Inferential Statistics

- Sampling and the normal distribution
- Z-scores
- Confidence levels and intervals
- Hypothesis testing
- Commonly used statistical methods

Inferential Statistics

- Descriptive statistics are useful on their own, but are also important as basis for making inferences from a sample of observations to characteristics of the population from which the sample came
  - Mean of a sample can be used to suggest the likely value of the mean of the population
  - Standard deviation of a sample can be used to suggest the likely value of the standard deviation of the population

Inferential Statistics

- Statistical inference
  - Enables use of one or more samples of observations to infer values of a population
  - Test hypotheses to determine if observed differences between groups or variables are real or occur simply by chance
  - Produces new information by making predictions and generalizations based on samples
  - Uses data to make estimates or inferences about population based upon sample
  - Parameter estimating - estimating a parameter based on sample

Inferential Statistics

- Basic assumption: by carefully following certain scientific procedures, one can make inferences about a large group of elements by studying a relatively small number selected from the larger group
- Based on random sampling, probability theory, normal distribution

Inferential Statistics

- Population and sample
  - Population or universe – group of people, things or events having at least one trait in common
  - Must be clearly defined, specifically delimited, carefully chosen
  - **Parameter** is any measure obtained by measuring entire population
- Sample – subset or subgroup of the population
  - **Statistic** is any measure obtained by measuring sample

Sampling

- Population: Set of entities in which researcher is interested, with purpose of being able to describe certain characteristics or make other predictive statements
- Why not study whole population?
  - Logistically impossible
  - Infinite populations
  - Future populations
  - Destructive
  - Too expensive
  - Too time-consuming

Sampling

- Sample: Subset of a population that is examined and from which inferences are drawn about characteristics of population
- Benefits of sampling:
  - Reduced cost
  - Results available sooner
  - Broader scope yields more information
  - Greater accuracy
    - More attention possible to each observation
      - For example, compare results from expending effort to collect 1000 observations in either of the following ways:
        - 1 variable observed on each of 1000 elements in the population
        - 10 variables observed on each of 100 elements in a sample

Sampling

- Much of inferential statistics is based on the sample:
  - Describing how accurately the sample represents the population
  - Making inferences about the population
  - Making predictions about the population
- Techniques of obtaining a valid sample are important

Sampling

- Determining sample size
  - Sample size is directly related to power and effect size
    - Power: the ability to detect "real" differences
      - Number of participants in a sample is directly related to the standard deviation of the sample data set
      - More participants → narrower the distribution → more likely any differences that exist will be detected
  - Effect size: strength of the association between variables and/or the strength (size) of the difference found (small = .20, medium = .50, and large = .80)
    - Greater the effect size → greater the power

Sampling

- How large a sample is necessary for adequate power?
  - Serious statistical texts have formulas to calculate actual and estimated effect sizes
  - Rules of thumb vary by type of test to be used
    - T test, analysis of variance: given a medium-to-large effect size, 30 participants per cell should lead to about 80% power
    - Correlation or regression: at least 50 participants, number increasing with larger numbers of independent variables
    - Chi-Square: no expected frequency in a cell should drop below 5, and overall sample should be at least 20

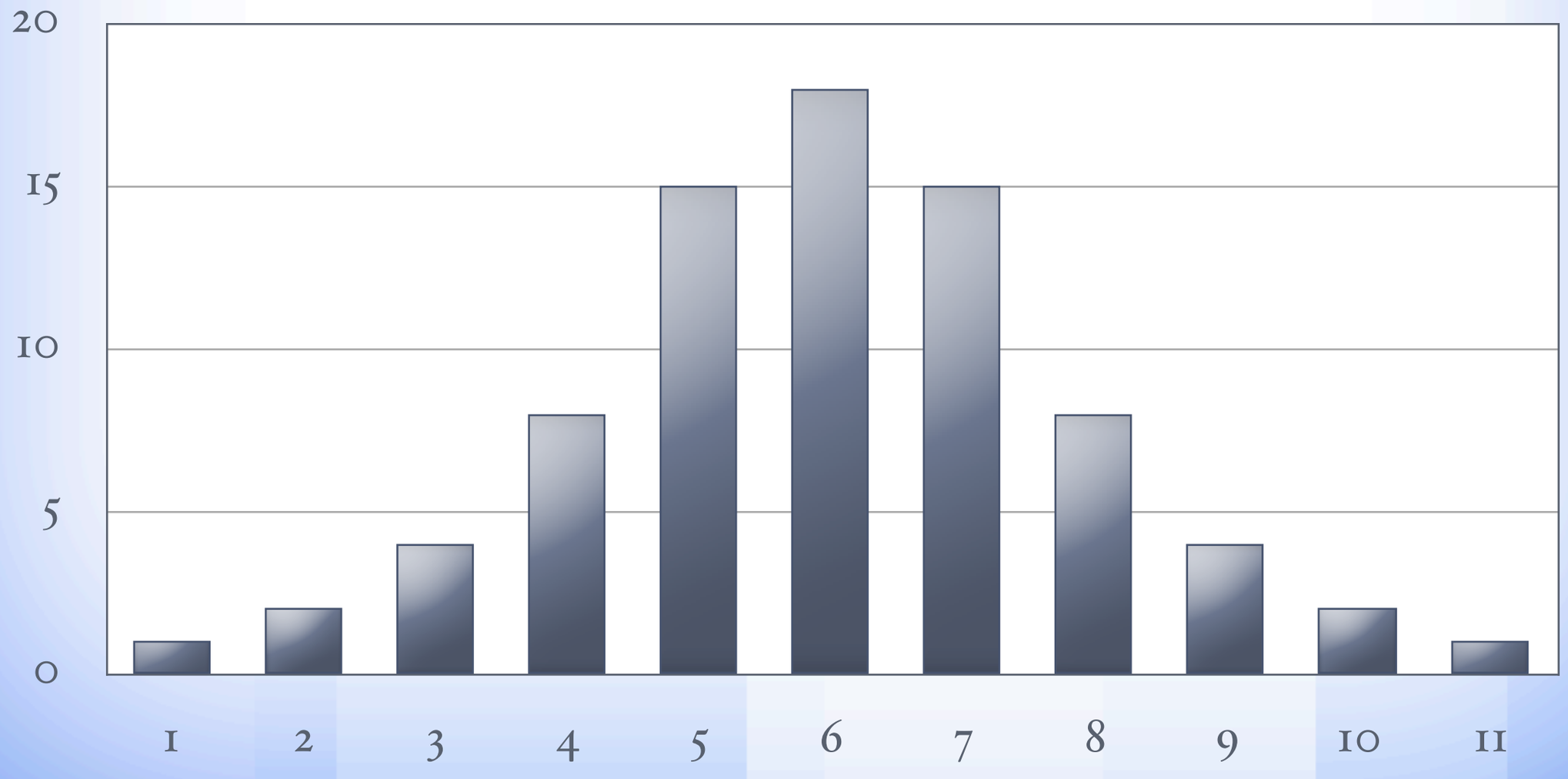
Normal Distribution

- Family of frequency distributions that have same general shape
  - Symmetrical
  - More scores in the middle than in the tails
  - Bell-shaped curve
- Height and spread of a normal distribution can be specified mathematically in terms of:
  - Mean
  - Standard deviation

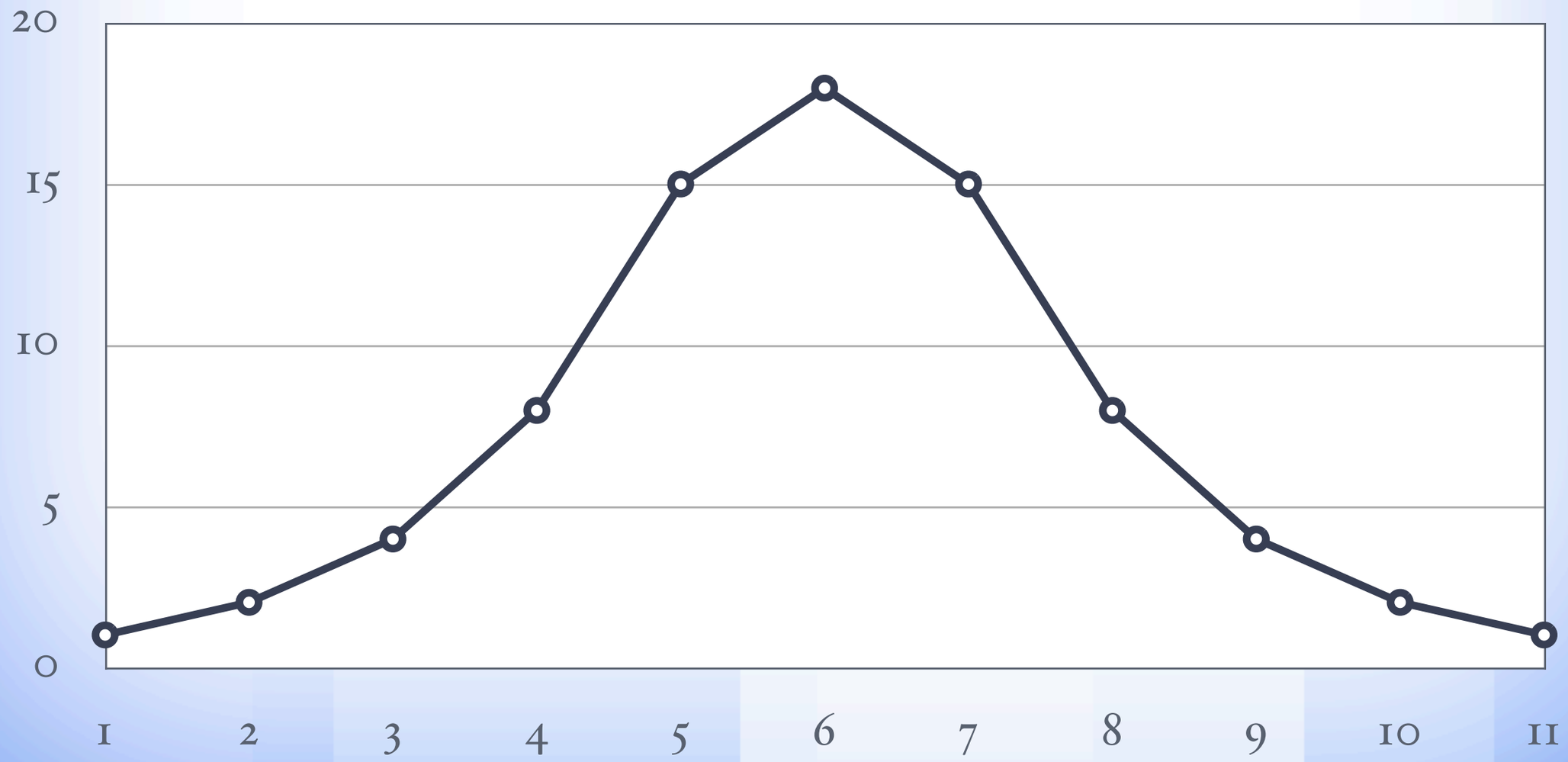
Normal Distribution

- If a set of data is normally distributed, it exhibits particular probability properties:
  - Mean, median, and mode are very similar
  - Curve is symmetrical on either side of mean value
  - Probability of any value being above, or below, the mean value is 0.5
  - Any individual value is more likely to be closer to the central tendency than the extremes of the curve
  - There is a constant relationship between the standard deviation and probability

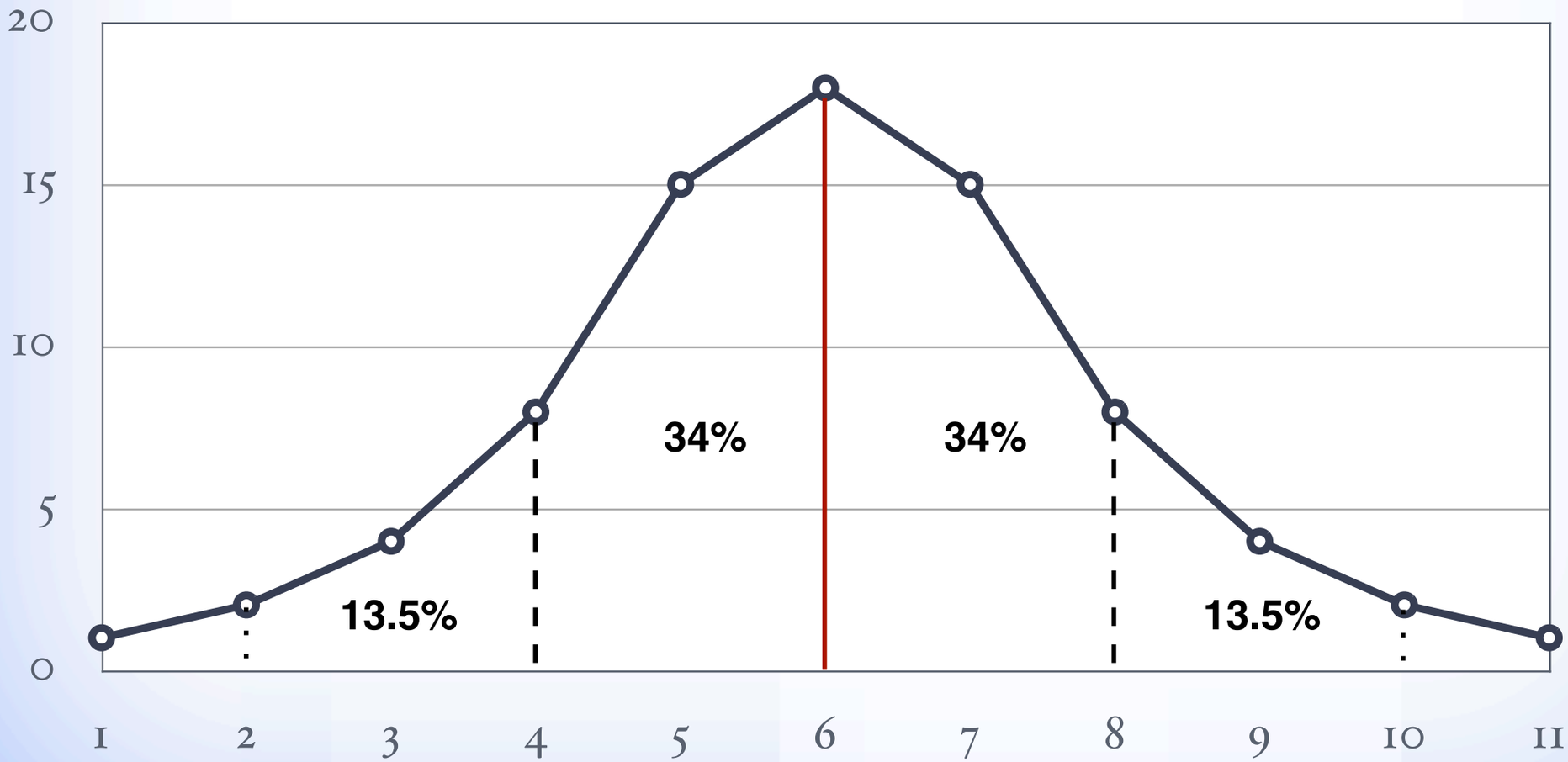
Normal Distribution



Normal Distribution



Normal Distribution



Z-Score

- Standard normal distribution has a mean of 0 and a standard deviation of 1
  - Any normally distributed data set, which may very likely have different mean and standard deviation values, can be transformed to the standard normal distribution through the use of z scores
- Z score: $z = \frac{x - \bar{x}}{s}$

where: x is a value from the original normal distribution

$\bar{x}$ is the mean of the original normal distribution

s is the standard deviation of the original normal distribution

- Z score thus becomes the number of standard deviations that a given value lies above or below the mean
- Relative value of score compared to the sample from which it comes

Z-Score

- Why convert data into Z scores?
  - Enables comparing data between different distributions or two different variables
  - Determines place or relative standing of a given score
  - Example: Mary got 78 on Math, 115 on Science, and 57 on English portions of a national achievement test - how did she perform relative to others?

$$z = \frac{x - \bar{x}}{s}$$

Math	Science	English
$\bar{x}$ = 75	$\bar{x}$ = 103	$\bar{x}$ = 52
S = 6	S = 14	S = 4
X = 78	X = 115	X = 57
z = 0.5	**z = 0.86**	**z = 1.25**

Z-Score

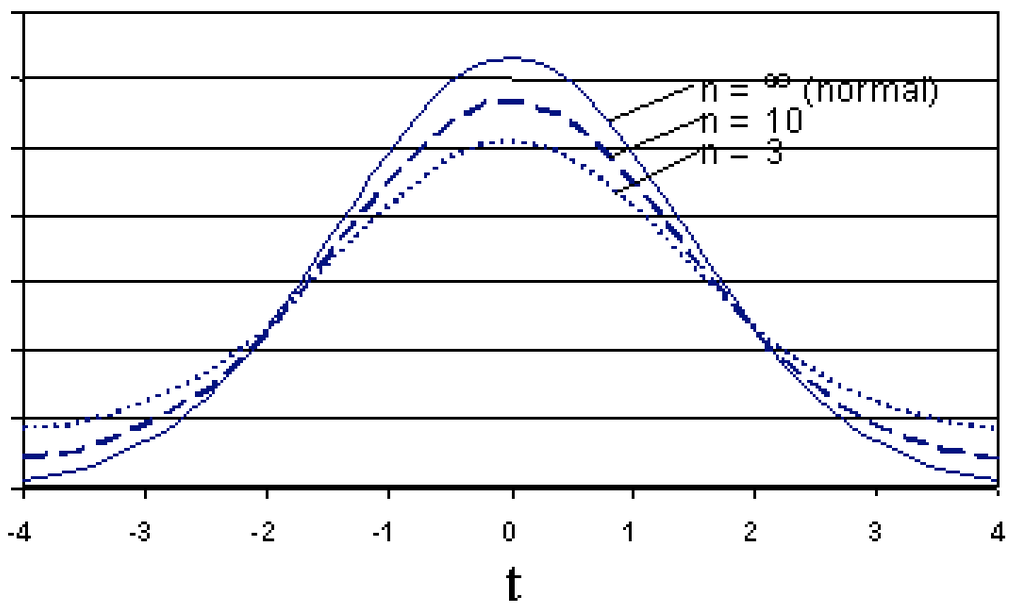
- Using standard normal distribution probability table enables stating z score as a relative percentage and to predict probability of any given score

- If z is positive, add .5 to table score to determine percentage below z score

Math	Science	English
$\bar{x}$ = 75	$\bar{x}$ = 103	$\bar{x}$ = 52
S = 6	S = 14	S = 4
X = 78	X = 115	X = 57
z = 0.5	z = 0.86	z = 1.25
0.1915	**0.2950**	**0.3944**
69.15%	**79.5%**	**89.44%**

T Distribution

- When samples are small (less than 30), the distribution of their means might not be normally distributed
  - Instead, they follow a slightly different distribution, called t-distribution
- Symmetrical like normal distribution but gets increasingly flat as sample size gets smaller



Confidence Levels and Intervals

- Assuming a truly unbiased and representative sample, it is still possible sample does not represent the parameters of the population, simply due to chance
- Any given sample is not likely to be exactly like the population
- Central limit theorem:
  - Collecting many samples will produce variations from population
  - But if you plot all the samples, they will show a normal distribution, and mean of the samples will be population mean (assuming large sample sizes)

Confidence Levels and Intervals

- So if you obtain a sample statistic, how do you know whether it truly represents the population parameter, or is simply a sample that varies by chance?
- Based on central limit theorem, you can state the statistic in terms of probability, in the form of **confidence level** and **confidence interval**

Confidence Levels and Intervals

- Confidence level or level of significance
  - Sample deviates from population according to a normal distribution
    - 99% of sample means lie within 3 standard deviations of population mean
    - 95% of sample means lie within 2 (1.96, to be exact) standard deviations of population mean
    - 68% (68.28%) of sample means lie within 1 standard deviation of population mean
- 95% level is most commonly used
- Higher confidence level used when costs of making incorrect decision based on stated results are high

Confidence Levels and Intervals

- Confidence interval
  - Based upon confidence level and standard deviation
  - Margin of error that you are confident (at stated confidence level) you won't be wrong when inferring or estimating the population mean from the sample mean
    - At 99% confidence level: $\bar{x} \pm 3SD$
    - At 95% confidence level: $\bar{x} \pm 1.96SD$
    - At 68% confidence level: $\bar{x} \pm 1SD$

Confidence Levels and Intervals

- Example: Mean age of sample is 33.2, SD = 2
  - We are 99% confident population is between 27.2 and 39.2
  - We are 95% confident population is between 29.3 and 37.1
  - We are 68% confident population is between 31.2 and 35.2
- As we raise the confidence level, we become "more and more sure of less and less"

Confidence Levels and Intervals

- Example: 45% of Americans believe in the Easter Bunny, with a margin of error of 3%
  - Translation: we are 95% sure that between 42% and 48% of Americans believe in the Easter Bunny
    - Mean = 45, SD = 1.5
    - 45% = the parameter inferred in the population
    - 95% = the confidence level (95% assumed since not stated)
    - 3% = the confidence interval

Hypothesis Testing

- Formulate two competing hypotheses
  - Null hypothesis: $H_O : \mu_1 = \mu_2$
    - μ_1 and μ_2 are means of the population from which samples hypothetically have been obtained
    - Always assumes no change or no difference between populations
    - Innocent until proven guilty
    - Sample means are drawn from same or identical populations
  - Alternative hypothesis: $H_1 : \mu_1$ not equal to μ_2
    - There is a difference between populations
    - Sample means are drawn from different populations

Hypothesis Testing

- Approach is to use statistics to state whether the null hypothesis is supported by the data:
  1. Formulate null hypothesis
  2. Calculate test statistic
  3. Determine probability (p value) of obtaining test statistic assuming hypothesis is true, usually by referring to a table of test-specific critical values
  4. Compare p value to pre-determined significance level (usually .05)
     - If p-value is smaller than significance level, can state that the data does not support null hypothesis (reject null hypothesis)
     - If p-value is greater than significance level, cannot reject null hypothesis
- Note that rejecting null hypothesis does not prove that the null hypothesis is false, just that the probability of it being true is very small

Hypothesis Testing

- Errors of Statistical Testing
  - Results are stated in probability - always a chance that when rejecting/not rejecting a null hypothesis the decision will be wrong
  - Error can happen in two ways:
    - **Type I error**: Null hypothesis H_o is correct and we reject it
      - Conclude there is a difference when there really isn't
      - Can occur only in situations where null hypothesis is rejected
    - **Type II error**: Null hypotheis H_o is wrong but we fail to reject it
      - Conclude there is no a difference when there really is
      - Can occur only in situations where null hypothesis is not rejected

Parametric Versus Non-Parametric Methods

- Parametric statistical methods assume:
  - Normal distribution
  - When comparing groups, scores exhibit similar variance or spread
  - Interval or ratio level data
- Non-parametric methods do not make assumptions about the sample of the population distribution
  - Data are categories or ranks (nominal or ordinal)
  - Usually less powerful
  - Need larger samples

T Test

- Parametric test to determine whether there is a significant difference between two sample means
- Two types of t-tests:
  - Two independent groups
    - Scores in each group are unrelated
    - Use independent t test
  - Matched pairs or repeated measures
    - Scores in groups are related
    - Use paired t test
    - More powerful than independent; higher probability of finding significant difference is one exists, because paired design reduces variability by eliminating individual differences

T Test

- Goal of t test: is there a difference in the populations based on data from samples of those populations?
- Based on two factors:
  - Sample mean difference
    - Difference between the two sample means
    - Larger the mean difference, more likely there is a difference between the two populations
  - Sample data variability
    - Greater variability reduces likelihood that the sample mean difference is result of a real difference

$$t_{calc} = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_{\bar{X}_1 - \bar{X}_2}}$$

T Test

- Example: The library is investigating two methods of teaching library skills to undergraduates. Twenty freshmen were randomly assigned to method COMPUTER and another twenty to method PAPER. After two weeks of instruction, the librarians computed the following from a library skills test.

Method	N	Mean	SD
COMPUTER	20	16.67	4.2
PAPER	20	13.75	5.1

T Test

- Hypothesis: $H_O : \mu_1 = \mu_2$

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}$$

  - Using formula, **t statistic is 2.07**
  - Need to determine probability that null hypothesis is true (p-value) based on t score → t table
    - T table requires degrees of freedom and significance level
      - Degrees of freedom: N1 + N2 - 2
      - Significance level: **0.05**
- Example: df = 20 + 20 -2 = **38**
- Look up critical value of t in T table for **df 30, t05**
  - t critical value = **2.024**
  - t = 2.07 critical value of t = 2.024

T Test

- **t = 2.07 critical value of t = 2.024**
- Do we accept or reject null hypothesis?
- Is there a statistically signficant difference between the samples?

T Test

- **t = 2.07 critical value of t = 2.024**
- Do we accept or reject null hypothesis?
- Is there a statistically signficant difference between the samples?
- We reject null hypothesis because t is greater than the critical value
  - We are confident at a 95% level that the difference in mean scores between methods is not due to chance

Method	N	Mean	SD
COMPUTER	20	16.67	4.2
PAPER	20	13.75	5.1

Chi-Square Test

- Basic purpose is similar to t test
  - Less powerful than t test
  - Why use? Wider applicability -- can deal with:
    - Samples from data sets that are not normally distributed
    - Interval and nominal data
    - More than two samples
      - One-sample X^2 test
      - Two-sample X^2test
      - Three or greater sample X^2tests

Chi-Square Test

- Performing chi-square test
  - Contingency table
    - Columns represent categories of one variable, rows categories of second variable
    - Entries in the cells indicate the frequency of cases for a particular row
    - Expected frequency is the theoretical frequency we would expect to see if the two variables were not related
    - Calculate expected frequencies for each cell = (Row marginal total x column marginal total) / grand total

	Yes	No	Total
Male	5 (15)	45 (35)	50
Female	25 (15)	25 (35)	50
Total	30	70	100

Chi-Square Test

- Null hypothesis – assumes the variables are not related
- Test statistic – chi-square score: X^2
  - Summarizes the discrepancies between the observed and the expected frequencies
- Determine the probability that the null hypothesis is true (the p-value) based on the test statistic
  - Refer to table of critical values of chi-square, using df and significance level
  - Larger the X^2 score, smaller the probability that the null hypothesis is true
- When the calculated X^2 score is equal to or greater than the critical value listed in the chi-square table, reject null hypothesis

Chi-Square Test

- Null hypothesis – assumes the variables are not related
- Example: $X^2 = 19.04$

$$x^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

  - Df = (#rows – 1) X (#columns-1), so df = (2 - 1) x (2 - 1)
- Look up critical value for df 1, p = .05 in chi square critical values table
  - Critical value = 3.84, $X^2 = 19.04$
  - When the calculated X^2 score is equal to or greater than the critical value listed in the chi-square table, reject null hypothesis

	Yes	No	Total
Male	5 (15)	45 (35)	50
Female	25 (15)	25 (35)	50
Total	30	70	100

Correlation

- Describes relationship between two sets of data
  - Measurement of association
  - Are variable X and variable Y related?
  - How are variable X and variable Y related?
  - How strongly are X and Y related?

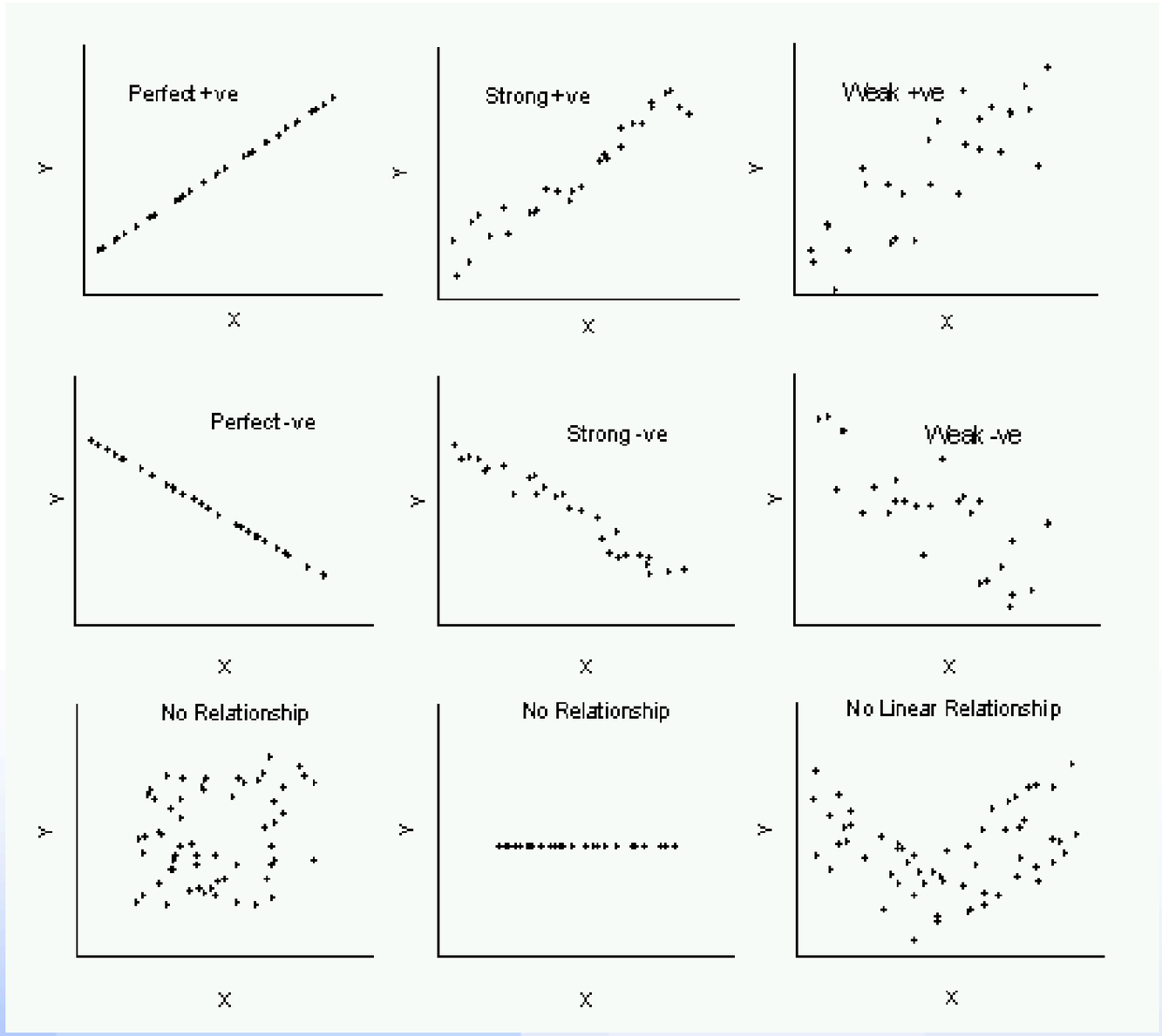
Correlation

- Positive correlation
  - Value of one variable increases as the value of other variable increases
    - Height and weight
    - Advertising and sales
    - Book borrowing frequency and frequency of book repair
- Negative correlation
  - Value of one variable decreases as the value of other variable increases
    - Highway speeders and state trooper vehicles
    - Recall and precision

Correlation

- Linear correlation: perfect correspondence between variables
  - As X increases by constant quantity, Y increases by constant quantity
  - Can be perfect positive (correlation coefficient = +1) or perfect negative correlation (correlation coefficient = -1)
- Curvilinear correlation: variables correlate along a curved line
  - Age and coordination
- No correlation: when one variable changes, the other shows no trend to change in a uniform way (correlation coefficient = 0)

Correlation



Correlation

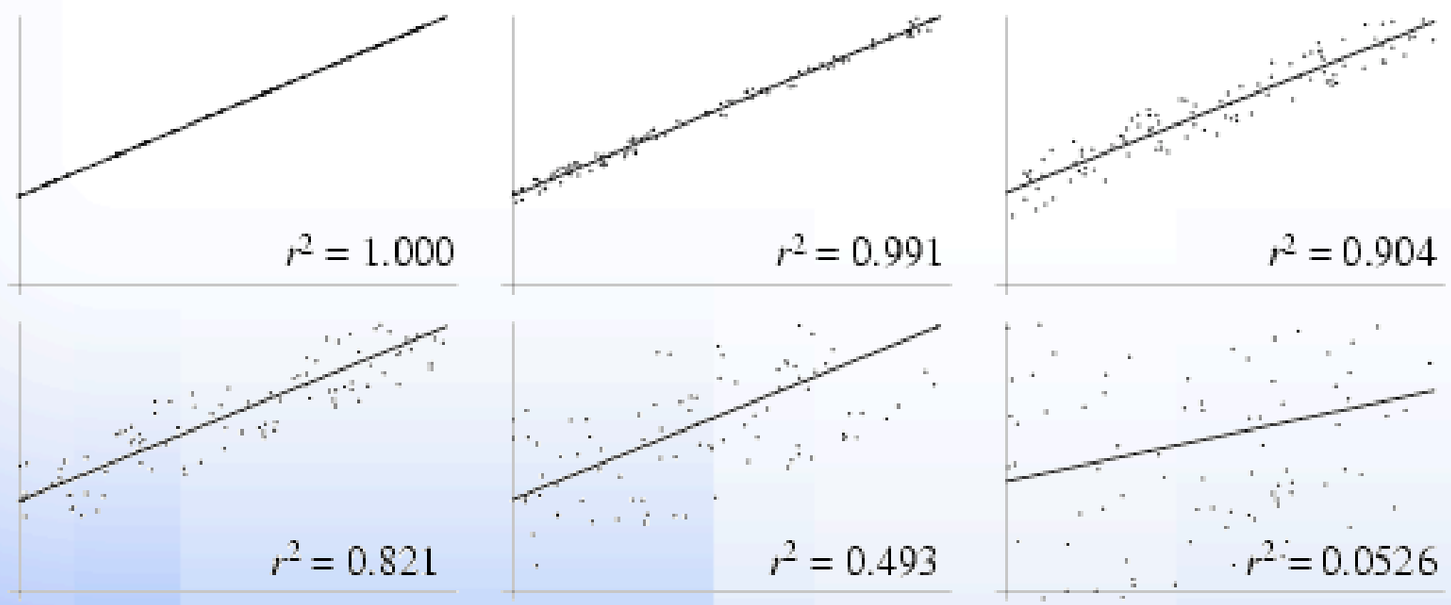
- Measuring strength of a relationship
  - Calculate correlation coefficients
  - Pearson's r

$$r = \frac{\Sigma z_x z_y}{N}$$

  - Derived from z scores of two distributions to be correlated
  - Interval or ratio data, assumes normal distribution of variables and linear relationship between variables
  - If assumptions are not met, can use version called Spearman's Rank Correlation Coefficient

Correlation

- Pearson r correlation coefficient ranges from -1 to 1
- Positive r means positive correlation and negative r means negative correlation
- Closer r is to 0, weaker the relationship



Correlation

- Testing significance of Pearson r
  - Pearson r coefficient by itself is a descriptive statistic
  - To conclude that the r coefficient indicates a relationship in the population, must test for signficance of r

Correlation

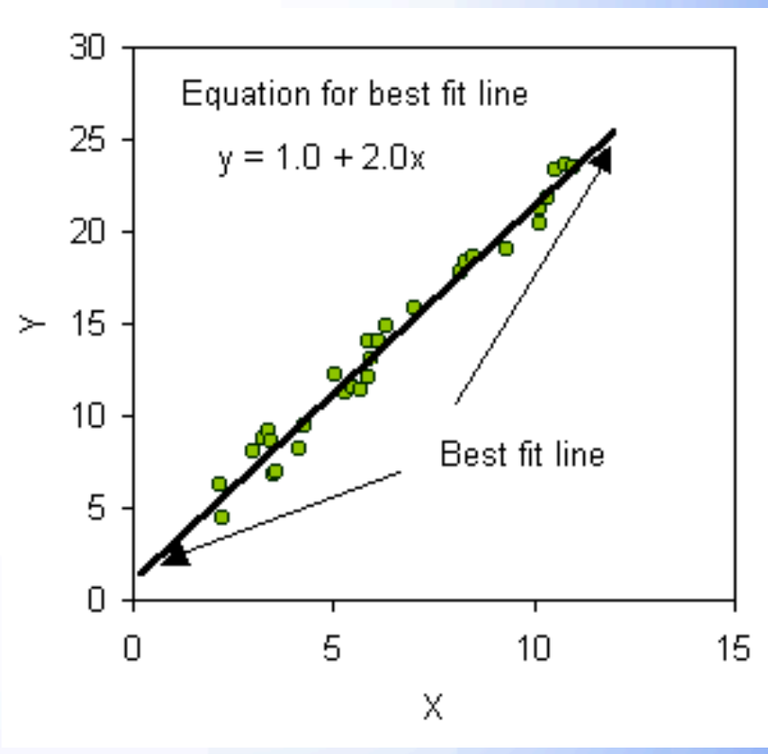
- Testing significance of Pearson r
  - Same basic procedure for hypothesis testing:
    - Formulate null hypothesis
    - Calculate test statistic
    - Determine probability (p value) of obtaining test statistic assuming hypothesis is true
      - Refer to table of critical values of r, based on degrees of freedom (number of pairs - 2)
    - Compare p value to pre-determined significance level (usually .05)
      - If p-value is smaller than significance level, can state that the data does not support null hypothesis (reject null hypothesis)
      - If p-value is greater than significance level, cannot reject null hypothesis

Correlation

- Correlation and causation
  - Does correlation mean causation?
    - Statistically significant correlation coefficient only means that there is a strong chance of a real relationship between variables, not just coincidence or chance
    - Does not mean that X caused Y or that Y caused X
    - Other factors, or a third variable Z could actually affect both X and Y
  - Even without causation, correlation is useful to know
  - Can make informed prediction of value of one variable based on value of another

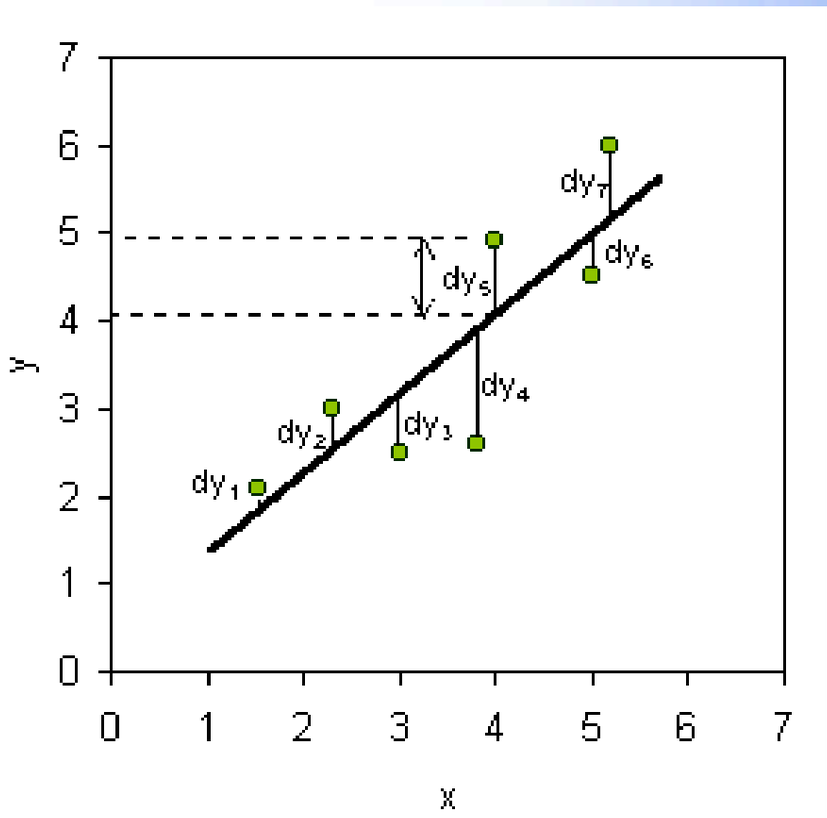
Regression

- Knowing two variables are correlated, how can we predict value of one variable based on the other?
- Regression analysis
  - Variety of forms, most common is linear regression
  - Goal is to find the best-fit straight line through data set
    - Single straight line that comes closest to all of data points in scatter plot
    - Y = a + bX



Regression

- Regression equation and regression line
  - Least-squares regression
    - Define a line such that the sum of the squared residuals are minimized
    - Residuals are difference between actual and predicted values -- observed y-values and the y-values predicted by solving for x in the regression equation
    - Using formula for least-squares, you end up with a regression equation that describes the best-fit line: y = 2.09 + 0.65x



Regression

- With regression equation, can input value for X and calculate the corresponding value for Y
- Calculating **accuracy** of prediction: coefficient of determination
- Square of correlation coefficient r: $\mathbf{r^2}$
- If r is .8, coefficient of determination $\mathbf{r^2}$ is .64, which means 64% of variation in Y can be attributed to X

Regression

- Multiple regression
  - Widely used research tool for anlayzing more complex situations
  - Predicts value of variable based on 2 or more other variables