

Basic Ideas in Probability and Statistics for Experimenters: Part II: Quantitative Work, Regression

*He uses statistics as a drunken man uses lamp-posts –
for support rather than for illumination ... A. Lang*

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
shivkuma@ecse.rpi.edu

<http://www.ecse.rpi.edu/Homepages/shivkuma>



- ❑ Quantitative examples of sample statistics, confidence interval for mean
- ❑ Introduction to Regression

- ❑ Also do the informal quiz handed out
- ❑ Reference: Chap 12, 13 (Jain), Chap 2-3 (Box,Hunter,Hunter),
- ❑ <http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html>
- ❑ Regression Applet:
<http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html>
- ❑ <http://www.statsoftinc.com/textbook/stmulreg.html>
- ❑ Tool: R-project (free statistical package)
 - ❑ <http://www.r-project.org/>

Independence

- $P(x \cap y) = 1/18(2x + y)$ for $x = 1,2$; and $y = 1,2$ and zero otherwise. Are the variables X and Y independent? Can you speculate why they are independent or dependent?

[Hint: $P(X)$ and $P(Y)$ can be formed by summing the above distribution (aka a joint distribution), since the combinations for specific values of x and y are mutually exclusive]

Independence

- $P(x \cap y) = 1/30(x^2 y)$ for $x = 1,2$; and $y = 1,2,3$ and zero otherwise. Are the variables X and Y independent? Can you speculate why they are independent or dependent?

Recall: Sampling Distribution

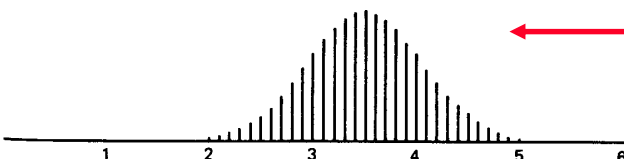
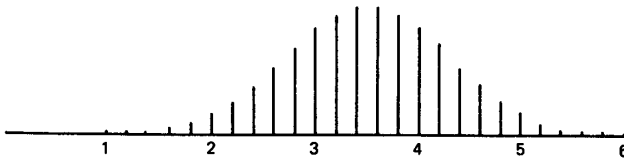
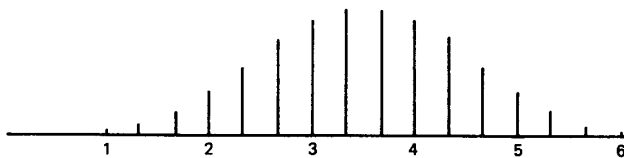
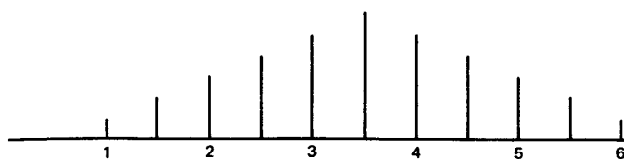
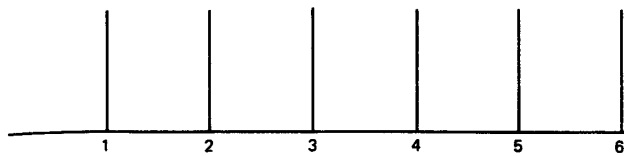


FIGURE 2.10. Distribution of average scores from throwing various numbers of dice.

Uniform distribution looks nothing like bell shaped (gaussian)! Large spread (σ)!

But the sampling distribution looks gaussian with smaller Standard deviation!

Sample mean $\sim N(\mu, \sigma / (n)^{0.5})$
i.e. the standard deviation of the sample mean (aka standard error) decreases with larger samples (n)

Shivkumar Kalyanaraman

Confidence Interval

- Sample mean: $\bar{x} \sim N(\mu, \sigma / (n)^{0.5})$
- The $100(1-\alpha)\%$ confidence interval is given by (if $n > 30$):

$$\{\bar{x} - z_{(1-\alpha/2)} s/(n)^{0.5}, \bar{x} + z_{(1-\alpha/2)} s/(n)^{0.5}\}$$
 - $z_{\beta} \sim N(0, 1)$; i.e. it is the unit normal distribution
 - $P\{(y - \mu)/\sigma \leq z_{\beta}\} = \beta$
- Eg 90% CI: $\{\bar{x} - z_{(0.95)} s/(n)^{0.5}, \bar{x} + z_{(0.95)} s/(n)^{0.5}\}$
- Refer to z-tables on pg 629, 630 (table A.2 or A.3)

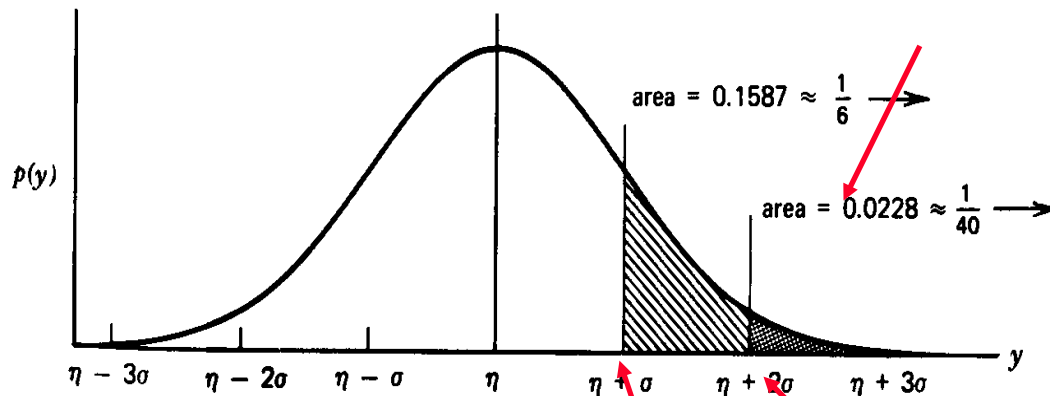


FIGURE 2.12. Tail areas of the normal distribution.

z=1

z=2

$$(\bar{x} - z_{1-\alpha/2} s / \sqrt{n}, \bar{x} + z_{1-\alpha/2} s / \sqrt{n})$$

$$z_{1-\alpha/2} = (1 - \alpha/2)\text{-quantile of } N(0,1)$$

Meaning of Confidence Interval

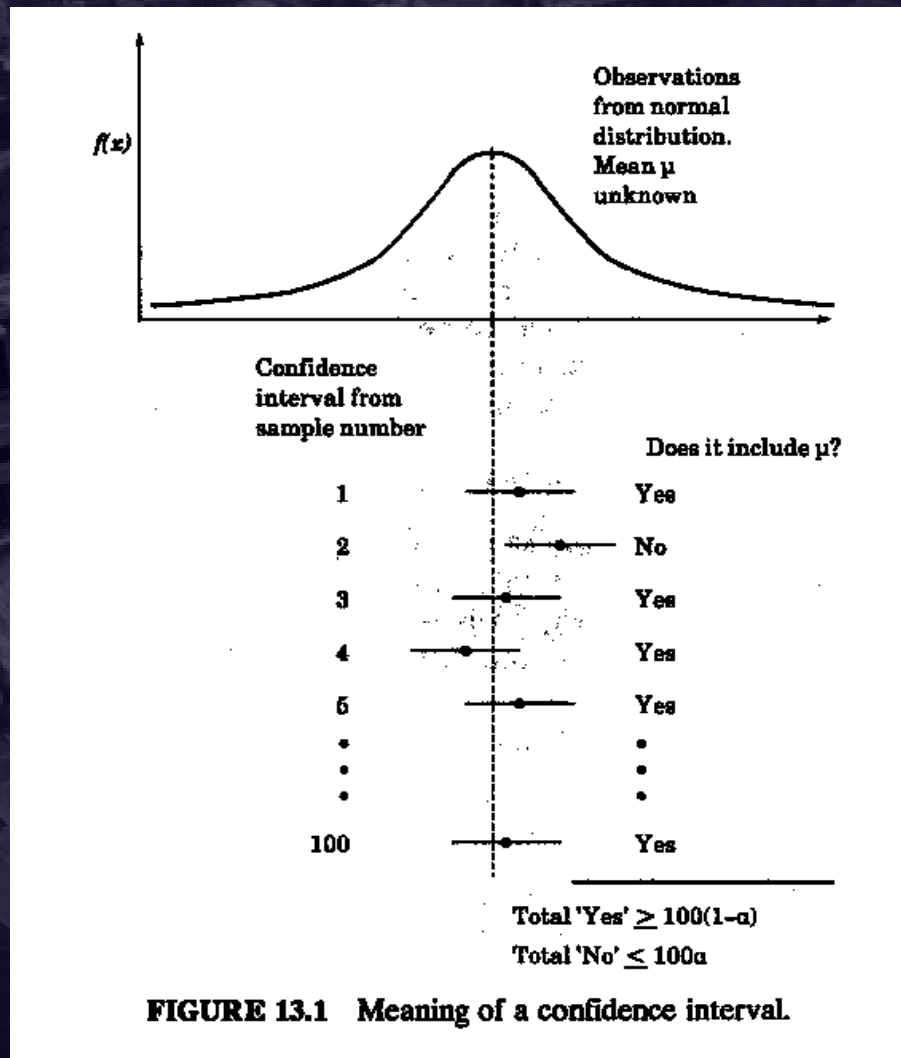


FIGURE 13.1 Meaning of a confidence interval.

http://www.math.csusb.edu/faculty/stanton/m262/confidence_means/confidence_means.html

(DEMO)

Ex: Sample Statistics, Confidence Interval

- Given: $n=32$ random RTT samples (in ms):
- {31, 42, 28, 51, 28, 44, 56, 39, 39, 27, 41, 36, 31, 45, 38, 29, 34, 33, 28, 45, 49, 53, 19, 37, 32, 41, 51, 32, 39, 48, 59, 42}
- 1. Find: sample mean (\bar{x}), median, mode, sample standard deviation (s), C.o.V., SIQR and 90% confidence interval (CI) & 95% CI for the *population* mean
- 2. Interpret your statistics *qualitatively*. I.e. what do they mean?
- Hint: Refer to the formulas in pg 197 and pg 219 of Jain's text (esp for s). Latter is reproduced in one of the slides

Box 13.1 Confidence Intervals

1. Given: A sample $\{x_1, x_2, \dots, x_n\}$ of n observations:

\bar{x} = sample mean; s = sample standard deviation

(a) Standard error of the sample mean: $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$

(b) $100(1 - \alpha)\%$ two-sided confidence interval for the mean:

$$\bar{x} \mp z_{1-\alpha/2} s / \sqrt{n}$$

$$\text{If } n \leq 30^\dagger: \bar{x} \mp t_{[1-\alpha/2; n-1]} s / \sqrt{n}$$

(c) $100(1 - \alpha)\%$ one-sided confidence interval for the mean:

$$(\bar{x}, \bar{x} + z_{1-\alpha} s / \sqrt{n}) \text{ or } (\bar{x} - z_{1-\alpha} s / \sqrt{n}, \bar{x})$$

$$\text{If } n \leq 30^\dagger: (\bar{x}, \bar{x} + t_{[1-\alpha; n-1]} s / \sqrt{n}) \text{ or } (\bar{x} - t_{[1-\alpha; n-1]} s / \sqrt{n}, \bar{x})$$

2. To compare two systems using unpaired observations:

(a) The standard error of the mean difference: $s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$

(b) The effective number of degrees of freedom:

$$\nu = \frac{(s_a^2/n_a + s_b^2/n_b)^2}{\frac{1}{n_a + 1} \left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b + 1} \left(\frac{s_b^2}{n_b}\right)^2} - 2$$

(c) The confidence interval for the mean difference:

$$(\bar{x}_a - \bar{x}_b) \mp t_{[1-\alpha/2; \nu]} s$$

3. If n_1 of the n observations belong to a certain class, the following statistics can be reported for the class:

(a) Proportion of the observations in the class: $p = \frac{n_1}{n}$

(b) $100(1 - \alpha)\%$ two-sided confidence interval for the proportion[‡]:

$$p \mp z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

(c) $100(1 - \alpha)\%$ one-sided confidence interval for the proportion[‡]:

$$\left(p, p + z_{1-\alpha} \sqrt{\frac{p(1-p)}{n}}\right) \quad \text{or} \quad \left(p - z_{1-\alpha} \sqrt{\frac{p(1-p)}{n}}, p\right)$$

[†] Only for samples from normal populations.

[‡] Provided $np \geq 10$.

t-distribution: for confidence intervals given few samples ($6 \leq n < 30$)

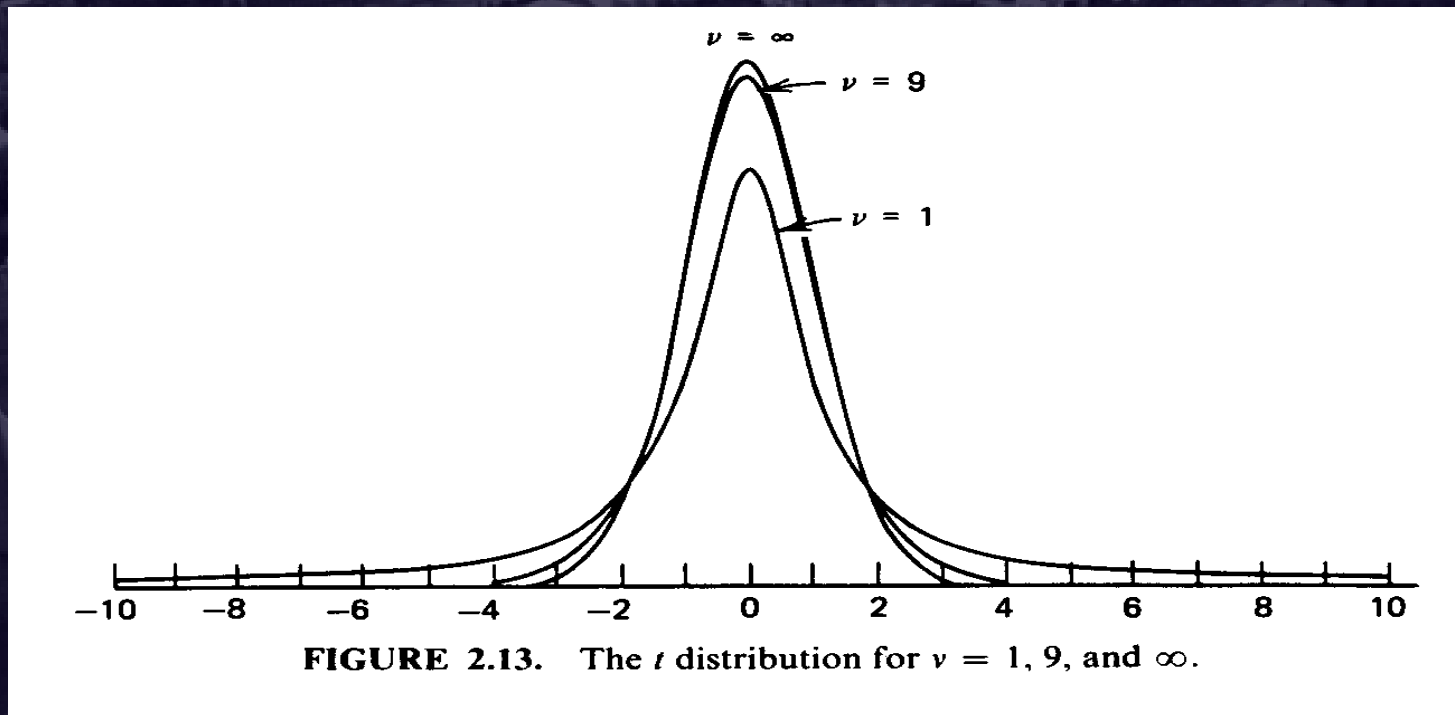


FIGURE 2.13. The t distribution for $\nu = 1, 9$, and ∞ .

- ❑ Idea: t-distribution with $n-1$ degrees of freedom approximates normal distribution for larger n ($n \geq 6$).
- ❑ t-distribution is a poor approximation for lower degrees of freedom (I.e. smaller number of samples than 6!)

t-distribution: confidence intervals

- The $100(1-\alpha)\%$ confidence interval is given by (if $n \leq 30$):

$$\{\bar{x} - t_{\{1-\alpha/2, n-1\}} s/(n)^{0.5}, \bar{x} + t_{\{1-\alpha/2, n-1\}} s/(n)^{0.5}\}$$

- Use t-distribution tables in pg 631, table A.4
- Eg: for $n = 7$, 90% CI:

$$\{\bar{x} - t_{\{0.95, 6\}} s/(n)^{0.5}, \bar{x} + t_{\{0.95, 6\}} s/(n)^{0.5}\}$$

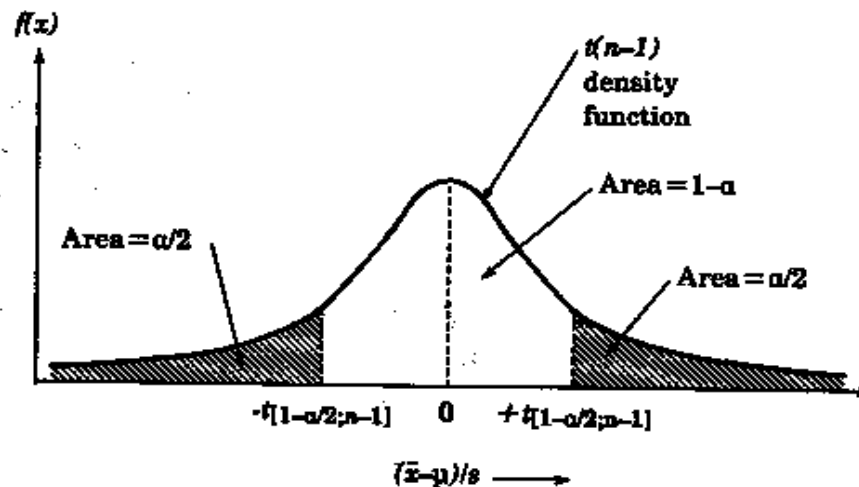
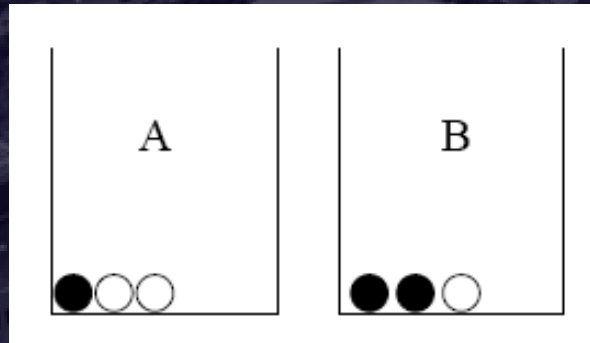


FIGURE 13.2 The ratio $(\bar{x} - \mu) / (s / \sqrt{n})$ for samples from normal populations follows a $t(n - 1)$ distribution.

Confidence Interval with few samples

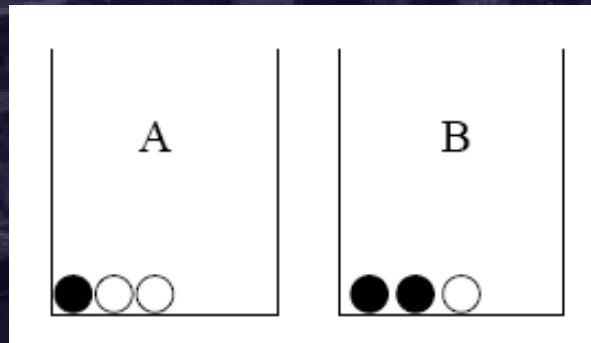
- Given: $n=10$ random RTT samples (in ms):
{31, 42, 28, 51, 28, 44, 56, 39, 39, 27}
- 1. Find: sample mean (\bar{x}), sample standard deviation (s) and 90% confidence interval (CI) & 95% CI for the *population* mean
- 2. Interpret this result relative to the earlier result using the normal distribution.
- Hint: Refer to the formulas in pg 197 and pg 219 of Jain's text (esp for s).

Likelihood Principle



- ❑ Experiment:
 - ❑ Pick Urn A or Urn B at random
 - ❑ Select a ball from that Urn.
- ❑ The ball is black.
- ❑ What is the probability that the selected Urn is A?

Likelihood Principle (Contd)

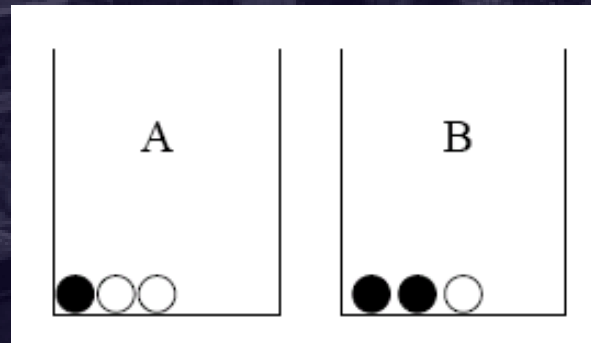


- Write out what you know!
- $P(\text{Black} \mid \text{UrnA}) = 1/3$
- $P(\text{Black} \mid \text{UrnB}) = 2/3$
- $P(\text{Urn A}) = P(\text{Urn B}) = 1/2$
- We want $P(\text{Urn A} \mid \text{Black})$.
- Gut feeling: Urn B is more likely than Urn A (given that the ball is black). But by how much?
- This is an inverse probability problem.
 - Make sure you understand the inverse nature of the conditional probabilities!
- Solution technique: Use Bayes Theorem.

Likelihood Principle (Contd)

- Bayes manipulations:
- $P(\text{Urn A} \mid \text{Black}) =$
 - $P(\text{Urn A and Black}) / P(\text{Black})$
- Decompose the numerator and denominator in terms of the probabilities we know.
- $P(\text{Urn A and Black}) = P(\text{Black} \mid \text{Urn A}) * P(\text{Urn A})$
- $P(\text{Black}) = P(\text{Black} \mid \text{Urn A}) * P(\text{Urn A}) + P(\text{Black} \mid \text{Urn B}) * P(\text{Urn B})$
- We know all these values (see prev page)! Plug in and crank.
- $P(\text{Urn A and Black}) = 1/3 * 1/2$
- $P(\text{Black}) = 1/3 * 1/2 + 2/3 * 1/2 = 1/2$
- $P(\text{Urn A and Black}) / P(\text{Black}) = 1/3 = 0.333$
- Notice that it matches our gut feeling that Urn A is less likely, once we have seen black.
- The information that the ball is black has CHANGED !
 - From $P(\text{Urn A}) = 0.5$ to $P(\text{Urn A} \mid \text{Black}) = 0.333$

Likelihood Principle



- ❑ Way of thinking...
- ❑ Hypotheses: Urn A or Urn B ?
- ❑ Observation: “Black”
- ❑ Prior probabilities: $P(\text{Urn A})$ and $P(\text{Urn B})$
- ❑ Likelihood of Black given choice of Urn: {aka *forward probability*}

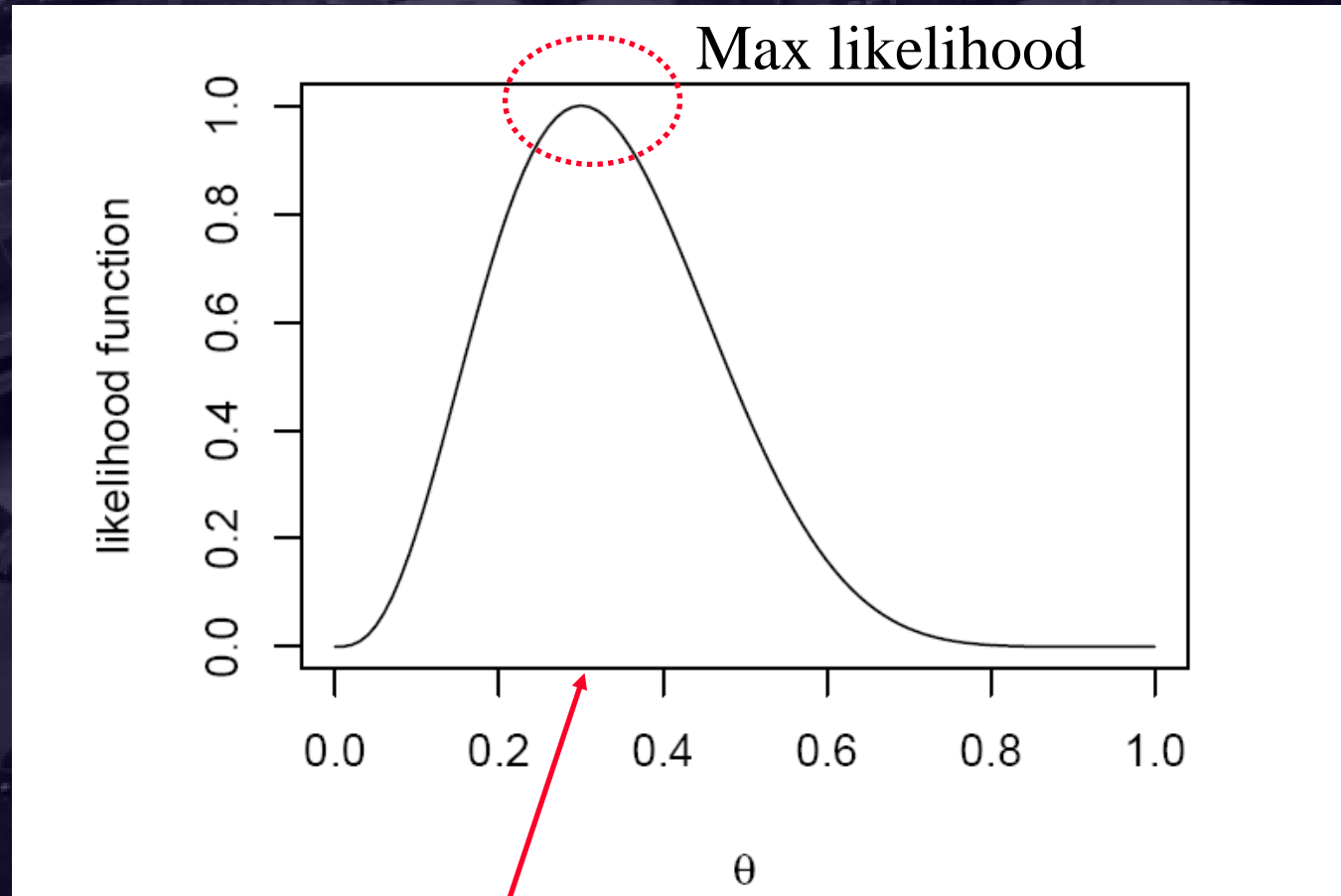
 - ❑ $P(\text{Black} | \text{Urn A})$ and $P(\text{Black} | \text{Urn B})$

- ❑ Posterior Probability: of each hypothesis given evidence
 - ❑ $P(\text{Urn A} | \text{Black})$ {aka *inverse probability*}
- ❑ Likelihood Principle (informal): All inferences depend ONLY on
 - ❑ The likelihoods $P(\text{Black} | \text{Urn A})$ and $P(\text{Black} | \text{Urn B})$, and
 - ❑ The priors $P(\text{Urn A})$ and $P(\text{Urn B})$
- ❑ Result is a probability (or distribution) model over the space of possible hypotheses.

Maximum Likelihood (intuition)

- Recall:
- $P(\text{Urn A} \mid \text{Black}) = P(\text{Urn A and Black}) / P(\text{Black}) = P(\text{Black} \mid \text{Urn A}) * P(\text{Urn A}) / P(\text{Black})$
- $P(\text{Urn?} \mid \text{Black})$ is maximized when $P(\text{Black} \mid \text{Urn?})$ is maximized.
 - Maximization over the hypotheses space (Urn A or Urn B)
- $P(\text{Black} \mid \text{Urn?}) = \text{“likelihood”}$
- \Rightarrow “Maximum Likelihood” approach to maximizing posterior probability

Maximum Likelihood: intuition



This hypothesis has the highest (maximum) likelihood of explaining the data observed

Maximum Likelihood (ML): mechanics

- **Independent Observations** (like Black): X_1, \dots, X_n
- **Hypothesis** θ
- **Likelihood Function**: $L(\theta) = P(X_1, \dots, X_n | \theta) = \prod_i P(X_i | \theta)$
 - {Independence => multiply individual likelihoods}
- **Log Likelihood** $LL(\theta) = \sum_i \log P(X_i | \theta)$
- **Maximum likelihood**: by taking derivative and setting to zero and solving for θ

$$\hat{\theta}_{ML}(x) = \arg \max_{\theta} P(x|\theta)$$
- **Maximum A Posteriori (MAP)**: if non-uniform prior probabilities/distributions
 - Optimization function

Back to Urn example

- In our urn example, we are asking:
 - Given the observed data “ball is black”...
 - ...which hypothesis (Urn A or Urn B) has the highest likelihood of explaining this observed data?
 - Ans from above analysis: Urn B
- Note: this does not give the posterior probability $P(\text{Urn A} \mid \text{Black})$, but quickly helps us choose the best hypothesis (Urn B) that would explain the data...

More examples: (biased coin etc)

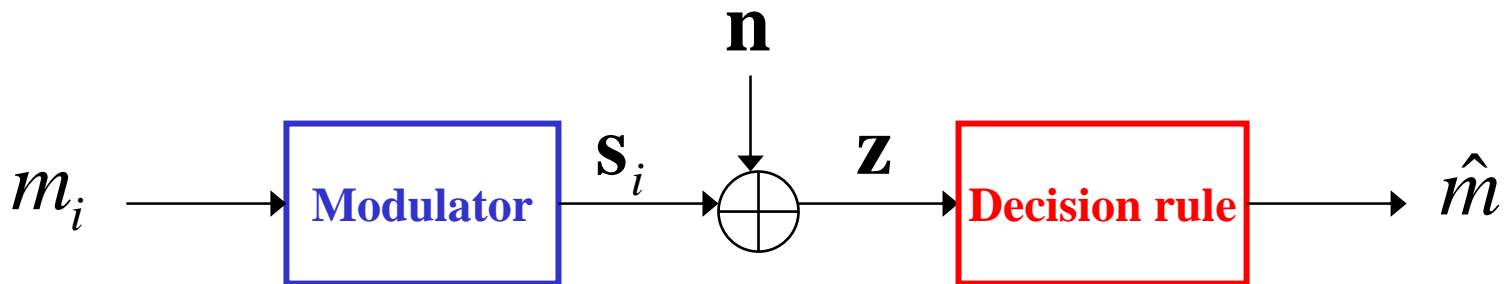
http://en.wikipedia.org/wiki/Maximum_likelihood

<http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html>

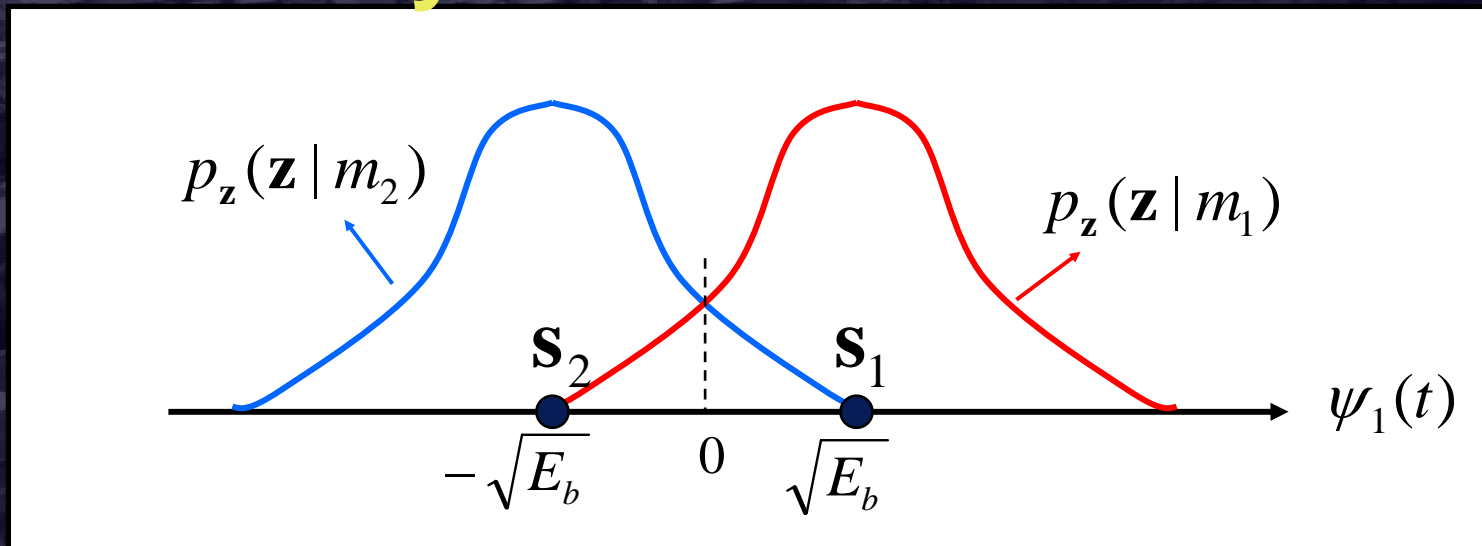
(chap 3)

Not Just Urns and Balls: Detection of signal in AWGN

- Detection problem:
 - Given the observation vector \mathbf{z} , perform a mapping from \mathbf{z} to an estimate \hat{m} of the transmitted symbol, m_i , such that the average probability of error in the decision is minimized.



Binary PAM + AWGN Noise



Signal s_1 or s_2 is sent. \mathbf{z} is received

Additive white gaussian noise (AWGN) \Rightarrow the likelihoods are

$p_z(\mathbf{z} | m_1)$ $p_z(\mathbf{z} | m_2)$ bell-shaped pdfs around s_1 and s_2

MLE \Rightarrow at any point on the x-axis, see which curve (blue or red) has a higher (maximum) value and select the corresponding signal (s_1 or s_2)

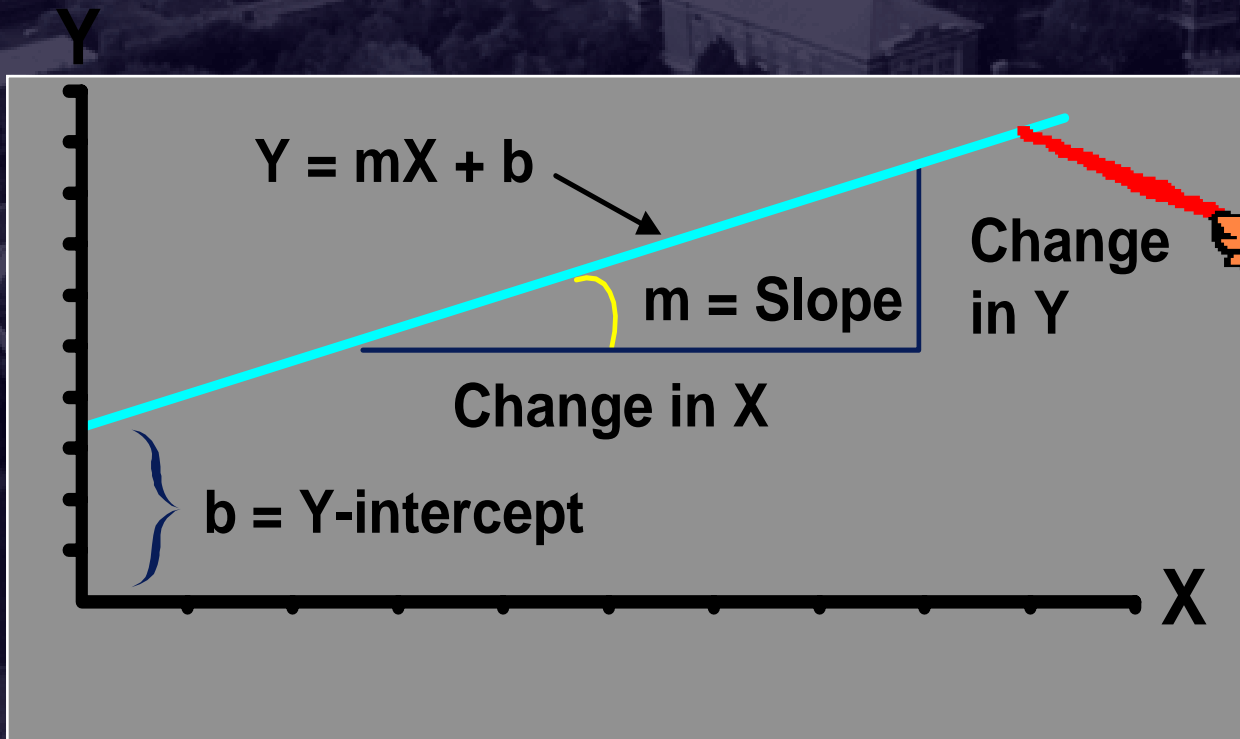
Unbiased vs Maximum Likelihood Estimation: final thoughts

- ❑ Max-likelihood picks the hypothesis most likely given the data or sample (and underlying prior assumptions).
- ❑ Unbiased estimator could have a high variance.
 - ❑ May not be able to find a consistent unbiased estimator each time
- ❑ Maximum likelihood provides a consistent approach to parameter estimation problems: comes from a standard optimization setup
 - ❑ This means that maximum likelihood estimates can be developed for a large variety of estimation situations.
 - ❑ MLEs become unbiased minimum variance estimators as the sample size increases
 - ❑ May be biased for small samples.

Linear Regression

- Goal: determine the relationship between two random variables X and Y .
- Example: X = height and Y =weight of a sample of adults.
- Linear regression attempts to explain this relationship with a straight line fit to the data, i.e. a linear model.
- The linear regression model *postulates* that
$$Y = a + bX + e$$
- Where the "*residual*" or "*error*" e is a random variable with mean = zero.
- The coefficients a and b are determined by the condition that the *sum of the square residuals* (i.e. the "energy" of residuals) is as small as possible (i.e. *minimized*).

Linear Equations



High School Teacher

Shivkumar Kalyanaraman ©1984-1994 T/Maker Co.

Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Population Y-Intercept (points to β_0)

Population Slope (points to β_1)

Random Error (points to ε_i)

Dependent (Response) Variable (points to Y_i)

Independent (Explanatory) Variable (points to X_i)

Population & Sample Regression Models

Population

Unknown Relationship ☺ \$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

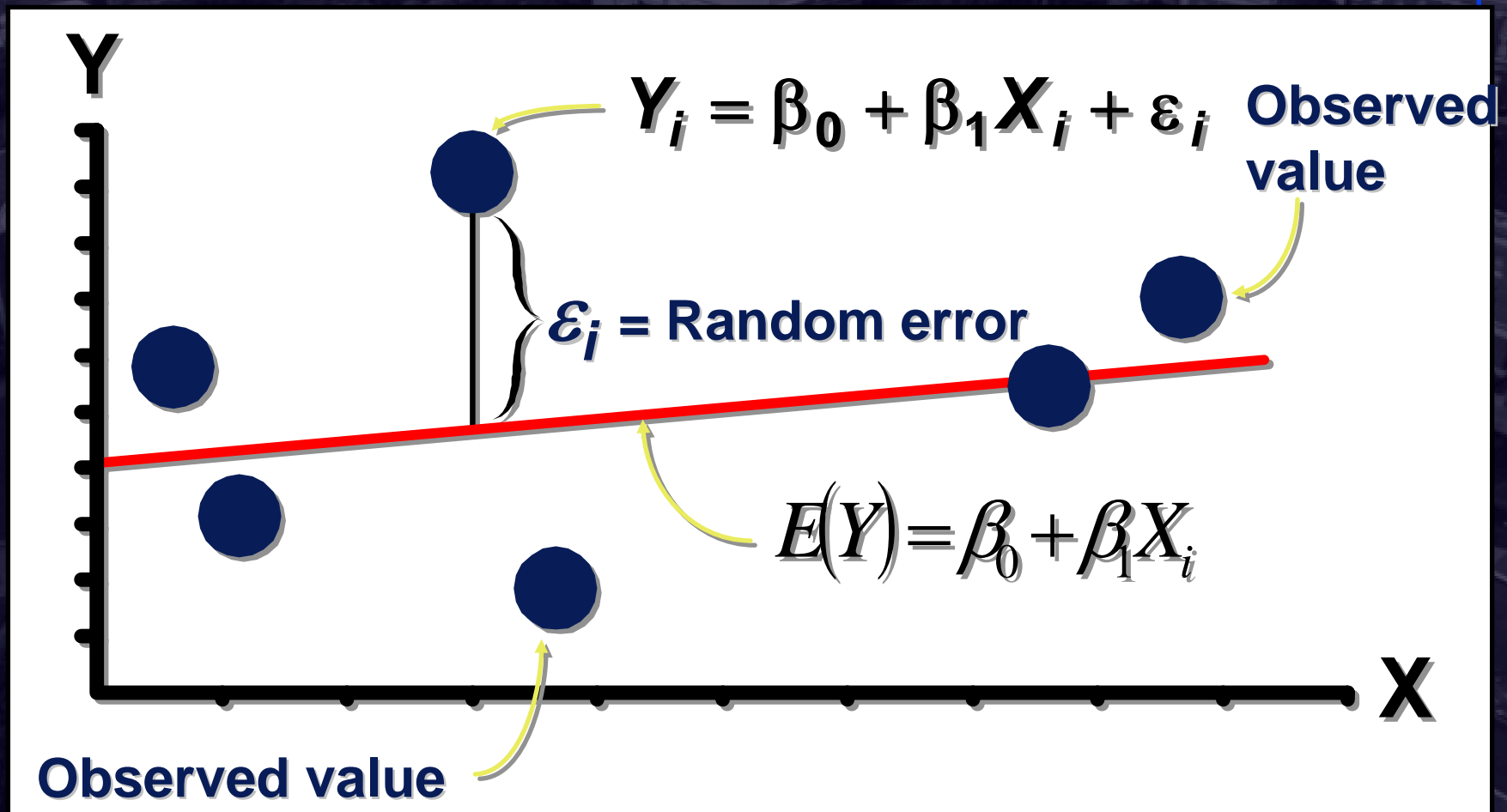


Random Sample

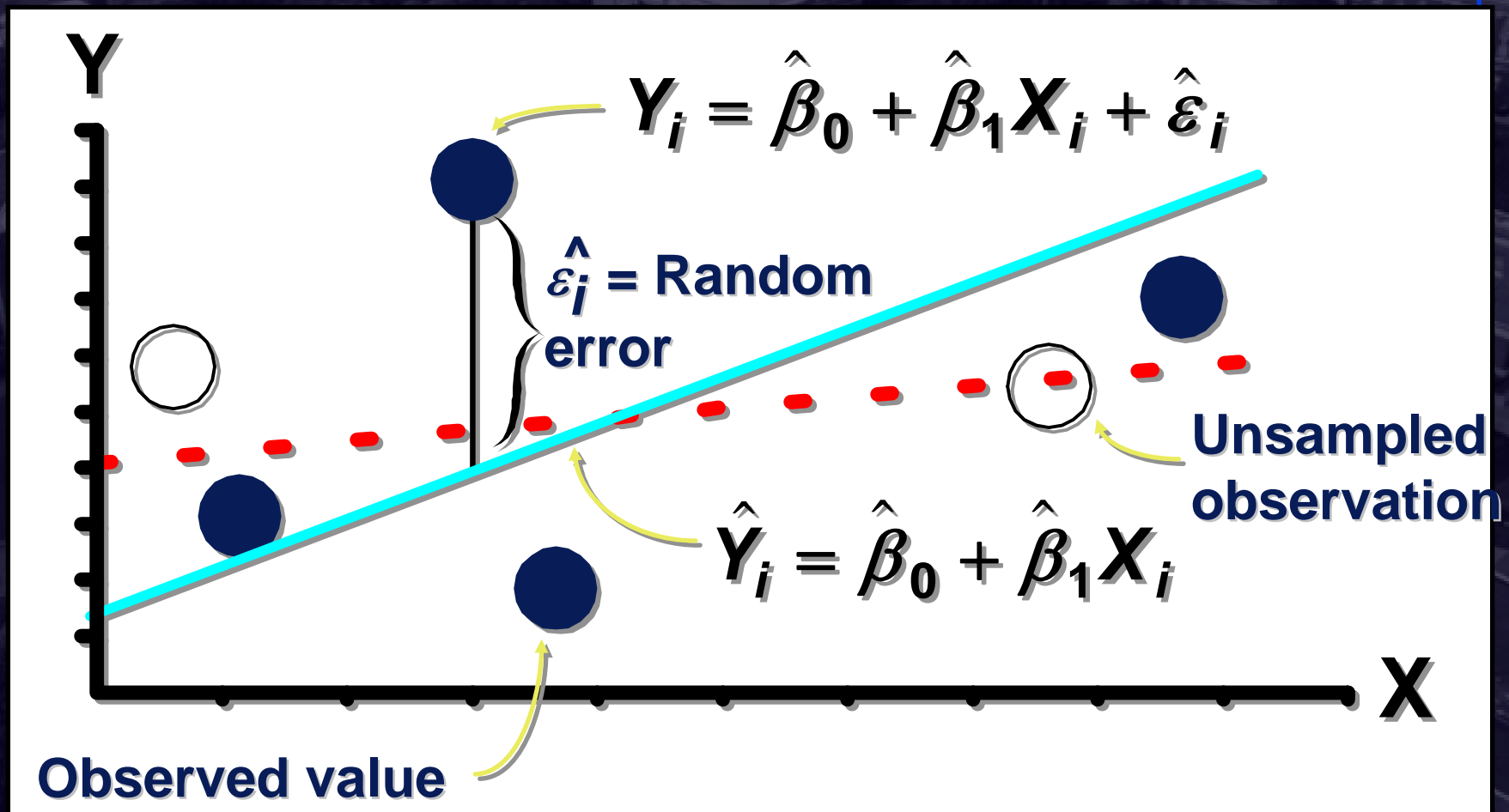
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$



Population Linear Regression Model



Sample Linear Regression Model



Demo: Regression Applet

- <http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html>

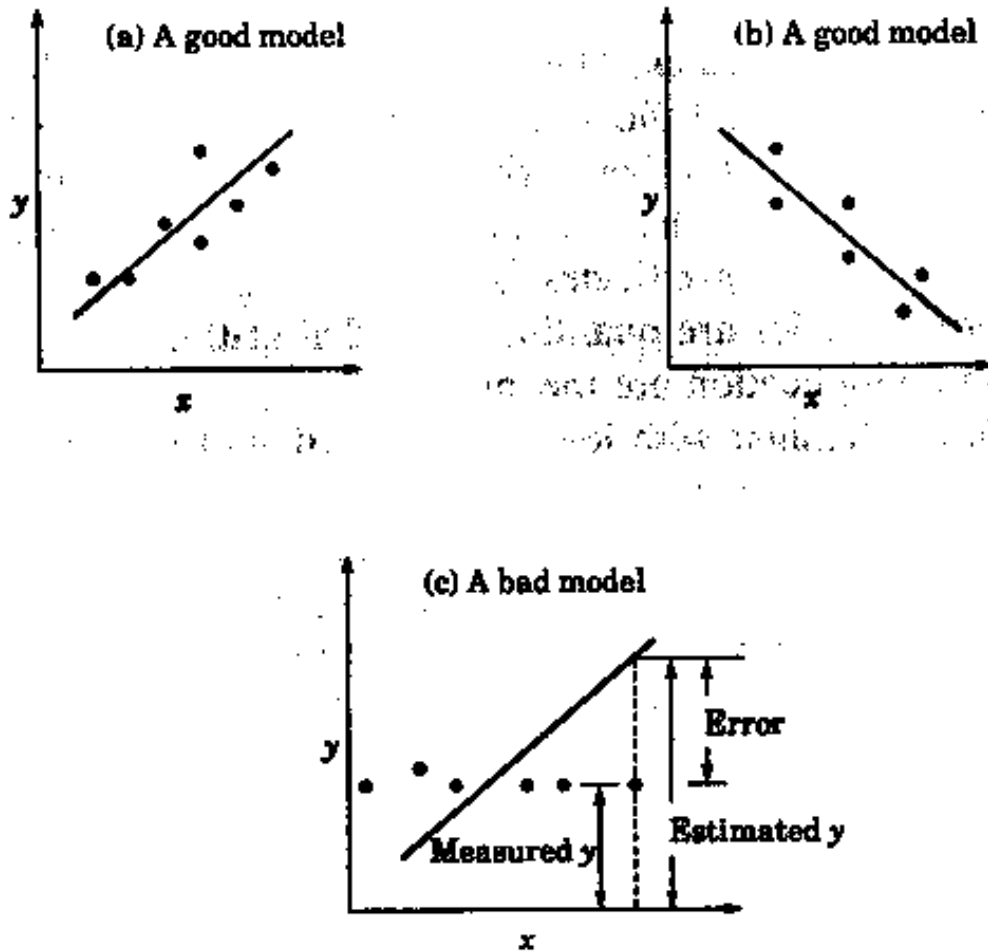


FIGURE 14.1 Good and bad regression models.

Regression Theory

- Model: $y_{\text{est}} = b_0 + b_1 x_i$
- Error: $e_i = y_i - y_{\text{est}}$
- Sum of Squared Errors (SSE): $\sum e_i^2 = \sum (y_i - b_0 + b_1 x_i)^2$
- Mean error (e_{avg}): $\sum e_i = \sum (y_i - b_0 + b_1 x_i)$

- Linear Regression problem:

Minimize: SSE

Subject to the constraint: $e_{\text{avg}} = 0$

- Solution: (I.e. regression coefficients)

- $b_1 = s_{xy}^2 / s_x^2 = \{ \sum xy - x_{\text{avg}} y_{\text{avg}} \} / \{ \sum x^2 - n(x_{\text{avg}})^2 \}$

- $b_0 = y_{\text{avg}} - b_1 x_{\text{avg}}$

Least Squares

- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum
 - *But* Positive Differences Off-Set Negative

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^n \hat{\epsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (SSE)

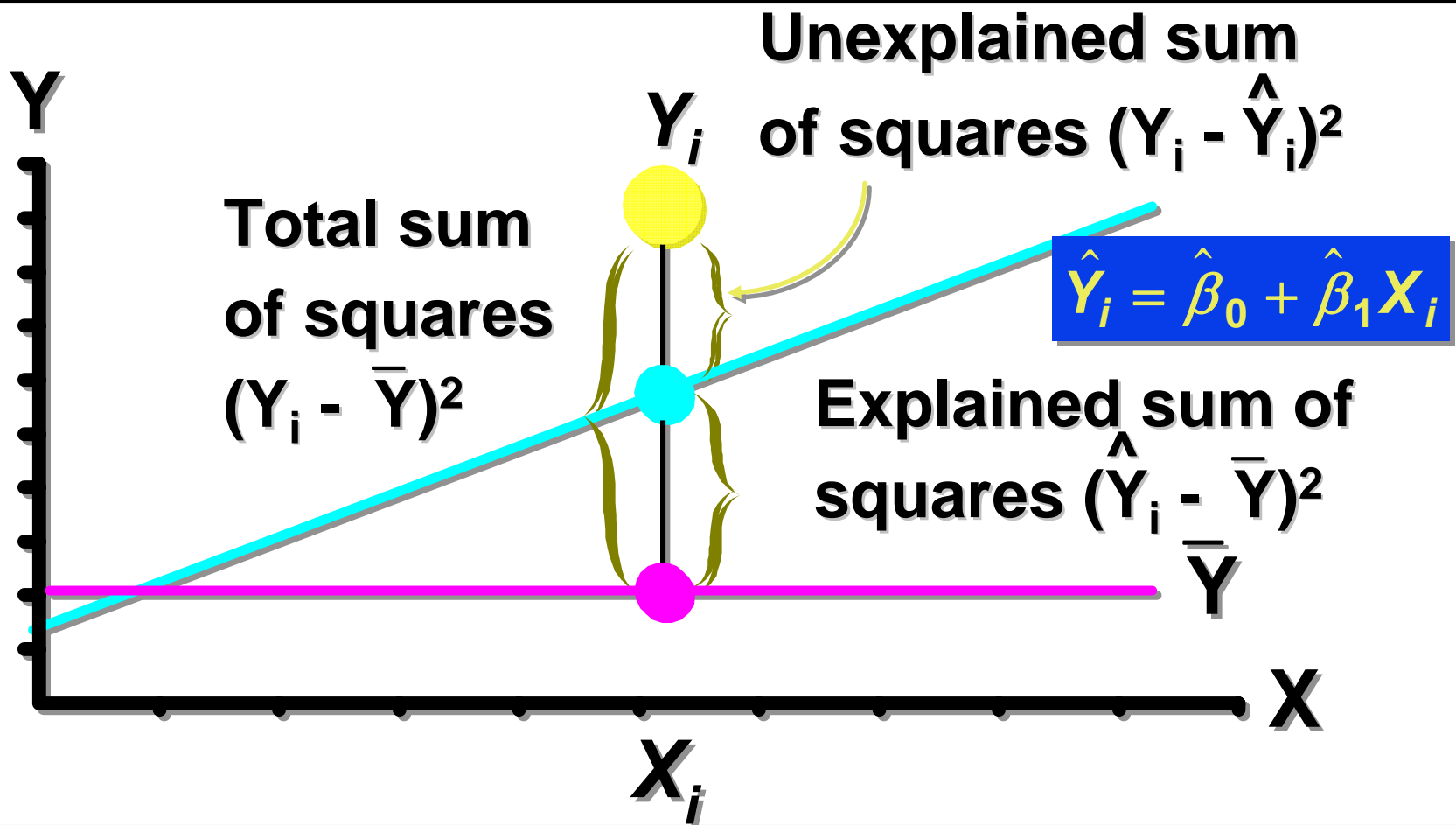
Random Error Variation

- ❑ 1. Variation of Actual Y from Predicted Y
- ❑ 2. Measured by Standard Error of Regression Model
 - ❑ Sample Standard Deviation of ε , \hat{s}
- ❑ 3. Affects Several Factors
 - ❑ Parameter Significance
 - ❑ Prediction Accuracy

Measures of Variation in Regression

- ❑ 1. Total Sum of Squares (SS_{yy})
 - ❑ Measures Variation of Observed Y_i Around the Mean \bar{Y}
- ❑ 2. Explained Variation (SSR)
 - ❑ Variation Due to Relationship Between X & Y
- ❑ 3. Unexplained Variation (SSE)
 - ❑ Variation Due to Other Factors

Variation Measures



Coefficient of Determination (R-squared)

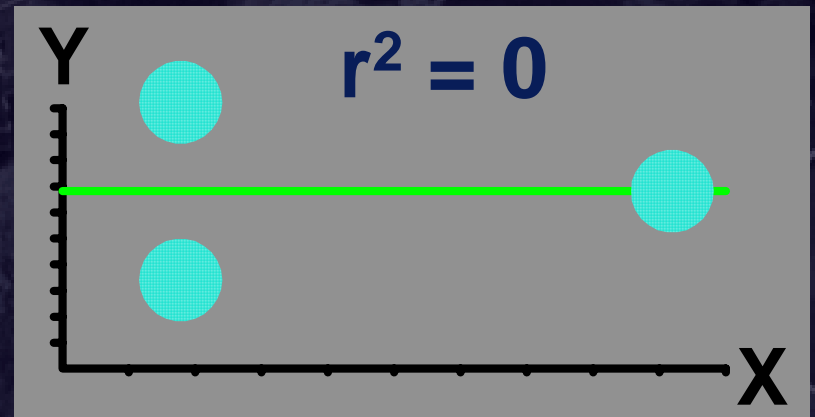
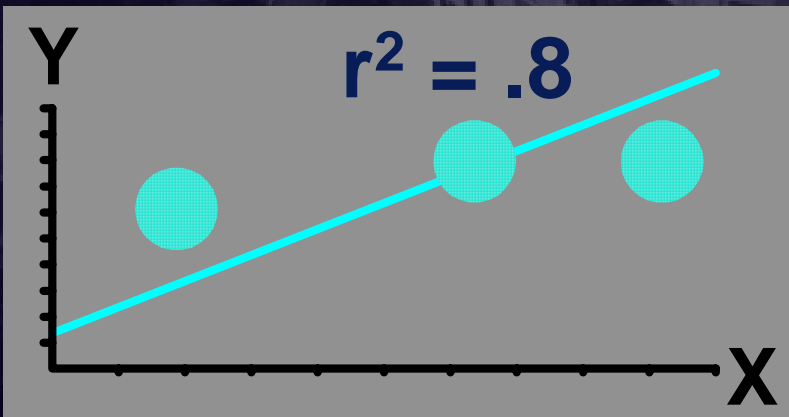
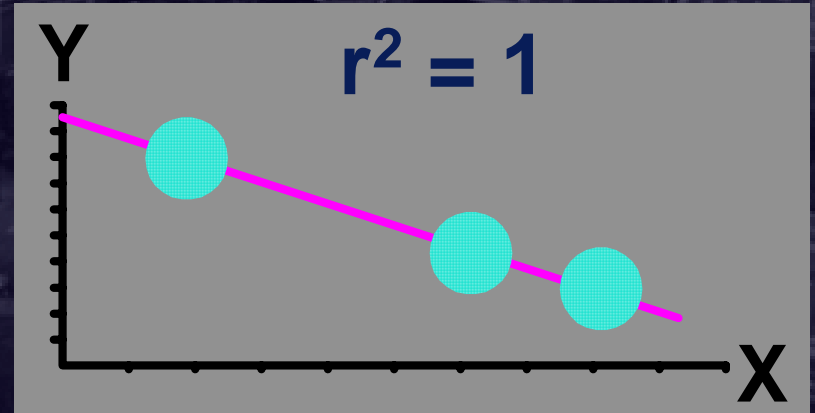
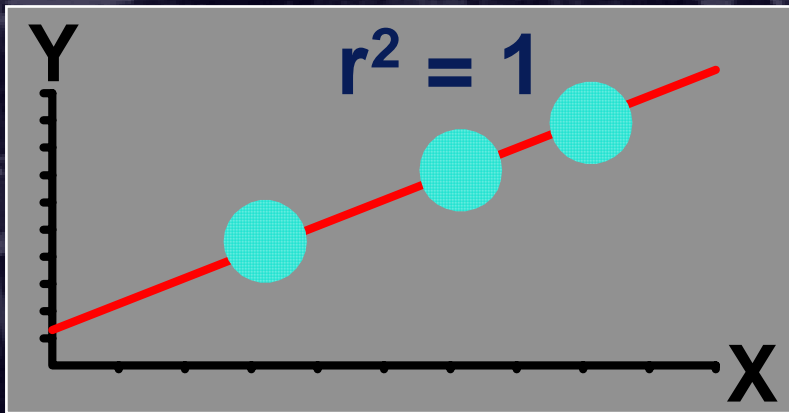
- 1. **Proportion** of Variation 'Explained' by Relationship Between X & Y

$$0 \leq r^2 \leq 1$$

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$
$$= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$



Coefficient of Determination Examples



Linear Regression Assumptions

- ❑ 1. Mean of Probability Distribution of Error Is 0
- ❑ 2. Probability Distribution of Error Has Constant Variance
- ❑ 3. Probability Distribution of Error is Normal
- ❑ 4. Errors Are Independent

Practical issues: Check linearity hypothesis with scatter diagram!

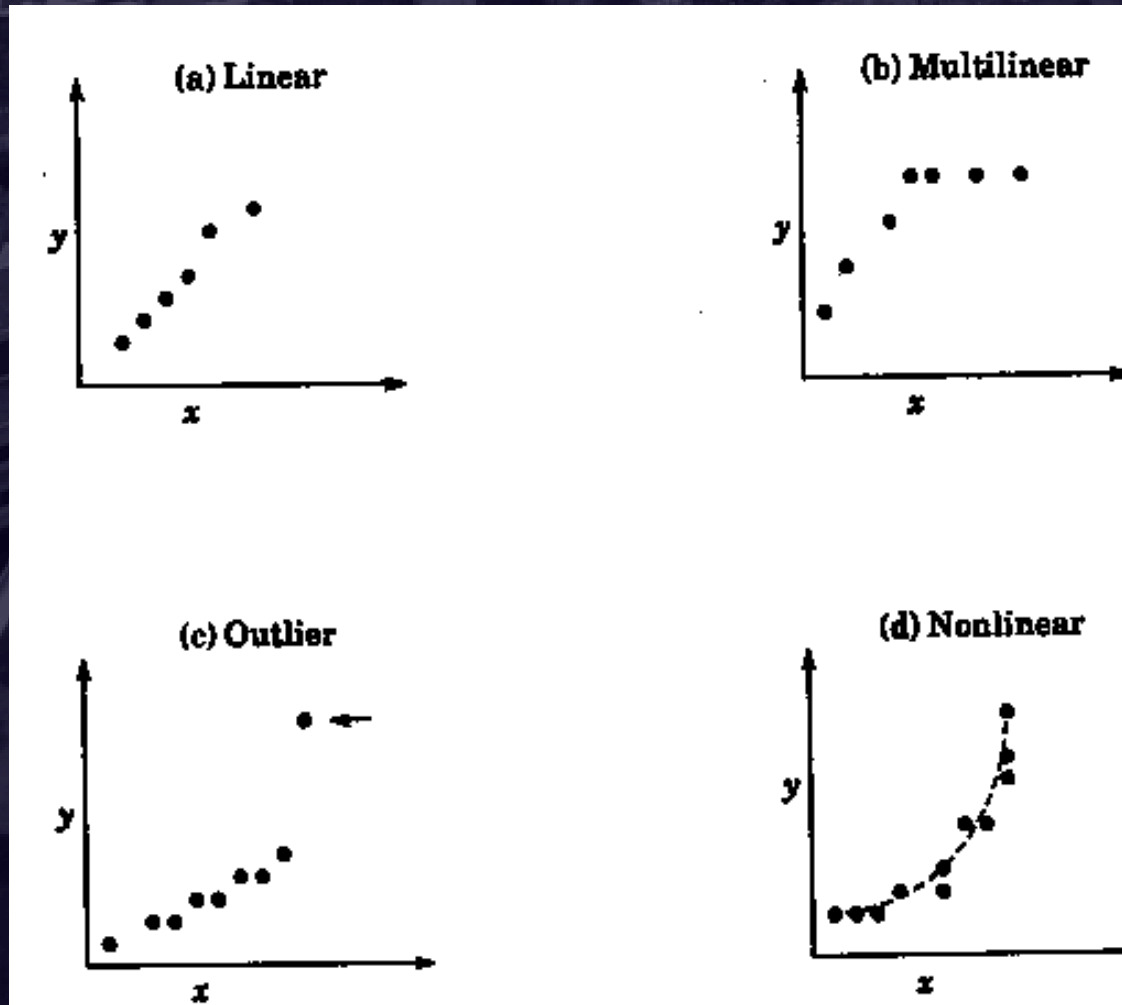


FIGURE 14.6 Possible patterns of scatter diagrams.

Practical issues: Check randomness & zero mean hypothesis for residuals!

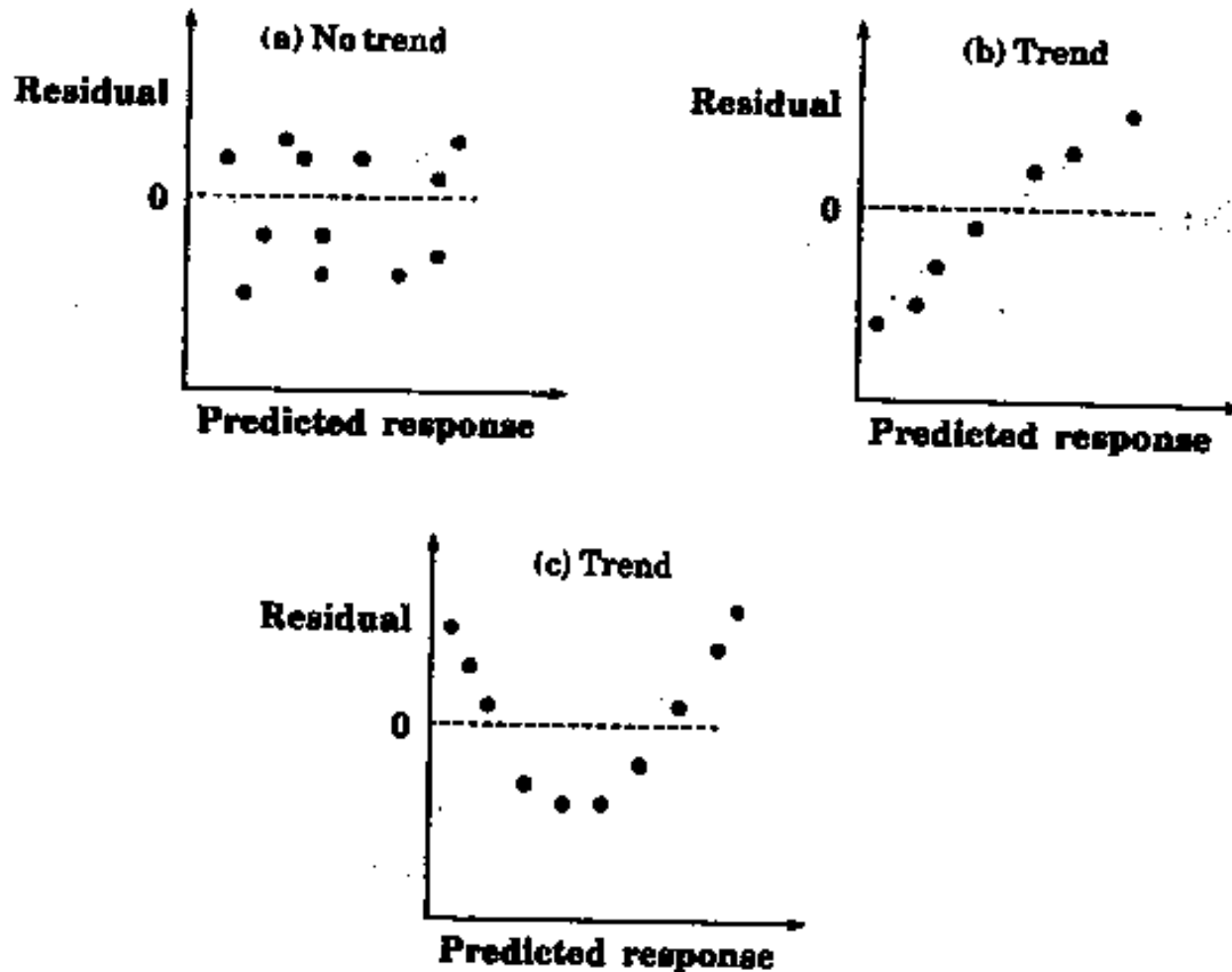


FIGURE 14.7 Possible patterns of residual versus predicted response graphs.

Practical issues: Does the regression indeed explain the variation?

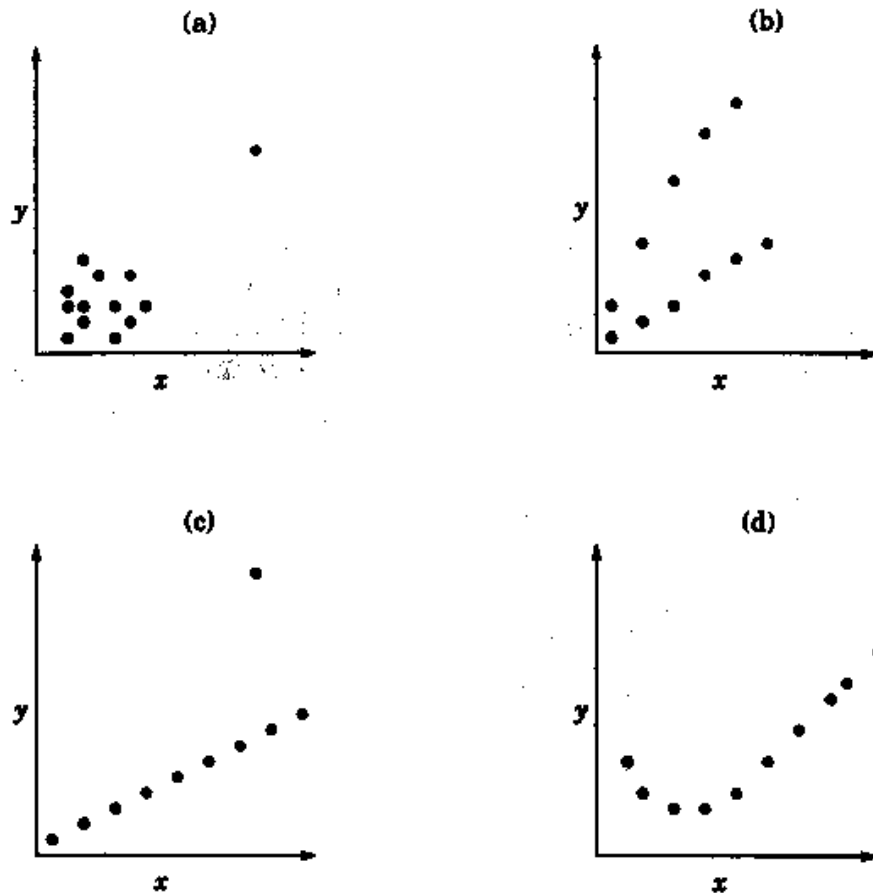


FIGURE 15.5 Examples of data that may give high coefficient of determination, but the linear model obtained may not represent the system correctly.

- Coefficient of determination (R^2) is a measure of the value of the regression (i.e. variation explained by the regression relative to simple second order statistics).
- But it can be misleading if scatter plot is not checked.

Non-linear regression: Make linear through transformation of samples!

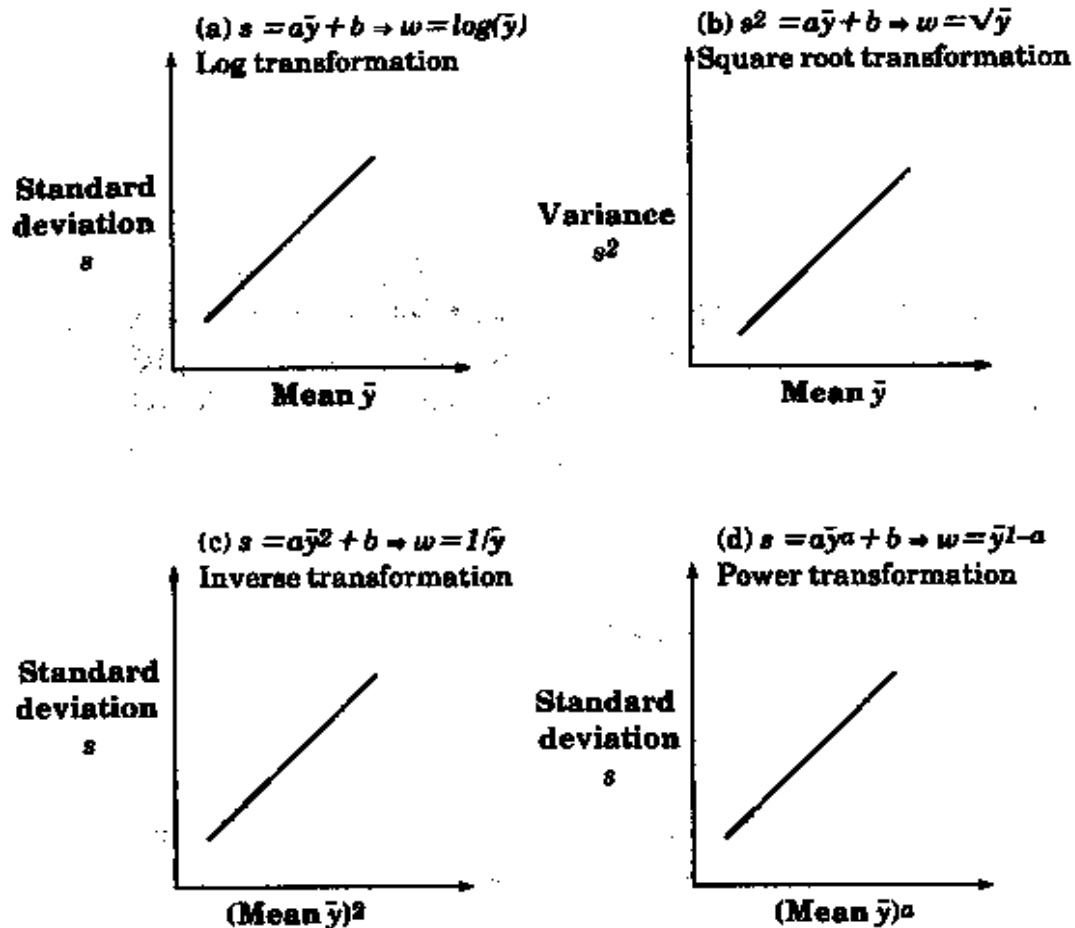


FIGURE 15.2 Standard deviation versus mean response graphs can be used to determine the transformation required.

CIs for Regression & Predictions

