Data: Information or Just a Bunch of Numbers?

Jim Burati

Clemson University, Civil Engineering Dept.

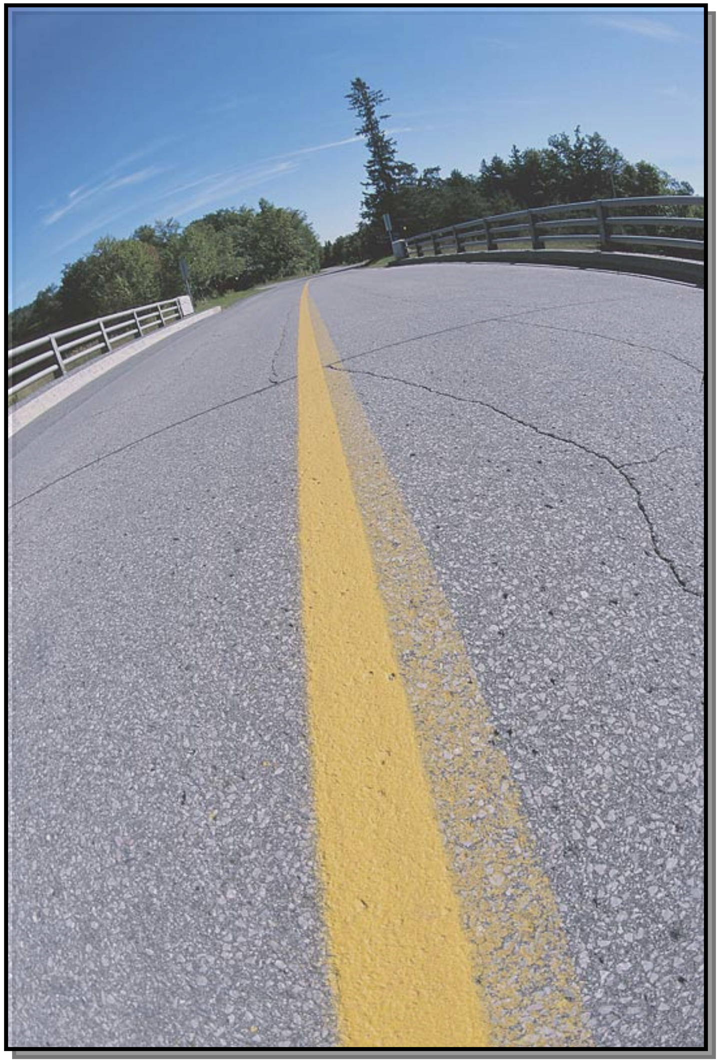
2005 Southeastern Pavement Management & Design Conference

June 21, 2005

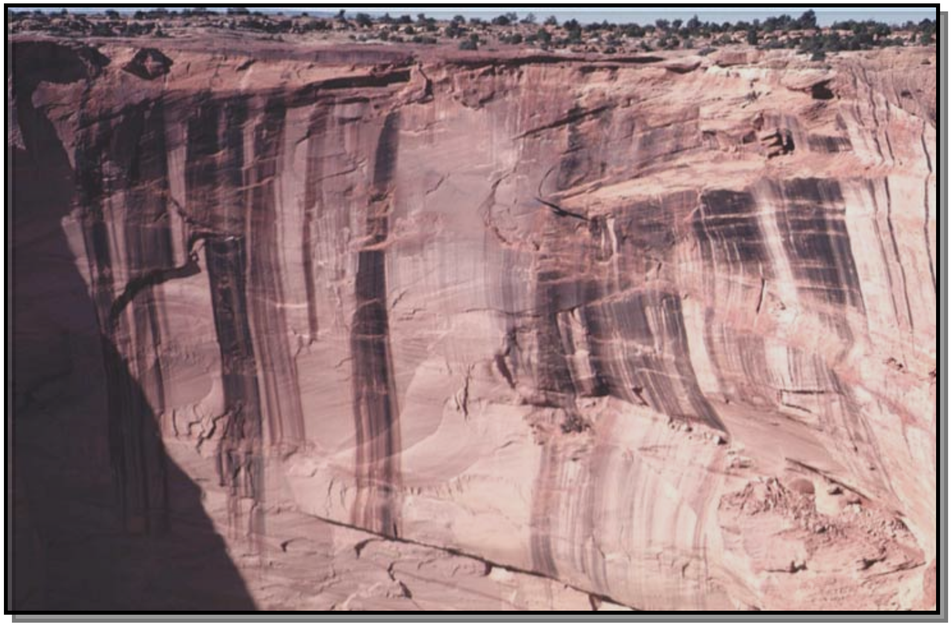
Topics

- Variability
- Data
- Estimating Parameters
- Probability Distributions
- Drawing Conclusions
- Regression Analysis?

Sources of Variability



Material

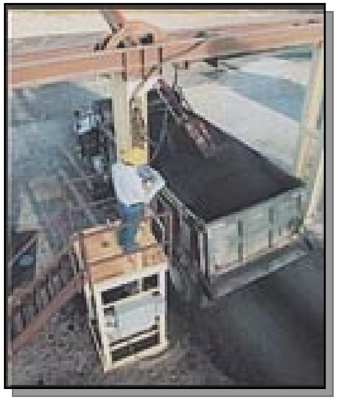


Sources of Variability

Process

Sources of Variability

Sampling



Truck?

Road?

Drive?

Walk?

Total Project?

Portion?

Sources of Variability

Testing

Variability

- Which variability do we obtain? Which do we need?
  - **Material variability.**
  - **Process variability.** $\Big\}\ \sigma_m^2$
  - **Sampling variability.** σ_s^2
  - **Testing variability.** σ_e^2

Overall Variability

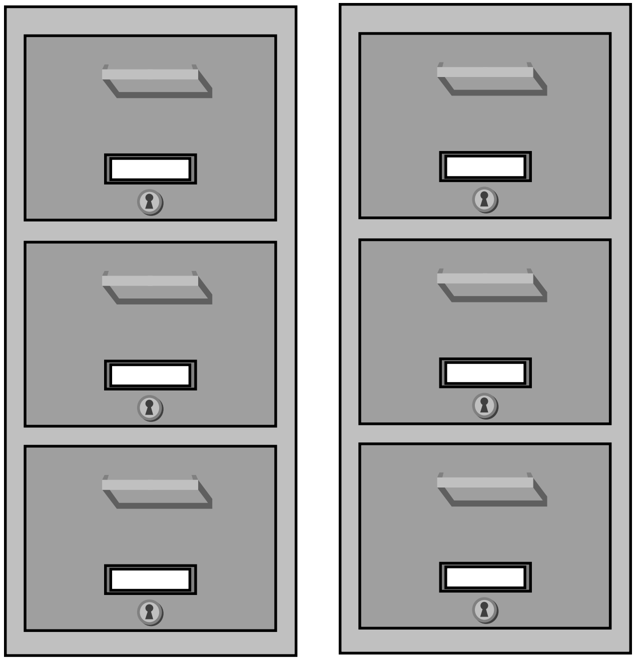
Obtain?

$$\sigma_o^2 = \sigma_m^2 + \sigma_s^2 + \sigma_e^2$$

Need?

$$\sigma_o^2 = \boxed{\sigma_m^2} + \sigma_s^2 + \sigma_e^2$$

Sources of Data



Be Wary of Historical Records

Sampling?

Biased Reporting?

Test Methods?

Sources of Data

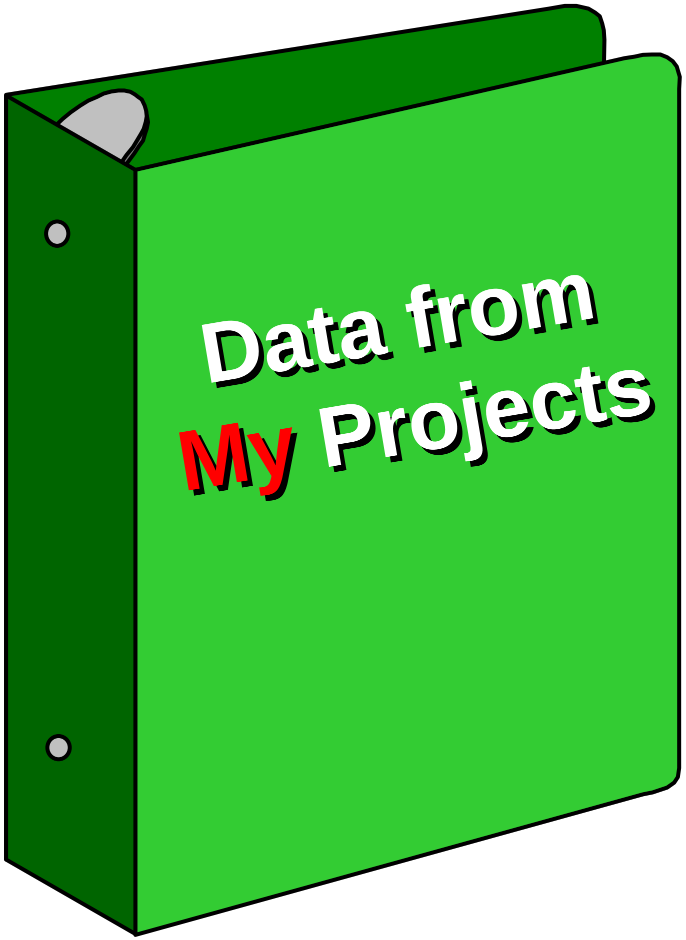
Data from Other States?

FHWA?

Materials?

Procedures?

Sources of Data



Random Sampling

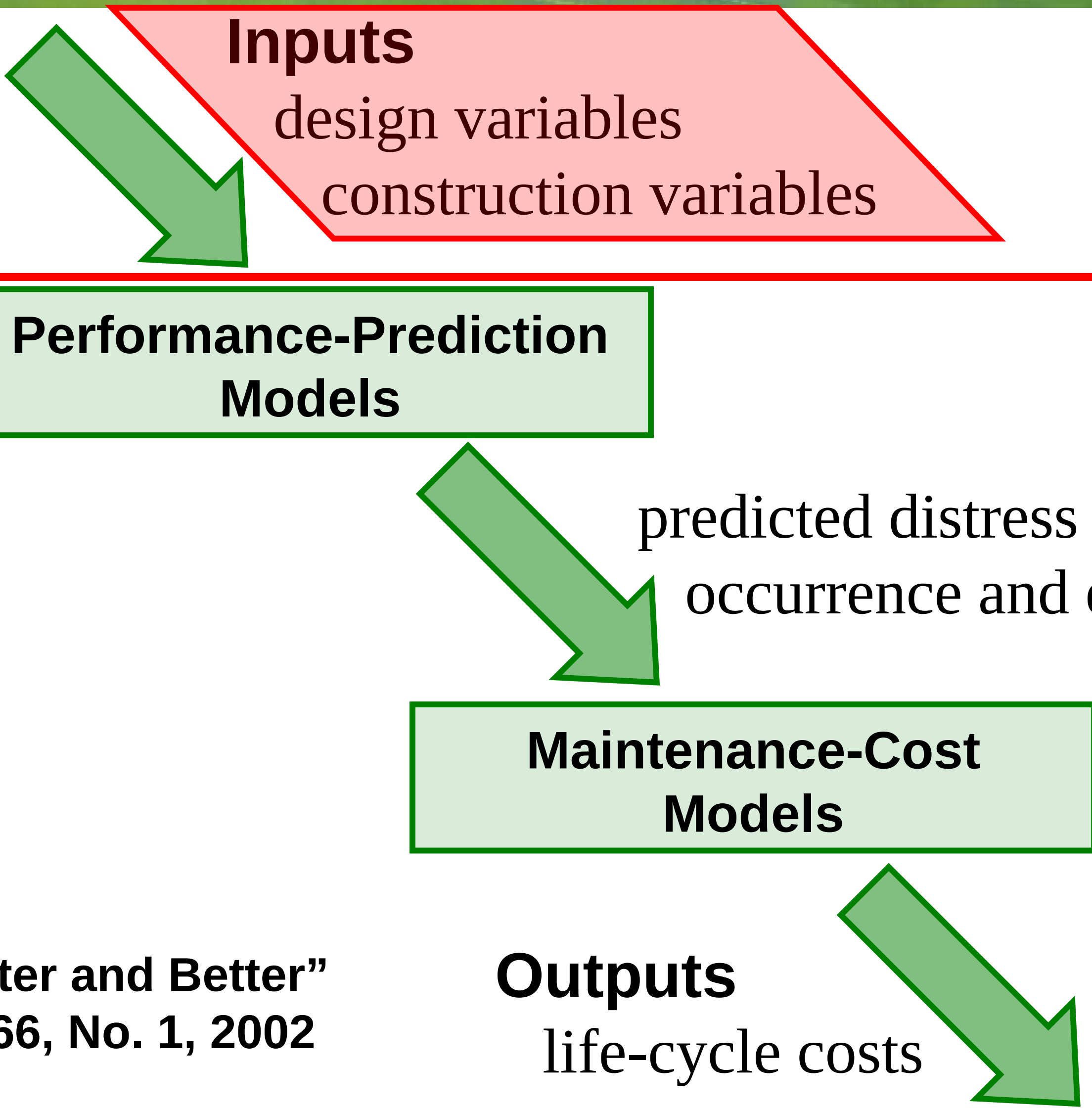
&

Unbiased Reporting

&

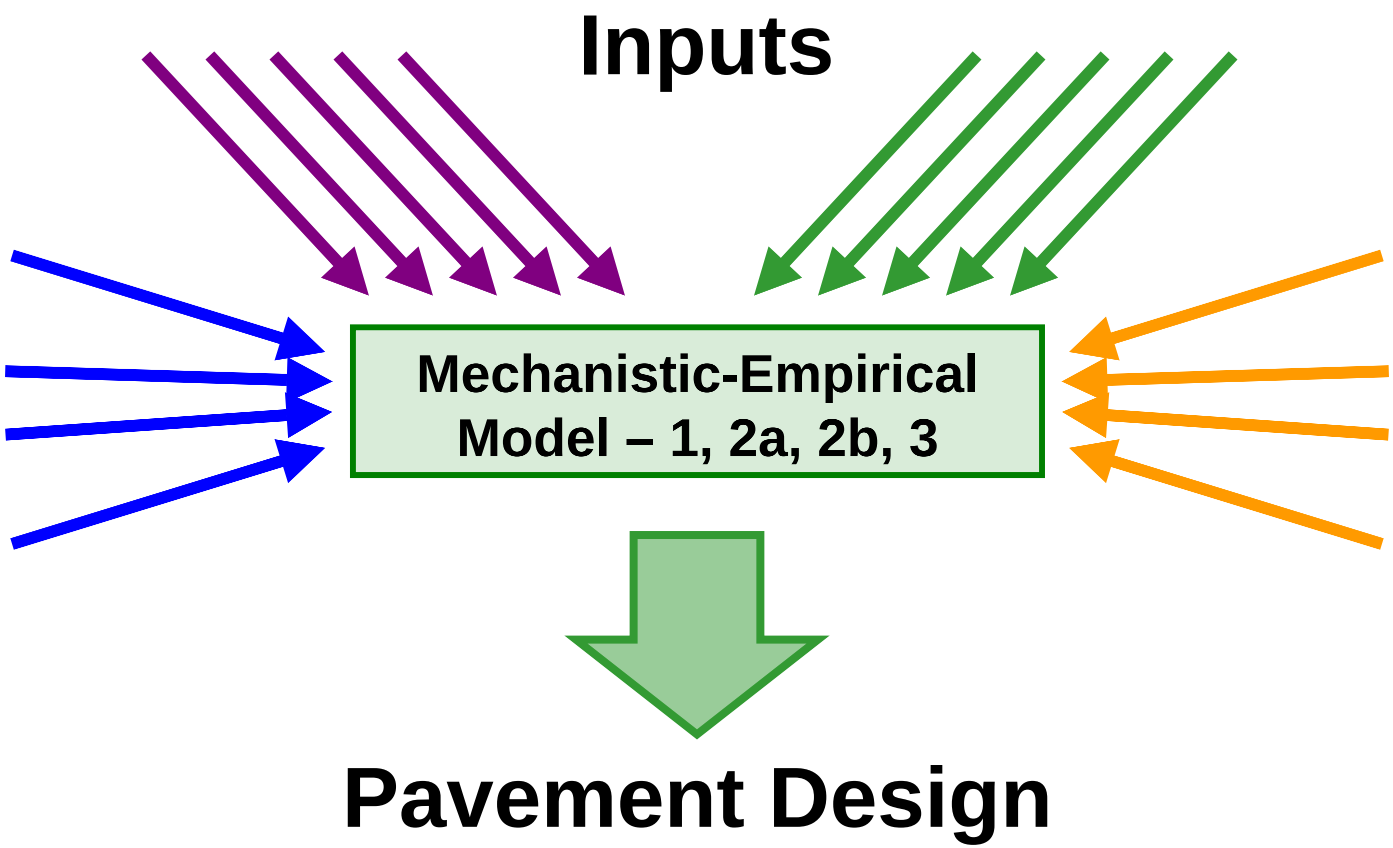
Same Procedures

Performance Related Specs



Peter Kopac
"Making Roads Better and Better"
***Public Roads,* Vol. 66, No. 1, 2002**

Mechanistic Empirical Model



What do we want to know?

→ **Center** ←

← **Spread** →

What Data Do We Need?

- Parameters
  - Mean.
  - Standard Deviation.
- Probability Distribution

Things are easy, right?

- **Parameters**
  - **Center? Mean.** *$\bar{x}$ for μ*
  - **Spread?**
    Standard Deviation *s for σ*
- **Probability Distributions**

Normal

Example

- Normal Population with:
  - **Mean = 100**
  - **Standard Deviation = 10**
- Wish to estimate the mean and standard deviation.

Estimating the Mean

- Unbiased estimate

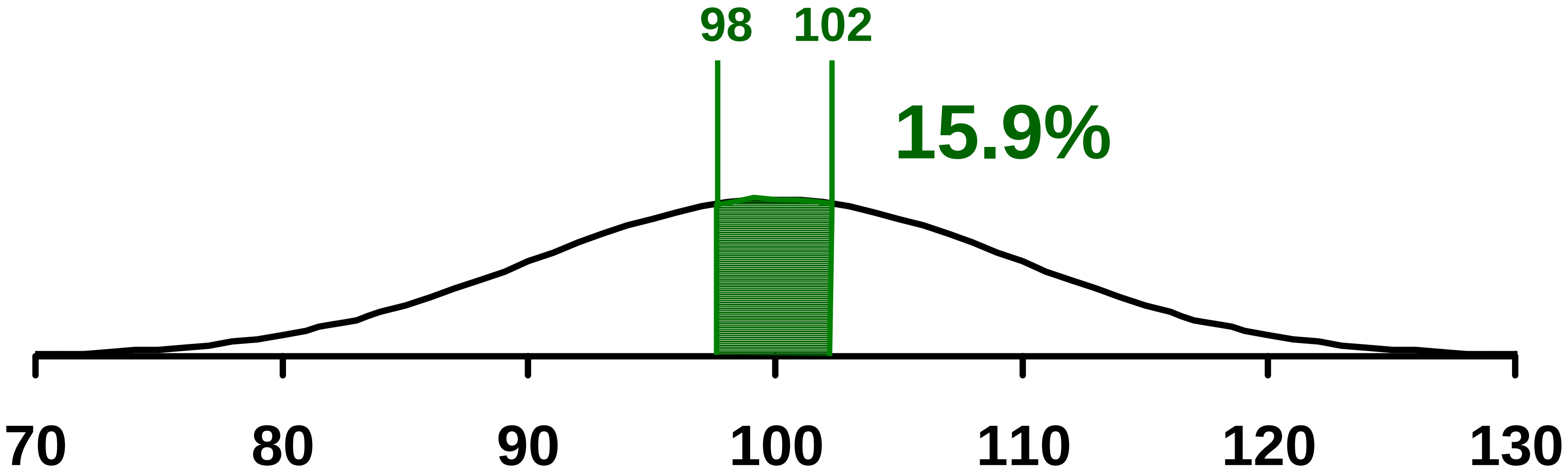
$$\bar{X} \text{ for } \mu$$

- Confidence Interval

Sample Size

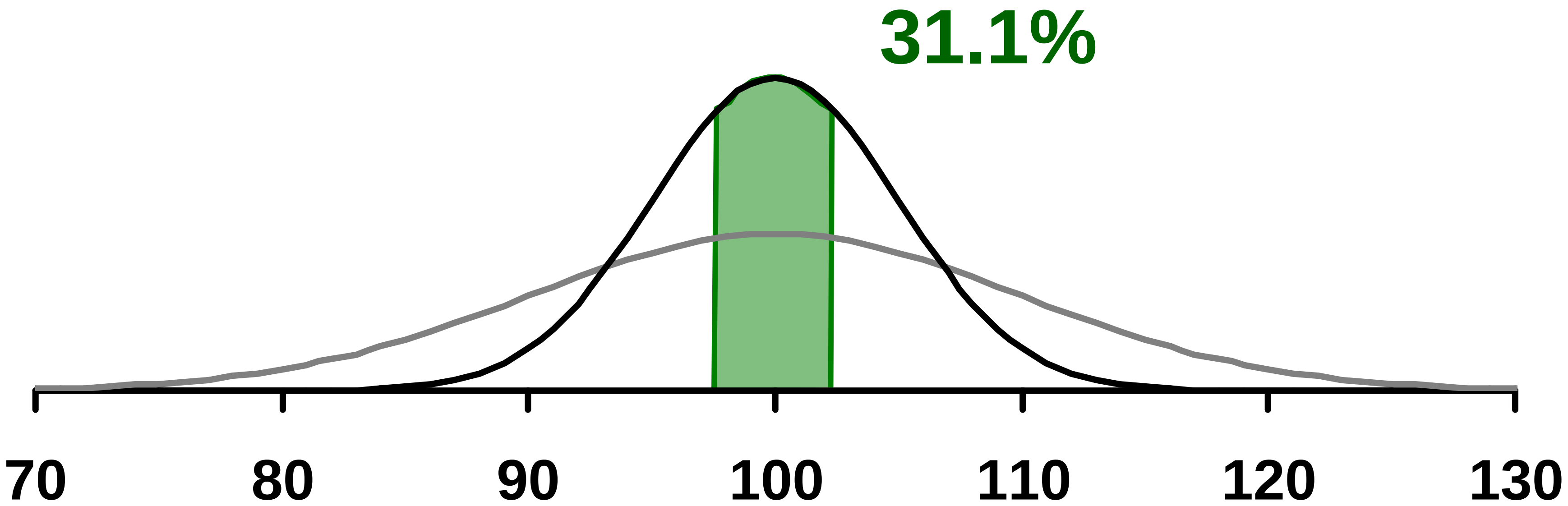
Central Limit Theorem, *n* = 1

$\sigma = 10$



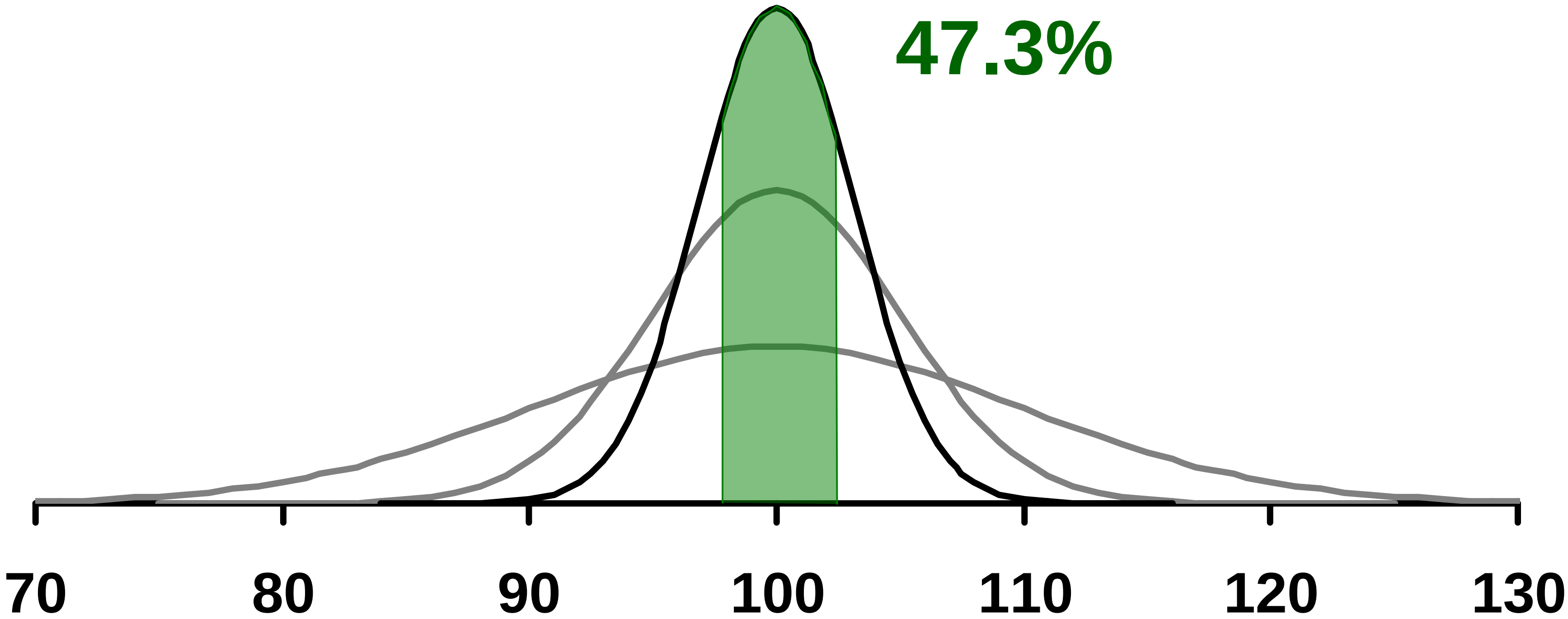
Central Limit Theorem, *n* = 4

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{4}} = 5$$



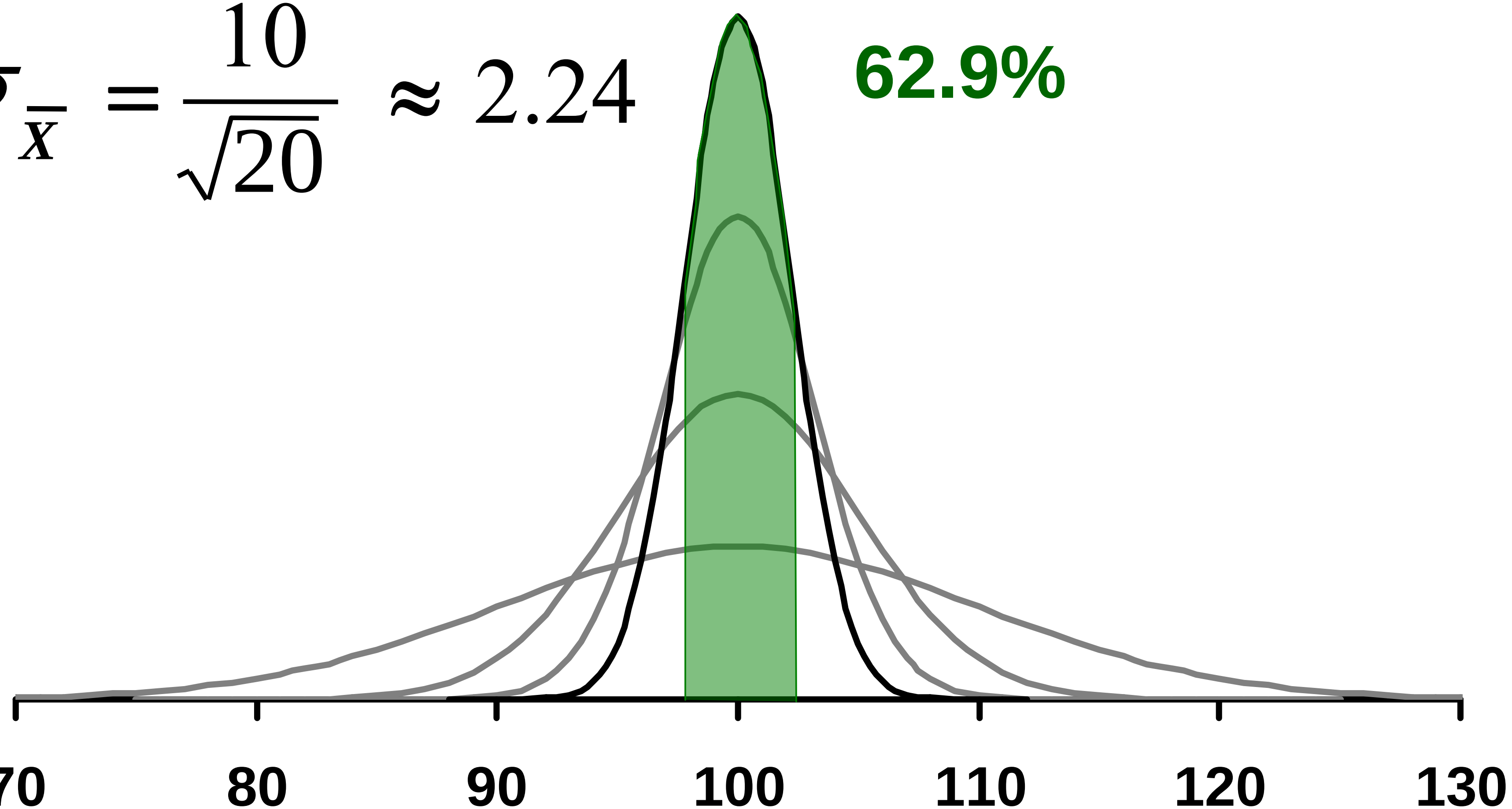
Central Limit Theorem, *n* = 10

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{10}} \cong 3.16$$



Central Limit Theorem, *n* = 20

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{20}} \approx 2.24$$



Estimating the Standard Deviation

- **Unbiased estimate**

$$s^2 \text{ for } \sigma^2 \qquad s \text{ for } \sigma \text{ ?}$$

- **Confidence Interval**

Sample Size

Estimator for σ?

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Sampling Distribution, s^2

$$\frac{(n-1)s^2}{\sigma^2} \Rightarrow \chi^2 - \text{distribution}$$

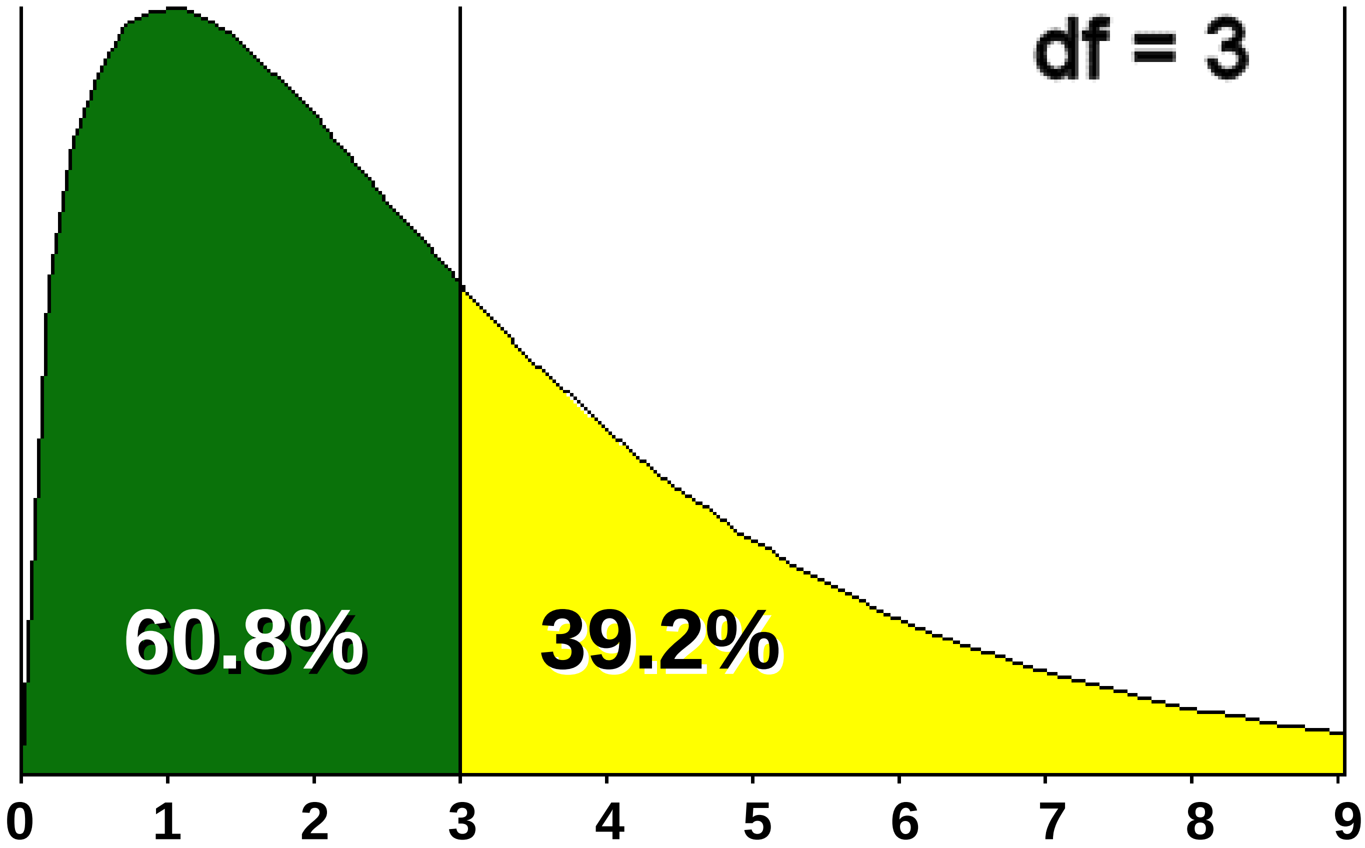
Since s^2 is an unbiased estimator for σ^2,

The average of s^2/σ^2 equals 1, so the mean of the χ^2 distribution = $(n - 1)$.

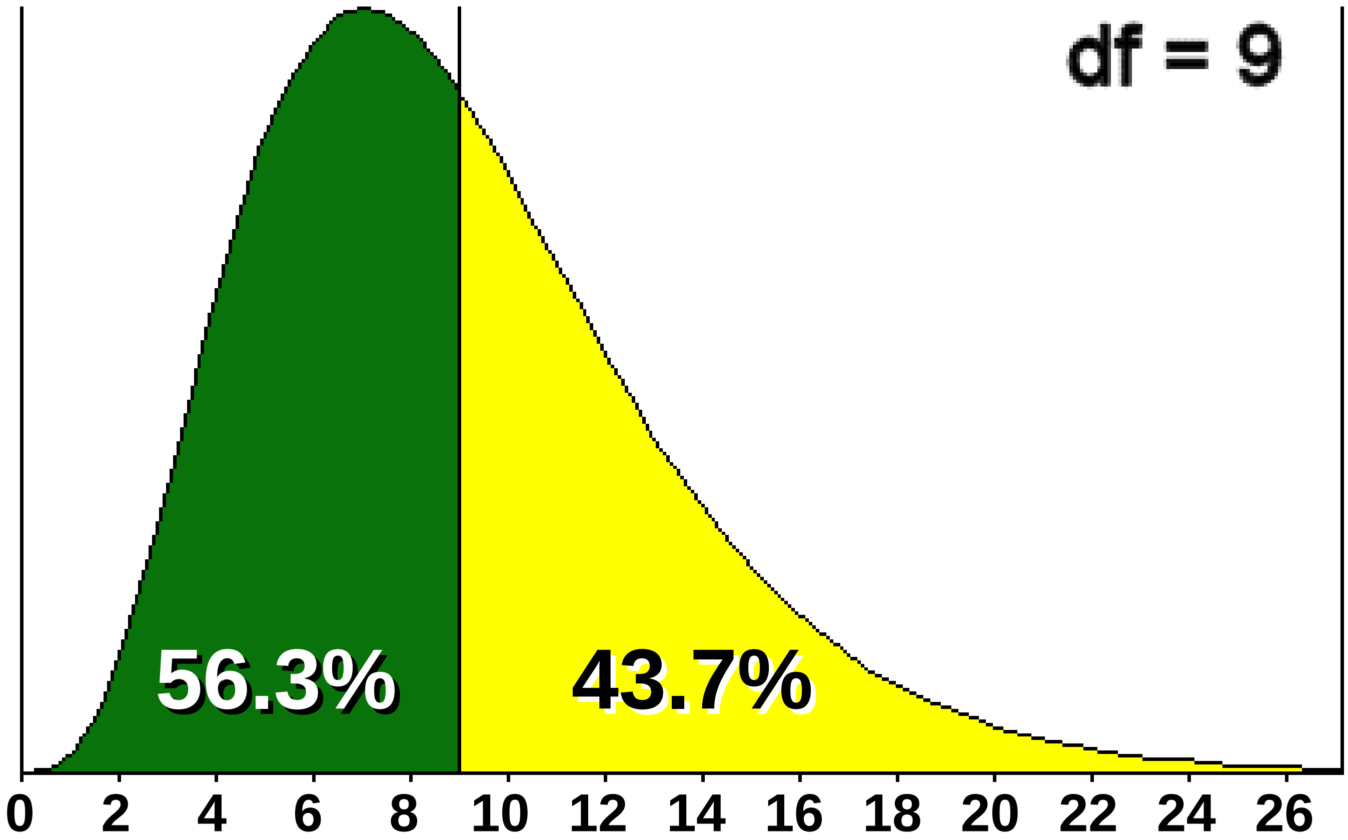
χ^2 Distribution

- Shape varies with degrees of freedom, i.e., $n - 1$.
- It is always positive.
- It is never symmetric.

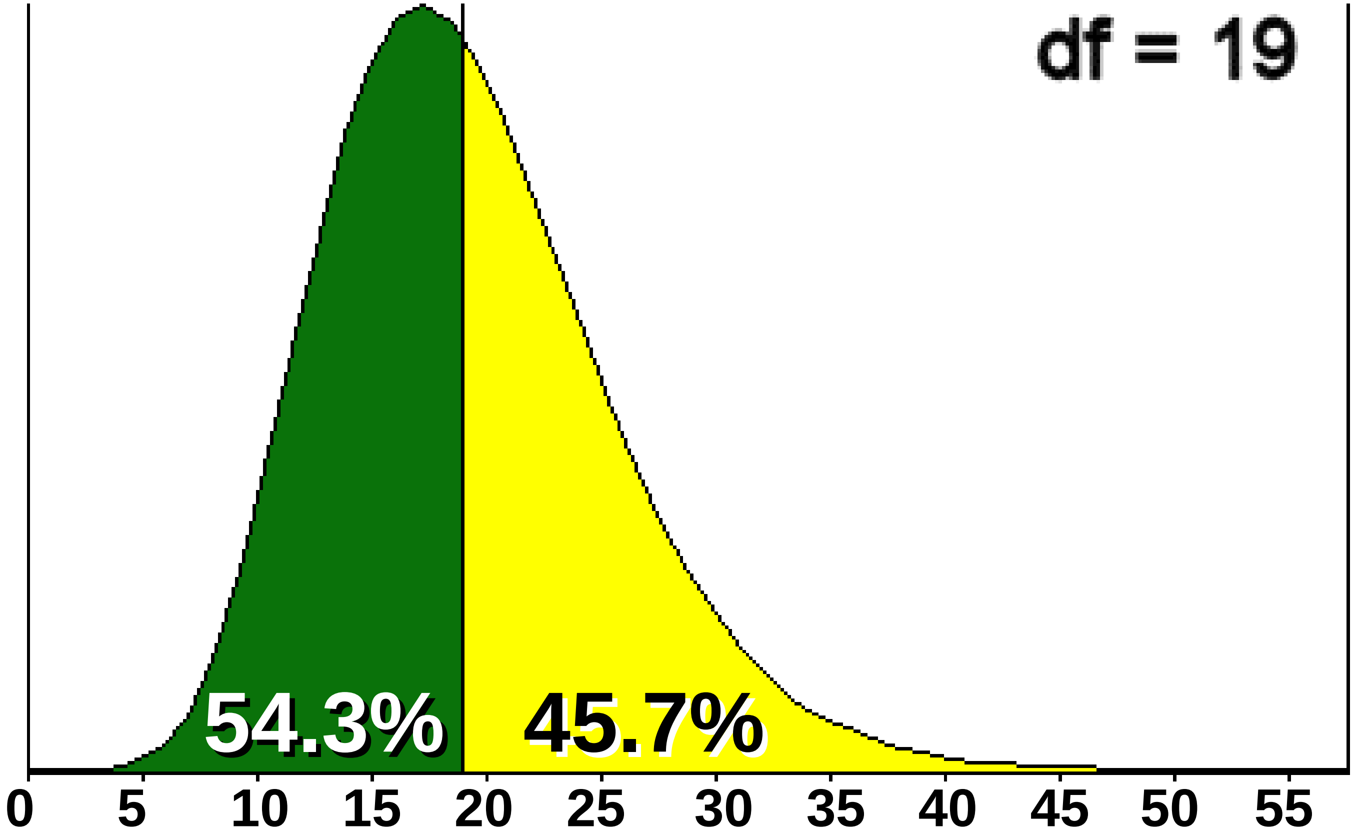
χ^2 Distribution, $n = 4$



χ^2 Distribution, n = 10



χ^2 Distribution, *n* = 20



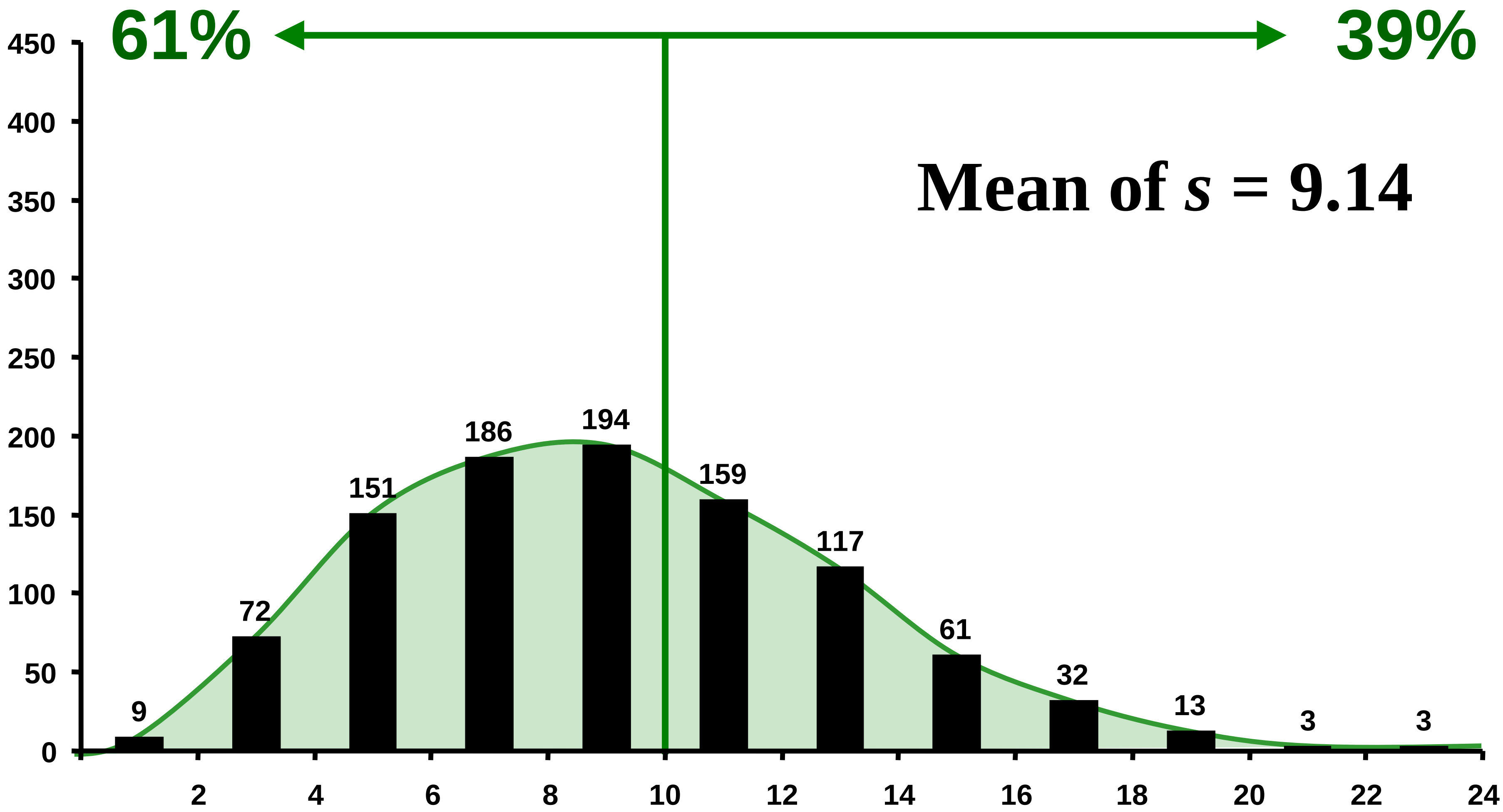
***Note*, for df = 200:**

51.3%	48.7%

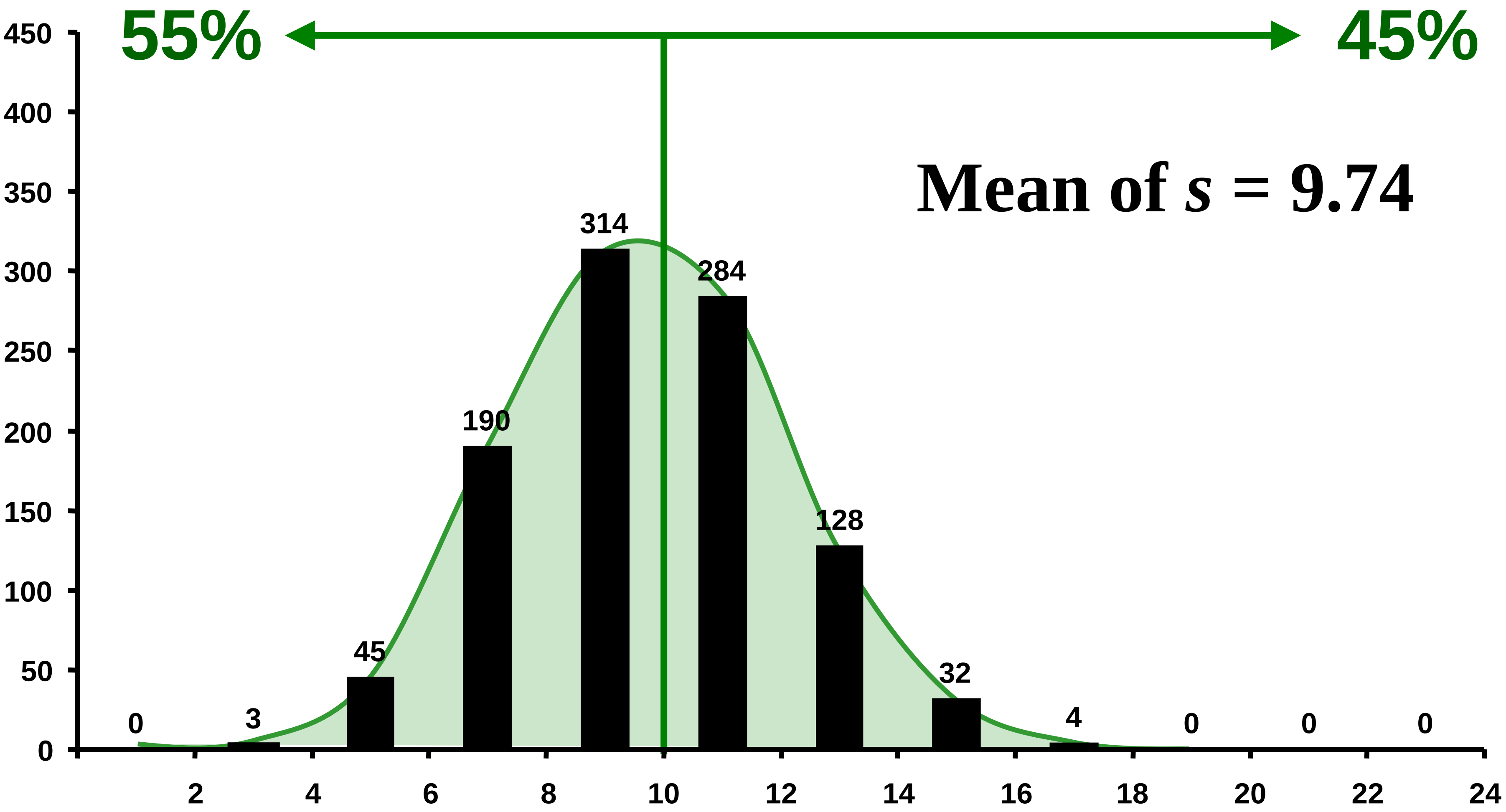
Sampling Example

- Normal Population with:
  - **Mean = 100**
  - **Std. Dev. = 10 (Variance = 100)**
- Draw 1000 Samples
  - ***n*** **= 4, 10, 20**
  - **Calculate each** ***s***

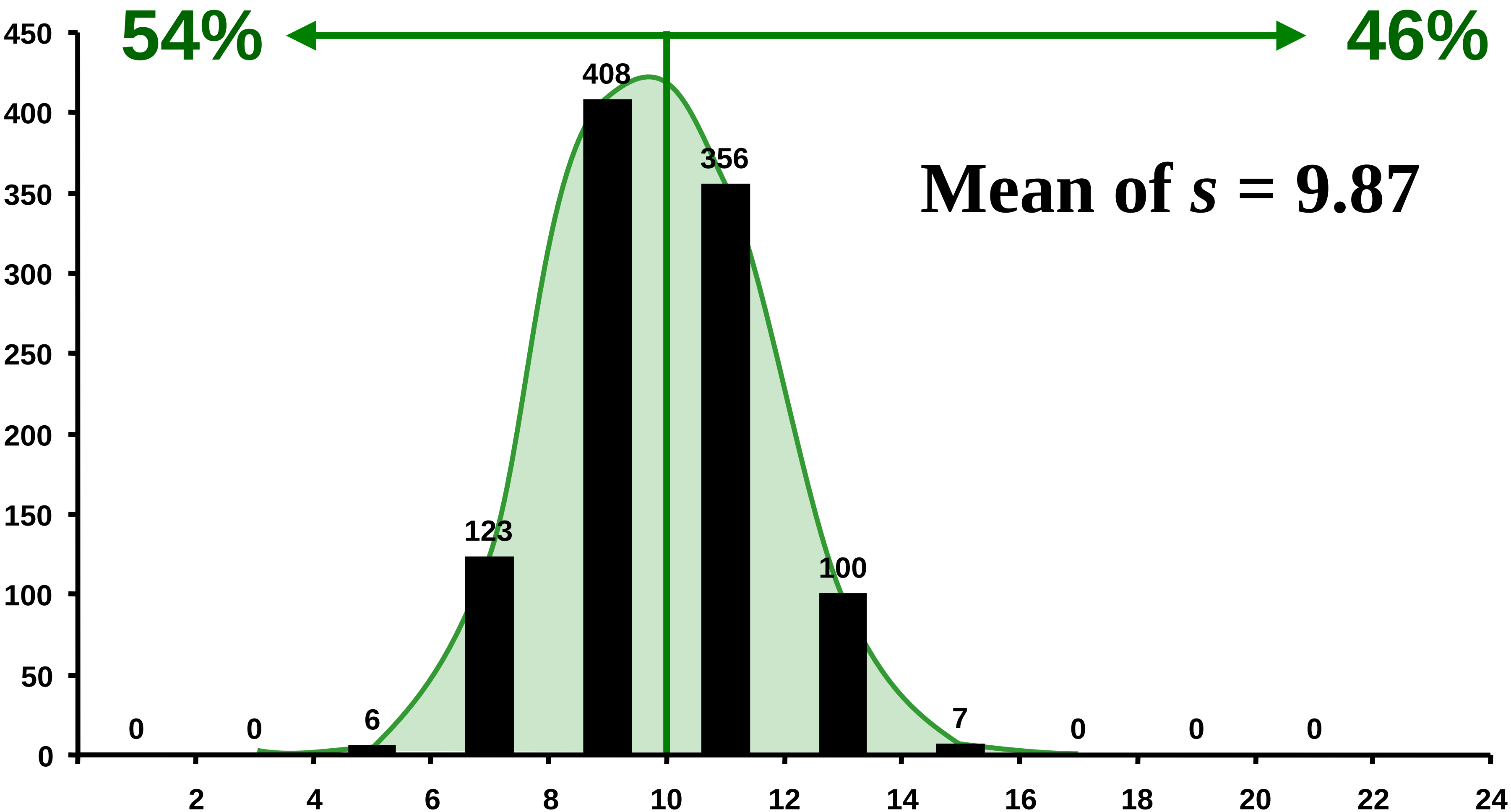
Estimating σ with s, $n = 4$



Estimating σ with s, $n = 10$



Estimating σ with s, n = 20



Estimators for σ ?

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Biased Low!

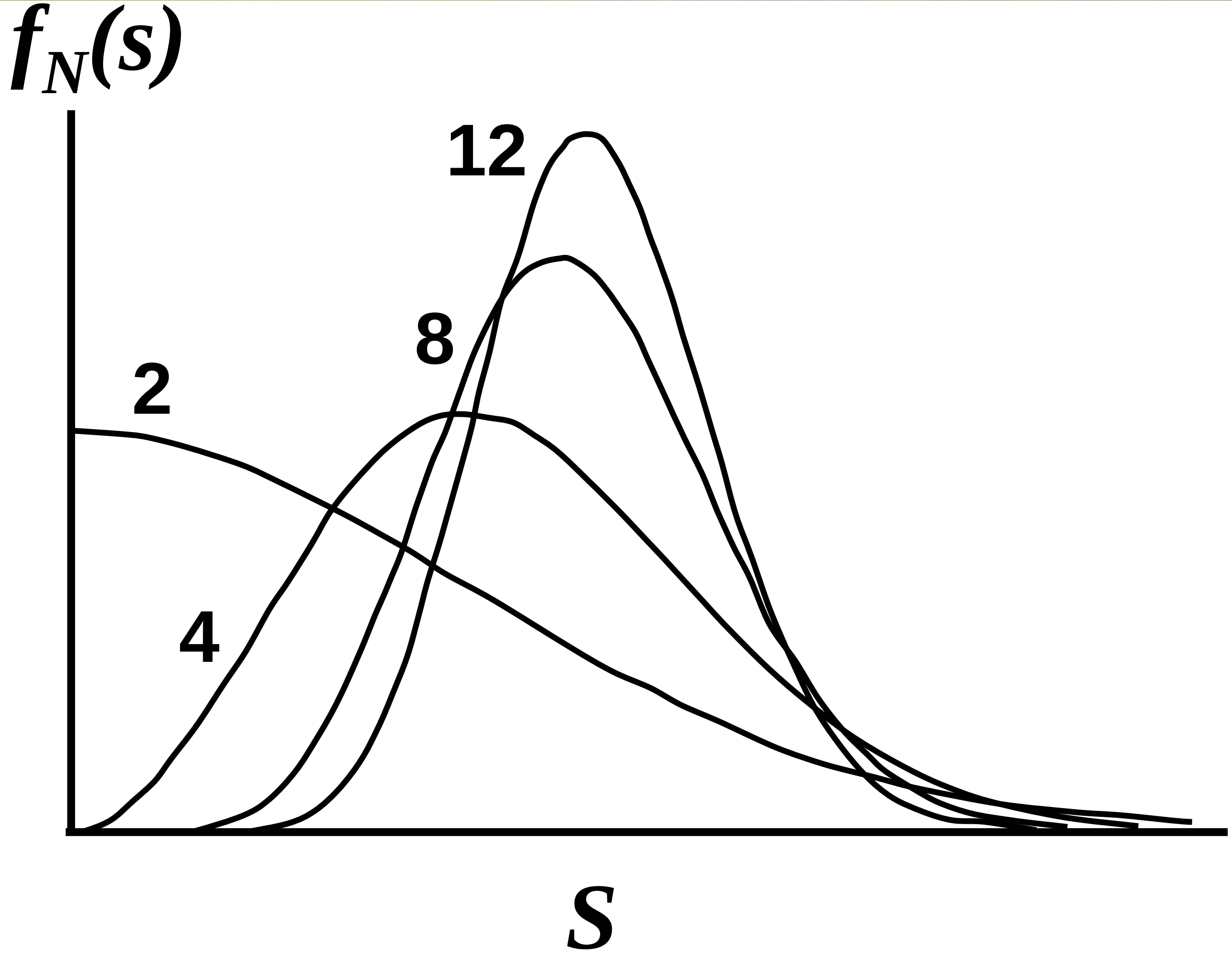
Bias < 1% when *n* > 26

Distribution of *s*, using N *d.f.*

$$s \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad \sigma^2 \equiv \frac{Ns^2}{N-1}$$

$$f_N(s) = 2\frac{\left(\frac{N}{2\sigma^2}\right)^{(N-1)/2}}{\Gamma\left(\frac{1}{2}(N-1)\right)} e^{-Ns^2/(2\sigma^2)} s^{N-2}$$

Distribution of *s*, using N *d.f.*



Unbiased Estimator for σ

$$(c(n) \text{ or } c_4)s \qquad \left[\frac{\Gamma\left(\frac{n-1}{2}\right)\sqrt{n-1}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2}}\right]s$$

Wow!

where $\Gamma(z) \equiv \int_0^\infty t^{z-1}e^{-t}\,dt$

Unbiased Estimator for σ

$$c(n) = \begin{cases} \dfrac{2^{n-2.5}\sqrt{n-1}}{(n-2)\sqrt{\dbinom{n-3}{\frac{n-3}{2}}}} & n \text{ odd} \\[2em] \dbinom{n-3}{\frac{n-2}{2}}\dfrac{\sqrt{\pi(n-1)}}{2^{n-2.5}} & n \text{ even}, n \ge 4 \end{cases}$$

Wow!

Unbiased Estimator for σ

$$c's$$

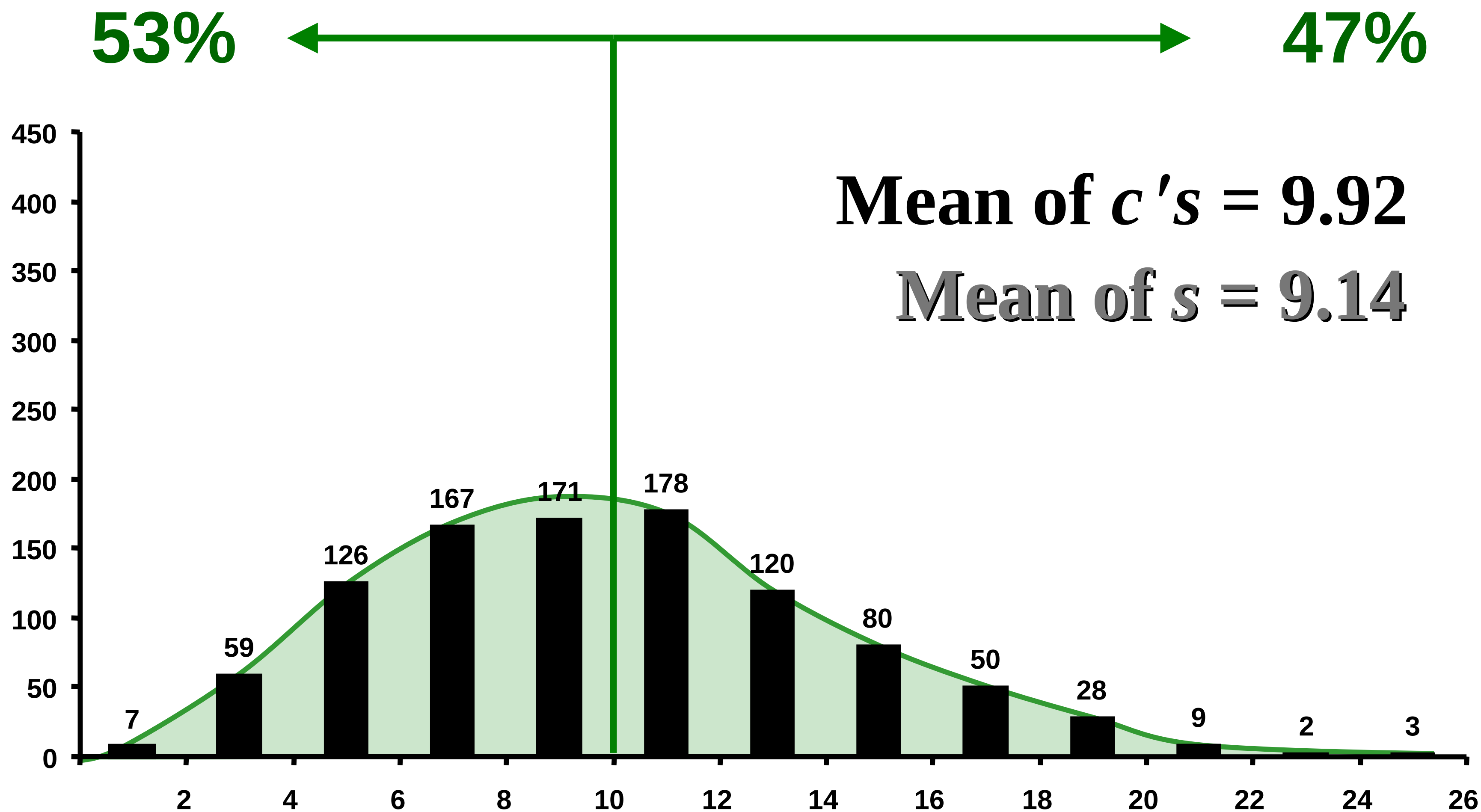
$$\left(\frac{n-0.75}{n-1}\right)s$$

Accurate to ¼ %
for all *n*.

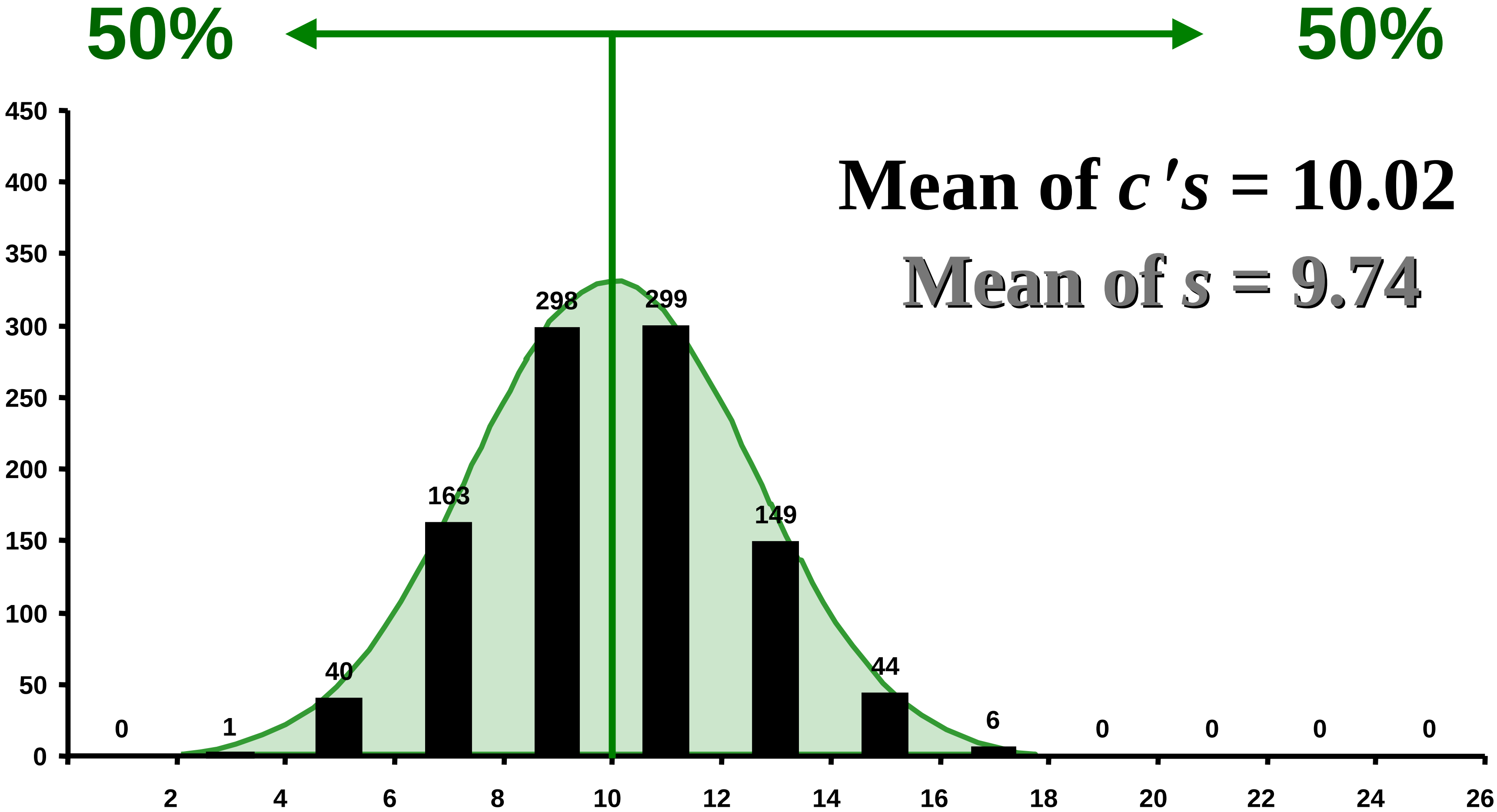
Sampling Example

- Normal Population with:
  - **Mean = 100**
  - **Std. Dev. = 10 (Variance = 100)**
- Draw 1000 Samples
  - $n = 4, 10, 20$
  - **Calculate each** $c's$

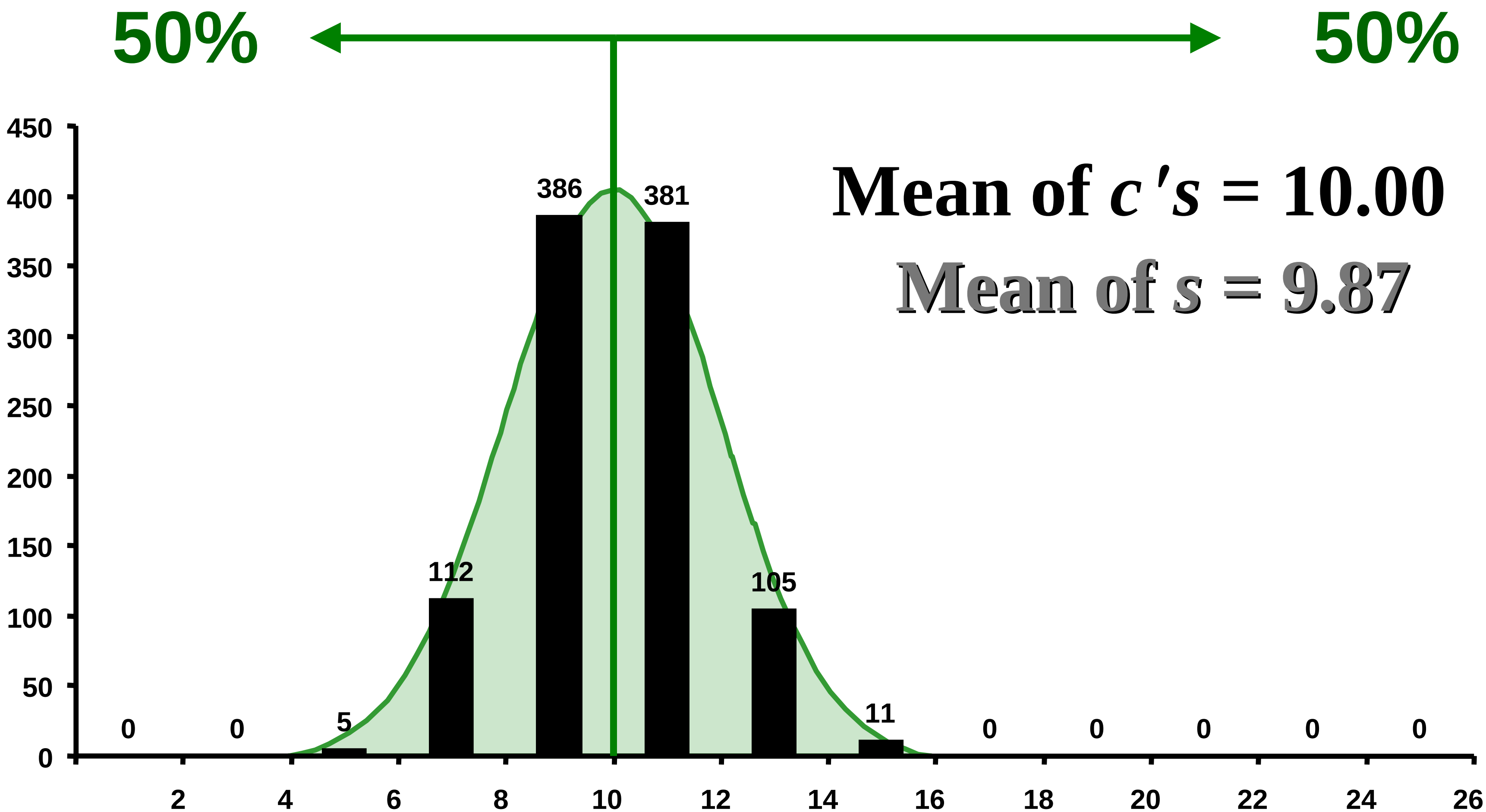
Estimating σ with $c's$, $n = 4$



Estimating σ with c's, n = 10



Estimating σ with $c's$, $n = 20$



Estimating the Standard Deviation

- **Unbiased estimate**

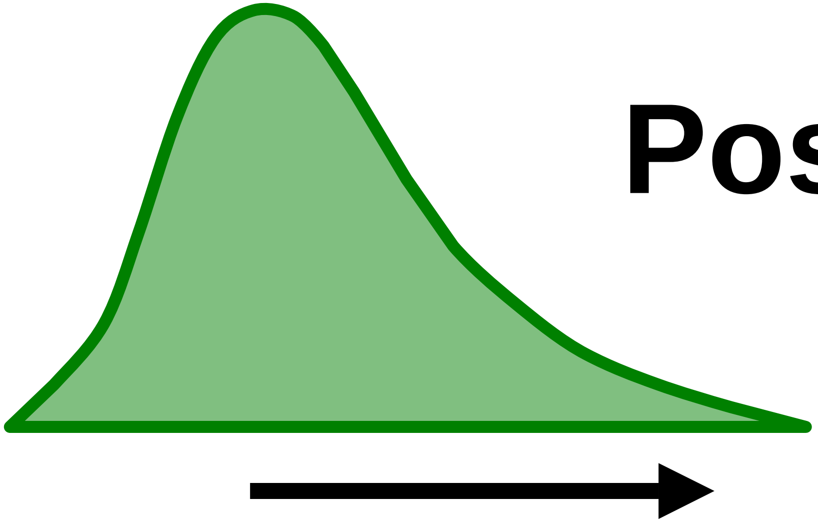
$$c's \quad versus \quad s$$

Probability Distribution

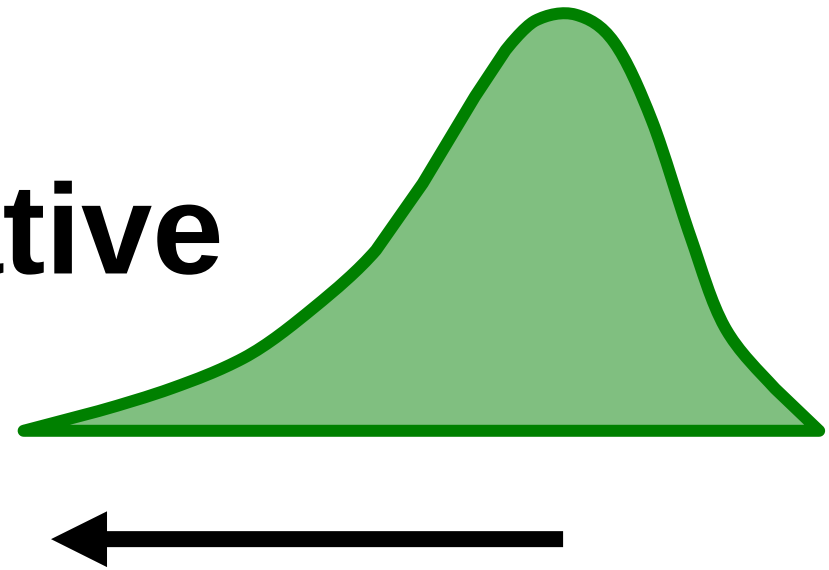
- Normality Assumption?
- Approximately Normal?
- Skewed? *(median)*
- Goodness of Fit to Normal?
  - **Histogram.**
  - **Normal probability paper.**
  - **Statistical test: ~~Chi Square, K–S?~~**
  - **Shapiro-Wilk, Anderson-Darling**

Skewed Distributions

Positive

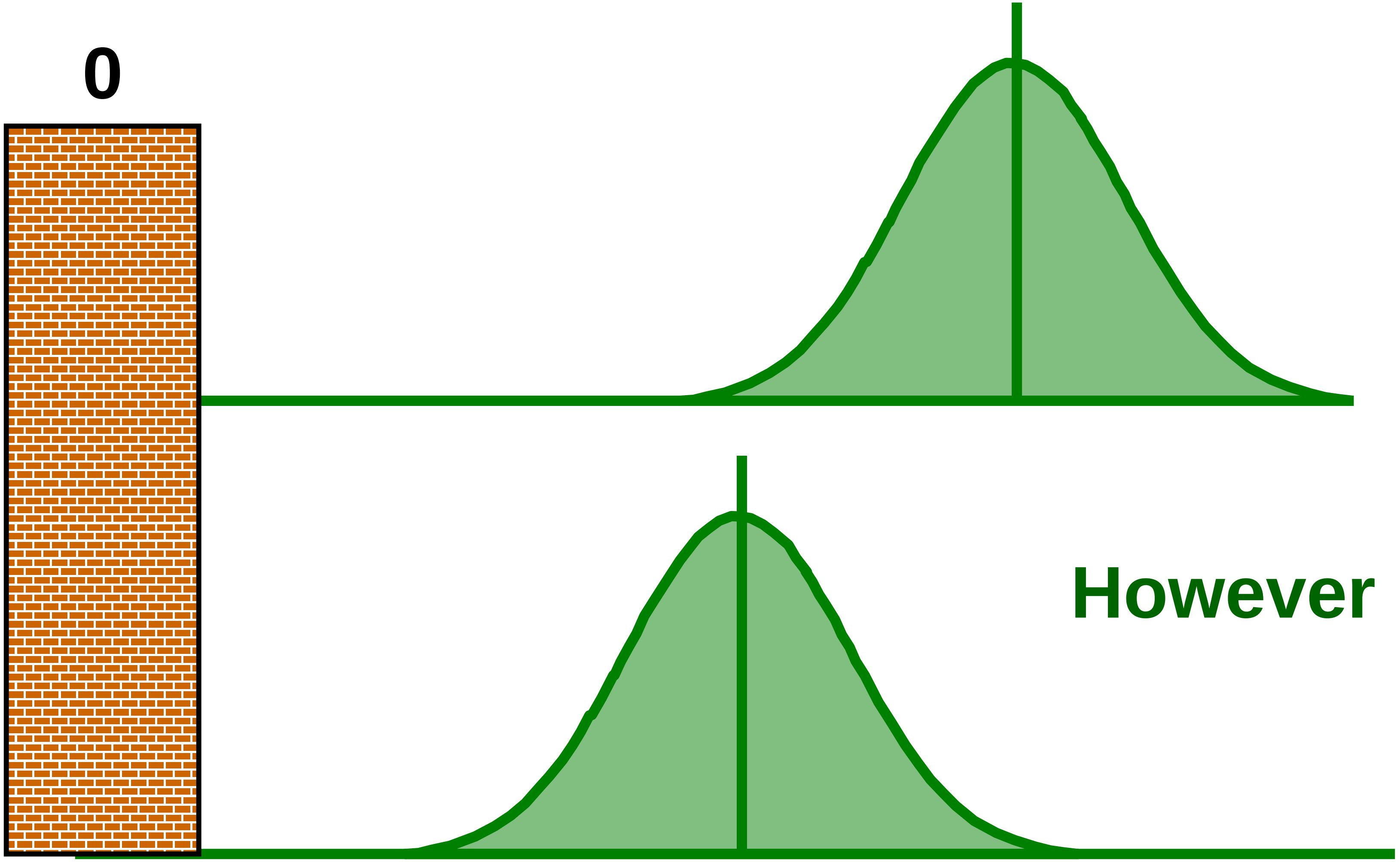


Negative

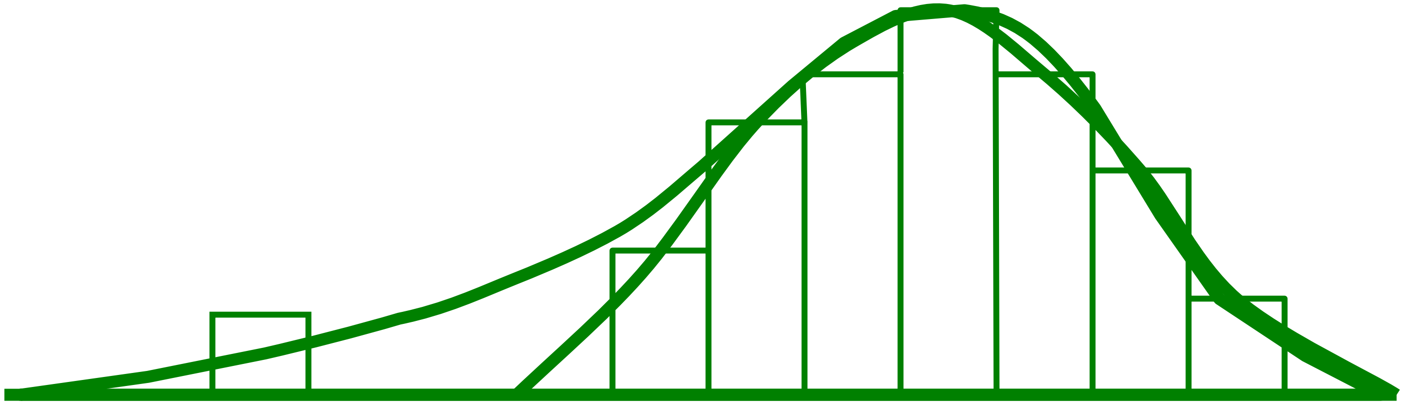


Pavement Materials

Few materials have skewed distributions.

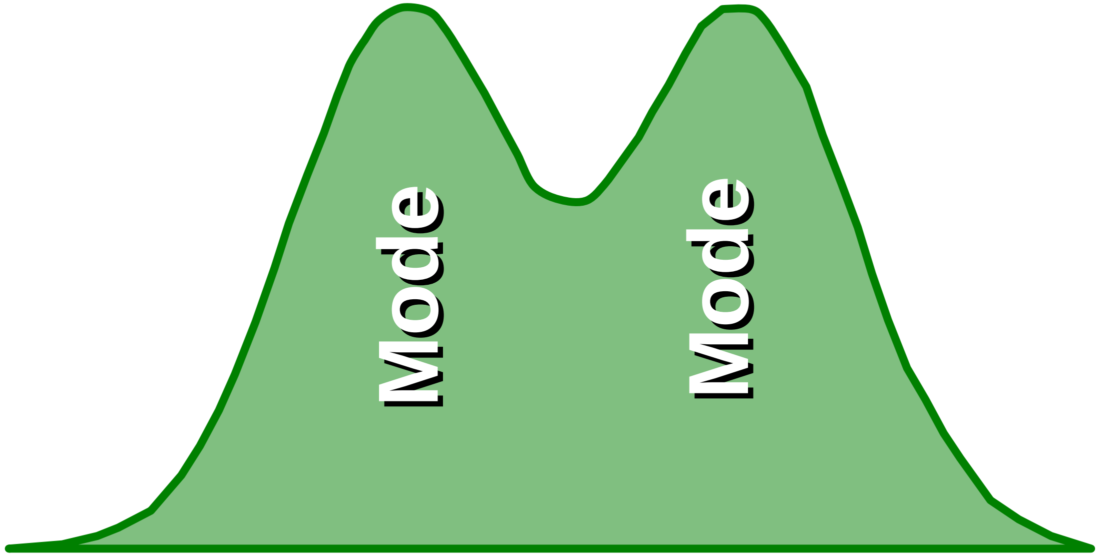


Outliers and Skewness

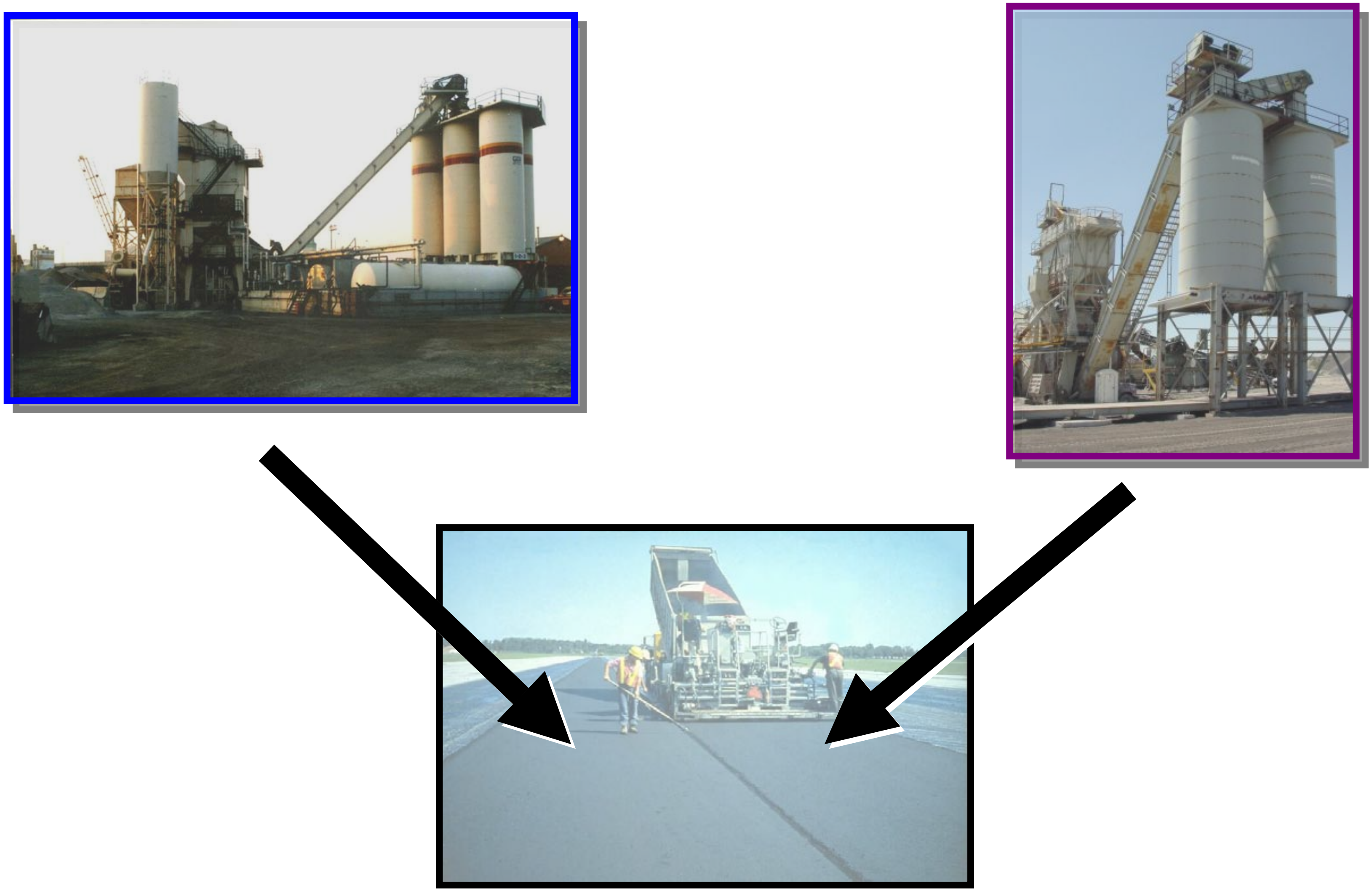


ASTM E-178

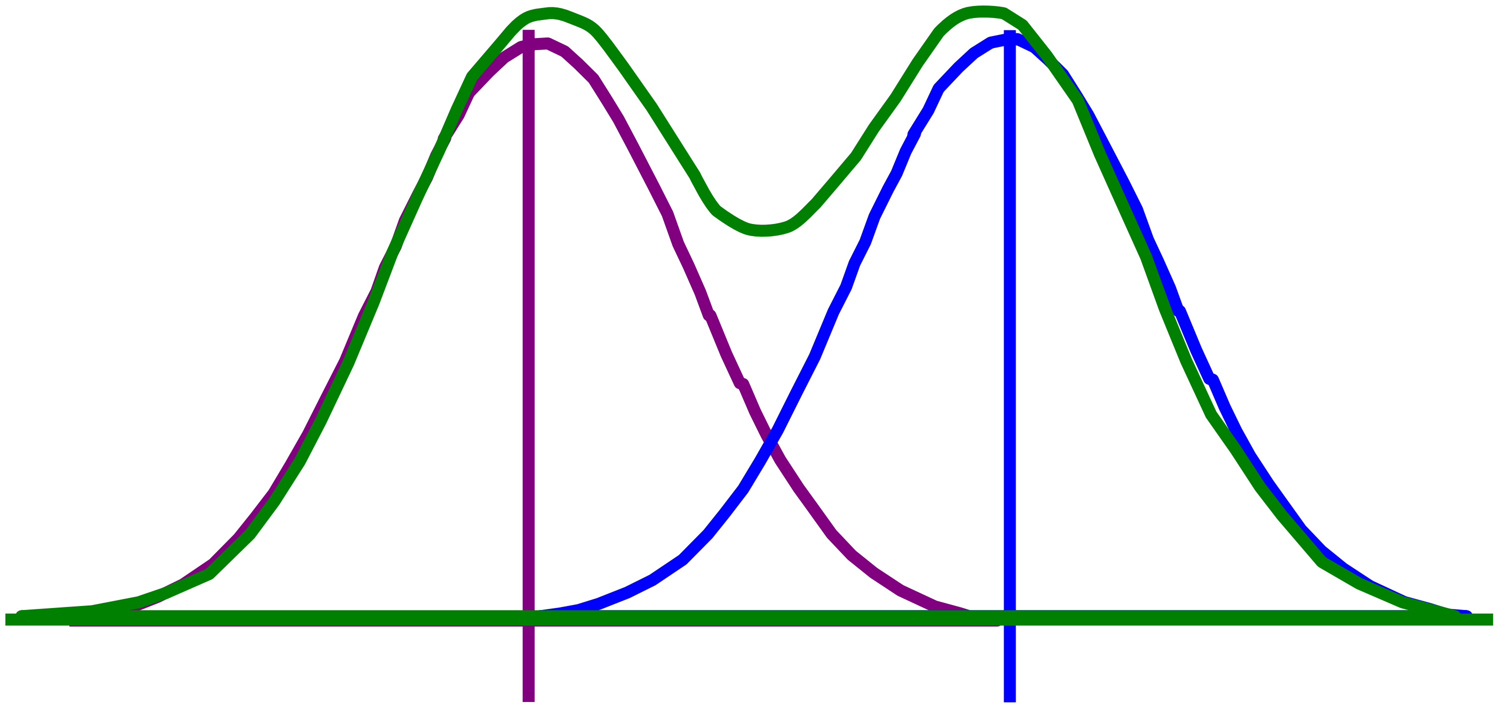
Bimodal Distributions



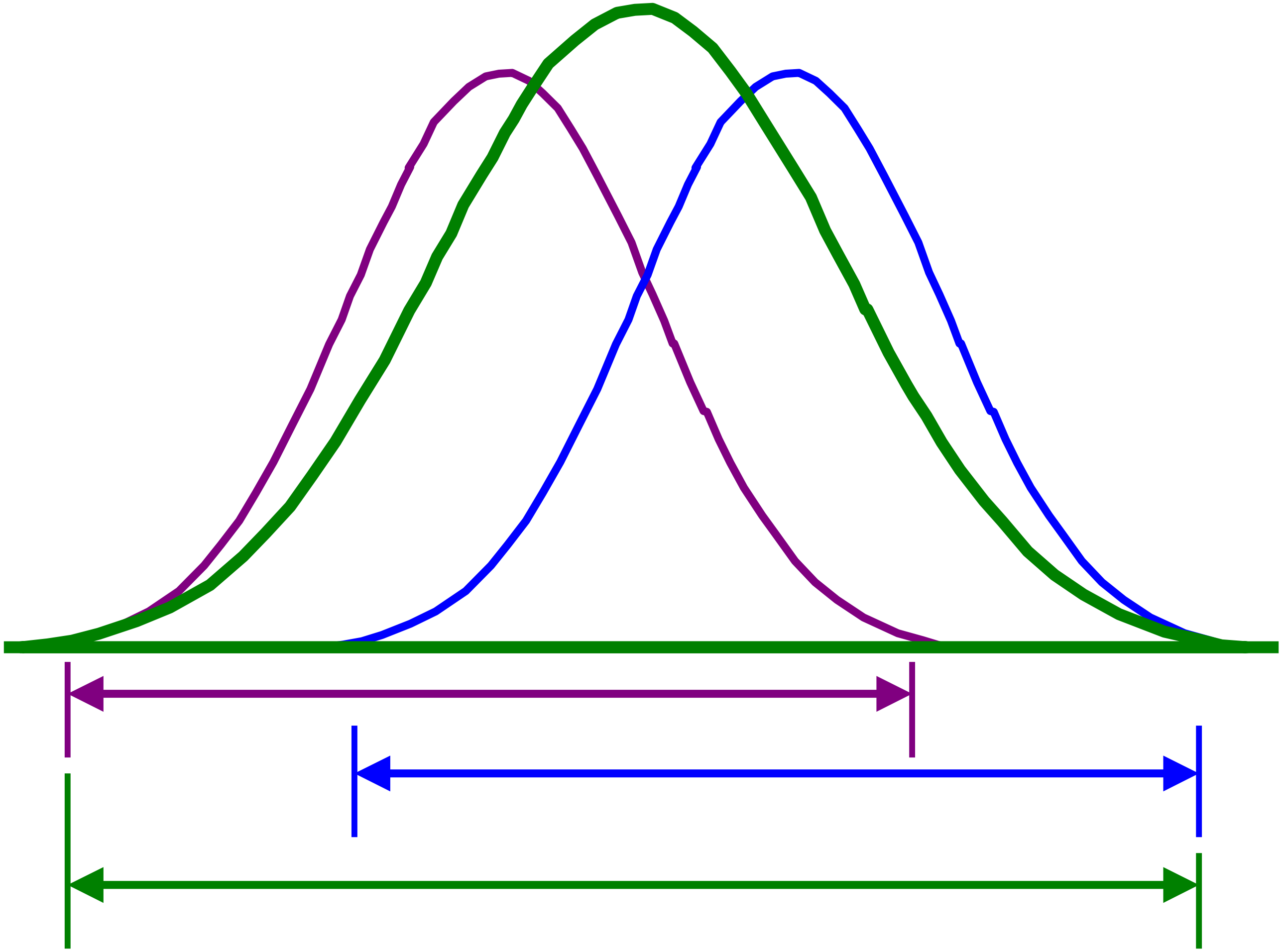
How?



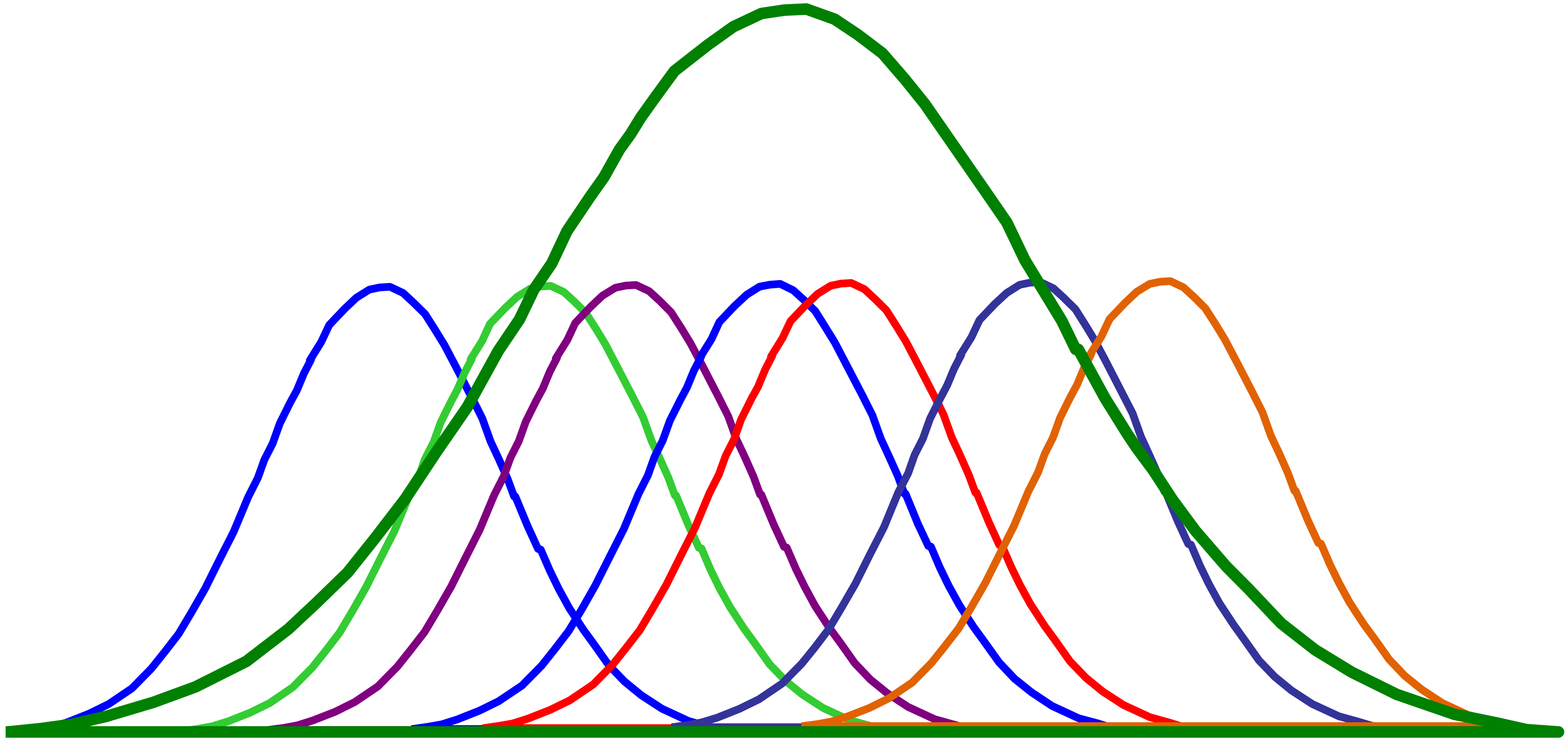
Obvious



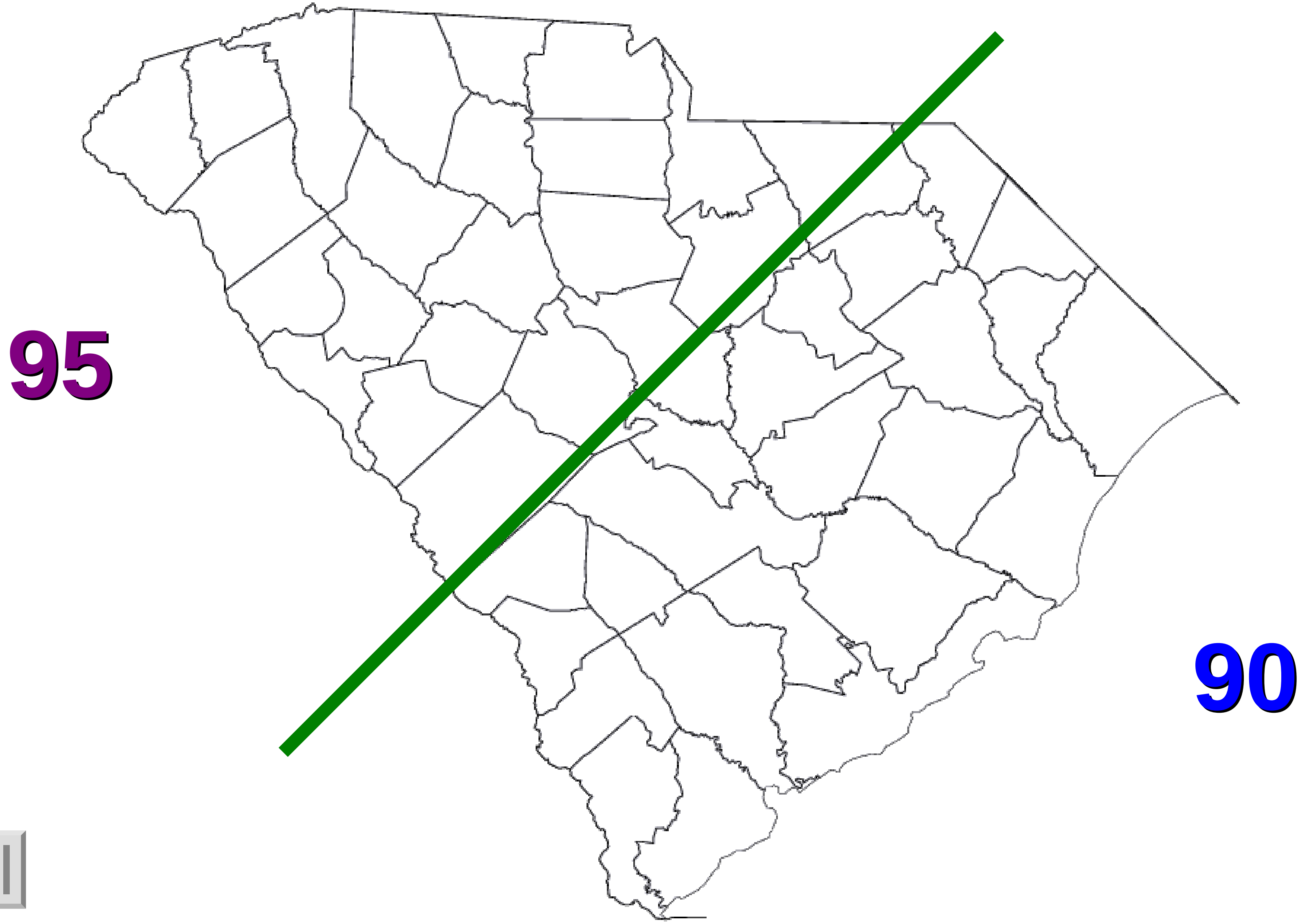
Not So Obvious?



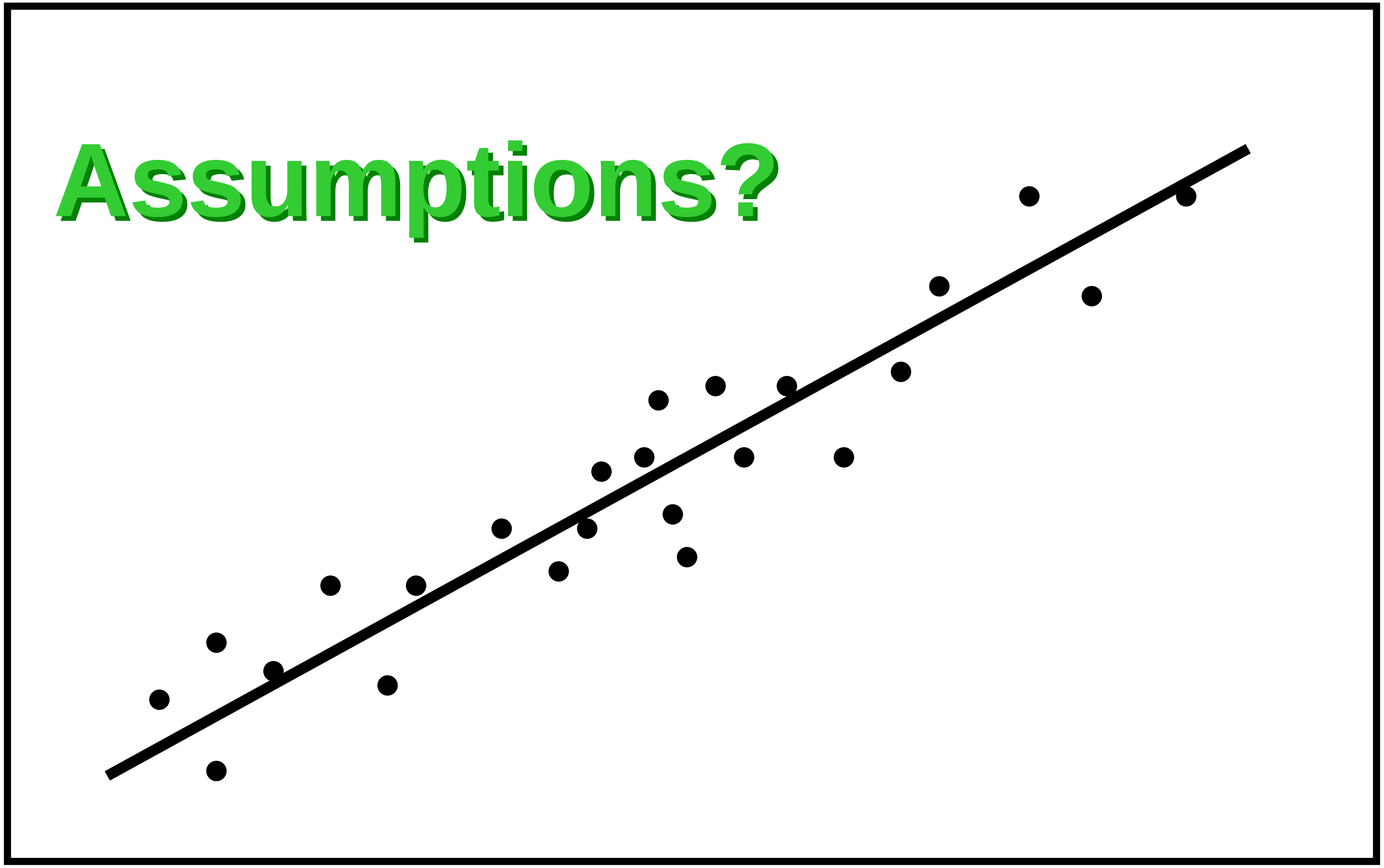
Multiple Raters?



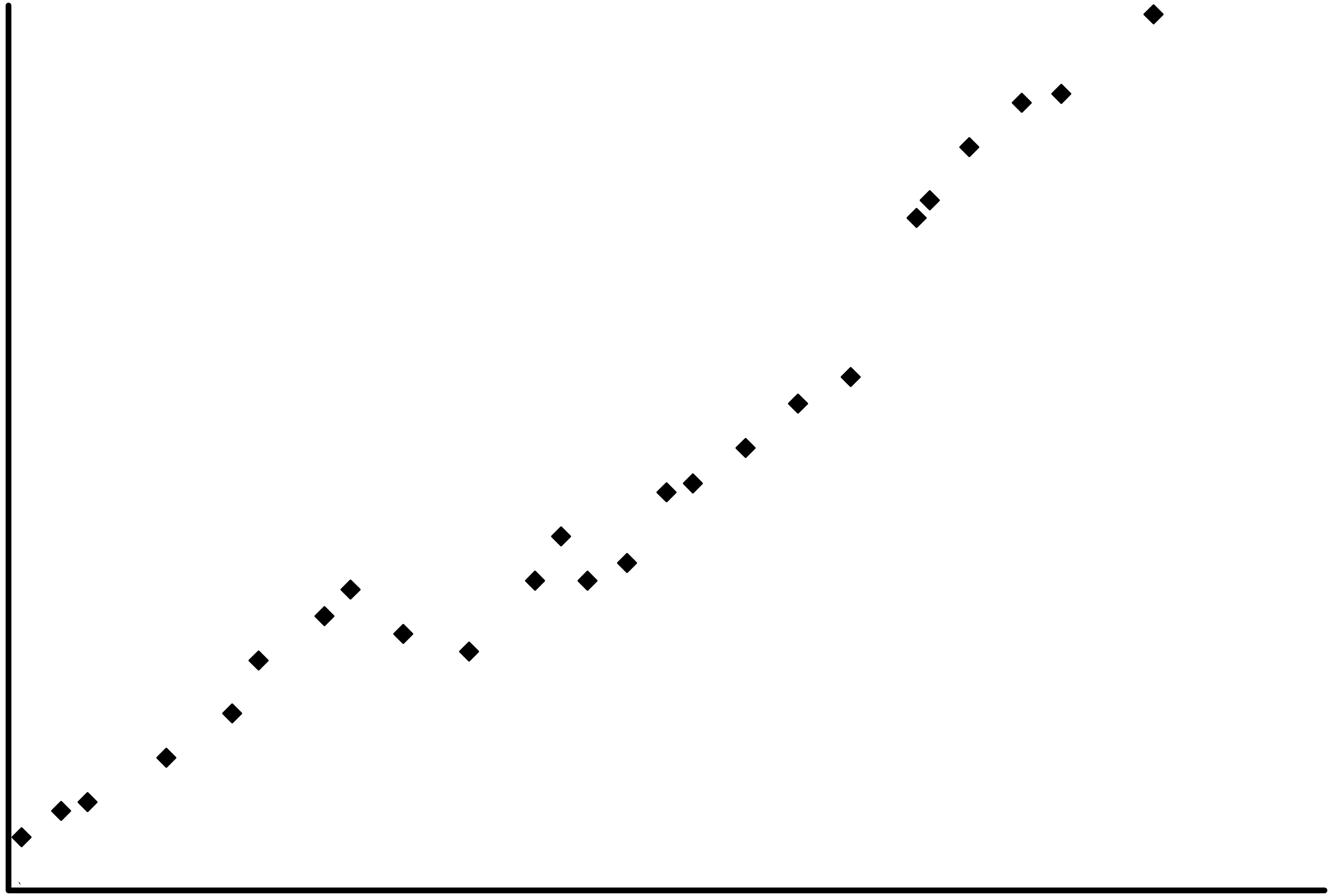
Drawing Conclusions from Data?



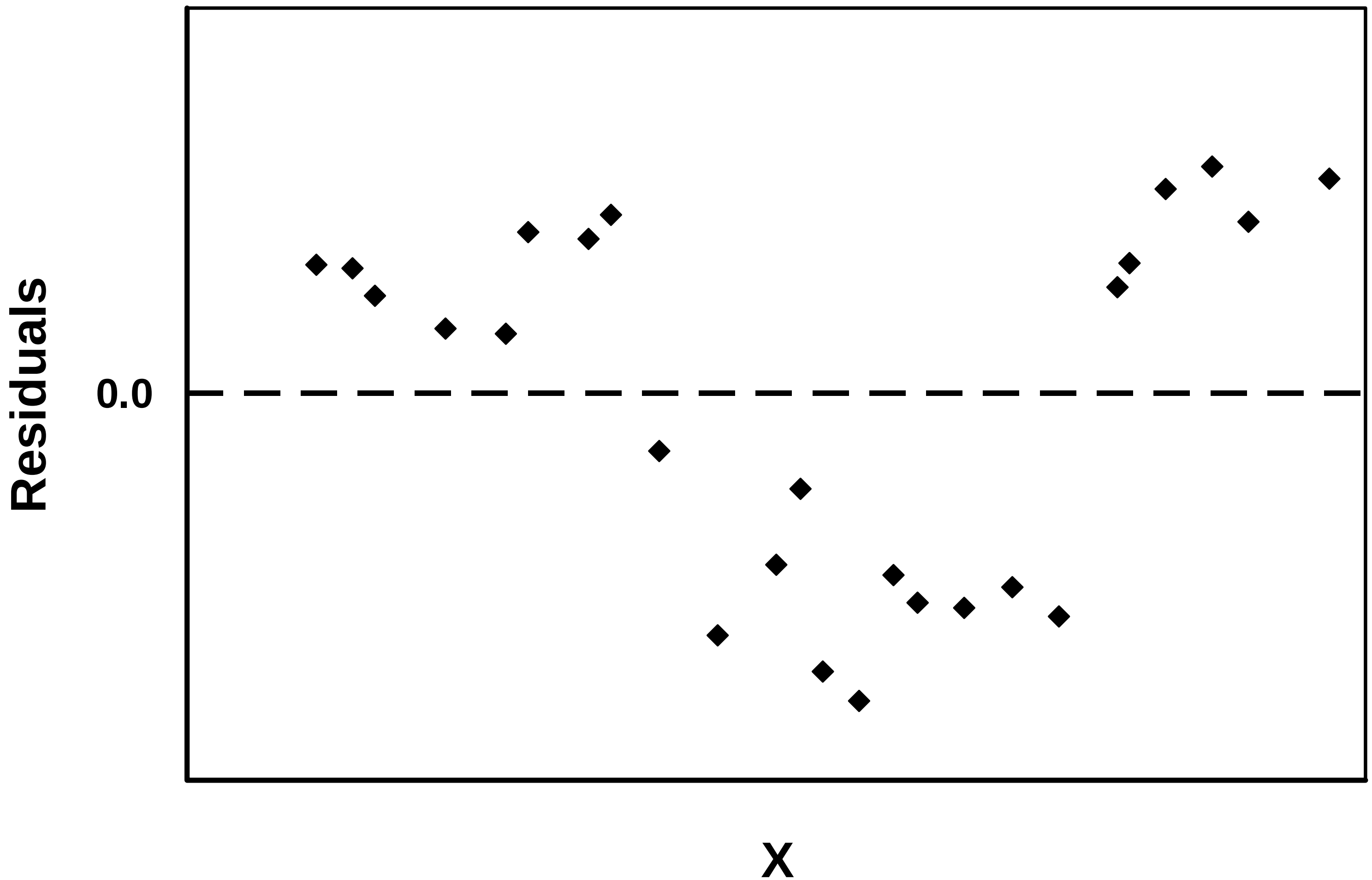
Regression Lines?



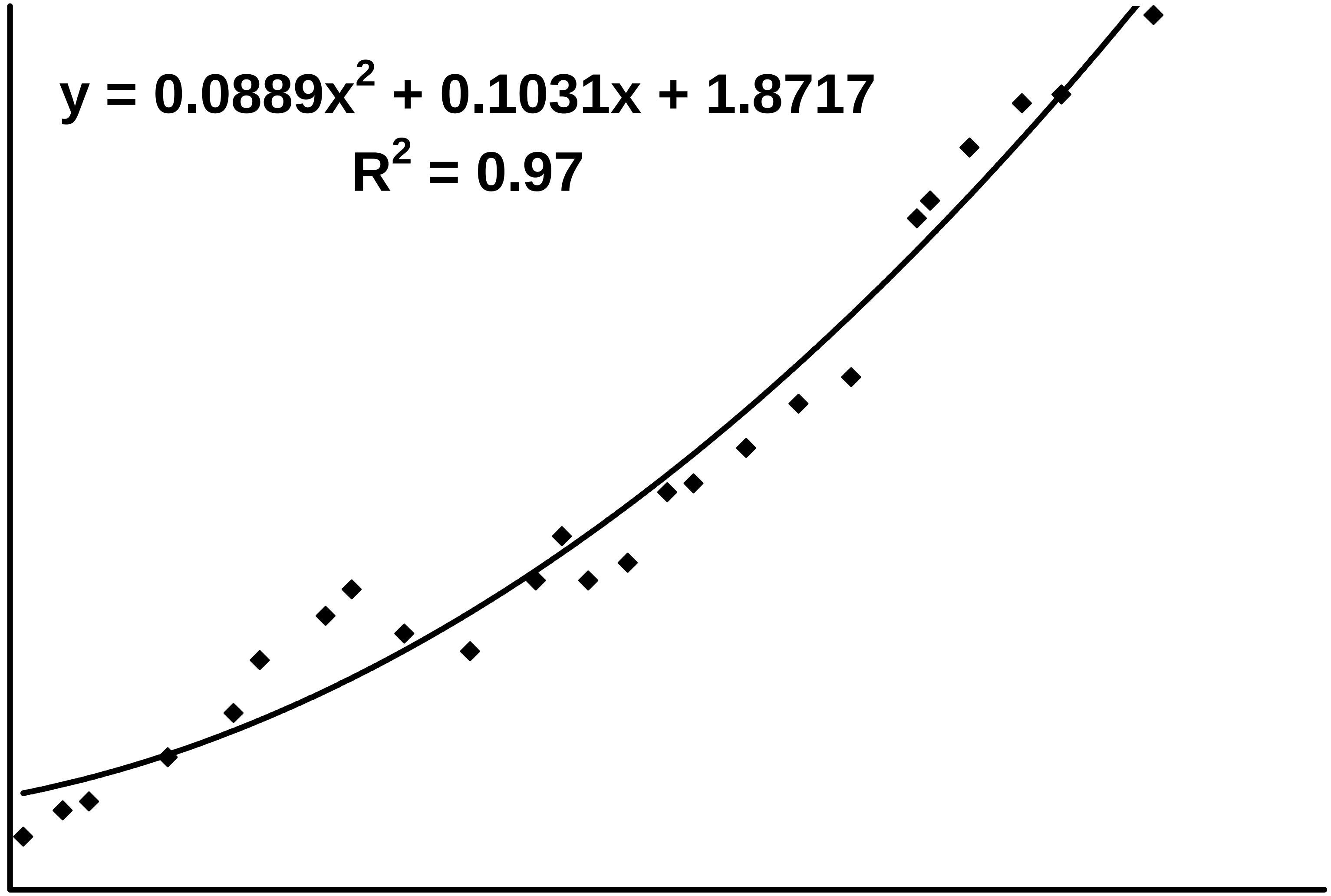
Regression & Normality?



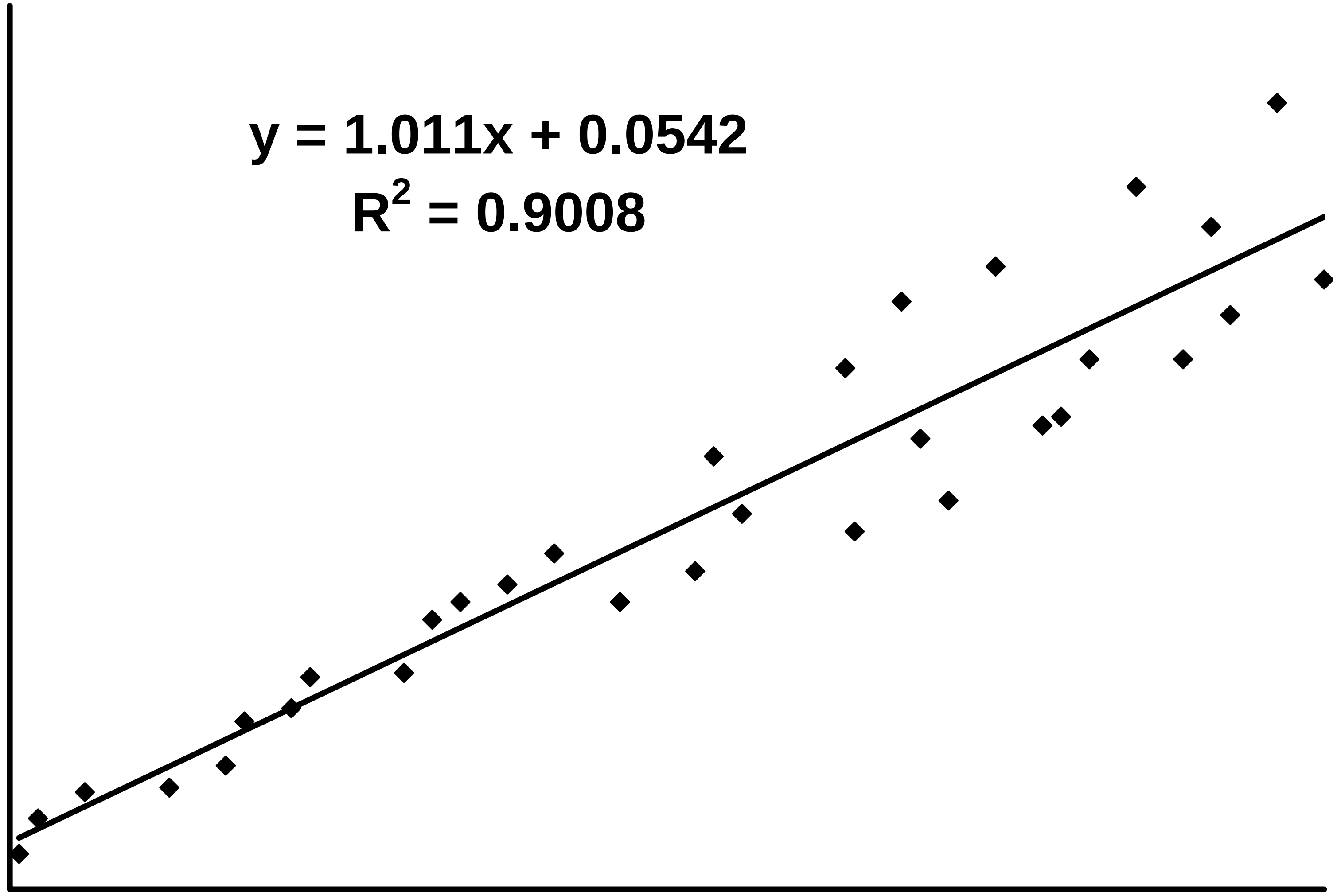
Residuals Normal?



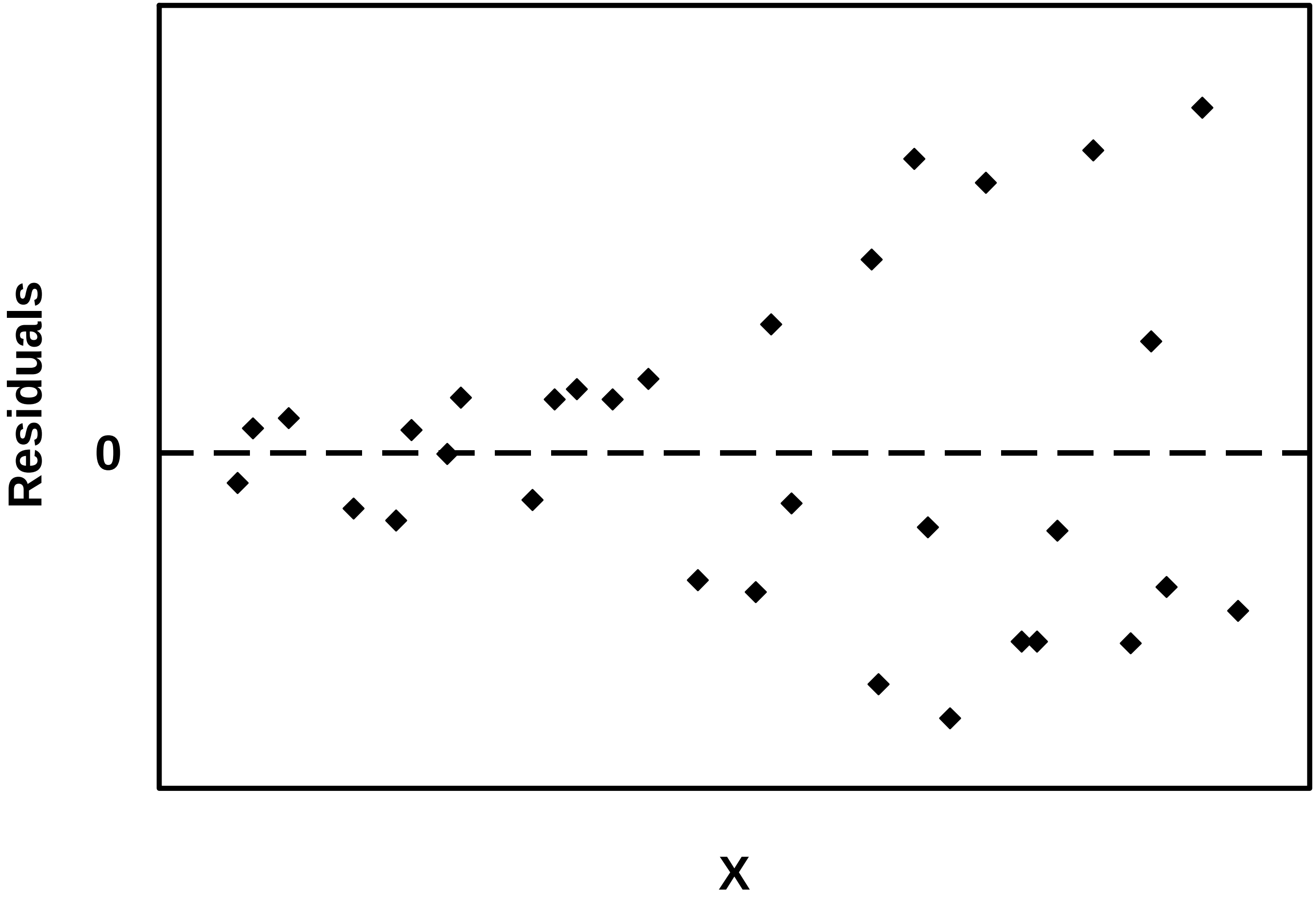
Correct Model?



Regression & Variability?



Regression & Variability?



Classic Example

For Each of Four (x , y) Data Sets:

$n = 11$

$\bar{x} = 9.0$

$\bar{y} = 7.5$

$s_x = 3.32$

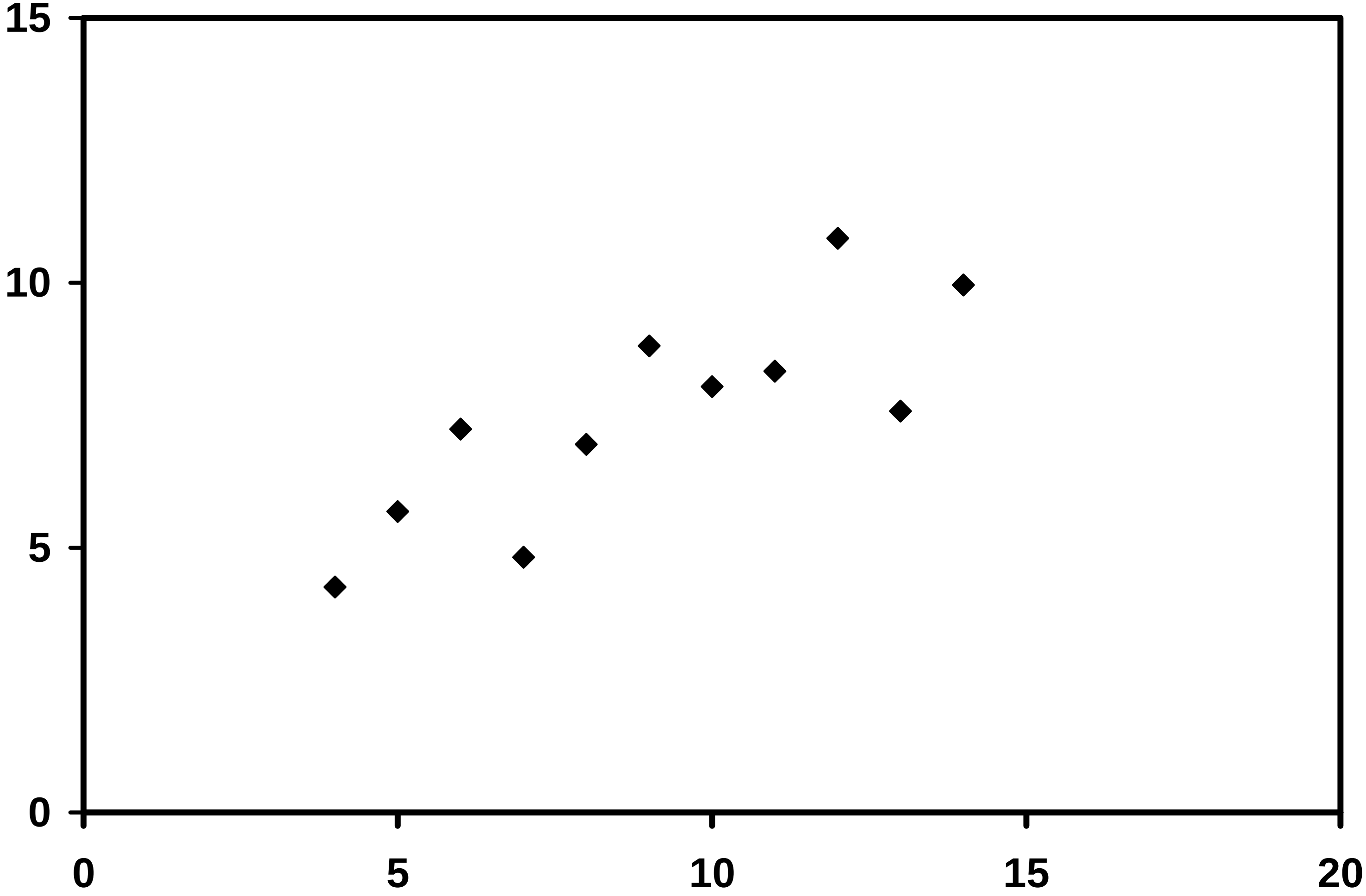
$s_y = 2.03$

$r = 0.82$

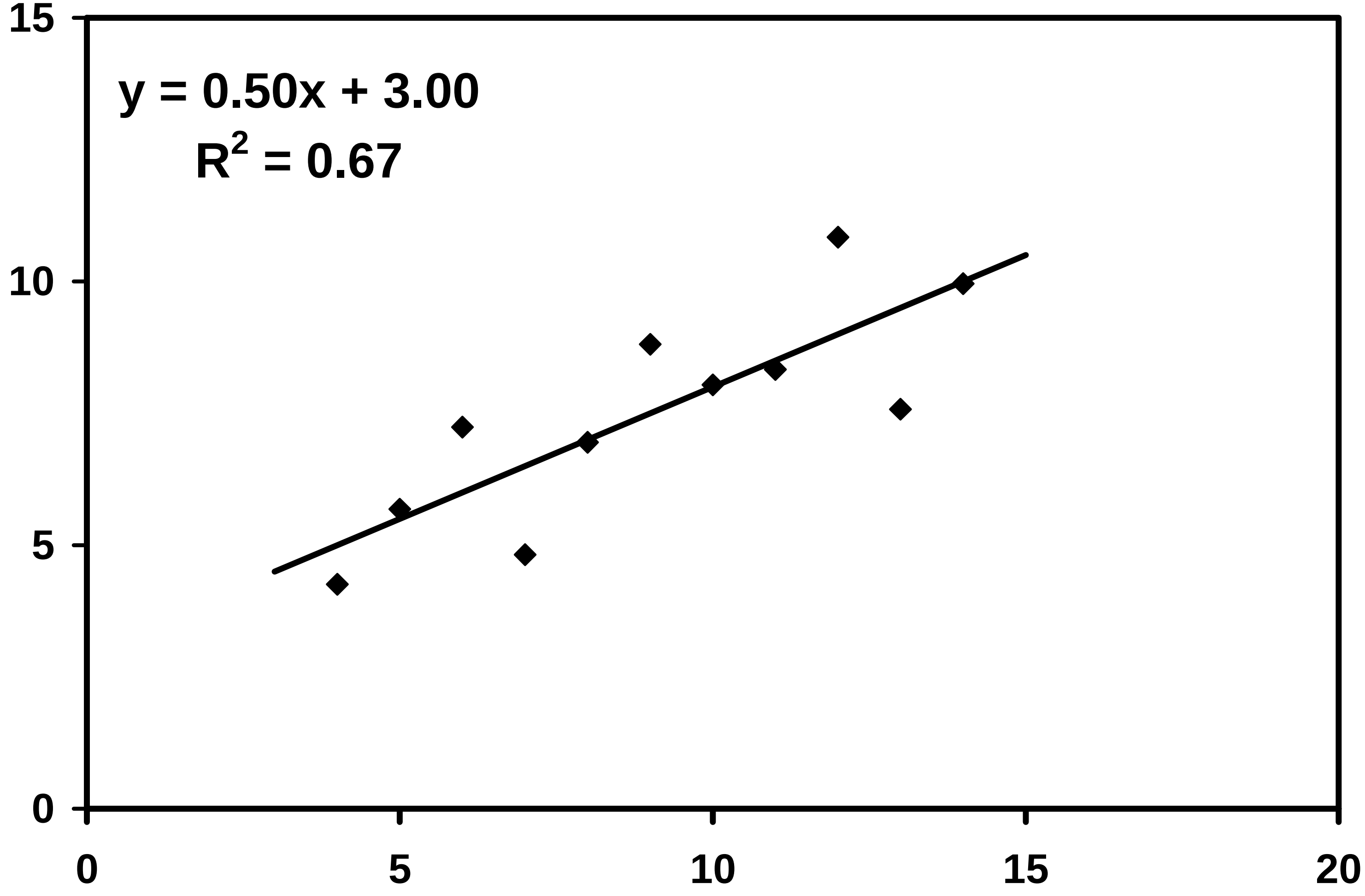
$y = 0.50x + 3.00$

$R^2 = 0.67$

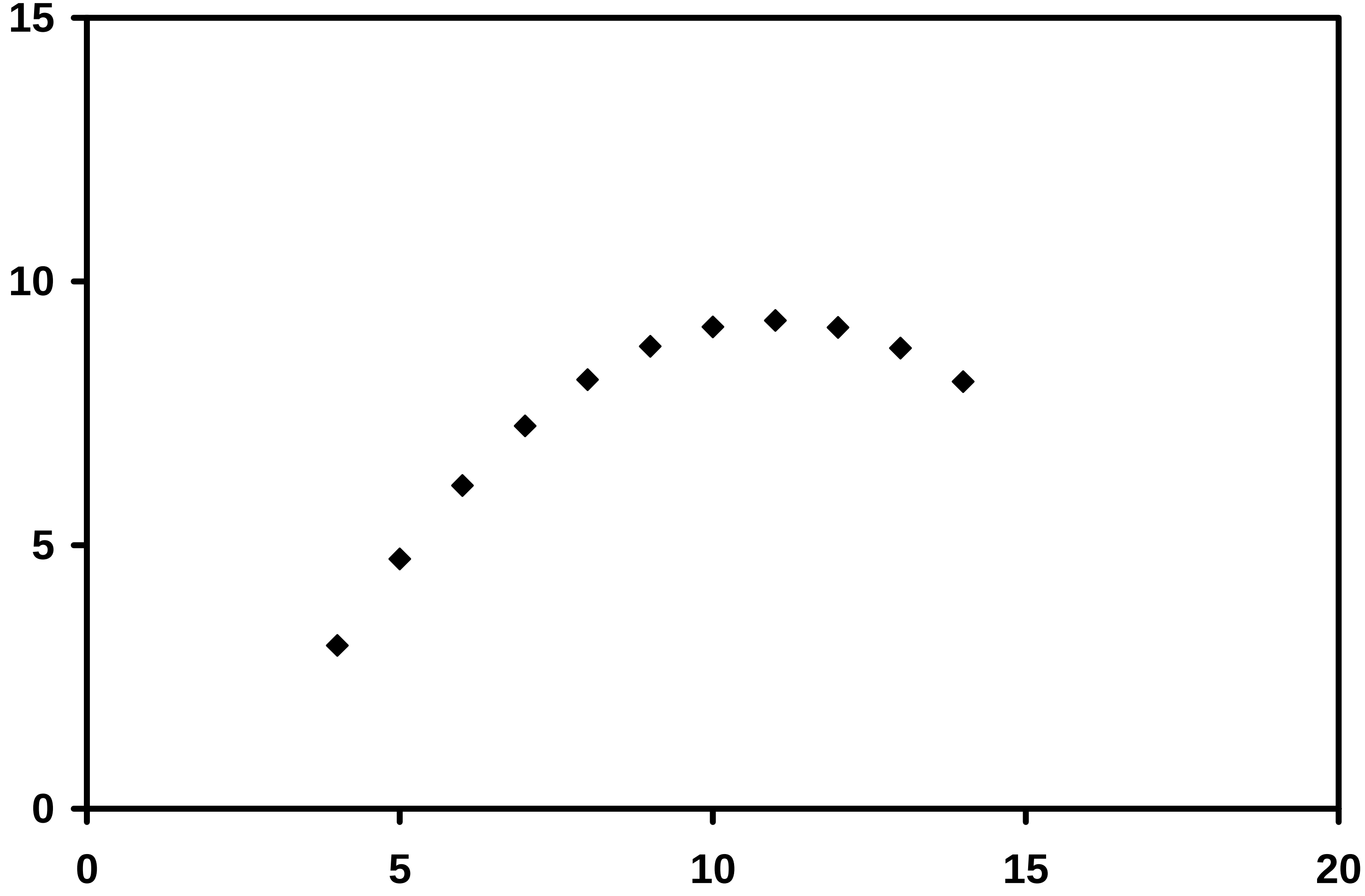
Data Set 1



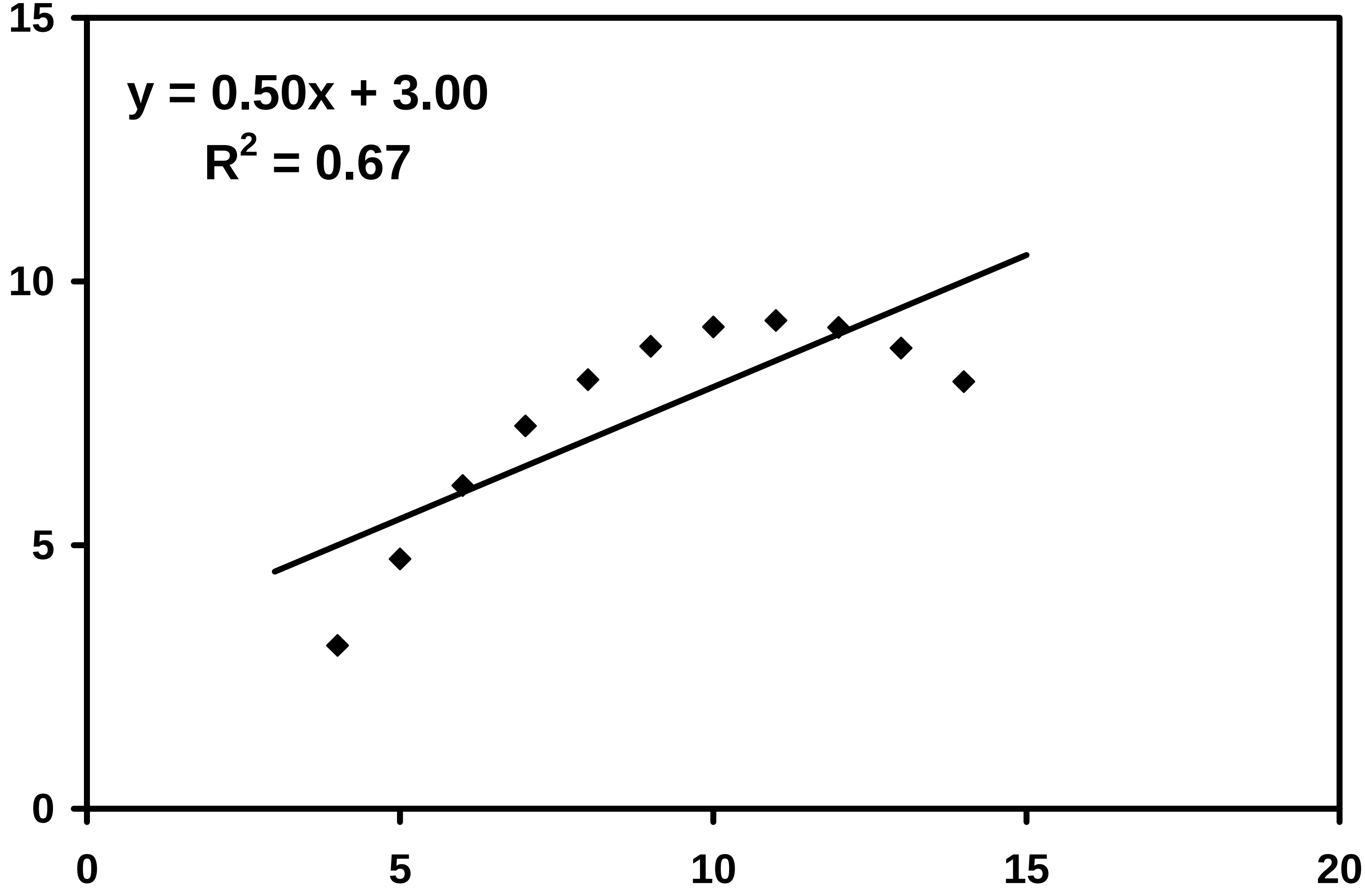
Data Set 1



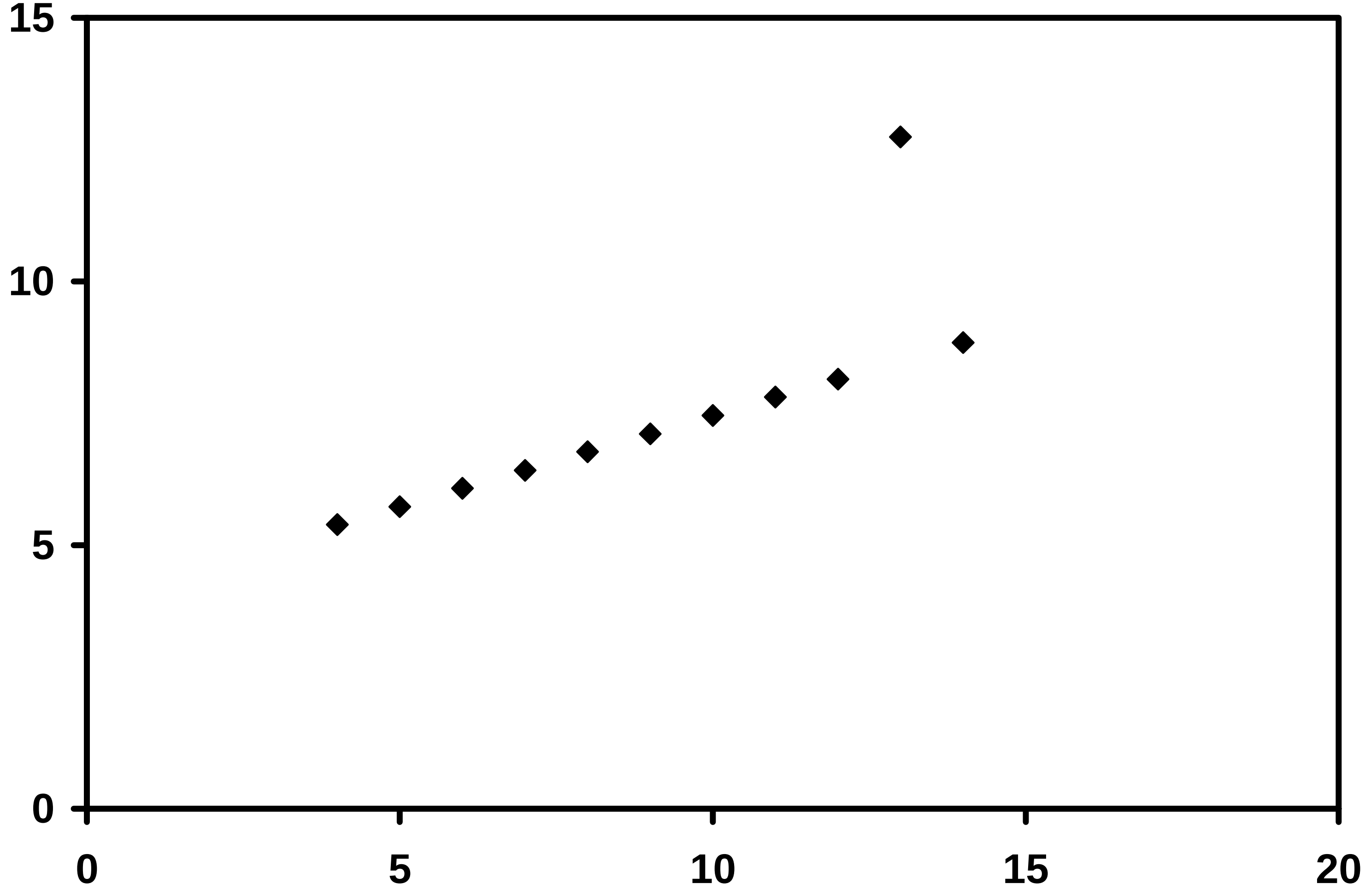
Data Set 2



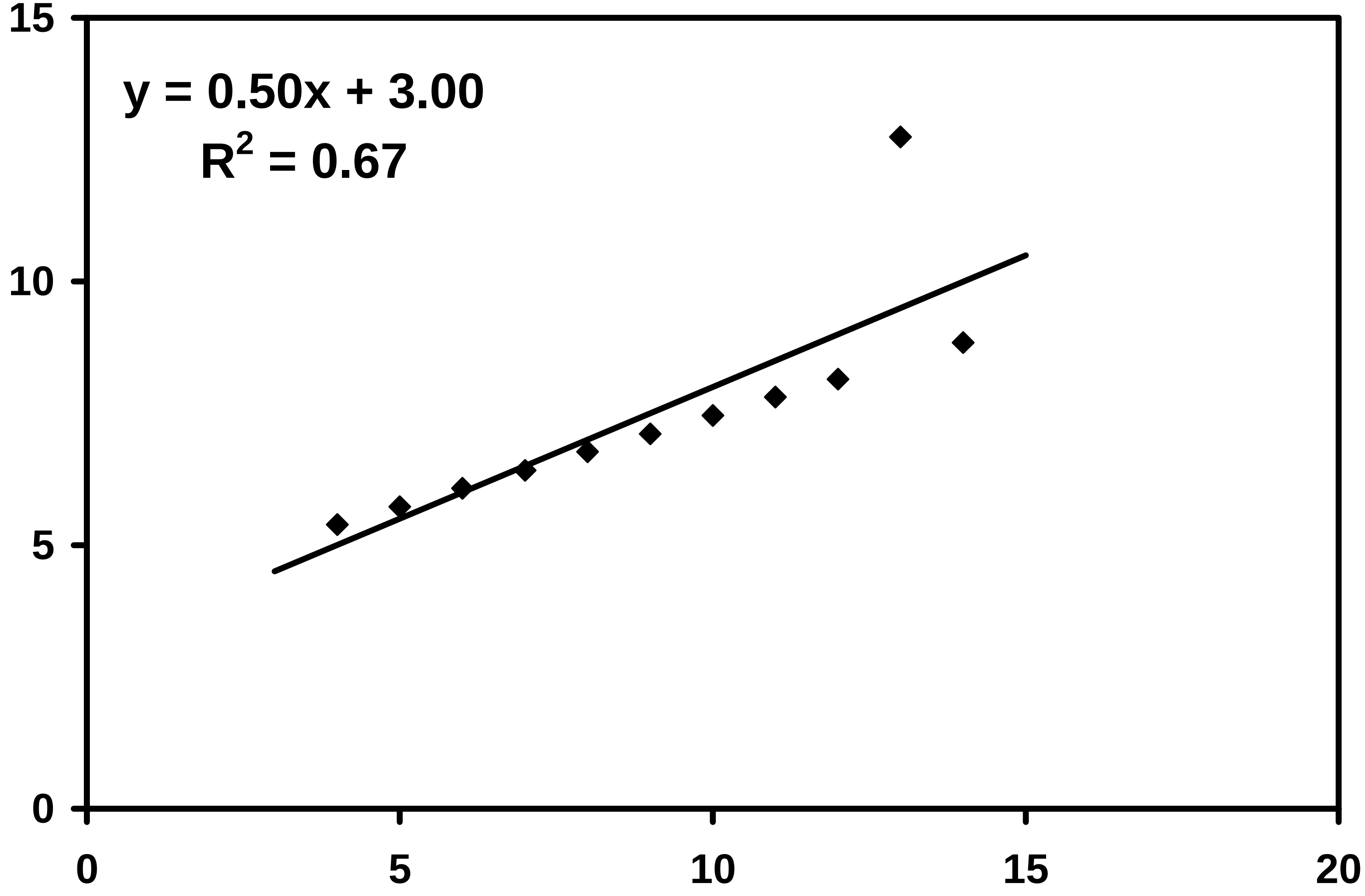
Data Set 2



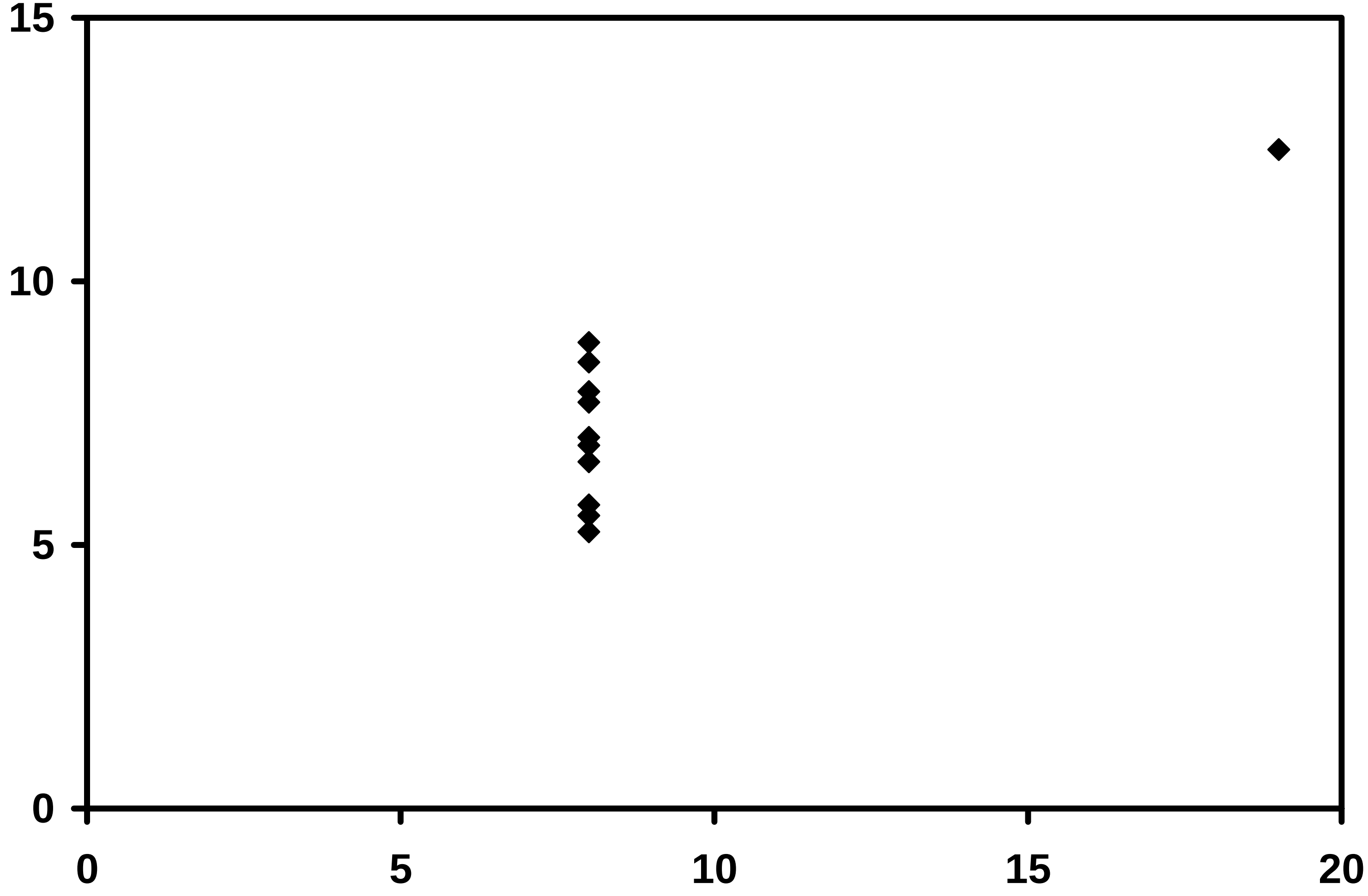
Data Set 3



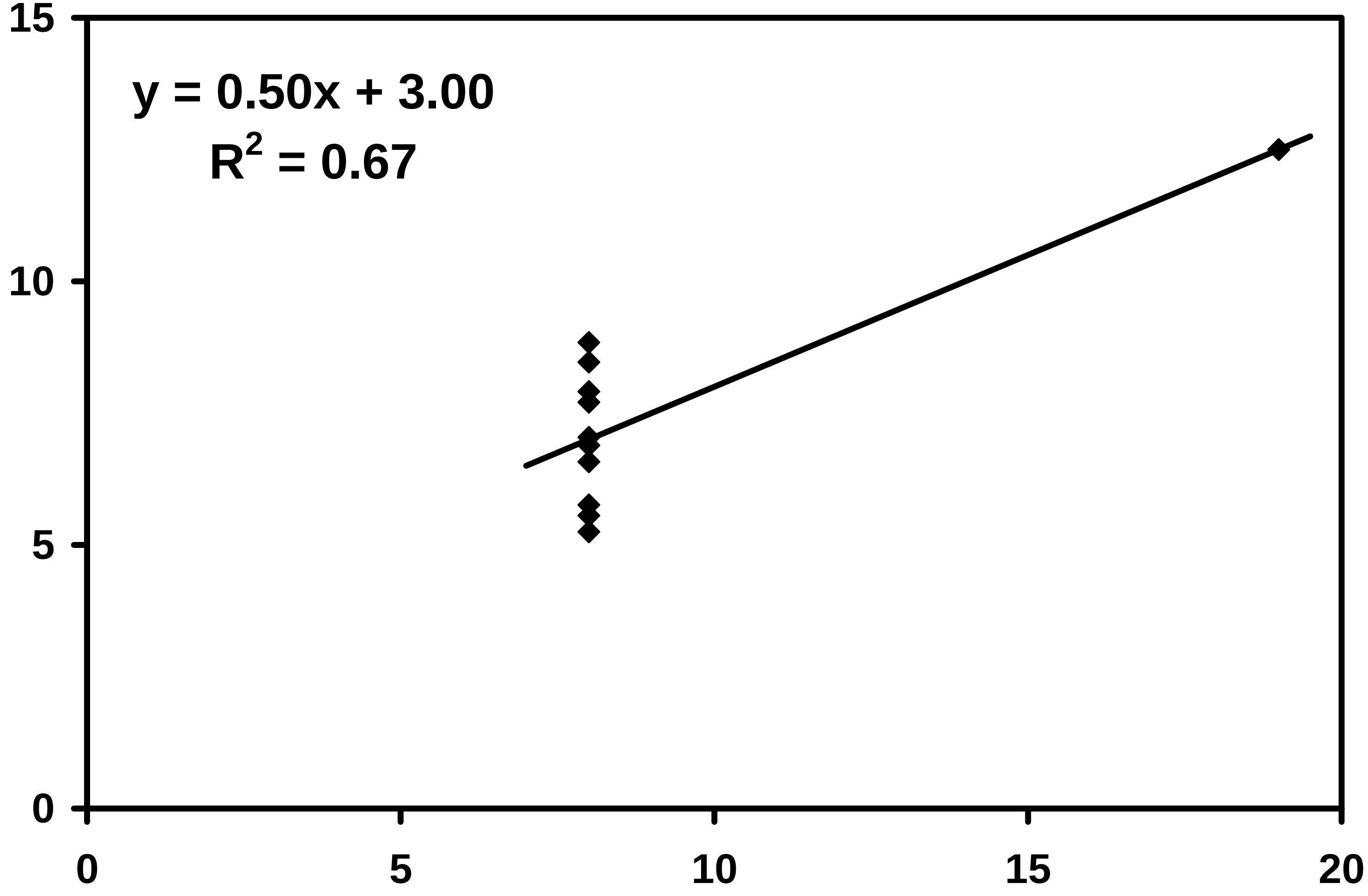
Data Set 3



Data Set 4



Data Set 4



Closing

- Getting data is easy.
- Getting valid data is not as easy.
- Analyses limited by quality of data.
- Implicit assumptions (e.g., normal).
- Data
  - Can produce meaningful decisions
  - Can be meaningless numbers
  - Can lead to erroneous conclusions.

Thanks for having me!

The End