Additional Hypothesis Tests Notes

Lt Col Ken Horton Professor Bradley Warner

28 July, 2020

Objectives

1) Conduct and interpret a hypothesis test for testing equality of two or more means; both permutation and using the F distribution.
2) Conduct and interpret a goodness of fit test using both Pearson's chi-squared to evaluate the independence between two categorical variables.
3) Conduct and interpret a hypothesis test for a equality of two variances.
4) Know and check assumptions for the tests in this lesson.

Introduction

The purpose of this lesson is to put all we learned in the block into perspective and then to also add a couple of new tests to demonstrate other statistical tests.

Remember that we have been using data to answer research questions. So far, we can do this with hypothesis tests or confidence intervals. There is a close link between these two methods. The key ideas have been to generate a single number metric to use in answering our question and then to obtain the sampling distribution of this metric.

In obtaining the sampling distribution we used permutation exact tests, probability models, mathematical models, and the bootstrap. Each of these had different assumptions and different areas that they can be applied. In some cases, several methods can be applied to the problem to get a since of the robustness to the assumptions.

Finding a single number metric has been difficult. For example, in the homework for last lesson, we wanted to determine if the prices of books at a campus bookstore were different from Amazon's prices. The metric we decided to use was the mean of the differences in prices. But is this the best way to answer the question? This metric has been used historically because of the need to use the t distribution. But there are other ways in which the prices of books can differ. Jack Welch was the CEO of GE for years and he made the claim that customers don't care about average but they do care about variability. The average temperature setting of your GE refrigerator could be off and you would adapt. However if the temperature had great variability, then you would be upset. So maybe metrics that incorporate variability might be good. In our bootstrap notes, we look at the ages of males and females in the HELP study. In the permutation test, we are assuming there is no difference in the distribution of ages between males and females. However, in the alternative we measured the difference using means. The means of these two populations could be equal but the distributions differ in other ways, for example variability. We could conduct a separate test for variances but we have to be careful about multiple comparisons, because the type 1 is inflated.

In addition, the use of the information in the data impacts the power of the test. In the golf ball example, when we used range as our metric, we did not have the same power as looking at the differences from expected values under the null hypothesis. There is some mathematical theory that leads to better estimators, they are called likelihood ratio tests, but this is beyond the scope of the class. What you can do, is create a simulation where you simulate data from the alternative hypothesis and then measure the power. This will give you a sense of the quality of your metric.

We will finish by examining problems with two variables. In the first case they will both be categorical but at least one of the categorical variables has more than two levels. In the second case, we will examine two variables where one is numeric and the other categorical. The categorical variable has more than two levels.

Categorical data

It is worth spending some time on common approaches to categorical data that you may come across. We have already dealt with categorical data to some extent. We have performed hypothesis tests and built confidence intervals for π, the population proportion of "success" in binary cases (support for a local measure in a vote). This is a single variable. Also, the golf ball example involved counts of four types of golf ball. This is considered categorical data because each observation is characterized by a qualitative value (type of ball). The data are summarized by counting how many balls in a sample belong to each type. This again was a single variable.

In another scenario, suppose we are presented with two qualitative variables and would like to know if they are correlated. For example, we have discussed methods for determining whether a coin could be fair. What if we wanted to know whether flipping the coin during the day or night changes the fairness of the coin? In this case, we have two categorical variables with two levels each: result of coin flip (heads vs tails) and time of day (day vs night). We have solved this problem by looking at a difference in probabilities of success. We also used a hypergeometric distribution.

We look at a scenario that involves categorical data, but where at least one variable has more than two levels. We will merely scratch the surface. You could take an entire course on statistical methods for categorical data. But with your inferential knowledge, you know the differences are only in the test statistic and distribution used to evaluate the hypotheses.

HELP example

Let's return to Health Evaluation and Linkage to Primary Care data set. We looked at the differences in ages between males and females, let's now do the same thing for the substance they were admitted to the program for.

There are three substances: alcohol, cocaine, and heroin. We'd like to know if there is evidence that these proportions differ for men and for women. In our data set, we observe modest differences.

```
tally( substance ~ sex, data = HELPrct,
format="prop", margins = TRUE)
```

```
##            sex
## substance    female      male
##   alcohol 0.3364486 0.4075145
##   cocaine 0.3831776 0.3208092
##   heroin  0.2803738 0.2716763
##   Total   1.0000000 1.0000000
```

But we need a test statistic.

Test statistic To help us understand the test statistic, let's simplify and use a simple example.

Suppose we have a 2 x 2 contingency table like the one below.

	Response 1	Response 2
Group 1	n_{11}	n_{12}
Group 2	n_{21}	n_{22}

If our null hypothesis is that the two variables are independent, a typical test statistic used is the Pearson chi-squared test statistic (X^2). Let e_{ij} be the expected count in the ith row and jth column under the null hypothesis:

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

But what is e_{ij}? What do we expect the count to be under H_0. To find this, we recognize that under H_0 (independence), joint probability is equal to the product of the marginal probabilities. Let π_{ij} be the probability of an outcome appearing in row i and column j. Our best guess at π_{ij} is $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$, where n is the total sample size. But under independence, $\pi_{ij} = \pi_{i+}\pi_{+j}$ where π_{i+} represents the total probability of ending up in row i and π_{+j} represents the total probability of ending up in column j. Note that π_{i+} is estimated by $\hat{\pi}_{i+}$ and

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n} = \frac{n_{i1} + n_{i2}}{n}$$

So, under H_0, our best guess for π_{ij} is:

$$\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}}{n}\frac{n_{+j}}{n}$$

So, under H_0,

$$e_{ij} = n\hat{\pi}_{ij} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

This may look too abstract, so let's break it down with an example, totally made up by the way.

Suppose we flip a coin 40 times during the day and 40 times at night and obtain the results below.

	Heads	Tails
Day	22	18
Night	17	23

To find the Pearson chi-squared (X^2), we need to figure out the expected value under H_0. Recall that under H_0 the two variables are independent. It's helpful to add the row and column totals prior to finding expected counts:

	Heads	Tails	Row Total
Day	22	18	40
Night	17	23	40
Column Total	39	41	80

Thus under independence, expected count is equal to the row sum multiplied by the column sum divided by the overall sum. So,

$$e_{11} = \frac{40 * 39}{80} = 19.5$$

Continuing in this fashion yields the following table of expected counts:

	Heads	Tails
Day	19.5	20.5
Night	19.5	20.5

Now we can find X^2:

$$X^2 = \frac{(22-19.5)^2}{19.5} + \frac{(17-19.5)^2}{19.5} + \frac{(18-20.5)^2}{20.5} + \frac{(23-20.5)^2}{20.5}$$

As you can probably tell, X^2 is essentially comparing the observed counts with the expected counts under H_0. The larger the difference between observed and expected, the larger the value of X^2. It is normalized by dividing by the expected counts since more data leads to a larger contribution. Under H_0, this statistic follows the chi-squared distribution with $(R-1)(C-1)$, in this case 1, degree of freedom (R is the number of rows and C is the number of columns).

p-value To find the Pearson chi-squared (X^2) in `R`:

```
e<-c(19.5,19.5,20.5,20.5)
o<-c(22,17,18,23)
x2<-sum(((o-e)^2)/e)

x2
```

```
## [1] 1.250782
```

```
1-pchisq(x2,1)
```

```
## [1] 0.2634032
```

Again, the p-value suggests there is not enough evidence to say these two variables are dependent.

Of course there is a

```
coin <- tibble(time = c(rep("Day",40),rep("Night",40)),
               result = c(rep(c("Heads","Tails"),c(22,18)),rep(c("Heads","Tails"),c(17,23))))
```

```
tally(~time+result,data=coin)
```

```
##        result
## time    Heads Tails
##   Day      22    18
##   Night    17    23
```

```
chisq.test(tally(~time+result,data=coin),correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tally(~time + result, data = coin)
## X-squared = 1.2508, df = 1, p-value = 0.2634
```

If you just want the test statistic, which we will for permutation tests, then use:

```
chisq(~time+result,data=coin)
```

```
## X.squared
##  1.250782
```

Extension to Larger Tables

The advantage of using the Pearson chi-squared is that it can be extended to larger contingency tables. Suppose we are comparing two categorical variables, one with r levels and the other with c levels. Then,

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

Under the null hypothesis of independence, X^2 follow the chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

Assumptions Note that to use this test statistic, the expected cell counts must be reasonably large. In fact, no e_{ij} should be less than 1 and no more than 20% of the e_{ij}'s should be less than 5. If this occurs, you should combine cells or look for a different test.

Permutation test

We will complete our analysis of the HELP data using a permutation test first.

First let write the hypotheses

H_0: The variables sex and substance are independent.
H_a: The variables sex and substance are dependent.

We will use the chi squared test statistic.

```
obs <- chisq(substance~sex,data=HELPrct)
obs
```

```
## X.squared
##  2.026361
```

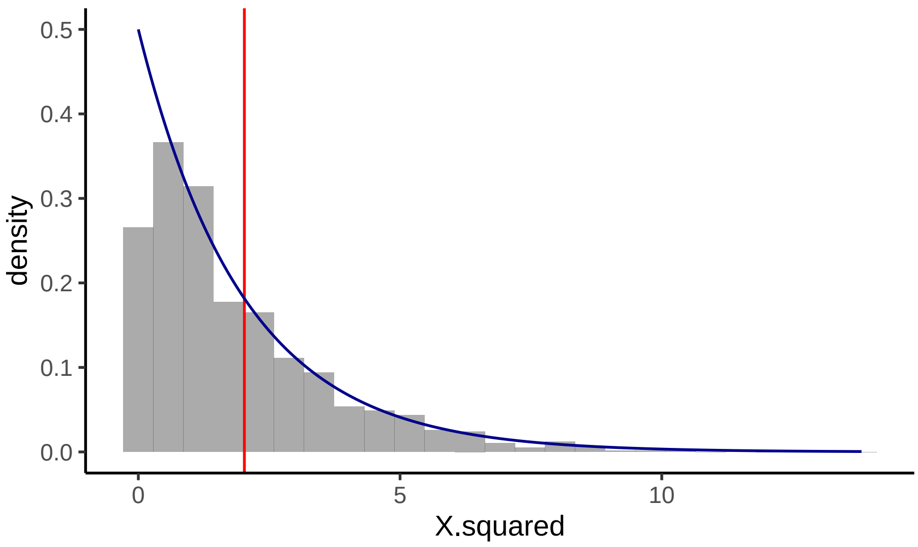
Next we will use a permutation randomization to find the sampling distribution.

```
set.seed(2720)
results <- do(1000)*chisq(substance~shuffle(sex),data=HELPrct)
```

A visual summary of the results helps us to gain some intuition about the p-value.

```
results %>%
  gf_dhistogram(~X.squared) %>%
  gf_vline(xintercept = obs,color="red") %>%
  gf_theme(theme_classic()) %>%
  gf_dist("chisq",df=2,color="darkblue")
```

```
## Warning: geom_vline(): Ignoring ‘mapping‘ because ‘xintercept‘ was provided.
```



We find the p-value using `prop1()`.

```
prop1((~X.squared>=obs),data=results)
```

```
## prop_TRUE
## 0.3536464
```

We don't double it because the chi-squared is a one sided test due to the fact that we squared the differences.

We fail to reject that the variables are independent.

Chi-squared test

We will jump straight to the function.

```
chisq.test(tally(substance~sex,data=HELPrct))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tally(substance ~ sex, data = HELPrct)
## X-squared = 2.0264, df = 2, p-value = 0.3631
```

We get a p-value very close to the one from the permutation test.

Notice that if the null hypothesis is true the test statistic is the minimum value of zero. We can't use a bootstrap confidence interval on this problem because zero will not be in the interval. It can be on the edge.

Numerical data

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance** (ANOVA) and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA with the F distribution:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

MLB batting performance

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (`OF`), infielder (`IF`), designated hitter (`DH`), and catcher (`C`). We will use a data set `mlbbat10` from the `openintro` package. The data is in the `mlb_obp.csv` which has been modified from the original data set to include only those with more than 200 at bats. The batting performance will be measured with the on-base percentage. The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

Read the data into `R`.

```
mlb_obp <- read_csv("data/mlb_obp.csv")
```

```
inspect(mlb_obp)
```

```
## Warning: ‘data_frame()‘ is deprecated as of tibble 1.1.0.
## Please use ‘tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

```
##
## categorical variables:
##       name     class levels   n missing
## 1 position character      4 327       0
##                                    distribution
## 1 IF (47.1%), OF (36.7%), C (11.9%) ...
##
## quantitative variables:
##      name   class   min    Q1 median     Q3   max     mean         sd   n
## ...1  obp numeric 0.174 0.309  0.331 0.3545 0.437 0.332159 0.03570249 327
##      missing
## ...1       0
```

Let's change `position` to a factor.

```
mlb_obp <- mlb_obp %>%
  mutate(position=as.factor(position))
```

```
favstats(obp~position,data=mlb_obp)
```

```
##   position   min      Q1 median      Q3   max      mean         sd   n missing
## 1        C 0.219 0.30000 0.3180 0.35700 0.405 0.3226154 0.04513175  39       0
## 2       DH 0.287 0.31625 0.3525 0.36950 0.412 0.3477857 0.03603669  14       0
## 3       IF 0.174 0.30800 0.3270 0.35275 0.437 0.3315260 0.03709504 154       0
## 4       OF 0.265 0.31475 0.3345 0.35300 0.411 0.3342500 0.02944394 120       0
```

The means for each group are pretty close to each other.

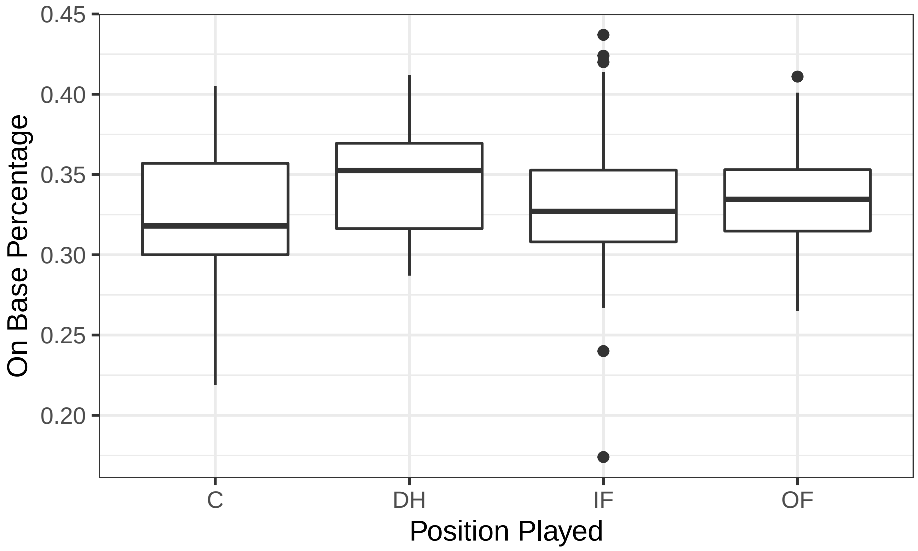
> **Exercise**: The null hypothesis under consideration is the following: $\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$. Write the null and corresponding alternative hypotheses in plain language.[1]

If we have all the data for the 2010 season, why do we need a hypothesis test? What is the population of interest?

If we are only making decisions or claims about the 2010 season, we do not need hypothesis testing. We can just use summary statistics. However, if we want to generalize to other years or other leagues, then we would need a hypothesis test.

> **Exercise**:
> Construct side-by-side boxplots.

```
mlb_obp %>%
  gf_boxplot(obp~position) %>%
  gf_labs(x="Position Played",y="On Base Percentage") %>%
  gf_theme(theme_bw())
```



[1] H_0: The average on-base percentage is equal across the four positions. H_A: The average on-base percentage varies across some (or all) groups.

The largest difference between the sample means is between the designated hitter and the catcher positions. Consider again the original hypotheses:

H_0: $\mu_{OF} = \mu_{IF} = \mu_{DH} = \mu_C$
H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of μ_{DH} and μ_C is statistically significant at a 0.05 significance level? The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally we would pick the groups with the large differences for the formal test, leading to an inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in means could have happened just by chance even if there was no difference in the respective population means.

Analysis of variance (ANOVA) and the F test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote,^[Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.] however, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error** (MSE)}, which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE are provided in the footnote[2]

[2] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total** (SST)} is computed as $SST = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$, where the sum is over all observations in the data set. Then we compute the **sum of squared errors** (SSE)} in one of two equivalent ways:

$$SSE = SST - SSG = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups. Using a permutation test, we could look at the difference in the mean squared errors.

We can use the F statistic to evaluate the hypotheses in what is called an **F test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. The F is really a ratio of chi-squared distributions. Using a permutation test we could look at the difference in variances.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

> The F statistic and the F test
> Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

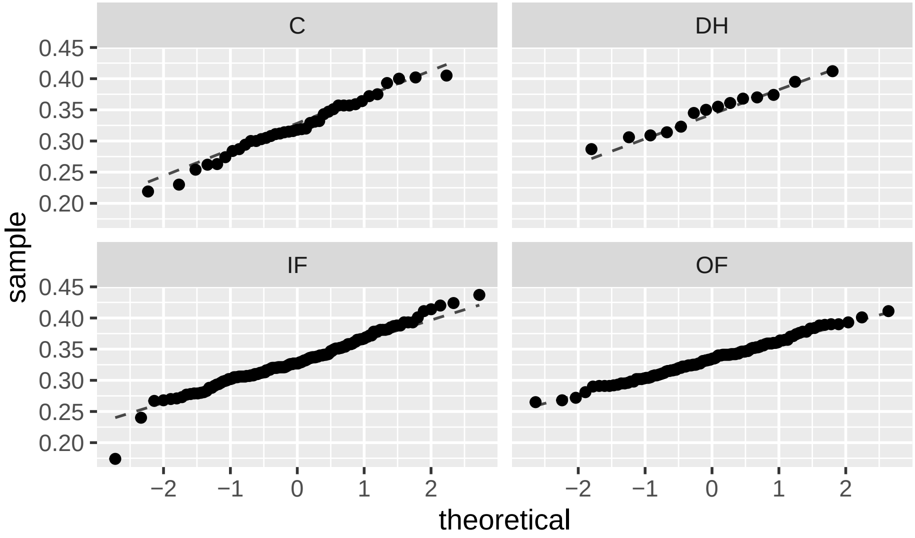
ANOVA We will use `R` to perform the calculations for the ANOVA. But let's check our assumptions first.

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

> **Independence**
> If the data are a simple random sample from less than 10% of the population, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). In our MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations. This is a bit of hand waving but remember independence is difficult to assess.

> **Approximately normal**
> As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small. The normal probability plots for each group of the MLB data are shown below; there is some deviation from normality for infielders, but this isn't a substantial concern since there are over 150 observations in that group and the outliers are not extreme. Sometimes in ANOVA there are so many groups or so few observations per group that checking normality for each group isn't reasonable. One solution is to combine the groups into one set of data. First calculate the **residuals** of the baseball data, which are calculated by taking the observed values and subtracting the corresponding group means. For example, an outfielder with OBP of 0.435 would have a residual of $0.435 - \bar{x}_{OF} = 0.082$. Then to check the normality condition, create a normal probability plot using all the residuals simultaneously.

```
mlb_obp %>%
  gf_qq(~obp|position) %>%
  gf_qqline()
```



Constant variance

The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes across the groups which we did previously. In this case, the variability is similar in the four groups but not identical. We also see in the output of `favstats` that the standard deviation varies a bit from one group to the next. Whether these differences are from natural variation is unclear, so we should report this uncertainty ofmeeting this assumption with the final results. The permutation test does not have this assumption and can be used to verify the results from the ANOVA.

In summary, independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

Let's write the hypotheses again.

H_0: The average on-base percentage is equal across the four positions.
H_A: The average on-base percentage varies across some (or all) groups.

The test statistic is the ratio of the between means variance and the pooled within group variance.

```
summary(aov(obp~position,data=mlb_obp))
```

```
##              Df Sum Sq  Mean Sq F value Pr(>F)
## position      3 0.0076 0.002519   1.994  0.115
## Residuals   323 0.4080 0.001263
```

The table contains all the information we need. It has the degrees of freedom, mean squared errors, test statistic, and p-value. The test statistic is 1.994, $\frac{0.002519}{0.001263} = 1.994$. The p-value is larger than 0.05, indicating the evidence is not strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data do not provide strong evidence that the average on-base percentage varies by player's primary field position.

The calculation of the p-value is

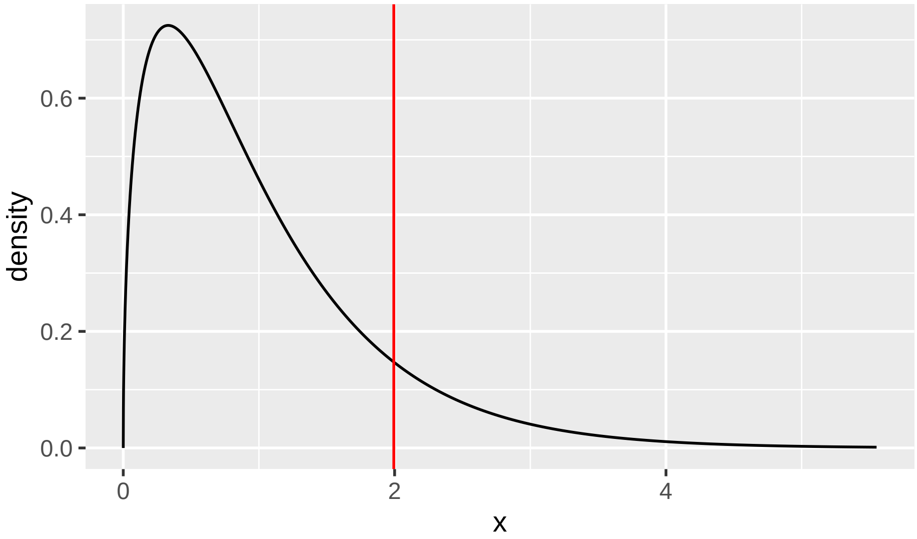
```
pf(1.994,3,323,lower.tail = FALSE)
```

```
## [1] 0.1147443
```

Let's look at a plot of the F distribution.

```
gf_dist("f",df1=3,df2=323) %>%
  gf_vline(xintercept = 1.994,color="red")
```

```
## Warning: geom_vline(): Ignoring ‘mapping‘ because ‘xintercept‘ was provided.
```



Permutation test We can repeat the same analysis using a permutation test. We will first run it using a ratio of variances and then for interest a difference in variances.

We need a way to extract the mean squared errors from the output. There is a package called `broom` that cleans up output from functions and makes them into dataframes.

```
library(broom)
```

```
aov(obp~position,data=mlb_obp) %>%
  tidy()
```

```
## # A tibble: 2 x 6
##   term         df   sumsq  meansq statistic p.value
##   <chr>     <dbl>   <dbl>   <dbl>     <dbl>   <dbl>
## 1 position      3 0.00756 0.00252      1.99   0.115
## 2 Residuals   323 0.408   0.00126     NA     NA
```

Let's summarize the values in the `meansq` column and develop our test statistic, we could just pull the statistic but we want to be able to generate a difference test statistic as well.

```r
aov(obp~position,data=mlb_obp) %>%
  tidy() %>%
  summarize(stat=meansq[1]/meansq[2])
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  1.99
```

Now we are ready. First get our test statistic using `pull()`.

```r
obs<-aov(obp~position,data=mlb_obp) %>%
  tidy() %>%
  summarize(stat=meansq[1]/meansq[2]) %>%
  pull()
obs
```

```
## [1] 1.994349
```

Let's put it into a function.

```r
f_stat <- function(x){
  aov(obp~shuffle(position),data=x) %>%
  tidy() %>%
  summarize(stat=meansq[1]/meansq[2]) %>%
  pull()
}
```

```r
f_stat(mlb_obp)
```

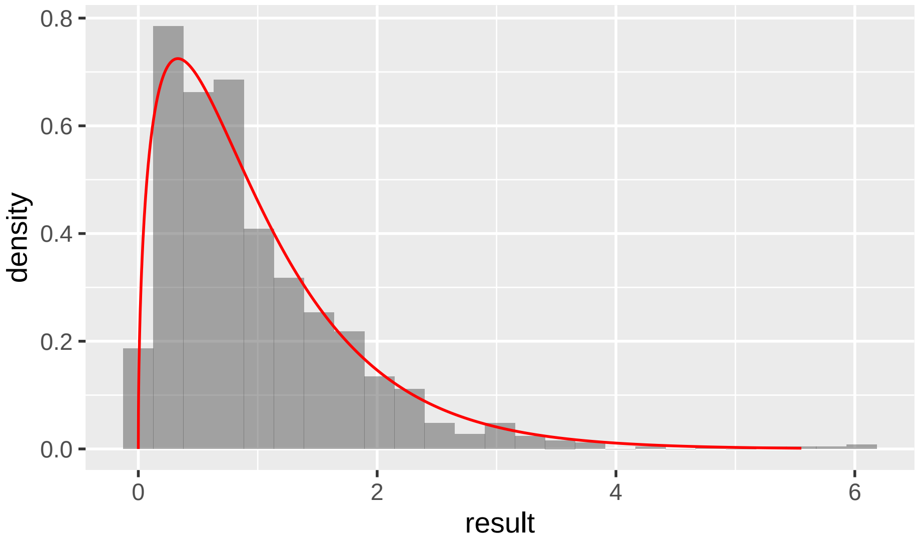
```
## [1] 0.4160649
```

Next, permute the position variable. There is an easier way to do this with the `purrr` package.

```r
set.seed(5321)
results<-do(1000)*(f_stat(mlb_obp))
```

That was slow in executing because we are using slow functions. We could optimize using `purrr`.

```r
results %>%
  gf_dhistogram(~result) %>%
  gf_dist("f",df1=3,df2=323,color="red")
```



The p-value is

```
prop1(~(result>=obs),results)
```

```
## prop_TRUE
## 0.0959041
```

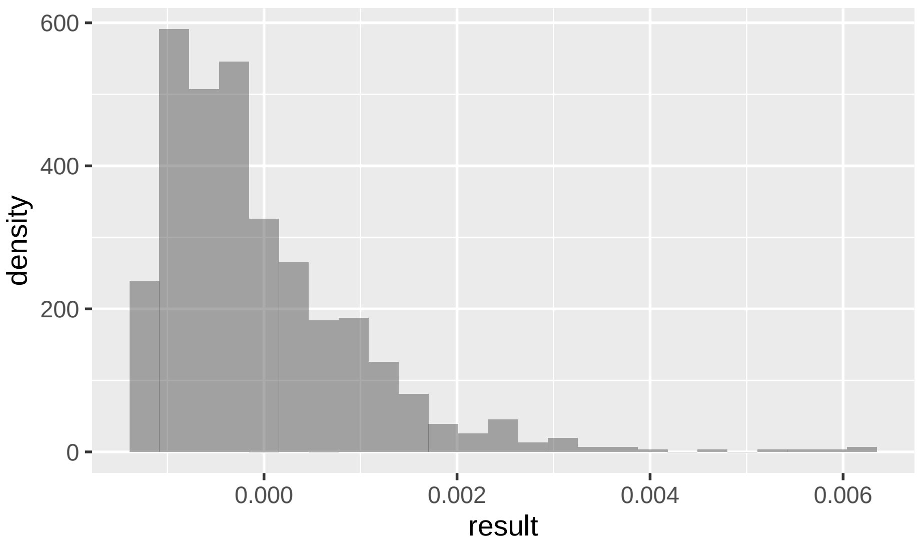
This is a similar p-value.

Repeat the analysis but use the difference in variance.

```
f_stat2 <- function(x){
  aov(obp~shuffle(position),data=x) %>%
  tidy() %>%
  summarize(stat=meansq[1]-meansq[2]) %>%
  pull(stat)
}
```

```
set.seed(5321)
results<-do(1000)*(f_stat2(mlb_obp))
```

```
results %>%
  gf_dhistogram(~result)
```



```
obs<-aov(obp~position,data=mlb_obp) %>%
  tidy() %>%
  summarize(stat=meansq[1]-meansq[2]) %>%
  pull(stat)
obs
```

```
## [1] 0.001255972
```

The p-value is

```
prop1(~(result>=obs),results)
```

```
## prop_TRUE
## 0.0959041
```

A similar p-value.

If we reject in the ANOVA test, we know there is a difference in at least one mean but we don't know which ones. How would you approach answering that question?

File Creation Information

- File creation date: 2020-07-28
- Windows version: Windows 10 x64 (build 18362)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.7.0
- `tidyverse` package version: 1.3.0