Module 2: Hypothesis Testing

# Reading

*Read*

* (Read first 2 pages) -- "The probable error of the mean" A. Student
* The ASA Statement on Statistical Significance and P-Values
* (OPTIONAL) -- Moving to a World Beyond “p < 0.05”
* (Cody) Chapter 7: One-Sample Tests
* (Cody) Chapter 8: Two-Sample Tests
* (OpenIntro) Chapter 7: Inference for numerical data

*Watch*

* Guinness, Student, and the History of the T Test -- <https://youtu.be/bqfcFCjaE1c>

Exercises:

* (Cody) Chapter 8, exercise #2 (pg. 90) [Bring completed problem code to residency meeting]

# Discussion for this week: (Response due by October 18, 2020)

*Prompt:* Read [“Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias”](https://doi.org/10.1111/cts.12178) and write a post addressing at least one fact, opinion, or conclusion presented in that article. Consider how ‘big data’ (really big samples) may have on statistical reasoning and how this could impact YOUR future work. Most fields today are working in the realm of sample sizes much larger than the Guinness brewery in 1908. Are the traditional statistical tests and methods still applicable today?

POST RESPONSE TO [#dan602\_discuss](https://stonehillmpsda21.slack.com/archives/C01A1MZUGDB) in Slack

# SAS Workshop: Hypothesis Testing

*Goal:* Use SAS Studio to perform statistically sound hypothesis testing

*Methods:*

Book Data: One Sample Testing

\*Refer to Chapter 7 (Cody) for help finding these options in SAS Studio\*

Adapted from Chapter 7 Exercise 4 and 5 from Cody (pg 72)

1. Set up a SAS Program file with the code: (Type, don’t copy)

data Difference;

call STREAMINIT(13579);

do Subj = 1 to 20;

Diff = 0.6 - rand('uniform');

output;

end;

run;

1. Run a 1-sample (1-way) t-test to determine if the difference scores (Diff) simulated here come from a population whose mean is zero. Report your results here. Are the data normal? Is it OK to use a t-test? Why or why not?
2. Edit the code above to run for 200 and 2000 iterations (e.g., Subj = 1 to 200). Does this change the results? Are the data normal? Is it OK to use a t-test? How does the p-value of the test you did the first time change?
3. If you decide for any of the above tests that the t test is NOT appropriate for these data what non-parametric (do not assume normality) test would you choose to report. Show this analysis and result for ONE of the above simulated tests.

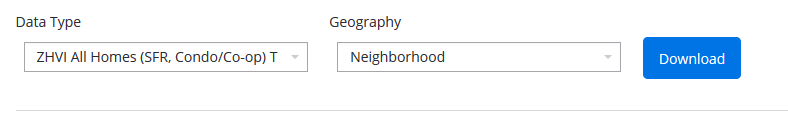
**Put all of the code into a single program file (if not already written that way) and save this file and turn it in with your answers to the questions above.**

Workshop: Zillow Home Value Index Data

1. Download data:   
     
   <https://www.zillow.com/research/data/>   
     
   **Zillow Home Value Index (ZHVI):** A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. *It reflects the typical value for homes in the 35th to 65th percentile range*. The raw version of that mid-tier ZHVI time series is also available.  
     
   Check out [this overview of ZHVI](https://www.zillow.com/research/zhvi-methodology-2019-highlights-26221/) and [a deep-dive into its methodology](https://www.zillow.com/research/zhvi-methodology-2019-deep-26226/).

Zillow publishes top-tier ZHVI ($, typical value for homes within the 65th to 95th percentile range for a given region) and bottom-tier ZHVI ($, typical value for homes that fall within the 5th to 35th percentile range for a given region).

**Download the ZHVI seasonally adjusted data for All homes at the Neighborhood level.**



\*This download will return a CSV file. Do not OPEN IT IN EXCEL.

IMPORT the data file into SAS Studio using your preferred library organization. Suggested use a course library (e.g., STATS\_IMP) as in past classes

FILENAME REFFILE 'C:/User/you/path/to/data\_folder/Neighborhood\_zhvi\_uc\_sfrcondo\_tier\_0.33\_0.67\_sm\_sa\_mon.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=STAT\_IMP.ZHVI;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=STAT\_IMP.ZHVI; RUN;

1. Filter and transform:  
     
   a. Filter for rows with ‘StateName EQ “MA”’ (type don’t copy)

proc sql noprint;

create table WORK.filter as select \* from STAT\_IMP.ZHVI where(StateName EQ

"MA");

quit;

b. Recode values to create a new column to split the data into two categories of “Metro Boston” and “NotBoston” (use the Metro column to categorize on “Boston-Cambridge-Newton”)

data WORK.RECODE;

length ISBOSTON $ 30;

set WORK.FILTER;

select (Metro);

when ('Boston-Cambridge-Newton') ISBOSTON='Boston';

otherwise ISBOSTON="NotBoston";

end;

run;

1. Set up a t-test to see if there is a difference in the ZHVI most recently reported (for 2020-08-31) between Boston and the rest of the state (Boston/NotBoston)

Is the assumption of normality met? Can you use the t-test?

On the OPTIONS tab choose to perform the **Wilcoxon rank-sum test** (a non-parametric test that works for non-normally distributed data). Does the p-value reported for the Wilcoxon test change any conclusions relative to the t-test?

**Turn in answers to questions and graphs ONLY. Keep program files for your reference.**

# Post-Class: Zillow Data exploration

OK… So it may not be surprising that home values are higher in Boston than the rest of the state. There are more compelling questions to ask of these data. Explore the data in the Zillow datasets at <https://www.zillow.com/research/data/>. Identify one other question that can be answered with these data using a one or two sample test. Then develop a SAS program that runs the appropriate data preparation and statistical tests.

Report your code and findings to the [#dan602\_discuss](https://stonehillmpsda21.slack.com/archives/C01A1MZUGDB) channel. Discuss your results and which tests ended up being appropriate to answer your question.

**Include a brief report of what you did and the results of the test you chose.**