Estimation and Statistical Tests

Instructors

John Warren
Office of Environmental Information, EPA

Barry Nussbaum
Office of Environmental Information, EPA

Schedule

8:00	Introduction
8:15	Estimation using confidence intervals
8:45	Estimation (class exercise)
9:15	Decision making and decision errors
9:45	*Break*
10:00	The t-test & sign test
10:45	Making decisions (class exercise)
11:15	Difference of two means tests
12:00	*Lunch*

Course Objectives

At the end of this course, you should be able to:

- **Make a confidence interval for estimation**
- **Explain how a statistical test works.**
- **Have confidence to apply a statistical test.**

Data obtained from Farquar River

20 samples taken along the Peabody Bank of the Farquar River, the local environmental group would like to know the average potassium level. They want us to be pretty sure of the result and want the error in the answer to be + or – roughly 2ppm.

Using a spread sheet we find the average of the observations to be 45.69 with an error of +/- 1.67ppm.

Estimation: Isn't is obvious?

- **What kind of average? Mean, median, mode?**
- **Where did 1.67 come from?**
- **How "pretty sure" are you about this answer?**
- **Was this a random sample and does it matter?**

Results from a random sample

- There were 30 random samples taken over a 1/8th acre site and analyzed for lead concentrations
- The sample mean was 374.131 mg/Kg, the sample standard deviation was 131.501 mg/Kg
- It looks like the site is less than 400 mg/Kg and so may suitable for human habitation, but is this overwhelming evidence the mean is less than 400mg/Kg?
- Use a statistical test to determine if this is so.

Using a *t*-test

$$t = \frac{\left(\bar{x} - \mu_o\right)}{s/\sqrt{n}}$$

$\bar{x}$ = the sample mean

μ_o : the value of interest stated in the null hypothesis

n = the number of samples taken

s = the standard deviation of the sample

Entering the values

$$t = \frac{(\bar{x} - \mu_o)}{s/\sqrt{n}} = \frac{(374.131 - 400)}{131.501/\sqrt{30}} \approx -1.07$$

The t-test has n – 1 degrees of freedom = 29

The Rejection Rules

The rule for rejection of the null hypothesis depends on the hypotheses set up:

Null hypothesis: mean => 400
Alt: hypothesis: mean < 400

Reject the null if t < - t(tables)

t(tables) is often called the critical value

Decision

As the calculated t is not less than - t(tables), the null is accepted and so there is not overwhelming evidence that the site mean is less than 400 mg/Kg.

Hypothesis test: was that clear?

- **Where did the "null hypothesis" come from?**
- **Where did "t(tables) come from?**
- **How did you get a truly "random sample"?**
- **What is this "degrees of freedom"**
- **What do you mean "overwhelming evidence"**
- **If the average was 374.131 why isn't the site clean?**

Estimation using Intervals

Estimation and intervals

- Interval estimation involves finding a estimate from the data and combining it with a "+/-" that reflects our uncertainty in the numerical value.

- For example:

  The estimated mean level of potassium

  25.7ppm +/- 2.0 ppm

Confidence Interval for a Mean

- From a text book comes the following information:
- A 100(1 – α)% 2-sided C.I.: $\overline{X} \pm t_{n-1,\,1-\alpha/2} \cdot s/\sqrt{n}$
- A 100(1 – α)% upper C.I.: $\overline{X} + t_{n-1,\,1-\alpha} \cdot s/\sqrt{n}$
- A 100(1 – α)% lower C.I.: $\overline{X} - t_{n-1,\,1-\alpha} \cdot s/\sqrt{n}$

What do the Terms Mean?

$\overline{X}$ = the sample mean
n = the number of samples taken
s = the standard deviation of the sample
$t_{n-1,\,1-\alpha/2}$ = Student's 2-tailed t, and is obtained from tables
$t_{n-1,\,1-\alpha}$ = Student's 1-tailed t, and is obtained from tables

$n - 1$ is often called the degrees of freedom for a t-test
α is often called the level of significance of a test
α is sometimes called the false positive rate
α is occasionally called the Type I error rate

The upper end of a confidence interval is called the UCL
The lower end of a confidence interval is called the LCL

Confidence Interval: An Example

In a study of green-gill fish from the Pnobscott Lake, a sample of 8 fish caught during the day had a mean weight of 84.30g with a standard deviation of 11.06g. Construct a 95% confidence interval for the true (but unknown) mean weight of green-gill fish.

$$\bar{X} \pm t_{n-1,\,1-\alpha/2} \cdot s/\sqrt{n}$$

$\bar{X}$ = 84.30, s = 11.06, $t_{n-1,\,1-\alpha/2}$ = 2.365

Therefore the interval is 84.30 +/- 2.365(11.06)/$\sqrt{8}$

Which can be written as 75.05 to 93.55

What Does a Confidence Interval Mean?

- A 95% C.I. means that this technique will create an interval that has a 0.95 chance of "capturing" the true (but unknown) population mean
- You will not know if you definitely have captured the unknown mean, just the chance you have.
- A large sample (n) leads to a small C.I., and vice versa
- Small variability leads to a small C.I., and vice versa
- Less certainty (1 – α) leads to a small C.I., and vice versa

Confidence Interval for a Proportion

- A 100(1 – α)% 2-sided C.I.: $p \pm t_{n-1,1-\alpha/2} \cdot \sqrt{p(1-p)} \Big/ \sqrt{n}$
- A 100(1 – α)% upper C.I.: $p + t_{n-1,1-\alpha} \cdot \sqrt{p(1-p)} \Big/ \sqrt{n}$
- A 100(1 – α)% lower C.I.: $p - t_{n-1,1-\alpha} \cdot \sqrt{p(1-p)} \Big/ \sqrt{n}$

Notice they are similar to those for a mean except the mean and standard deviation have been replaced by "p" which is the sample proportion

Prediction Interval

- This concerns a single future event
- It assumes the same set of circumstances that were used to generate a confidence interval

$$\bar{X} \pm t_{n-1,1-\alpha/2} \cdot s(1+1/\sqrt{n})$$

- The prediction interval includes all the variability from the confidence interval plus an extra amount of variability as it is going into the future.

Prediction Interval: An Example

Is it likely to obtain a green-gill fish over 120g in weight when we resample the Lake? Using the same green-gill summary data, the prediction interval is:

$84.30 \pm 2.365(11.06)(1 + 1/\sqrt{8}) = 48.89 \text{ to } 119.71$

Note that the weight being asked for exceeds the upper prediction limit.
Conclude it is not likely to obtain such a heavy fish next year as it exceeds the upper prediction limit.

Tolerance Interval

- A tolerance interval specifies a region that contains a certain proportion of the population with some specified confidence
- For example: A 90% tolerance interval for 95% of the target population is 17.7 to 23.8
- This is interpreted as we are 90% sure that the interval 17.7 to 23.8 contains 95% of the target population
- The formula is complicated and almost never computed manually. It is included with many soft-ware statistics.
- They can be one or two sided.

Tolerance Interval: An Example

- For the green-gill fish in Pnobscott Lake, what would be the upper tolerance limit such that we could be 95% sure that 90% of the weights are below this value?
- The existing data is used to obtain the estimate 112.38
- We are 95% confident that 90% of the green-gill fish in this lake are less than 112.38g.

Conclusions

- Interval estimation has 3 types: Confidence Interval, Prediction Interval, and Tolerance Interval
- A confidence interval combines an estimate from the sample with precision and need for a degree of certainty
- Be careful when reading a confidence interval. Sometimes it is written as $\bar{X} \pm t_{n-1,\,1-\alpha/2} \cdot \text{s.e.}$ instead of $\bar{X} \pm t_{n-1,\,1-\alpha/2} \cdot s/\sqrt{n}$. This is because some writers prefer to express a confidence interval as "estimate +/- standard error of the estimate", the latter part being abbreviated to simply s.e. In most cases, "standard error" means "standard error of the mean" and this is the standard deviation (s) divided by the square root of sample size ($\sqrt{n}$).

Class Exercise – Estimation and Confidence Intervals

Participant's Exercise

- **Calculate a 95% 2-sided C.I. from a data set of 16 random samples taken from a potential Brownfield site. The samples were analyzed for Tri-chloro-benzolate by a properly certified laboratory operating under standard conditions.**
- **The data set is in the handout**

Analysis of 60 Confidence Intervals

- Consider the C.I.s from samples A through HHH
- Note that all the intervals look reasonable
- Suppose it is now revealed that the true mean (previously unknown) is 20.00 ppb
- Which samples led to the wrong conclusion?

So What Does This Imply?

- Samples C, Q, CCC all miss the true value 20.00
- Out of 60 samples, 3 missed, 57 were correct
- 57/60 = 95% which (by coincidence!) matches the intended 95% part of the Confidence Interval statement
- When you take a sample and construct the Confidence Interval, you don't know whether you will be the one that will miss, only a probability you have (i.e. 5%)

Decision Making and Decision Errors

Systematic planning is an Agency requirement

- Description of project goal, objectives, and schedule
- Identification of type of data and the link to project goal
- Decision on type, quality, and quantity of data needed
- Specification of acceptance or performance criteria
- Development of a sampling plan and QA/QC requirements
- Identification of sponsoring organization and personnel
- Preliminary description of how the data will be analyzed

and this is implemented through the DQO Process

Seven Steps of the DQO Process

The Data Quality Objectives Process is the Agency's recommended method for systematic planning:

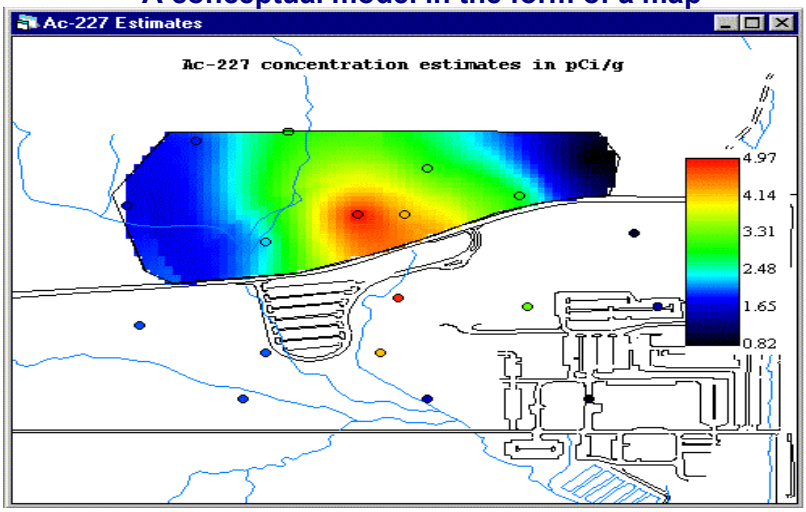
1. State the problem
2. Identify the goal of the study
3. Identify information inputs
4. Define the boundaries of the study.
5. Develop the analytical approach
6. Specify performance or acceptance criteria
7. Optimize the plan for obtaining data

What do you want to achieve with the DQO Process?

- A good conceptual model of the problem
- A firm basis for taking action
- Meaningful documentation of potential decision errors

A Conceptual Model is essential

A conceptual model in the form of a map



Step 6 of the DQO Process: Specify performance or acceptance criteria

- Define an "if/then" statement that incorporates the action level and the alternatives. *For example: If the mean level of contaminate exceeds the action level, then proceed to clean up*

- Turn the "if/then" statement into a test of hypothesis. This test uses the parameter of interest, the action level, and a baseline assumption determined by the regulation or decision maker.

- Baseline assumption: Null Hypothesis
  Alternative assumption: Alternative Hypothesis

- The Null Hypothesis is assumed to be true unless over-whelming evidence (data) indicates it must be false.

After data has been collected: the structure of a hypothesis test

- In general, a hypothesis test has 5 steps:
  - Statement of the null hypothesis
  - Statement of the alternative hypothesis
  - Calculation of the test statistic (based on the collected data)
  - Calculation of the critical value
  - Hypothesis test conclusion

Statement of Hypotheses

- Null hypothesis – statement of current beliefs. For example:
  - The average copper concentration is at most 400 ppm.
  - Site concentrations are not elevated above background.
- Alternative hypothesis – statement of interest for the project. For example:
  - The average copper concentration is more than 400 ppm.
  - The site concentrations are elevated above background.

Decision errors and plain English

Baseline assumption: the program is in compliance

False Rejection Decision Error
(False Positive, F(+), Type I Error)

- **Deciding program is not in compliance when it is.**
- **An overreaction to a situation.**
- **Wasted resources, unnecessary expenditure.**

False Acceptance decision Error
(False Negative, F(-), Type II Error)

- **Deciding program is in compliance when it is not.**
- **A missed opportunity for correction.**
- **Allowing a hazard to public health or the ecosystem.**

Decision False Rejection/Acceptance

Null Hypothesis: True mean equal or below standard

Alternative: True mean level above standard

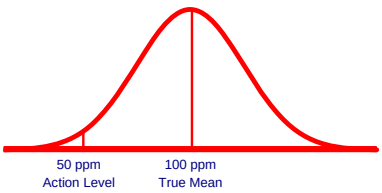
		In Actuality	
		Below Standard	***Above Standard***
Decision based on a small sample	***Below Standard***	Correct	F(a) β
	Above Standard	F(r) α	Correct

Specifying limits on decision errors

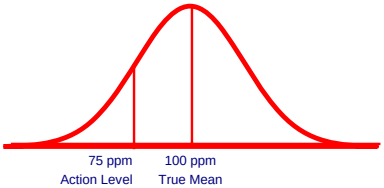
- Determine consequences of each decision error:
  - Health risks
  - Ecological risks
  - Political risks
  - Social risks
  - Resource risks
- Set quantitative limits on false rejection and false acceptance errors by considering the consequences of these potential decision errors.
- The smaller the tolerance of such risks, the larger the number of samples needed to meet these tolerances.

Magnitude of decision errors

Chance of making potential decision errors depends how close the true mean is to the action level



If the true mean is much greater than the action level, few low readings will occur. Therefore only a small chance of reaching a wrong conclusion.



If the true mean is close to the action level, many low readings will occur. Therefore erroneous conclusions are much more likely.

Calculation of the Test Statistic

- The test statistic is a measure of the difference between the observed result and the expected result (as defined by the null hypothesis).
- In general, the test statistic is
  - Parametric test: a standardized value of the sample version of the parameter of interest.
  - Nonparametric test: a standardized value of a function of the ranks.
- If the observed result is very different from the expected result, then the test statistic will be large (in absolute value), which will lead to a rejection of the null hypothesis.

Calculation of the Critical Value and the Hypothesis Test Conclusion

- The critical value is calculated from the probability distribution of the test statistic assuming the null hypothesis is true.
- Using the false rejection error rate, the critical value defines a point which if the observed test statistic is more extreme than that point, then we reject the null hypothesis.

Hypothesis Test Error Rates and Power

- Part of the DQO process is to set two statistical error rates:
  - false rejection error rate: The chance of rejecting the null hypothesis when it is true.
  - false acceptance error rate: The chance of accepting the null hypothesis when it is false.
- Power is the probability of rejecting the null hypothesis if it is actually false and is equal to one minus the false acceptance error rate.

Necessary Sample Size

- If the we accept the null hypothesis, then it is important to determine if the sample size was adequate to achieve the desired power or false acceptance error rate.
- Several tests have simple formulae for determining adequate sample size, while in other instances more complex methods are required.

Parametric Versus Nonparametric Hypothesis Tests

- Parametric tests make stronger assumptions about the underlying population distribution and have more power than a nonparametric equivalent.
- Parametric tests generally use the data values to compute test statistics where nonparametric tests use data ranks.
- Determining which specific test to use depends not only upon project goals, but also upon the underlying population distribution.

Conclusions

- Statistical hypothesis testing provides a means of quantitative decision making for DQA.
- The choice of hypothesis test is dependent upon project goals as well as the underlying distribution of interest.
- While hypothesis tests provide a decision making mechanism, the testing conclusion should be used in conjunction with the science behind the problem to draw an appropriate conclusion.

The t-test and Sign Test

Steps of a Hypothesis Test

- In general, a hypothesis test has 5 steps:
  - Statement of the null hypothesis
  - Statement of the alternative hypothesis
  - Calculation of the test statistic (based on the collected data)
  - Calculation of the critical value
  - Hypothesis test conclusion

Parametric Versus Nonparametric Hypothesis Tests

- Parametric tests make stronger assumptions about the underlying population distribution and have more power than a nonparametric equivalent.
- Parametric tests generally use the data values to compute test statistics where nonparametric tests use data ranks.
- Determining which specific test to use depends not only upon project goals, but also upon the underlying population distribution.

One-Sample t-test: Description

- **Purpose**: A parametric test for a difference between a population mean and a fixed threshold.
- **Data**: A simple or systematic random sample, $x_1,\ldots,x_n$, from the population of interest. The sample may contain composite data but can't be a mixture of both.
- **Assumptions**: The data are independently acquired and come from an approximately normal distribution.
- **Limitations and Robustness**: This test is robust against the population distribution deviating moderately from normality. However, it is not robust against outliers and has difficulty handling non-detects.

One-Sample t-test: Instructions

- **STEP 1 – Null Hypothesis:** $H_0 : \mu = C$

  μ is the population mean
  C is the threshold

- **STEP 2 – Alternative Hypothesis:**

  i) $H_A : \mu > C$ (upper - tail test)
  ii) $H_A : \mu < C$ (lower - tail test)
  iii) $H_A : \mu \neq C$ (two - tail test)

- **STEP 3 – Test Statistic:**

$$t = \frac{\overline{X} - C}{s/\sqrt{n}}$$

$\overline{X}$ is the sample mean
s is the sample std dev
n is the sample size

One-Sample t-test: Instructions (cont.)

- **STEP 4 – Critical Value: Use a Student's-*t* table or statistical software to find:**

  i) $t_{n-1,\,1-\alpha}$
  ii) $-t_{n-1,\,1-\alpha}$
  iii) $t_{n-1,\,1-\alpha/2}$

- **STEP 5 – Conclusion:**

  i) If $t > t_{n-1,\,1-\alpha}$, then reject H_0.
  ii) If $t < -t_{n-1,\,1-\alpha}$, then reject H_0.
  iii) If $|t| > t_{n-1,\,1-\alpha/2}$, then reject H_0.

One-Sample t-test: Background

- We wish to determine if there is evidence that the mean lead concentration for a particular site is greater than 100 ppm.
- The decision maker has specified a 0.05 false rejection error rate (α), and a 0.20 false acceptance error rate (β) at 110 ppm (μ_1).
- A random sample of 9 locations yields (in ppm):

87.3	108.4	109.9	109.5	103.3
118.2	91.6	96.7	113.2	

One-Sample t-test: First 3 Steps

- STEP 1 – Null Hypothesis: $H_0: \mu \leq 100$
- STEP 2 – Alternative Hypothesis: $H_A: \mu > 100$
- STEP 3 – Test Statistic:

$$t = \frac{\overline{X} - C}{s/\sqrt{n}} = \frac{104.2 - 100}{10.3/\sqrt{9}} = 1.22$$

One-Sample t-test: Next 2 Steps

- **STEP 4 – Critical Value: Use a Student's-*t* table or statistical software to find:**

$$t_{n-1,\,1-\alpha} = t_{8,\,0.95} = 1.86$$

- **STEP 5 – Conclusion: Since the test statistic (t = 1.22) is less than the critical value ($t_{n-1,\,1-\alpha}$ = 1.86), we accept the null hypothesis that the mean lead concentration is at most 100 ppm.**

One-Sample t-test: Final Step

- **Required Sample Size: Since the null hypothesis was accepted, it is necessary to determine if the sample size is large enough to satisfy the false acceptance error rate. Since,**

$$n \geq \frac{s^2\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{\left(\mu_1-\mathrm{C}\right)^2}+\frac{z_{1-\alpha}^2}{2}=\frac{10.3^2(1.645+0.842)^2}{(100-110)^2}+\frac{1.645^2}{2}=7.9$$

the formula indicates 8 samples would have been necessary and we actually had a sample size 9, it would seem that the false acceptance error rate has been satisfied.

One-Sample t-test: Comments

- **Was the test straight-forward?** Yes, we can do the mechanical procedure with little problem – in fact this is often done automatically by a canned program.
- **Was it obvious where α and β came from?** In this case they were stated – very rare in practice. You have to choose these yourself as there really are no "default" values. Always state what you choose to avoid confusion.
- **Were the assumptions obvious?** No. The t-test is often performed with no regard as to whether it is appropriate or not. The assumptions of independence of data, approximate normality, and integrity of data are extremely important.

One-Sample Sign Test: Description

- **Purpose**: A nonparametric test for a difference between the population median and a fixed threshold.
- **Data**: A simple or systematic random sample, $x_1,\ldots,x_n$, from the population of interest. The sample may or may not contain compositing.
- **Assumptions**: The Sign test can be used for any underlying population distribution.
- **Limitations and Robustness**: The Sign test has less power than the one-sample *t*-test or the Wilcoxon Signed Rank test, but makes no distributional assumptions and can handle non-detects if the detection limit is below the threshold.

One-Sample Sign Test: Instructions

- **STEP 1 – Null Hypothesis:** $H_0 : \text{median} = C$

- **STEP 2 – Alternative Hypothesis:**

  i) $H_A : \text{median} > C$ (upper-tail test)

  ii) $H_A : \text{median} < C$ (lower-tail test)

  iii) $H_A : \text{median} \neq C$ (two-tail test)

- **STEP 3 – Test Statistic:**

  $B = \#$ of data points greater than the threshold, C.

One-Sample Sign Test: Instructions (cont.)

- **STEP 4 – Critical Value: Use the Critical Values for the Sign Test table to find:**

  i) $B_{\text{upper}}(n, 2\alpha)$

  ii) $B_{\text{lower}}(n, 2\alpha) - 1$

  iii) $B_{\text{lower}}(n, \alpha) - 1$ and $B_{\text{upper}}(n, \alpha)$

- **STEP 5 – Conclusion:**

  i) If $B \geq B_{\text{upper}}(n, 2\alpha)$, then reject H_0.

  ii) If $B \leq B_{\text{lower}}(n, 2\alpha) - 1$, then reject H_0.

  iii) If $B \geq B_{\text{upper}}(n, \alpha)$ or $B \leq B_{\text{lower}}(n, \alpha) - 1$, then reject H_0.

One-Sample Sign Test: Background

- **After site remediation, it is desired to determine if the median arsenic concentration meets the 10 ppm threshold. Random samples are provided in the table below.**

Arsenic	9.7	10.4	10.9	8.9	5.7	11.6	3.3	5.2	7.9	<1

- **The decision maker has specified an α = 0.10 (false rejection error rate).**

One-Sample Sign Test: First 3 Steps

- **STEP 1 – Null Hypothesis:**

$$H_0: \text{median} \geq 10$$

- **STEP 2 – Alternative Hypothesis:**

$$H_A: \text{median} < 10$$

- **STEP 3 – Test Statistic:**

$$B = \#\text{ of data points greater than the threshold} = 3$$

One-Sample Sign Test: Next 2 Steps

- **STEP 4 – Critical Value: With a sample size of 10 and a false rejection error rate of 0.10, the Critical Values for the Sign Test table gives:**

$$B_{\text{lower}}(n, 2\alpha) - 1 = B_{\text{lower}}(10, 0.20) - 1 = 2$$

- **STEP 5 – Conclusion: Since the test statistic (B = 3) is greater than the critical value, we accept the null hypothesis that the median arsenic concentration is at least 10 ppm. Therefore, site remediation did not meet the threshold.**

One-Sample Sign Test: Comments

- **Was the test straight-forward? Yes, all we did was count the number that exceeded the hypothesized median. The notation was a little strange however but gives the procedure a certain mystique.**
- **Was it obvious where α and β came from? In this case α was stated but not β. In fact when using a sign test it is rare that β is ever mentioned. You usually have to choose the α yourself as there really are no "default" values. Again, always state what you choose to avoid confusion.**
- **Were the assumptions obvious? No. The sign test usually works with medians but can also be used for means. In the later case the t-test is usually a better choice as it is more powerful for a symmetric distribution of data.**

Conclusions

- Statistical hypothesis testing provides an easy and defensible way of making quantitative decision making.
- The choice of hypothesis test is dependent upon project goals as well as the underlying distribution of interest: the t-test needs approximate normality of data.
- If the data are approximately normal, the t-test is much more powerful than the Sign test. If the underlying distribution of data is far from normal, the Sign test is better than the t-test.
- There are other nonparametric tests such as Wilcoxon signed rank that are more powerful than the Sign test.

Class Exercise – Decision Making

Participant's Exercise

- **Determine if the median level of Tri-chloro-benzolate is definitely greater than 18 ppb using a data set of 16 random samples taken from a potential Brownfield site. The samples were analyzed for Tri-chloro-benzolate by a properly certified laboratory operating under standard operating conditions.**
- **The data set is in the handout**

Analysis of 60 Decisions

- Consider the decisions from samples A through HHH
- Note that all the medians look reasonable
- Notice that samples I, Q, S, MM, AAA, and CCC were all significant at the 10% level.
- Suppose it is now revealed that the true median (previously unknown) is actually 18.0 ppb
- What has happened here?

Discussion

- Out of 60 samples, 37 had medians above 18.0 (and 23 were lower)
- 6 samples showed significance (10% of the 60 samples)
- You only have one sample to consider and 10% level of significance means that, on the average, you run a 0.10 chance of saying there is a significant increase when in fact there is not.

The Two Sample t-test and Wilcoxon Test

Steps of a Hypothesis Test

- In general, a hypothesis test has 5 steps:
  - Statement of the null hypothesis
  - Statement of the alternative hypothesis
  - Calculation of the test statistic (based on the collected data)
  - Calculation of the critical value
  - Hypothesis test conclusion

Statement of Hypotheses

- Null hypothesis – statement of current beliefs. For example:
  - The average copper concentration is at most 400 ppm.
  - Site concentrations are not elevated above background.
- Alternative hypothesis – statement of interest for the project. For example:
  - The average copper concentration is more than 400 ppm.
  - The site concentrations are elevated above background.

Calculation of the Test Statistic

- The test statistic is a measure of the difference between the observed result and the expected result (as defined by the null hypothesis).
- In general, the test statistic is
  - Parametric test: a standardized value of the sample version of the parameter of interest.
  - Nonparametric test: a standardized value of a function of the ranks.
- If the observed result is very different from the expected result, then the test statistic will be large (in absolute value), which will lead to a rejection of the null hypothesis.

Calculation of the Critical Value and the Hypothesis Test Conclusion

- The critical value is calculated from the probability distribution of the test statistic assuming the null hypothesis is true.
- Using the false rejection error rate, the critical value defines a point which if the observed test statistic is more extreme than that point, then we reject the null hypothesis.

Hypothesis Test Error Rates and Power

- Part of the DQO process is to set two statistical error rates:
  - false rejection error rate: The chance of rejecting the null hypothesis when it is true.
  - false acceptance error rate: The chance of accepting the null hypothesis when it is false.
- Power is the probability of rejecting the null hypothesis if it is actually false and is equal to one minus the false acceptance error rate.

Necessary Sample Size

- If the we accept the null hypothesis, then it is important to determine if the sample size was adequate to achieve the desired power or false acceptance error rate.
- Several tests have simple formulae for determining adequate sample size, while in other instances more complex methods are required.

Parametric Versus Nonparametric Hypothesis Tests

- Parametric tests make stronger assumptions about the underlying population distribution and have more power than a nonparametric equivalent.
  t-test assumptions versus Sign test assumptions
- Parametric tests use the data values to compute test statistics where nonparametric tests use data ranks.
  t-test used actual values, Sign test used simple numbers
- Determining which specific test to use depends on project goal and also the underlying population distribution.
  Mean versus Median, assumption of independence

Two-Sample t-test: Description

- **Purpose**: A parametric test for a difference between two population means when the population variances are assumed to be equal.
- **Data**: Independent, Simple or systematic random samples from two populations – $x_1, x_2, \ldots, x_m$, and $y_1, y_2, \ldots, y_n$.
- **Assumptions**: The populations are independent and approximately normally distributed with approximately equal variances.
- **Limitations and Robustness**: The two-sample t-test is robust against moderate violations of the normality assumption. However, the test is not robust against outliers.

Two-Sample t-test: Instructions

- **STEP 1 – Null Hypothesis**: $H_0 : \mu_X - \mu_Y = \delta_0$
- **STEP 2 – Alternative Hypothesis**:

$i)$ $H_A : \mu_X - \mu_Y > \delta_0$ (upper - tail test)

$ii)$ $H_A : \mu_X - \mu_Y < \delta_0$ (lower - tail test)

$iii)$ $H_A : \mu_X - \mu_Y \neq \delta_0$ (two - tail test)

- **STEP 3 – Test Statistic**:

$$t = \frac{(\bar{X} - \bar{Y}) - \delta_0}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \qquad \text{where } s_p^2 = \frac{(m-1)\cdot s_x^2 + (n-1)\cdot s_y^2}{m+n-2}$$

Two-Sample t-test: Instructions (cont.)

- **STEP 4 – Critical Value: Use a Student's-*t* table or statistical software to find:**

$$i) \quad t_{m+n-2,\,1-\alpha}$$
$$ii) \quad -t_{m+n-2,\,1-\alpha}$$
$$iii) \quad t_{m+n-2,\,1-\alpha/2}$$

- **STEP 5 – Conclusion:**

$i)$ If $t > t_{m+n-2,\,1-\alpha}$, then reject H_0.

$ii)$ If $t < -t_{m+n-2,\,1-\alpha}$, then reject H_0.

$iii)$ If $|t| > t_{m+n-2,\,1-\alpha/2}$, then reject H_0.

Two-Sample t-test: Example 1

- **At a hazardous waste site, cadmium concentrations from an impacted area were compared to those from a reference area.**
- **Test if the average concentration of the impacted area is elevated above reference.**
- **The false rejection error rate was set at 5%.**

Impacted	10.2	9.1	10.4	9.5	7.7	8.5	7.4	
Reference	3.0	4.4	5.8	1.3	5.7	7.1	7.6	7.1

Two-Sample t-test: First 3 Steps

- **STEP 1 – Null Hypothesis:** $H_0: \mu_1 - \mu_2 \leq 0$
- **STEP 2 – Alternative Hypothesis:** $H_A: \mu_1 - \mu_2 > 0$
- **STEP 3 – Test Statistic:**

$$s_p = \sqrt{\frac{6 \cdot 1.36 + 7 \cdot 4.89}{7 + 8 - 2}} = 1.77$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{8.97 - 5.25}{1.77 \cdot \sqrt{\frac{1}{7} + \frac{1}{8}}} = 4.06$$

Two-Sample t-test: Next 2 steps

- **STEP 4 – Critical Value: Use a Student's-*t* table or statistical software to find:**

$$t_{m+n-2,\ 1-\alpha} = t_{13,\ 0.95} = 1.77$$

- **STEP 5 – Conclusion: Since the test statistic (t = 4.06) is greater than the critical value ($t_{m+n-2,\ 1-\alpha}$ = 1.77), we reject the null hypothesis that the mean concentrations of the reference and impacted areas are equal.**

Two-Sample t-test: Comments

- **Was the test straight-forward?** Yes, we can do the mechanical procedure with little problem. Only difficulty was the increase in greek letters.
- **Was it obvious where α and β came from?** In this case the α value was stated (false rejection was set at 5%) but β was not. In many cases this is because canned software can compute a power curve for potential alternatives (i.e. multiple β values). Always state what α-level you choose to avoid confusion.
- **Were the assumptions obvious?** No. The t-test is often performed with no regard as to whether it is appropriate or not. The assumptions of independence of data, approximate normality, and integrity of data are extremely important.

Wilcoxon Rank Sum Test: Description

- **Purpose**: A nonparametric test for a difference between two population means.
- **Data**: Independent, Simple or systematic random samples from two populations – $x_1, x_2, \ldots, x_m$, and $y_1, y_2, \ldots, y_n$.
- **Assumptions**: The two underlying distributions are assumed to have approximately the same shape (variance).
- **Limitations and Robustness**: The test may produce misleading results if there are many tied data values. If non-detects are present, then a statistician should be consulted on the potential use of Gehan ranking.

Wilcoxon Rank Sum Test: Instructions

- **STEP 1 – Null Hypothesis:** $\mathrm{H}_0 : \mu_X - \mu_Y = 0$

- **STEP 2 – Alternative Hypothesis:**

  i) $\mathrm{H}_\mathrm{A} : \mu_X - \mu_Y > 0$ (upper - tail test)
  ii) $\mathrm{H}_\mathrm{A} : \mu_X - \mu_Y < 0$ (lower - tail test)
  iii) $\mathrm{H}_\mathrm{A} : \mu_X - \mu_Y \neq 0$ (two - tail test)

- **Computations: Rank the pooled data from smallest to largest assigning average rank to ties. Sum the ranks of the first population and denote this by R_1.**

Wilcoxon Rank Sum Test: Instructions (cont.)

- **STEP 3 – Test Statistic:** $W = R_1 - \dfrac{m(m+1)}{2}$

- **STEP 4 – Critical Value: Use the Critical Values for the Wilcoxon Rank Sum Test table to find:**

  i) $mn - w_\alpha$
  ii) w_α
  iii) $mn - w_{\alpha/2}$ and $w_{\alpha/2}$

- **STEP 5 – Conclusion:**

  i) If $W \geq mn - w_\alpha$, then reject H_0.
  ii) If $W \leq w_\alpha$, then reject H_0.
  iii) If $W \geq mn - w_{\alpha/2}$ or $W \leq w_{\alpha/2}$, then reject H_0.

Wilcoxon Rank Sum Test: Example 2

- At a hazardous waste site, chromium concentrations from an impacted area were compared to those from a reference area.
- Test if the average concentration of the impacted area is elevated above reference.
- The false rejection error rate was set at 10%.

Impacted	17	23	26	5	13	13	12	
Reference	16	20	5	4	8	10	7	3

Wilcoxon Rank Sum Test: First 2 Steps

- STEP 1 – Null Hypothesis: $H_0: \mu_1 - \mu_2 \leq 0$
- STEP 2 – Alternative Hypothesis: $H_A: \mu_1 - \mu_2 > 0$
- Computations: The ordered pooled data and their ranks are (impacted area denoted by *):

Pooled Data	3	4	5*	5	7	8	10	12*	13*	13*	16	17*	20	23*	26*
Rank	1	2	3.5*	3.5	5	6	7	8*	9.5*	9.5*	11	12*	13	14*	15*

The sum of the ranks of the impacted area is:

$$R_1 = 3.5 + 8 + 9.5 + 9.5 + 12 + 14 + 15 = 71.5$$

Wilcoxon Rank Sum Test: Next 3 Steps

- **STEP 3 – Test Statistic:**

$$W = R_1 - \frac{m(m+1)}{2} = 71.5 - \frac{7\cdot(7+1)}{2} = 43.5$$

- **STEP 4 – Critical Value: From the Critical Values for the Wilcoxon Rank Sum Test table,**

$$mn - w_{0.10} = 7\cdot 8 - 16 = 40$$

- **STEP 5 – Conclusion: Since the test statistic (*W* = 43.5) is greater than the critical value, we reject the null hypothesis that the mean concentrations of the reference and impacted areas are equal.**

Wilcoxon Rank Sum Test: Comments

- **Was the test straight-forward? Yes, all we did was combine the data, rank from largest to smallest (or smallest to largest, it doesn't matter) add the ranks for the first population and compare with tables.**
- **Was it obvious where α and β came from? In this case α was stated but not β. In fact when using a sign test it is rare that β is ever mentioned. You usually have to choose the α yourself as there really are no "default" values. Again, always state what you choose to avoid confusion.**
- **Were the assumptions obvious? No but the Wilcoxon test works with both means and medians. If there is a large sample from each population a normal-based approximation can be used.**

Conclusions

- **Statistical hypothesis testing provides an easy and defensible way of making quantitative decision making.**
- **If the data are approximately normal, the Two-Sample t-test is more powerful than the Wilcoxon test. As many applications involve the comparison to some background and non-normality is very common, the Wilcoxon test is probably better than the t-test.**
- **Further advice on statistical tests is to be found in *Data Quality Assessment: Statistical Tools for Practitioners (EPA QA/G-9S)***

Conclusions

Statistical software

- Many different types available, all are good
- Freeware to be found on the Quality Staff site:
  www.epa.gov/quality/qa_links.html
- Some Agency sponsored:
  ProUCL: www.epa.gov/esd/tsc/software.htm
  GiSdT: www.gisdt.org
- Easy to use website: www.statpages.org

Was that clear? A bit better!

- **Where did the "null hypothesis" come from?**
  Consideration of why a test was needed
- **Where did "t(tables) come from?**
  Special tables giving critical values for decisions
- **How did you get a truly "random sample"?**
  Through careful use of systematic planning
- **What is this "degrees of freedom"**
  t-tests have this; n-1 for 1-sample, n+m-2 for 2-sample
- **What do you mean "overwhelming evidence"**
  Probability of occurrence less than alpha
- **If the average was 374.131 why isn't the site clean?**
  It's not so small that it could not have happened by chance