

Statistical Issues in Road Safety: Uncertainty, Variability, Sampling

Shrikant I. Bangdiwala, PhD
McMaster University
Hamilton, Canada



I

IIT-Delhi 2019 Dec 07

Why statistics in road safety research?

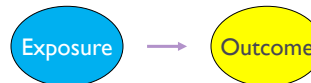
Our questions are not simple:

▶ When and how accidents occur?

▶ Understanding a situation → observe & estimate

▶ Why accidents occur?

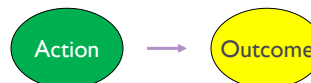
▶ Understanding relationships → observe & estimate;
association models



▶ What can affect occurrence of accidents?

▶ Evaluation of actions → experimental studies; intervene and
then observe & estimate;

→ test effectiveness



▶ 2

IIT-Delhi 2019 Dec 07

Why statistics in road safety research?

Road safety and traffic issues are complex:

- ▶ When and how accidents occur?
 - ▶ Multiple inter-connected factors
 - ▶ Why accidents occur?
 - ▶ Multiple factors may be associated; but causal relationship?
 - ▶ What is the 'risk' of occurrence? – probability, chance, usually not 0 or 1
 - ▶ What can affect occurrence of accidents?
 - ▶ Variability in exposures and in probabilities of occurrence
 - ▶ Need proper experimental designs
- **Uncertainties; Multiplicities**

▶ 3

IIT-Delhi 2019 Dec 07

Statistics – definitions I & II

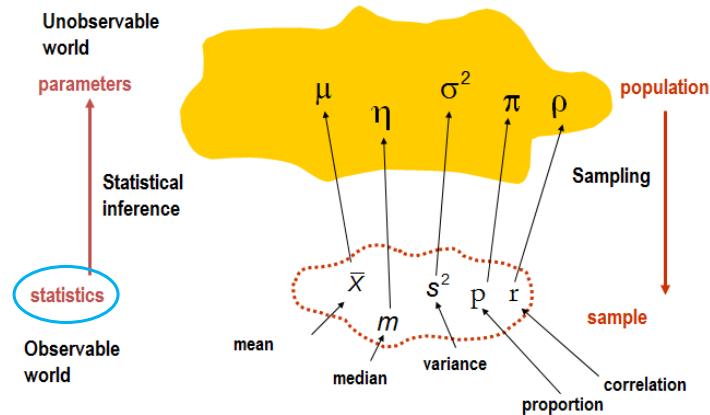
- ▶ Old definition – measurements of the state: 'stat' & 'ics'
 - ▶ Summarized → description of the population
 - ▶ Still used today:
 - ▶ Census – e.g. injury surveillance, FARS, IRTAD
 - ▶ Counts:
 - Police records of reported crashes – e.g. FIR
 - All hospitalizations due to trauma
 - All insurance claims for injuries/deaths
- ▶ Definition based on how to misuse/abuse information
 - ▶ A way to...
 - ▶ A way to...
 - ▶ Over-emphasis on 'significance' and 'p-values'

▶ 4

IIT-Delhi 2019 Dec 07

Statistics – definition III

- ▶ Scientific definition – measurements on a sample from the population



▶ 5

IIT-Delhi 2019 Dec 07

Role of **statistics** in addressing our questions

- ▶ Addressing our research questions in the face of **uncertainty**
 - ▶ Inherent variability in what we are studying
 - ▶ Incompleteness of information from sampling
 - ▶ Role of chance
- ▶ Statistics is the methodological science that allows for the understanding of **quantitative information** in the midst of **uncertainty**
 - ▶ Quantify it, Understand it, Reduce it, Control it
 - ▶ Probability (risk) models
 - ▶ Descriptive analyses
 - ▶ Controlled studies
 - ▶ Regression models

▶ 6

IIT-Delhi 2019 Dec 07

Statistics – collective & individual risks

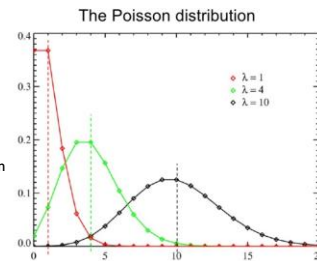
- ▶ The # of accidents in a given space/area L & time period T follows a Poisson distribution

$$\Pr(\# = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\text{where } \lambda = [\lambda_0 LT]$$



Siméon Denis Poisson
(1781-1840)



- ▶ The instantaneous conditional probability of being in an accident at a given time point t = hazard

$$h(t|X) = \frac{\Pr(\text{accident} = \text{yes in } \Delta t | X, \text{ not in accident before})}{\Delta t}$$

▶ 7

IIT-Delhi 2019 Dec 07

Modeling risks

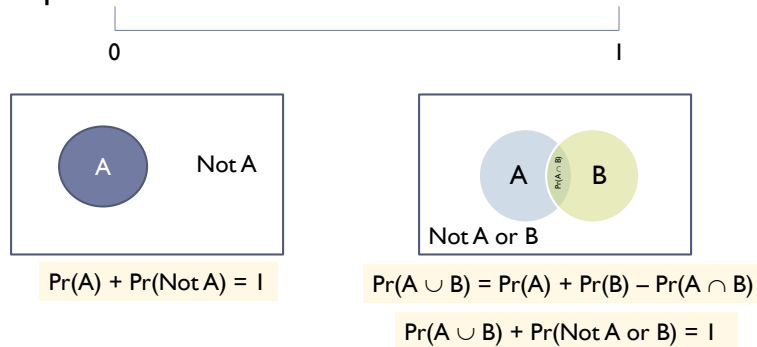
- ▶ We want to understand risks
- ▶ We need to control uncertainty in the estimation of the risks
 - ▶ Risk model of a trend in a given locality – mathematical functions
 - ▶ Risk models in multiple individuals or localities – statistical models
- ▶ Statistical methods are concerned with
 - ▶ ways to ‘control’ uncertainty
 - ▶ reduce variability
 - ▶ reduce sampling uncertainty
 → to understand estimates of risks (we provide bounds based on uncertainty) or relationships among quantitative factors and risks in a population

▶ 8

IIT-Delhi 2019 Dec 07

Probabilities are not well understood

- ▶ A probability is a theoretical mathematical concept
 - ▶ Derived from theoretical postulates – ‘updated’ with data [Bayes]
 - ▶ ‘Estimated’ from data – frequency approach
- ▶ Properties



▶ 9

IIT-Delhi 2019 Dec 07

Probabilities are not well understood

- ▶ A probability is a prediction in the future, it does not provide a ‘certainty’

What is the probability of electrocution?



▶ 10

IIT-Delhi 2019 Dec 07

Probabilities are not well understood

Relative risks of driving under different scenarios against not using phone

- Talking on a handheld phone
- Talking on a hands-free phone
- Drunk with BAC=0.10%
- Texting or reading email
- Talking with an adult passenger

Cell phones and crash risk

SOURCE: UNC HIGHWAY SAFETY RESEARCH CENTER DTH/NATASHA SMITH

▶ 11 IIT-Delhi 2019 Dec 07

Probabilities are not well understood

- ▶ Probabilities of being in a crash are low

- ▶ But the expected loss is HIGH:

$$E(L) = \text{Pr}(\text{crash}) * L(\text{per crash}) * \text{Exposure}(t)$$

▶ 12 IIT-Delhi 2019 Dec 07

Uncertainty

- ▶ When we estimate 'risks' as a probability – we do it with **uncertainty!!**
 - ▶ Instantaneous conditional risk \rightarrow hazard function $h(t|X)$
 - ▶ Number of accidents/victims \rightarrow distribution function (e.g. Poisson model, Negative binomial model, ...)
 - ▶ **Example:** Delhi pedestrian risks – from individual to collective
 - ▶ Individual risk is very low $\sim 0.00007 = 7 * 10^{-5}$ [how obtained?]
 - ▶ Collective risk is high since exposure is high 13,000,000 exposed [who is 'exposed?']
 - ▶ \rightarrow expect 910 pedestrian fatalities

▶ 13

IIT-Delhi 2019 Dec 07

Uncertainty

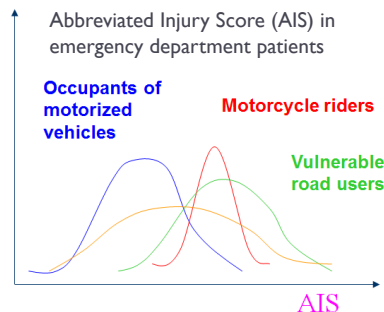
- ▶ When we estimate 'risks' – we do it with **uncertainty!!**
- \rightarrow Addressing our research questions in the face of **uncertainty**
 - ▶ Inherent variability in what we are studying
 - ▶ Incompleteness of information from sampling
 - ▶ Role of chance
 - ▶ Also: measurement error in everything we study !
 - ▶ Estimating numerator: outcomes
 - ▶ Estimating denominator: exposures

▶ 14

IIT-Delhi 2019 Dec 07

The study of variability

- ▶ Every crash is so particularly, uniquely different
- ▶ Statisticians do NOT study individual crashes or persons, but study groups of crashes or persons
 - ▶ The behavior of the group is called the 'distribution' of the behavior
- ▶ Researchers focus on the central tendency (mean, median, mode)
- ▶ Statisticians focus on the variability (variance, range)



▶ 15

IIT-Delhi 2019 Dec 07

Incompleteness → Uncertainty

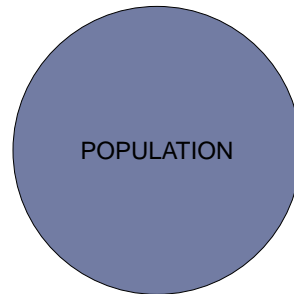
- ▶ In order to understand a situation → must study several occurrences
 - ▶ HOW MANY?
- ▶ Since we cannot usually study ALL situations, we study an incomplete subset
 - ▶ A 'sample' is never complete, leading to uncertainty
 - ▶ How **representative** is it of the complete set?

▶ 16

IIT-Delhi 2019 Dec 07

Why do we have uncertainty?

- ▶ Uncertainty from variability & incompleteness



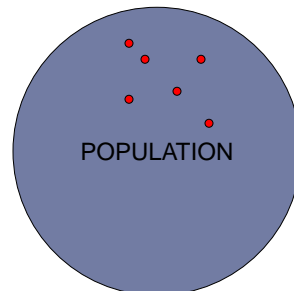
Assume we want to study a population

▶ 17

IIT-Delhi 2019 Dec 07

Why do we have uncertainty?

- ▶ Uncertainty from variability & incompleteness



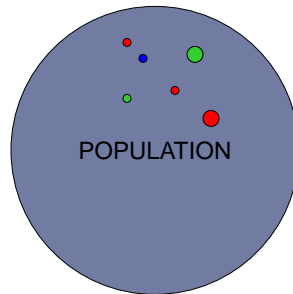
If all in a population are exactly the same, then we need to study _____

▶ 18

IIT-Delhi 2019 Dec 07

Why do we have uncertainty?

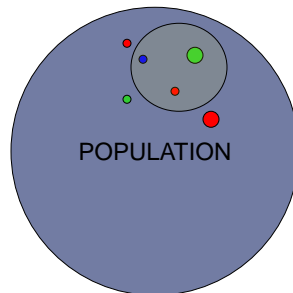
► Uncertainty from variability & incompleteness



Subjects in a population are NOT exactly the same, so then we need to study _____

Why do we have uncertainty?

► Uncertainty from variability & incompleteness



We sample a few →
We have observed an incomplete part of the population

Q1: Is the sample representative?

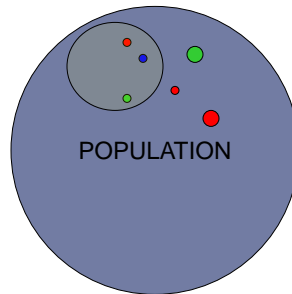
Q2: Is the sample size adequate?

Why do we have uncertainty?

► Uncertainty from chance

We sample a few →
Chance gave us the following sample

Q1: Is the sample representative?



► 21

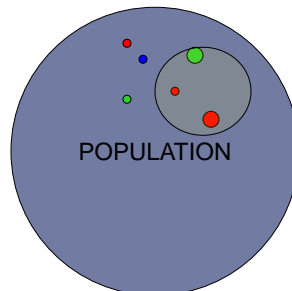
IIT-Delhi 2019 Dec 07

Why do we have uncertainty?

► Uncertainty from chance

We sample a few →
Chance gave us the following sample

Q1: Is the sample representative?



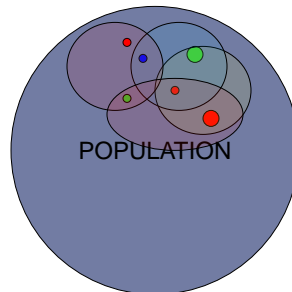
► 22

IIT-Delhi 2019 Dec 07

Why do we have uncertainty?

► Uncertainty from sampling

20 possible samples of size 3 – all equally likely to happen



We usually take only 1 sample
→
Chance gives 1 of many possible

The one we get is 'the luck of the draw' !!

We use it to 'guess' at the population, but we are never certain!

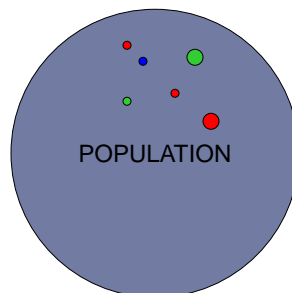
► 23

IIT-Delhi 2019 Dec 07

Uncertainty

► How can we eliminate the uncertainty?

- Reduce: stratified sampling
- Eliminate: study the entire population!



→Census; all medical records; all car crashes, ...

→There is no need for statistics, except for summarizing information

...but, \$\$\$ and often impractical or impossible!

► 24

IIT-Delhi 2019 Dec 07

Sampling process

- ▶ How do we select the sample?
 - ▶ Criteria
 - ▶ Sample should be 'like the population' → **representative**
 - ▶ Sample should be selected without introducing personal biases → objective
 - ▶ Sample should provide a 'correct estimate' of the population parameter → unbiased
 - ▶ Sample should provide a 'precise estimate' of the population parameter → 'adequate' size
- 'Probability' sample = we know the probability of selection of each person in the population

▶ 25

IIT-Delhi 2019 Dec 07

Sampling process

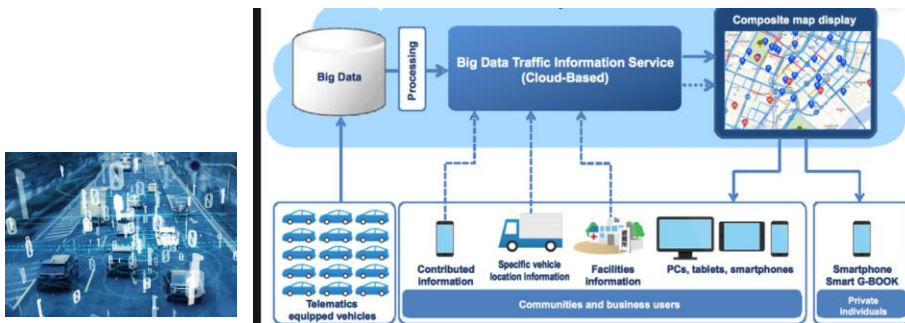
- ▶ 'Probability' samples
 - ▶ Simple random sample
 - ▶ Systematic random sample
 - ▶ Stratified random sample
 - ▶ Cluster random sample
 - ▶ Area random sample
 - ▶ Complex multi-stage probability sample
- ▶ What about 'purposely selected' sample?
 - ▶ Convenience sample = garbage sample
 - ▶ 'internet' sample ?
- ▶ What about not sampling and studying the entire population?

▶ 26

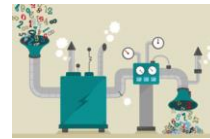
IIT-Delhi 2019 Dec 07

What about BIG data?

▶ Traffic information systems in 'Smart Cities'



- ▶ Remember data needs to be unbiased, high quality: Garbage in, Garbage out



▶ 27

IIT-Delhi 2019 Dec 07

Other sources of uncertainty

Imprecision

- ▶ Systematic errors – biases
 - ▶ Systematic **measurement** errors
 - ▶ Recall bias
 - ▶ Observer (instrument) bias
 - ▶ Data sources have different quality – classification bias
 - ▶ Systematic **sampling** errors
 - ▶ Selection biases
 - ▶ Data sources – different coverage
 - ▶ Non-response bias – missing data
- ▶ Random errors
 - ▶ Variation due to measurement
 - ▶ Variation due to sampling chance !

▶ 28

IIT-Delhi 2019 Dec 07

How can statistics help us?

- ▶ Statistics helps understand the behavior of quantitative data in **GROUPS**
 - ▶ In a **population**, we want to know:
 - ▶ Behavior of a single variable at a given time point – **risks**
 - ▶ Behavior of single variable over time - **trends**
 - ▶ Behavior of multiple variables – **relationships**
 - ▶ In a **sample** from the population, we are able to obtain:
 - ▶ Behavior of a single variable at a given time point – **estimation**
 - ▶ Behavior of single variable over time – **time series analyses**
 - ▶ Behavior of multiple variables – **regression models**

Research questions in Road safety

- ▶ What are the effects on risks of doing X?
 - ▶ X = decisions in engineering, planning, regulation & policy; education, ...
- ▶ Examine links between variables/factors and safety risks
- ▶ Themes
 - ▶ Accident analysis and prevention
 - ▶ Behavioral and social issues
 - ▶ Trauma care services
 - ▶ Legal and compliance issues

→ relationships

Unique issues in injury research

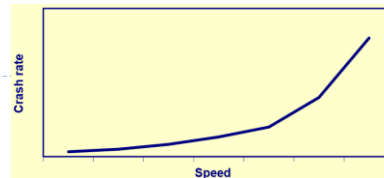
- ▶ Non-constant exposure → impact on appropriateness of indicators
- ▶ Counting rare events → impact on demonstrating effects and distributional models
- ▶ Multiple factors → complexities
- ▶ Intervening on the extreme cases → ‘regression to the mean’
- ▶ Study design options → observational vs experimental

▶ 31

IIT-Delhi 2019 Dec 07

Exercise

- ▶ Research Question:
Do lower speeds lead to safer roads?



- ▶ How do we answer this question?
 - ▶ What type of study?
 - ▶ How we define ‘lower’? How do we define ‘safer’?
 - ▶ Who or what do we study? How many?
 - ▶ Who or what do we compare results to? How many?
 - ▶ What data do we collect? How do we measure it? When do we measure? For how long do we measure?
 - ▶ What is a meaningful relationship?
 - ▶ How can we know if what we observe could have been due to chance?

▶ 32

IIT-Delhi 2019 Dec 07