PUBH5769 (Biostatistics II)

Description of datasets used in the course

Throughout this course, you will use several example datasets that are based on real datasets.
A file of type filename.sas is a SAS program file that contains SAS syntax (ie commands) and sometimes may also include data.
A file of type filename.dat is a plain text data file in free-field format.
A file of type filename.sas7bdat is a SAS dataset.

The specific files provided to you (downloadable from LMS) are listed below. This document provides a brief description of each dataset (including a list of the variables in the dataset) and the SAS Proc Means summary of the dataset.

Data set	**Files**
Diabetes data	diabet.dat diabet.sas
Community survey data	survey.sas7bdat
US city air pollution data	so2city.sas7bdat
BCG case-control data	bcg.sas7bdat
Endometrial cancer case-control data	endomet.sas7bdat
Grouped smoking death cohort data	smokdth.sas7bdat
Lymphoma survival data	lymphoma.sas7bdat

Diabetes dataset

The diabetes dataset relates to baseline and mortality follow-up data for a cohort of 498 persons with diabetes.

AGE	age in years
SEX	0 = female 1 = male
DURATION	duration of diabetes in years
TREAT	1 = insulin injections 2 = tablets 3 = special diet
PDW	percent of desirable weight
SBP	systolic blood pressure in mmHg
HAEM	glycosylated haemoglobin in mmol/L
DIED5	0 = have not died within 5 years 1 = died within 5 years

Variable	N	Mean	Std Dev	Minimum	Maximum
age	498	63.873	9.550	40.000	80.000
sex	498	0.554	0.498	0.000	1.000
duration	498	8.092	6.426	1.000	39.000
treat	498	1.859	0.672	1.000	3.000
pdw	498	123.594	14.336	76.600	149.900
sbp	498	152.442	21.560	98.000	198.000
haem	498	10.652	2.022	6.300	18.800
died5	498	0.215	0.411	0.000	1.000

Note that the mean of died5 = proportion of values equal to 1.

Community survey data

This dataset (adapted from real data) relates to baseline (in 1981) and mortality follow-up data (to 1995) for the 1,552 participants aged 40-69 years who participated in a Western Australian health survey in 1981. The variables in the **survey.sas7bdat** SAS dataset are:

AGE	age in years
ALCGRAMS	alcohol consumed per week in grams
ANGINA	ever had angina (0 = no, 1 = yes)
ASTHMA	ever had asthma (0 = no, 1 = yes)
BMI	body mass index (weight/height^2) in kg/m^2
BRONCH	ever had bronchitis (0 = no, 1 = yes)
CHOL	blood cholesterol in mmol/L
CIGSDAY	number of cigarettes smoked per day
CVDCENS	died from cardiovascular disease (0 = no, 1 = yes)
DBP	diastolic blood pressure in mmHg
DIABETES	have diabetes (0 = no, 1 = yes)
DRINKING	alcohol consumption categories (1 = never, 2 = ex, 3= < 20 grams/day 4 = 20 – 60 grams/day, 5 = > 60 grams/day)
DTHCENS	died during follow-up (0 = no, 1 = yes)
DYSPNOEA	get shortness of breath (0 = never, 1 = hurrying or walking up a hill 2 = 1 and when walking with people of same age on level ground 3 = 1, 2 and when walking at my own pace on level ground)
EXERCISE	number of days exercise per week
FEV	forced expiratory volume in 1 second in L
FVC	forced vital capacity in L
HAYFEVER	ever had hayfever (0 = no, 1 = yes)
HEIGHT	height in m
MARITAL	marital status (1 = single, 2 = divorce, widowed, separated, 3 = married or de facto)
MYOCARD	ever had myocardial infarction (0 = no, 1 = yes)
OCCUP	occupation (1 = professional, 2 = farmer, 3 = manual 4 = home duties/unemployed/pensioner)
RXHYPER	on treatment for hypertension (0 = no, 1 = yes)
SBP	systolic blood pressure in mmHg
SEX	sex (0 = male, 1 = female)
SMOKING	smoking status (1 = never, 2 = ex, 3 = <15 cigarettes/day, 4 = 15+ cigarettes/day)
STENCHD	coronary heart disease (0 = no, 1 = possible, 2 = definite)
SURVTIME	follow-up time from survey date in years
WEIGHT	weight in kg
YEARSMOK	years of smoking
GPVISITS	Number of visits to GP in last 12 months

Variable	N	Mean	Std Dev	Minimum	Maximum
age	1552	55.816	8.486	40.020	69.980
alcgrams	1552	89.470	156.437	0.000	2450.000
angina	1552	0.037	0.188	0.000	1.000
asthma	1552	0.075	0.263	0.000	1.000
bmi	1552	26.100	3.963	16.800	45.300
bronch	1552	0.193	0.395	0.000	1.000
chol	1552	6.151	1.169	3.010	14.410
cigsday	1552	3.233	7.968	0.000	70.000
cvdcens	1552	0.050	0.219	0.000	1.000
dbp	1552	79.259	11.857	30.000	178.000
diabetes	1552	0.025	0.157	0.000	1.000
drinking	1552	2.624	1.081	1.000	5.000
dthcens	1552	0.114	0.318	0.000	1.000
dyspnoea	1552	0.387	0.752	0.000	3.000
exercise	1552	3.110	2.932	0.000	7.000
fev	1552	2.623	0.818	0.300	5.900
fvc	1552	3.490	0.961	1.000	7.200
hayfever	1552	0.189	0.392	0.000	1.000
height	1552	1.665	0.086	1.440	1.920
marital	1552	2.860	0.415	1.000	3.000
myocard	1552	0.017	0.128	0.000	1.000
occup	1552	3.229	1.134	1.000	4.000
rxhyper	1552	0.188	0.390	0.000	1.000
sbp	1552	131.182	19.392	84.000	223.000
sex	1552	0.557	0.497	0.000	1.000
smoking	1552	1.800	1.004	1.000	4.000
stenchd	1552	0.348	0.639	0.000	2.000
survtime	1552	13.368	2.459	0.340	14.120
weight	1552	72.558	13.258	37.800	126.000
yearsmok	1552	14.135	16.764	0.000	56.000
gpvisits	1552	1.104	0.911	0.000	5.000

US city air pollution data

These data on 40 cities come from an air pollution study in the USA. The SO2 level (micrograms/cubic metre), average annual temperature in F, the number of factories employing at least 20 persons, the population in thousands, and the average annual rainfall in inches were obtained for the 41 cities. The variables in the **so2city.sas7bdat** SAS dataset are CITY, SO2, TEMP, FACT, POP and RAIN. Note that City = Dallas has no SO2 value so SAS will not use Dallas for fitting any models.

Variable	N	Mean	Std Dev	Minimum	Maximum
so2	40	30.58	23.53	8.00	110.00
temp	41	55.76	7.23	43.50	75.50
fact	41	462.37	563.18	35.00	3344.00
pop	41	608.61	579.11	71.00	3369.00
rain	41	36.77	11.77	7.05	59.80

BCG case-control data

These grouped data come from a (stratified) case-control study of the association between BCG vaccination and leprosy. All male and female cases of leprosy aged 10 – 39 years were examined for the presence or absence of a BCG scar. The cases were grouped according to age into categories 10 – 19, 20 – 29 and 30 – 39 years. Controls were selected at random from the 'healthy' population so that, for each sex, there were four times as many controls as cases in each age group. The controls were also examined for the presence or absence of a BCG scar.

		Cases			Controls		
	Age	BCG Scar			BCG Scar		
Sex	Group	Absent	Present	All	Absent	Present	All
Men	10 – 19	24	28	52	63	145	208
	20 – 29	32	16	48	96	96	192
	30 – 39	25	3	28	92	20	112
Women	10 – 19	20	22	42	50	118	168
	20 – 29	24	14	38	80	72	152
	30 – 39	22	3	25	82	18	100

This is a stratified case-control study involving 6 strata defined by combinations of Sex and Agegroup. The variables in the **bcg.sas7bdat** SAS dataset are stratum (a number from 1 to 6 corresponding to the six age by sex strata), sex (1 = men, 2 = women), agegroup (1 = 10-19, 2 = 20-29, 3 = 30-39), bcg (0 = no, 1 = yes), cases, total (ie cases+controls).

Variable	N	Mean	Std Dev	Minimum	Maximum
stratum	12	3.5000000	1.7837652	1.0000000	6.0000000
sex	12	1.5000000	0.5222330	1.0000000	2.0000000
agegroup	12	2.0000000	0.8528029	1.0000000	3.0000000
bcg	12	0.5000000	0.5222330	0	1.0000000
cases	12	19.4166667	9.0398947	3.0000000	32.0000000
total	12	97.0833333	44.2214640	21.0000000	173.0000000

Endometrial cancer case-control data

This is a very old but well known matched 1-to-1 case-control study of risk factors for endometrial cancer. The investigators identified 63 cases of endometrial cancer occurring in a retirement community in Los Angeles from 1971 to 1975. Each case was matched to a control living in the retirement village at the time the case was diagnosed. The controls were the same age as the case, the same marital status as the case, entered the retirement community at approximately the same time as the case and who had not had a hysterectomy prior to the time the case was diagnosed (and who was therefore still at risk for the disease). The variables in the dataset **endomet.sas7bdat** are:

AGEGROUP	0 = < 70 years of age 1 = 70 + years of age
CASE	0 = No (ie control) 1 = Yes
ESTROGEN	history of oestrogen use 0 = No 1 = Yes
GALLBD	history of gall bladder disease 0 = No 1 = Yes
HYPERTEN	history of hypertension 0 = No 1 = Yes
MATCHSET	matched set (a number from 1 to 63)

Variable	N	Mean	Std Dev	Minimum	Maximum
MATCHSET	126	32.0000000	18.2568343	1.0000000	63.0000000
CASE	126	0.5000000	0.5019960	0	1.0000000
AGEGROUP	126	0.5555556	0.4988877	0	1.0000000
GALLBD	126	0.2063492	0.4062996	0	1.0000000
HYPERTEN	126	0.3730159	0.4855368	0	1.0000000
ESTROGEN	126	0.6825397	0.4673464	0	1.0000000

Grouped smoking death cohort data

This data set contains grouped data on death rates for a cohort of male employees in a certain industry by smoking status at start of employment.

	Non-smokers		Smokers	
Age Group	Deaths	Person-yrs	Deaths	Person-yrs
35 – 44	2	18,790	32	52,407
45 – 54	12	10,673	104	43,248
55 – 64	28	5,710	206	28,612
65 – 74	28	2,585	186	12,663
75 – 84	31	1,462	102	5,317
Total	101	39,220	630	142,247

The variables in the dataset **smokdth.sas7bdat** are agegroup (1 to 5), smoker (0 = no, 1 = yes), deaths, pyr (total person-years of follow-up) and logpyr, ie ln(pyr).

Variable	N	Mean	Std Dev	Minimum	Maximum
agegroup	10	3.0000000	1.4907120	1.0000000	5.0000000
smoker	10	0.5000000	0.5270463	0	1.0000000
deaths	10	73.1000000	73.4218405	2.0000000	206.0000000
pyr	10	18146.70	17762.40	1462.00	52407.00
logpyr	10	9.2739756	1.1856071	7.2875606	10.8667954

Lymphoma survival data

This dataset contains the survival times (ie time from diagnosis to death) in months for a cohort of newly diagnosed lymphoma patients. There are two groups defined by different types of presenting symptoms.

The variables in the dataset **lymphoma.sas7bdat** are symptoms (1 or 2), months (follow-up time in months), and censor (0 indicates that the patient has not died by the end of follow-up whereas a censor value of 1 indicates that the patient has died).

Variable	N	Mean	Std Dev	Minimum	Maximum
symptoms	64	1.4843750	0.5037065	1.0000000	2.0000000
months	64	35.3437500	23.1529344	2.5000000	75.7000000
censor	64	0.4531250	0.5017331	0	1.0000000