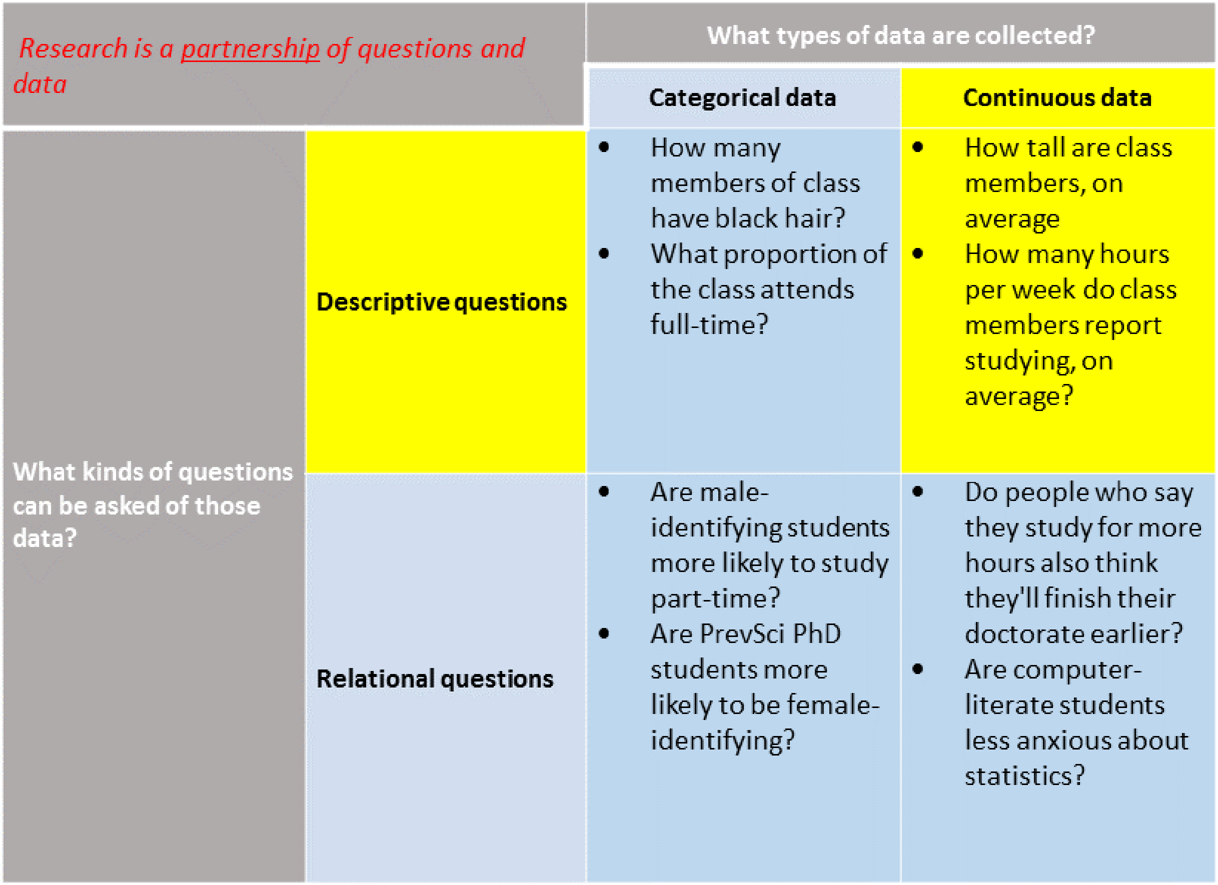
Transformations and *z*-scores

EDUC 641: Unit 3 Part 2

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap



Research is a partnership of questions and data		What types of data are collected?	
		Categorical data	**Continuous data**
What kinds of questions can be asked of those data?	**Descriptive questions**	• How many members of class have black hair? • What proportion of the class attends full-time?	• How tall are class members, on average • How many hours per week do class members report studying, on average?
	Relational questions	• Are male-identifying students more likely to study part-time? • Are PrevSci PhD students more likely to be female-identifying?	• Do people who say they study for more hours also think they'll finish their doctorate earlier? • Are computer-literate students less anxious about statistics?

Class goals

- Construct a standardized or z-score and explain its substantive meaning
- Use a z-transformation to compare distributions, observations within distributions and interpret outlying values
- Be prepared for future use of z-transformations in analysis

A "standard" deviation

The standard deviation (s) represents the **positive square root of the variance.**[1]

$$s = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{N}}$$

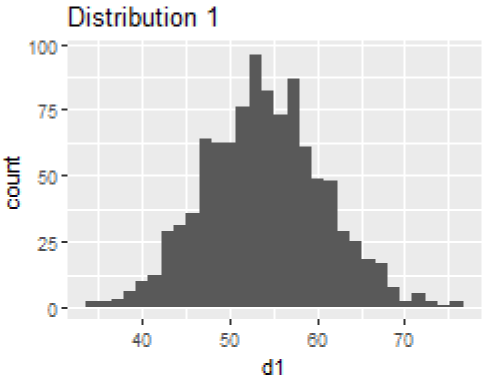
Steps:

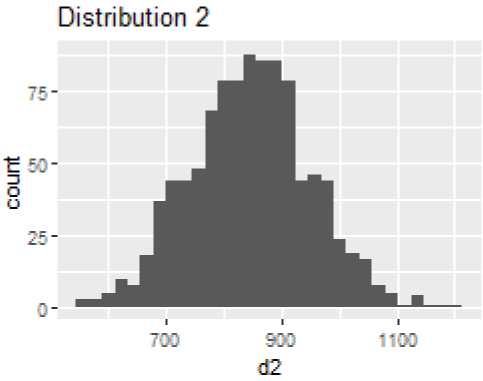
1. Subtract the mean from each observation in your data (this number is the deviation from the mean)
2. Square each resulting difference
3. Add up all of the squared deviations
4. Divide by the total number of observations
5. Take the square root $\rightarrow$ standard deviation

[1] This is actually not quite right. When calculating a sample statistic of the variance or standard deviation, the denominator in the above equation is actually *N*-1. We will learn why when we get to *degrees of freedom* in the next unit.

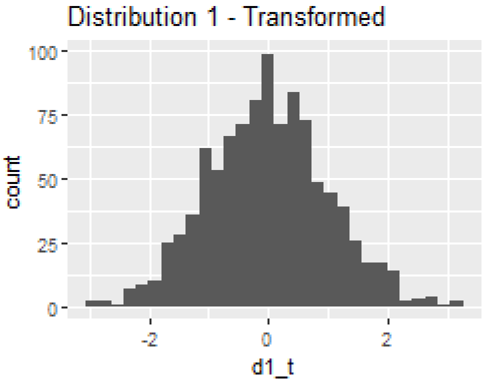
A common metric

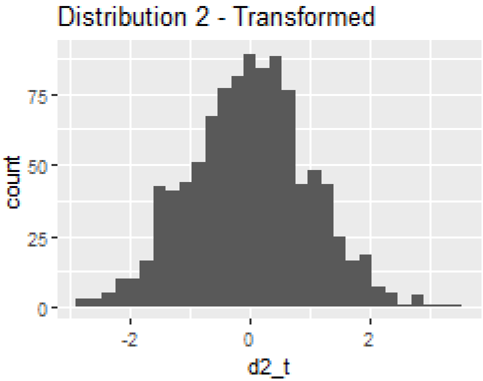
- A distribution can have any mean and any (positive) standard deviation.





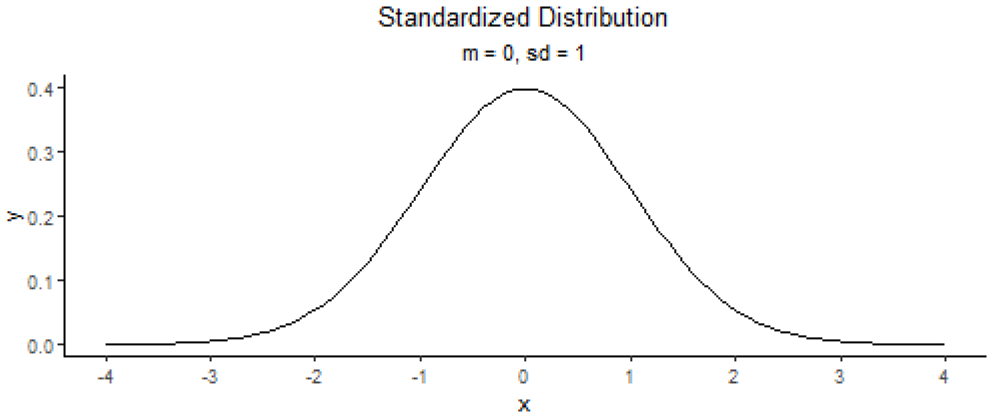
- Sometimes it is helpful to "standardize" a distribution to a common mean and standard deviation so we can more easily compare them (and understand outlying values).





Z-transformations

- The most common transformation is a z-transformation.
- A z-transformation re-scales the distribution to a mean (μ) of 0 and a standard deviation (σ) of 1.



Z-transformations

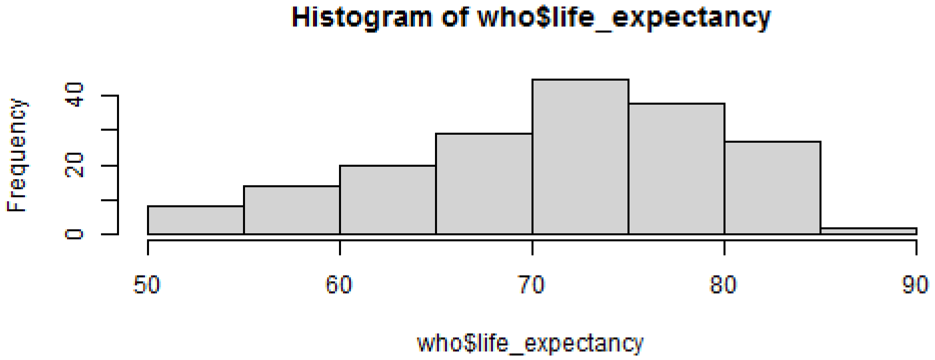
- Any score and distribution can be standardized using a simple algorithm.
- Each observation (i) is transformed into a **z-score** using the following formula:

$$z_i = \frac{x_i - \mu}{\sigma}$$

- A z-score is calculated by **subtracting the mean** from each value and **dividing by the standard deviation**.
- An observation's z-score value is equal to its distance from the mean, in standard deviation units.
- Some fun facts about z-scores
  - $\Sigma z_i = 0$
  - $\Sigma z_i^2 = N$

Transformed distributions

Here is a histogram of our life expectancy data.

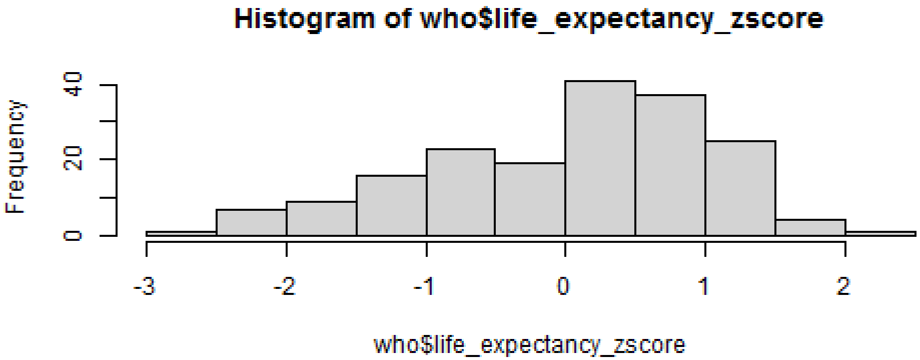


We are going to create a new variable called `life_expectancy_zscore` using the formula described on the previous slide.

```
who$life_expectancy_zscore ←
  (who$life_expectancy - mean(who$life_expectancy)) /
        sd(who$life_expectancy)
```

The new distribution

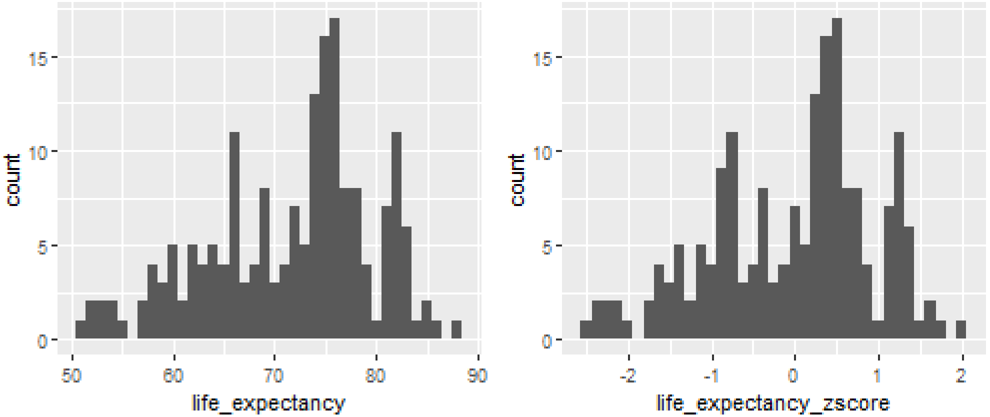
```
## Histogram of the new z-scores
hist(who$life_expectancy_zscore)
```



We now have a mean of 0 and standard deviation of 1.

"Transforming" vs. "normalizing"

An important note about standardizing a distribution is that it changes the mean and standard deviation, but **does not change the overall shape.**



You try

> Given the following set of observed value (75, 74, 66, 78, 73, 78), perform a z-transformation. What are the resulting z-scores?

How has this helped?

So we started with the promise that "transforming" (or "standardizing") a distribution would help us to better understand the "distance" that a given observation is from the center of the distribution and that z-scores allow us to compare across units of measurement.

Let's say we are interested in the life expectancy in a particular country and how this compares to both the average life expectancy and the distribution of life expectancies. For convenience, say Canada:

```r
mean(subset(who$life_expectancy,
            who$region == "Canada"))
```

```
## [1] 82
```

```r
mean(who$life_expectancy)
```

```
## [1] 71.63934
```

How different is life expectancy in Canada compared to our sample average?

Ok, but how different are these two numbers? And how different is Canada from the life-expectancy sample mean as compared to its difference from countries' average years of schooling?

How has this helped?

Life expectancy:

```
mean(subset(who$life_expectancy,
            who$region == "Canada"))
```

```
## [1] 82
```

```
mean(who$life_expectancy)
```

```
## [1] 71.63934
```

Canadian schooling:

```
## [1] 16.3
```

Average schooling:

```
## [1] 12.92717
```

Comparing on common metric

Now let's compare z-scores

```
mean(subset(who$life_expectancy_zscore,
            who$region == "Canada"))
```

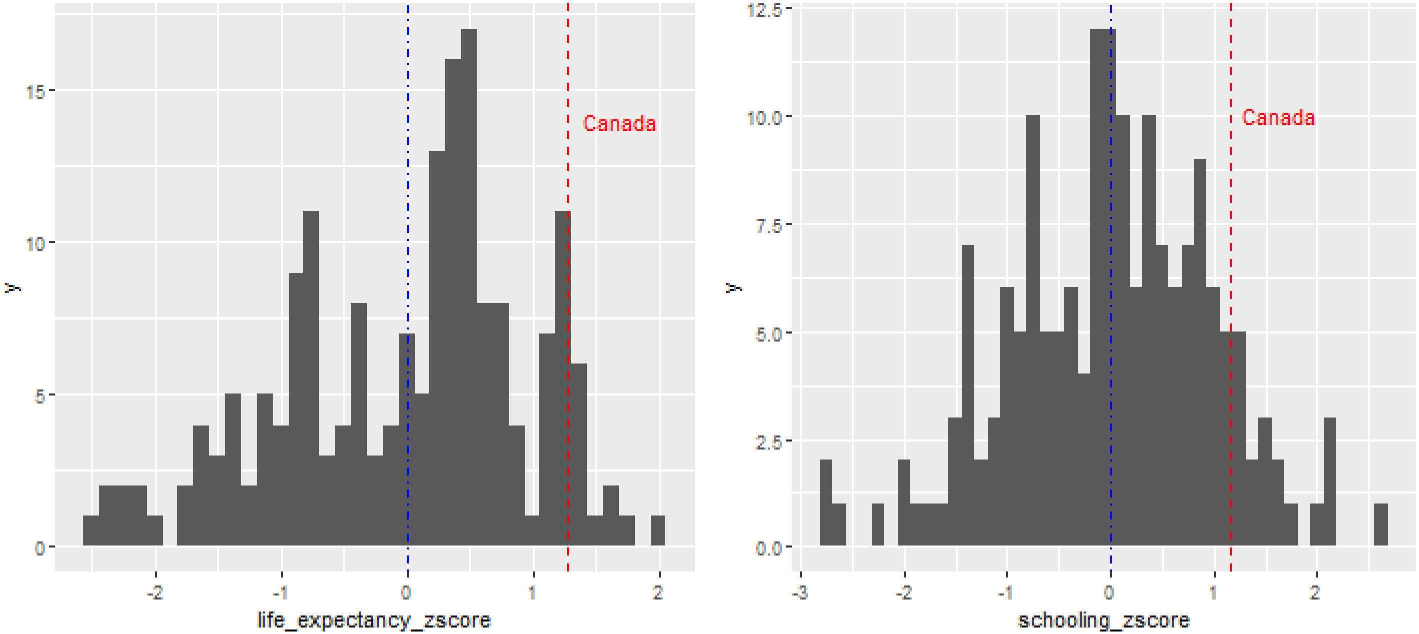
```
## [1] 1.271178
```

```
mean(subset(who$schooling_zscore,
            who$region == "Canada"))
```

```
## [1] 1.158107
```

Is Canada more unusual with respect to its schooling or life expectancy?

Comparing on common metric



Outliers

Compare the raw life expectancy to the standardized ones to get a better sense of outlying values:

```
mean(who$life_expectancy, na.rm=T)
```

```
## [1] 71.63934
```

```
head(sort(who$life_expectancy))
```

```
## [1] 51 52 52 53 53 54
```

```
tail(sort(who$life_expectancy))
```

```
## [1] 83 84 85 85 86 88
```

Are these extreme values a lot or a little away from the mean, given the rest of the distribution?

Outliers

Compare the raw life expectancy to the standardized ones to get a better sense of outlying values:

```
head(sort(who$life_expectancy_zscore))
```

```
## [1] -2.532299 -2.409606 -2.409606 -2.286913 -2.286913 -2.164220
```

```
tail(sort(who$life_expectancy_zscore))
```

```
## [1] 1.393870 1.516563 1.639256 1.639256 1.761949 2.007334
```

Effect sizes

Careful[1] standardization of continuous variables will permit:

- Common understanding of any individual observation's distance from the center of the distribution, across variables
- Ease of identifying outlying values
- Ability to understand the standard normal distribution (next!)
- Conduct a z-test (next!)
- Calculation of magnitude of continuous relationships in a common metric known as the **effect size**[2]

[1] "Careful" because the distribution within which you standardize the variable has important implications for the transformation and the resulting analysis you will do.

[2] *Further thoughts for those interested*: the **correlation coefficient** is a standardized effect size which can be used communicate the strength of a relationship. We will examine the correlation coefficient and the related concept of **effect size** further in EDUC 643 this winter.

Mid-term SES results

Response rate: 38 percent (11/29)

Quantitative results

Generally positive: (>=70% rate as beneficial)

- Inclusivity
- Support from instructors
- Feedback provided
- Quality of course materials
- Communication
- Organization
- Relevance of content
- Assignments/projects
- Accessibility

Generally insufficient: (<70% rate as beneficial)

- Level of challenge
- Clarity of assignment instructions/grading
- Active learning
- Student interaction

There are diverging opinions within each of these categories, and so important to attend to ways in which these broad-stroke patterns are not true for all individuals.

Qualitative results

Helpful:

- Take feedback & adjust
- Explanation of concepts
- Feedback on how to improve
- Response to students' questions
- Generally supportive
- Readings and class website

Need improvement/suggestions:

- Demeanor/expression in class
- 5 minutes to ask general questions
- Follow existing formats for learning R/more time learning R/more focus on R
- Clarity of quizzes
- More explicit/direct instruction
- Over-explanation of concepts → confusion

Action steps

1. Time for stats/programming questions (as much as possible)
2. Practice explanation of concepts beforehand to improve clarity
3. Improve quiz structure
4. Increase opportunities for practice and student-student interaction

Maintain primary focus of course on developing (applied) statistical and analytic toolkit with a secondary focus on application of these skills in the R programming language, following the syllabus as approved by your advisors and program directors via the College of Education curriculum Committee.

Synthesis and wrap-up

Class goals

- Construct a standardized or z-score and explain its substantive meaning
- Use a z-transformation to compare distributions, observations within distributions and interpret outlying values
- Be prepared for future use of z-transformations in analysis

To Dos

Quiz 3

- Opens 3:45pm on Oct. 27, closes at 5pm on Oct. 28

Assignments

- Assignment #3 due November 7, 11:59pm