hw_lect23_1:

Consider the data from an example in a past lecture where a survey of students in 390 yielded the following data:
17 students like lab
48 do not like lab
15 have no opinion.
Suppose I believed that the proportion of students in each of the 3 categories (like, no-like, no-opinion) was equal. Does this data contradict that belief? Let alpha=0.05.

Let π_1 = True proportion of students who like lab.
π_2 = " " " " do not like Lab.
π_3 = " " " " have no opinion.

Then H_0: $\pi_1 = \frac{1}{3}$, $\pi_2 = \frac{1}{3}$, $\pi_3 = \frac{1}{3}$

H_1: At least 1 of These is wrong.

$$\chi^2_{obs} = \sum^3 \frac{(obs_i - exp_i)^2}{exp} = \sum^3_i \frac{(n\pi_{0i} - n_i)^2}{n\pi_{0i}}$$

$n = \sum n_i = 17+48+15 = 80$

$$= \frac{(80(\frac{1}{3}) - 17)^2}{80(\frac{1}{3})} + \frac{(80(\frac{1}{3}) - 48)^2}{80(\frac{1}{3})} + \frac{(80(\frac{1}{3}) - 15)^2}{80(\frac{1}{3})}$$

$$= \frac{1}{80(\frac{1}{3})}\left[(26.7-17)^2 + (26.7-48)^2 + (26.7-15)^2\right]$$

large
See below.

$$= \frac{1}{26.7}\left[94.09 + \boxed{453.69} + 136.89\right] = \underline{25.64}$$

p-value = $pr(\chi^2 > 25.64)$ $< .001$ df = 3-1 = 2 Table VII

So Because p-value $< \alpha$ (at $\alpha = .05$)
We can reject H_0 in favor of H_1 (at least one of The 3 proportions is not $\frac{1}{3}$), ie. evidence contradicts The belief.

In English: At least 1 of The props is not 1/3.

FYI: Diagnosis: The Categ. with The biggest deviation from $\frac{1}{3}$ is The "do not like lab" group.

hw_lect23_2 (by hand):
A sample of 210 Bell computers has 56 defectives. Theory suggests that a third of all Bell computers should be defective. Does this data contradict the theory (at alpha=0.05)? Specfcially,

a) Do a z-test (p.2 above),

b) Do a chi-squared test with k=2 categories. Hint: The pi's (and pi_0's) of the k categories must sum to 1.

c) Are the conclusions in a and b consistent?

a) Let π = proportion of defectives in all Bell comps.

$H_0: \pi = 1/3$
$H_1: \pi \neq 1/3$

Data says $p = 56/210 = 0.267$

This is a 1-sample, 2-sided/tailed test of a proportion.

If H_0 = True, then $z_{obs} = \dfrac{(56/210) - (1/3)}{\sqrt{\frac{1}{210}\left(\frac{1}{3}\right)\left(1-\frac{1}{3}\right)}} = -2.05$

$\therefore$ p-value $= 2\,\text{prob}(z < -2.05) = 2(0.0202) = \boxed{0.04}$

p-value $< \alpha \Rightarrow$ Reject H_0 in favor of H_1, i.e. Data does not support the theory.

b) This time let

π_1 = proportion of defectives in all Bell computers.

π_2 = " " non-defs " " " " .

H_0: $\pi_1 = 1/3$, $\pi_2 = 2/3$
H_1: At least one is wrong

$$\left.\begin{array}{l} H_0: \pi_1 = 1/3,\ \pi_2 = 2/3 \\ H_1: \text{At least one is wrong} \end{array}\right\} \Rightarrow \pi_{01} = 1/3,\ \pi_{02} = 2/3$$

	Cat. 1	Cat. 2
Expected counts :	$\frac{1}{3}(210) = 70$	$\frac{2}{3}(210) = 140$
observed counts :	56	$210 - 56 = 154$

$$X^2_{obs.} = \frac{(70-56)^2}{70} + \frac{(140-154)^2}{140} = 2.8 + 1.4 = 4.20$$

Table $\underline{\text{VII}}$ with $df = 2-1 = 1 \Rightarrow 0.045 < p\text{-value} < 0.040$

In fact, Table $\underline{\text{VII}}$ suggests that p-value $\sim \boxed{0.04}$

c) Yup! The 2-sided z-test gives p-value = 0.04
The chi-squared test gives p-value $\sim$ 0.04

Also note that $X^2_{obs} = 4.20$ turns out to be equal to $z^2_{obs} = (-2.05)^2 = 4.20$.
This is not accidental: The chi-squared distr. is related to the square of a variable that's normally distributed.

hw_lec24_1:
The following data refer to the melting temperature, y (in some unit), of a certain material at four different pressures, x (in some unit).

```
Pressure     Temperature
----------------------------
1.6      59.5, 53.3, 56.8, 63.1, 58.7
3.8      55.2, 59.1, 52.8, 54.4
6.0      51.7, 48.4, 53.9, 49.0
10.2     44.6, 48.5, 41.0, 47.3, 46.1
```

a) Make a comparative boxplot of y for the four pressure levels. By R.

```
x = c(1.6, 1.6, 1.6, 1.6, 1.6, 3.8, 3.8, 3.8, 3.8, 6.0, 6.0, 6.0, 6.0, 10.2, 10.2, 10.2, 10.2, 10.2)
y = c(59.5, 53.3, 56.8, 63.1, 58.7, 55.2, 59.1, 52.8, 54.5, 51.7, 48.8, 53.9, 49.0, 44.6, 48.5, 41.0, 47.3, 46.1)
boxplot(y~x, data = dat)    # This was covered in the lab material for 1-way ANOVA.
```

b) Based on the above boxplot, would you say there is a difference in the mean melting temperature for at least 2 of the pressure levels?

Yes, it looks like the mean melting temperature may be different for the 1st and 3rd pressure levels, 1st and 4th levels, 2nd and 4th levels, and perhaps even for the 3rd and 4th levels.

c) At alpha = 0.05, is there evidence that the mean melting temperature at the at least 2 of the four pressure levels are different? Report the p-value, and state the conclusion clearly. By R; see prelab to see how to do 1-way ANOVA in R.

```
aov.1 = aov(y ~ as.factor(x))
summary(aov.1)        # pvalue = 5.87e-05

#                  Df Sum Sq Mean Sq F value   Pr(>F)
# as.factor(x)  3  456.5    152.17     17.11     5.87e-05
# Residuals    14  124.5    8.89
```

Since the p-value (5.87e-05) < alpha, we reject H0 (that the 4 means are equal) in favor of H1 (at least two of the means are different).

d) Write code to compute the above p-value "by hand," i.e. without using aov() or lm(), but using the basic formulas for SS_between, SS_within, etc.

There are many ways, but here is one:

```
y1 = mean(y[x==1.6]);
y2 = mean(y[x==3.8]) ;
y3 = mean(y[x==6.0]) ;
y4 = mean(y[x==10.2]);
ybarbar = mean(y)

v1 = var(y[x==1.6]);
v2 = var(y[x==3.8]);
v3 = var(y[x==6.0]);
v4 = var(y[x==10.2])

n1 = length(y[x==1.6]);
n2 = length(y[x==3.8]);
n3 = length(y[x==6.0]);
n4 = length(y[x==10.2]);
n = length(y)
```

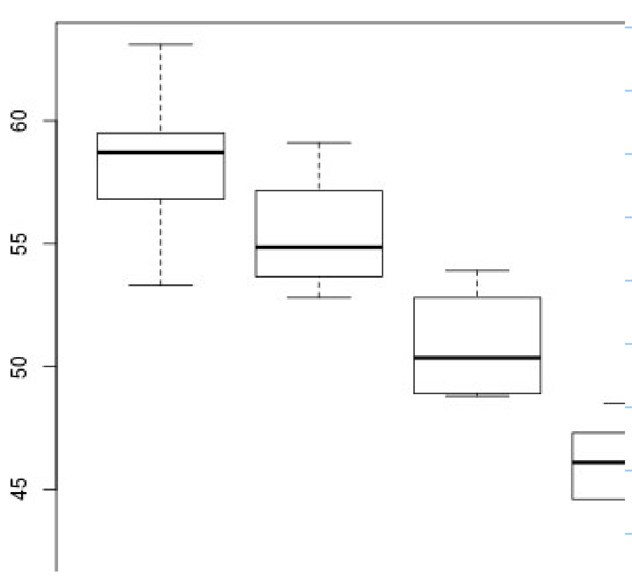
```
k = 4
SS_between = n1*(y1-ybarbar)^2 + n2*(y2-ybarbar)^2 + n3*(y3-ybarbar)^2 + n4*(y4-ybarbar)^2   # 456.5048
SS_within = (n1-1)*v1 + (n2-1)*v2 + (n3-1)*v3 + (n4-1)*v4        # 124.498
MS_between = SS_between/(k-1)                          # 152.1683
MS_within = SS_within/(n-k)                            # 8.892714
F_obs = MS_between/MS_within                           # 17.11157
pf(F_obs, df1=(k-1), df2=(n-k), lower.tail=F)  # 5.874561e-05
```

e) After (or before) a 1-way ANOVA test, one should check the two assumptions that the y's are normally distributed within each group, and with the same variance. To that end, make a plot that shows four qqplots (one for each pressure level) superimposed onto a single figure; make sure that the four qqplots have different colors. Are the 4 qqplots reasonably straight, and do they have approximately equal slopes? Hint: in the first call to plot(), use xlim=c(-2,2) and ylim=range(y).). Use the "By hand" method for making qqplots.

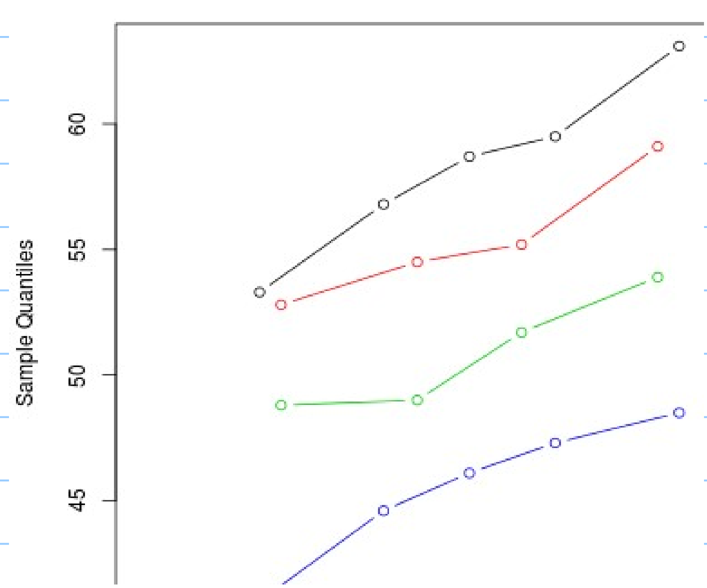
```
xx = y[x==1.6]
N = length(xx)
X = seq(.5/N, 1-.5/N, length=N)
Q = qnorm(X)
plot(Q, sort(xx), type="b", ylim = range(y), xlim=c(-2,2), xlab="Theoretical Quantiles", ylab="Sample Quantiles")

for(i in 2:4){
  if(i==2){ xx = y[x==3.8] }
  if(i==3){ xx = y[x==6.0] }
  if(i==4){ xx = y[x==10.2] }
N = length(xx)
X = seq(.5/N, 1-.5/N, length=N)
Q = qnorm(X)
points(Q, sort(xx), type="b", col=i )
}
```

The qqplots all look pretty straight (i.e., suggesting normality), and the slopes are pretty similar too (i.e., suggesting equal variances).

a)



e)



hw_lect24_2
In a problem dealing with flow rate (y) and pressure-drop (x) across filters, it is known that y= -0.12 + 0.095 x. Note: this is the true "fit" to the population. Suppose it is also known that sigma_epsilon = 0.025. Now, IF we were to make repeated observations of y when x=10, what's the prob. of a flow rate exceeding 0.835?

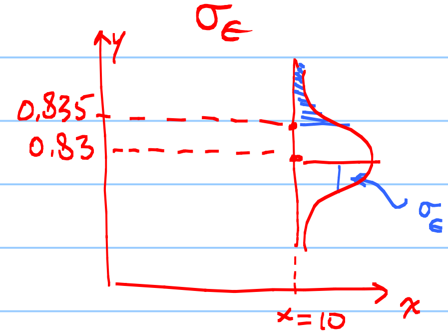
random y at x=10

$$? = \text{prob}(y > 0.835 \mid x = 10)$$

$$= \text{prob}\left(\frac{y - (\overbrace{\text{true mean of } y \text{ at } x=10}^{=y(x)})}{\sigma_\epsilon} > \frac{0.835 - (-0.12 + 0.095(10))}{\sigma_\epsilon} \right)$$

$$= \text{prob}\left(z > \underbrace{\frac{0.835 - 0.83}{.025}}_{0.2} \right) = 1 - 0.5793$$

$$= \underline{0.4207}$$



hw_lect25_1
We have learned that if p-value < alpha, then there's evidence to reject H0 in favor of H1. For the test of model utility, p-value = pr(F > F_obs). So, for that p-value to be less than alpha, F_obs must be larger than some critical value.

a) At alpha=0.05, find that critical value of F_obs for a multiple regression problem involving four betas, and 30 cases.

b) Find the critical value of R^2 (above which p-value < alpha). Hint: The F-ratio appearing in the test of model utility depends on R^2 of the model. So, if you know the critical value of F (as in part a), then you know the critical value of R^2.

Moral: Like all other tests we have studied, the reject/no-reject decision can be based in the critical value of some statistic, i.e. without a p-value. For the test of model utility, the decision can be made by comparing F_obs with some critical value (e.g. found in part a), or even by comparing R^2_obs with its critical value (e.g. found in part b).

a) $\text{p-value} = pr(F > F_{obs}) = 0.05 \implies F_{obs} = 2.76$

(from Table $\underline{\text{VIII}}$, $df = (4, 25)$)

b) $\frac{R^2}{1-R^2} \cdot \frac{25}{4} = 2.76 \implies \frac{R^2}{1-R^2} = \frac{2.76(4)}{25} \implies \frac{1-R^2}{R^2} = \frac{25}{4(2.76)}$

$\frac{1}{R^2} - 1 = \frac{25}{4(2.76)} \implies \frac{1}{R^2} = 1 + \frac{25}{4(2.76)} = \frac{4(2.76)+25}{4(2.76)}$

$R^2 = \frac{4(2.76)}{4(2.76)+25} = 0.306$

Moral: Eventhough the p-value is computed from F_{obs}, it is possible to perform the test of model utility using F_{obs} itself (ie. without a p-value), or even R^2_{obs} itself. Infact you can even do the test using s_e, because recall that R^2 is basically SS_{exp}, and s_e is essentially $SS_{unexp.}$, and $SS_{exp} + SS_{unexp} = $ fixed constant.