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:

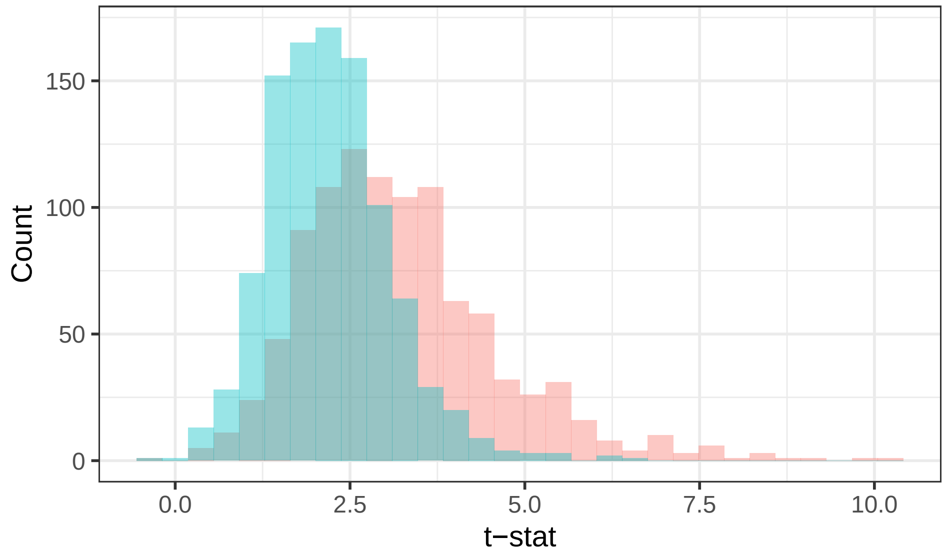
1. Distribution of T_n^2. Hint: Think about the representation of T_n.

2. Distribution of T^{-2}

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

Simulations

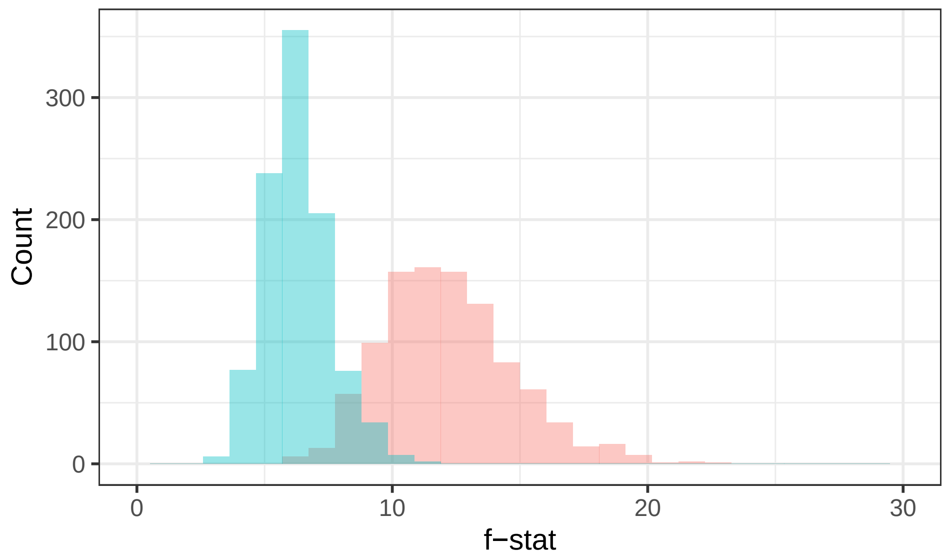
1. Let $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Then, let $X_{i,1} = X_i + \epsilon_{i,1}$ and $X_{i,2} = X_i + \beta + \epsilon_{i,2}$ with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Suppose we simulate many paired and unpaired t-tests for the difference in the mean of the $X_{i,1}$s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



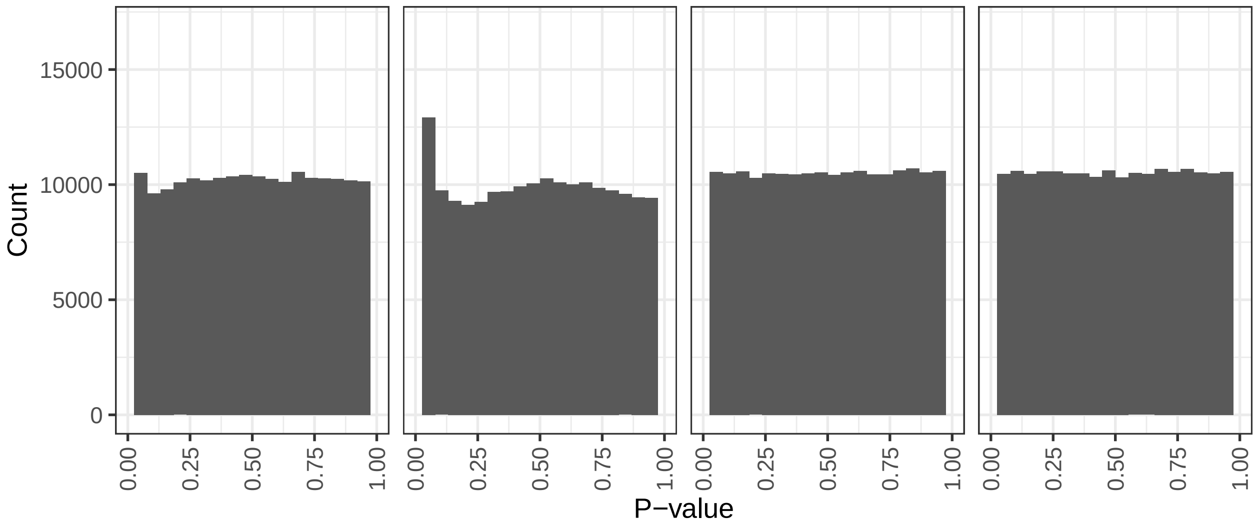
2. Suppose we have some β_i for $i \in \{1, ..., n_\beta\}$ that are not equal. Let $X_{i,j} = \beta_i + \epsilon_{i,j}$ for $j = 1$ to n with $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$. We want to test whether $\beta_1 = \beta_2 = \cdots = \beta_{n_\beta}$. We'll run a simulation in which we consider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

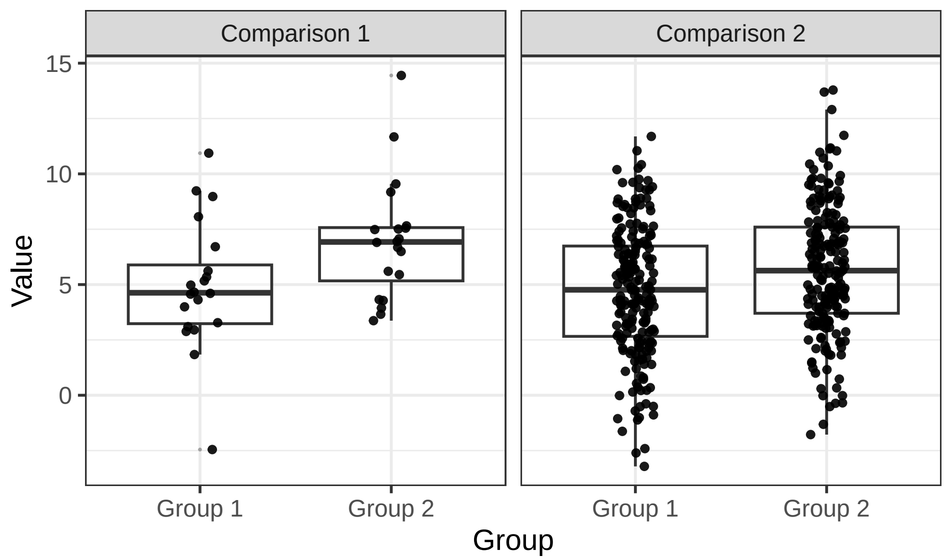
We'll run an ANOVA in each case and repeat this many times. Which color is which case?



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.



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)?



Variance by decomposition

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

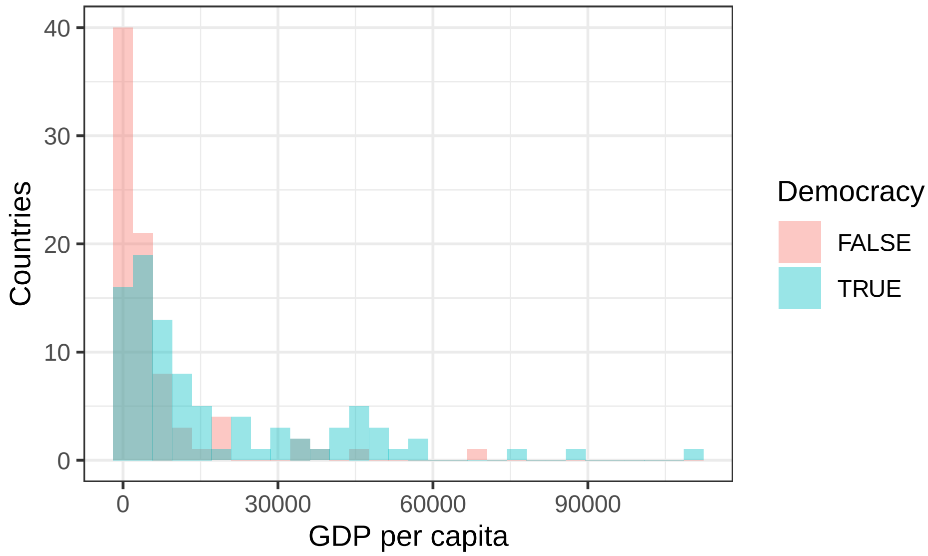
1. Find the variance of $X|r$ by using the variance of a known distribution (See 3.9.2 in the Stat 110 book for a hint).

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.)

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?



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
```

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.

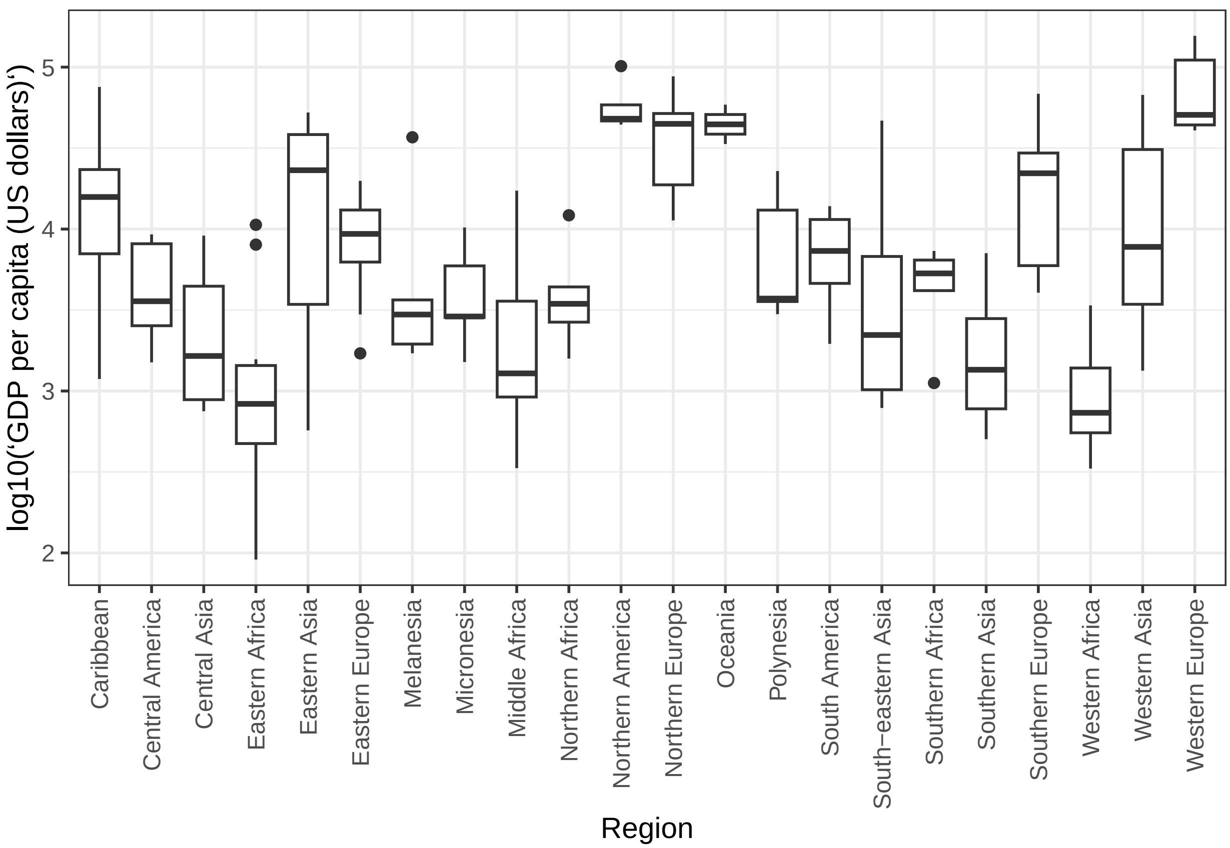
```
##
##  Welch Two Sample t-test
##
## data:  log(dem_gdps) and log(nondem_gdps)
## t = 5.8451, df = 169.64, p-value = 2.533e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8196649 1.6556519
## sample estimates:
## mean of x mean of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

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?

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
```

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
```