Practice Exam #2 - Sketch Solution

Directions

- Answer each question using *no more than specified number of sentences* and not attempt to avoid these guidelines by using run-on sentences. Answers that are unnecessarily verbose may result in point loss.
- Do not include superfluous information in your answers, you may be penalized if you make an inaccurate statement even if you go on to provide a correct answer. Your answers should be clear, concise, and include only what is needed to answer the question that was asked.

Formula Sheet

Standard Errors:

Statistic	Standard Error	Conditions
$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ and $n(1-p) \geq 10$
$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	normal population or $n \geq 30$
$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1,2\}$
$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	normal populations or $n_1 \geq 30$ and $n_2 \geq 30$

Definitions:

- Odds: frequency of an event divided by the frequency of the complimentary event (ie: A/B)
- Risk: proportion of the time that an event occurs (ie: A/(A+B))
- Type I error: Rejecting H_0 when it is true
- Type II error: Not rejecting H_0 when it is false

Confidence Interval Calibration Values:

Distribution	c for 95%	c for 99%
Normal	1.96	2.58
$t(df = 5)$	2.57	4.03
$t(df = 25)$	2.06	2.79

Section 1 - True/False

Directions: Clearly indicate whether each of the following statements is true or false. You do not need to explain your reasoning and you will not receive a better score for doing so.

1. The main difference between an *experimental study* and an *observational study* is whether the researchers manipulated the explanatory variable.

- TRUE - an observational design is characterized by the researchers simply observing the study's subjects without any intervention, while an experiment involves actively manipulating at least one variable.

2. In a retrospective, case-control study investigating oral cancer the researchers obtain their data by recruiting a sample of participants with oral cancer and an entirely separate sample of participants without oral cancer.

- TRUE - in a retrospective case-control design subjects are selected into the study separately based upon whether they experienced the outcome of interest (cases) or did not (controls).

3. For categorical outcomes, odds ratios and relative risks are generally considered a more useful measure of effect size than differences in proportions when the sample size is large.

- FALSE - odds ratios and relative risks are preferred when the outcome is rare (small proportions), which doesn't necessarily have anything to do with the sample size.

4. Consider a random sample $n = 10$ games played by an NFL team where the sample odds of the team winning are 4. These sample odds suggest that the team lost exactly 2 games in the sample.

- TRUE - the sample odds are the ratio of how often a certain outcome (winning in this example) is observed relative to how often it is not observed (losing). Since the team played 10 games they must have won 8 and lost 2 to have odds of 4.

5. The null hypothesis in a Chi-squared Goodness Fit test is always set up to suggest each category as being equally likely within the population.

- FALSE - the proportions in the null distribution can different (see the Alameda County jury pool example from Lab 8)

6. When finding the p-value in a Chi-squared test we only consider the area of the null distribution corresponding to values greater than the observed test statistic, even when we are interested in differences that could go in either direction.

- TRUE - because the numerator in the Chi-squared test statistic is squaring the difference between observed and expected counts any deviations from expected will make the test statistic a larger positive number, regardless of whether the observed count is higher than expected, or lower than expected.

7. In one-way ANOVA, rejecting the null hypothesis suggests that the observed data do not follow a Normal distribution.

- FALSE - One-way ANOVA assumes the residuals are Normally distributed, but this is not the null hypothesis. Rejecting the null hypothesis suggests the alternative model is superior, meaning that at least one group mean differs from the others.

8. In one-way ANOVA, the fit of the null and alternative models are measured using the sum of their squared residuals.

- TRUE - a residual measures the deviation between an observed outcome and a model's corresponding predicted outcome. Summing the squared residuals provides a single number that expresses the relative goodness of fit of a model, as when the residuals are closer to zero the model's predictions are more accurate.

9. Lowering the threshold for statistical significance from $\alpha = 0.05$ to $\alpha = 0.01$ will increase the likelihood of a hypothesis test producing a Type II error.

- TRUE - this change makes it more difficult to the reject H_0, even in situations where H_0 is incorrect and should be rejected.

10. When performing many hypothesis tests as part of a single experiment, family-wise Type I error rate control procedures, such as the Bonferroni adjustment, typically result in more Type II errors than false discovery rate control procedures.

- TRUE - family-wise Type I error control does something that is more stringent (limits the probability of *any* type I errors) which means that fewer hypotheses will be rejected, which in turn will lead to a higher likelihood of type II errors.

11. Suppose a scientifically rigorous study reports a 95% confidence interval estimate for the mean cholesterol level of US adults as (202.4, 225.6). This interval suggests 95% of the US adult population has a cholesterol level between 202.4 and 225.6.

- FALSE - the confidence interval is estimating the population's mean, not the values of individual members of the population. Additionally, the confidence level describes the long-run success rate of the procedure used to make the interval; it doesn't say anything about the distribution of cases in the population.

12. Bootstrapping is a way to increase the sample size of a study by resampling the observed cases with replacement.

- FALSE - bootstrapping is a way to construct valid confidence intervals in scenarios where it is difficult to identify an appropriate probability model. Bootstrap samples should be the same size as the original sample in order to capture the correct amount of sampling variability. Remember that the goal of bootstrapping is to mimic the variability that is expected when a sample of a certain size is drawn from a population.

13. Consider a 95% confidence interval estimate for ρ, the correlation between two variables in a population. If a sample of $n = 50$ cases produces an interval estimate of (0.01, 0.26) then we'd expect a hypothesis test of $H_0 : \rho = 0$ to produce a p-value less than 0.05.

- TRUE - the null value of 0 is not contained within in the confidence interval, which implies the p-value will be less than one minus the confidence level. This is because both methods (confidence intervals and hypothesis tests) are using the same underlying probability model, but they are focusing on complimentary parts of it (the middle $P\%$ in confidence intervals, vs. the tails in hypothesis testing).

Section 2 - Conceptual Questions

Directions: Answer each question in about 3-sentences. Do not include unnecessary details, as you will be penalized for any inaccurate statements, regardless of whether they are relevant or not. Aim to clearly answer the question that was posed, not to demonstrate your knowledge of related topics.

1. In your own words, explain the concept of the "confidence level" in confidence interval estimation. That is, what does the confidence level tell us about an interval? What inherent problem in interval estimation does the "confidence level" address?

- Confidence level describes the long-run success rate of the procedure used to produce the interval estimate.
- It is a way of adding meaning to the interval's margin of error. Without it we wouldn't know how to interpret the margin of error, and we wouldn't be able to comprehend how the interval is balancing the goals of accuracy (containing the truth) and precision (small enough range to be informative)

2. Suppose you and one of your friends are trying to estimate the proportion of Grinnell students who are double majors. You each take a representative sample of students and construct a 95% confidence using a statistically valid procedure. However, you sampled $n = 60$ students but your friend only sampled $n = 30$ students. Whose interval is more likely to contain the true proportion? Explain your reasoning.

- Both intervals are equally likely to contain the truth because it is given that both are statistically valid 95% confidence intervals. The same confidence level implies the same long-run success rate. My interval will have a narrower width though.

3. Consider a Chi-squared test of independence where we are interested in determining whether there is an association between three different exposure groups and three different health outcomes. In your own words, explain how you'd go about finding the expected counts within each group under the null hypothesis of this test.

- Under the null hypothesis of independence, the likelihood of a certain outcome category does not depend upon the value of the explanatory variable (ie: which group a subject belongs to). So, you can calculate the overall, pooled proportion of each outcome (ie: proportions ignoring the explanatory variable) and then multiply those proportions by the size of each group created by the explanatory variable. For example, if 40% of the sample, regardless of group, have outcome "A", and group 1 has 100 cases, then we'd expect 40 cases in group 1 to have the "A" outcome under independence.

4. Log-transformations are often used on the outcome variable prior to performing one-way ANOVA. In your own words, explain the purpose of a log-transformation in this context. That is, what are the reasons why someone might log-transform their outcome variable prior to performing an ANOVA test?

- ANOVA assumes the residuals follow a Normal distribution, which tends not to happen when the outcome variable is right-skewed. Log-transformations address right-skew by compressing the distance between large values, which tends to make the residuals more Normally distributed. Log-transformations are also popular because the results are more easily interpreted than other data transformations that alleviate right-skew.

Section 3 - Application #1

Researchers in Florida collected data from a random sample of $n = 26$ lakes across the state, recording the median mercury level of large mouth bass (ppm) in each lake. The average mercury level in the sampled lakes was 0.55 ppm, and the standard deviation was 0.53 ppm.

Part A: Consider the calculation of a 99% confidence interval estimate for the average mercury level in all lakes in Florida using this sample. Provide both components of the margin of error for this interval.

- From the front page, $c = 2.79$ for 99% confidence and a t-distribution with 25 degrees of freedom. We estimate $SE = s/\sqrt{n} = 0.53/\sqrt{26} = 0.104$

Part B: Provide the 99% confidence interval estimate for the average mercury level in all lakes in Florida using the margin of error you calculated in Part A.

- $0.55 \pm 2.79 \cdot 0.104 = (0.26, 0.84)$

Part C: The EPA has a maximum acceptable level of 1.0 ppm for the mercury concentration in predatory and bottom-dwelling species of fish, with levels lower than this being deemed "safe". Does the confidence interval you calculated in Part B suggest that it is plausible that lakes in Florida, on average, have unsafe levels of mercury according to the EPA?

- No, it is not plausible that the average level of mecurcy is Florida lakes is unsafe since the entire 99% CI is below the maximum acceptable level. An average concentration about 1.0 is not plausible according to the CI.

Part D: Assuming everything else remains unchanged, which of the following will *decrease* the width of the interval you calculated in Part B.

- i) An additional 20 lakes are sampled, bringing the total sample size to $n = 46$
  - Yes, this will decrease the width due to a smaller SE (there will be less sampling variability in the sample mean)
- ii) You change the confidence level to 95%
  - Yes, this will decrease the width due to a smaller c (the interval doesn't need to be as likely to succeed)
- iii) Lakes are sampled until the interval's margin of error reaches 0.1
  - Yes, the previous MOE was 0.29, so sampling until the MOE is 0.1 will result in a much narrower CI

Part E: It is possible that *sampling bias* causes the confidence interval you calculated in Part B to fail to contain the true mercury level of all lakes in Florida? Briefly explain your reasoning.

- No, the researchers took a random sample of lakes in Florida. There might be other types of bias at play (ie: measurement bias), but the selection of cases into the sample did not introduce any bias into the estimate.

Section 4 - Application #2

We've previously analyzed the "flicker" data set, where researchers measured the critical flicker frequency (the variable `Flicker`), which is the highest frequency where a person can differentiate between a solid and flickering light source, for subjects with three different eye colors: blue, brown, and green (the variable `Color`). The researchers hypothesized that eye color was associated with critical flicker frequency.

Below are some descriptive statistics from this study that you should use throughout this question:

```r
## Overall summary
flicker %>%
  summarize(mean_flicker = mean(Flicker),
            median_flicker = median(Flicker),
            sd_flicker = sd(Flicker),
            n = n())
```

```
##   mean_flicker median_flicker sd_flicker  n
## 1     26.75263           26.8   1.845526 19
```

```r
## Summary by eye color
flicker %>%
  group_by(Color) %>%
  summarize(mean_flicker = mean(Flicker),
            median_flicker = median(Flicker),
            sd_flicker = sd(Flicker),
            n = n())
```

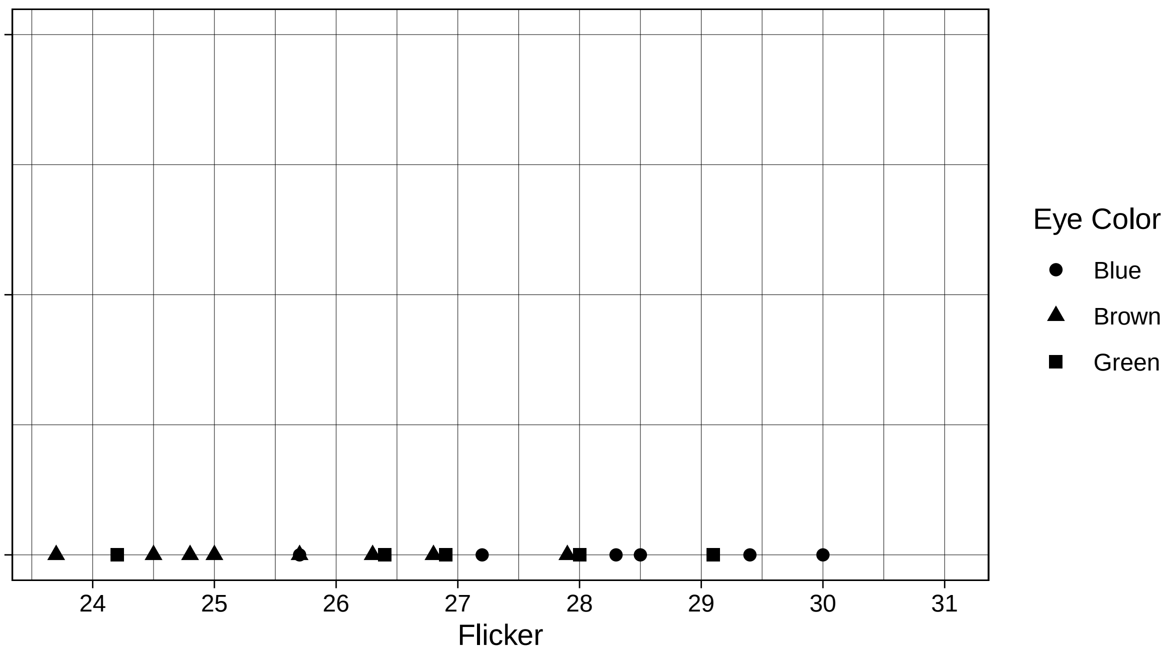
```
## # A tibble: 3 x 5
##   Color mean_flicker median_flicker sd_flicker     n
##   <chr>        <dbl>          <dbl>      <dbl> <int>
## 1 Blue          28.2           28.4       1.53     6
## 2 Brown         25.6           25.4       1.37     8
## 3 Green         26.9           26.9       1.84     5
```

Part A: One-way ANOVA can be expressed as a comparison of statistical models. With this in mind, briefly describe the *null model* using either words or statistical symbols (or both).

- $y_i = \mu + \epsilon_i$, where μ is a single, overall mean of the population, and ϵ_i is a random error from a Normal distribution.

Part B: The *null model* in one-way ANOVA involves using a certain probability distribution to model the data. Sketch this distribution on the graph below. Be careful where you place the center of the distribution.

- There should be Normal distribution should be centered at the overall mean.



Part C: One-way ANOVA compares models using sums of squares, which involve the squared deviations between observed and predicted values (ie: squared residuals). More specifically, $SS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. With this is in mind, what is the *residual* for the observation with the highest observed critical flicker frequency *under the null model.* Use the data shown on the x-axis of the graph given above.

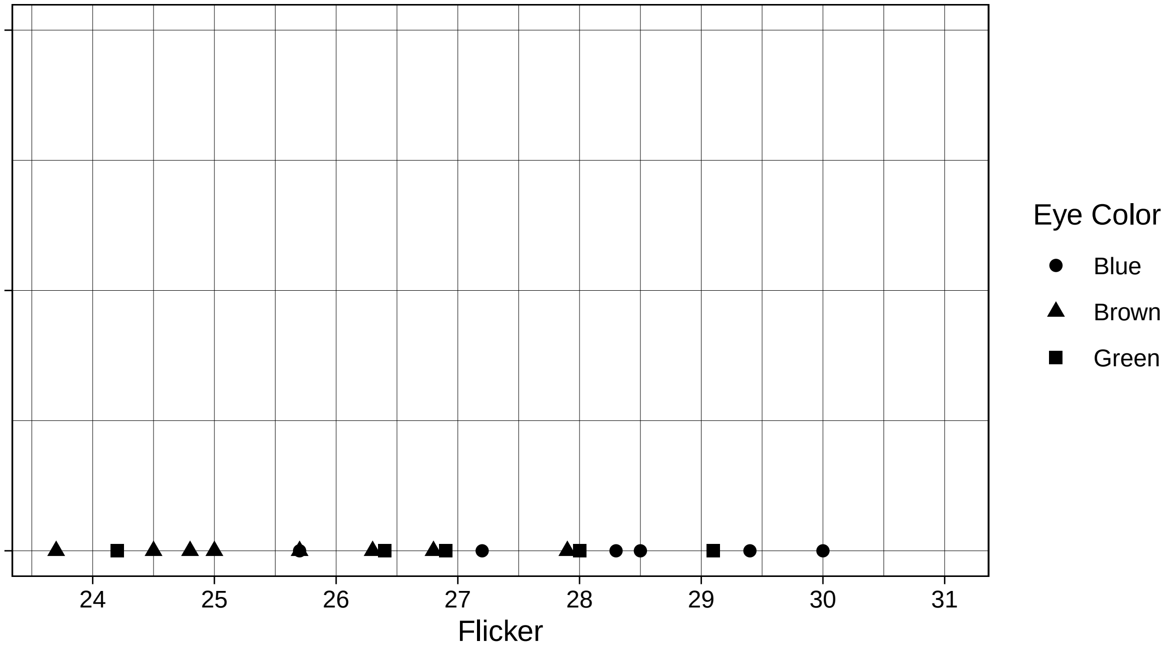
- 30-26.75 = 3.25

Part D: Now briefly describe the *alternative model* involved in one-way ANOVA using either words or statistical symbols (or both).

- $y_i = \mu_i + \epsilon_i$, where μ_i is the group-specific mean of the group that the i^{th} case belongs to and ϵ_i is a random error from a Normal distribution.

Part E: Similar to Part B, sketch the *alternative* model involved in one-way ANOVA on the graph given below:

- There should be three Normal distributions with the same amount of spread, each centered at a group-specific mean.



Part F: Similar to Part C, what is the *residual* for the observation with the highest observed critical flicker frequency *under the alternative model.*

- 30-28.16667 = 1.83333

Part G: The sum of squared residuals for the null model, SST, in this application is 61.31. Do you expect the sum of squared residuals for the alternative model to be larger, smaller, or approximately equal to this value? State "larger", "smaller", or "approximately equal" and briefly explain your reasoning.

- Smaller - the group-specific means are much closer to the cases in those groups, so the residuals will tend to be smaller when group-specific means are used to make predictions. Smaller residuals implies a smaller sum of squared residuals.

Part H: The F-statistic and p-value for the one-way ANOVA that analyzes the relationship between `Color` and `Flicker` are $F = 4.8$ and $p = 0.023$ respectively. Based upon these results, provide a brief conclusion in regard to the researcher's hypothesis described in the introduction of Question 2. You may assume the conditions of the test have been met.

- We find that there is moderately strong evidence (p=0.023) that eye color and critical flicker frequency are associated.

Part I: Shown below are the results of post-hoc pairwise testing for the ANOVA model described throughout this application. Briefly explain how these results allow you to make a more precise conclusion than the one you made in Part H.

```
model = aov(Flicker ~ Color, data = flicker)
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Flicker ~ Color, data = flicker)
##
## $Color
##                   diff        lwr        upr     p adj
## Brown-Blue  -2.595833 -4.7621317 -0.4295349 0.0181490
## Green-Blue  -1.263333 -3.6922387  1.1655720 0.3934460
## Green-Brown  1.332500 -0.9542388  3.6192388 0.3155339
```

- In Part H we had evidence that there was an association between eye color and critical flicker frequency, but we could not say which eye colors differed from each other. Using post-hoc testing we can see that the association is driven by blue-eyed individuals having higher critical flicker frequencies than brown eyed individuals.