

Data Analysis Exam #1

Write your answers on separate paper. You may do the problems in any order, just identify them so that we can find your answers. Remember that a few problems may be very difficult and that there may be easier problems later in the exam. The goal is to maximize your points, not to get every question correct.

Problem #1

This semester we have a large contingent of students from IBG who are interested in pharmacological research. They have persevered despite some frustrating problems getting computer connections between IBG and samiam. To reward this perseverance, we have a models question relevant to pharmacological research. The questions involve important model comparisons that are relevant in everyone's research.

A lot has been learned about drug receptors by using antagonists that block the effect of other drugs. In one kind of study, a given concentration B of an antagonist is given and then the dose of the primary drug (the agonist) is increased to produce some specified level of effect. How much the primary drug had to be increased to produce the same effect is known as the drug ratio or DR. You will be happy that we are skipping a lot of chemical kinetic equations and cutting directly to the chase: If molecules of the drug and the antagonist compete one-for-one for receptor binding sites, then

$$\log(DR - 1) = K + \log(B)$$

where K is a constant to be estimated from the data. We will simplify this equation by assuming we have the following two variables available for analysis:

$$\text{LOGDR} = \log(\text{DR}-1)$$

$$\text{LOGB} = \log(B)$$

Use these two variables, with perhaps other variables you define based on them, to specify the Model A/Model C comparisons you would use to answer the following questions. The analysis of such drug data has the special name of "Schild analysis" in the drug literature, but the following questions demonstrate that it is nothing more than specific Model A/Model C comparisons.

A. The competition model says that we ought to be able to predict LOGDR with LOGB. Specify the models that test that assertion.

B. The one-for-one competition model says that the slope ought to be one, while slopes greater than one suggest that multiple drugs are binding to the same receptors and slopes less than one might suggest agonist uptake processes. What model comparison asks whether the slope differs from one?

C. The constant K is based on the "equilibrium dissociation constant for the antagonist-receptor complex." We don't have to know what that means. What is important is that there are often theoretical means for predicting what that constant ought to be, and hence what K ought to be. Suppose theory predicted that for a particular drug and antagonist, that $K = 0.3$. What models would test this prediction?

D. Terrence Kenakin in his book *Pharmacologic Analysis of Drug-Receptor Interaction* recommends that "it is preferable to use as wide a range [of concentrations of the antagonist B] as possible." What is the statistical basis for this recommendation?

E. Kenakin also warns that "small random changes in tissue sensitivities have greater effects on low as opposed to high dose ratios [low vs. high DR]." In other words, noise might have greater effects on the observed DR at low values than at high values. What possible violation of a statistical assumption is he warning us about?

Problem #2

Data were gathered on life expectancy statistics from the 40 largest countries in the world for the year 1990. The following variables were available on each country:

FLIFEXP	Life expectancy of Females
MLIFEXP	Life expectancy of Males
TVPER	Number of televisions per 100 people
DOCPER	Number of physicians per 100 people

From the first two variables, the following additional variable was computed:

$$\text{FMDIF} = \text{FLIFEXP} - \text{MLIFEXP}$$

Various univariate and simple regression statistics were generated by SAS with the output on the following pages. Use these results to answer the following questions.

A. Examine the univariate statistics and the box and quantile - quantile plots from SAS/INSIGHT for FLIFEXP. Write one or two sentences (no more!) that discuss how the distribution of this variable departs from a normal distribution.

B. What is the 95% confidence interval for the mean value of FMDIF? Write two sentences (no more!) that tell us what this confidence interval means.

C. Test the null hypothesis that female life expectancy is no longer than male life expectancy in these data. That is, is the difference between female and male life expectancies different from zero? (Provide Models C and A in their "beta" forms and in their estimated forms, SSE(A) and SSE(C), PRE, F*, a statistical conclusion, and a one sentence news summary.)

D. A goal of the United Nations has been to bring life expectancy up to 70 years. Test the null hypothesis that this has been achieved for females in these countries. (Provide Models C and A in their "beta" forms and in their estimated forms, SSE(A) and SSE(C), PRE, F*, a statistical conclusion, and a one sentence news summary.)

E. Provide brief interpretations for the intercept and the slope in the model where FLIFEXP is regressed on TVPER.

F. Examine the bivariate regression plot from SAS/INSIGHT for the FLIFEXP by TVPER relationship. Write a sentence or two that summarizes your reaction to this plot

G. Can we reliably predict FLIFEXP from DOCPER (note: not TVPER)? (Provide Models C and A in both "beta" and estimated forms, SSE(A) and SSE(C), PRE, F*, and a one sentence news summary.)

H. A subset of these 40 countries are in Africa. A researcher wants to examine the relationship between FLIFEXP and DOCPER only for these African nations. Give two reasons why this test will have less power for detecting the relationship than the test in (G) above.

I. In the context of a model where predictions of FLIFEXP are made conditional on the number of physicians per 100 people, test whether female life expectancy differs from the U.N. goal of 70 years. (Provide Models C and A, SSE(A) and SSE(C), PRE, F*, a statistical conclusion, and a one sentence news summary.)

J. Explain why the test in (I) is or is not more powerful than the test in (D).

The SAS System
Univariate Procedure

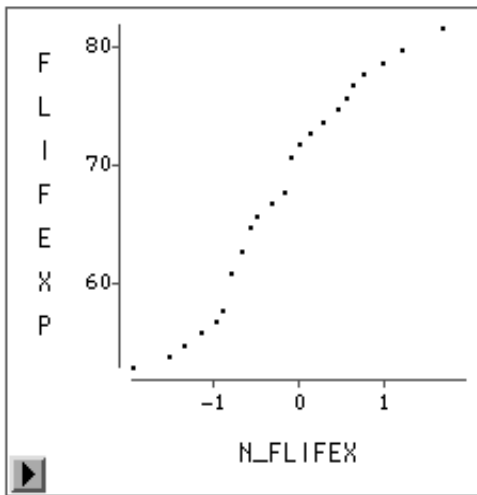
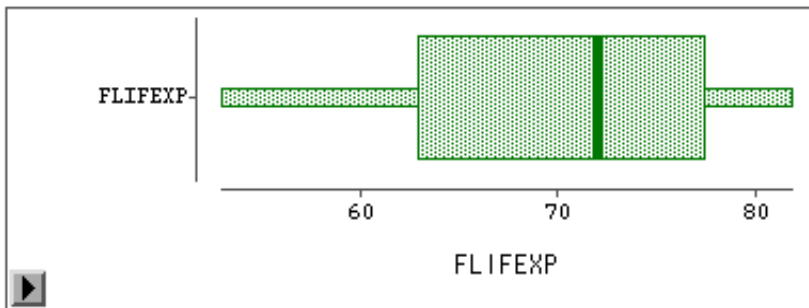
Variable=FLIFEXP

Moments

N	40	Sum Wgts	40
Mean	69.575	Sum	2783
Std Dev	9.162038	Variance	83.94295
Skewness	-0.42316	Kurtosis	-1.01873
USS	196901	CSS	3273.775
CV	13.16858	Std Mean	1.448645
T:Mean=0	48.02762	Pr> T	0.0001
Num ^= 0	40	Num > 0	40
M(Sign)	20	Pr>= M	0.0001
Sgn Rank	410	Pr>= S	0.0001

Quantiles(Def=5)

100% Max	82	99%	82
75% Q3	77.5	95%	82
50% Med	72	90%	81
25% Q1	63	10%	55.5
0% Min	53	5%	53.5
		1%	53
Range	29		
Q3-Q1	14.5		
Mode	67		



The SAS System
Univariate Procedure

Variable=MLIFEXP

Moments

N	40	Sum Wgts	40
Mean	64.5	Sum	2580
Std Dev	7.421383	Variance	55.07692
Skewness	-0.40788	Kurtosis	-0.75042
USS	168558	CSS	2148
CV	11.50602	Std Mean	1.173424
T:Mean=0	54.96736	Pr> T	0.0001
Num ^= 0	40	Num > 0	40
M(Sign)	20	Pr>= M	0.0001
Sgn Rank	410	Pr>= S	0.0001

Quantiles(Def=5)

100% Max	76	99%	76
75% Q3	70	95%	75
50% Med	66	90%	73.5
25% Q1	59.5	10%	52.5
0% Min	50	5%	51
		1%	50
Range	26		
Q3-Q1	10.5		
Mode	68		

The SAS System
Univariate Procedure

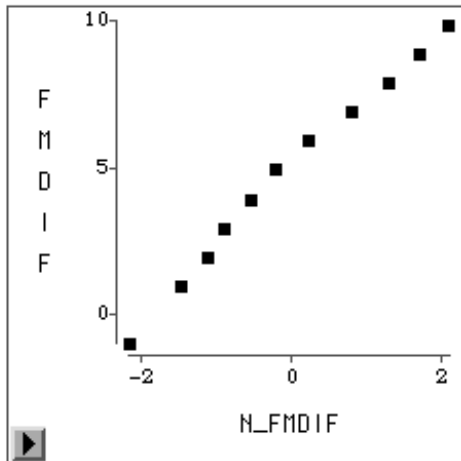
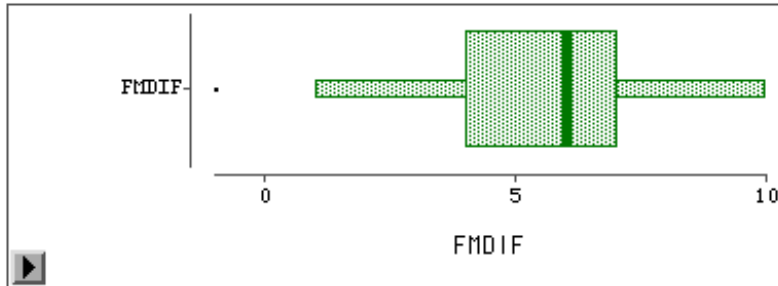
Variable=FMDIF

Moments

N	40	Sum Wgts	40
Mean	5.075	Sum	203
Std Dev	2.421988	Variance	5.866026
Skewness	-0.47967	Kurtosis	0.016649
USS	1259	CSS	228.775
CV	47.7239	Std Mean	0.38295
T:Mean=0	13.25239	Pr> T	0.0001
Num ^= 0	40	Num > 0	39
M(Sign)	19	Pr>= M	0.0001
Sgn Rank	407	Pr>= S	0.0001

Quantiles(Def=5)

100% Max	10	99%	10
75% Q3	7	95%	8.5
50% Med	6	90%	8
25% Q1	4	10%	1
0% Min	-1	5%	1
		1%	-1
Range	11		
Q3-Q1	3		
Mode	6		



Moments			
N	40.0000	Sum Wgts	40.0000
Mean	5.0750	Sum	203.0000
Std Dev	2.4220	Variance	5.8660
Skewness	-0.4797	Kurtosis	0.0166
USS	1259.0000	CSS	228.7750
CU	47.7239	Std Mean	0.3829

Quantiles			
100% Max	10.0000	99.0%	10.0000
75% Q3	7.0000	97.5%	9.5000
50% Med	6.0000	95.0%	8.5000
25% Q1	4.0000	90.0%	8.0000
0% Min	-1.0000	10.0%	1.0000
Range	11.0000	5.0%	1.0000
Q3-Q1	3.0000	2.5%	0
Mode	6.0000	1.0%	-1.0000

Confidence Interval for Mean			
Mean	Level (%)	Lower Limit	Upper Limit
5.0750	95.0000	4.3004	5.8496

The SAS System
Univariate Procedure

Variable=TVPER

Moments

N	38	Sum Wgts	38
Mean	19.18839	Sum	729.1589
Std Dev	18.22724	Variance	332.2323
Skewness	1.26345	Kurtosis	1.631166
USS	26283.98	CSS	12292.6
CV	94.99097	Std Mean	2.956849
T:Mean=0	6.489474	Pr> T	0.0001
Num ^= 0	38	Num > 0	38
M(Sign)	19	Pr>= M	0.0001
Sgn Rank	370.5	Pr>= S	0.0001

Quantiles(Def=5)

100% Max	76.92308	99%	76.92308
75% Q3	31.25	95%	58.82353
50% Med	15.90909	90%	38.46154
25% Q1	4.347826	10%	1.041667
0% Min	0.168919	5%	0.198807
		1%	0.168919
Range	76.75416		
Q3-Q1	26.90217		
Mode	38.46154		

The SAS System
Univariate Procedure

Variable=DOCPER

Moments

N	40	Sum Wgts	40
Mean	0.139549	Sum	5.581969
Std Dev	0.124679	Variance	0.015545
Skewness	0.888642	Kurtosis	0.002427
USS	1.385211	CSS	0.606252
CV	89.34429	Std Mean	0.019714
T:Mean=0	7.078858	Pr> T	0.0001
Num ^= 0	40	Num > 0	40
M(Sign)	20	Pr>= M	0.0001
Sgn Rank	410	Pr>= S	0.0001

Quantiles(Def=5)

100% Max	0.442478	99%	0.442478
75% Q3	0.215525	95%	0.407642
50% Med	0.101026	90%	0.326327
25% Q1	0.030497	10%	0.01055
0% Min	0.002728	5%	0.004138
		1%	0.002728
Range	0.43975		
Q3-Q1	0.185028		
Mode	0.27027		

The SAS System

Model: MODEL1
 Dependent Variable: FLIFEXP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1679.80147	1679.80147	51.405	0.0001
Error	36	1176.40906	32.67803		
C Total	37	2856.21053			

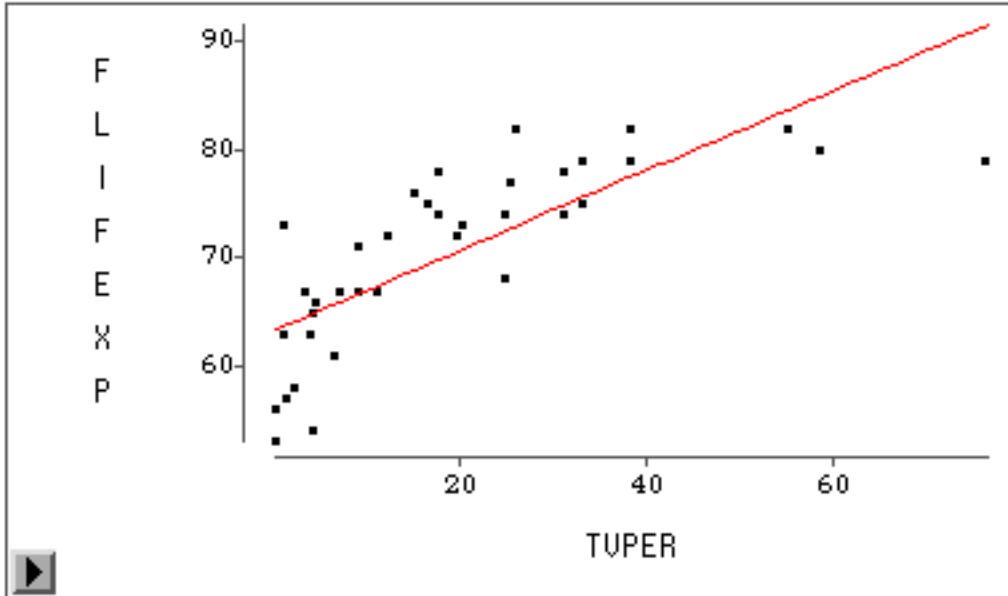
Root MSE	5.71647	R-square	0.5881
Dep Mean	70.31579	Adj R-sq	0.5767
C.V.	8.12971		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	63.222531	1.35600075	46.624	0.0001
TUPER	1	0.369664	0.05155920	7.170	0.0001

Model Equation

FLIFEXP = 63.2225 + 0.3697 TUPER



Dependent Variable: FLIFEXP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1745.24120	1745.24120	43.387	0.0001
Error	38	1528.53380	40.22457		
C Total	39	3273.77500			
Root MSE	6.34228	R-square	0.5331		
Dep Mean	69.57500	Adj R-sq	0.5208		
C.V.	9.11575				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	62.087642	1.51581851	40.960	0.0001
DOCPER	1	53.653890	8.14552739	6.587	0.0001