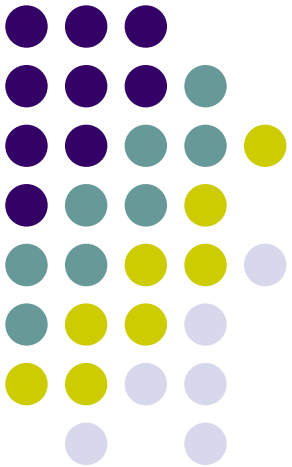
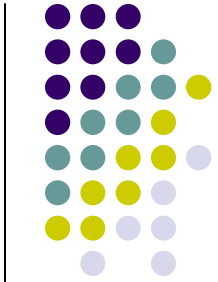
Statistics for Biologists
a brief review…

Herb E. Schellhorn

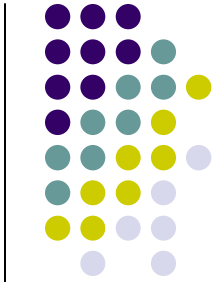


Statistics



- Why do we need statistics?
- What concepts are important?
- How are statistics commonly misused?

Outline



- History
- Types of Statistics
- Simple Statistics
- T-tests
- Transformations
- Assumptions
- Statistics and scientific reasoning
- Statistics and Excel-practical considerations

Non-Statistical arguments used to support conclusions in Science…

- "I ***think*** there is a difference"
- "I ***feel*** that there is a difference"
- "I ***really believe*** that my data is telling me there is a difference"

A Short History of Statistics

Statistics and Beer

- Beer – invented in Egypt and China over 5000 years ago
- Modern statistics was invented 100 years ago…

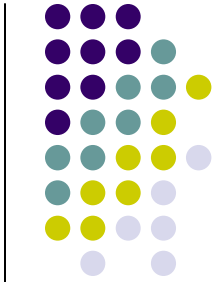
…**in a brewery!**

A Short History of Statistics

The Beer Story….

- Modern hypothesis-driven statistics was developed by William Gosset to monitor beer quality at Guinness breweries (~1908)
- Used statistical tests to measure production lot quality by assaying small samples.
- Not surprising that use of statistics is misunderstood...many of the giants of statistics did not initially appreciate Gosset's work (e.g. Karl Pearson)
- From Gosset's early work, R.A. Fisher developed the modern t-test…in the 1930's

Types of Statistics



- Parametric
  - Random variable is assumed to have an underlying distribution
  - e.g. weights in a given population is *normally* distributed
- Non-Parametric
  - Random variable does not have an assumed distribution
  - e.g. frequency values used in genotype analysis do not have an assumed distribution.
  - X^2 tests are an example of a non-parametric test.

Some things you should now understand and be able to apply…

- Probability Analysis-particularly Bionomial
- t-tests (various types)
- Linear Regression

Statistics-an overlooked measure –Co-efficient of variation (CV)

- Co-efficient of variation.
  - Is the standard deviation divided by the mean.
  - Good measure of the quality of a dataset

  - e.g. Pipetting errors
  - Use a pipettor to liquot 5 x 100 ul samples

$$\sqrt{\frac{n\sum x^2-\left(\sum x\right)^2}{n(n-1)}}$$

Coefficient of variation

Actual values

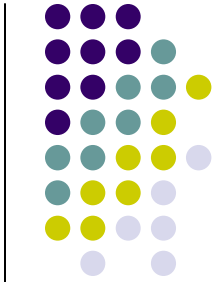
98.1
99.2
100.3
100.2
100.0

Mean (x) = (98.1+99.2+100.3+100.2+100.0)/5
= **99.56 ul**

Standard Deviation (SD) = **0.923 ul**

Coefficient of Variation = 0.923/99.56
= 0.00928
= 0.93%

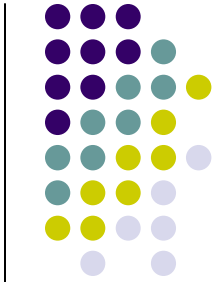
T-tests



- Paired and unpaired
- Equal sample variances, unequal sample variances

- Paired comparison with equal variance is the most powerful
- Unpaired, unequal variance is the least powerful

Transformations…



- Data transformations are used to restore “assumption” of equal variance in a a given test.

Log transformation of Microarray data

- Many sample means compared in a microarray will have very unequal variances.

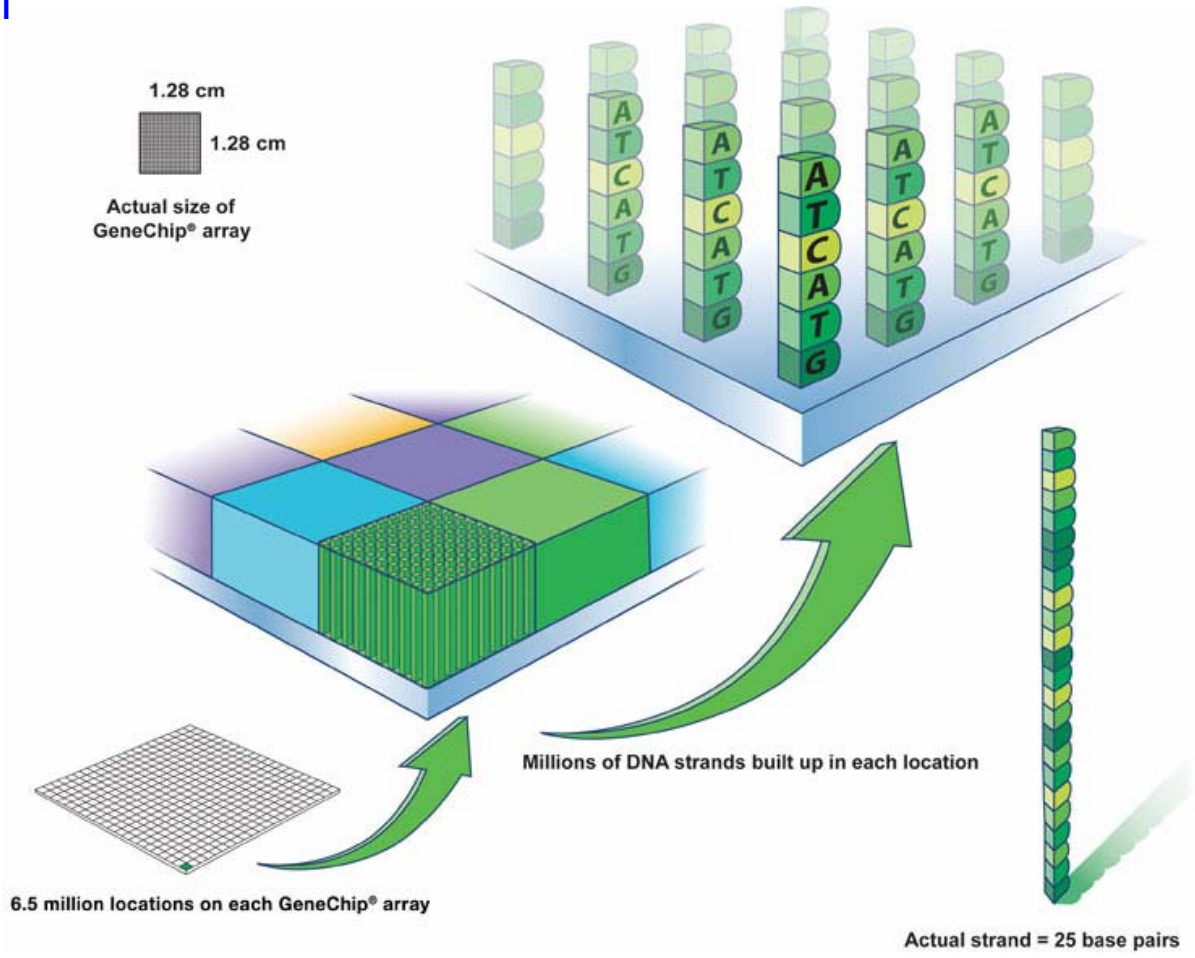
- This can happen quite often in analysis on microarray data

Microarrays

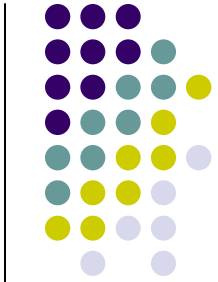
- Contain thousands of oligonucleotides (probes) that hybridize to sample RNA (or DNA)
- Usually highly redundant (e.g. Affymetrix E. coli chip has 300,000 probes for ~5000 genes

Microarrays

Cartoon depicting hybridization of tagged probes to Affymetrix GeneChip® microarray. Image courtesy of Affymetrix.



Microarrays-Analysis



Two basic levels of analysis

Probe level

- quality control,
- basic normalization,
- used to calculate gene responses

Gene level

- Used to compare expression levels between treatments

Most use gene level comparisons

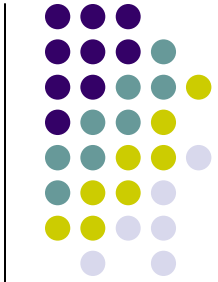
t-tests-Review

- t-tests are used to determine significance of an observed difference between two means
- there are several types of t-tests…
  - **Paired**
  - **Unpaired**
    - Equal variances
    - Unequal variances

    The "power" of these tests vary alot because of differences in assumptions we make prior to doing the test.

Making valid assumptions increases the power of tests.

t-tests



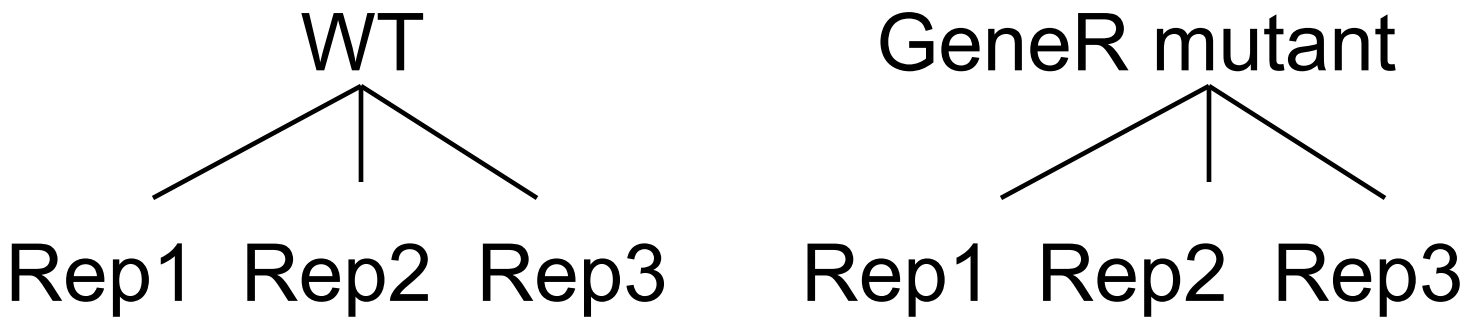
Some possible assumptions…

(1) **Homoscadasticity**-equality of standard errors (or variance)

(2) Normality of distribution of variables

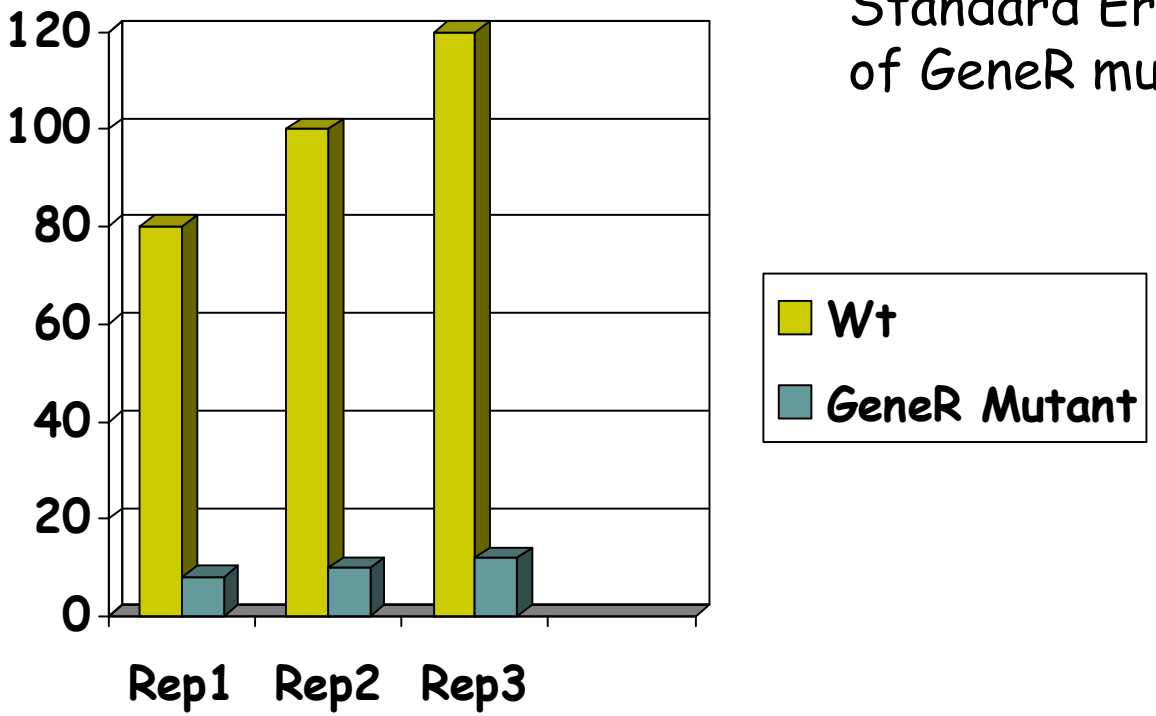
(3) Variables are dependent on one another (paired).

T-Tests and Microarrays



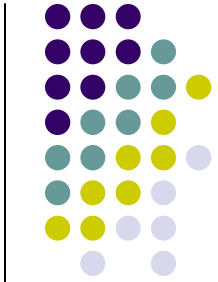
Can Plot results for GeneX…

Microarrays-Transformation



Standard Error of Wt is 10x that of GeneR mutant->unequal variance

Microarrays-Transformation



T-tests based on means with equal variance are much more powerful than those in which variance is considered unequal.

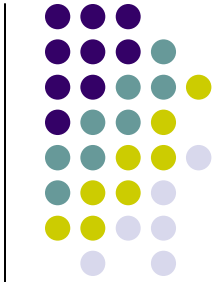
Even though the variance in microarray data is not equal there is a pattern→ the error is proportional to the response of the variable.

We can use this to log transform the data.

Microarray response (X) – not normally distributed –> **Weak T-test**

Log (X) – normally distributed –> **Robust T-Test**

Microarrays



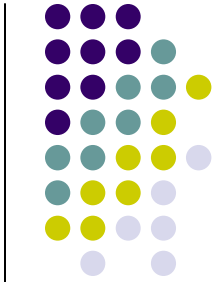
Common Example

- want to compare the expression of GeneX in the presence and absence of a suspected regulator (GeneR)
- How?

Make a GeneR mutant

Compare three independent replicates of wt vs three replicates of a GeneR mutant.

Microarrays



Common Example

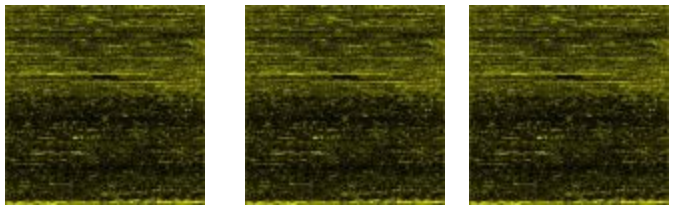
- want to compare the expression of GeneX in the presence and absence of a suspected regulator (GeneR)
- How?

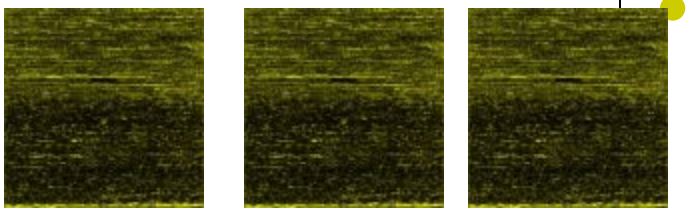
Make a GeneR mutant

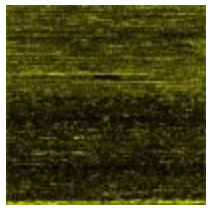
Compare three independent replicates of wt vs three replicates of a GeneR mutant.

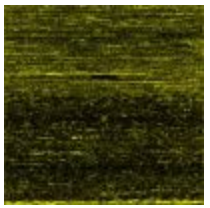
RpoS+

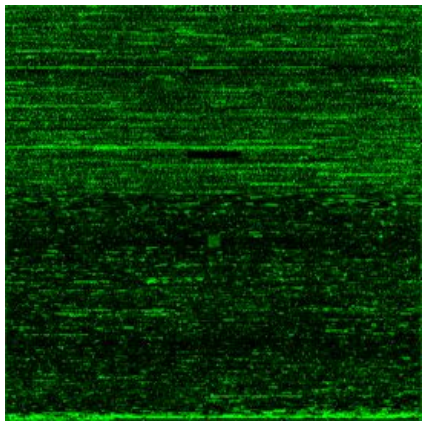
RpoS-

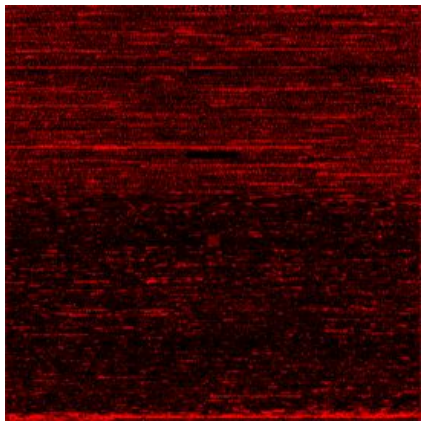




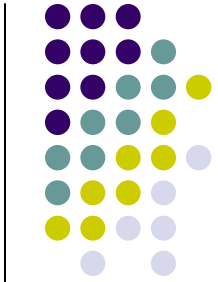








Microarrays



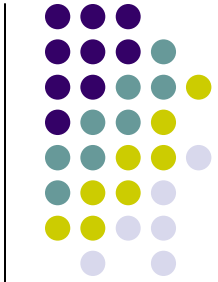
Common Example

- want to compare the expression of GeneX in the presence and absence of a suspected regulator (GeneR)
- How?

Make a GeneR mutant

Compare three independent replicates of wt vs three replicates of the GeneR mutant.

Microarrays-Normalization



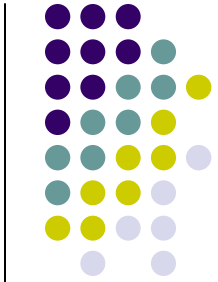
Why Normalize?

- Remove systemic differences in response (e.g. high background levels on one chip)
- Adjust for differing efficiencies in hybridization, cDNA labelling
- Can standardize against known reference gene set (e.g. a set of genes known to be invariant).
- Many other possibilities..

Considerations

- Best to decide on method in advance of analysis to eliminate experimenter bias

Microarrays-Steps in analysis



Normalize data

↓

Log transform data

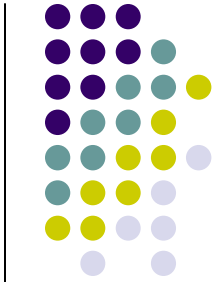
↓

Use median rather than mean of data

↓

Perform T Tests

Microarrays-Type I error problems

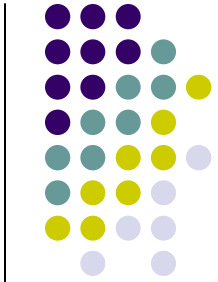


Recall that at a 5% level of significance, there is a 1/20 chance that we call a difference signficant when, in fact, it is not-→ Type I error.

Why is this a special problem in microarray analysis?

->Large number of comparisons made in a typical experiment practically guarantees Type I errors

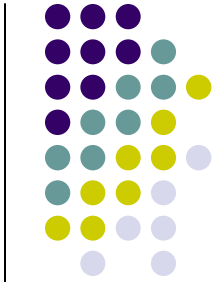
Microarrays-Type I error problems



How can we reduce possibility of Type I errors?

(1)Use a priori hypotheses whenever possible
(2)Use a more stringent level of significance than 5%
(3)Verify conclusions using other technologies

Statistics are…



- A mathematical extension of scientific reasoning…

- And therefore essential for all scientists to understand…

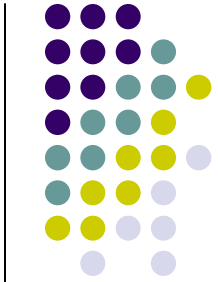
Statistical tests

Possible outcomes

1. A result is significant
2. A result is not significant
3. The test used is not powerful (robust) enough to determine if a difference is significant

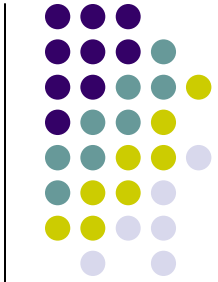
   ->Can use a more powerful test or increase replication

Statistical tests



- All statistical tests make assumptions
- More assumptions → Greater power
- Fewer assumptions → Less powerful
- Is there a downside in making more assumptions?

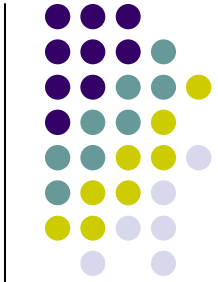
Statistical test assumptions



Yes!!!

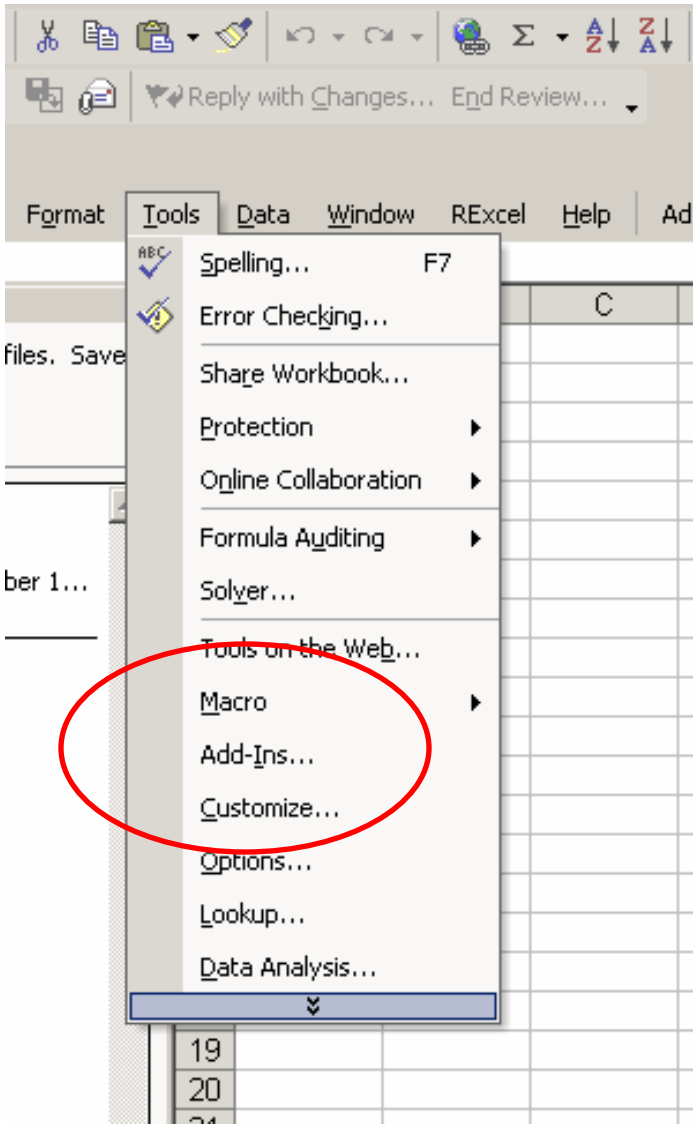
The stated assumptions may be incorrect leading to invalid conclusions…

Statistics in Excel

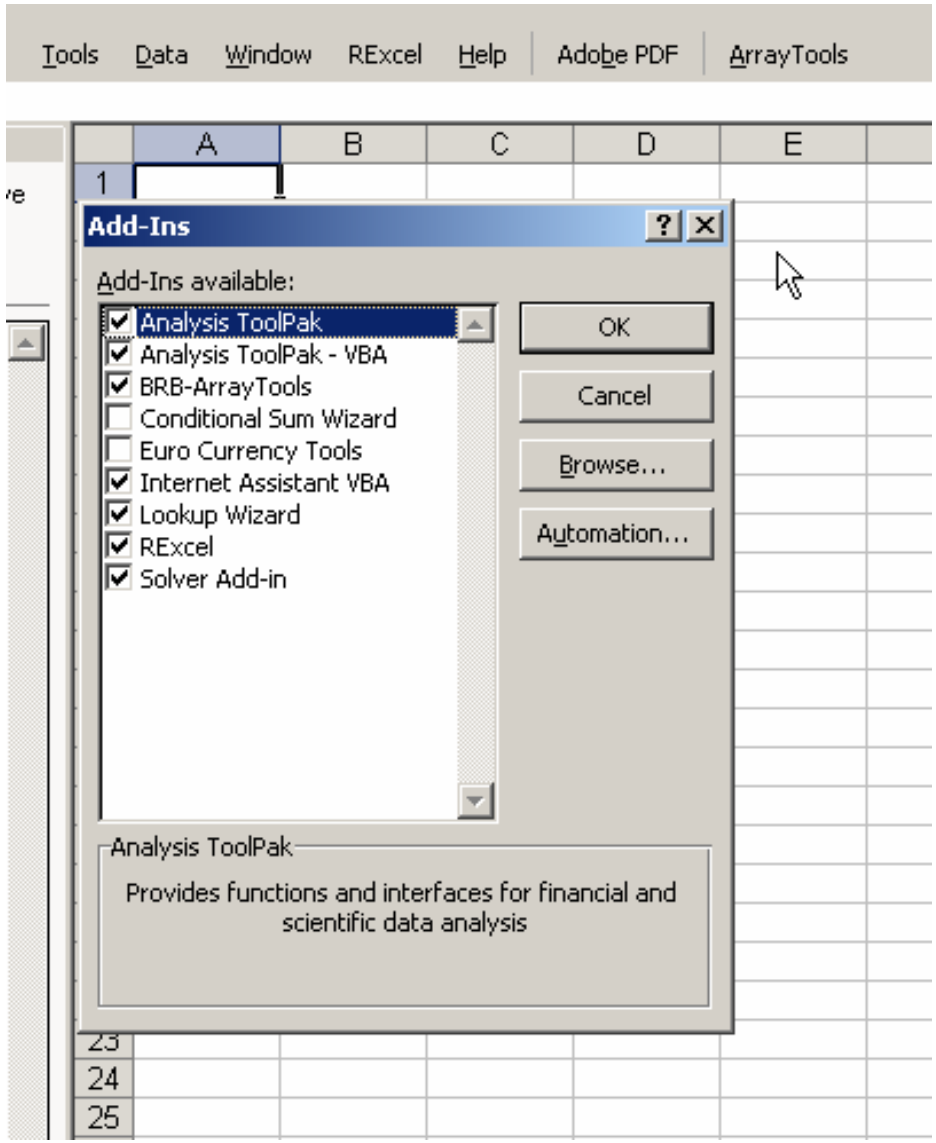


- Can use functions to determine basic statistics but this is limited.
- To use all the statistical capabilities within Excel, you must install the Analysis ToolPak Add-in.
  - This can do regression analysis, ANOVA, t-tests

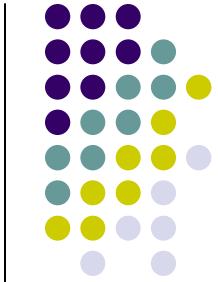
Statistics in Excel



Statistics in Excel



Statistics in Excel



Perform a statistical analysis

1. On the **Tools** menu, click **Data Analysis**.

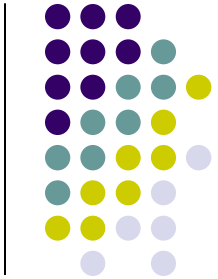
   If **Data Analysis** is not available, load the Analysis ToolPak.

   ▶ How?

2. In the **Data Analysis** dialog box, click the name of the analysis tool you want to use, then click **OK**.
3. In the dialog box for the tool you selected, set the analysis options you want.

   You can use the **Help** button on the dialog box to get more information about the options.

Statistics in Excel



About statistical analysis tools

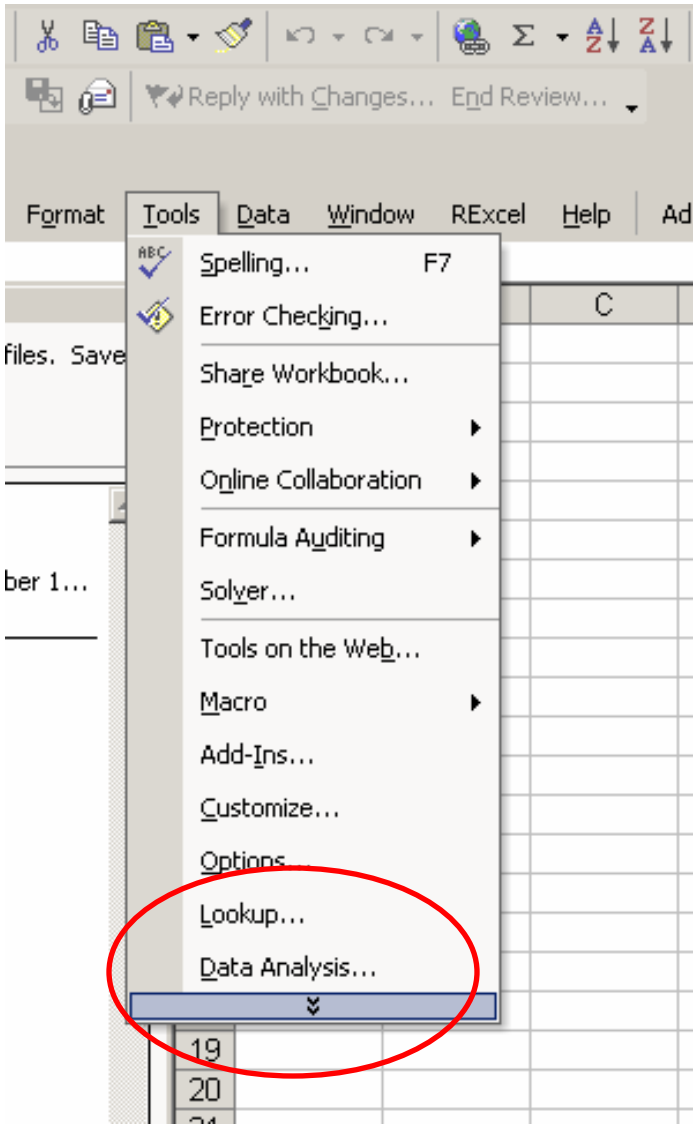
Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that you can use to save steps when you develop complex statistical or engineering analyses. You provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Related worksheet functions Excel provides many other statistical, financial, and engineering worksheet functions. Some of the statistical functions are built-in and others become available when you install the Analysis ToolPak.

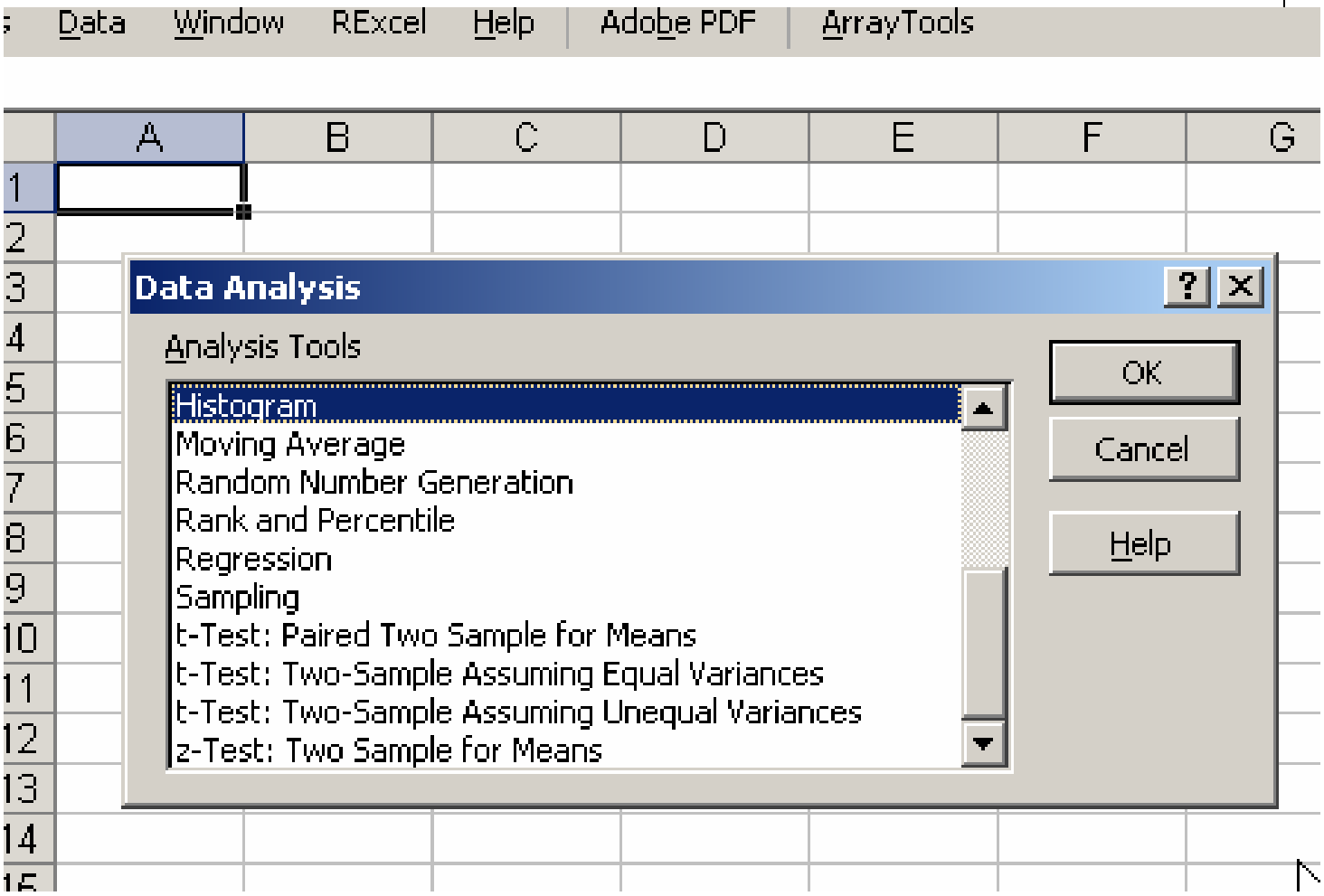
Accessing the data analysis tools The Analysis ToolPak includes the tools described below. To access these tools, click **Data Analysis** on the **Tools** menu. If the **Data Analysis** command is not available, you need to load the Analysis ToolPak add-in program.

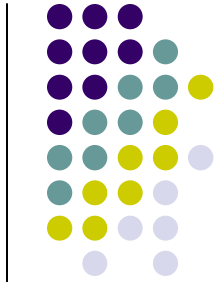
- Anova
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test
- z-Test

Statistics in Excel



Statistics in Excel





The End…