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 (t-test, z-test, ANOVA, F-test)

Country demographics

We'll start by making last week's exploratory data analysis a bit more precise. These problems will deal with a data set of country-level statistics from UNdata and Varieties of Democracy.

1. We speculated that the Western African and Eastern African countries probably did not have a significant difference in means. Perform a formal t-test for the difference in population means between Western African and Eastern African countries. Recall that a formal test includes (1) the hypotheses, (2) the test statistic, (3) the p-value, and (4) the conclusion in the context of the problem.

```
##
##  Welch Two Sample t-test
##
## data:  west_african and east_african
## t = 0.12188, df = 24.688, p-value = 0.904
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.98061  22.49249
## sample estimates:
## mean of x mean of y
##  18.39294  17.13700
```

Let μ_0 be the mean population of countries in Western Africa and μ_1 be the mean population of countries in Eastern Africa. We are testing H_0: $\mu_0 = \mu_1$ vs H_a : the two means are different. We get a t-statistic of 0.12 with 24.7 degrees of freedom for a two-sided t-test, which corresponds to a p-value of $0.90 > 0.05$, so we fail to reject the null and do not have sufficient evidence to conclude that the mean populations of Western African and Eastern African countries are different. (The confidence interval for the difference in means is $(-20.0, 22.5)$, which includes 0, consistent with the t-test.)

2. Perform a formal z-test for the difference in the proportions of the populations that are nurses or midwives in the US versus the UK in 2010.

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(us_nurses_midwives, uk_nurses_midwives) out of c(us_pop, uk_pop)
## X-squared = 57941, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.003585282 0.003637892
## sample estimates:
##      prop 1      prop 2
## 0.012504975 0.008893388
```

Let p_{US} be the proportion of the population that are nurses or midwives in the US and p_{UK} be the equivalent proportion in the UK. We want to test $H_0 : p_{\text{US}} = p_{\text{UK}}$ vs. H_a : they are not equal. We get a z-statistic of $\sqrt{57941} = 240.7$ which gives a p-value less than 2.2×10^{-16}, so we reject the null and conclude that the proportion of the population that are nurses and midwives is significantly higher in the US. Note that `prop.test` performs a χ^2 test, so you need to take the square root of the test statistic to get the z-test test statistic.

3. Suppose we wanted to test whether there was a change in the mean number of doctors per country between 2019 and 2020 (e.g., in response to COVID-19). What would be a good way to do so?

We should use a paired t-test with pairing by country to 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.

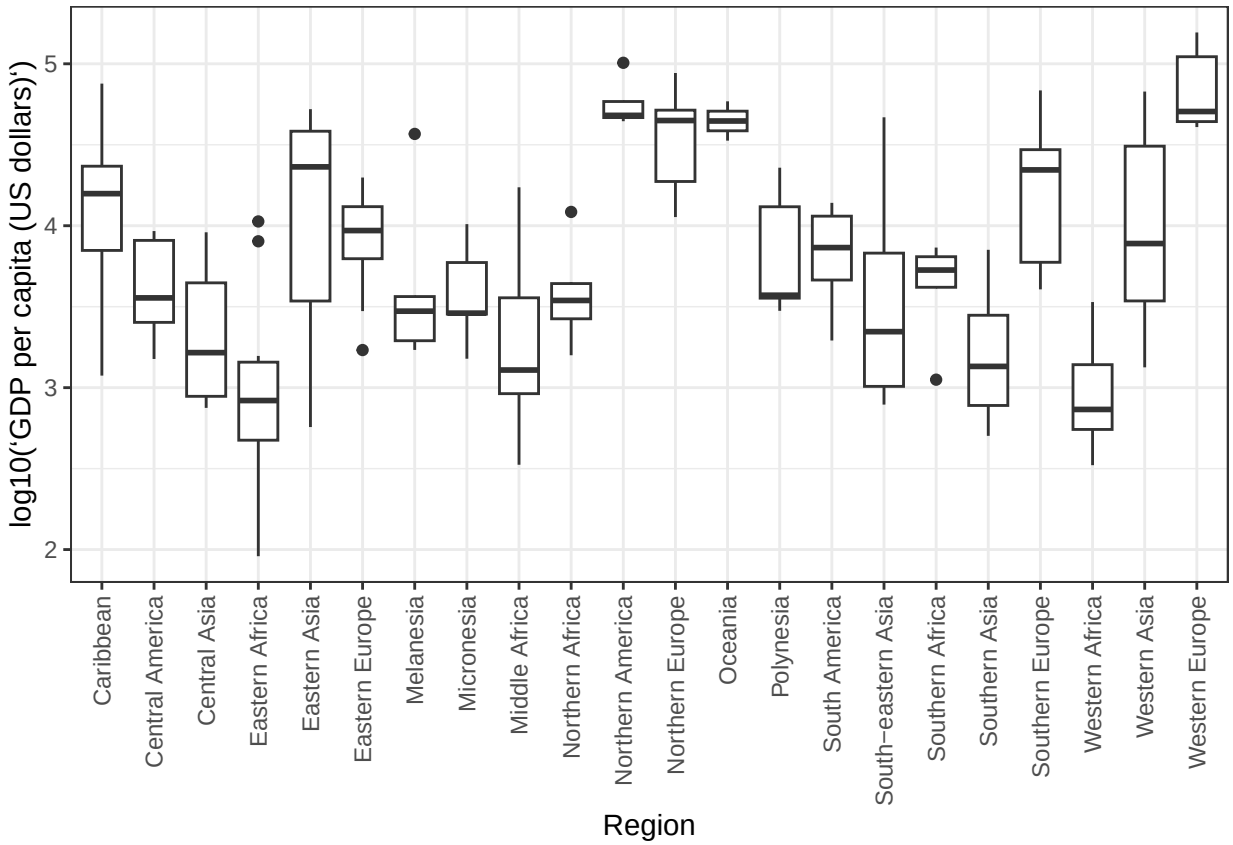
```
##
##  One Sample t-test
##
## data:  joined_df$`2020 doctors` - joined_df$`2019 doctors`
## t = 1.9379, df = 14, p-value = 0.07307
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -370.2314 7306.4981
## sample estimates:
## mean of x
##  3468.133
```

4. 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: $\mu_1 = ... = \mu_{22}$ 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.

5. Comment on the assumptions of the test.




```
##                Region Variance Number of countries
## 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.

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_n^2 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_1^2 distribution and this is our representation of the F random variable.

2. Distribution of T^{-2}

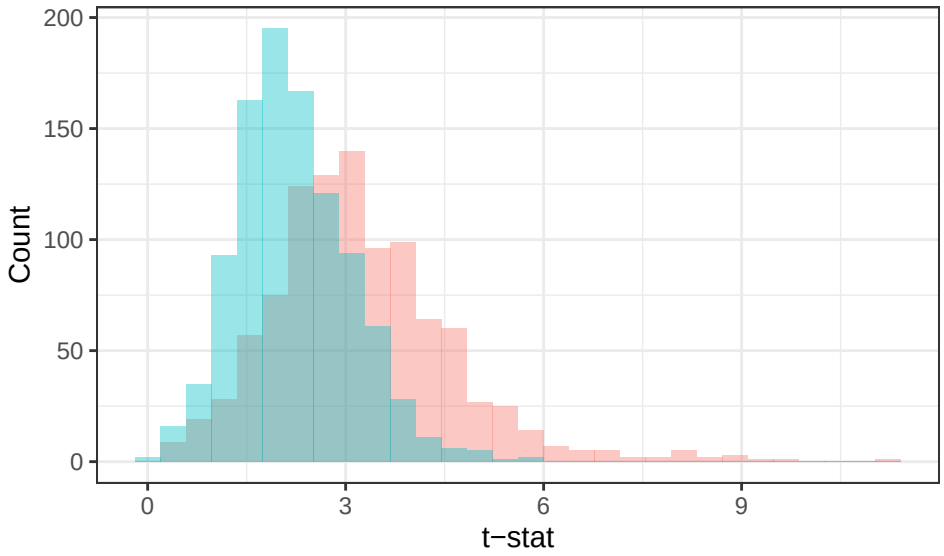
$$T_n^{-2} = F_{1,n}^{-1} = F_{n,1}$$

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

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

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?

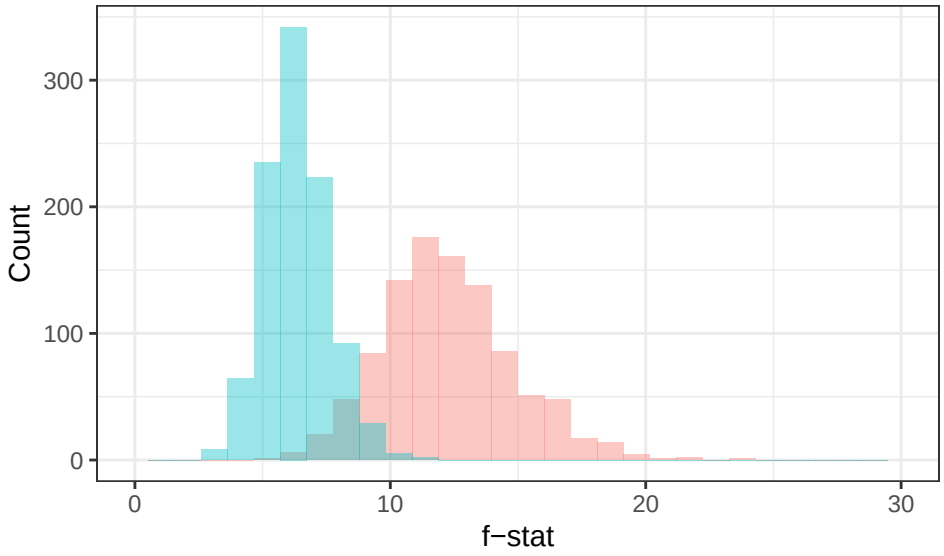


Because this is the proper set-up for a paired t-test and there is a difference in means ($\beta \neq 0$), the paired t-test will have higher power and therefore larger t statistics. Thus, the paired t-test is red.

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?



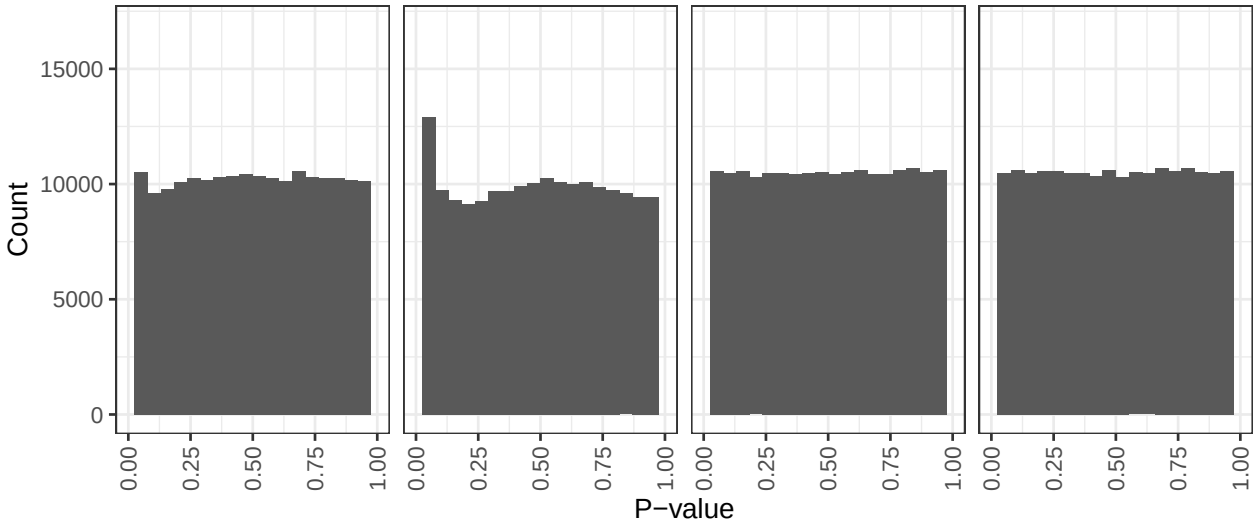
Since the group means in subdivided groups will be similar, the sum of squares within and between groups

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):

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

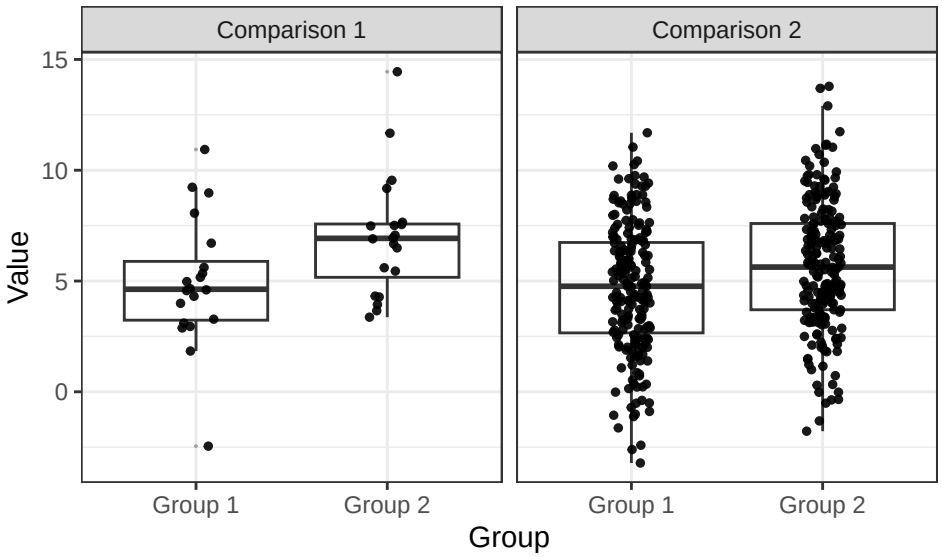
As k doubles, the f-statistic decreases. Thus, the blue is the subdivided groups.

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 $\sqrt{40/400} = 0.316$ 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
```