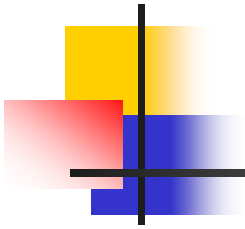
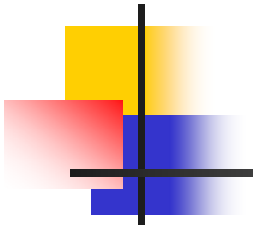


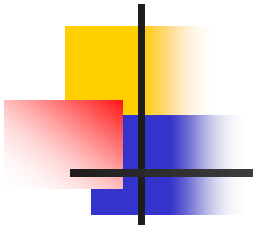
Reasoning with Uncertainty

Statistics and Hypothesis Testing



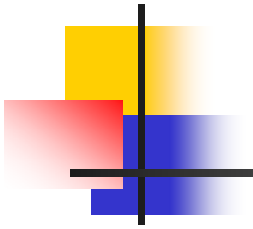
Hypothesis Testing

- Hypothesis testing is a statistical method used to evaluate if a particular hypothesis about data resulting from an experiment is reasonable.
  - Uses statistics to represent the data
    - Value of the data
    - Distribution of the data
  - Determine how likely it is that a given hypothesis about the data is correct



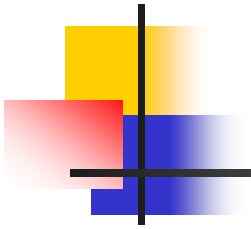
Statistics

- Statistics attempt to represent the important characteristics of a set of data items (or of a probability distribution) and the uncertainty contained in the set (or the distribution).
  - Statistics represent different attributes of the probability distribution represented by the data
  - Statistics are aimed at making it possible to analyze the data based on its important characteristics



Statistics

- A number of important statistics can be used to characterize a data set (or a population from which the data items are drawn)
  - Mean
  - Median
  - Mode
  - Variance
  - Standard deviation



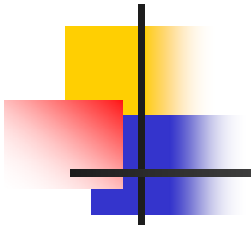
Mean

- The arithmetic mean μ represents the average value of data set $\{X_i\}$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

  - The arithmetic mean is the expected value of a random variable, i.e. the expected value of a data item drawn at random from a population

$$\mu = E[X]$$



Median and Mode

- The median m is the middle of a distribution

$$\left|\{X_i \mid X_i \leq m\}\right| = \left|\{X_i \mid X_i \geq m\}\right|$$

- The mode of a distribution is the most frequently (i.e. most likely) value

Variance and Standard Deviation

- The variance σ^2 represents the spread of a distribution

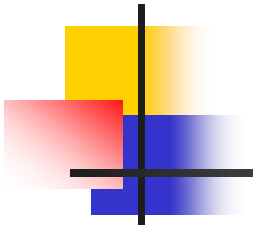
$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

  - In a data set $\{X_i\}$ an unbiased estimate s^2 for the variance can be calculated as

$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N-1}$$

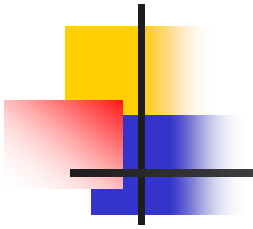
    - $N-1$ is often called the number of degrees of freedom of the data set

- The standard deviation σ is the square root of the variance
  - In the case of a sample set, s is often referred to as standard error



Examples

- Statistics of a distribution
  - http://www.ruf.rice.edu/~lane/stat_sim/descriptive/index.html



Hypothesis Testing

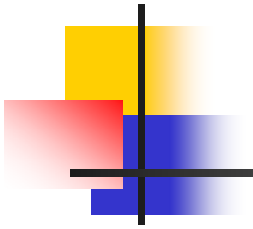
- Hypothesis testing is aimed at establishing if a particular hypothesis about a set of observations (data) should be trusted
  - Example:
    - The average and variance of the body height of the population of a country is

$$\mu = 1.7 \quad , \quad \sigma^2 = 0.01$$

    - In a different country a set of 10 people are randomly selected and measured resulting in the following data set with mean $\overline{X} = 1.776$:

$$\{1.8, 1.9, 1.92, 1.75, 1.7, 1.77, 1.82, 1.75, 1.65, 1.7\}$$

    - Can we conclude that people in this second country are on the average taller (average height μ_X) than people in the first one ?



Hypothesis Testing

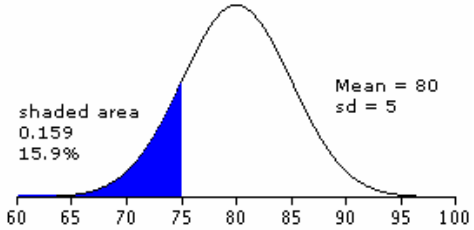
- To be able to trust in a hypothesis on statistical data we have to make sure that the data set could not be the result of random chance
  - In the example the hypothesis would be:

    $H: \quad \mu_X > \mu$

  - To determine if the hypothesis has a base we have to make sure that we do not accept it if the data could be the result of random chance
    - What is the likelihood that the data could be obtained by randomly sampling 10 items from the distribution in the first country ?

Percentiles

- To determine the likelihood that a data item could come from a distribution we have to be able to determine percentiles
  - A data item belongs to the n^{th} percentile if the likelihood to obtain a value that is equal to the data item or even further away from the distribution mean is greater or equal to $n\%$



  - For certain distributions (e.g. normal distribution) percentiles can be relatively easily calculated
    - http://davidmlane.com/hyperstat/z_table.html

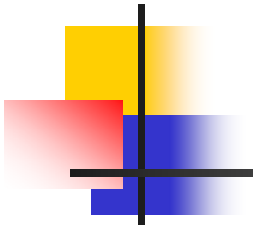
Percentiles in Normal Distributions

- The percentile in a normal distribution is a function of the distance of the data value from the mean and of the standard deviation

$$z = \frac{X - \mu}{\sigma}$$

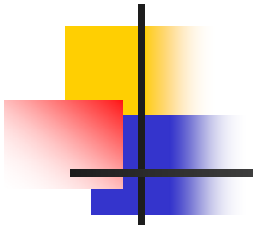
z	Area from −∞ to z
-3.0	.0013
-2.5	.0062
-2.0	.0227
-1.5	.0668
-1.0	.1587
-0.5	.3085
0.0	.5000
0.5	.6915
1.0	.8413
1.5	.9332
2.0	.9772
2.5	.9938
3.0	.9987

  - E.g. a data value that is more than 1.5 standard deviations larger than the mean of the distribution occurs only with probability *0.0668*



Percentiles

- If the distribution of the population is normal, the z-value and the z-table allow to compute how likely it would be to randomly draw the particular data value (or one even further from the mean)
  - If the likelihood is not very small, then we should not assume that the data value is significant different from the value of the distribution
- Percentiles for general, skewed distributions are difficult to derive
  - Attempt to formulate hypothesis on a statistic for which the distribution is approximately normal



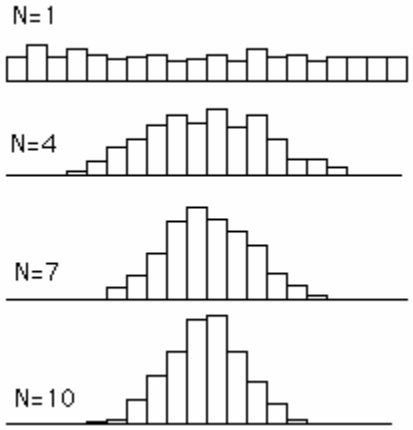
Sampling Distributions

- Sample Distribution: The probability distribution representing individual data items
- Sampling Distribution: The probability distribution of a statistic calculated from a set of randomly drawn data items
  - Sampling Distribution of the mean: The distribution of the means of random data samples of size n
    - For a sample distribution with mean μ and standard deviation σ the mean μ_s and standard deviation σ_s of the sampling distribution of the mean over n samples is:

$$\mu_s = \mu \ , \ \sigma_s = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

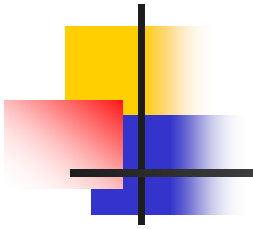
- For any sample distribution with mean μ and standard deviation σ, the sampling distribution of the mean approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as n becomes larger
  - Percentiles for the sampling distribution of the mean are easier to compute than for the sample distribution.



  - http://onlinestatbook.com/simulations/CLT/clt.html

Logic of Hypothesis Testing

- The goal of hypothesis testing is to establish the viability of a hypothesis about a parameter of the population (often the mean)
  - Define hypothesis (also called alternative hypothesis)
    - E.g.: $H_A: \mu_X > \mu$
  - Set up Null hypothesis (i.e. the “opposite” of the hypothesis)
    - E.g.: $H_0: \mu_X = \mu$
  - Compute the percentile and thus the likelihood of the Null hypothesis
  - If the Null hypothesis has more than a small likelihood, the data does not significantly support the hypothesis (since it could also represent the Null hypothesis)
    - Usually thresholds or 5% or smaller are used

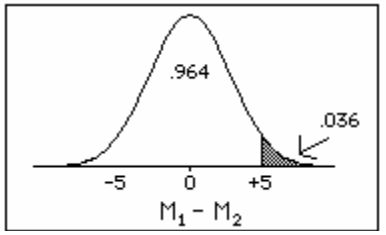


Logic of Hypothesis Testing

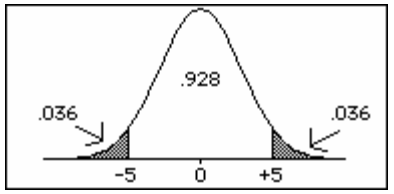
- If the Null hypothesis' likelihood (i.e. the likelihood to obtain data at least as extreme) is smaller than the significance level, the Null hypothesis can be rejected
  - Rejection implies that the Null hypothesis is discarded in favor of the alternative hypothesis and the result is considered significant
    - Note that a p-value less than 5% for the Null hypothesis does NOT imply a likelihood of 95% for the alternative hypothesis.
    - Note that it is NOT possible to show that the Null hypothesis is correct. Failure to reject the Null hypothesis does NOT imply acceptance of the Null hypothesis but rather that no significant conclusion could be drawn from the test

One-Tailed vs. Two-Tailed Tests

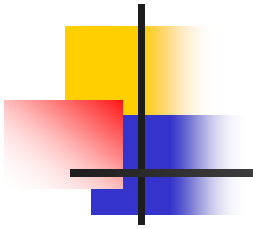
- Depending on the hypotheses we might be interested to know how the likelihood to generate data that is more extreme than the test data in a particular direction (e.g. the likelihood of it being larger than or equal to the given data) or in any direction (i.e. that it is further from the mean than the given data)
  - If we are only interested in data on one end of the distribution we perform a one-tailed test, i.e. we only count the percentile at one end of the distribution



  - If we are interested in both sides, we perform a two-tailed test which computes the percentile at both ends



  - If we are not sure we should choose a two-tailed test (which is more stringent)



The Z Test

- The Z Test is the most basic hypothesis test to evaluate a hypothesis relating an unknown distribution (with mean μ_X) from which a known sample set $\{X_i\}$ of size n with mean $\overline{X}$ was randomly drawn to a population with sample distribution with mean μ and standard deviation σ
  - Assumes that the sampling distribution of the means is normal
    - Either the sample distribution is normal or the sample size is very large
  - Example Hypotheses:

$$H_0 : \mu_H = \mu$$

$$H_A : \mu_H > \mu$$

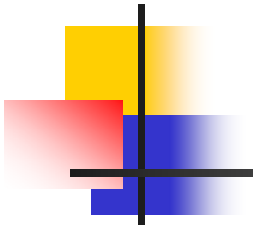
  - Compute z-value:

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

  - Translate z-value to p-value and evaluate significance
    - Translation usually uses z-table. E.g. p = 2.5% -> z=1.96

Z Test With Unknown Variance

- If the standard deviation of the population is unknown we can make the assumption that the population and the data set have come from populations with the same standard deviation
  - Use standard error s of the sample set to estimate standard deviation of the sampling distribution $$\sigma_s = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$
  - Compute z-value: $$z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$
  - Translate z-value to p-value and evaluate significance
    - Translation usually uses z-table. E.g. p = 2.5% -> z=1.96



Student's t Distribution

- If the sample size is small and the form of the sample distribution is unknown a normal distribution might not be the correct distribution for the sampling distribution of the mean
  - Student's t distribution addresses this by increasing the spread of the distribution as the sample size decreases
    - For large sample sizes Student's t approximates the normal distribution arbitrarily well
    - For small sample sizes Student's t models the deviations in the variance estimates
  - http://www.econtools.com/jevons/java/Graphics2D/tDist.html

The t Test

- The t test operates in the same way as the Z test but uses Student's t distribution instead of the normal distribution
  - Example Hypotheses: $H_0 : \mu_H = \mu$

    $H_A : \mu_H \neq \mu$
  - Compute t-value: $t_{n-1} = \frac{\overline{X} - \mu}{s/\sqrt{n}}$
  - Translate t-value to the corresponding p-value (percentile) according to the Student's t distribution for sample size n and evaluate significance
    - Translation usually uses t-table. E.g. p = 2.5% -> t_9=2.26

The t Test

- The t test should be used whenever the sample size is smaller than approximately 30
- Example:
  - The average and variance of the body height of the population of a country is $\mu = 1.7 \quad , \quad \sigma^2 = 0.01$
  - In a different country a set of 10 people are randomly selected and measured resulting in the following data set with mean $\overline{X} = 1.776$:
    $\{1.8, 1.9, 1.92, 1.75, 1.7, 1.77, 1.82, 1.75, 1.65, 1.7\}$
  - Can we conclude that people in this second country are on the average taller (average height μ_X) than people in the first one ?
    - Hypotheses: $H_0 : \mu_X = \mu \quad , \quad H_A : \mu_X > \mu$
    - T value: $t_9 = \frac{1.776 - 1.7}{0.1/\sqrt{10}} = 2.403 > 2.26$
    - Reject Null hypothesis in favor of alternative hypothesis.
      - People in the second country are on average taller than in the first country

Two-Sample t Test

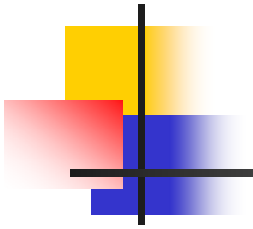
- A two-sample test is to compare two samples to see whether they come from the same or different distributions
  - E.g.: Does algorithm 1 perform better than algorithm 2 based on a set of experiments performed with each
  - Since no population standard deviation or mean is available, the standard error from the two samples is pooled to obtain an estimate of the standard deviation of the difference between the two sample distributions

$$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  - Example Hypotheses: $H_0 : \mu_1 = \mu_2 \;\; , \;\; H_A : \mu_1 \neq \mu_2$
  - Compute t-value:

$$t_{n_1+n_2-2} = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)}{\sigma_{\overline{X_1}-\overline{X_2}}} = \frac{\overline{X_1} - \overline{X_2}}{\sigma_{\overline{X_1}-\overline{X_2}}}$$

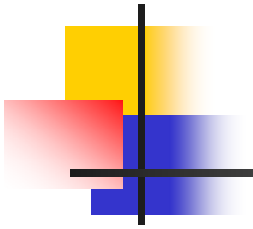
  - Translate t-value to the corresponding p-value (percentile) according to the Student's t distribution for n_1+n_2-2 degrees of freedom and evaluate significance



Paired Sample t Test

- A paired sample test is used to compare two sample sets that have a different common variable that should be controlled for to see whether they come from the same or different distributions
  - E.g.: Does algorithm 1 perform better than algorithm 2 based on their performance on a specific set of problems (the same problems for both)
  - A paired sample test avoids the variance caused by the controlled variable (e.g. the specific problem the algorithm is applied to) by establishing the sampling distribution over the differences in the value between paired data items from both sets $\{X_\delta^{(i)} = X_1^{(i)} - X_2^{(i)}\}$
  - Example Hypotheses: $H_0 : \mu_\delta = 0 \;,\; H_A : \mu_\delta > 0$
  - Compute t-value: $$t_{n-1} = \frac{\overline{X}_\delta - \mu_\delta}{s_\delta / \sqrt{n}}$$
  - Translate t-value to the corresponding p-value (percentile) according to the Student's t distribution for sample size n and evaluate significance

Paired Sample vs. Two-Sample Test



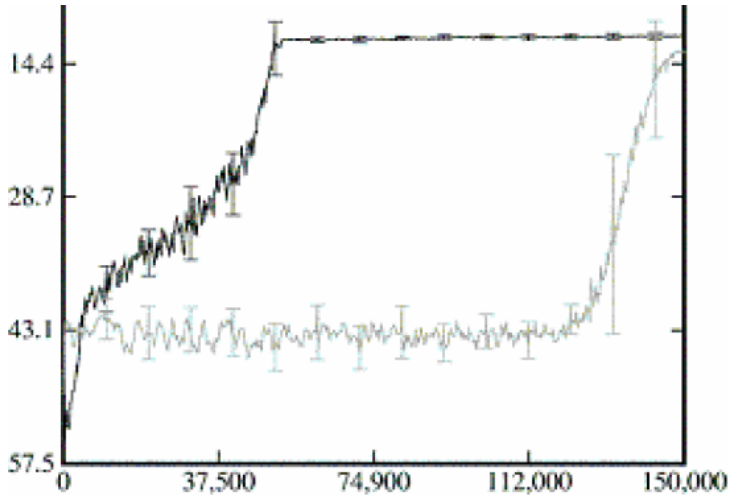
- The paired sample test is preferable whenever an additional variable is known which produces variations in the data items
  - Paired sample test often has smaller standard deviations because of the avoided variance
- If no conditional variable that would pair individual samples together is known to be relevant, the two-sample test is most of the time better because it uses more samples

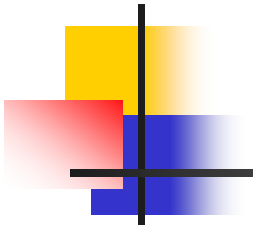
Confidence Intervals

- Confidence intervals on the means of data points (or curves) indicate intervals for which, if a data point from a different sample falls within it, a significance test would not succeed to reject the Null hypothesis.
  - E.g.: The performance for system 1 is significantly better than the performance of system 2 if the performance values lie outside the confidence intervals.
  - A $(1\text{-}\alpha)\%$ confidence interval around a data point $\overline{X}$ would cover all values for which the t-value with respect to $\overline{X}$ would have a p-value below $\alpha\%$
    - Confidence interval bounds: $\left[\overline{X} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right]$
  - http://www.math.csusb.edu/faculty/stanton/m262/confidence_means/confidence_means.html

Confidence Intervals

- When presenting and comparing performance data (and making statements regarding performance differences) either significance test should be performed or error bars (confidence intervals) should be presented with the data
  - Error bars illustrate the significance of the difference between two performance measures
    - Error bars usually either represent $(1-\alpha)\%$ confidence intervals or are of size σ





Significance Testing

- To be able to make statements comparing performance derived from experiments it is necessary to show that the differences are not the result of chance
- Benefits
  - Significance tests are a flexible way to evaluate if a hypothesis about the sampling mean (or some similar statistics) has significant support
  - Significance tests can be applied without complete knowledge of the distributions underlying the problem
- Problems:
  - Significance tests only reject the Null hypothesis
    - No direct proof of the hypothesis
  - Significance tests are difficult when trying to evaluate hypotheses that are not involving the mean