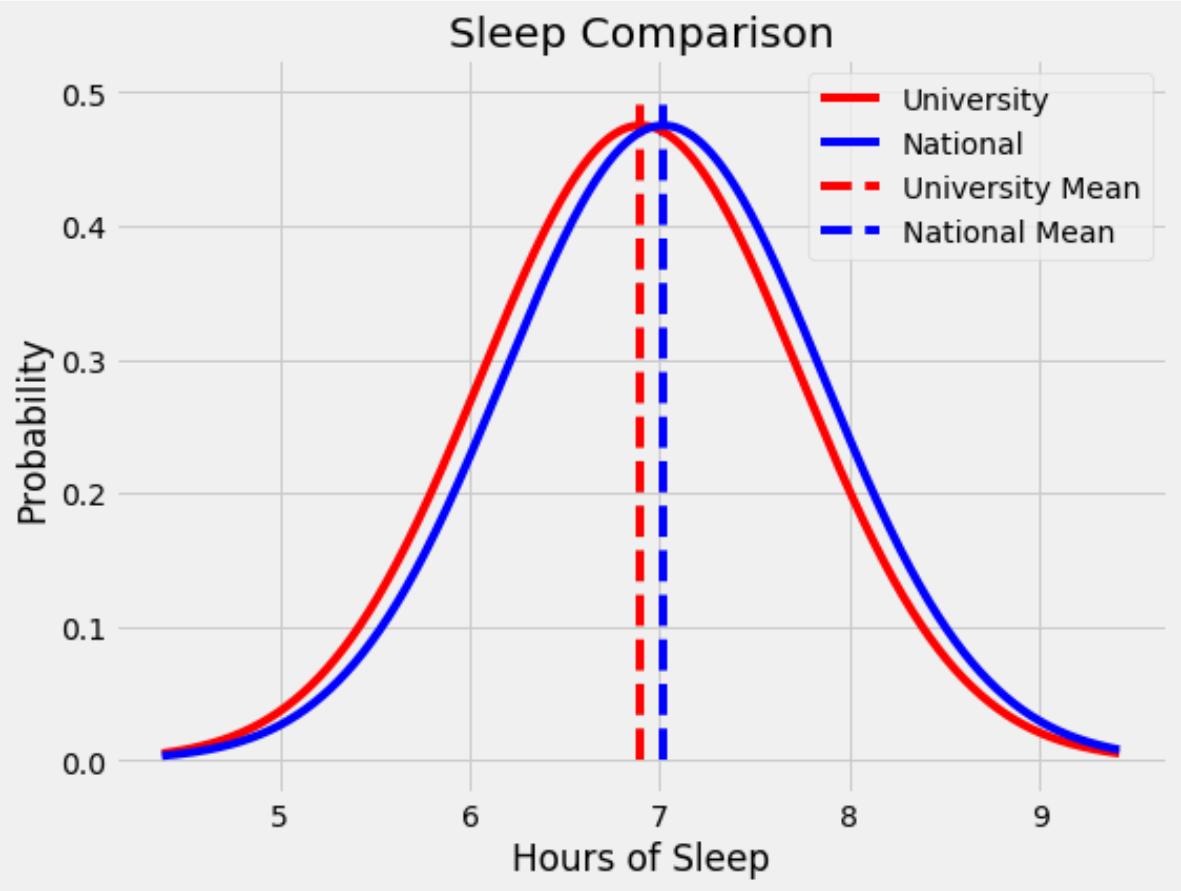
Statistical Significance

Aleksandr Fisher

You are getting very sleepy...

- Students at Haverford average 6.80 hours of sleep per night
- National college average of 7.02 hours
- You have to decide if this is a serious issue

Sleep Comparison



What it mean to prove something with data?

- Statistical Significance is built on a few simple ideas:
  - hypothesis testing,
  - the normal distribution
  - p values.

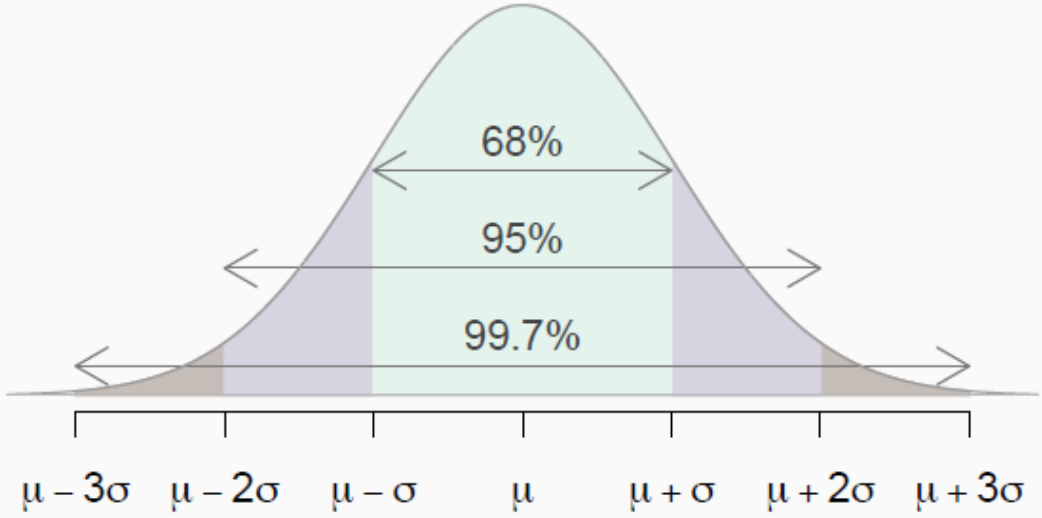
Hypothesis testing

- The "hypothesis" refers to initial belief about the situation before the study
  - **Alternative Hypothesis:** The average amount of sleep by students at our university is below the national average for college student.
  - **Null Hypothesis:** The average amount of sleep by students at our university is not below the national average for college students.
- This is an example of a one-sided hypothesis test because we are concerned with a change in only one direction

Normal Distribution

- The normal distribution is used to represent how data from a process is distributed defined by
  - the mean, given the Greek letter μ (mu)
  - the standard deviation, given the letter σ (sigma)

Normal Distribution



Normal Distribution

- We can determine how anomalous a data point is based on how many standard deviations it is from the mean
  - 68% of data is within $\pm$ 1 standard deviations from the mean
  - 95% of data is within $\pm$ 2 standard deviations from the mean
  - 99.7% of data is within $\pm$ 3 standard deviations from the mean

Example

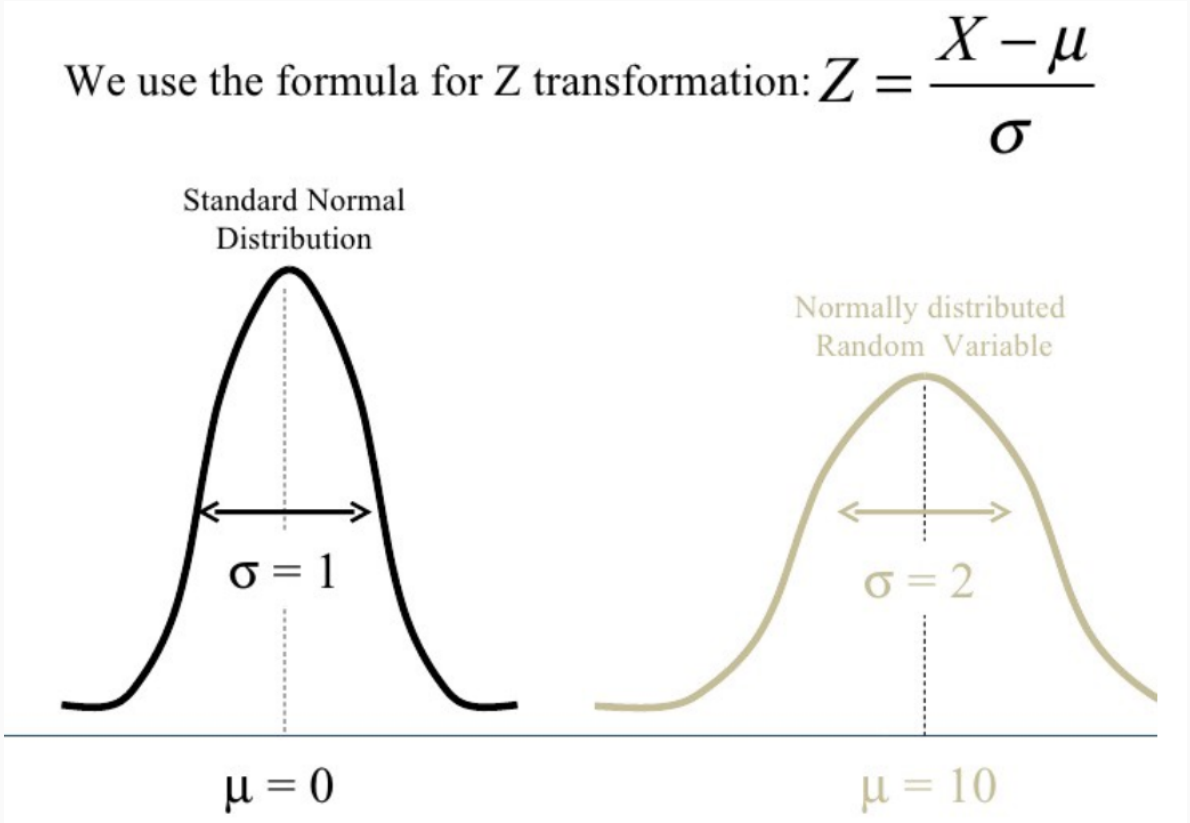
- Average female height in the US is 65 inches with a standard deviation of 4 inches.
- If we meet a new acquaintance who is 73 inches tall, we can say she is two standard deviations above the mean and is in the tallest 2.5% of females
  - 2.5% of females will be shorter than $\mu - 2\sigma$ (57 in) and 2.5% will be taller than $\mu + 2\sigma$.
- Instead of saying our data is two standard deviations from the mean, we assess it in terms of a z-score

Z-Score

- Conversion to a z-score is done by subtracting the mean of the distribution from the data point and dividing by the standard deviation
- The higher or lower the z-score, the more unlikely the result is to happen by chance and the more likely the result is meaningful

Z-Score

We use the formula for Z transformation: $Z = \frac{X - \mu}{\sigma}$



P-Values

- A p-value is the probability of observing results at least as extreme as those measured when the null hypothesis is true. . .
  - The p-value is NOT the probability the claim is true.
  - The p-value is NOT the probability the null hypothesis is true.

P-Values

- The p-value is actually the probability of getting a sample like ours, or more extreme than ours IF the null hypothesis is true.
- So, we assume the null hypothesis is true and then determine how "strange" our sample really is.
- If it is not that strange (a large p-value) then we don't change our mind about the null hypothesis.
- As the p-value gets smaller, we start wondering if the null really is true and well maybe we should change our minds (and reject the null hypothesis).

P-Values

- Whether or not the result can be called statistically significant depends on the p-value (known as alpha) we establish for significance before we begin the experiment
- If the observed p-value is less than alpha, then the results are statistically significant.
- We need to choose alpha before the experiment because if we waited until after, we could just select a number that proves our results are significant no matter what the data shows

THE 0.05

- Most commonly used value is 0.05, corresponding to a 5% chance the results occurred at random
- If you ran the experiment 100 times — again, assuming the null hypothesis is true — you'd see these same numbers (or more extreme results) five times.
    - R.A. Fischer, the father of modern statistics, choose a p-value of 0.05 for indeterminate reasons and it stuck)!

But Really Why 0.05?

- https://www.openintro.org/book/stat/why05/

Summary

- As a summary so far, we have covered three ideas:
- **Hypothesis Testing:** A technique used to test a theory
- **Normal Distribution:** An approximate representation of the data in a hypothesis test.
- **p-value:** The probability a result at least as extreme at that observed would have occurred if the null hypothesis is true

Back to Haverford Students' late nights

- Students across the country average 7.02 hours of sleep per night according to the National Sleep Foundation
- In a poll of 200 students at Haverford the average hours of sleep per night was 6.90 hours with a standard deviation of 0.84 hours.
- Our alternative hypothesis is the average sleep of students at Haverford is below the national average for college students.
- We will use an alpha value of 0.05 which means the results are significant f the p-value is below 0.05.

First we need a z-score

- subtracting the population mean (the national average) from our measured value and dividing by the standard deviation over the square root of the number of samples.

$$\frac{x - \bar{x}}{\sigma / \sqrt{n}}$$

- When you are estimating the standard error, SE, for the mean (the SE is the standard deviation of the means of samples), the larger your sample size, the smaller the standard deviation. for example, if you took a sample of 200, you would be much more likely to get close to the true mean than if you took a sample of 2. In other words, the larger your "n", the smaller the standard deviation.

First we need a z-score

$$\frac{6.90 - .7.02}{0.84/\sqrt{200}} = -2.03$$

- The z-score is called our test-statistic. Once we have a test-statistic, we can use a table or a programming language such as R to calculate the p-value.

```
# Calculate the results
z_score = (6.90 - 7.02) / (0.84 / sqrt(200))
p_value = pnorm(z_score)
# Print our results
sprintf('The p-value is %f for a z-score of %f.', p_value, z_score)
```

```
## [1] "The p-value is 0.021676 for a z-score of -2.020305."
```

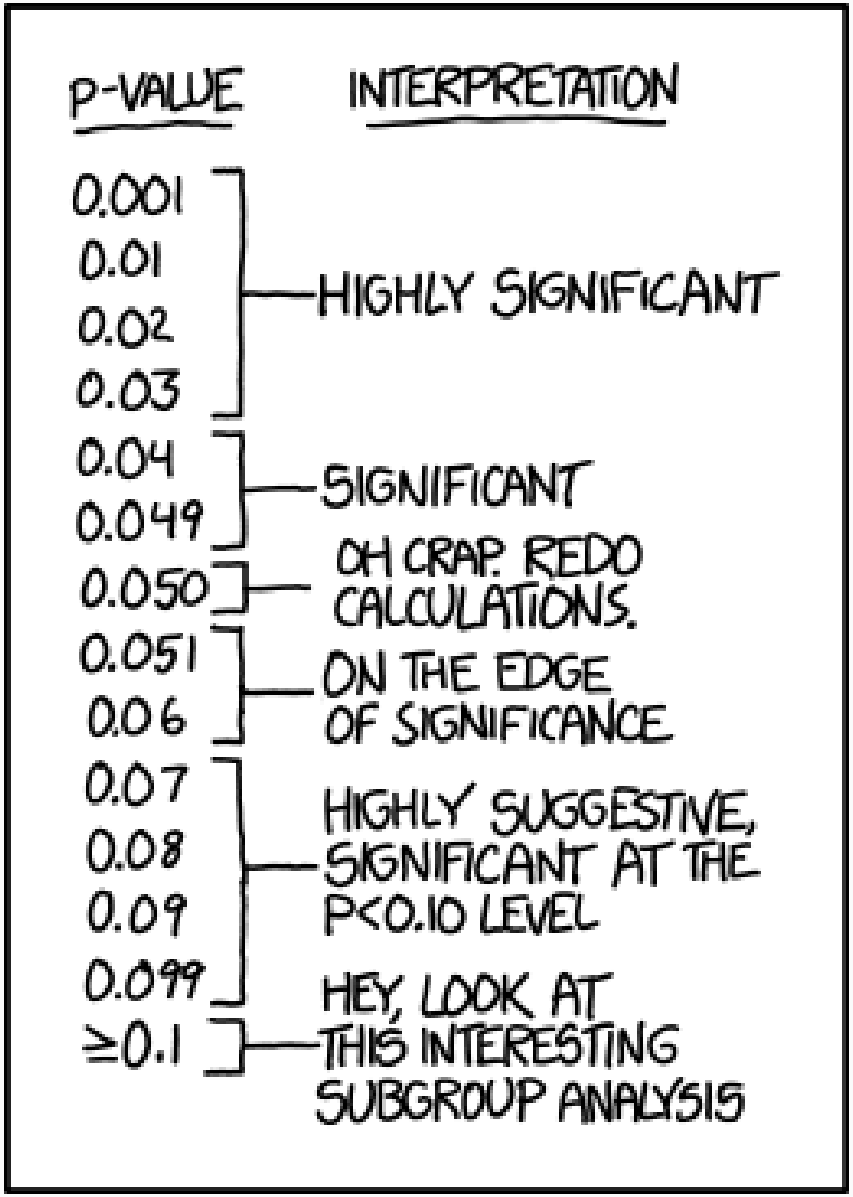
So what do we know?

- Based on the p-value of 0.02116, we can reject the null hypothesis. (Statisticians like us to say reject the null rather than accept the alternative.)
- There is statistically significant evidence our students get less sleep on average than college students in the US at a significance level of 0.05. The p-value shows there is a 2.12% chance that our results occurred because of random noise.
- Notice that our p-value, 0.02116, would not be significant if we had used a threshold of 0.01.

Some thoughts on p-values

- We should think about the p-value and the sample size in addition to the conclusion.
- might have statistical significance, but that does mean it is practically meaningful.
- This was an observational study, which means there is only evidence for correlation and not causation.

P-hacking


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

BIG Debates on p-values

- Replication crises
- Publication bias
- Propose a change to $P < 0.005$
    - fewer false positive
- Rejecting the null doesn't tell you anything about the mechanism
- It doesn't tell you if the experiment is well designed, or well controlled for, or if the results have been cherry-picked.

The case against p-values

- A famous 2015 paper in Science attempted to replicate 100 findings published in a prominent psychological journal. Only 39 percent passed
- Studies that yielded highly significant results (less than p=.01) are more likely to reproduce than those that are just barely significant at the .05 level.
- The increased burden of proof — the proposal authors hope — would nudge labs into adopting other practices science reformers have been calling for, such as data sharing and thinking more long-term about their work.

The case against p$<$.005

- High standards could impede scholars with low budgets.
- It keeps scientific communities fixated on p-values

How else to evaluate good social science

- Concentrating on effect sizes (how big of a difference does an intervention make, and is it practically meaningful?)
- Confidence intervals (what's the range of doubt built into any given answer?)
- Whether a result is novel study or a replication (put some more weight into a theory many labs have looked into)
- Whether a study's design was preregistered (so that authors can't manipulate their results post-test), and that the underlying data is freely accessible (so anyone can check the math)
- There are also alternative statistical techniques — like Bayesian analysis — that, in some ways, more directly evaluate a study's results

For more see...

- Slides based on:
  - https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687
- P-values:
  - https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-
  - https://www.vox.com/latest-news/2019/3/22/18275913/statistical-significance-p-values-explained
  - https://warwick.ac.uk/fac/soc/economics/staff/vetroeger/publications/pvaluedebate_vt1.pdf