Statistics 104 – Autumn 2004

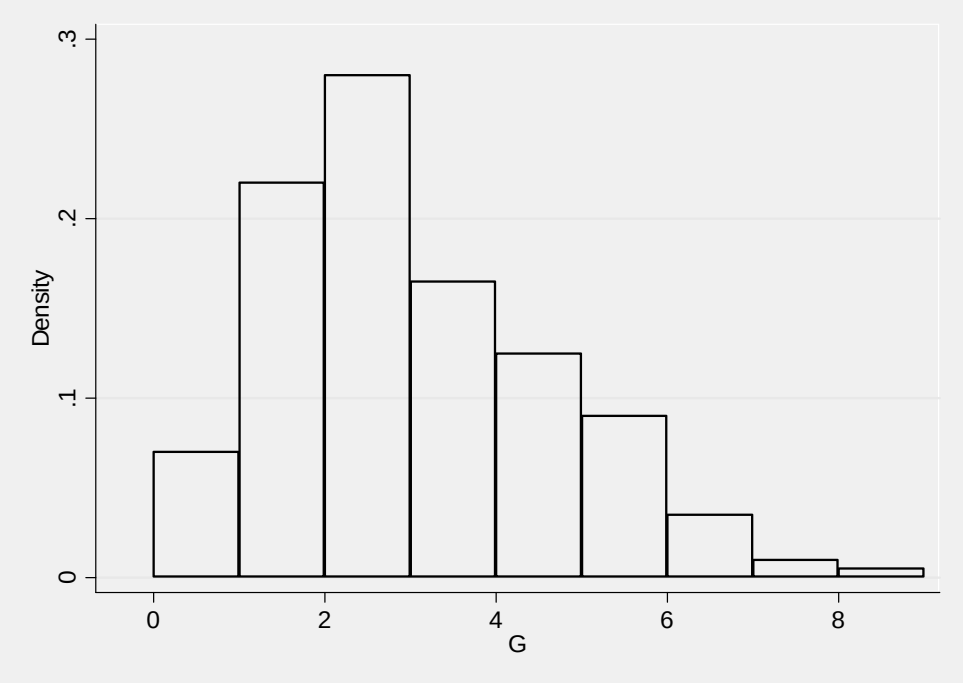
Practice Midterm Examination 1 Solutions

1. (10 points) Indicate which of the following statements are true and briefly, for each of the others, show why they are false. You may simply correct the given statement as a way of showing why.

   a) (2 points) If you recode a set of observations on the variable X by $Z = 2 - 3X$, then the mean of Z will be $\overline{Z} = 2 - 3\overline{X}$ and the standard deviation will be $s_Z = 2 - 3s_X$.

   False. It should be $s_Z = 3s_X$. Note that it can't be $s_Z = -3s_X$ since standard deviations must be positive

   b) (2 points) In the following histogram, the median is approximately equal to the mean.



   False. Since the distribution is skewed right, the median will be less than the mean.

   c) (2 points) For a data set which is approximately normally distributed, we would expect to find about 3 out of 1000 observations more than two standard deviations from the mean.

   False. There are a number of ways that the statement could be corrected. The two most popular were:

   about 3 out of 1000 observations more than **three** standard deviations from the mean

or

about **50** out of 1000 observations more than two standard deviations from the mean.

d) (2 points) A correlation of zero between two variables implies that there is no relationship between them.

False. $r = 0$ implies there is no **linear** relationship. You could still have a strong nonlinear relationship. For example, let X take values -100, -99, …, 99, 100 and let Y be X^2. If you were to calculate r for this dataset, you would get exactly even though there is a strong nonlinear relationship in the data.

e) (2 points) A correlation of –1 means the points falls on a straight line and one variable can perfectly predict the other.

True. A correlation of 1 or -1 can only occur when all the points fall on a straight line and that line can be determined by regression. In this situation, all points in the dataset will fall on this exactly. A number of people commented that the statement is false since $r = -1$ doesn't mean there is proof of causation (which is true) and if is no causation you can't predict (definitely false). Much of applied statistics is based on making predictions based on observed correlations. For example, it has been shown there is a strong association between median teacher salaries (in dollars) and sales of alcoholic beverages (in dollars) in year. Even though increasing teacher salaries doesn't cause alcohol sales to go up (at least no very much) you can still make useful prediction of what might happen to alcohol sales if teacher salaries changed by a certain amount. You can do this since there are a number of common factors (such as the inflation rate) which drive both variables.

2. (12 points) The data below are annual average CO_2 readings from the Mauna Loa Observatory in Hawaii over the years 1980 – 1988.

Year	CO_2
1980	338.4
1981	339.5
1982	340.8
1983	342.8
1984	344.3
1985	345.7
1986	346.9
1987	348.6
1988	351.2

A linear regression model was fit to the data with the output below.

```
regress CO2 Year

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =  907.55
       Model |    146.0164     1    146.0164           Prob > F      =  0.0000
    Residual |  1.12623777     7  .160891109           R-squared     =  0.9923
-------------+------------------------------           Adj R-squared =  0.9913
       Total |  147.142638     8  18.3928297           Root MSE      =  .40111

------------------------------------------------------------------------------
         CO2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Year |   1.560002   .0517834    30.13   0.000     1.437554     1.68245
       _cons |    -2750.8   102.7383   -26.77   0.000    -2993.737   -2507.862
------------------------------------------------------------------------------
```

a) (3 points) Use the regression equation to predict the average CO_2 for 1980.

fit = 1.56 * 1980 -2750.8 = 338

b) (3 points) What is the residual for 1980?

residual = obs – fit = 338.4 – 338 = 0.4

c) (2 points) What is the numerical value of the correlation coefficient *r*? Please give your answer rounded to three decimal places, i.e. 0.755 or –0.623.

$r = \sqrt{r^2} = \sqrt{0.9923} = 0.996$

d) (4 points) Suppose you were asked to predict CO_2 for the year 2025 based on the above regression output. Is this likely to be a good prediction? Why or why not?

This prediction is likely to be poor. This would be an example of extrapolation, where prediction is being done outside the range of the data. It could be that the relationship between CO_2 and time could change (which should happen if the Kyoto accords have any effect). Also any model used for prediction is only approximate. If you get outside the range of your data, that approximation could break down, leading to poor predictions.

3. (13 points) The data in the table and the output below are taken from a study entitled, "Smoking During Pregnancy and Lactation and Its Effects on Breast Milk Volume" (*American Journal of Clinical Nutrition*, 1991, 1011-1016). The data give milk volume, expressed in grams per day (g/day). The purpose of the study was to determine whether the amount of breast milk that a mother can produce is affected by smoking cigarettes.

Smokers	621	793	593	545	753	655	895	767	714	598	693
Nonsmokers	947	945	1086	1202	973	981	930	745	903	899	961

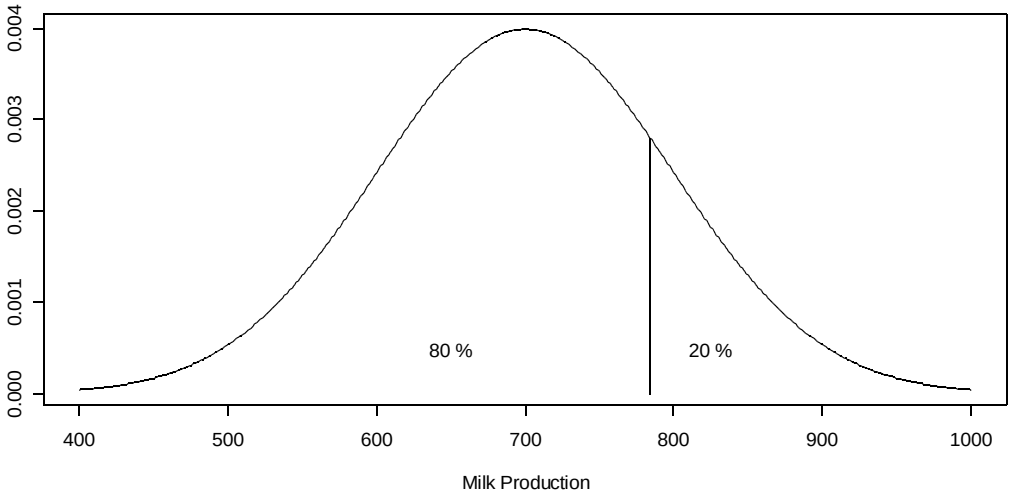
```
Variable     N   Mean    Median  StDev    Min      Max      Q1       Q3
Smokers     11  693.4     693.0  103.9  545.0    895.0   598.0    767.0
Nonsmokers  11  961.1     XXXXX  113.8  745.0   1202.0   903.0    981.0
```

a) (3 points) Give the median and interquartile range for the **Non-smoking** mothers.

Median = 6^{th} ordered value = 947 g/day

IRQ = Q3 – Q1 = 981 – 903 = 78 g/day

b) (4 points) Assuming that the data for smokers follows a normal distribution with mean μ = 700 and standard deviation s = 100, how much milk was supplied by the mothers with the top 20% in volume.



Want x^* s.t. P[$X = x^*$] = 0.2, which is equivalent to P[$X = x^*$] = 0.8

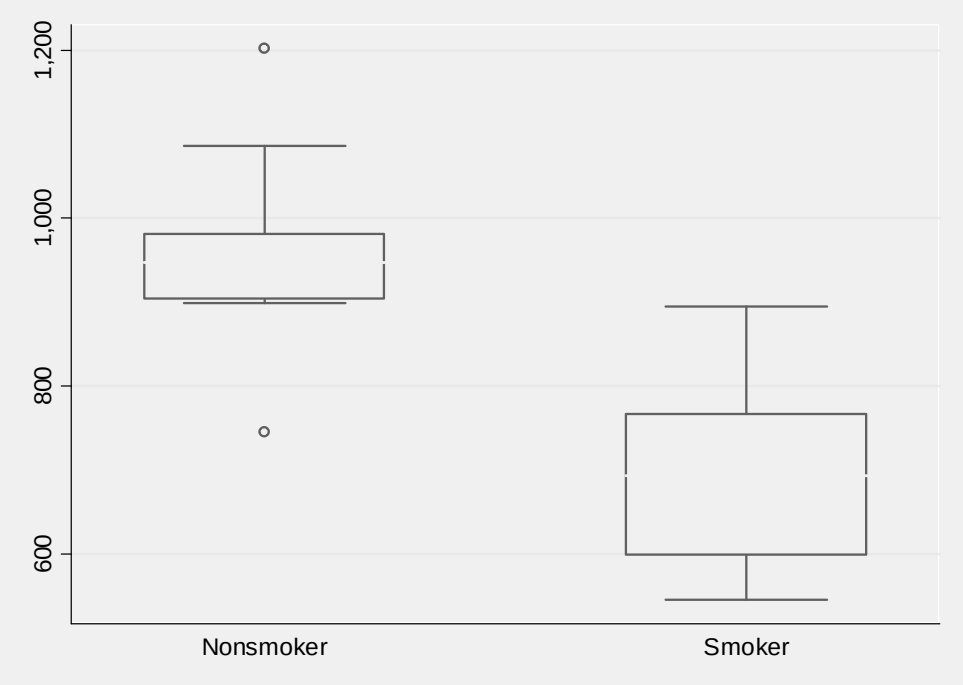
For a standard normal P[$Z = 0.84$] = 0.8 so $x^* = 100 \times 0.84 + 700 = 784$

So the top 20% of milk producers supplied at least 784 g/day

c) (3 points) Suppose that the observation in the smoking group of 545 g/day (the smallest observed) was recorded incorrectly as 45 g/day. How would this error effect the summary statistics? (You do not need to do any calculations, just describe generally what will change and how.)

If 45 g/day were used instead of 545 g/day, the calculated mean would be less (by 45.45 g/day = 500/11) and the standard deviation would be larger (by 116 g/day – you need to calculate the standard deviation for both versions of the datasets to figure this out). The median, Q1, Q3, and IQR would be unchanged.

d) (3 points) Describe in two or three sentences the conclusions that he investigators might draw from this study. Below is a side by side box plot of the data.



Smoker tend to produce much less milk than non-smokers, 267.7 g/day on average. In fact, for this data set, all but one non-smoker produces more milk than the smokers.

4. (15 points) In an experiment on the behaviour of young children, each child is placed in an area with four toys. The response of interest is the number of toys that the child plays with. Past experiments with many children have shown that the probability distribution of the number X of toys played with is as follows:

X	0	1	2	3	4
$P[X = x] = p_x$	0.05	0.15	???	0.30	0.10

a) (2 points) What is the probability that a child plays with exactly 2 toys?

$P[X = 2] = 1 - (0.05 + 0.15 + 0.30 + 0.10) = 0.4$

b) (2 points) What is the probability that a child plays with less than 2 toys?

$P[X < 2] = P[X = 1] = 0.05 + 0.15 = 0.2$

c) (3 points) What is the probability that a child plays with no toys, given that you are told that the child played with less than 2 toys?

$P[X = 0 \mid X < 2] = P[X = 0 \ \& \ X < 2] / P[X < 2] = [X = 0] / P[X < 2] = 0.05 / 0.2 = 0.25$

d) (3 points) Assuming children are independent, what is the probability that 2 children both play with 3 toys?

$P[X_1 = 3 \ \& \ X_2 = 3] = P[X_1 = 3] \times P[X_2 = 3] = 0.3^2 = 0.09$

e) (5 points) Find the mean μ of X.

$\mu = 0 \times 0.05 + 1 \times 0.15 + 2 \times 0.4 + 3 \times 0.3 + 4 \times 0.1 = 0 + 0.15 + 0.8 + 0.9 + 0.4 = 2.25$