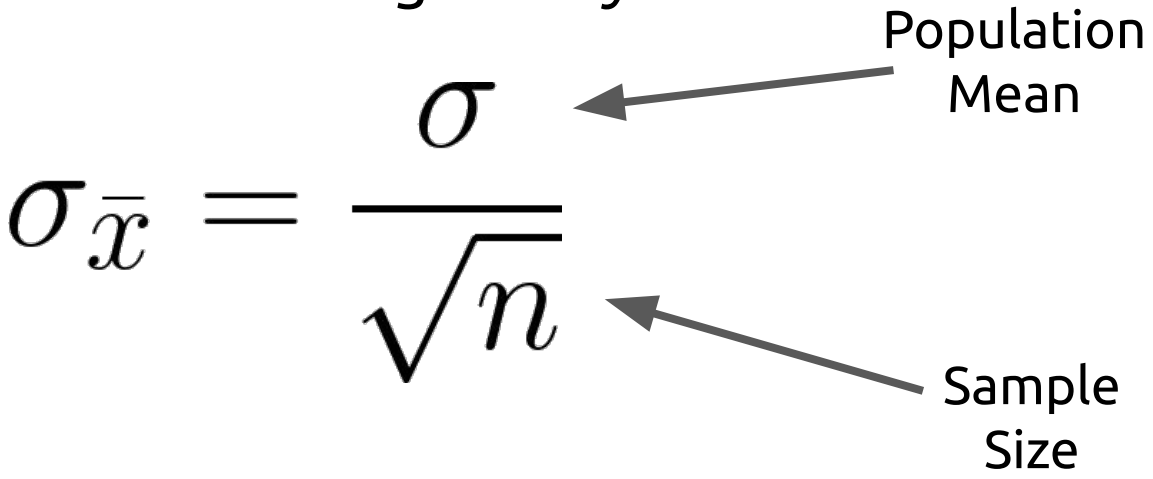
Estimation, Part 2

Two Cautionary Data Analysis Tales

Introduction

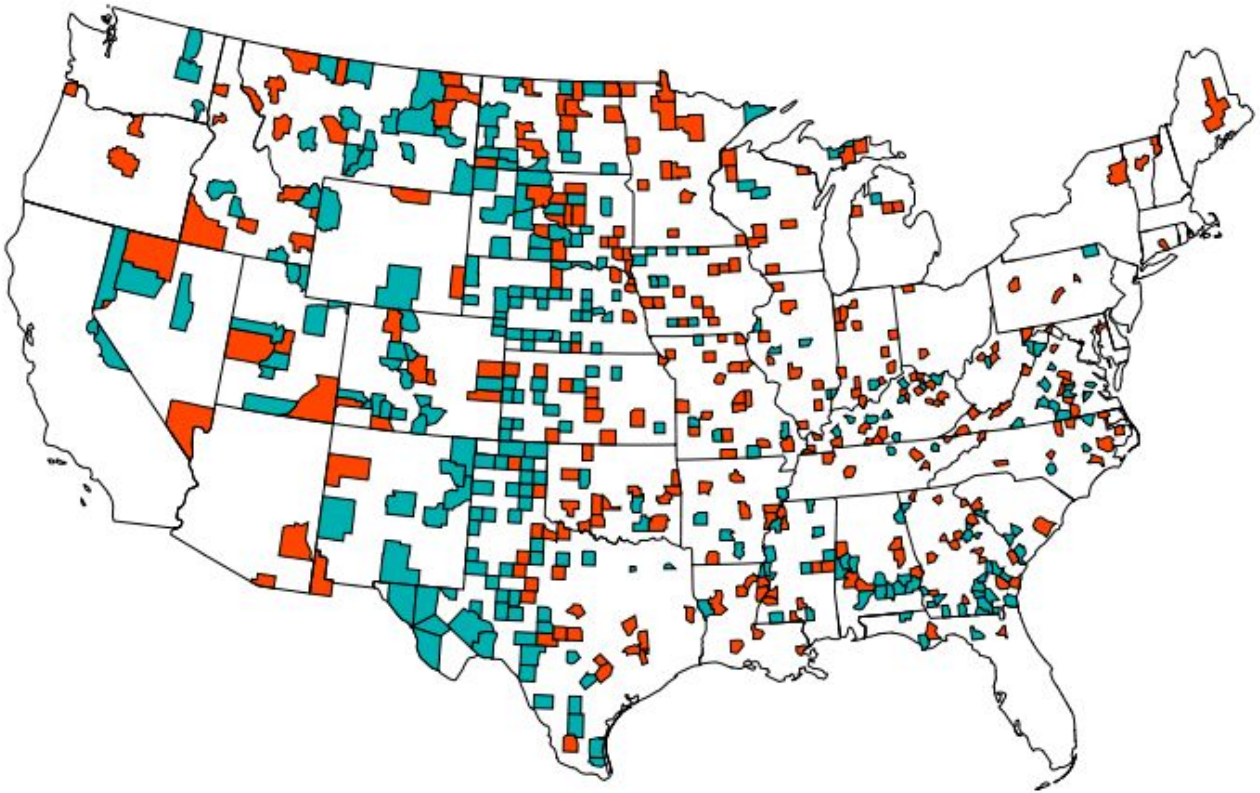
The margin of error when creating a confidence interval is determined largely by the *standard error of the mean*.

The standard error of the mean is given by



Estimation

Borrowing heavily from Howard Wainer (http://www-stat.wharton.upenn.edu/~hwainer/Readings/Most%20Dangerous%20eqn.pdf), let's explore some of the implications of the formula for the standard error of the mean.



Map of the US counties with the highest (in red) and lowest (in teal) rates of kidney cancer

Source: http://www-stat.wharton.upenn.edu/~hwainer/Readings/Most%20Dangerous%20eqn.pdf

Estimation

For counties with the highest rates of kidney cancer:

- Rural
- Southern, Midwestern, or Western counties

Potential causes for high rates of kidney cancer:

- poverty
- lack of access to medical care
- overall less healthy lifestyle

NASHVILLE SOFTWARE SCHOOL

Estimation

For counties with the lowest rates of kidney cancer:

- Rural
- Southern, Midwestern, or Western counties

Potential causes for low rates of kidney cancer:

- cleaner air
- cleaner water
- access to fresh foods

Estimation

So which is it?

Rural lifestyle lead to *higher* rates or rural lifestyle leads to *lower* rates?

Looking at the map, you can see that there are counties with the highest rates adjacent to those with the lowest rates!

Estimation

Let’s reframe this question in terms of the standard error of the mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Estimation

Let’s assume that the rural/urban dynamic plays no part in determining a person’s risk of developing kidney cancer.

Each county can be viewed as a “sample” from the US population.

We can calculate the cancer rate within each county to get an estimate of the overall US kidney cancer rate.

Estimation

How much variance can we expect in these estimates?

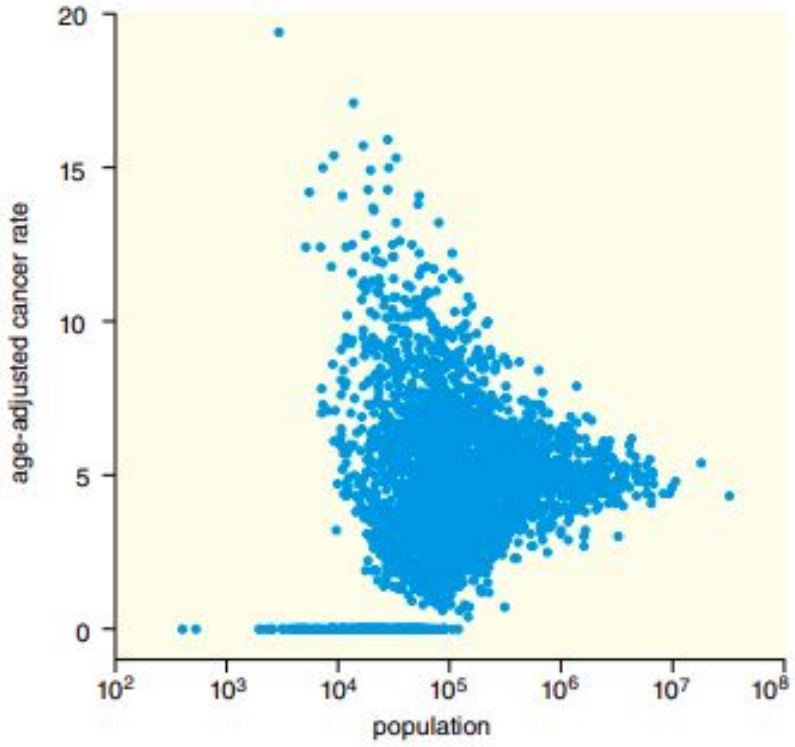
Let's return to our old friend, the equation for the standard error of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

For small sample size (rural counties), we get a larger standard error, hence larger variance in estimates.

For a large sample size (urban counties), we get a smaller standard error, hence less variation.

Estimation



A scatterplot of the population of all counties against their rate of kidney cancer shows a large amount of variation for smaller counties and a much smaller amount of variation for large counties, exactly as expected.

Source: http://www-stat.wharton.upenn.edu/~hwainer/Readings/Most%20Dangerous%20eqn.pdf

NASHVILLE SOFTWARE SCHOOL

Estimation

Moral: When studying aggregate statistics (means, rates) across differently-sized observations, make sure to take into account the amount of variance you would expect based on the formula for the standard error.

Part 2: Beware Regression to the Mean

Imagine that you have a team of 40 salespeople.

Wanting to identify traits of your best performers, you identify your top 5 salespeople in terms of sales conversion rate for the last year.

You decide to hold up these 5 salespeople as examples to aspire to.

Part 2: Beware Regression to the Mean

You keep track of these 5 salespeople over the next year.

The next year, this group's average conversion rate is much closer to the overall conversion rate company-wide.

Perhaps the pressure of being made into an example was too much...

Or is there some other reason for the slump?

Part 2: Beware Regression to the Mean

Let’s reframe the problem now in terms of quarters.

You buy a roll of 40 quarters and want to identify those quarters which are best at landing on heads when flipped.

You flip each quarter in the roll 100 times and pull out the 5 which landed on heads the greatest number of times.

NASHVILLE SOFTWARE SCHOOL

Part 2: Beware Regression to the Mean

Now, if we flip each of these “top 5 performing” coins 100 more times, what would you expect for the average number of heads?

Probably somewhere close to 50, which is the overall average.

There isn’t anything special about these coins, it’s simply sampling error.

Part 2: Beware Regression to the Mean

Moral: Keep **regression to the mean** in mind when doing data analysis, particularly when studying a phenomena with a great deal of variability.