

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 Friday 9/29

Introductions (again)

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

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

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

1. Name three tests we've learned so far that would not be applicable for comparing the X s and Y s.
2. We'll be comparing the Type I and II error for the following tests: a one 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?

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

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

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

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

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?

Slope independent of outcome mean

In this problem, we'll show that the slope in a linear regression ($\hat{\beta}_1$) is independent of the mean outcome (\bar{Y}). Suppose we have pairs (X_i, Y_i) for $i \in \{1, \dots, n\}$.

1. Recall that in a simple linear regression we assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with X_i known and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The vector $(\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y})^T$ has a multivariate Normal distribution. Find the covariance of \bar{Y} and $Y_i - \bar{Y}$.

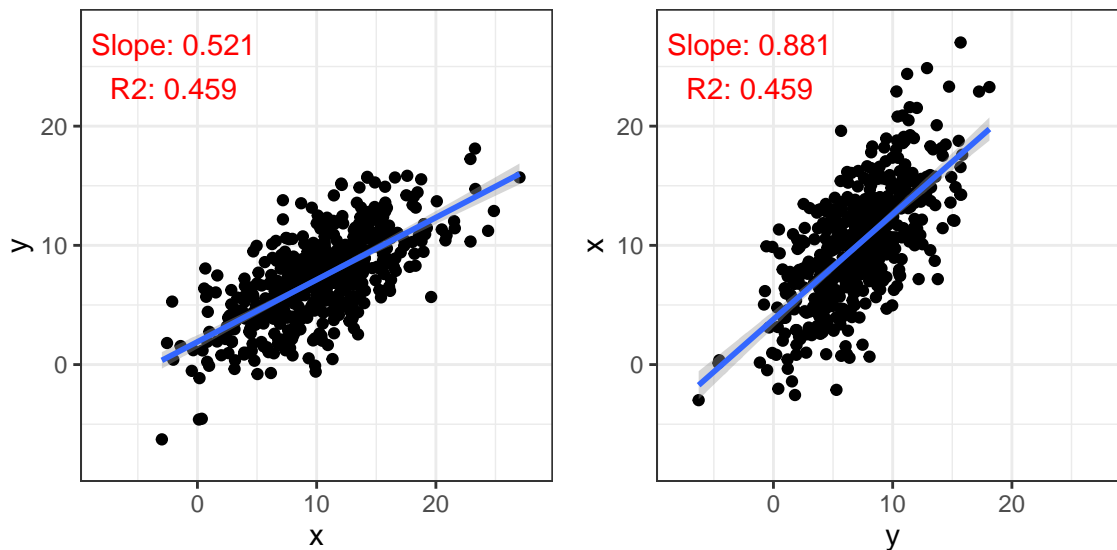
2. What does this imply about \bar{Y} and all the $Y_i - \bar{Y}$?

3. What does this say about the relationship between \bar{Y} and $\hat{\beta}_1$? Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Rule of thumb

Suppose we have n pairs of (X_i, Y_i) and we regress Y on X to get a slope $\hat{\beta}_1$ and X on Y to get a slope $\hat{\beta}'_1$. At first glance, it might seem like the $\hat{\beta}_1 = 1/\hat{\beta}'_1$. However, as you can see in the plots below, this is wrong.



1. Why is this wrong?
2. In the rest of the problem, we'll try to find the proper relationship between the two slopes. Recall that when regressing Y on X , we have

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Consider our simple regression with the estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and consider the flipped regression estimators

$$\hat{\beta}'_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \hat{\beta}'_0 = \bar{X} - \hat{\beta}'_1 \bar{Y}$$

Find an expression for $\hat{\beta}'_1$ in terms of $\hat{\beta}_1$.

3. Solve for R^2 in terms of $\hat{\beta}_1$ and $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. You may use the fact that

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

(See my Stat 111 section 6 notes for why this is the case in simple linear regression.)

4. Use this to write an expression for $\hat{\beta}_1'$ in terms of R^2 and $\hat{\beta}_1$.

Real data linear model

These problems will deal with a dataset of country-level statistics from [UNdata](#) and [Varieties of Democracy](#).

1. Suppose we want to know the relationship between log 2010 GDP per capita and the 2010 life expectancy for females at birth. Interpret the following output.

```
##
## Call:
## lm(formula = `Life expectancy at birth for females (years)` ~
##     log2(`GDP per capita (US dollars)`), data = countries_2010)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3076  -2.2879   0.8095   3.4083  12.2565
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   31.3250     2.2198   14.11  <2e-16 ***
## log2(`GDP per capita (US dollars)`)  3.3499     0.1753   19.11  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.622 on 207 degrees of freedom
## (23 observations deleted due to missingness)
## Multiple R-squared:  0.6382, Adjusted R-squared:  0.6365
## F-statistic: 365.2 on 1 and 207 DF,  p-value: < 2.2e-16
```

2. Suppose we read this result in a paper but what we actually cared about was the regression of log2 GDP per capita on female life expectancy at birth. What can we conclude about this alternative regression?