If $A$ is a real symmetric $n \times n$ matrix then show that $A$ is idempotent if and only if $A = P P^T$, where $P$ is a $n \times r$ matrix such that

$$P^T P = I_r$$

and $r$ is the rank of $A$.

Solution-Math 310 -MS Exam, Spring 2007-Majumdar

“If” Part: $A^2 = P P^T P P^T = P P^T = A$

“Only if part”: Spectral Decomposition: $A = Q \Delta Q^T$ where $QQ^T = Q^T Q = I_n$ where $\Delta$ is diagonal. Since the eigen values of $A$ are 0 or 1 we can write

$$\Delta = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$

where $r =$rank $(A)$.

So

$$A = \begin{bmatrix} P & P_1 \end{bmatrix} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} P^T \\ P_1^T \end{bmatrix} = PP^T.$$

Writing $Q = [P; P_1]$ we have

$$I = Q^T Q = \begin{bmatrix} P^T \\ P_1^T \end{bmatrix} \begin{bmatrix} P & P_1 \end{bmatrix} = \begin{bmatrix} PP^T & PP_1^T \\ P_1^T P & P_1^T P_1 \end{bmatrix} \Rightarrow P^T P = I_r.$$

Q.E.D
Let $f$ and $g$ be continuous on $[0, 1]$, and suppose that $f(0) < g(0)$ and $f(1) > g(1)$. Show that there exists $x$ in $(0, 1)$ such that $f(x) = g(x)$. Deduce that the equation

$$\frac{x + 1}{3} = \sin \frac{\pi x}{2}$$

has a solution in $(0, 1)$.

**Solution to Math 313 Problem -MS Exam, Spring 2007-Miescke**

The function $f - g$ is continuous, and $(f - g)(0) < 0$, $(f - g)(1) > 0$. Hence, by the intermediate value theorem, there exists $c$ in $(0, 1)$ such that $(f - g)(c) = 0$.

For the given equation, use

$$f(x) = \sin \frac{\pi x}{2} \quad \text{and} \quad g(x) = \frac{x + 1}{3},$$

where $x$ is in $[0, 1]$. Q.E.D
The joint p.d.f. of random variables $X$ and $Y$ are given below. Compute the correlation coefficient of $X, Y$. Are they independent?

$$f(x, y) = \begin{cases} 
2, & \text{if } 0 < x < 1, \ 0 < y < 1, \ x + y < 1 \\
0, & \text{otherwise}.
\end{cases}$$

Solution to Stat 401 problem- MS Exam, Spring 2007-Majumdar

$$E(X) = 2 \int_{0}^{1} \int_{0}^{(1-x)} xdydx = 2 \int_{0}^{1} x(1-x)dx = \frac{1}{3}$$

Similarly $E(Y) = \frac{1}{3}$.

$$E(X^2) = 2 \int_{0}^{1} \int_{0}^{(1-x)} x^2dydx = 2 \int_{0}^{1} x^2(1-x)dx = \frac{1}{6} = E(Y^2).$$

$$E(XY) = 2 \int_{0}^{1} \int_{0}^{(1-x)} xydydx = \frac{1}{12}$$

$$V(X) = E(X^2) - E(X)^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18} = V(Y)$$

$$\text{cov}(X, Y) = \frac{1}{12} - \frac{1}{9} = -\frac{1}{36}$$

$$\rho = \text{Corr} (X, Y) = \frac{-\frac{1}{36}}{\sqrt{\frac{1}{18} \times \frac{1}{18}}} = -\frac{1}{2}$$

Since $\rho \neq 0$ $X, Y$ are not independent. Q.E.D

Stat 411 - (Chapter 5,6,7):-MS Exam, Spring 2007

Suppose $X_1, \ldots, X_n$ are iid with the pdf

$$f(x; \theta) = \frac{2x}{\theta^2}, \ 0 < x \leq \theta$$

(a) Find the mle $\hat{\theta}$ for $\theta$.

(b) Find the mle for the median of the distribution.
(a) The likelihood function

\[ L(\theta) = \prod_{i=1}^{n} \frac{2x_i}{\theta^2} = \theta^{-2n} \cdot 2^n \prod_{i=1}^{n} x_i, \quad 0 < x_1, \ldots, x_n \leq \theta \]

The log likelihood function

\[ l(\theta) = \log L(\theta) = -2n \log \theta + n \log 2 + \log \left( \prod_{i=1}^{n} x_i \right) \]

The first partial of \( l \) is

\[ \frac{\partial l(\theta)}{\partial \theta} = -\frac{2n}{\theta} < 0, \quad \text{for} \; \theta \geq \max\{x_1, \ldots, x_n\} > 0 \]

which implies that

\[ \hat{\theta} = \max\{X_1, \ldots, X_n\} \]

is the mle for \( \theta \).

(b) Because the distribution is continuous, the median \( m \) is the constant satisfying

\[ \int_{0}^{m} \frac{2x}{\theta^2} dx = \frac{1}{2} \]

Note that \( 0 < m < \theta \) and

\[ \int_{0}^{m} \frac{2x}{\theta^2} dx = \frac{x^2}{\theta^2} \bigg|_{0}^{m} = \frac{m^2}{\theta^2} \]

Then the median

\[ m = \sqrt{\frac{\theta^2}{2}} = \frac{\theta}{\sqrt{2}} \]

By part (a), the mle for \( \theta \) is \( \hat{\theta} = \max\{X_1, \ldots, X_n\} \). So the mle for the median is

\[ \frac{1}{\sqrt{2}} \max\{X_1, \ldots, X_n\} \]

Q.E.D
Let $X = (X_1, \ldots, X_n)'$ denote a random sample from the distribution $N(\theta, 1)$ that has the pdf

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right), \quad -\infty < x < \infty.$$ 

It is desired to test the hypothesis $H_0 : \theta = 0$ against the alternative hypothesis $H_1 : \theta = 1$.

(a) Show that the likelihood ratio $L(\theta = 0; X)/L(\theta = 1; X)$ is based upon the statistic $Y = \sum_{i=1}^{n} X_i$.

(b) If $n = 16$, find the best critical region of size $\alpha = 0.05$ for the hypothesis test. (Hint: Use the normal table attached.)

(c) If $n = 16$, find the power of the test in part (b).

Solution to Stat 411 Problem - (Chapter 8,9):-MS Exam, Spring 2007-Yang

(a) Proof: The likelihood function of $\theta$ given $x = (x_1, \ldots, x_n)'$ is

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2} \left[ \sum_{i=1}^{n} x_i^2 - 2\theta \sum_{i=1}^{n} x_i + n\theta^2 \right] \right\}.$$ 

Therefore the likelihood ratio

$$\frac{L(\theta = 0; X)}{L(\theta = 1; X)} = \exp\left\{ -\frac{n}{2} \left[ \sum_{i=1}^{n} X_i + \frac{n}{2} \right] \right\}.$$ 

So it is based upon the statistic $Y = \sum_{i=1}^{n} X_i$.

(b) For any positive constant $k$,

$$\frac{L(\theta = 0; x)}{L(\theta = 1; x)} \leq k \iff \sum_{i=1}^{n} x_i \geq \frac{n}{2} - \log k = c.$$ 

By the Neyman-Pearson Theorem, the best critical region $C$ takes the form of $\{(x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i \geq c\}$, where the constant $c$ is determined by

$$P_{H_0}(X \in C) = \alpha.$$ 

If $n = 16$, $Y \sim N(0,16)$ under the null hypothesis. So $Y/4 \sim N(0,1)$ and
\[0.05 = P_{H_0}(Y \geq c) = P(Y/4 \geq c/4)\]

By the normal table attached, \(c/4 = (1.64 + 1.65)/2 = 1.645\). So the best critical region is

\[C = \left\{(x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i \geq 6.58\right\}\]

(c) If \(n = 16\), under the alternative hypothesis \(H_1 : \theta = 1, Y \sim N(16,16)\). Then \(Z = (Y - 16)/4 = Y/4 - 4 \sim N(0,1)\).

The power is

\[P_{H_1}(Y \geq 6.58) = P(Z \geq -2.355)\]
\[= P(Z \leq 2.355)\]
\[= 0.95(0.9906 + 0.9909)\]
\[= 0.9907\]

Q.E.D


The following are the number of minutes it took 10 workers in a factory to complete a certain task, one before and one after each worker had received a special training for completing this task.

<table>
<thead>
<tr>
<th>Worker</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>After Training</td>
<td>4.6</td>
<td>5.8</td>
<td>6.9</td>
<td>7.2</td>
<td>8.2</td>
<td>6.5</td>
<td>6.9</td>
<td>7.2</td>
<td>7.3</td>
<td>8.7</td>
</tr>
<tr>
<td>Before Training</td>
<td>8.1</td>
<td>7.3</td>
<td>7.2</td>
<td>7.7</td>
<td>8.2</td>
<td>7.2</td>
<td>6.3</td>
<td>6.8</td>
<td>8.5</td>
<td>8.8</td>
</tr>
</tbody>
</table>

(a) Explain the model under which both, the Sign Test and the Wilcoxon Signed-Rank Test, can be used.

(b) Specify the appropriate null hypothesis and alternative.

(c) Compute the test statistic of the Sign Test for the given data.

(d) Compute the test statistic of the Wilcoxon Signed-Rank Test for the given data.

(e) Under what changes of the model could you still use the Sign Test, but not the Wilcoxon Signed-Rank Test?

Solution to Stat 416 Problem- MS Exam, Spring 2007. -Miescke

(a) \(Y_i - X_i, i = 1,\ldots,n\), are independent. \(Y_i - X_i\) has a c.d.f. \(F_i\) with a density \(f_i\) that is symmetric about \(\theta, i = 1,\ldots,n\).

(b) \(H_0: \theta = 0\) versus \(H_A: \theta > 0\), where the \(Y_i\) are from "Before Training".

(c) \(B = \#\{i : Y_i > X_i, i = 1,\ldots,n\} = 7\), and \(n\) is reduced 9 (Worker 5).
The sum of the positive signed ranks is \( T^+ = 37 \), and \( n \) is reduced to 9 (Worker 5).

(e) If the densities \( f_i \) are not all symmetric then only the Sign Test can be used.

Q.E.D

[7] Stat 431 - MS Exam, Spring 2007,

A random sample of size \( n \) has been drawn from a finite population of \( N \) units by adopting SRSWOR method and the sample so drawn is denoted by "s". Subsequently, a sub-sample \( s' \) of size \( n' \) \(< n \) is drawn from the chosen sample \( s \), again by adopting SRSWOR method. Denote by \( \bar{y}_s \) and \( \bar{y}_{s'} \) the respective sample means for a study variable \( Y \) based on the two samples. Assume:

\[
E[\bar{y}_s] = \bar{Y}
\]

\[
Var[\bar{y}_s] = S^2\left[\frac{1}{n} - \frac{1}{N}\right] \text{ where } S^2 \text{ denotes the population variance based on } Y-\text{values.}
\]

(a). Show that \( \bar{y}_{s'} \) also serves as an unbiased estimate for the population mean \( \bar{Y} \).

(b). Compute \( Var(\bar{y}_{s'}) \) and \( Cov(\bar{y}_s, \bar{y}_{s'}) \)

(c). Find the best linear combination of \( \bar{y}_s \) and \( \bar{y}_{s'} \) to serve as a pooled estimate of \( \bar{Y} \).

You may use conditional expectation arguments. Note that a subsample (\( s' \)) behaves like an induced sample with reference to the subpopulation captured by the original sample (\( s \)).

Solution to Stat 431 Problem -MS Exam, Spring 2007, Hedayat

Essential Steps.

(a) \( E[\bar{y}_{s'}; \text{given } s] = E_2[\bar{y}_{s'}] = \bar{y}_s \) and hence \( E[\bar{y}_{s'}] = E_1E_2 = E[\bar{y}(s)] = \bar{Y} \).

(b) \( Var[\bar{y}_{s'}] = V_1E_2 + E_1V_2 \)

\[
= S^2\left[\frac{1}{n} - \frac{1}{N}\right] + S^2\left[\frac{1}{n'} - \frac{1}{n}\right] = S^2\left[\frac{1}{n'} - \frac{1}{n}\right]
\]

(c) \( E[\bar{y}_s, \bar{y}_{s'}] = E_1E_2 = E[\bar{y}_s^2] = \bar{Y}^2 + Var[\bar{y}_s] \)

Therefore, \( Cov[\bar{y}_s, \bar{y}_{s'}] = Var[\bar{y}_s] = S^2\left[\frac{1}{n} - \frac{1}{N}\right] \)

The rest follows from a standard result in pooling two dependent estimates.

Q.E.D
A Markov chain $X_0, X_1, \ldots$ has the transition probability matrix

\[
\begin{pmatrix}
0 & 1 & 2 \\
0 & .3 & .2 & .5 \\
1 & .5 & .1 & .4 \\
2 & 0 & 0 & 1
\end{pmatrix}
\]

and is known to start in state $X_0 = 0$. Let $T = \min\{n \geq 0 : X_n = 2\}$. Find the probability that $T$ is an odd number.

**Solution to Stat 461 Problem - MS Exam, Spring 2007**-El-Neweihi

Let $u_0 = P(T \text{ odd}/X_0 = 0), \quad u_1 = P(T \text{ is odd}/X_0 = 1)$.

By the first step analysis

\[
u_0 = (.3)(1 - u_0) + (.2)(1 - u_1) + .5 \\
u_1 = (.5)(1 - u_0) + (.1)(1 - u_1) + .4
\]

Solving for $u_0$ we get $u_0 = \frac{25}{37}$. Q.E.D

---

Consider a normal population $X \sim N(\mu, \sigma^2 = 20)$. To test $H_0 : \mu = 25$ against $H_1 : \mu < 25$. Determine the sample size $n$ and the rejection region $\{\bar{X} < c\}$ to satisfy the significance level 0.025 and power level 0.975 at $\mu = 23$. [Given: $\Phi(1.96) = 0.975$.]

**Solution to Problem [9]:Stat 481 - MS Exam, Spring 2007**-Wang

The significance level and power level provide the following equations

\[
\begin{align*}
\alpha &= P\{ \text{Reject } H_0 | H_0 \text{ is true.} \} \\
1 - \beta &= P\{ \text{Reject } H_0 | H_0 \text{ is false.} \} \\
P(\bar{X} < c | \mu = 25) &= 0.025 \\
P(\bar{X} < c | \mu = 23) &= 0.975.
\end{align*}
\]

\[
\begin{align*}
\Phi\left(\frac{c - 25}{\sigma \sqrt{n}}\right) &= 0.025 \\
\Phi\left(\frac{c - 23}{\sigma \sqrt{n}}\right) &= 0.975
\end{align*}
\]

\[
\Rightarrow \begin{cases}
\frac{c - 25}{\sqrt{20/n}} = -1.96 \\
\frac{c - 23}{\sqrt{20/n}} = 1.96
\end{cases}
\]

Hence, we have sample size $n = 77, c = 24$ and the rejection region $\{\bar{X} < 24\}$. 

---
A multiple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, was used to fit $n = 20$ data points. It is calculated that $SSTO = 200$, $SSR = 66$.

(1). Construct the ANOVA table.
(2). Calculate and interpret the coefficient of the determination $R^2$.
(3). Test $H_0 : \beta_1 = \beta_2 = 0$ at the $\alpha = 0.05$ significance level.
(4). Explain the meaning of the coefficient $\beta_1$. Find the statistic and its sampling distribution for an individual test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

[Given: $F(0.05, 2, 17) = 3.59$]

Solution to Problem [10]: Stat 481 - MS Exam, Spring 2007 - Wang

(1). ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>66</td>
<td>33</td>
<td>4.19</td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>134</td>
<td>7.88</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(2). The coefficient of the determination

$$R^2 = \frac{SSR}{SSTO} = 33\%$$

It means that there are 33% of variation of the response explained by the linear regression model, or the variation in the data is reduced by 33% after introducing the two predictors.

(3) The $F$-statistic in the ANOVA table $F = 4.19 > F(0.05, 2, 17) = 3.59$. Therefore, we will reject the null hypothesis at level 0.05, i.e. at least one predictor contribute significantly to explain the variation in the response.

(4). The regression coefficient $\beta_1$ is the effect on mean response for a unit increase in predictor variable $X_1$, while holding the other predictor $X_2$ constant.

Use $t$-test for individual test $H_0 : \beta_1 = 0$ and the statistic is

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ and $s(\hat{\beta}_1)$ are the least square estimate and its standard error of linear coefficient $\beta_1$. Under $H_0$, the test statistic $t_{\hat{\beta}_1} \sim t(n - p - 1)$ i.e. $t_{\hat{\beta}_1} \sim t(17)$. 
