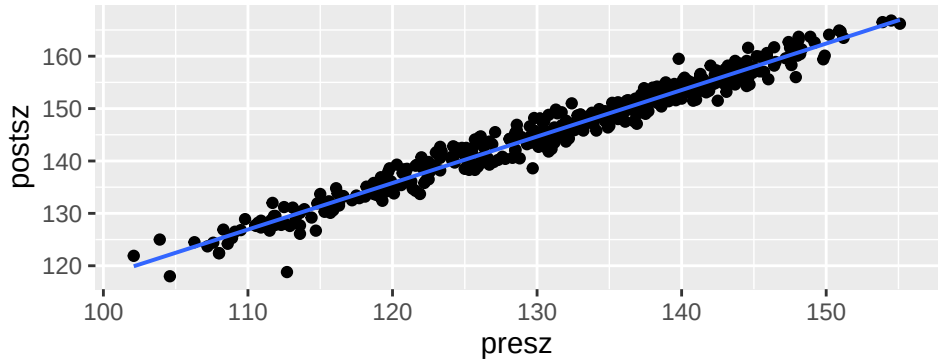
1. (a) One way to obtain a scatterplot with regression line:

```
gf_point(postsz ~ presz, data=bigCrabs) %>% gf_lm()
```



(b) Using the `lm()` command, we have

```
lm.result <- lm(postsz ~ presz, data=bigCrabs)
lm.result


Call:
lm(formula = postsz ~ presz, data = bigCrabs)

Coefficients:
(Intercept)        presz
    29.3082       0.8874
```

This shows the best-fit linear model to be

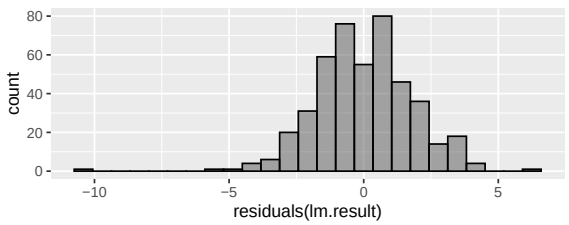
$$\widehat{\text{postsz}} = 29.31 + (0.89)(\text{presz}).$$

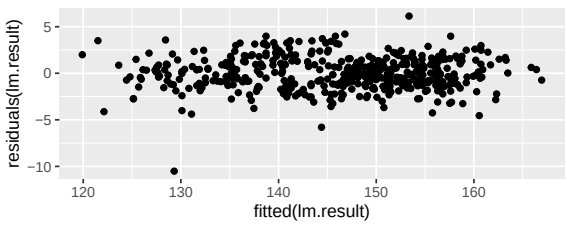
In particular, the slope is 0.89, indicating that for each additional millimeter in `presz` value, the `postsz` value increases an average of about 0.89 millimeter.

(c) The commands

```
gf_histogram( ~ residuals(lm.result), color="black" )
gf_point(residuals(lm.result) ~ fitted(lm.result))
```

produce a histogram of residuals (on the left) and a plot of residuals vs. fitted-values (on the right).





With the exception of a single extreme residual, the histogram appears to be plausibly normal centered at 0, and the plot on the right appears unpatterned in any alarming way. As far as we have investigated things, it appears the model assumptions hold.

(d) We have predicted `postsz`: $29.31 + (0.89)(125) = 140.56$ mm.

(e) The hypotheses of the model utility test may be stated in several ways. Perhaps the most general way is to write the null hypothesis as "no linear model with `presz` as explanatory variable is useful in predicting `postsz` values," with alternative that "such a model is useful." Two other ways to state it are

$$\mathbf{H}_0\colon\ \beta_1 = 0, \qquad \mathbf{H}_a\colon\ \beta_1 \neq 0,$$

and

$$\mathbf{H}_0\colon\ \rho = 0, \qquad \mathbf{H}_a\colon\ \rho \neq 0.$$

Taking the correlation

```
cor(postsz ~ presz, data=bigCrabs)
```

```
[1] 0.9840395
```

as our test statistic r, we obtain the following t-score

$$t \;=\; \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \;=\; \frac{(0.984)\sqrt{453-2}}{\sqrt{1-(0.984)^2}} \doteq 117.3$$

along with corresponding P-value

```
2 * (1 - pt(117.3, df=451))
```

```
[1] 0
```

which is so small as to be indistinguishable from 0. Hence, we reject the null hypothesis in favor of the alternative, that the model is a useful one for predicting `postsz` values.

(f) The correlation, as calculated above, is 0.984. Thus, the R^2 value is $(0.984)^2 = 0.968$, which means that almost 97% of the variation seen in `postsz` values has been explained by our model.

(g) From the output

```
summary(lm.result)$sigma     # extra "$sigma" reduces output to the one desired number
```

```
[1] 1.782577
```

we see the `Residual standard error` is 1.783. We wish to count how many residuals are below $(-2)\times(1.783)$ or above $2\times(1.783)$, and see if the ratio of this count to the total number of rows in `bigCrabs` is about 5%. A little more sophisticated (and direct) way to do this than sorting the data is this:

```
nrow(bigCrabs)      # counts the total number of big crabs
```

```
[1] 453
```

```
sum(abs(residuals(lm.result)) >= 2*1.783)  /453   # proportion of residuals within 2 RSEs
```

```
[1] 0.03752759
```

We have 3.8% of residuals outside 2 RSEs of 0, which is pretty close to 5%. Recall that

```
pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

or about 95.4%, of observations generally lie within 2 standard deviations of the mean (0 here), and we have come pretty close to that, with $1 - 0.038 = 0.962$.

Similarly, we expect about 68% of observed values to lie within a single standard deviation of the mean (0). In the data, the percentage is

```
sum(abs(residuals(lm.result)) <= 1.783) / 453
```

```
[1] 0.7019868
```

and 70.2% is fairly similar to 68%.

(h) *(5 pts)* Give a 90% prediction interval for crabs with premolt size 125 mm. Write an interpretation, in context, of this interval. We create our predictor function first, and then use it to get the desired prediction interval:

```
postszPredictor = makeFun(lm.result)
postszPredictor(presz = 125, interval="prediction", level=0.9)
```

```
       fit      lwr      upr
1 140.2279 137.2854 143.1703
```

From the output, we have 90% prediction interval $[127.29, 143.17]$. We expect (i.e., with 90% confidence) the next crab we see with a `presz` measurement of 125 mm to have a `postsz` measurement inside this interval.

(i) *(5 pts)* Since each row in `bigCrabs` provides multiple measurements on a single crab, the premolt and postmolt sizes may be viewed as **matched pairs** data. Produce a 95% confidence interval for the mean difference `postsz - presz` of these sizes among the bigger crabs. Following the process for a *matched-pairs t*-test, we look at mean and standard deviation for the **difference**, on a case-by-case basis, of `postsz` and `presz` measurements:

```
favstats( bigCrabs$postsz - bigCrabs$presz )
```

```
 min   Q1 median   Q3  max     mean       sd   n missing
 6.1 12.9   14.4 15.9 21.1 14.49735 2.175516 453       0
```

Using the appropriate critical value for sample size $n = 453$

```
qt(0.975, df=452)
```

```
[1] 1.965226
```

we have

```
14.497 + c(-1,1)*1.965*2.176/sqrt(453)
```

```
[1] 14.2961 14.6979
```

or a 95% CI $[14.3, 14.7]$.

2. (a) From

```
gradeCounts <- c(98,163,153,52,27)
sum(gradeCounts)
```

```
[1] 493
```

```
gradeCounts / 493
```

```
[1] 0.19878296 0.33062880 0.31034483 0.10547667 0.05476673
```

we have that the proportion of students who earned As is 0.199 (or 19.9%). For Bs, Cs, Ds and Fs the proportions are 0.331, 0.31, 0.105 and 0.548, respectively.

(b) Our hypotheses are these:

$\mathbf{H}_0$: $p_{As} = 0.2$, $p_{Bs} = 0.35$, $p_{Cs} = 0.25$, $p_{Ds} = 0.15$, $p_{Fs} = 0.05$

$\mathbf{H}_a$: At least one of these population proportions is unequal to its stated value.

Some relevant commands (though, perhaps, unlike the ones you use):

```
nullProportions = c(.2, .35, .25, .15, .05)
expectedCounts = nullProportions*493     # creates a vector of the expected counts
expectedCounts

[1]  98.60 172.55 123.25  73.95  24.65
```

From this we see the expected counts all exceed 5, so we feel justified using a chi-square distribution with $5 - 1 = 4$ degrees of freedom to calculate the approximate P-value.

```
sum( (gradeCounts - expectedCounts)^2 / expectedCounts )   # compute chi-sq stat

[1] 14.45253

1 - pchisq(14.45, df=4)     # compute corresponding P-value

[1] 0.005989055
```

This P-value is significant, even at the 1% level, so we reject the null hypothesis in favor of the alternative. Whether or not the professors at this university are shooting for Calc II grades to be distributed according to proportions put forth in the null hypothesis, we reject that they are succeeding.

(c) The population may be viewed as consisting of all possible grades the particular collection of professors who taught Calc II in that semester might assign to college students taking the course. In "allowing" 493 students to take the course and tabulating the grades they earned, we have sampled from this population.

3. (a) Our hypotheses:

$\mathbf{H}_0$: There is no association between "enthusiasm for video games" and a student's "expected grade"

$\mathbf{H}_a$: The two variables, "enthusiasm for video games" and "expected grade", are associated

There is no shame in calculating all the expected counts "by hand". But most any of the calculations we wish to do *can* be performed in RStudio. What follows are commands that achieve the various steps.

We build a table like the one provided:

```
freqTab <- as.table(rbind(c(10,14,6), c(13,32,14)))
rownames(freqTab) = c("A","B or C")
colnames(freqTab) = c("strongly like", "like somewhat", "disinterested")
```

The expected counts are among the output of the `chisq.test()` command. That command does more than we want at this point in that, by default, it uses a chi-square distribution to calculate P-values, and until we view the expected counts, we reserve judgment as to whether this is appropriate. Thus, we ask specifically for the expected counts, and avoid viewing the complete output.

```
chisq.test(freqTab)$expected

       strongly like like somewhat disinterested
A           7.752809      15.50562      6.741573
B or C     15.247191      30.49438     13.258427
```

All of these are at least 5, so we may proceed without using a randomization distribution. The χ^2 statistic (which, again, you may compute by hand):

```
chisq.test(freqTab)$statistic

X-squared
 1.326146
```

and, knowing $df = (3-1)(2-1) = 2$, we obtain the P-value

```
1 - pchisq(1.326, df=2)

[1] 0.5153031
```

We fail to reject the null hypothesis. The data is consistent with the null hypothesis that these two variables are not associated.

(b) The modified table looks like this:

Expected Grade	Strong liking	Less enthusiastic	Total
A	10	20	30
B or C	13	46	59
Total	23	66	89

We may test for an association between variables using a 2-proportion test. Using "enthusiasm for video games" as the explanatory variable, we may use p_S to denote the proportion of students, among those who strongly like video games, expecting an A in Calc II; similarly, p_L will denote the proportion of students, among those who are less enthusiastic, expecting an A. (As the notation suggests, these are population proportions.) The usual way, in a 2-proportion test, for us to write the "association" hypotheses is

$$\mathbf{H}_0\colon\ p_S - p_L = 0, \qquad \mathbf{H}_a\colon\ p_s - p_L \neq 0.$$

We have $\hat{p}_S = 10/23 \doteq 0.435$ and $\hat{p}_L = 20/66 \doteq 0.303$. The pooled proportion is

$$\hat{p} \;=\; \frac{10+20}{23+69} \;=\; \frac{30}{89} \;\doteq\; 0.337,$$

yielding standard error

$$\text{SE} \;=\; \sqrt{(0.337)(1-0.337)\left(\frac{1}{23}+\frac{1}{66}\right)} \;\doteq\; 0.114.$$

Thus, we obtain P-value by calculating areas in a normal distribution $\mathsf{Norm}(0, 0.114)$ associated with the test statistic $\hat{p}_S - \hat{p}_L \doteq 0.132$:

```
2 * (1 - pnorm(.132, 0, .114))

[1] 0.246907
```

Or, we get the same result if we standardize:

$$z = \frac{(\hat{p}_S - \hat{p}_L) - 0}{\text{SE}} = \frac{0.132}{0.114} = 1.158,$$

```
2 * (1 - pnorm(1.158))

[1] 0.246864
```

This *P*-value has not employed **continuity correction**. The `prop.test()` command does employ it:

```
prop.test(c(10,20), n=c(23,66))


2-sample test for equality of proportions with continuity correction

data:  c out of c10 out of 2320 out of 66
X-squared = 0.80095, df = 1, p-value = 0.3708
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1285116  0.3920162
sample estimates:
   prop 1    prop 2
0.4347826 0.3030303
```

and produces an even higher *P*-value (0.371 as opposed to 0.247). The observed difference is not significant, not even at the 10% level. So, we fail to reject the null hypothesis.

4. (a) One can find out more about the set by typing either of these commands

```
?red.cell.folate
help(red.cell.folate)
```

Along with getting some specifics about the variables `folate` and `ventilation`, we see that the book *Practical Statistics for Medical Research*, by D.G. Altman, served as the source for the creator of this data set.

(b) We look at the number of instances from each group:

```
tally(~ventilation, data=red.cell.folate)

ventilation
N2O+O2,24h  N2O+O2,op     O2,24h
         8          9          5
```

and see that there are nowhere near 30 observations from each group. It would be difficult, with such small samples, to confirm (with any sort of plot, even a quantile-quantile plot) that the three populations are normally distributed. If there were other reasons (known to the researchers) to believe the populations are normal, then one might not worry about normality. As it is, we proceed with some trepidation.

Looking at the three standard deviations,

```
sd(folate ~ ventilation, data=red.cell.folate)

N2O+O2,24h  N2O+O2,op     O2,24h
  58.71709   37.12180   33.75648
```

we see that the ratio of largest to smallest is $58.72/33.76 \doteq 1.74$, not exceeding 2. That rule of thumb, at least, is met.

(c) We have hypotheses

$\mathbf{H}_0$: all (population) group means are equal

$\mathbf{H}_a$: at least one (population) group mean differs from another

If we proceed as if part (b) gave no cause for worry (about using an F-distribution to determine the P-value),

```
anova(lm(folate ~ ventilation, data=red.cell.folate))
```

```
Analysis of Variance Table

Response: folate
            Df Sum Sq Mean Sq F value  Pr(>F)
ventilation  2  15516  7757.9  3.7113 0.04359 *
Residuals   19  39716  2090.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We might, instead, use randomization. Our sample (test) statistic is the same F-value, 3.7113, as computed above. We select many randomization samples (ones that incorporate the assumption `folate` and `ventilation` are independent), computing an F-statistic for each one

```
manyFstats <- do(5000) * anova( lm( folate ~ shuffle(ventilation), data=red.cell.folate ) )$F[1]
head(manyFstats)
```

```
     result
1 0.3430726
2 1.1355590
3 1.2768383
4 1.0563790
5 0.3678615
6 0.4456326
```

and look at the relative frequency of an F-statistic as extreme as ours:

```
nrow( filter(manyFstats, result >= 3.711) ) / 5000
```

```
[1] 0.0434
```

The P-values, 0.044 and 0.047, from the two methods, are both small enough to reject the null hypothesis in favor of the alternative at the 5% level.

(d) Now that we have rejected $\mathbf{H_0}$, we carry out pairwise comparisons:

```
TukeyHSD(aov(folate ~ ventilation, data=red.cell.folate))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = folate ~ ventilation, data = red.cell.folate)

$ventilation
                          diff        lwr      upr     p adj
N2O+O2,op-N2O+O2,24h -60.18056 -116.61904 -3.74207 0.0354792
O2,24h-N2O+O2,24h    -38.62500 -104.84037 27.59037 0.3214767
O2,24h-N2O+O2,op      21.55556  -43.22951 86.34062 0.6802018
```

There is a significant difference (again, at the 5% level) between group means for the `N2O+O2,op` and `N2O+O2,24h` groups. No other pair of groups appears to have had a significant difference in means.