

Stat 3011 Midterm 1

Problem 1

(a) The sample mean is

$$\frac{15 + 8 + 30 + 41 + 33 + 150 + 61 + 46 + 38 + 44}{10} = 46.6$$

(b) The numbers in sorted order are

8 15 30 33 38 41 44 46 61 150

The median is the middle number if the sample size is odd. Here the sample size (10) is even, and the median is the average of the middle two numbers $(38 + 41)/2 = 39.5$.

(c) To find the 10% trimmed mean, throw 10% of the numbers (10% of 10 is one) from each end (of the sorted list) and calculate the mean of the rest

$$\frac{15 + 30 + 33 + 38 + 41 + 44 + 46 + 61}{8} = 38.5$$

(d) There are two good answers, one involving the notion of “outlier” and the other involving the notion of “asymmetry.”

Answer 1: The \$150,000 income is very different from the rest. If this is considered an “outlier” a robust estimate of center like the median or a trimmed mean should be used instead of the mean, because robust estimates are less sensitive to outliers.

Answer 2: Incomes have a distribution with a long left tail so the mean, median, and trimmed mean estimate different quantities. The median is a better indication of the “typical” income. Note that half of the numbers are below the median (by definition), but 80% of the numbers are below the mean. That’s not very representative of the “typical” income. A similar analysis applies to the trimmed mean, which is close to the median and far from the (untrimmed) mean.

Problem 2

Correlation only indicates that random variables tend to vary together. It cannot indicate which causes which. More precisely, if x and y are correlated random variables, this is consistent with three different scenarios

- x causes y ,
- y causes x , and

- y and x are both caused by some third variable z .

There is no way to distinguish among the three cases from correlation alone.

Regression is just correlation in another guise, as is apparent from the formula

$$b = r \frac{s_y}{s_x}$$

for the slope of the regression line. Thus regression is no more useful than correlation at indicating causal relationships among variables. The slogan for this is “regression is for prediction, not explanation.”

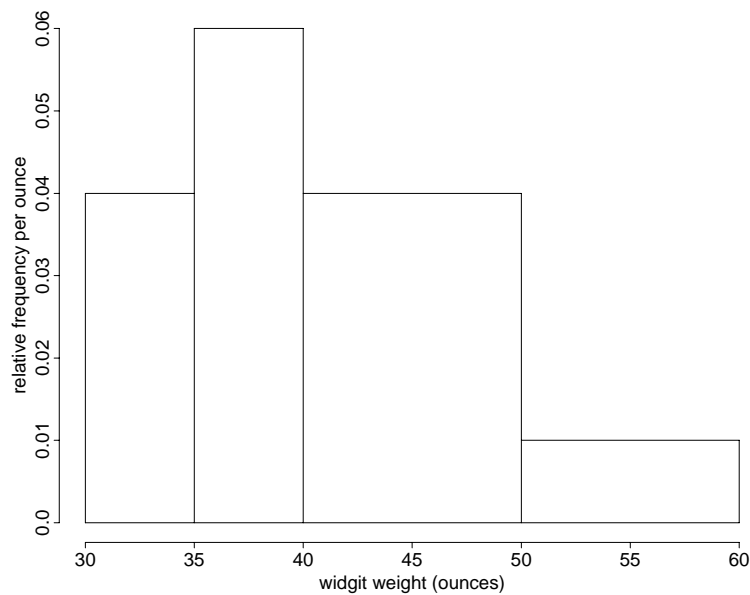
Problem 3

(a) The answers to this part are given in the first four columns of the table below. The last two columns relate to the next part.

Class	Frequency	Cumulative		Width	Height
		Relative Frequency	Relative Frequency		
30 – < 35	20	0.20	0.20	5	0.04
35 – < 40	30	0.30	0.50	5	0.06
40 – < 50	40	0.40	0.90	10	0.04
50 – < 60	10	0.10	1.00	10	0.01

(b) Because the bin widths are not all equal we must use the rule height = relative frequency/width, giving the heights in the last column of the table in part (a).

The histogram is



Problem 4

(a) The first thing to note is that x , the predictor variable, is the score on the first midterm, thus $\bar{x} = 85$ and $s_x = 6$, and that y , the response variable, is the score on the second midterm, thus $\bar{y} = 83$ and $s_y = 6$. If you get this backwards, you, of course, mess up the whole problem.

The equations needed to calculate the regression coefficients here are

$$b = r \frac{s_y}{s_x}$$
$$a = \bar{y} - b\bar{x}$$

plugging in the numbers in the problem gives

$$b = 0.6 \cdot \frac{8}{6} = 0.8$$
$$a = 83 - 0.8 \times 85 = 15$$

Thus the least-squares regression equation is

$$y = a + bx = 15 + 0.8 \cdot x$$

(b) The predicted value of y for an x value of 80 is

$$y = 15 + 0.8 \times 80 = 79$$

(c) The coefficient of determination $r^2 = 0.6^2 = 0.36$.

Problem 5

(a) By the multiplication rule

$$\begin{aligned} P(\text{four last more than 1000 hours}) &= P(\text{one lasts more than 1000 hours})^4 \\ &= .25^4 \\ &= 0.0039. \end{aligned}$$

(b) The addition rule is not appropriate because it is possible for several light bulbs to last more than 1000 hours.

Thus we try the complement rule. The outcome “at least one lasts more than 1000 hours” is the complementary outcome of “four last less than 1000 hours”, and the probability of that outcome can be calculated by the multiplication rule

$$P(\text{four last less than 1000 hours}) = P(\text{one lasts less than 1000 hours})^4$$

if we knew the probability on the right hand side, but that can also be done using the complement rule. Thus we get done with two applications of the

complement rule and one application of the multiplication rule.

$$\begin{aligned}P(\text{one lasts less than 1000 hours}) &= 1 - P(\text{one lasts more than 1000 hours}) \\ &= 1 - 0.25 \\ &= 0.75\end{aligned}$$

Then

$$\begin{aligned}P(\text{four last less than 1000 hours}) &= P(\text{one lasts less than 1000 hours})^4 \\ &= 0.75^4 \\ &= 0.3164\end{aligned}$$

Finally

$$\begin{aligned}P(\text{at least one lasts more than 1000 hours}) \\ &= 1 - P(\text{four last less than 1000 hours}) \\ &= 1 - 0.3164 \\ &= 0.6836\end{aligned}$$

Problem 6

(a) Between $P(z < -2.345) = .0096$ and $P(z < -2.35) = .0094$, about half way between, which is .0095.

(b) $P(.67 < z < .89) = P(z < .89) - P(z < .67) = .8133 - .7486 = .0647$.

(c) This is the same question as “What is the z^* such that $P(z < z^*) = 0.20$ ”? We can extract the following from the normal distribution table

z^*	$P(z < z^*)$
-0.84	.2005
-0.85	.1977

We can see that the probability we want to look up is much closer to the top row than the bottom, so the z^* is perhaps -0.842 . With the computer we can do

```
qnorm(0.20)
[1] -0.8416212
```

and get the answer to six significant figures. (Of course, you couldn't do that during the exam.)