Statistical Tests

INFO 370

Learning Objectives

Consider how to develop **segregation metrics** for assignment 2

Describe **t-tests** as a method for measuring differences in means across groups

Notebook part 2: Do male and female students **perform similarly** in an Informatics course?

Segregation

How can we measure how segregated a city is?

	GEO.display.label	pop	pop.white	pop.not.white	pct.white	pct.not.white
1	Census Tract 101, Baltimore city, Maryland	1378	1276	102	0.9259796...	0.07402032
2	Census Tract 102, Baltimore city, Maryland	1262	1082	180	0.8573692...	0.14263074
3	Census Tract 103, Baltimore city, Maryland	965	847	118	0.8777202...	0.12227979
4	Census Tract 104, Baltimore city, Maryland	1196	1146	50	0.9581939...	0.04180602
5	Census Tract 105, Baltimore city, Maryland	951	759	192	0.7981072...	0.20189274
6	Census Tract 201, Baltimore city, Maryland	871	697	174	0.8002296...	0.19977038
7	Census Tract 202, Baltimore city, Maryland	791	617	174	0.7800252...	0.21997472
8	Census Tract 203, Baltimore city, Maryland	1646	1427	219	0.8669501...	0.13304982
9	Census Tract 301, Baltimore city, Maryland	879	269	610	0.3060295...	0.69397042
10	Census Tract 302, Baltimore city, Maryland	887	533	354	0.6009019...	0.39909808
11	Census Tract 401, Baltimore city, Maryland	2046	1363	683	0.6661779...	0.33382209
12	Census Tract 402, Baltimore city, Maryland	574	262	312	0.4564459...	0.54355401

Census tract data snapshot

Visualizing the Racial Divide (link)

measures-of-segregation

Statistical Tests

You should currently feel comfortable...

Computing **summary statistics** on a sample or population

Using z-score to assess **relative performance**

Calculating and using **p-values** to confirm or deny hypothesis

- Given that the *null hypothesis is true*, there is only an N% chance of having observed the data *by chance alone*.

Understand the **probability** of something based on the **distribution** of the event

Today's additional skill is using the **t-test** to assess the significance of differences in means across groups.

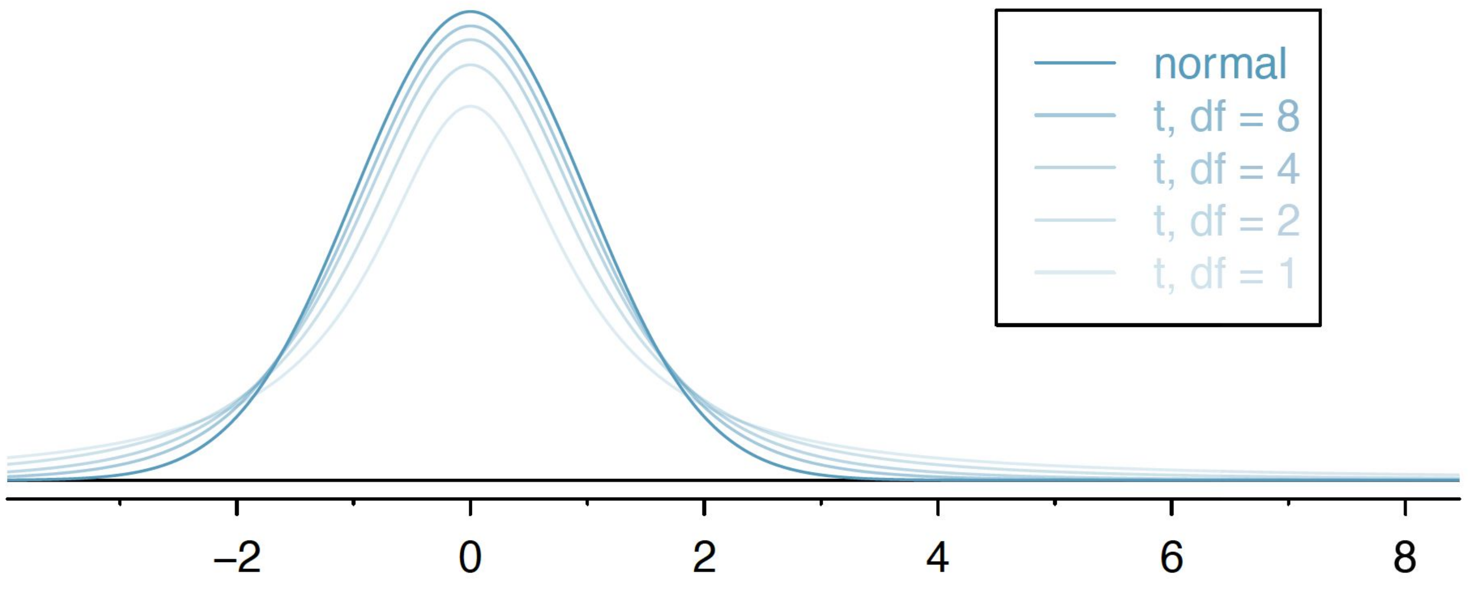

Figure 5.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal model.

The t-distribution (open-intro-stats)

We are 95% confident that the average mercury content is between X and Y.

What are the steps to solve for X and Y?

Figure 5.6: A Risso's dolphin.

Photo by Mike Baird (www.bairdphotos.com). CC BY 2.0 license.

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[3] The data are summarized in Table 5.7. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Computing one-sample confidence intervals using the t-distribution

	Two-sided α											
	.50	*.20*	*.10*	*.05*	*.02*	*.01*	*.005*	*.002*	*.001*	*.0005*	*.0002*	*.0001*
df												
1	1.00	3.08	6.31	12.71	31.82	63.66	127.32	318.31	636.62	1273.24	3183.10	6366.20
2	.82	1.89	2.92	4.30	6.96	9.22	14.09	22.33	31.60	44.70	70.70	99.99
3	.76	1.64	2.35	3.18	4.54	5.84	7.45	10.21	12.92	16.33	22.20	28.00
4	.74	1.53	2.13	2.78	3.75	4.60	5.60	7.17	8.61	10.31	13.03	15.54
5	.73	1.48	2.02	2.57	3.37	4.03	4.77	5.89	6.87	7.98	9.68	11.18
6	.72	1.44	1.94	2.45	3.14	3.71	4.32	5.21	5.96	6.79	8.02	9.08
7	.71	1.42	1.90	2.37	3.00	3.50	4.03	4.79	5.41	6.08	7.06	7.88
8	.71	1.40	1.86	2.31	2.90	3.36	3.83	4.50	5.04	5.62	6.44	7.12
9	.70	1.38	1.83	2.26	2.82	3.25	3.69	4.30	4.78	5.29	6.01	6.59
10	.70	1.37	1.81	2.23	2.76	3.17	3.58	4.14	4.59	5.05	5.69	6.21
11	.70	1.36	1.80	2.20	2.72	3.11	3.50	4.03	4.44	4.86	5.45	5.92
12	.70	1.36	1.78	2.18	2.68	3.06	3.43	3.93	4.32	4.72	5.26	5.69
13	.69	1.35	1.77	2.16	2.65	3.01	3.37	3.85	4.22	4.60	5.11	5.51
14	.69	1.35	1.76	2.15	2.63	2.98	3.33	3.79	4.14	4.50	4.99	5.36
15	.69	1.34	1.75	2.13	2.60	2.95	3.29	3.73	4.07	4.42	4.88	5.24
16	.69	1.34	1.75	2.12	2.58	2.92	3.25	3.69	4.02	4.35	4.79	5.13
17	.69	1.33	1.74	2.11	2.57	2.90	3.22	3.65	3.97	4.29	4.71	5.04

Write a full sentence to describe what the number 2.18 means

Using t-tests to assess differences in means

1. Calculate the mean of each group (group by // summarize)
2. Compute the difference in your means (m1 - m2)
3. Test if that value is significantly different from 0 using standard errors

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t-distribution and the standard error

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.16}$$

when each sample mean can itself be modeled using a t-distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

Paired v.s. un-paired data

Different methods for computing the standard error for paired v.s. un-paired data

How would you classify these examples?

- Comparing incomes across two companies?
- Comparing price of goods at two stores?
- Grades of each student on assignment 1 and assignment 2?

Unpaired Data

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Paired Data

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

class/t-tests

Code along!

```
# Set the upstream remote for the 'class' repo, if you haven't already
git remote add upstream https://github.com/info370a-w18/class.git

# Pull in the remote changes for the *master* branch
git checkout master # make sure you're on master
git pull upstream master

# Pull in the remote changes for the *complete* branch
git checkout complete # make sure you're on complete
git pull upstream complete

# Checkout the master branch to start doing the code-along
git checkout master
```

Upcoming...

a2-data-wrangling due ***Tuesday before class***

Notebook 3 due **Friday night** (notably shorter, more time in class/lab)

Next week:

- Linear and Poisson Regression