MEASURING INEQUALITY: RACE AND GENDER IN SCIENCE

ROBIN ANDREASEN, ASSOCIATE PROFESSOR, COGNITIVE SCIENCE

HEATHER DOTY, ASSISTANT PROFESSOR, MECHANICAL ENGINEERING

UNIVERSITY OF DELAWARE

MEASURING INEQUALITY: RACE AND GENDER IN SCIENCE

FORTHCOMING IN TED RICHARDS AND KEVIN ELLIOTT (EDS)

EXPLORING INDUCTIVE RISK

OXFORD UNIVERSITY PRESS


NSF ADVANCE
INSTITUTIONAL TRANSFORMATION

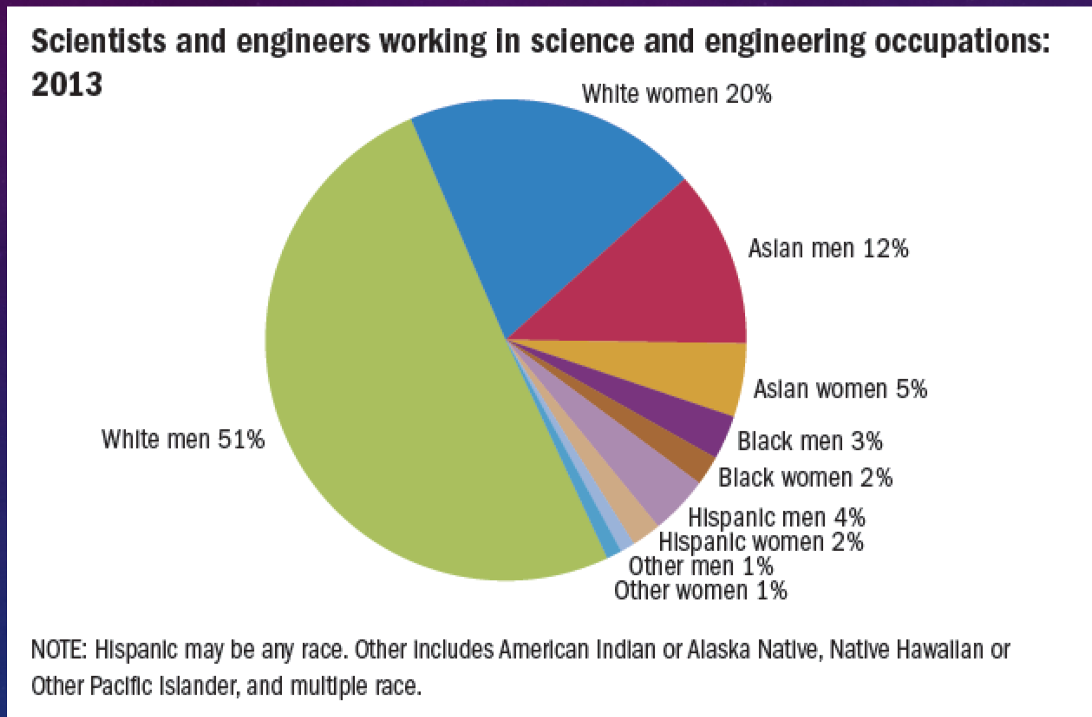
…to increase the *representation and advancement* of Women, and women of color, in academic science and engineering careers

…to address *various aspects of STEM academic culture and institutional structure* that may differentially affect women faculty and women faculty of color.

NSF ADVANCE-IT HRD 1409472

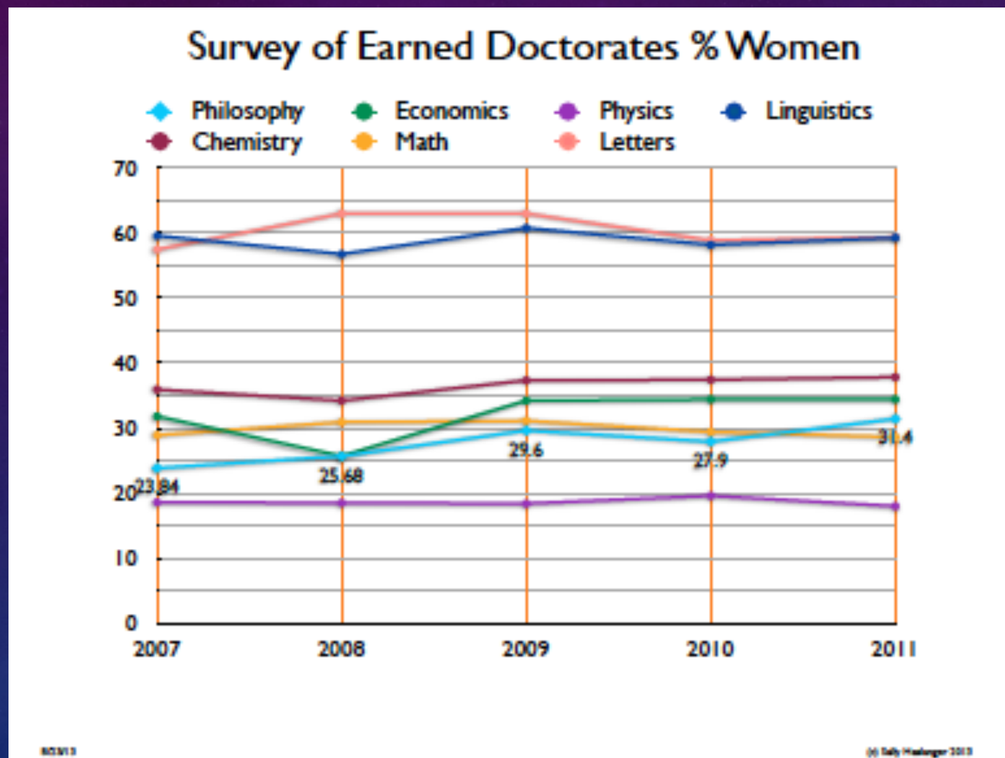
UD ADVANCE
This way up.

DIVERSITY DATA SCIENCE & ENGINEERING

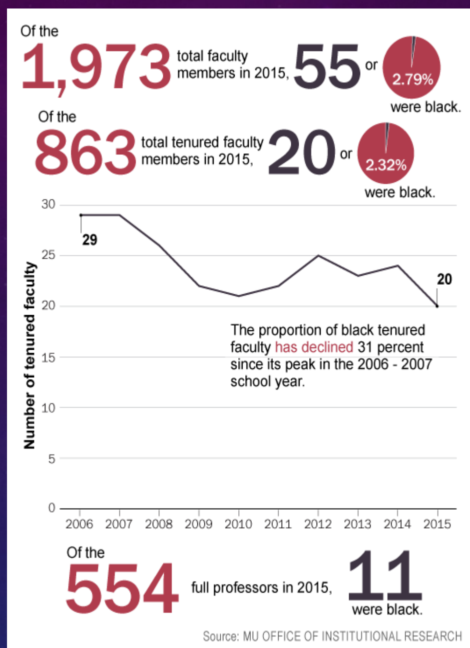


NSF, 2015, Women, Minorities, and Persons with Disabilities, nsf.gov

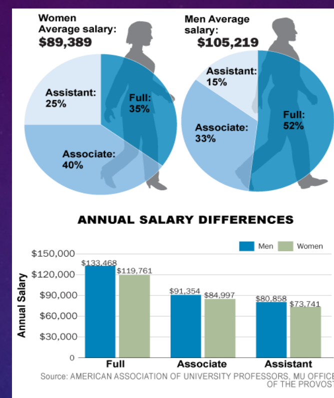
WOMEN IN PHILOSOPHY EARNED PHDS



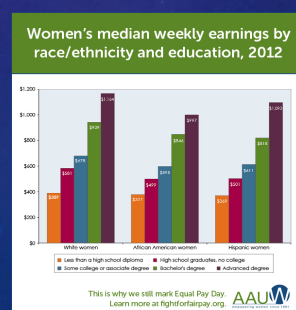
Other Disparities



Retention

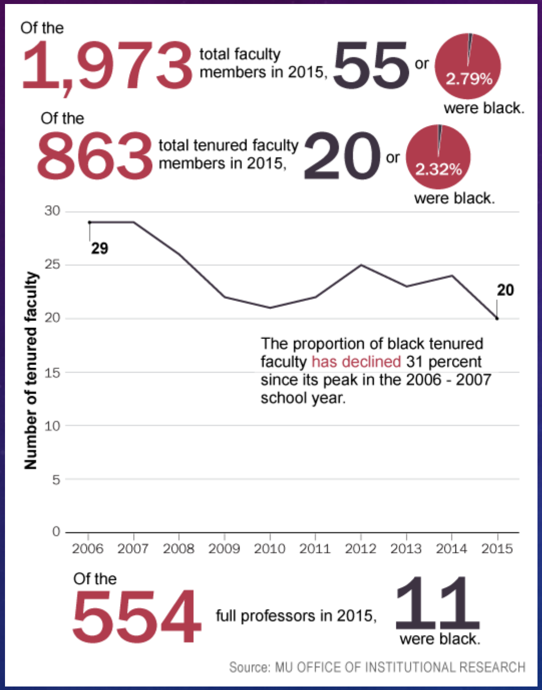


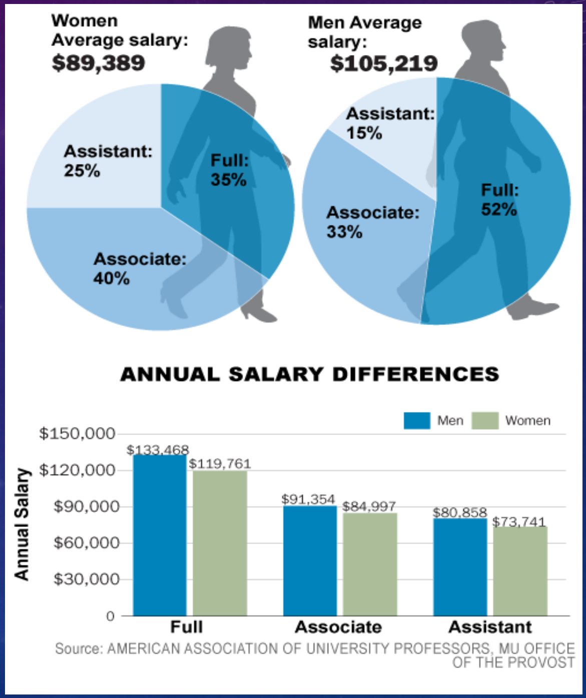
Salary



Awards & Leadership

IMPORTANCE OF DATA FOR EFFECTING CHANGE





FT T/TT FACULTY IN S&E BY GENDER & URM STATUS UNIVERSITY OF DELAWARE, 2014

College	%F	#F	% URM*	# URM
Agriculture	27.9%	19	2.9%	2
Arts & Sciences - Natural Sciences	25.0%	37	5.4%	8
Earth, Ocean, & Environment	23.5%	12	0.0%	0
Engineering	16.7%	21	3.9%	5

Source: University of Delaware, Office of Institutional Research

URM:
- Black
- Hispanic
- Latina/Latino

Today's Talk: We will take a closer look at some methodological difficulties with collecting and reporting diversity data.

I. Traditional statistical methods may sometimes raise the risk of false negatives in hypothesis testing – and inequalities may be allowed to persist.

Traditional Statistical Methods

- Largely driven by significance testing
- Implicit endorsement of a value free ideal.

F Negative: Concluding that there is no inequity, when a true disparity exists.

II. In some cases, non-epistemic (social, normative, pragmatic) values may have a role to play.

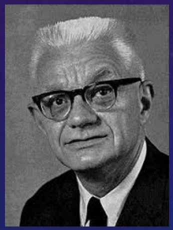
Inductive Risk: The risk of erroneously accepting a false hypothesis (false-positive) or erroneously rejecting a true one (false-negative), due to the probabilistic nature of hypothesis testing.

- With the exception of hypotheses that are deductively confirmed by the evidence, there is a component of hypothesis testing that goes beyond the evidence.

ARGUMENT FROM INDUCTIVE RISK

Under certain circumstance, non-epistemic values (social, normative, pragmatic) have a legitimate role in the scientific testing process.

Richard Rudner 1953



Carl Hempel 1965

Heather Douglas, 2000; 2009

What are those circumstances?

- There must be considerations of inductive risk
- Scientists must rely on non-evidential standards to fill the gap between evidence and acceptance.
- When there are non-epistemic consequences associated with hypothesis testing, non-epistemic values have a legitimate role to play.

Examples, Heather Douglas (2000; 2009)

- Setting a confidence level for hypothesis acceptance. P-values, for example.
- Evidence Characterization, Rat Liver Tumors
- Interpretation of Results, Are there thresholds?

Dioxin and Rat Liver Slides (Douglas, 2009)

- Expert pathologists examine slides
- An inherent element of judgment
- Inductive Risk

  False Positive: A non-cancerous lesion is deemed cancerous.

  False Negative: A cancerous lesion is deemed non-cancerous.

- Potential Non-Epistemic Consequences & a Role for Non-Epistemic Values

  False Positives: Over-regulation and associated costs.

  False Negatives: Under-regulation and increased incidents of cancer.

General Structure, Argument from Inductive Risk

- Inherent element of judgment.

  Even when a scientist has collected ample data using reliable methods and has gone as far as one can go with the data, there remains some room for judgment.

- Scientists must decide whether to err in the direction of F positives or F negatives.
- There are non-epistemic consequences associated with choice.

Today's Talk: We will take a closer look at some methodological difficulties with collecting and reporting diversity data.

I. Traditional statistical methods may sometimes raise the risk of false negatives in hypothesis testing – and inequalities may be allowed to persist.

   Traditional Statistical Methods

   - Largely driven by significance testing
   - Implicit endorsement of a value free ideal.

   F Negative: Concluding that there is no inequity, when a true disparity exists.

II. In some cases, non-epistemic (social, normative, pragmatic) values may have a role to play.

KAMINSKI & GEISLER, 2012

Science AAAS

Are there gender differences in retention and promotion among S&E faculty?

- Longitudinal study (1990 – 2009)
- Large data set: 2966 S&E Faculty at 14 Institutions
- Retention. Year of hire as assistant prof. → Year of departure from same institution.
- Promotion. % promoted to associate prof. & time to promotion to full prof.
- Analyzed data by gender both within and across disciplines.

Conclusion: A broad view of gender parity in retention & promotion S&E faculty.

Q. To what extent does K&G's conclusion of relative parity depend on their methods.

We found no problems with their results, *given their choice of methods. But…*

Methodological choices can make a difference in establishing significance or lack thereof.

Ex.: Choice of significance Test

Ex.: Operationalizing Variables

There are alternatives to significance testing for measuring disparities.

Outline:

- We will look at two examples.
- In each case, there is an inherent element of judgment.
- There are also non-epistemic consequences associated with hypothesis testing.
- A role for non-epistemic values.

STATISTICAL SIGNIFICANCE VS. 4/5 RULE

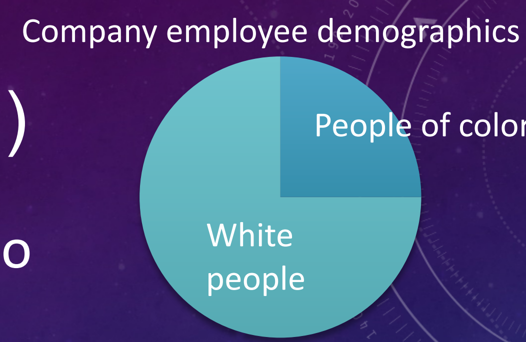
Statistical Significance: Is there a "clear" difference between groups?

- Easier to establish stat. significant difference in large samples

4/5 Rule: Is the difference large enough to matter?

- Used in the law to determine if employment outcomes for a minority group is different enough from that of a majority group to be actionable.

EXAMPLE: 4/5 RULE (80% RULE)



- Compares success rate of minority group to success rate of majority group.
- Say, 75% of white employees who apply for a promotion receive it, but only 50% of the people of color do.
- Is this disparity in employment outcome large enough to be actionable?
- .5/.75 = .67
- .67 < .80 , so **yes**, it could be considered actionable.

STATISTICAL SIGNIFICANCE VS. 4/5 RULE

Possible Scenarios:

A. A disparity is not statistically significant, but large enough to be actionable under 4/5 rule (more likely to happen with small samples; possibly also with larger samples that have small subsamples within).

B. A disparity is statistically significant, but too small to be actionable under 4/5 rule (more likely to happen with large samples).

EXAMPLE: 4/5 RULE AND K&G

- We compared rates of retention for men and women faculty in K&G's data set.
- "retained" = still at hiring institution in 2009, the study's end.
- We looked at specific disciplines.
- In no discipline was the difference in retention between men and women statistically significant.

EXAMPLE: 4/5 RULE AND K&G

* Indicates a disparity actionable under 4/5 rule

Discipline	# Faculty	# W	% W	% Retained, W	% Retained, M	W's retention as % of M's
Electrical Eng.	194	16	8.0%	31.3%	51.1%	61.25% *
Phys/Astr	171	19	11.0%	36.8%	49.3%	74.6% *
Mech. Eng.	164	26	16%	46.2%	59.4%	77.8% *
Chemistry	109	18	16.5%	50.0%	54.9%	91.1%
Math	194	36	19%	25.0	34.2	73.1% *
Comp. Sci.	169	34	20%	47.1	45.9	>100%
Civil Eng.	121	28	23%	39.3	47.3	83.1%
Biology	143	42	29%	52.4%	51.5%	>100%
Chemical Eng.	71	22	31%	31.8%	44.9%	70.8% *

In none of these cases is the difference in retention between men and women statistically significant.

WHICH TEST SHOULD POLICY MAKERS RELY ON?

- It's a judgment call. Relying on statistical significance is more likely to tend toward false negative – and the persistence of the status quo.
- Relying on the 4/5 rule is more likely to tend toward false positives – and potentially expensive remedies.
- Data set supports either outcome –the decision of which method to use goes beyond the evidence.
- Non-epistemic values may have an important and legitimate role to play.

OPERATIONALIZATION OF SCIENTIFIC VARIABLES

- Kaminski and Geisler conclude that there is parity between men and women STEM faculty based on their study.
- However, they measure retention in just one way (median time to departure).
- There are other ways to operationalize the variable ‘retention,’ which may lead to different results.
- K&G don’t tell the whole story in their paper.

PERCENT WOMEN VS. PERCENT MEN DEPARTED

- As in the previous example, we considered “retained” = still at hiring institution in 2009, the study’s end.
- Aggregating across disciplines, we found that 47.2% of the men but only 41% of the women were retained.
- This disparity was statistically significant ($p \leq .05$, Pearson’s Chi-Square and Fisher’s Exact Test).

OPERATIONALIZATION AND INDUCTIVE RISK

- K&G operationalized retention one way and found no disparity. If they are wrong, it is a false negative.
- We operationalized retention a different way and found a disparity. If we are wrong, it is a false positive.

WHAT SHOULD A SCIENTIST DO?

- Neither measure captures all of ‘retention.’
- For complex variables, any choice of measurement is likely to fall short in terms of providing sufficient evidence for hypothesis acceptance (or rejection).
- Multiple measures may capture more aspects of the variable, but the risk remains of failing to provide a complete picture.

SHOULD VALUES PLAY A ROLE?

- Sure. A scientist may use non-epistemic values to choose which aspect of retention she is more interested in (median time to departure or percentage retained). But…
- This use of values is not justified by consideration of inductive risk.
  - The role for values in the argument from inductive risk is to aid scientific judgment in cases where one can go no further with the data.

WHAT MIGHT K&G HAVE DONE?

- They could have looked at both median time to departure and percent retained and reported both results.
- They could have acknowledged in their paper that their study analyzed one one facet of retention – and their broad view of parity relates only to the time men and women science faculty stay in one position.

CONCLUSION

- We have shown examples from social-science literature where a scientist is in the position to make a choice that is not easily decided empirically.
- Because of the nature of the research questions being asked, there are non-epistemic consequences associated with error in hypothesis acceptance/ rejection.
- These choices are inherently value laden and, thus, traditional statistics (i.e., value-free science that relies on significance testing alone) can lead to errors and the potential for inequity.
- Non-epistemic values may have an important and legitimate role to play as scientists choose.

BETTER WAYS TO STUDY SMALL SAMPLES?

- Common suggestion: combine qualitative and quantitative data. This is good, but is it good enough in all cases?
- We find that scientists and university administrators often react more strongly to quantitative results. How can quantitative data analysis methods improve in the case of small subsamples?