The role of statistics in promoting human rights:
Without data, all you are is just another person with an opinion

AAAS Science and Human Rights
Mary W. Gray
January 11, 2111

mgray@american.edu
Why statistics?

There are three kinds of lies ...lies, damn lies, and statistics.

Benjamin Disraeli

Numbers do not lie, but they have the propensity to tell the truth with intent to deceive.

The death of one man is a tragedy. The death of millions is a statistic. Stalin, comment to Churchill at Potsdam, 1945

If you want to inspire confidence, give plenty of statistics. It does not matter that they should be accurate, or even intelligible, as long as there is enough of them.

Lewis Carroll

Far better an approximate answer to the right question, than the exact answer to the wrong question.

John Tukey
Why use statistics?

What can we say and do to help human rights researchers and activists?

Statistics can provide the information needed to answer questions
where should a new clinic be located

Statistics can suggest questions that should be asked
what is the prevalence of infant mortality

Statistics can tell us how probable certain results are or will be
is it credible that figures offered by the government are correct

Statistics can assist us to make predictions or estimates
outcome of an election, immunization coverage

Statistics can help us come to conclusions
what needs to be done about a problem
Can you believe what you read in the newspapers?

**What Happened to Skepticism?**

*New York Times* front page headline 21 May 2009:  
“1 in 7 Detainees Rejoined Jihad, Pentagon Finds”

The first paragraph reported that an unreleased Pentagon study had concluded that about one in seven of the 534 prisoners already transferred abroad from Guantanamo had “returned to terrorism or militant activity.”

The article failed to distinguish between former prisoners suspected of new acts of terrorism—more than half the cases—and those supposedly confirmed to have rejoined jihad against the West. Had only confirmed cases been considered, 1 in 7 would be changed to 1 in 20.

Now the recidivism rate is reduced from 14% to 5%. But is it recidivism at all? Were any of the released inmates ever engaged in terrorist or militant activity before their imprisonment? Or were they radicalized while in Guantanamo?

So, clearly the *New York Times* reporters and editors should have been skeptical, but so should the readers—and in fact, many were and the paper’s Public Editor ended up writing a critical column on the issue.
“Could this be why not?”


---

IT’S A CHRISTIAN DIASPORA—ONE OF THE LARGEST IN HISTORY!

IT’S HARD TO DETERMINE PRECISE NUMBERS, BUT AS MANY AS 600,000 CHRISTIANS MAY HAVE BEEN DRIVEN FROM THEIR HOMES!

BUT YOU ADMIT NO ONE KNOWS. SO IT COULD JUST AS EASILY BE 600 PEOPLE AS 600,000!

IT COULD?

I’LL JUST REPORT BOTH FIGURES—LET OUR VIEWERS DECIDE!
If it seems too good/bad to be true, it probably isn’t

The Kuwaiti incubators story.

Were there so many premature babies in Kuwait in 1991 that 312 incubators were in operation?
More skepticism: “The Deadly Toll of Abortion by Amateurs”

This was the headline on 2 June 2009. The story reported that for every 100,000 births in Tanzania, 950 women die. But nowhere do we see how many of the deaths are attributable to botched abortions.

We are told there are 19 million “unsafe” abortions a year and 70,000 deaths attributable to them (presumably worldwide). This is, we learn further, is 13% of all maternal deaths.

World population 6,800,000,000
women 15 to 49 1,760,000,000
Rate of “unsafe” abortions worldwide 19/1760 = 1%
Is it credible that each year 1 woman in every 100 worldwide has an “unsafe” abortion?

By extension of the worldwide figures on maternal deaths, we might conjecture that 123 (of every 100,000 births) of the maternal deaths in Tanzania arise from “unsafe” abortions if Tanzania is an “average” country. But is it? Or are all 950 deaths referenced above due to unsafe abortions, making the situation particularly dire there?
Maternal deaths and unsafe abortions, continued

In what may be intended as an explanation of the situation in Tanzania, we are told that the use of contraception is 25% in Tanzania, but 39% in Kenya, and 60% in South Africa (where abortion is legal). But no mortality rate is given for these countries.

So what are we to conclude about “unsafe” abortion, use of contraception and death rates?

Well, “unsafe” abortions are not good, in Tanzania or anywhere else, but without statistics, I’m just another person with an opinion.
Advice to researchers:
Some sources of data

Online

www.worldbank.org

www.unstats.un.org/demographic

www.stat.can.gc.can. www.statistics.gov.uk

www.who.int/research/en/

www.nber.org/links/data.html

University data sources, e.g., University of Michigan

www.lib.umich.edu/govdocs/stforeign.html#comp

This source lists general international sources as well as data sources for individual countries
Other sources of data

Census data

- Generally collected by government agencies (so look for bias)
- Worldwide compilations by IGOs and others

Financial, budget, accounts records
Church records, Land records

Survey data

- Governments, IGOs, NGOs
- Academic researchers
- Commercial entities

Observational data

- Retrospective case studies

Experimental data

- Clinical trials
Considerations when seeking data

What is the age of the data?
Where did they come from?
In what medium were they originally produced?
What is the geographical coverage of the data?
Does the data seem logical and consistent?
In what format are the data kept?
How were the data checked?
Why were the data compiled?
What is the reliability of the provider?
From what population do the data come?
When to give up

Consider the possibility that there is no data set that contains exactly the information you need in exactly the form you need it.

In trying to compare data, face the fact that because of different time frames, geographic limits, different categories and different definitions of terms, it may be impossible to integrate data from various sources.

Sample the data sets to check for compatibility, credibility, etc. If there are glaring inconsistencies, you probably need to abandon the data set.

Is it possible to construct your own data set?
In theory, census data includes all of a population.

- Undercount
- Overcount

Surveys rely on sampling to make estimates about a population.

- A sampling frame is required

**Probability samples**
- Simple random sample
- Systematic sample
- Stratified sample
- Cluster sample
- Multi-stage sample

**Non-probability samples**
- Volunteer sample
- Convenience sample
Sampling

Why sample?

To estimate something about a population
To make a prediction
To test a hypothesis

A **simple random sample** of size $n$ is a sample such that any subset of $n$ elements is equally likely to be selected.

A **stratified sample** consists of random samples from several strata of a population.

A **cluster sample** consists of samples from randomly selected subsets of a population.
Sampling and non-sampling error

Sampling error in making estimates about the population from a sample or in testing a hypothesis results from the process of sampling itself and can be controlled, but not eliminated.

Non-sampling error

- Selection bias
- Question(er) bias
- Response bias
  - Involuntary
  - Voluntary
Sample size and variation

Too small a sample: population differences cannot be detected or estimates will not be very accurate. The smaller the sample the larger the sampling error.

Too large a sample: costly (in money and in people) and may focus on differences too small to be of practical significance.

The larger the sample the larger the sampling error.

If the sample is not too large or too small, the sampling error does not depend on the size of the population.

But variation within the sample is also critical—the larger the variation, the larger the sampling error and the smaller the variation, the smaller the sampling error.
What might a data set look like?

<table>
<thead>
<tr>
<th>Score</th>
<th>gender</th>
<th>study</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

*Score* represents the scores on an exam administered to 36 individuals.

*Gender* is represented by 0 for males and 1 for females for computational purposes.

*Study* represents the number of hours that each individual studied for the exam.
Organizing data

Each line represents an observation
- Individuals
- Groups of people (averages)
- Countries
- Provinces
- Regions
- Cities
- Period of time

Each column represents values of a variable

Kinds of variables
- Categorical (qualitative)—gender, country, “dummy” variables
- Ordered—Likert scales (but be careful)
- Quantitative—age, income, number of newspapers read
Centrality measures

Mean
sum of all observations/number of observations ("average")

Median
if there are an odd number of ordered observations
median = middle observation

if there are an even number of ordered observations
median = mean the two middle observations

Mode
the most frequent observation
Normal curve

Data distribution

SD
Average tax cut

1. Bill Gates $999,010

2. Taxpayer 2 10

1000. Taxpayer 1000 10
$1,000,000

Average = $1,000,000/1000 = $1000

Bill Gates is an “outlier.”
From description to inference

From descriptive statistics we would like to make inferences

What can we say about a population from sample data?

How likely is an event?

Test an hypothesis about a population

Can we make predictions?

Outcome of election, need for facilities

What is the relation between variables?

Correlation v. causation
Distinguish between

**Uncertainty** is a lack of knowledge about specific factors, models, parameters, measurement, sampling, systematic errors, oversimplification of real world processes, misspecification of model structure, inappropriate proxy variables, descriptive or aggregation errors, misjudgment, incomplete analysis

**Variability** arises from real heterogeneity, diversity, results of natural random processes
Summary advice: Know when you need a statistician and get one as early as possible

The Role of a Statistician in a Project

Participate early and often to provide guidance in designing studies and collecting data
- Formulate the questions to be asked
- Consider the data needed to answer the questions
- Resist unrealistic expectations

Monitor the execution of studies and collection of data to provide a basis for accountability and cost-effectiveness
- From existing sources—creating methods to improve the quality of available data
- Through retrospective case studies
- Through surveys—sampling plan

Grapple with the data
- Clean the data

Design and carry out the analysis

Interpret the data and results and draw conclusions
Pitfalls for which statisticians may share responsibility

1) Failure to gain the support of key decision-makers
2) Unrealistic goals and expectations
3) Failure to develop a clear map of the process
4) Building of scenarios that are not credible
5) Inappropriate time frames and scopes
6) Failure to design a survey process to address the questions asked
Useful references

Software

SPSS         SAS         STATA         R

Basic text


More advanced


Sampling


Applications

Ball, Scheuren, Seltzer and Spirr, “Multiple or N-system estimates of the number of political killings in Guatemala,” *ASA Proceedings in Social Statistics*, pp. 156-160, 1999