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 2 due Saturday 9/30

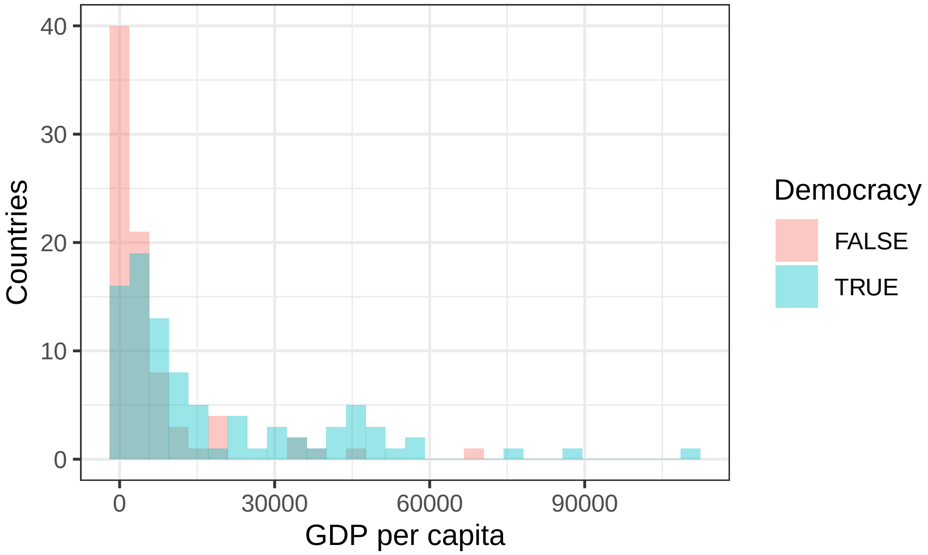
Introductions (again)

- Name
- One question or thought related to lecture last week (ranks, bootstrap, randomization)

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, a log-transformed t-test, or a permutation test would all be reasonable. The rank-sum test, log-transformed t-test, and permutation 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: $(\exp(0.82), \exp(1.66)) = (2.27, 5.24)$.

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,

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

Everything everywhere all at once (all two-sample continuous comparisons)

Let $X_1, \ldots X_{n_1} \sim \text{Exp}(1/\mu_1)$ and $Y_1, \ldots Y_{n_2} \sim \text{Exp}(1/\mu_2)$.

1. Name three tests we've learned so far that would not be applicable for comparing the Xs and Ys.

ANOVA (only two groups), paired t-test (different number of observations in each group), proportion test (not proportions)

2. We'll be comparing the Type I and II error for the following tests: a two sample t test, a log-transformed t-test, a rank-based test, and a permutation test. We'll consider two scenarios:

- First, $n_1 = 5$, $n_2 = 15$ with $\mu_1 = \mu_2 = 5$ when calculating the Type I error rate and $\mu_1 = 5$ and $\mu_2 = 3$ when calculating the Type II error rate.
- Second, the same as before but with $n_1 = 20$.

Why should we use $\mu_1 = \mu_2$ when calculating the Type I error rate but $\mu_1 \neq \mu_2$ when calculating the Type II error rate?

A type I error is when we reject the null but we should have retained it. We should retain the null when μ_1 and μ_2 are indeed equal. A type II error is when we fail to reject the null when we should have rejected it. Therefore, the means must be different to make this error.

3. Compare the results. What has the highest power? Which maintain their nominal false positive rates? Which test is best in which situations?

Test (First scenario)	Type I	Type II
t test	0.074	0.949
log t test	0.061	0.840
Rank test	0.043	0.885
Permutation test	0.053	0.780

Test (Second scenario)	Type I	Type II
t test	0.044	0.696
log t test	0.044	0.796
Rank test	0.047	0.774
Permutation test	0.051	0.743

In the first scenario where we have a very small sample size, the t-test is slightly above its nominal Type I error rate (0.05), and its type II error is the highest in the group (which means it has the lowest power). However, its assumptions are severely violated by using a very right-skewed distribution with only 5 data points. The other tests are closer to their nominal Type I error rates (0.05) with the permutation test having the highest power.

In the second scenario where we have reasonable sample sizes, all the tests have about the same Type I error rate (near 0.05), but the t test has the highest power. The takeaway is that even with skewed distributions, as long as the sample size is at least moderate, the t test will have the highest power while maintaining its nominal Type I error rate.

4. What assumptions do we need for each test and what hypotheses are we testing?

- t-test
  - Assumptions: independence within and between groups and normality (or a large sample size)
  - Hypotheses: H_0 : Means are equal; H_a: Means are not equal.

- Log-transformed t-test
  - Assumptions: independence, symmetry once transformed, and normality once transformed (or a large sample size)
  - Hypotheses: H_0 : Ratio of medians is 1; H_a : Ratio of medians isn't 1.
- Rank-based test
  - Assumptions: independence; $n_1, n_2 \geq 10$
  - Hypotheses: H_0 : The two distributions are the same, H_a: The two distributions are different.
- Permutation test
  - Assumptions: independence
  - Hypotheses: H_0 : The distribution of outcomes is not related to group status. H_a : It is related.

5. The following simulation uses the same set-ups as above to calculate a t-based confidence interval for the difference in means, a t-based confidence interval for the ratio of medians, a percentile bootstrap interval for the difference in means, and a reversed percentile bootstrap interval for the difference in means. Shown below are the coverage probability and interval width for each. Comment on the results.

Interval (First scenario)	Coverage probability (means different)	Interval width (means different)	Coverage probability (means same)	Interval width (means same)
t interval	0.902	10.83	0.935	11.38
Transformed t interval	0.939	8.71	0.942	5.25
Percentile bootstrap	0.821	7.30	0.860	8.33
Rev. Perc. bootstrap	0.824	7.30	0.876	8.33

Interval (Second scenario)	Coverage probability (means different)	Interval width (means different)	Coverage probability (means same)	Interval width (means same)
t interval	0.951	5.40	0.953	6.81
Transformed t interval	0.953	3.80	0.949	2.28
Percentile bootstrap	0.915	4.92	0.915	6.18
Rev. Perc. bootstrap	0.925	4.92	0.932	6.18

When one of the samples is very small and there's a difference, all of the intervals except the interval for the ratio of medians show coverage probabilities quite a bit below their nominal levels. Interestingly, the percentile methods are even worse than the t interval. When the means are the same, all the methods perform a bit better. The interval widths correlate with the coverage probability: smaller intervals have lower coverage probabilities.

When the samples are larger, the coverage probabilities are closer to their nominal levels, but the percentile methods still give intervals that are too narrow and therefore have slightly lower coverage probabilities. Note that in both scenarios the transformed t interval is trying to capture something different than the other intervals, explaining its smaller width.

6. Based on the results above, which is the best confidence interval to use?

The t or transformed t have coverage probabilities closest to the nominal confidence level, so we should use those. Shorter intervals would be nice if the intervals maintained the same coverage probabilities, but otherwise they're just not calibrated and therefore not useful.

7. Imagine now that we wanted to construct a confidence interval for μ_1 by using a studentized bootstrap interval. If we knew the data were distributed exponentially, what's one small change we could make to the confidence interval for μ_1 so that the interval is equally as wide or narrower while keeping the same confidence level?

Make the lower bound 0 if it's ever negative in the studentized bootstrap interval.