AP Stats Frappys

name:

1. The Behavioral Risk Factor Surveillance System is an ongoing health survey system that tracks health conditions and risk behaviors in the United States. In one of their studies, a random sample of 8,866 adults answered the question "Do you consume five or more servings of fruits and vegetables per day?" The data are summarized by response and by age-group in the frequency table below.

Age-Group (years)	Yes	No	Total
18–34	231	741	972
35–54	669	2,242	2,911
55 or older	1,291	3,692	4,983
Total	2,191	6,675	8,866

data

Do the data provide convincing statistical evidence that there is an association between age-group and whether or not a person consumes five or more servings of fruits and vegetables per day for adults in the United States?

H_0: Consuming five or more servings of fruits and vegetables is independent of age group.

H_A: Consuming five or more servings of fruits and vegetables is not independent of age group.

CONDITIONS

✓ count data

✓ SRS – states "a random sample"

✓ expected counts ≥ 5
(checking [B] after analysis)
all counts in B ≥ 5

Enter data in matrix [A] and perform a χ^2-Test of independence

$\chi^2 = 8.9835$

p-value = .0112

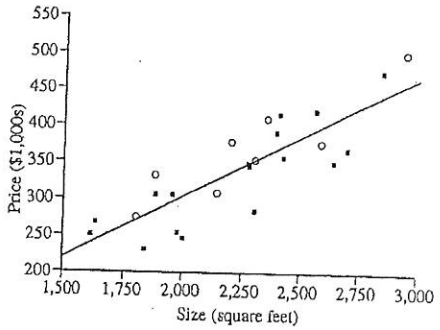
df = 2

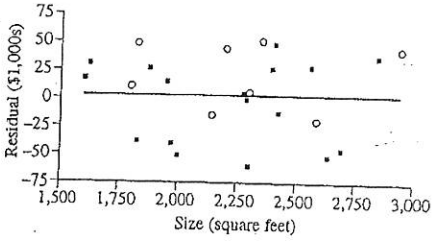
With $\alpha = .05$ p-value = .0112 is low, so we reject H_0.
We do have sufficient statistical evidence to conclude that consuming five or more servings of fruits and vegetables is not independent of age group.

4)

A real estate agent is interested in developing a model to estimate the prices of houses in a particular part of a large city. She takes a random sample of 25 recent sales and, for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below and on the next page.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	−18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	−23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	−45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	−46
244	2,000	no	−57
347	2,274	no	1
345	2,279	no	−2
282	2,300	no	−69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	−19
419	2,560	no	26
348	2,639	no	−58
365	2,701	no	−52
474	2,849	no	33





Linear Fit
Price = −28.144 + 0.165 Size

Summary of Fit
RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	−28.144	48.259	−0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

1) Write the LSRL. Identify all variables

$\hat{price} = -28.144 + 0.165(size)$

price = price of house, in $1000s
size = size of house, in ft^2

2) Is there association between square footage and price? Conduct the appropriate hypothesis test.

$H_0: \beta = 0$ (no association between size and price)
$H_A: \beta \neq 0$ (is an association between size and price)

CONDITIONS

✓ straight enough (scatterplot provided)
✓ residuals: no pattern or fanning (residual plot provided)
✓ residuals Nearly Normal (enter last column resids in calc. for histogram)

(good enough)

Using Software Output
p-value on inference for slope is $p < .0001$
with $\alpha = .05$, $p < .0001$ is low so we reject H_0.
We do have sufficient statistical evidence to conclude that there is an association between house size and sale price.

3) Explain what s_b and s are in the context of the problem.

- $s_b = .0213$ $1000/ft^2

If we took many different samples and found LSRLs for each the slope, b, would vary due to natural sampling variation. We expect the standard deviation of these slopes to be .0213 $1000/ft^2

- $s = 37.3959$ ($1000) (or $37,396)

From doing 1-Var stats on the residuals column, we find the standard deviation of the residuals. In this sample the error between the predicted house sale price and the actual price is $37,396