Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 1 due Friday 9/22

Introductions (again)

- Name
- One question or thought related to lecture last week (ANOVA, F-test, ranks)

Manipulating new distributions

Let $T_n \sim t_n$. Find the following:

1. Distribution of T_n^2. Hint: Think about the representation of T_n.

We can represent T_n as $Z/\sqrt{V_n/n}$ where Z is a standard Normal and V_n is a χ^2_n random variable. Then,

$$T_n^2 = Z^2/(V_n/n) = (Z^2/1)/(V_n/n) = F_{1,n}$$

since Z^2 has a χ^2_1 distribution and this is our representation of the F random variable.

2. Distribution of T^{-2}

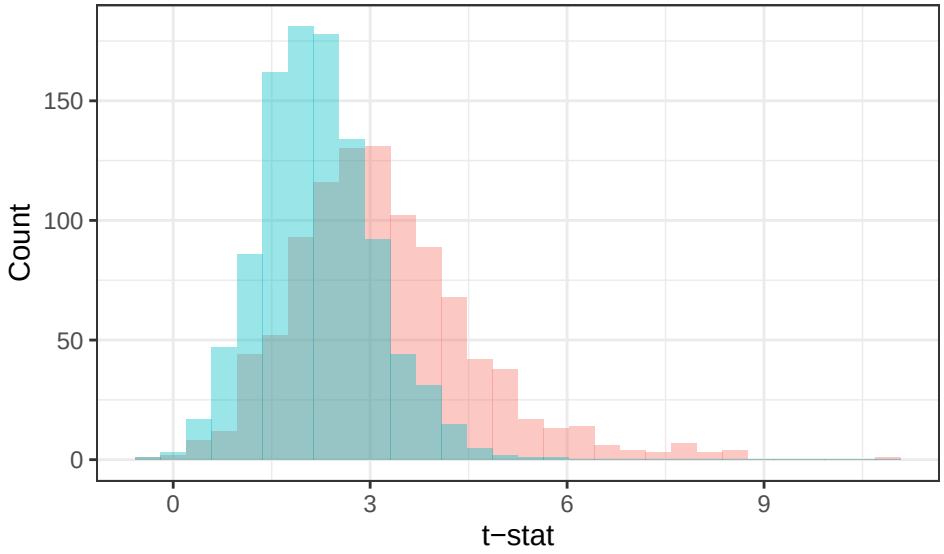
$$T_n^{-2} = F_{1,n}^{-1} = F_{n,1}$$

3. Let $X_1, ..., X_n \sim \text{Expo}(\alpha)$. Find the k (in terms of α) such that $k\sum_{i=1}^n X_i \sim \chi^2_{2n}$.

$$2\alpha \sum_{i=1}^n X_i \sim \text{Gamma}(n, 1/2) \sim \chi^2_{2n} \implies k = 2\alpha$$

Simulations

1. Let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Then, let $X_{i,1} = X_i + \epsilon_{i,1}$ and $X_{i,2} = X_i + \beta + \epsilon_{i,2}$ with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Suppose we simulate many paired and unpaired t-tests for the difference in the mean of the $X_{i,1}$s vs. the mean of the $X_{i,2}$s. If β is non-zero, which color is the paired t-test?

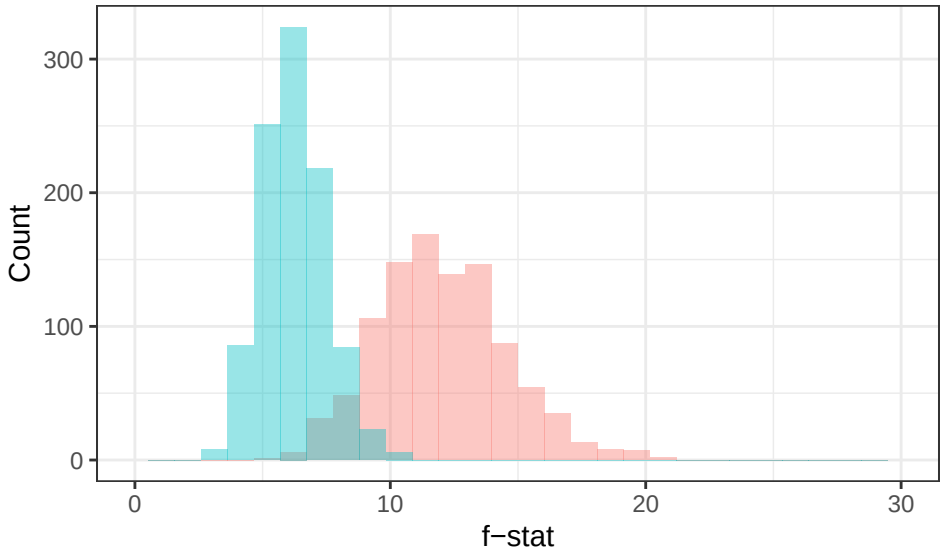


Because this is the proper set-up for a paired t-test and there is a difference in means ($\beta \neq 0$), the paired t-test will have higher power and therefore larger t statistics. Thus, the paired t-test is red.

2. Suppose we have some β_i for $i \in \{1, ..., n_\beta\}$ that are not equal. Let $X_{i,j} = \beta_i + \epsilon_{i,j}$ for $j = 1$ to n with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. We want to test whether $\beta_1 = \beta_2 = \cdots = \beta_{n_\beta}$. We'll run a simulation in which we consider two cases:

- In the first case, we use the proper groupings of the $X_{i,j}$; that is, there are n observations in each group, all with the same β_i.
- In the second case, we'll subdivide each of these groups into 2 so that there are $n/2$ observations in each group with two groups for each β_i.

We'll run an ANOVA in each case and repeat this many times. Which color is which case?



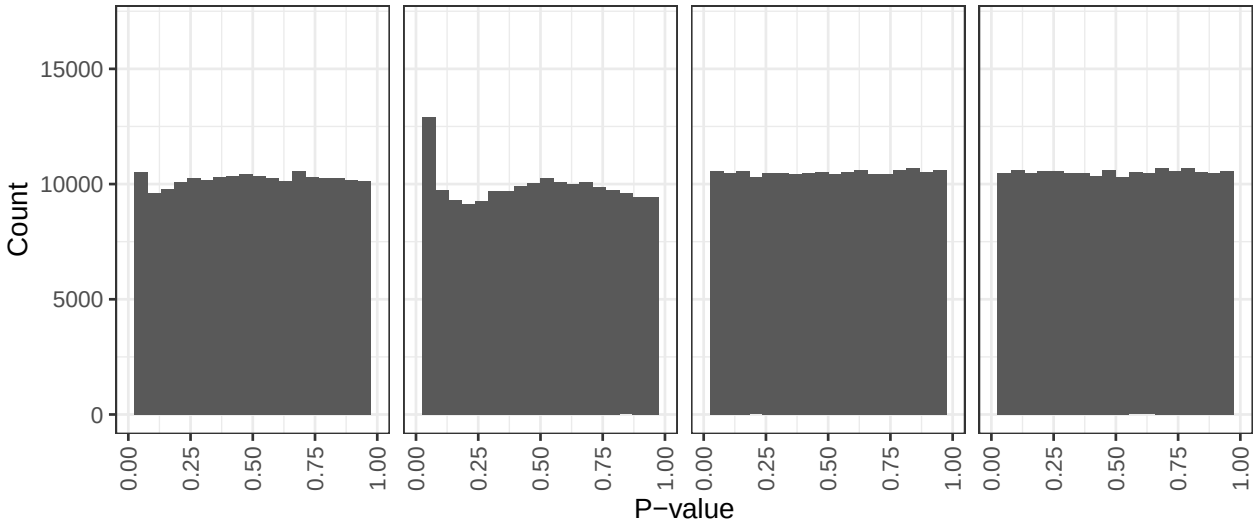
Since the group means in subdivided groups will be similar, the sum of squares within and between groups

doesn't change much. However, by increasing the number of groups, we're increasing the degrees of freedom in the between-group part and decreasing the degrees of freedom in the within-group part (k increases):

$$F = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_i - \bar{Y})^2/(k-1)}{\sum_{i=1}^{K}(n_i - 1)S_i^2/(n-k)}$$

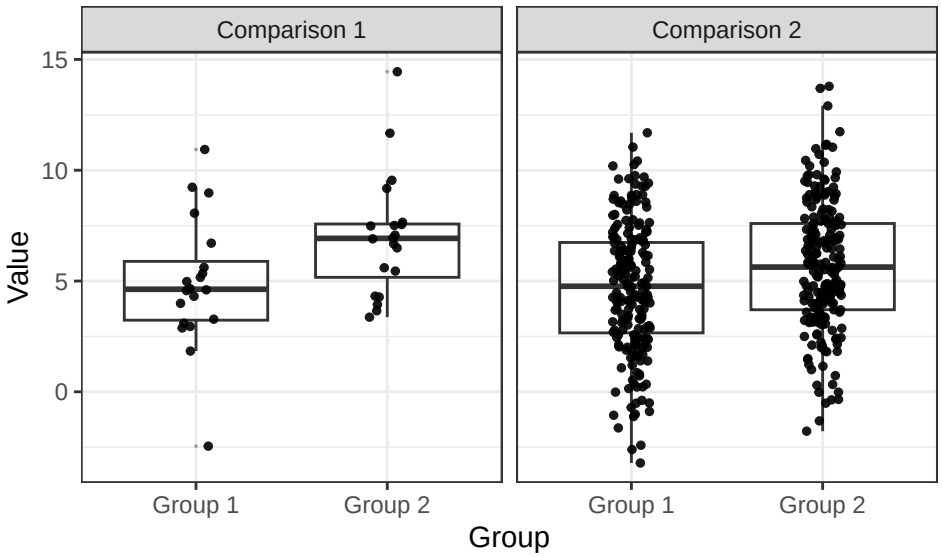
As k doubles, the f-statistic decreases. Thus, the blue is the subdivided groups.

3. Let $X_i \sim \mathcal{N}(0,1)$ for i from 1 to n. Let $Y_i \sim -1 + \text{Expo}(1)$ for i from 1 to n. Suppose we conduct a two-sided, one-sample t-test for $H_0 : \mu = 0$ vs. $H_a : \mu \neq 0$ and record the p-value. The plots below show p-values from simulations repeating this many times for the two distributions and $n = 5$ or $n = 20$. Identify which is which.



The point of this question is to notice that when the t-statistic has the t distribution, the p-values will be uniform. The t-statistic will have the t distribution by definition if the observations are Normal, so the third and fourth histograms can be either number of Normals. For the exponential distribution, the mean is 0, but the t distribution of the t-statistic relies on the Central Limit Theorem, so the larger n will give more uniform p-values. Thus, the first histogram is the exponential case with $n = 20$, and the second is the exponential distribution with $n = 5$.

4. Which of the two comparisons do you expect to have the lower p-value? The one with a larger difference in sample means or the one with more data points (40 vs 400)?



Even though the standard deviations are the same in both, what matters is the standard error, which is much lower in the second because of the increased sample size. The difference in sample means in the second comparison is about half that of the first comparison. However, the standard error of the second comparison is about $1/\sqrt{10} = 0.316$ times the standard error of the first comparison, so the t-statistic is about 1.58 times as large in the second comparison. Therefore, the second will have the lower p-value.

```
# Comparison 1
t.test(df$values[df$group == "Group 1" & df$comparison == "Comparison 1"],
       df$values[df$group == "Group 2" & df$comparison == "Comparison 1"])$p.value
```

```
## [1] 0.02832484
```

```
# Comparison 2
t.test(df$values[df$group == "Group 1" & df$comparison == "Comparison 2"],
       df$values[df$group == "Group 2" & df$comparison == "Comparison 2"])$p.value
```

```
## [1] 0.000780156
```

Variance by decomposition

Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$. Let $X + Y = r$.

1. Find the variance of $X|r$ by using the variance of a known distribution (See 3.9.2 in the Stat 110 book for a hint).

$X|r \sim \text{HGeom}(n, m, r)$, so by the variance of the hypergeometric we have:

$$\text{Var}(X|r) = \frac{n+m-r}{n+m-1} \cdot \frac{nr}{n+m} \cdot \frac{m}{n+m}$$

2. Find the variance of $X|r$ by using the fact that $\text{Var}(X + Y|r) = 0$ and treating X and Y as the sum of Bernoulli random variables. Verify that the two answers are the same. (Hint: Once you get to the Bernoulli random variables, think about how knowing the sum is r makes p irrelevant.)

First, note that the variance of a constant is 0, so $\text{Var}(X + Y|r) = 0$. Each of X and Y can be decomposed into Bernoullis, and each of these have the same variance and covariance by symmetry. Therefore,

$$0 = \text{Var}(X + Y|r) = (n + m)\text{Var}(I_1|r) + 2\binom{n+m}{2}\text{Cov}(I_1, I_2|r)$$

Then, we can solve for the covariance: $\text{Cov}(I_1, I_2|r) = -\text{Var}(I_1|r)/(n + m - 1)$. Conditioning on r, $P(I_1 = 1|r) = \frac{r}{n+m}$, so

$$\text{Var}(I_1|r) = \frac{r}{n+m} \cdot \frac{n+m-r}{n+m}$$

and

$$\text{Cov}(I_1, I_2|r) = -\frac{r}{n+m} \cdot \frac{n+m-r}{(n+m)(n+m-1)}$$

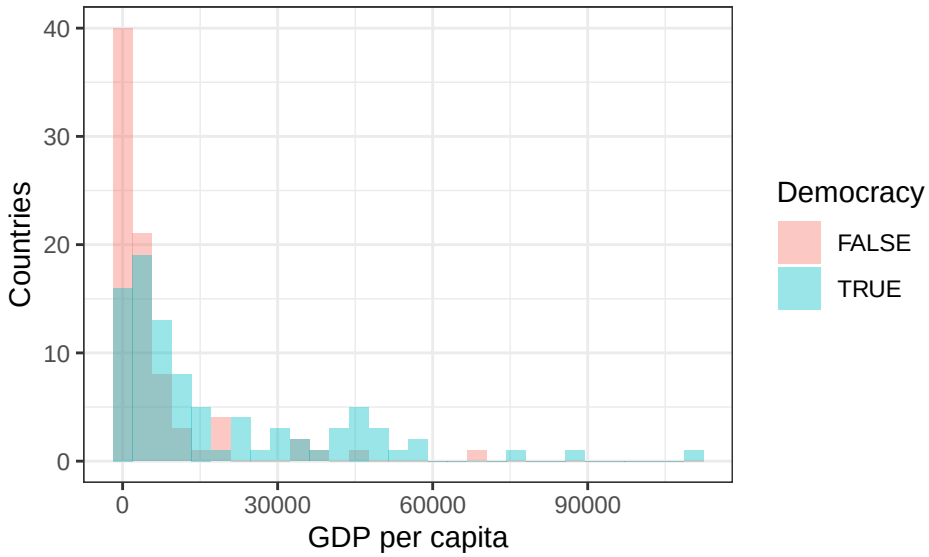
Then, building up X again,

$$\begin{aligned}
\text{Var}(X|r) &= \text{Var}(\sum_{i=1}^{n} I_i) \\
&= n \cdot \frac{r}{n+m} \cdot \frac{n+m-r}{n+m} - 2\binom{n}{2}\frac{r}{n+m} \cdot \frac{n+m-r}{(n+m-1)(n+m)} \\
&= \frac{nr(n+m-r)}{(n+m)^2}\left[1 - \frac{n-1}{n+m-1}\right] \\
&= \frac{nmr(n+m-r)}{(n+m)^2(n+m-1)}
\end{aligned}$$

Hypothesis testing on real data

These problems will deal with a dataset of country-level statistics from UNdata and Varieties of Democracy.

1. Suppose we want to test for a difference in mean 2010 GDP per capita between democracies and non-democracies. The following plots show the distributions. Which tests would be valid?



An unpaired t-test, a rank-sum test, or a log-transformed t-test would all be reasonable. The rank-sum test and log-transformed t-test would account for the fact that the data is not Normally distributed. However, we have enough data points that the sample means will be approximately normally distributed, so an unpaired t-test could work as well.

2. Perform a formal rank-sum test for the difference in GDP per capita between democracies and non-democracies.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dem_gdps and nondem_gdps
## W = 5443, p-value = 7.754e-08
## alternative hypothesis: true location shift is not equal to 0
```

Our hypotheses are H_0 : the distributions of 2010 GDP per capita are the same between democracies and non-democracies versus H_a : they are different. We get a test statistic of $W = 5443$ and a p-value of 7.75×10^{-8}, so we reject the null and conclude that democracies tend to have higher GDPs per capita.

3. Perform a formal log-transformed t-test for the difference in GDP per capita between democracies and non-democracies. Give a 95% confidence interval for the ratio of medians.

```
##
##  Welch Two Sample t-test
##
## data:  log(dem_gdps) and log(nondem_gdps)
## t = 5.8451, df = 169.64, p-value = 2.533e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8196649 1.6556519
## sample estimates:
## mean of x mean of y
##  9.015952  7.778294
```

Let μ_0 be the mean log GDP per capita of non-democracies and μ_1 be the mean log GDP per capita of democracies. We are testing H_0: $\mu_0 = \mu_1$ vs H_a : the two means are different. We get a t-statistic of -5.85 with 169.6 degrees of freedom, which corresponds to a p-value of 2.5×10^{-8}. Therefore, we reject the null and conclude that democratic countries have higher average log GDP per capita. To find a 95% confidence interval for the ratio of medians, we can exponentiate the current interval: $(\exp(0.82), \exp(1.66)) = (2.27, 5.24)$.

4. Suppose we wanted to test whether there was a difference in the mean number of doctors per country between 2019 and 2020. What would be a good way to do so?

We should use a paired t-test with pairing by country. We can check whether the mean difference in countries' doctor numbers between 2020 and 2019 is significantly non-zero. The t-test shows that the difference is not quite significant, but only 15 countries have data for both 2020 and 2019, so our test might just have not had enough power.

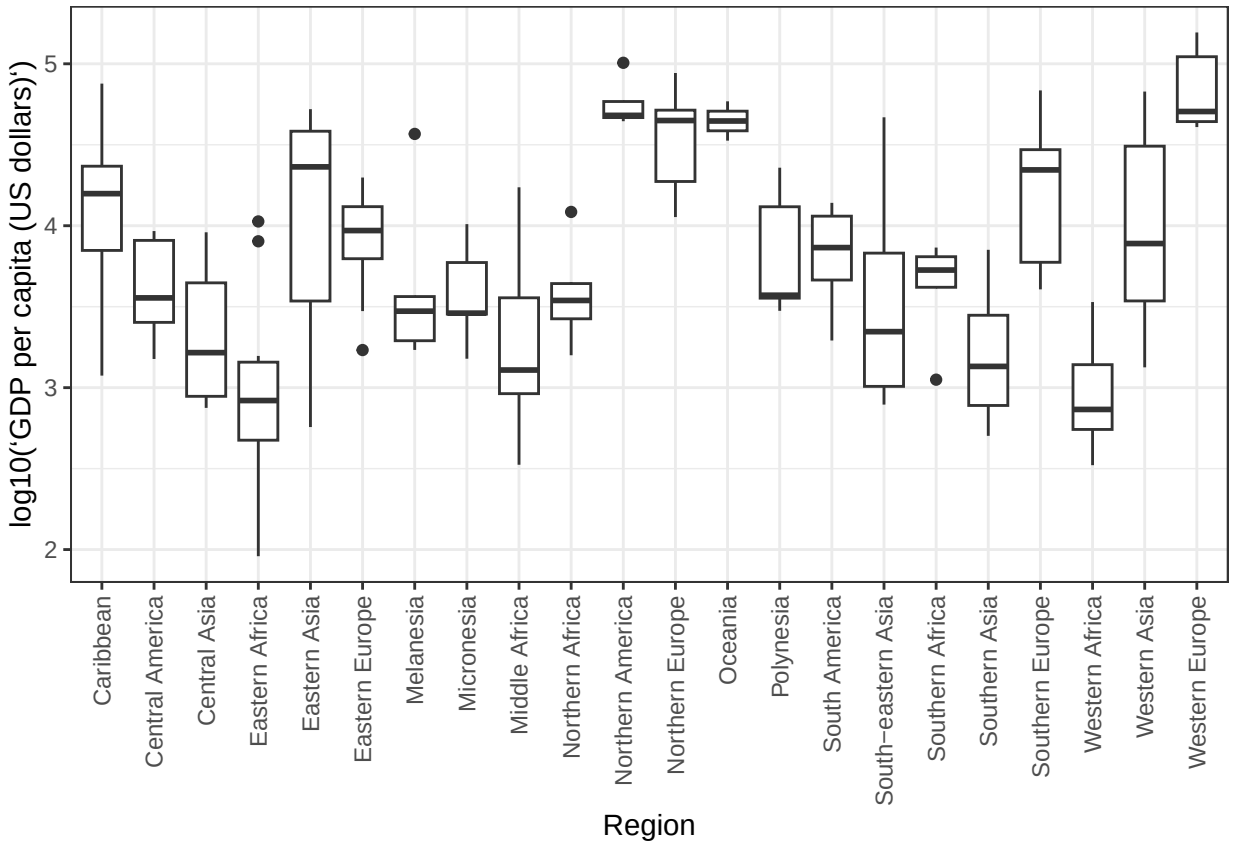
```
##
##  One Sample t-test
##
## data:  joined_df$`2020 doctors` - joined_df$`2019 doctors`
## t = 1.9379, df = 14, p-value = 0.07307
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -370.2314 7306.4981
## sample estimates:
## mean of x
##  3468.133
```

5. Perform a formal analysis of variance for the difference in 2010 log GDP per capita by world region.

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Region       21 6.977e+10 3.322e+09   12.84 <2e-16 ***
## Residuals   187 4.839e+10 2.588e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 23 observations deleted due to missingness
```

Let μ_k be the mean GDP per capita of countries in region k (indexed arbitrarily). We want to test H_0: $\mu_0 = \mu_1 = ... = \mu_k$ vs H_a: the means are not all equal. We get an F-statistic of 12.84 for 21 and 187 degrees of freedom for a p-value of less than 2×10^{-16}, suggesting that the mean GDPs per capita of different regions are not equal.

6. Comment on the assumptions of the test.




```
##                Region Variance Number
## 1           Caribbean     0.80     22
## 2     Central America     0.49      8
## 3        Central Asia     1.15      5
## 4      Eastern Africa     1.28     18
## 5        Eastern Asia     2.95      7
## 6      Eastern Europe     0.60     10
## 7           Melanesia     1.56      5
## 8          Micronesia     0.55      5
## 9       Middle Africa     1.72      9
## 10    Northern Africa     0.47      6
## 11   Northern America     0.15      4
## 12    Northern Europe     0.52     10
## 13            Oceania     0.16      2
## 14          Polynesia     0.84      5
## 15      South America     0.35     12
## 16 South-eastern Asia     2.17     11
## 17    Southern Africa     0.57      5
## 18      Southern Asia     0.96      9
## 19    Southern Europe     0.89     14
## 20     Western Africa     0.43     16
## 21       Western Asia     1.44     17
## 22     Western Europe     0.31      9
```

The GDPs are about symmetric, so normality is a reasonable assumption. The variances are quite different though (from 0.15 to 2.95; well beyond a 2x difference), so the equal variance assumption is violated. Independence probably does not hold either because the countries likely trade with each other, causing their

GDPs to be correlated both within and between groups.