Some late-semester inference practice problems

TLS

2024-04-17

Problem 1

```
ssurv = read.csv("https://scofield.site/teaching/data/csv/ssurv.csv")
names(ssurv)
```

```
##  [1] "sex"            "class"          "gpa"            "height"
##  [5] "pulse"          "childrank"      "numchildren"    "haircut"
##  [9] "randomnum"      "speedtickets"   "cds"            "smoker"
## [13] "hourssleep"     "selfhandedness" "momhandedness"  "dadhandedness"
## [17] "region"         "oncampus"       "cupscoffee"     "birthday"
## [21] "overtwenty"
```

The relevant variables to this problem are called `region` and `smoker`. A closer look at the data confirms they are both categorical. The research question appears to be tied to determining if these variables are independent or associated. That is, we will conduct a chi-square test for association.

$$\mathbf{H}_0 : \text{the variables region and smoker are independent}$$

$$\mathbf{H}_a : \text{the variables are associated}$$

We can make a two-way table using `tally()`

```
tally(smoker ~ region, data=ssurv)
```

```
##        region
## smoker      Rural Suburban Urban
##           0     1        2     1
##   Non     1    49      173    35
##   Smoke   0     0       15     3
```

but there is evidence, here, of missing responses from those surveyed. Let's filter out any blank responses in the two variables, and then rebuild this table:

```
filteredSsurv = subset(ssurv, region!="" & smoker!="")
regionSmokingTable = tally(smoker ~ region, data=filteredSsurv)
regionSmokingTable
```

```
##        region
## smoker  Rural Suburban Urban
##   Non      49      173    35
##   Smoke     0       15     3
```

We calculate the expected counts, either by hand (row total times column total divided by grand total), or using software:

```
chisq.test(regionSmokingTable)$expected
```

```
## Warning in chisq.test(regionSmokingTable): Chi-squared approximation may be
## incorrect
```

```
##        region
## smoker      Rural  Suburban     Urban
##   Non   45.792727 175.69455 35.512727
##   Smoke  3.207273  12.30545  2.487273
```

These expected counts will be used, along with the observed counts from the original two-way table, to calculate a χ^2-statistic. Noting that **not all of these counts are at least 5**, we will not trust a chi-square distribution to produce an accurate P-value. Instead, we will ask `chisq.test()` to simulate a P-value:

```
chisq.test(regionSmokingTable, simulate.p.value = TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  regionSmokingTable
## X-squared = 4.1764, df = NA, p-value = 0.1089
```

Our P-value is not significant even at the 10% level, so we fail to reject the null hypothesis. While we have not proved that students smoke (or not) in the same proportions across all regions, our data does not present us with strong enough evidence to refute that.

Problem 2

I will read in the data and look at the first few rows, maybe just because that gets me thinking about what sort of data I have.

```
ironContent = read.csv("https://scofield.site/teaching/data/csv/ips5e/ironContent.csv")
head(ironContent)
```

```
##    typepot g iron
## 1 Aluminum 1 1.77
## 2 Aluminum 1 2.36
## 3 Aluminum 1 1.96
## 4 Aluminum 1 2.14
## 5     Clay 2 2.27
## 6     Clay 2 1.28
```

Note that the iron content is in a quantitative variable called `iron`, and the pot material is in `typepot`. Our original research question might be answered by assessing the evidence that population mean amounts of iron are different across the three groups of pots represented here; that is, we will do an ANOVA test. First, we are concerned with whether it is appropriate.

```
favstats(iron ~ typepot, data=ironContent)
```

```
##    typepot  min     Q1 median    Q3  max   mean        sd n missing
## 1 Aluminum 1.77 1.9125  2.050 2.195 2.36 2.0575 0.2519755 4       0
## 2     Clay 1.28 2.0225  2.375 2.530 2.68 2.1775 0.6213091 4       0
## 3     Iron 4.06 4.1800  4.695 5.195 5.27 4.6800 0.6282781 4       0
```

The number of cases in each group are very small. We must hope that the populations they come from have (near) bell-shaped distributions for iron measurements. Morever, the ratio of sample standard deviations $0.6283/0.2520 > 2$. While I will conduct the ANOVA test, this does present some concerns about conclusions we might draw from it.

```r
anova(lm(iron ~ typepot, data=ironContent))
```

```
## Analysis of Variance Table
##
## Response: iron
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## typepot    2 17.5392  8.7696  31.162 9.006e-05 ***
## Residuals  9  2.5327  0.2814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value, if we trust it, leads to rejection of

$$\mathbf{H}_0 : \mu_{\text{clay}} = \mu_{\text{aluminum}} = \mu_{\text{iron}}$$

in favor of the alternative, that at least one of these group's has a different mean than another. The follow-up test

```r
TukeyHSD(iron ~ typepot, data=ironContent)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $typepot
##                  diff        lwr      upr     p adj
## Clay-Aluminum 0.1200 -0.9273123 1.167312 0.9454849
## Iron-Aluminum 2.6225  1.5751877 3.669812 0.0001688
## Iron-Clay     2.5025  1.4551877 3.549812 0.0002412
```

suggests the mean for iron is different from both the mean for clay and the mean for aluminum. We do not have similar evidence to refute that the means for clay and aluminum are different. All of this, of course, is valid when the model assumptions are in place and, since we could not verify those model assumptions, we might suggest follow-up studies to bolster the results stated here.

Problem 3

We want to test whether our random number generator is producing numbers in $(0, 1)$ uniformly. Given the methodology of splitting results off into 5 categories, all of equal width, we would want the data to fit a probability model where each category is equally-likely. That is, I am suggesting a goodness-of-fit test with hypotheses

$$\mathbf{H}_0 : p_{[0,0.2]} = p_{(0.2,0.4]} = p_{(0.4,0.6]} = p_{(0.6,0.8]} = p_{(0.8,1.0]} = \frac{1}{5},$$

and an alternative that at least one of these proposed proportions is incorrect.

For ease of calculation, I put observed and expected counts into lists:

```r
observed = c(114, 92, 108, 101, 85)
nullProportions = rep(1/5, 5)
expected = 500 * nullProportions
expected
```

```
## [1] 100 100 100 100 100
```

Our expected counts are all above 5. We will proceed with the calculation of the χ^2-statistic and a P-value using the chi-square distribution with df $= 4$.

```
chisqStat = sum((observed - expected)^2 / expected)
chisqStat
```

```
## [1] 5.5
```

```
1 - pchisq(chisqStat, df=4)
```

```
## [1] 0.2397295
```

This, again, is not a significant result (at the 10% level). So, while we have not proved our random number generator selects numbers in these 5 categories equally-often, we do not have evidence to refute it.

Problem 4

Let's obtain the data and look over the first few rows.

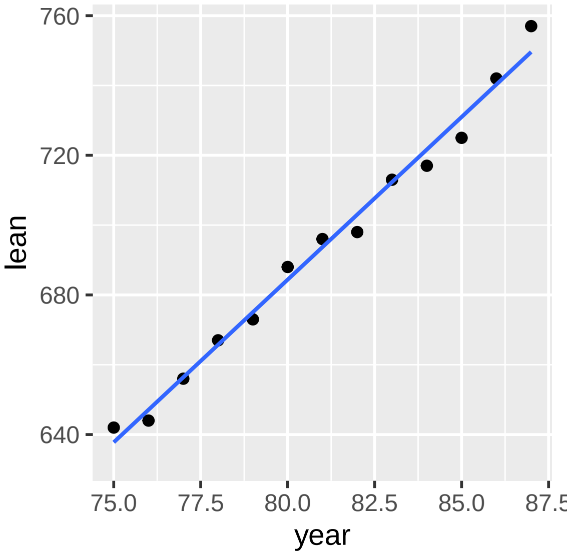
```
tower = read.csv("https://scofield.site/teaching/data/csv/ips5e/leaningTowerPisa.csv")
head(tower)
```

```
##   year lean
## 1   75  642
## 2   76  644
## 3   77  656
## 4   78  667
## 5   79  673
## 6   80  688
```

Now, let's do a scatterplot to see if a linear association seems plausible.

```
gf_point(lean ~ year, data=tower) |> gf_lm()
```

```
## Warning: Using the `size` aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the `linewidth` aesthetic instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



I will skip, here, the residual plots (the ones I generated using `mplot()`) and go straight to doing inference. (That is, I'm doing inference without having checked the conditions that "bless" it.)

The question I want to address is the model utility test. For, if the model were not useful (notwithstanding the possible deception of a nonzero b_1 from the sample), then both β_1 and ρ would be zero. I'll use the former as my method for stating the null hypothesis for model utility:

$$\mathbf{H}_0 : \beta_1 = 0 \qquad \text{vs.} \qquad \mathbf{H}_a : \beta_1 \neq 0.$$

The standardized t-statistic (as well as the F-statistic) are reported in the output of

```
summary(lm(lean ~ year, data=tower))
```

```
##
## Call:
## lm(formula = lean ~ year, data = tower)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9670 -3.0989  0.6703  2.3077  7.3956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.1209    25.1298  -2.432   0.0333 *
## year          9.3187     0.3099  30.069  6.5e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.181 on 11 degrees of freedom
## Multiple R-squared:  0.988,  Adjusted R-squared:  0.9869
## F-statistic: 904.1 on 1 and 11 DF,  p-value: 6.503e-12
```

And we can confirm the P-value reported by R using either one of these test statistics:

```
2*(1 - pt(30.069, df=11))
```

```
## [1] 6.502354e-12
```

```
1 - pf(904.1, df1=1, df2=11)
```

```
## [1] 6.504131e-12
```

Either way, we reject the null hypothesis in favor of $\beta_1 \neq 0$, confirming that the linear model is useful.