

Chem 351
Midterm Exam
Take-Home Portion

Complete each of the following problems. For number-crunching you may choose any combination of a handheld calculator, Excel or the program R.

Begin each problem by explaining what statistical analysis you are going to use and why it is appropriate, followed by your calculations. Be sure to fully and completely annotate your work; partial credit cannot be assigned for incorrect answers unless your work is traceable. In general this means you should turn in annotated Excel worksheets and/or cut and paste into your answers relevant output from R. A narrative explaining your number-crunching is essential. Your narrative and annotations must clearly lead me through your work. Each problem must begin on a separate page.

Where appropriate, be sure to state null and alternative hypotheses and to state clearly your conclusions. Be critical when evaluating the results of statistical tests. Do you trust the results or are there reasons for finding them suspicious?

You are free to use your textbook, the library, web resources, previous problem sets and your notes and handouts while working on this exam. You are not free to discuss any portion of this exam with other students or with faculty members other than the instructor. This applies to the programs R and Excel as well. All questions about the exam or about the use of R or Excel must be directed to the instructor.

R files containing the data for each problem are located in your personal folder on the course's I-drive. If you wish to work with the data in Excel, you can create comma-separated value files using the following syntax:

```
write.csv(X, file = "filename.csv")
```

where X is an object and filename.csv is your choice for the file's name (the extension ".csv" is required). You can open such files in Excel by choosing 'Open' from the menu bar and selecting 'All Files' for the file type.

Exams are due on Monday, March 19th at the beginning of class. This is an absolute deadline. Late exams will carry a penalty of 10 points per day, beginning with the start of class.

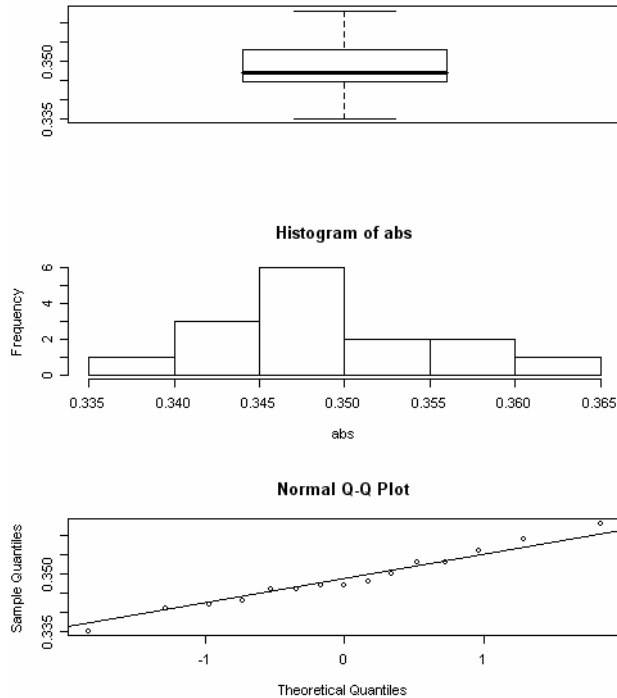
The take-home portion of the midterm is worth 67% of the exam's total value.

Good Luck!

Problem One. As a test of an analyst's skill with volumetric glassware, she was asked to prepare several dilutions of a stock solution of Cu^{2+} and to measure the absorbance of each solution. The results are in the file "ProblemOne.RData." Examine this data and evaluate whether it is normally distributed. Your consideration of this must be both quantitative and qualitative.

Answer: Three visual approaches to examining the data are a box plot, a histogram and a Q-Q plot. None of the three plots provides compelling evidence to suggest that the data are not normally distributed. There is some small evidence of a skew to the data in that the median is in the lower half of the IQR (as shown by the box), but this is not unusual for relatively small data sets.

Two quantitative approaches to examining the data are the skewness and kurtosis, which are, respectively, 0.210 and -0.686, both of which should be zero for normally distributed data. The skewness value is small and consistent with the conclusions about skew from the plots. The kurtosis value suggests that distribution is a bit too flat, but, again, for a small number of data points, this is not, by itself, particularly compelling evidence. Together, the visual and quantitative evidence suggest that the data probably are normally distributed.



Assuming that the mean and standard deviation provide good approximations for μ and σ , what is the probability that her next dilution will produce an absorbance of more than 0.350? Of less than 0.340? Of between 0.345 and 0.355?

Answer: The mean and standard deviation are 0.3486 and 0.007308, respectively. To find the probability of obtaining an absorbance of more than 0.350, we find the probability of obtaining a result of less than 0.350 using the command **pnorm** and subtract from 1.00; this gives a result of 0.424, or 42.4%.

To find the probability of obtaining a result of less than 0.340, we use the command **pnorm** without subtracting from 1.00, obtaining a result of 0.120, or 12.0%.

To find the probability of obtaining an absorbance between 0.345 and 0.355, we find the probability of obtaining a result less than 0.355 and the probability of obtaining a result

of less than 0.345. The difference between the two gives the probability for which we are looking. The two probabilities are 0.809 and 0.311, and the difference is 0.498, or 49.8%.

The stock solution had an absorbance of 1.318 and the analyst diluted it using a 25-mL volumetric pipet and a 100-mL volumetric flask. Is there any evidence, at a probability level of 0.05, that there is determinate error in her dilutions? What assumption must you make in answering this question?

Answer: If we assume that Beer's law applies, then the dilution of the sample should lead to a true absorbance of 0.329₅. The experimental mean value of 0.348₆ for 15 measurements can be evaluated with respect to the true mean using a t-test with

$$H_0: \bar{X} = \mu \quad H_A: \bar{X} \neq \mu$$

A t-test gives indicates that the probability of the null hypothesis being correct is 8.027×10^{-8} , which is much smaller than 0.05; therefore, we reject the null hypothesis and conclude that there is determinate error in the analyst's dilutions.

Problem Two. To evaluate two methods for extracting lead from soil, samples were collected from 20 locations and returned to the lab. Each sample was homogenized to a consistent particle size and then split into two parts. One part was extracted using a solution of dilute HCl and the other part was extracted using a solution of dilute EDTA. Each extract was then analyzed by atomic absorption spectroscopy. Results, as $\mu\text{g Pb/g}$ soil, are in the file "ProblemTwo.RData." Determine, at $\alpha = 0.05$, if there is a significant difference in the methods of extracting lead from soil.

Answer: Because each sample is from a separate location and is divided into two parts, each extracted by a different method, the data are paired. A two-tailed analysis with a paired t-test using

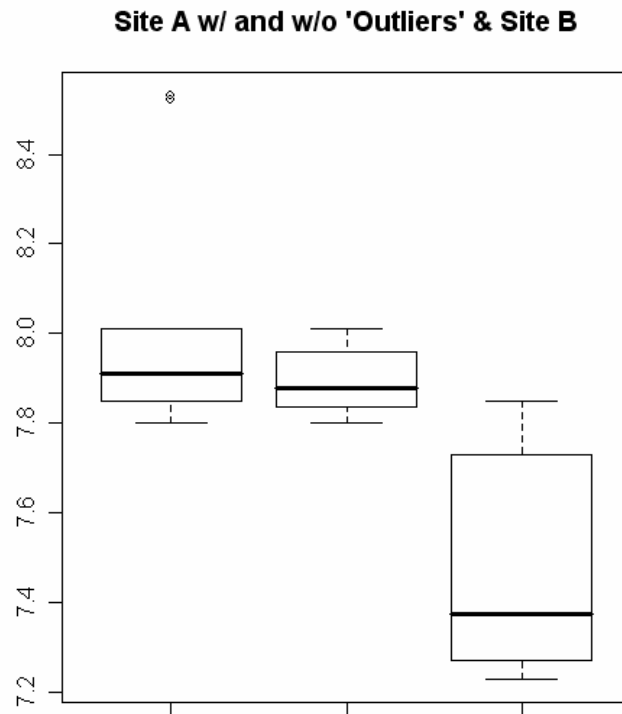
$$H_0: d = 0 \quad H_A: d \neq 0$$

where d is the difference between results for the two extraction methods shows that the probability for accepting the null hypothesis is 0.07047. Because this probability exceeds the stated level of 0.05, we retain H_0 and conclude that there is no evidence for a difference between the two extraction methods.

Problem Three. Soil pH is an important measurement in environmental studies of soils. One method for measuring a soil's pH is to combine a sample with H_2O at a specified solid:solution ratio, shake the sample for a specified period of time, allow the soil to settle and then measure the pH of the solution overlying the soil. Two samples of soil from different locations are collected and returned to the lab. Ten subsamples are taken from each of the two samples, and the soil pH measured for each subsample. The

resulting pH values are in the file “ProblemThree.RData.” Determine, at $\alpha = 0.05$, if there is a significant difference in the pH values.

Answer: This problem points out some of the difficulties in working with relatively small sets of data. Examining the raw data suggests that Site A might have a couple of outliers (see box plots to the right). On the other hand, if the samples are homogeneous, then there is every reason to believe that the variances for the two sites should reflect the variance in the method for determining pH and, therefore, should be similar; removing the outliers from Site A leads to a significant difference in variance (as confirmed by an F-test; results not shown). Since there is no particular reason to believe that the underlying data are not normally distributed, an unpaired t-test is suitable; however, a Wilcoxon test also is appropriate. The t-test requires first determining if variances can be pooled, which is confirmed by an F-test ($H_0: s^2_{\text{SiteA}} = s^2_{\text{SiteB}}, H_A: s^2_{\text{SiteA}} \neq s^2_{\text{SiteB}}, p = 0.726$). The t-test ($H_0: \bar{X}_{\text{SiteA}} = \bar{X}_{\text{SiteB}}, H_A: \bar{X}_{\text{SiteA}} \neq \bar{X}_{\text{SiteB}}, p = 0.0001902$) shows that the difference between the pH at Site A and Site B is significant. A Wilcoxon test shows a similar result, yielding a p-value of 0.0005745.



Problem Four. A determination of Vitamin E in salad oil was carried out using a standard spectrophotometric procedure that includes an extraction of the vitamin from the oil and by an HPLC separation with spectrophotometric detection. The data in the file “ProblemFour.RData” gives results, as % by mass, for several replicate samples drawn from a single bottle of salad oil. Determine, at $\alpha = 0.05$, if there is a significant difference between the methods.

Answer: This is a more straightforward analysis than for the previous problem as there is no suggestion that the data are not from a normal distribution. The comparison requires an unpaired t-test ($H_0: \bar{X}_{\text{Extract}} = \bar{X}_{\text{HPLC}}, H_A: \bar{X}_{\text{Extract}} \neq \bar{X}_{\text{HPLC}}$), which, in turn, requires an F-test on the variances ($H_0: s^2_{\text{Extract}} = s^2_{\text{HPLC}}, H_A: s^2_{\text{Extract}} \neq s^2_{\text{HPLC}}$). The F-test yields a probability of 0.3363, which means we retain the null hypothesis of equal

variances. The t-test yields a probability of 0.00432, which is less than 0.05; thus, we reject the null hypothesis and have evidence for a significant difference between the two methods.

Problem Five. To evaluate the interlaboratory reproducibility of an analytical method for determining the purity of 1-bromopropane, a pesticide used against nematodes, samples were sent to four laboratories with instructions to carry out an analysis and to report results. The data are in the file “ProblemFive.RData.” Determine, at $\alpha = 0.05$, if there is a significant difference between the labs. If there is a significant difference, then evaluate the source(s) of the difference.

Answer: This problem calls for a one-way analysis of variance, with the treatments being the four laboratories. There are 22 total measurements, giving 21 total degrees of freedom. The between treatment variance is 36.530 with three degrees of freedom and the within treatment variance is 0.568 with 18 degrees of freedom. A one-tailed F-test of

$$H_0: MS_{\text{between}} = MS_{\text{within}} \quad H_A: MS_{\text{between}} > MS_{\text{within}}$$

gives an F-ratio of 64.284 and a probability of 8.153×10^{-10} , indicating that there is a significant difference between the laboratories. To evaluate the source of the difference we use the TukeyHSD test, which shows significant differences (p-values < 0.05) between all but Labs A and B.

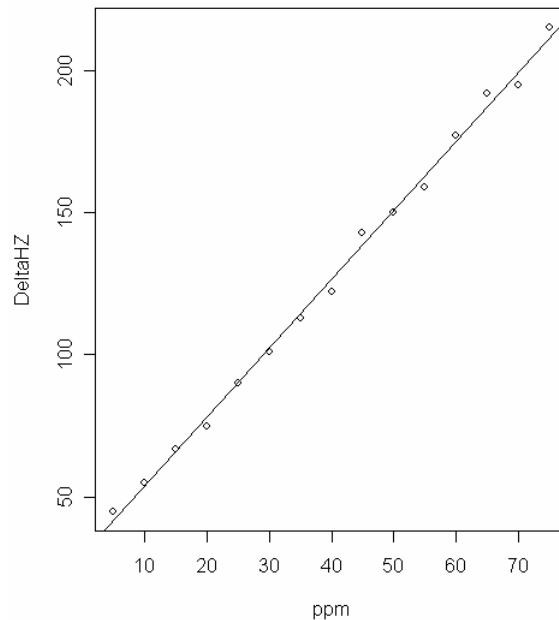
Problem Six. To study the effect of temperature and time on a reaction’s percent yield, a chemist ran three trials for each combination of three reaction times and three reaction temperatures. The data are in the file “ProblemSix.RData” with object names of DxxTxx for the duration and temperature; thus, for example, the object D20T50 gives percent yields for the combination of a 20 minute reaction time and a temperature of 50°C. Evaluate the significance of reaction time and temperature on the reaction’s percent yield.

Answer: This problem requires a two-way analysis of variance with the treatments being the duration (time) and the temperature. There are 27 total measurements, giving 26 total degrees of freedom. The data need to be placed in a data frame with columns for the values and codes for the durations and temperatures. For the later two, the codes must be non-numeric. The variance between durations is 46.259 with two degrees of freedom. The variance between temperatures is 61.815 with two degrees of freedom. There also is an interaction between duration and temperature, which has a variance of 75.815 with four degrees of freedom. The within variance, which is due to random error, is 3.677 with 18 degrees of freedom. The effects of the duration, temperature and interaction are significant with F-ratios (and p-values) of 12.616 (0.000375), 16.859 (7.484×10^{-5}) and 20.677 (1.562×10^{-6}) respectively.

Problem Seven. Piezoelectric sensors are devices that can be made to vibrate at a characteristic frequency, which depends upon the sensor’s composition. Such materials can serve as a balance because the addition of a very small mass changes the sensor’s

characteristic frequency. For example, a sensitive sensor for SO₂ uses a piezoelectric material coated with a thin film that absorbs SO₂; the more SO₂ that absorbs, the greater the change in mass and the greater the shift in the sensor's characteristic frequency. Typical calibration data is provided in the file "ProblemSeven.RData." Construct a regression model for this data using the model $\Delta\text{Hertz} = \beta_0 + \beta_1 \times \text{ppm}$. Previous work with the same sensor gave b_0 as 25.176 and b_1 as 2.372. Is there evidence, at $\alpha = 0.05$, that the new calibration curve differs from the previous calibration curve?

Answer: A plot of the data and the regression model are shown in the figure below. The regression model is $\Delta\text{Hertz} = 29.514 + 2.427 \times \text{ppm}$ with a correlation coefficient of 0.9963 and an F-ratio of 3817 (p-value < 2.2×10^{-16}), both of which suggest that there is a correlation between the dependent and independent variables. Visual examination of the residuals, which are of similar magnitude and show no determinate trend, suggests that there is no reason to doubt that the model is appropriate.



Evaluating whether the old and new models are different requires a t-test of the intercepts and slopes. In this case we will take the old values for b_0 and b_1 of 25.176 and 2.372, respectively, as being the true values and use two-tailed t-tests of

$$H_0: b_{0,\text{new}} = 25.176 \quad H_A: b_{0,\text{new}} \neq 25.176$$

$$H_0: b_{1,\text{new}} = 2.372 \quad H_A: b_{1,\text{new}} \neq 2.372$$

to evaluate whether the differences are significant. The value of t_{exp} for the intercept is 2.429, which, for 13 degrees of freedom, has a probability of 0.0304. We can, therefore, conclude that the new intercept is significantly different than the old intercept. The value of t_{exp} for the slope is 1.403, which, for 13 degrees of freedom, has a probability of 0.1839; thus, there is no evidence for a significant difference between the old and new slopes.

Problem Eight. Gas chromatography with detection by mass spectrometry (GC-MS) is a common quantitative analytical method for dimethylsulfide (DMS). A series of 10 standards containing known concentrations of DMS (in ng/L) were prepared and injected onto the instrument's column, and the area under the resulting chromatographic peak was recorded. Construct a suitable regression model for this calibration using the data in the file "ProblemEight.RData."

Answer: A plot of the data shows that a straight-line is inappropriate and that a quadratic (second-order polynomial) equation might suffice. Fitting the model

$$\text{PeakArea} = \beta_0 + \beta_1 \text{DMS} + \beta_2 \text{DMS}^2$$

gives estimates for β_0 , β_1 and β_2 of 0.02786, 4.999 and 0.2500, respectively. The correlation coefficient of 1 and the F-ratio of 1.26×10^{10} ($p\text{-value} < 2.2 \times 10^{-16}$) suggest that there is a strong correlation between PeakArea and DMS using this model. A plot of the data and the regression model shows small residual errors without any obvious trend that might suggest a poor model, a fact that we can confirm using residual plots, Q-Q plots and leverage plots.

