**Applied Geoscience Data Analysis Using Matlab**

*17 Sept 2013*  
(Lecture 8)

**Hypothesis Testing**  
KS goodness-of-fit test, t-test, f-test  
bootstrap approach

*Sections 3.6-3.8 Trauth*

---

**Steps**

1. Apply some formula to the data to a compute a statistic that informs us about the null hypothesis.

2. Find where that value falls in a probability distribution computed on the basis of the “null hypothesis.”

3. If it falls in an unlikely spot (on distribution tail), conclude that we have disproven the null hypothesis.

---

**Hypothesis Testing**

**Goals**

Generate a statement of comparison between a sample and a known population or between two samples.

This statement, the null hypothesis $H_0$, is usually formulated so that it will be **rejected** (“proven” false).

— **Examples of a null hypothesis**: The chromium content of sample A is the same at that from sample B.  
$H_0 : Cr_A = Cr_B$

---

**Hypothesis Testing**

**Example – Chi-squared Goodness-of-Fit**

**Steps**

1. Apply some formula to the data to a compute a statistic that informs us about the null hypothesis.

The value of the test-statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where

- $X^2$ is the test statistic that asymptotically approaches a $\chi^2$ distribution.
- $O_i$ is an observed frequency.
- $E_i$ is an expected (theoretical) frequency, asserted by the null hypothesis;
- $n$ is the number of possible outcomes of each event.
2. Find where that value falls in a probability distribution computed on the basis of the “null hypothesis.”

3. If it falls in an unlikely spot (on distribution tail), conclude that we have disproven the null hypothesis with a given level of confidence.

**K-S (Kolmogorov-Smirnov) Goodness-of-Fit Test**
- Used for comparing two distributions, via empirical CDF’s
- H₀ is that they are the same distribution
- Non-parametric Test (as was the chi-squared)
- Does not need to be performed on binned data

**The test statistic is:**
\[ \text{max}[F_1(x) - F_2(x)] \]

\( F_1(x) \) is the CDF of x₁
\( F_2(x) \) is the CDF of x₂

**Example from last lecture organicmatter_one.txt**
\[ \text{chi2}=\text{sum}((\text{CNexp-CNobs}).^2./\text{CNexp}) = 8.95 \]

**Critical Value**
\[ \text{chi2inv}(0.95,5) = 11.07; \quad 95 \% \text{ confidence} \]

**Cannot Reject**

**Critical Values for KS-test**

Let \( n \) be the sample size. Then the following critical values are used for the test:

For significance level \( \alpha = 0.10 \) (where \( \alpha = 1 - p \); or the probability or being true is less than 0.10)
a critical value of \( 1.22/V(n) \) is used.

For significance level \( \alpha = 0.05 \),
a critical value of \( 1.36/V(n) \) is used.

For significance level \( \alpha = 0.01 \),
a critical value of \( 1.63/V(n) \) is used.
The example of organic matter....
OGM=load('organicmatter_one.txt');
sOGM=sort(OGM);
obsCDF=(1:length(sOGM))./length(sOGM);
stairs(sOGM, obsCDF,'g')

Expected value is evaluated at each sorted OMG data value
expCDF=normcdf(sOGM,mean(OGM),std(OGM));
stairs(sOGM,expCDF,'r','LineWidth',2);

DIF=abs(expCDF-obsCDF');
maxDIF=max(DIF); % 0.0757
i=find(DIF==max(DIF))
plot(sOGM(i),0,'x')
kscrit=1.36/
sqrt(length(sOGM)) %0.1756

Parametric Hypothesis Tests
- Assume Certain Distribution (e.g., normal)
- Can be sensitive, provided assumptions are met.

Today we will look at two common tests that assume normally distributed data:

t-test
- test if the mean of a sample is different than that of a populations with known mean

F-test
- test if the variances of two samples are different
**T-test:** Test if a sample is derived from a population with a given mean and unknown variance

- $x$-bar = mean of our sample
- $SE$ = standard error of our sample
- $\mu$ = mean of population we're testing against

- Assumes similar variances
- Normal distributions

$$t = \frac{x - \mu}{SE}$$

$SE = \frac{\sigma}{\sqrt{n}}$

Reject the null hypothesis (means of the two are equal) if $t > t_c$.

$tinv(P, df) \ % \ find \ critical \ value$

**EXAMPLE:** The worldwide population of *Composita* (a type of shell) has mean size of **14.2 mm**.

We collect a sample of 6 shells from a population: $x = [33.9625 \ 31.3565 \ 23.5298 \ 33.5815 \ 31.8559 \ 23.3230]$.

**Question:** Is it likely that the population that the samples came from is the same as the other (global) population or have we discovered a new population of Composita?

**Formulate the null hypothesis:** $H_0$

- $\mu_{our \ population} = \mu_{worldwide}$

$SE = \text{std}(x)/\sqrt{n} = 1.99; \ \text{mean}(x) = 29.60$

$$t = \frac{x - \mu}{SE} = \frac{29.6 - 14.2}{1.99} = 7.7387$$

If $t$ is large enough then we can reject the NULL hypothesis

$\text{tinv(0.95,5) \ % \ find \ critical \ value \ at \ 95\% \ conf. = 2.0150}$

$t > t_{\text{critical}}$

**To specific level of significance:**
- Need calculated $t$ and d.f. to look up critical value.
- Since the mean is used in calculating the standard error, the number degrees is n-1 = 5

$tinv(0.95,5) \ % \ find \ critical \ value \ at \ 95\% \ conf. = 2.0150$

Can reject null hypothesis at 99.95% conf. level OR 0.005 significance

<table>
<thead>
<tr>
<th>SIGNIFICANCE LEVELS ($\alpha$)</th>
<th>df</th>
<th>0.01</th>
<th>0.025</th>
<th>0.05</th>
<th>0.075</th>
<th>0.10</th>
<th>0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.324920</td>
<td>1.000000</td>
<td>3.077684</td>
<td>6.313745</td>
<td>12.70620</td>
<td>31.8205</td>
<td>63.65674</td>
</tr>
<tr>
<td>2</td>
<td>0.288675</td>
<td>0.816497</td>
<td>1.684890</td>
<td>2.919200</td>
<td>4.30265</td>
<td>6.96456</td>
<td>9.92445</td>
</tr>
<tr>
<td>3</td>
<td>0.276671</td>
<td>0.764802</td>
<td>1.637514</td>
<td>2.575826</td>
<td>3.84145</td>
<td>5.84107</td>
<td>8.84145</td>
</tr>
<tr>
<td>4</td>
<td>0.270722</td>
<td>0.740697</td>
<td>1.533206</td>
<td>2.230882</td>
<td>3.21861</td>
<td>4.54040</td>
<td>7.37485</td>
</tr>
<tr>
<td>5</td>
<td>0.267181</td>
<td>0.726687</td>
<td>1.475884</td>
<td>2.015048</td>
<td>2.57058</td>
<td>3.84145</td>
<td>5.50188</td>
</tr>
</tbody>
</table>
\[ t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \]

where \( \bar{X} \) is the sample mean, \( \mu \) is the hypothesized population mean, \( s \) is the sample standard deviation, and \( n \) is the sample size. Under the null hypothesis, the test statistic will have Student's t distribution with \( n - 1 \) degrees of freedom.

\[ [h,p] = ttest(\ldots) \]

returns a 100*(1 - alpha)%

confidence interval on the population mean, or on the difference of population means for a paired test.

What if:
\[ x = [10.4623 \ 11.3760 \ 9.2422 \ 14.0254 \ 17.9509 \ 16.8862] \]

\[ [h,p]=ttest(x,14.2) \]

H = 0 (we can NOT reject the hypothesis at the 95% confidence level... they might be from the same population)

p = 0.5725 (large probability of \( H_0 \) being true... i.e., the data are sampled from a population with mean of 14.2)

\[ t = \frac{\bar{a} - \bar{b}}{\sqrt{\frac{\sum (a_i - \bar{a})^2}{n_a} + \frac{\sum (b_i - \bar{b})^2}{n_b}}} \]

The denominator makes a "pooled" estimate of the Standard Deviation.

\[ d.f= n_a+n_b-2 \]

a-bar= mean(a)  b-bar=mean(b)
In Matlab

h=1 reject
h=0 cannot reject

In Matlab, the `ttest2` function is used to perform a t-test of the hypothesis that two independent samples, in the vectors X and Y, come from distributions with equal means, and returns the result of the test in H. H=0 indicates that the null hypothesis ("means are equal") cannot be rejected at the 5% significance level. H=1 indicates that the null hypothesis can be rejected at the 5% level. The data are assumed to come from normal distributions with unknown, but equal, variances. X and Y can have different lengths.

\[ H = \text{ttest2}(X,Y) \]

\[ \{H,P\} = \text{ttest2}(X,Y) \]

\[ \{H,P,CI\} = \text{ttest2}(X,Y) \]

Are these means different?

```matlab
>> mean(corg1) = 25.5816
>> mean(corg2) = 25.1507
```

Our null hypothesis is that they are not different.

```matlab
[h,p,ci] = ttest2(corg1,corg2,0.05)
```

- \( h = 0 \) (can NOT reject the null hypothesis that samples come from a population with the same mean at 95% confidence level)
- \( p = 0.0745 \) (7.45% chance Ho is true)
- \( \text{ci} = -0.0433 \text{ to } 0.9053 \) (confidence interval for the true difference of population means.)

Organic matter example...

```matlab
load('organicmatter_two.mat');

[n1,x1] = hist(corg1);
[n2,x2] = hist(corg2);

h1 = bar(x1,n1);
hold on
h2 = bar(x2,n2);

set(h1,'FaceColor','none','EdgeColor','r')
set(h2,'FaceColor','none','EdgeColor','b')
```

New dataset

```matlab
load('organicmatter_three.mat');
```
% two means are 25.4012   24.1963

\[ [H,P,ci] = \text{ttest2(corg1,corg2,0.05)} = \]
\[ H = 1 \text{ (reject the Null hypothesis that they are equal with 95% confidence)} \]
\[ P = 6.1138\times10^{-6} \]
\[ ci = [0.0711, 1.7086] \text{ (95% confidence interval for the true difference of population means.)} \]

% test

\[ [H,P,ci] = \text{ttest2(corg1,corg2,0.000005)} \]
\[ H = 0 \text{ (can NOT reject at this level)} \]

\[
\hat{F} = \frac{S_a^2}{S_b^2} \quad \text{Assuming var(a) > var(b)}
\]
\[ \text{d.f.}_a = \text{length(a)-1} \]
\[ \text{d.f.}_b = \text{length(b)-1} \]

Again, if F is bigger than some critical value (that depends on d.f) we reject the null hypothesis that the two variances are equal.

>> [H,P,CI] = \text{vartest2(corg1,corg2)}

\[ H = 0 \text{ (can NOT reject the null hypothesis that the variances are equal at 95% confidence level).} \]
\[ P = 0.7787 \text{ Probability that } H_0 \text{ is true} \]
\[ CI = [0.6429, 1.8018] \text{ (95% confidence interval for the true ratio var(X)/var(Y).)} \]
**Available Hypothesis Tests**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2sample</td>
<td>Mean equality test. Tests if two independent samples come from the same distribution, against the alternative that they come from different distributions.</td>
</tr>
<tr>
<td>chisq</td>
<td>Chi-square goodness-of-fit test. Tests if a sample comes from a specified distribution, against the alternative that it does not come from that distribution.</td>
</tr>
<tr>
<td>ranksum</td>
<td>Wilcoxon rank sum test. Tests if the null distribution of a linear regression is independent, against the alternative that there is auto-correlation among them.</td>
</tr>
<tr>
<td>jarque</td>
<td>Jarque-Bera test. Tests if a sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.</td>
</tr>
<tr>
<td>linreg</td>
<td>Linear regression test. Tests if the residuals from a linear regression are independent, against the alternative that there is auto-correlation among them.</td>
</tr>
<tr>
<td>scores</td>
<td>One-sample Kolmogorov-Smirnov test. Tests if a sample comes from a continuous distribution with specified parameters, against the alternative that it does not come from that distribution.</td>
</tr>
<tr>
<td>smear</td>
<td>Two-sample Kolmogorov-Smirnov test. Tests if two samples come from the same continuous distributions, against the alternative that they do not come from the same distribution.</td>
</tr>
<tr>
<td>ttest</td>
<td>T-test. Tests if a sample comes from a continuous distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.</td>
</tr>
<tr>
<td>ttest2</td>
<td>Two-sample t-test. Tests if two independent samples come from normal distributions with unknown but equal (or, optionally, unequal) variances and the same mean, against the alternative that the means are unequal.</td>
</tr>
<tr>
<td>varshared</td>
<td>One-sample chi-square variance test. Tests if a sample comes from a normal distribution with specified variance, against the alternative that it comes from a normal distribution with a different variance.</td>
</tr>
<tr>
<td>varshared2</td>
<td>Two-sample F-test for equal variances. Tests if two independent samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.</td>
</tr>
<tr>
<td>varshunt</td>
<td>Bartlett's multiple sample test for equal variances. Tests if multiple samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.</td>
</tr>
<tr>
<td>ztest</td>
<td>One-sample z-test. Tests if a sample comes from a normal distribution with unknown variance and specified mean, against the alternative that it does not have that mean.</td>
</tr>
</tbody>
</table>

---

**Bootstrap hypothesis testing (non-parametric alternative)**

- Example, test that the means of two datasets are the same.

```matlab
load('organicmatter_two.mat');
% let combine the two dataset together
pooled_corg = cat(2, corg1, corg2);
pooled_corg = randsample(pooled_corg, 120); % randomize them for good measure, without replacement

n = length(corg1); % both happen to be 60 points long here
k = 1000; % take 1000 bootstrap samples.
sa = zeros(k, n); % zeros.m allows to preallocate memory
sb = sa;
for ii = 1:k
    % do this k times
    % random samples with replacement (true == replacement)
sa(ii, :) = randsample(pooled_corg, true); % take n=60 random samples
    sb(ii, :) = randsample(pooled_corg, true); % take n=60 random samples
end

d = mean(sa, 2) - mean(sb, 2); % difference the means in each row
% make a relative frequency histogram
[nn, xx] = hist(d, 20); figure; bar(xx, nn/sum(nn));
```
\texttt{[nn, xx] = hist(d,20); figure; bar(xx,nn/\text{sum(nn))};)

\texttt{obs\_meandiff=mean(corg1)-mean(corg2);}
\texttt{hold on; plot([obs\_meandiff,obs\_meandiff],[0, \text{max(nn/\text{sum(nn))}]},'r')}

\textbf{How to get a p-value:}
\texttt{Nexceed=\text{length(find(d>obs\_meandiff))};} % number exceeding observed
\texttt{p=Nexceed/k; \% probability of exceeding}
\texttt{p = 0.039}

\textbf{Compare bootstrap results with earlier ttest2 results.}

\texttt{>> mean(corg1)}
\texttt{25.5816}
\texttt{>> mean(corg2)}
\texttt{25.1507}

\texttt{[h,p,ci] = ttest2(corg1,corg2,0.05)}

\texttt{h = 0} \hspace{1cm} \text{(can NOT reject the null hypothesis that samples come from a population with the same mean at 95\% confidence level)}

\texttt{p = 0.0745 \hspace{1cm} (7.45\% chance Ho is true)}

\texttt{ci = -0.0433 \hspace{0.5cm} 0.9053 \hspace{0.5cm} (confidence interval for the true difference of population means.)}