BST 140.752
Problem Set 3

1 Hypothesis testing

1. A special study is conducted to test the hypothesis that persons with glaucoma have higher blood pressure than average. Two hundred subjects with glaucoma are recruited with a sample mean systolic blood pressure of $140mm$ and a sample standard deviation of $25mm$. (Do not use a computer for this problem.)

   a. Construct a 95% confidence interval for the mean systolic blood pressure among persons with glaucoma. Do you need to assume normality? Explain.

   b. If the average systolic blood pressure for persons without glaucoma of comparable age is $130mm$. Is there statistical evidence that the blood pressure is elevated? Perform the relevant test and report a P-value.

2. Consider the previous question.

   a. Make a probabilistic argument that the interval

   $$\left[\bar{X} - z_{.95}\frac{s}{\sqrt{n}}, \quad \infty\right]$$

   is a 95% *lower bound* for μ.

3. Suppose we wish to estimate the concentration μg/mℓ of a specific dose of ampicillin in the urine. We recruit 25 volunteers and find that they have sample mean concentration of 7.0 μg/m ℓ with sample standard deviation 3.0 μ g/mℓ. Let us assume that the underlying population distribution of concentrations is normally distributed.

   a. Find a 90% confidence interval for the population mean concentration.

   b. How large a sample would be needed to insure that the length of the confidence interval is 0.5 μ g/mℓ if it is assumed that the sample standard deviation remains at 3.0 μ g/m ℓ?

4. A study of blood alcohol levels (mg/100 ml) at post mortem examination from traffic accident victims involved taking one blood sample from the leg, A, and another from the heart, B. The results were:

Case	A	B	Case	A	B
1	44	44	11	265	277
2	265	269	12	27	39
3	250	256	13	68	84
4	153	154	14	230	228
5	88	83	15	180	187
6	180	185	16	149	155
7	35	36	17	286	290
8	494	502	18	72	80
9	249	249	19	39	50
10	204	208	20	272	271

Test whether or not the mean blood alcohol level differs between the heart and the leg. Give the appropriate null and alternative hypotheses. Give the relevant P-value. Interpret your results, state your assumptions.

5. Forced expiratory volume FEV is a standard measure of pulmonary function. We would expect that any reasonable measure of pulmonary function would reflect the fact that a person's pulmonary function declines with age after age 20. Suppose we test this hypothesis by looking at 10 nonsmoking males ages 35-39, heights 68-72 inches and measure their FEV initially and then once again 2 years later. We obtain this data.

Person	Year 0 FEV (L)	Year 2 FEV (L)	Person	Year 0 FEV (L)	Year 2 FEV (L)
1	3.22	2.95	6	3.25	3.20
2	4.06	3.75	7	4.20	3.90
3	3.85	4.00	8	3.05	2.76
4	3.50	3.42	9	2.86	2.75
5	2.80	2.77	10	3.50	3.32

   a. Preform and interpret the relevant test. Give the appropriate null and alternative hypotheses. Intepret your results, state your assumptions and give a P-value.

6. Another aspect of the preceding study involves looking at the effect of smoking on baseline pulmonary function and on change in pulmonary function over time. We must be careful since FEV depends on many factors, particularly age and height. Suppose we have a comparable group of 15 men in the same age and height group who are smokers and we measure their FEV at year 0. The data are given (For purposes of this exercise assume equal variance where appropriate).

Person	FEV Year 0 (L)	FEV Year 2 (L)	Person	FEV Year 0 (L)	FEV Year 2 (L)
1	2.85	2.88	9	2.76	3.02
2	3.32	3.40	10	3.00	3.08
3	3.01	3.02	11	3.26	3.00
4	2.95	2.84	12	2.84	3.40
5	2.78	2.75	13	2.50	2.59
6	2.86	3.20	14	3.59	3.29
7	2.78	2.96	15	3.30	3.32
8	2.90	2.74			

Test the hypothesis that the change in FEV is equivalent between non-smokers and smokers. State relevant assumptions and interpret your result. Give the relevant P-value.

7. Suppose that systolic blood pressures were taken on 16 oral contraceptive users and 16 controls at baseline and again then two years later. The average difference from follow-up SBP to the baseline (followup - baseline) was $11\ mmHg$ for oral contraceptive users and $4\ mmHg$ for controls. The corresponding standard deviations of the differences was $20\ mmHg$ for OC users and $28\ mmHg$ for controls.

   a. Calculate and interpret a 95% confidence interval for the **relative** change in systolic blood pressure for oral contraceptive users; assume normality on the log scale.

   b. Does the change in SBP over the two year period appear to differ between oral contraceptive users and controls? Perform the relevant hypothesis test and interpret. Give a P-value. Assume normality and a common variance.

2 Inference and estimation in linear models

1. Consider the linear model $Y = X\beta + \epsilon$ and $\epsilon \sim N(0, \sigma^2 I)$. Do the following

   A. Derive the ML estimate, $\hat{\beta}$.

   B. Derive the variance of the ML estimate.

   C. Show that $\hat{\beta}$ is independent of the residual vector, e.

2. Let β_j be an element of β from the previous problem. Let $\hat{SE}_{\hat{\beta}_j}$ be the standard error of $\hat{\beta}_j$, the ML estimate of β_j. Argue that $(\hat{\beta}_j - \beta_j)/\hat{SE}_{\hat{\beta}_j}$ follows a T distribution with $n-p$ degrees of freedom. Use this to create a confidence interval for $\hat{\beta}_j$.

3. Let $Y_{ij} = \mu_i + \epsilon_{ij}$ for $i = 1, 2$ and $j = 1, \ldots, J_i$ where the $\epsilon_{ij} \sim N(0, \sigma^2)$ are iid. Show that the unbiased estimate of σ^2 is the so-called pooled variance estimate, $S_p^2 = \frac{1}{J_1+J_2-2}\{(J_1 - 1)S_1^2 + (J_2 - 1)S_2^2\}$ where S_i^2 is the standard variance estimate within group i. Derive a T confidence interval for $\mu_1 - \mu_2$ and test of $\mu_1 = \mu_2$.

4. Derive the variance estimate from the previous problem of $i = 1, \ldots I$. Derive an overall F test for the hypothesis that $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$ versus the alternative that at least two are unequal. Argue that this F test compares the variation between the groups to that within the groups. (This is called the general ANOVA F test.)

5. Let $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \Sigma)$.

   A. Argue that for any W, including I, $\hat{\beta}(W) = (X'WX)^{-1}X'WY$ is an unbiased estimate of β.

   B. What is the variance of $\hat{\beta}(W)$?

   C. Argue that $\hat{\beta}(\Sigma^{-1})$ is the ML estimate if Σ were known.

   D. Use the previous result to calculate the MLE of $\mu = (\mu_1, \ldots, \mu_I)'$ when $Y_{ij} = \mu_i + \epsilon_{ij}$ where the ϵ_{ij} are independent Gaussians with mean 0 and variance σ_i^2.

6. Consider the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Argue that the variance of $\hat{\beta}_1$ is minimized with the variance in the observed X_i is maximized. Ergo, the lowest variance estimate is obtained with what pattern in the X_i?

7. Consider the general linear hypothesis $H_0 : K\beta = m$ versus $H_a : K\beta \neq m$. Go through a careful development of the general F test.

8. Derive the T confidence interval, T-test and F test associated with the hypothesis $H_0 : \beta_k = 0$ versus $H_a : \beta_k \neq 0$ where β_k is a component of β.

9. Give a proof that if $X = [X_1 X_2]$ with X_1 as $1 \times n$ and X_2 is $(p-1) \times n$ where $Y = X\beta + \epsilon$ and $\beta = [\beta_1 \beta_2]'$ with β_1 as 1×1 then $\hat{\beta}_1 = e_{y|X_1} e_{X_1|X_2} / < e_{X_1|X_2}, e_{X_1|X_2} >$ where $e_{A|B} = (I - B(B'B)^{-1}B')A$.

10. Show that the T confidence interval for $\beta_1 - \beta_2$ for the model $Y_{ij} = \beta_i + \epsilon_{ij}$ for $i = 1, 2$ and $j = 1, \ldots, J_i$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ is $\bar{Y}_1 - \bar{Y}_2 \pm t_{1-\alpha/2, J_1+J_2-2} S_p \sqrt{\frac{1}{J_1} + \frac{1}{J_2}}$. (This is the standard interval given for two group differences in introductory statistics classes).

3 Coding and data analysis exercises

1. Perform the following simulation. Randomly simulate $1,000$ sample means of size 16 from a normal distribution with means 5 and variances 1. Calculate $1,000$ test statistics for a test of $H_0 : \mu = 5$ versus $H_a : \mu < 5$. Using these test statistics calculate $1,000$ P-values for this test. Plot a histogram of the P-values. Note, this exercise demonstrates the fact that the distribution of P-values is uniform.

2. Here we will verify that standardized means of iid normal data follow Gossett's t distribution. Randomly generate $1,000 \times 20$ normals with mean 5 and variance 2. Place these results in a matrix with $1,000$ rows. Using two `apply` statements on the matrix, create two vectors, one of the sample mean from each row and one of the sample standard deviation from each row. From these $1,000$ means and standard deviations, create $1,000$ t statistics. Now use

R's `rt` function to directly generate $1,000$ t random variables with 19 df. Use R's `qqplot` function to plot the quantiles of the constructed t random variables versus R's t random variables. Do the quantiles agree? Describe why they should.

3. Here we will verify the chi-squared result. Simulate $1,000$ sample variances of 20 observations from a normal distribution with mean 5 and variance 2. Convert these sample variances so that they should be chi-squared random variables with 19 degrees of freedom. Now simulate $1,000$ random chi-squared variables with 19 degrees of freedom using R's `rchisq` function. Use R's `qqplot` function to plot the quantiles of the constructed chi-squared random variables versus those of R's random chi-squared variables. Do the quantiles agree? Describe why they should.

4. Download the data at

   `http://dl.dropbox.com/u/95701/751.2/takeHomeData.zip`

   The data documentation are as follows The data are obtained from the Sleep Heart Health Study, though having been modified for the exercise and for data confidentiality, so that numbers from this data set will not match with published numbers from the same study.

```
Variables:
1. tst  total sleep time in hours;
2. events  number of sleep related events over the night
3. meds  was the subject on anti-hypertensive medications (1=yes, 0=no).
4. sbp  systolic blood pressure mmHg
5. dbp  diastolic blood pressure mmHg
6. age  age in years
7. bmi  body mass index kg / m2
8. race  (1=w 2=b 3=Nat Am/Alaskan 4=Asian/PI 5=hisp/Mex Amer 6=other)
9. gender  (1 =male, 0=female)
10. alcohol  (number of drinks per week)
11. smoke  (ever smoked cigarettes, at least 20 packs in a lifetime, 1=yes, 0=no)
12. Waist/Hip ratio
```

   Of interest is the rate of respiratory disturbances (events) per hours slept. Per common practice in the field, the manuscripts referenced above refer to this rate as the apnea/hypopnea index (AHI) or respiratory disturbance index (RDI).

   A. Fit a linear model to consider the relationship between the Log(respiratory disturbance index + 1) (response) and BMI (predictor).
   B. Create exploratory graphics to investigate the relationship.
   C. Test the hypothesis that RDI is associated with BMI.
   D. Create and interpret relevant confidence intervals.
   E. Create relevant residual plots.

F. Investigate missing data patterns.

Use the article "Association of sleep disordered breathing, sleep apnea and hypertension in a large, community based study" as your guide for including confounders.

5. Download the data from

   `https://dl.dropboxusercontent.com/u/95701/teams.zip`

   Some documentation for the data can be found at:

   `https://dl.dropboxusercontent.com/u/95701/teamsDocumentation.docx`

   Answer the following

   A. What team performance predictor variables are most useful for predicting winning percentage (games won divided by games played) among teams since 1970?

6. Write an R function, myLM. The function should take in a Y vector and X matrix and do the following:

   A. Check to make sure that X and Y have the right format (matrix and vector respectively), have the right dimensions, have no missing, NA or Inf, are numeric and that X is full rank.

   B. Return a list with the following information:

      i. The least squares estimate of the associated linear model.
      ii. R squared.
      iii. A T table (estimate, standard error, t statistics, P-value).
      iv. A residual vector.
      v. A vector of fitted values.
      vi. A vector hat diagonals.

   C. Test it out on models that you fit for the previous problem and make sure that your numbers agree with that of R's lm.

7. Go to Leonard Stefanski's web page on residuals plots. Read the associated American Statistician article. Reproduce at the figures for the "Correct figure", "X marks the spot" and "Homer Simpson" examples using your myLM function.