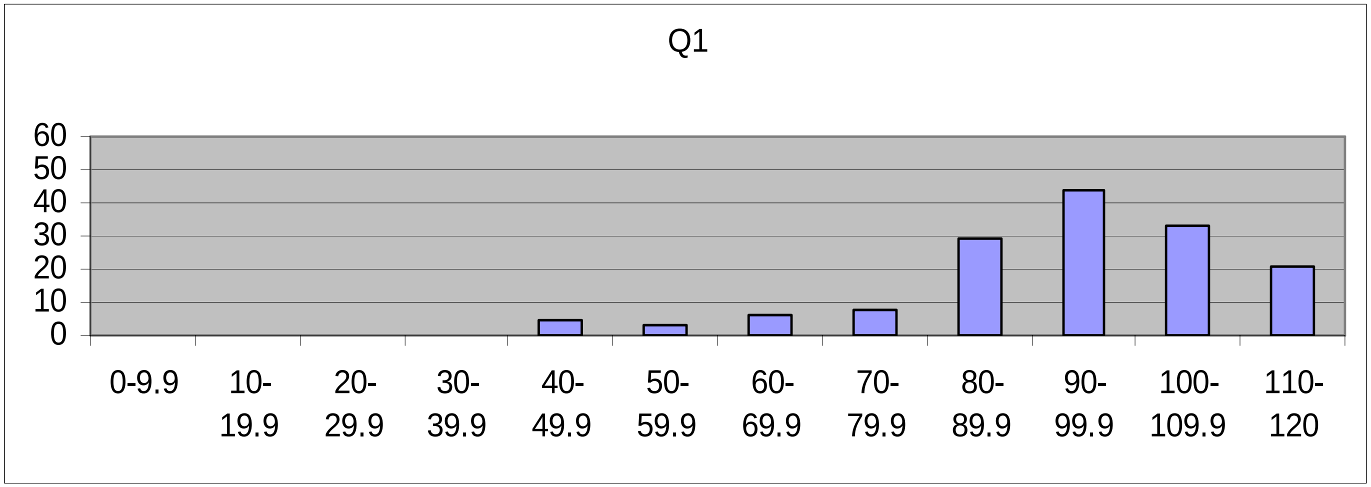
  - Find confidence interval:

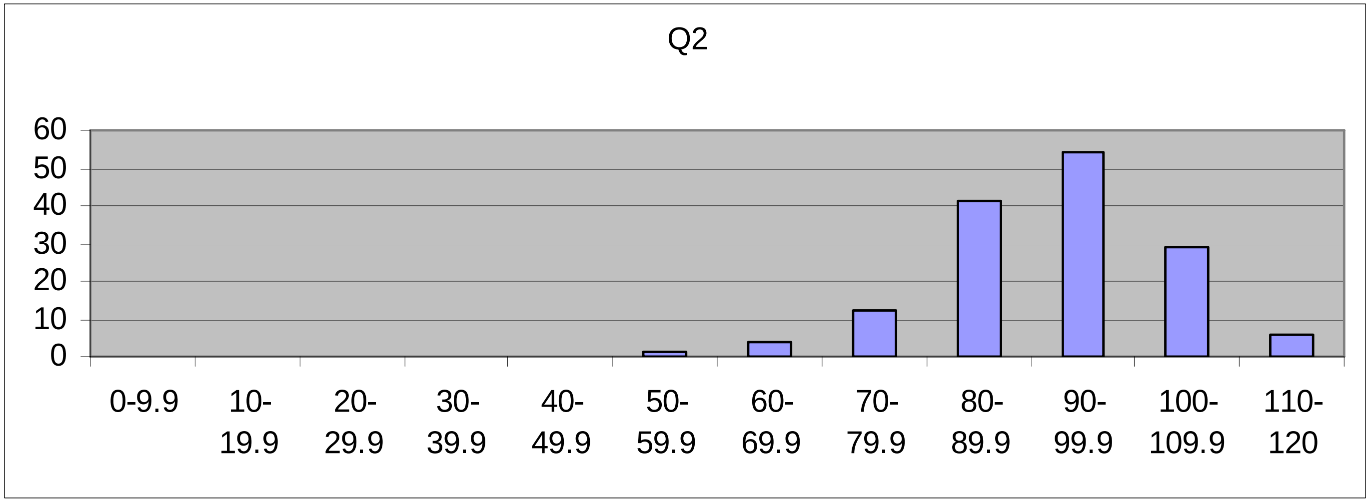
$$\overline{x} - \frac{z_{\alpha/2}S}{\sqrt{n}} \le \mu \le \overline{x} + \frac{z_{\alpha/2}S}{\sqrt{n}} = 80 - \frac{2.326\cdot 3}{\sqrt{50}} \le \mu \le 80 + \frac{2.326\cdot 3}{\sqrt{50}}$$

$$79.013 \le \mu \le 80.987$$

Identifying Relationships Between Data

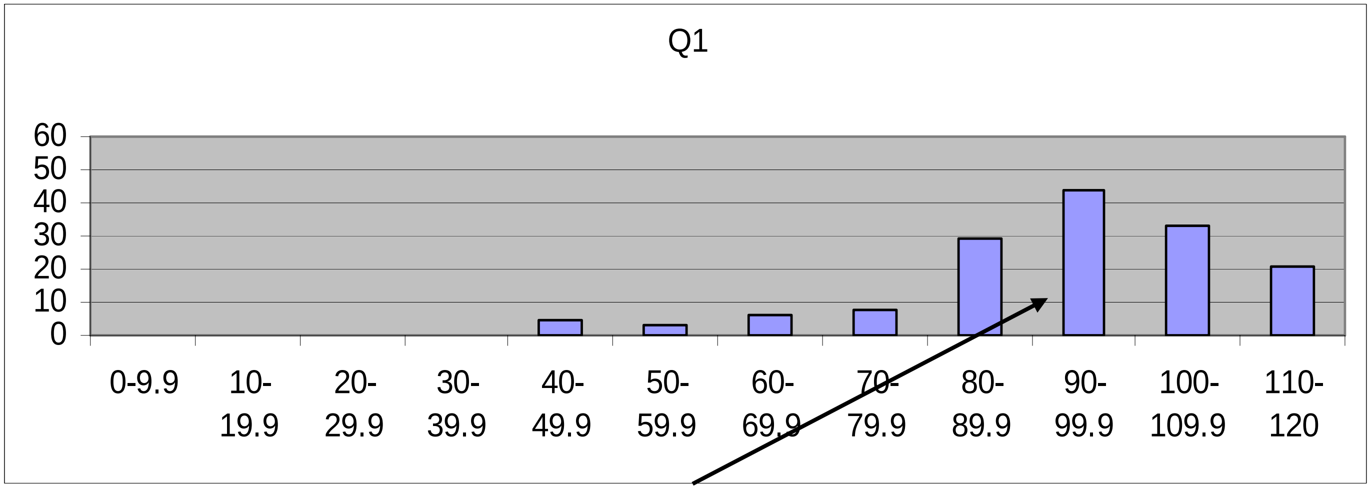
- Consider the Quiz1 and Quiz2 grades in E232-Spring07:



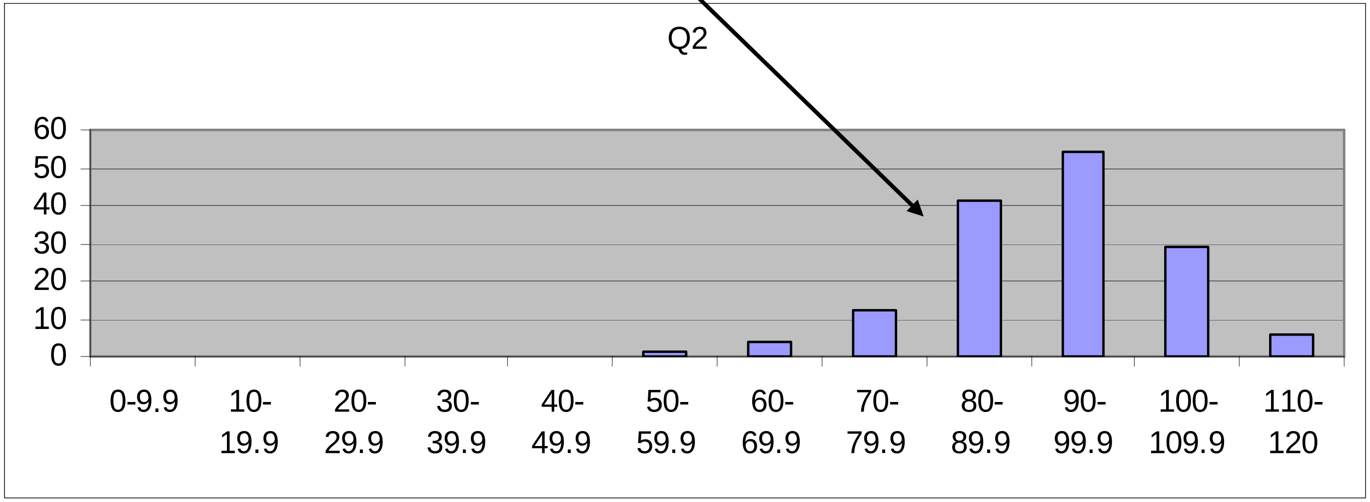


Identifying Relationships Between Data

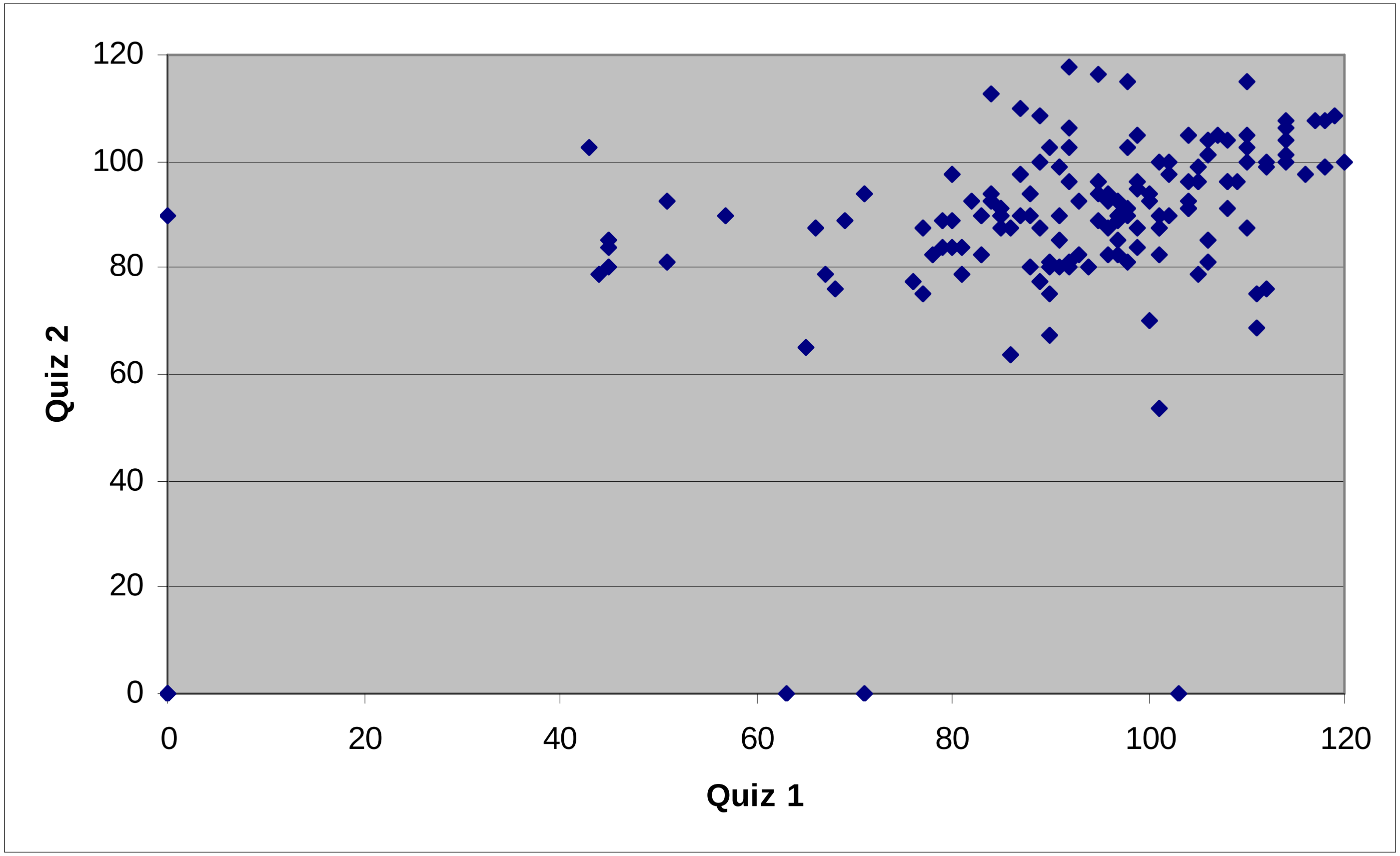
- Consider the Quiz1 and Quiz2 grades in E232-Spring07: How are these related?



Distribution shape is about the same

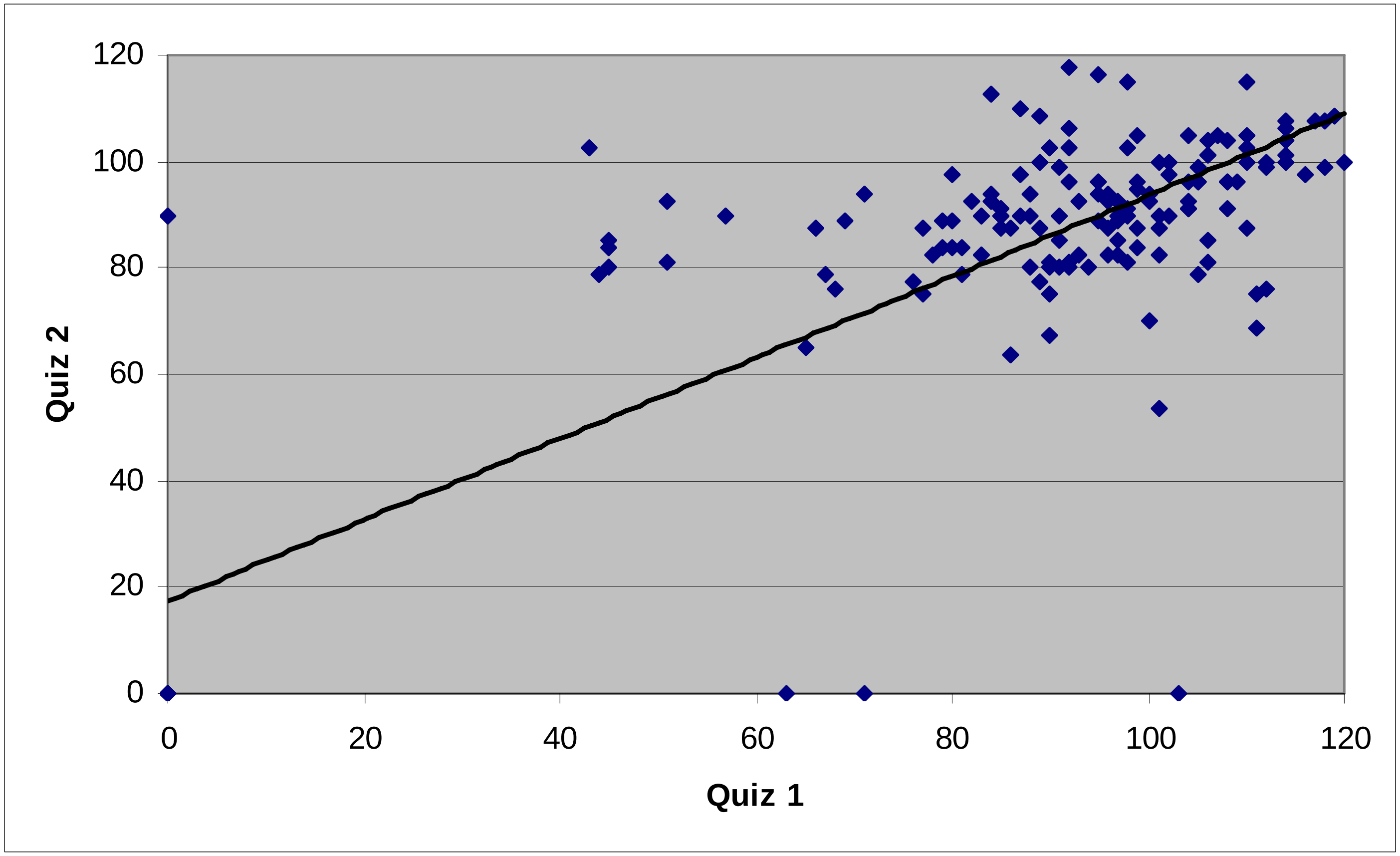


X-Y Scatter Plot



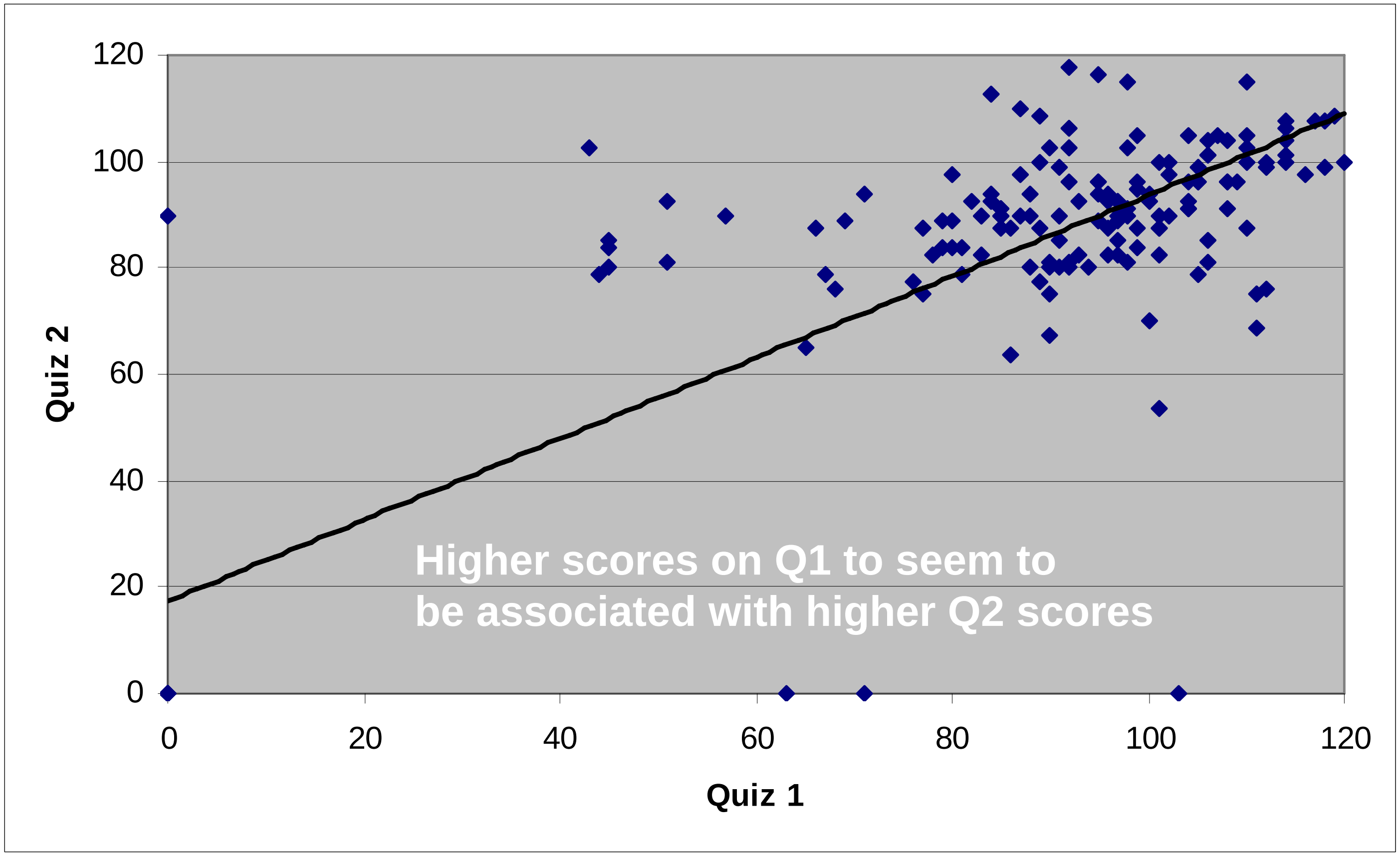
X-Y Scatter Plot

- Fit a trendline to available data



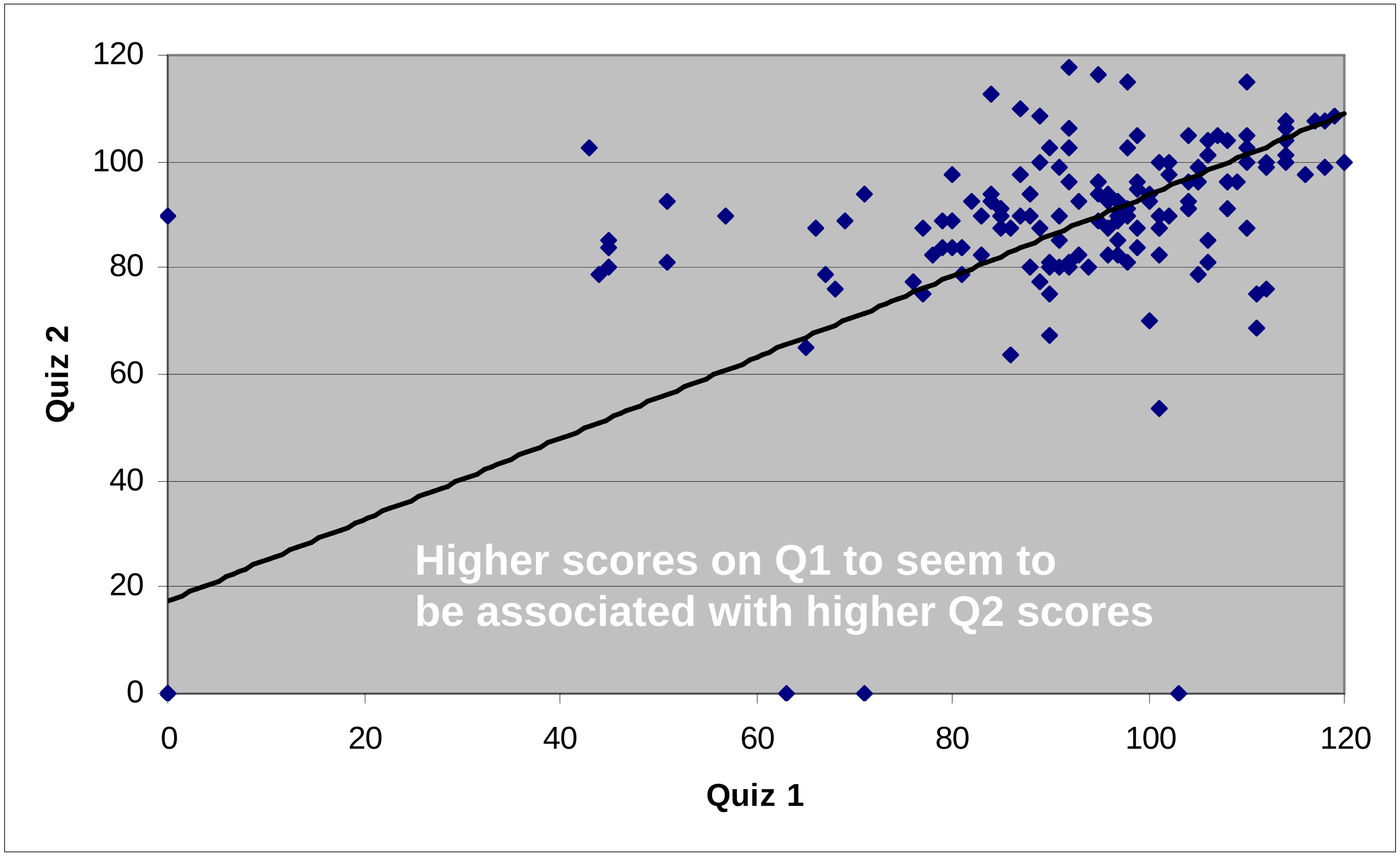
X-Y Scatter Plot

- Fit a trendline to available data



X-Y Scatter Plot

- Fit a trendline to available data – how can we compute the similarity of data values?



Correlation Coefficient

- x_i and y_i are random variables (x=Q1 grades, y=Q2 grades)

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\left(\sum \left(x_i - \overline{x}\right)^2\right) \cdot \left(\sum \left(y_i - \overline{y}\right)^2\right)}}$$

Correlation Coefficient

- x_i and y_i are random variables (x=Q1 grades, y=Q2 grades)

$$\left\langle A,B\right\rangle = \frac{1}{T}\int_{t}^{t+T} A(x)B(x)dx$$

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\left(\sum\left(x_i - \overline{x}\right)^2\right)\cdot\left(\sum\left(y_i - \overline{y}\right)^2\right)}}$$

Variance of x

Variance of y

Characteristics of Correlation Coefficient

$$-1 \le r_{xy} \le 1$$

x&y are closely related, but opposite in sign (anticorrelated)

x&y are closely related

$r_{xy} = 0$: x & y are unrelated

Characteristics of Correlation Coefficient

$$-1 \le r_{xy} \le 1$$

x&y are closely related, but opposite in sign (anticorrelated)

x&y are closely related

$r_{xy} = 0$: x & y are unrelated

For Q1 and Q2 data, $r_{xy} = .63$

Characteristics of Correlation Coefficient

$$-1 \le r_{xy} \le 1$$

x&y are closely related, but opposite in sign (anticorrelated)

x&y are closely related

$r_{xy} = 0$: x & y are unrelated

For Q1 and Q2 data, $r_{xy} = .63$

Does this indicate significant correlation?

Computing Critical Values of r_{xy}

- See Table 6.9:

α values

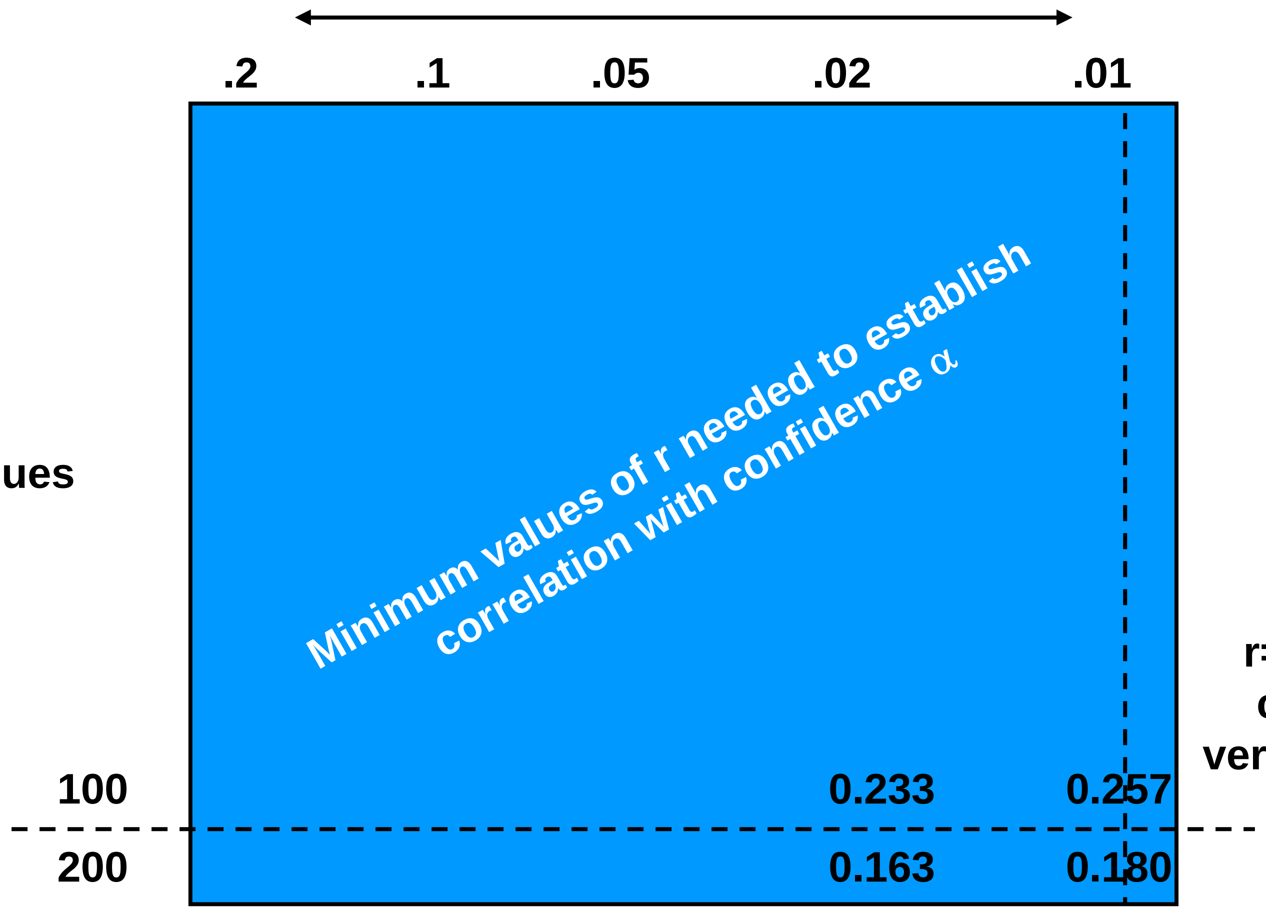
n values

	.2	**.1**	**.05**	**.02**	**.01**
100				**0.233**	**0.257**
200				**0.163**	**0.180**

Minimum values of r needed to establish correlation with confidence α

Computing Critical Values of r_{xy}

- See Table 6.9:



α values

n values

Minimum values of r needed to establish correlation with confidence α

	.2	**.1**	**.05**	**.02**	**.01**
100				**0.233**	**0.257**
200				**0.163**	**0.180**

r=.63 establishes correlation with very high confidence