Statistics 2000, Section 001, Final (300 Points)

December 12, 2001, Dr. Jürgen Symanzik

Your Name: _____

First look at all 6 questions. Then start with the question that looks easiest to you. Continue with a more difficult question. Try to answer as many questions as possible in these 110 minutes.

Note that you will obtain at least partial credit if you indicate a correct formula but your final result is incorrect. If you just rely on your calculator without indicating the formula that should be used and your result is incorrect, you will obtain no credit at all for this part of a question.

Question 1: Short Answers (**60 Points**)

1. Let $z_i = \frac{x_i - \overline{x}}{s}, i = 1, \ldots, n$, be the z–scores for $n \geq 2$ arbitrary numbers $x_1, x_2, \ldots, x_n$ that are not all equal. Which of the following statements is correct: (**10 Points**)

   (a) The correlation coefficient r between x and z must be negative since I subtract $\overline{x}$ from each x_i.

   (b) The correlation coefficient r between x and z is somewhere between -0.99 and -0.01.

   (c) The correlation coefficient r between x and z is somewhere between 0.01 and 0.7.

   (d) The correlation coefficient r between x and z is somewhere between 0.7 and 0.99.

   only (e) The correlation coefficient r between x and z is exactly +1.

   (f) The correlation coefficient r between x and z is exactly 0.

   (g) The correlation coefficient r between x and z is exactly -1.

   each z_i is a linear transformation of x_i; so the points must fall on a straight line; since the slope is $\frac{1}{s} > 0$ (s is > 0 itself), the line must be increasing; (e) is the only correct answer

   -6 if more than 1 answer, including (e)

2. Determine the slope and the y–intercept of the lines whose equations are given as: **(12 Points)**

(a) $3x - 8y = 4$ $\Leftrightarrow -8y = 4 - 3x$

$\Leftrightarrow y = -\frac{1}{2} + \frac{3}{8}x$

-2 if "-" sign missing

-2 if slope = a·x

Slope = $\frac{3}{8}$ (3)

y–intercept = $-\frac{1}{2}$ // (3)

(b) $4y + x + 5 = 0$ $\Leftrightarrow 4y = -5 - x$

$\Leftrightarrow y = -\frac{5}{4} - \frac{1}{4}x$

Slope = $-\frac{1}{4}$ (3)

y–intercept = $-\frac{5}{4}$ // (3)

3. A foreign lottery has a game called "4 out of 59". You make a selection of 4 numbers between 1 and 59 and you win the big prize if exactly these numbers are drawn in the weekly drawing. How many different combinations are possible to select 4 out of 59 numbers? Calculate this value! **(10 Points)**

combinations: $\binom{59}{4} = \frac{59!}{4!(59-4)!} = \frac{59!}{4!\,55!}$

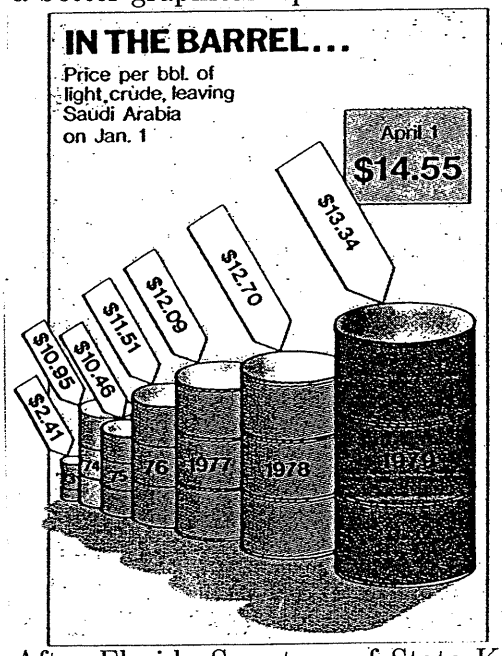
$= \frac{59 \cdot 58 \cdot 57 \cdot 56 \cdot 55!}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 55!}$

$= 455,126$ //

-5 if not the correct (exact) integer value

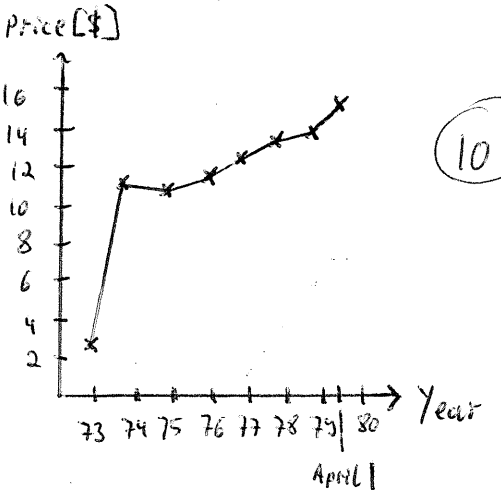
6

4. What is wrong with this graphic from "Time" (April 9, 1979, p. 57), reprinted in Edward R. Tufte's book "The Visual Display of Quantitative Information". Provide a better graphical representation of the same data. **(16 Points)**



wrong: volume (a 3-dimensional representation) has been used to represent a 1-dim variable (price) (6)

better graphic:



(10)

-3 at April 1, 79 missing

5. After Florida Secretary of State Katherine Harris certified the Florida results in the 2000 presidential election as a 537–vote Bush win, a Salt Lake City TV station held a call–in poll. The station asked its viewers to call in with their answer to the question:

"Do you believe Al Gore should stop trying to overturn legally certified votes in Florida and acknowledge that he has lost?"

4,237 people phoned in, and 3,482 said "yes". The station then announced that "82% of Americans believe that Al Gore should concede defeat."

Identify at least three problems with this survey and the announced results. **(12 Points)**

Problems: (4) each

i, call-in poll = voluntary response sample, i.e., only people that have a strong opinion are likely to be calling

ii, people living in or near Salt Lake City not representative of all "Americans"

iii, no margin of error given

iv, probability reported is not necessarily identical to probability obtained from answers, i.e.,

$P(\text{stop trying to overturn votes } \underline{\text{AND}} \text{ acknowledge loss}) = 0.82 \neq P(\text{acknowledge loss})$

✓ suggestive wording

Question 2: Normal Distribution (**50 Points**)

Part I:

1. Let Z be a standard Normal variable, i.e., $Z \sim N(0,1)$, and X be a Normal variable with mean $\mu = -4$ and variance $\sigma^2 = 4$, i.e., $X \sim N(-4, 2^2)$. Determine the following: (**30 Points, i.e., 5 Points each**)

(a) $P(Z < -2.2) =$ 0.0139 //

- 3 if negative probability (but otherwise correct)

- 2 if prob. not obtained

(b) $P(-2.6 < Z < 1.6) =$ $0.9452 - 0.0047$

$= 0.9405$ //

(c) $P(X < -2.2) =$ $P\left(\frac{X-(-4)}{2} < \frac{-2.2-(-4)}{2}\right)$

$= P(Z < 0.9)$

$= 0.8159$ //

(d) $P(-2.6 < X < 1.6) =$ $P\left(\frac{-2.6-(-4)}{2} < \frac{X-(-4)}{2} < \frac{1.6-(-4)}{2}\right)$

$= P(0.7 < Z < 2.8)$

$= 0.9974 - 0.7580$

$= 0.2394$ //

(e) Find a number # so that $P(Z > \#) = 0.35$

$\Rightarrow P(Z \leq \#) = 0.65$ Starting at the body of the table, we find that # falls between 0.38 and 0.39

(f) Find a number # so that $P(X > \#) = 0.35$

Since $Z = \frac{X-(-4)}{2} = \frac{X+4}{2}$, we can solve this equation for X and get

$X = 2Z - 4 \approx \begin{cases} 2 \cdot 0.38 - 4 \\ 2 \cdot 0.39 - 4 \end{cases} \approx \begin{cases} -3.24 \\ -3.22 \end{cases}$

So # falls between -3.22 and -3.24 in this case

Part II:

The rate of return on stock indexes (which combine many individual stocks) is approximately Normally distributed. Since 1945, the Standard & Poor's 500 index has had a mean yearly return of 11.8%, with a standard deviation of 16.6%. Take this Normal distribution to be the distribution of yearly returns over a long period. $X \sim N(11.8, 16.6^2)$

1. In what range do the middle 95% of all yearly returns lie? **(7 Points)**

$P(-\tilde{\#} \le Z \le \tilde{\#}) = 0.95$ — −3 if incorrect use of empirical rule

$\Leftrightarrow P(Z \le -\tilde{\#}) = \frac{0.05}{2} = 0.025$

From body of the table, $\tilde{\#} = 1.96$ and $-\tilde{\#} = -1.96$

Since $Z = \frac{X-11.8}{16.6}$, it is $X = 16.6Z + 11.8$, i.e., $\# = 16.6 \cdot (1.96) + 11.8 = 44.34$ //

$-\# = 16.6 \cdot (-1.96) + 11.8 = -20.74$ //

2. The market is down for the year if the return on the indexes is less than zero. In what percent of years is the market down? **(7 Points)**

Alternatively, by empirical rule:

$\mu \pm 2 \cdot 5 = 11.8 \pm 2 \cdot 16.6$

$= (-21.4, 45.0)$

$P(X \le 0) = P\left(\frac{X-11.8}{16.6} \le \frac{0-11.8}{16.6}\right)$

$= P(Z \le -0.71)$

$= 0.2389$ //

i.e., in about 23.9%

3. In what percent of years does the index gain 25% or more? **(6 Points)**

$P(X \ge 25) = P\left(\frac{X-11.8}{16.6} \ge \frac{25-11.8}{16.6}\right)$

$= P(Z \ge 0.795)$ — −3 if not 1−P(...)

$= 1 - P(Z < 0.795)$

$\approx 1 - 0.7866$

≈ 0.2134 //

i.e., in about 21.3%

Question 3: Linear Regression & Correlation (**50 Points**)

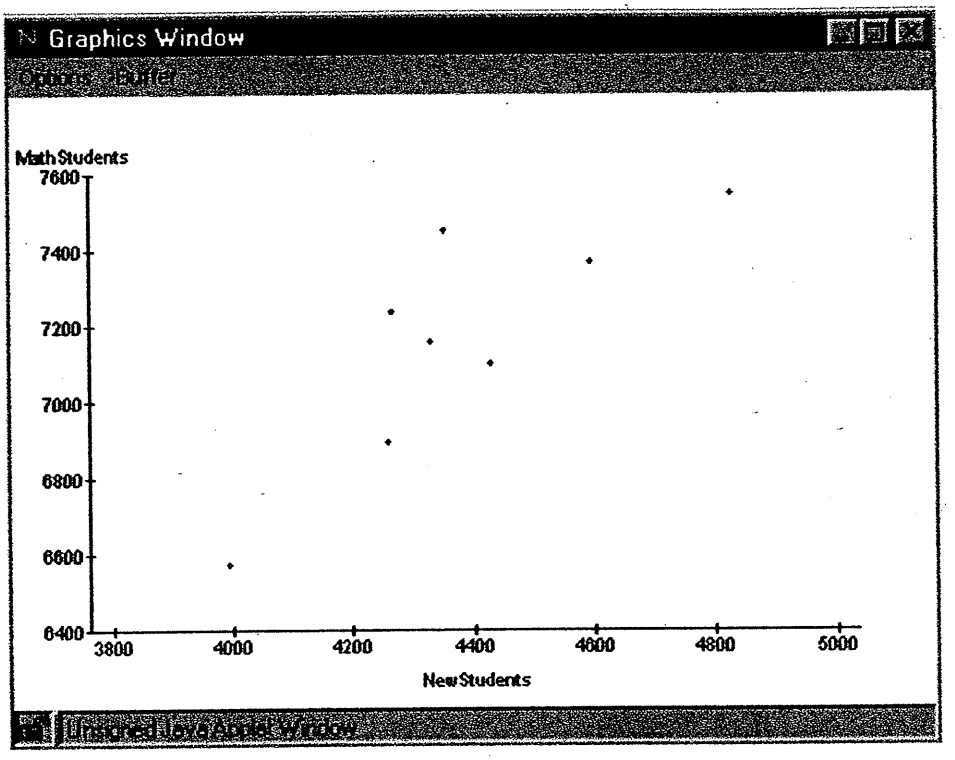
The mathematics department of a large state university would like to use the number of freshman entering the university (NewStudents $= x$) to predict the number of students who will sign up for freshman–level math courses (MathStudents $= y$) in the fall semester. Data for the years 1986 through 1993 are given below as they appear in the WebStat main window. The fourth data column shows the Residuals related to the simple linear regression equation obtained on the next page.

N WebStat 2.0

WebStat Data Stat Graphics Help

	Year	NewStudents	MathStudents	Residuals
1	1986	4595	7364	-20.442984
2	1987	4827	7547	-92.82977
3	1988	4427	7099	-114.30083
4	1989	4258	6894	-139.09235
5	1990	3995	6572	-180.64957
6	1991	4330	7156	46.13244
7	1992	4265	7232	191.44839
8	1993	4351	7450	317.7397

The scatterplot of entering students (NewStudents) vs students taking freshman–level math courses (MathStudents) is shown below:



We used WebStat to fit a simple linear (least squares) regression to the data. Here is the numerical output:

```
Results:

Simple linear regression results:

Independent variable: NewStudents
Dependent variable: MathStudents
Sample size: 8
Correlation coefficient: 0.8333
(See fitted line plot in Graphics Panel.)
Residuals stored in column Residuals

Estimate of sigma: 188.94853

Parameter      Estimate     Std. Err.    DF   Tstat        Pval
Intercept      2492.6917    1267.1991    6    1.9670876    0.0484
NewStudents    1.0663223    0.2888466    6    3.691656     0.0051
```

1. Indicate the exact values for slope, y–intercept, the regression equation, and the correlation coefficient obtained from WebStat.

slope: $b_1 = 1.0663223$ (3)

y-intercept: $b_0 = 2492.6917$ (3)

regression equation: $\hat{y} = b_0 + b_1 x = 2492.6917 + 1.0663223 \cdot x$ (3)

correlation: $r = 0.8333$ (3)

Provide an interpretation of Pearson's correlation coefficient r between entering students and students taking freshman–level math courses for this given data set. **(15 Points)**

There is a reasonably strong linear positive relationship between NewStudents and MathStudents. In a scatterplot the data points fall close to a straight rising line. (3)

2. Based on your equation in (1.), what is the predicted number of students taking freshman–level math courses when the number of entering students is 4,200?

$x = 4200 \Rightarrow \hat{y} = 2492.6917 + 1.0663223 \cdot 4200$

≈ 6971 students (6)

And how many students taking freshman–level math courses would you predict when the number of entering students is 10,000? Explain! **(15 Points)** -6 if only predicted

If $x = 10,000$, $\hat{y} = 2492.6917 + 1.0663223 \cdot 10000 \approx 13,156$ students.

However, 10,000 is twice as much as the highest number of admissions ever was. So, this is a considerable extrapolation and the predicted value of 13,156 is basically useless. The university may have changed its admission rules and this may effect the number of students taking Mathcourses in an unpredictable way. (9)

-3 if no clear statement whether useful or useless predicted value

3. Draw 2 different residual plots for this data set. One should contain the lurking variable. Is there any visible pattern in any of the your 2 residual plots? If there is a pattern, desribe the pattern. Finally, argue whether the least squares regression line describes the relationship between entering students and students taking freshman–level math courses reasonably well. **(20 Points)**

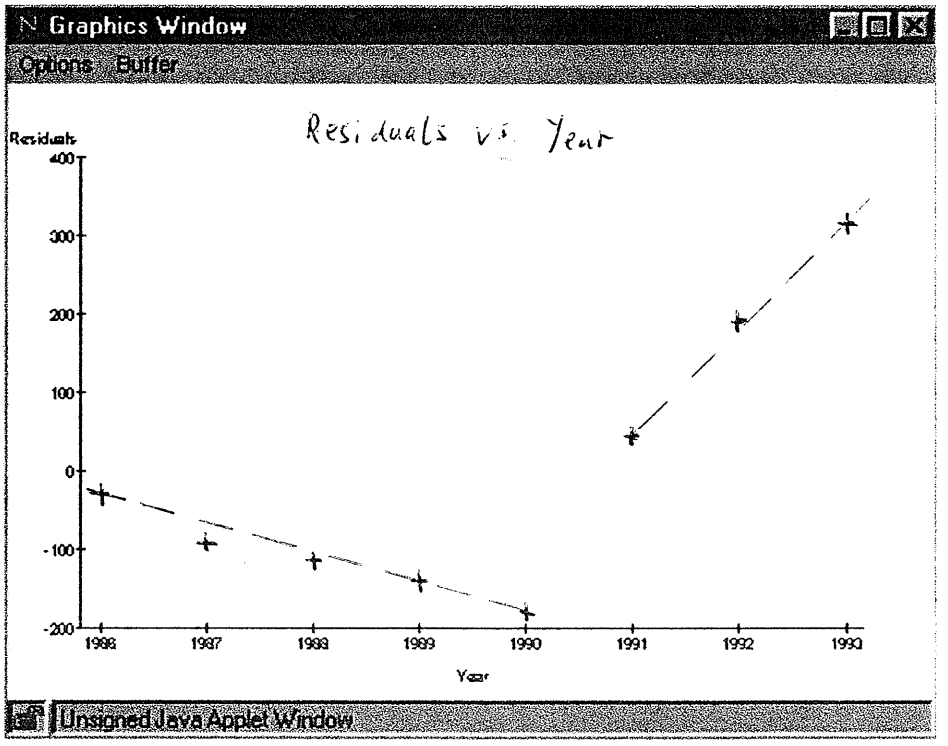
see next page

3. Since we know one additional (lurking) variable (Year), we should definitely plot the residuals e_i vs the Year.

The second residual plot should be a plot of e_i vs x_i, $\hat{y}_i$, or y_i.

Here are 2 residual plots: (-5) if Residuals vs. Year not plotted

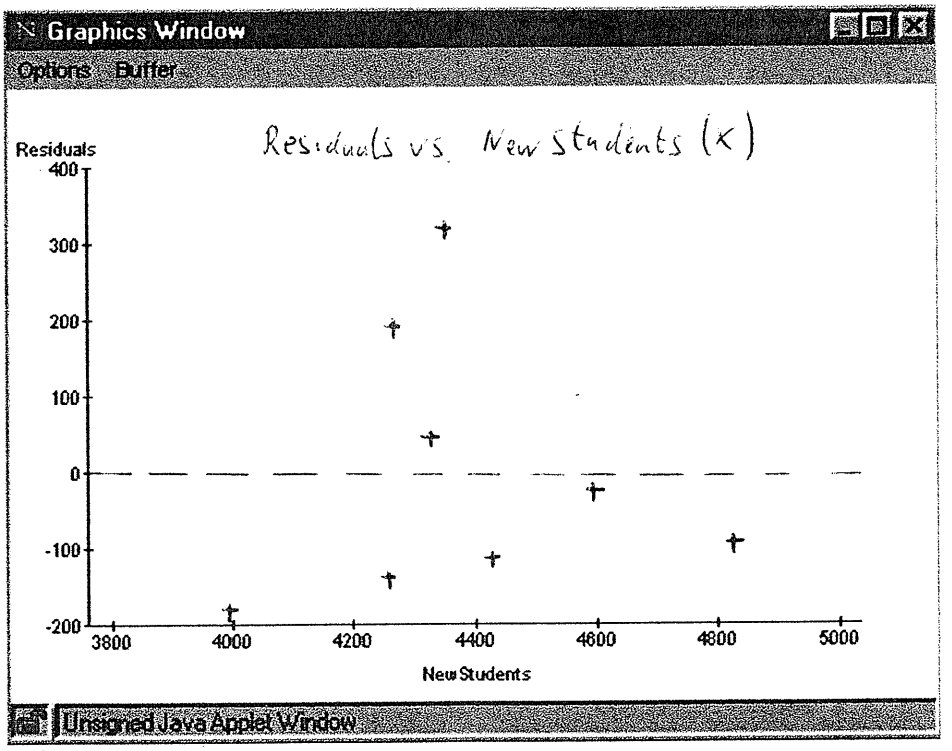
(5)



The plot Residual vs. Year shows a decrease of the residuals between 1986 to 1990 (all $<<0$) and an increase of the residuals between 1991 to 1993 (all $>>0$). (5)

The plot Residuals vs New Students shows highest ($>>0$) residuals for New Students between 4265 and 4351 (which relates to the years 1991 to 1993) and low ($<<0$) residuals everywhere else.

(5)



Overall, these residual plots suggest that the regression line does not describe the relationship very well. It seems that a structural change occurred in 1991. Could it be that some departments have changed their requirements with respect to Math courses, i.e., ordering their students to take additional Math courses ??

Question 4: Probability Distributions (**50 Points**)

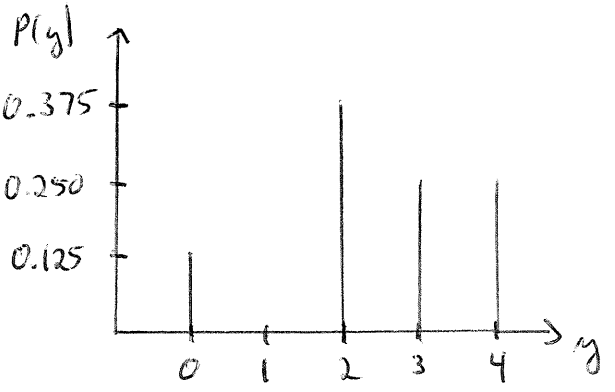
In a recent Statistics class, the following grade points (representing full–letter grades) have been reported for the 40 students that participated in this class:

# students	grade points y	$p(y)$ (5)→1.	$F(y) = P(Y \le y)$ (5)→2.	$y \cdot P(y)$ (5)→4.	$y^2 \cdot P(y)$ (5)→5.
5	0.0	$5/40 = 0.125$	0.125	0	0
0	1.0	$0/40 = 0$	0.125	0	0
15	2.0	$15/40 = 0.375$	0.500	0.75	1.50
10	3.0	$10/40 = 0.25$	0.750	0.75	2.25
10	4.0	$10/40 = 0.25$	1.000	1.00	4.00
40					

We introduce a random variable Y that represents the grade points for this class.

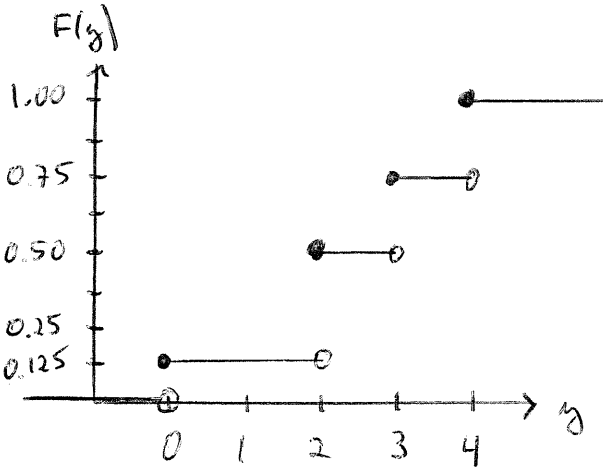
1. Add the values for $p(y)$, i.e., the probability distribution for Y, to the table above and draw a spike graph of $p(y)$. (**10 Points**)

(5)



2. Add the values for $F(y)$, i.e., the cumulative probability function (or cumulative distribution function, cdf) for Y, to the table above and draw a graph of $F(y)$. (**15 Points**)

(10)



−2 if part below 0 missing

3. Indicate $P(Y \leq 3.0)$? **(5 Points)**

$$P(Y \leq 3.0) = \frac{30}{40} = 0.75$$

4. Calculate $\mu = E(Y)$, i.e., the mean (expected value) of Y. **(10 Points)**

$$\mu = E(Y) = \sum_i y_i P(y_i) = 0 + 0 + 0.75 + 0.75 + 1.00$$
$$= 2.50 \qquad (5)$$

5. Calculate $\sigma^2 = Var(Y)$, i.e., the variance of Y. **(10 Points)**

$$\sigma^2 = Var(Y) = \sum_i y_i^2 P(y_i) - \mu^2$$
$$= 0 + 0 + 1.50 + 2.25 + 4.00 - (2.50)^2$$
$$= 7.75 - 6.25$$
$$= 1.50 \qquad (5)$$

Question 5: Binomial Probability Distributions (**50 Points**)

In a previous quiz, 23 out of 25 Business Stat students answered a particular exam question correctly. Let us assume that in the upcoming academic year all new Business Stat students have to answer this same question in one of their quizzes. Based on previous years, the upper enrollment limit of 300 students has been obtained each year so we can safely assume that there will be exactly 300 students again during the next year that will attend this class.

Let X be the random variable that describes the number of students that will correctly answer this question in the next year. It is safe to asssume that students' ability to answer a particular question does not change over time.

1. Indicate the probability distribution of X. Complete the following formula that relates to the probability distribution of X: (**5 Points**)

$X \sim \text{Bin}(n, p)$

$= \text{Bin}\left(300, \frac{23}{25}\right)$

$= \text{Bin}(300, 0.92)$

- 3 if Bin (or B) missing

- 2 if wrong parameters

2. Calculate the mean of X ($\mu = E(X)$) and the variance of X ($\sigma^2 = Var(X)$). (**15 Points**)

$\mu = E(X) = n \cdot p = 300 \cdot 0.92 = 276$ (7.5)

$\sigma^2 = Var(X) = n \cdot p(1-p) = 300 \cdot 0.92 \cdot 0.08 = 22.08$ (7.5)

$\sigma = \sqrt{n \cdot p(1-p)} = \sqrt{22.08} \approx 4.70$

3. What is the **exact** probability that 290 students will answer this question correctly? Be as efficient in your calculations as possible. **(10 Points)**

$$P(X = 290) = \binom{300}{290} 0.92^{290} \cdot 0.08^{10} \quad (6)$$

$$= \frac{300!}{290!\ 10!} \cdot 0.92^{290} \cdot 0.08^{10}$$

$$= \frac{300 \cdot 299 \cdot 298 \cdot \ldots \cdot 292 \cdot 291}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot 9 \cdot 10} \cdot 0.92^{290} \cdot 0.08^{10} \quad (2)$$

$$\approx (1.398 \cdot 10^{18}) \cdot 0.92^{290} \cdot 0.08^{10}$$

$$\approx 4.73 \cdot 10^{-4} \approx 0.000473, \text{ i.e. around } 0.047\% \quad (2)$$

4. What is the probability that 200 through 280 students will answer this question correctly? You may use an **approximation** if **appropriate** but need to check the required conditions first. **(20 Points)**

Check: $n \cdot p = 300 \cdot 0.92 = 276 \geq 10$ ✓

$n \cdot (1-p) = 300 \cdot 0.08 = 24 \geq 10$ ✓ (5)

We can use the De Moivre-Laplace limit theorem:

$$X \overset{\text{approx}}{\sim} N\left(np, \left(\sqrt{n \cdot p \cdot (1-p)}\right)^2\right)$$

$$= N\left(276, \left(\sqrt{22.08}\right)^2\right) \quad (5)$$

$$P(200 \leq X \leq 280) = P\left(\frac{200-276}{\sqrt{22.08}} \leq \frac{X-276}{\sqrt{22.08}} \leq \frac{280-276}{\sqrt{22.08}}\right)$$

$$= P(-16.17 \leq Z \leq 0.85)$$

$$= 0.8023 - 0$$

$$= 0.8023, \text{ i.e. around } 80.2\%. \quad (10)$$

−2 for each incorrect table value

Question 6: Tests of Significance (**40 Points**)

1. Nutrition experts recommend that one's daily diet contain a minimum of 20 grams of fiber. The director of a summer camp for teenagers wants to show that the camp provides meals that exceed this amount. What null and alternative hypotheses should be tested? (**6 Points**)

$H_0: \mu = 20$ vs $H_A: \mu > 20$

-2 if H_0 & H_A flipped
-3 if only H_0 or H_A

2. Suppose a competitor believes that the camp's claim is an exaggeration, and that the camp is not meeting the nutritional needs of its participants with regard to fiber content. In particular, the competitor suspects that teenagers are provided a daily average of fewer than 20 grams of fiber. What set of hypotheses would the competitor be interested in testing? (**6 Points**)

$H_0: \mu = 20$ vs $H_A: \mu < 20$

3. Suppose an unbiased third party is only interested in determining if the camp's mean daily amount of fiber differs from 20 grams. It has no preconceived notion as to whether the actual mean is more or less than this figure. What set of hypotheses would this independent group want to test? (**6 Points**)

$H_0: \mu = 20$ vs $H_A: \mu \neq 20$

4. Suppose the unbiased third party is examining the fiber contents of meals on an annual basis. From previous years, it is known that the standard deviation is $\sigma = 1.5$. Based on a sample of 5 meals, the unbiased third party obtains a sample mean $\overline{x} = 21.8$ grams of fiber.

   Indicate the test statistic and calculate the p–value for the set of hypotheses specified in (3.). Can you reject H_0 at the 5% level of significance? (**22 Points**)

Test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{21.8 - 20}{1.5/\sqrt{5}} = 2.68$$ (8)

p-value:

$$p = 2 \cdot P(z \leq -2.68) = 2 \cdot 0.0037 = 0.0074 < 0.05$$ (7)

Reject H_0 at the 5% level of significance – there is strong evidence that the food does not contain 20 grams of fiber. (7)