Learning objectives

- Understand the role of statistics in the study of road injuries
  - Quantify events and behaviors
  - Identify risk factors
  - Understand relationships
  - Evaluate actions
- Understand the foundational principles of the science of statistics
  - Variability
  - Uncertainty
- Understand the two basic processes of statistics
  - Sampling
  - Statistical inference
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

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**
Statistics – definition I

- Old definition – measurements of the state: ‘stat’ & ‘ics’
  - Summarized → description of the population
  - Still used today:
    - Census – e.g. injury surveillance, Fatality Analysis Reporting System (FARS), International Road Traffic and Accident Database (IRTAD)
    - Counts:
      - Police records of all reported crashes
      - All hospitalizations due to trauma
      - All insurance claims for injuries/deaths
  - In India – difficult:

Statistics – definition II

- Scientific definition – measurements on a sample to understand a population (with uncertainty)
What do statistics do?

- 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**

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
  - Random error (imprecision) in what we observe
  - Systematic error (bias) in what we observe – e.g. under-reporting
  - Role of chance

- Statistics is the methodological science that allows for the understanding of **quantitative information** in the midst of uncertainty – a way of thinking about the uncertainty
  - Quantify it, Understand it, Reduce it, Control it
    - Probability (risk) models → quantify it and understand it by theory
    - Descriptive analyses → quantify it and understand it by analysis
    - Controlled studies → reduce uncertainty by design
    - Regression models → control it and understand it by analysis
The # of accidents in a given space/area L & time period T may follow a Poisson distribution
\[
\Pr(\# = k) = \frac{\lambda^k e^{-\lambda}}{k!}
\]
where \( \lambda = [\lambda_0 LT] \)

The instantaneous conditional probability (hazard) of being in an accident at a given time point t is called hazard, and may follow a Weibull, Gompertz or other ‘skewed’ extreme value distribution
\[
h(t|X) = \frac{Pr(\text{accident} = \text{yes in} \Delta t | X, \text{not in accident before})}{\Delta t}
\]

Theoretical probability models may or may not perfectly fit the observed data

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
\( \Rightarrow \) to understand estimates of risks (we provide bounds based on uncertainty) or relationships among quantitative factors and risks in a population
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

$$\text{Pr}(A) + \text{Pr}(\text{Not } A) = 1$$

$$\text{Pr}(A \cup B) = \text{Pr}(A) + \text{Pr}(B) - \text{Pr}(A \cap B)$$

$$\text{Pr}(A \cup B) + \text{Pr}(\text{Not } A \text{ or } B) = 1$$

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?

What is the probability of getting killed in Russian roulette?
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

Probabilities are not well understood

Probabilities of being in a crash are low

But the expected loss is HIGH:

\[ E(L) = \Pr(\text{crash}) \times L(\text{per crash}) \times \text{Exposure(t)} \]
Uncertainty

- When we estimate ‘risks’ as a probability – we do it with **uncertainty!!**
  - Cumulative individual/collective risk over time $\rightarrow$ risk=probability
  - Instantaneous conditional risk $\rightarrow$ hazard $h(t|X)$
  - 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 $\sim 0.00007 = 7 \times 10^{-5}$ [how obtained?]
  - Collective risk is high since exposure is high 13,000,000 exposed [who is ‘exposed’?]
  - $\rightarrow$ expect 910 pedestrian fatalities

---

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)
Incompleteness $\Rightarrow$ Uncertainty

- In order to understand a situation $\Rightarrow$ 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?

Why do we have uncertainty?

- Uncertainty from variability & incompleteness

Assume we want to study a population
Why do we have uncertainty?

- Uncertainty from variability & incompleteness

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

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?
Why do we have uncertainty?

- Uncertainty from chance
  
  We sample a few → Chance gave us the following sample

  Q1: Is the sample representative?

- Uncertainty from sampling
  
  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!
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!

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
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 'purposively selected' sample?
  - Convenience sample = garbage sample
  - 'internet' sample?

- What about not sampling and studying the entire population?

---

What about BIG data?

- Traffic information systems in 'Smart Cities'

- Remember data needs to be unbiased, high quality: Garbage in, Garbage out
Levels of information

- Quality of information – completeness, coverage, accuracy, precision
- Depth of information

Other sources of uncertainty

Imprecision

- Systematic errors – biases
  - Systematic measurement errors
    - Recall bias
    - Observer (instrument) bias
    - Data sources have different quality – classification bias
    - Biased underreporting
  - Systematic sampling errors
    - Selection biases
    - Data sources – different coverage
    - Non-response bias – missing data

Random errors

- Variation due to measurement
- Random underreporting
- Variation due to sampling chance!
Research questions in Road safety

- What are the effects on risks of doing X?
  - X = decisions in engineering, planning, regulation & policy; environment; 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
  - Models will be a based on a combination of physics and empirical data
- Multiple factors → complexities
- Intervening on the extreme cases → ‘regression to the mean’
- Study design options → observational vs experimental
### Study designs

- **Descriptive**
  - Case report
  - Case series
  - Descriptive epidemiological study
  - Ethnographic study
  - Other

- **Analytic**
  - Experimental
  - Randomized controlled trial
  - Non-randomized (Quasi-experimental)

- **Observational**
  - Case-control study
  - Case-crossover
  - Cross-sectional study
  - Longitudinal designs
  - Cohort study
  - Other

### Statistics – recall definition I

- **Original definition – measurements of the State**
  - STAT = state
  - _ICS = measurement of

- **Workers**
- **Taxes**

Why do governments care about the ‘statistics’?

- Plan mobility systems (→ ₹ required!)
  - road network
  - public transport network
- Plan health care systems (→ ₹ required!)
  - ambulance network
  - emergency care centers
  - trauma centers
- Plan interventions to reduce problem (and reduce ₹ !)

Desirable characteristics of statistics

- Obtainable (measurable, observable)
- Complete (i.e. not missing)
- Coverage (i.e. all people, places and times)
- Accurate (i.e. not biased)
- Precise (i.e. not ‘noisy’)
- Informative (i.e. worthwhile to collect/study)
Common statistics

- Number of deaths due to road transport
- Number of hospitalizations due to road transport
- Number of minor injuries due to road transport
- Number of crashes/collisions
- Number of drivers wearing seat belts
- Number of drivers driving impaired
  - Due to alcohol
  - Due to drugs
  - Due to tiredness/sleepiness
  - Due to distraction (e.g. using mobile phone; radio; other occupant)

Absolute counts of binary events

Statistics that are comparable

- If interested in deaths due to road transport

\[ \text{Risk} = \frac{\text{# of deaths due to road transport}}{\text{# of individuals exposed to road transport}} \]

- Risk = probability of occurrence of a binary event

Note: Often incorrectly called ‘rate’
- Rates involve a time period

Relative proportions of binary events
Risk vs Rate

- Incidence = occurrence of new cases in the population
- **Rate** = the number of new cases divided by total person-time of risk
  - Ranges between 0 to infinity
  - Numerator is number of cases
  - Denominator is total individuals and their time at risk (person-time)
- **Risk** = the proportion of the population who become cases over a fixed time interval (**average probability of becoming a case over a fixed time interval**)
  - Ranges between 0 to 1
  - Numerator is number of cases
  - Denominator is total number of individuals

Example #1:
Cohort study of crashes in 3 years

<table>
<thead>
<tr>
<th>Crashed</th>
<th>Yes (1)</th>
<th>No (0)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure – use of arterial highway Yes (1)</td>
<td>13</td>
<td>4987</td>
<td>5000</td>
</tr>
<tr>
<td>No (0)</td>
<td>7</td>
<td>9993</td>
<td>10000</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>14980</td>
<td>15000</td>
</tr>
</tbody>
</table>

- Overall risk of crash: \( p = \frac{20}{15000} = .0013 \)
- Risk of crash in arterial highway: \( \hat{p}_1 = \frac{13}{5000} = .0026 \)
- Risk of crash in non-arterial highway: \( \hat{p}_2 = \frac{7}{10000} = .0007 \)
Measures of effect

1. Risk difference
   \[ \hat{p}_1 - \hat{p}_2 = 0.019 \]

2. Relative risk
   \[ \frac{\hat{p}_1}{\hat{p}_2} = 3.71 \]

3. Odds ratio
   \[ \frac{\frac{\hat{p}_1}{1 - \hat{p}_1}}{\frac{\hat{p}_2}{1 - \hat{p}_2}} = \frac{\hat{p}_1(1 - \hat{p}_2)}{(1 - \hat{p}_1)\hat{p}_2} = 3.72 \]

Different study designs permit obtaining the answers to different research questions

- **Is X associated with Y?**
  - X is binary
  - Y is binary
  - X and Y are 'correlated' \( \rightarrow \) retrospective studies, cross-sectional studies
  - Odds Ratio (OR)

- **Is X a risk factor for Y?**
  - X is binary
  - Y is binary
  - X and Y are 'correlated' and X must be known before Y \( \rightarrow \) prospective studies
  - Relative Risk (RR)

- **Does X cause Y?**
  - X is binary
  - Y is binary
  - X and Y are 'correlated' and X must be known before Y and If X occurs, then Y occurs but if X does not occur, then Y does not occur \( \rightarrow \) prospective experimental studies
  - Relative Risk (RR)
Risk vs Odds

- Risk is a probability while Odds is a disparity

**Risk vs Odds**

<table>
<thead>
<tr>
<th>Total</th>
<th>Risk</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>YES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative risk vs Odds Ratio

**Relative Risk**

**Odds Ratio**

11/30/2023
Interpreting relative measures

- If OR, RR is > 1
  - \(=1.5 \rightarrow 1.5 \text { times more likely} \ldots\)
  - \(=2.0 \rightarrow \text{Twice as likely} \ldots \text{or double the} \ldots\)

- If OR, RR = 1
  - Just as likely \ldots \text{or no relationship}

- If OR, RR is < 1, easier to interpret reciprocal
  - \(=0.75 \rightarrow 1/(0.75) = 1.33 \rightarrow 1.33 \text { times less likely} \ldots\)
  - \(=0.50 \rightarrow 1/(0.50) = 2.00 \rightarrow \text{Twice less likely or 2 times less likely} \ldots\)

Bangladesh (2017) Interpretation of relative effects. Int J Injury Control & Safety Promotion

Interpreting relative measures

- RR = 2 is ‘equivalent’ to RR = 0.50
- RR = 5 is ‘equivalent’ to RR = 0.20
- RR = 10 is ‘equivalent’ to RR = 0.10
Some Comments on RR and OR

- If the study is prospective, RR is preferred
- In case-control study, should use OR
  - RR is biased
- If disease is rare (less than 10%), then OR is similar in magnitude to RR

Example #1:

**Cohort study** of crashes in 3 years

exposure (regular driving in arterial highway) and outcome (involved in crash)

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>13</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>Noncases</td>
<td>4987</td>
<td>9993</td>
<td>14980</td>
</tr>
<tr>
<td>Total</td>
<td>5000</td>
<td>10000</td>
<td>15000</td>
</tr>
<tr>
<td>Risk</td>
<td>.0026</td>
<td>.0007</td>
<td>.0013333</td>
</tr>
<tr>
<td>Pt. Est.</td>
<td></td>
<td>[95% Conf. Interval]</td>
<td></td>
</tr>
<tr>
<td>Risk difference</td>
<td>.0019</td>
<td>.0003963</td>
<td>.0034037</td>
</tr>
<tr>
<td>Risk ratio</td>
<td>3.714286</td>
<td>1.482854</td>
<td>9.303627</td>
</tr>
</tbody>
</table>

OR = 3.72 !!
Example #2: Case-control study

exposure (regular driving in arterial highway) and outcome (involved in crash)

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
<th>Total</th>
<th>Proportion Exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>129</td>
<td>221</td>
<td>350</td>
<td>0.3686</td>
</tr>
<tr>
<td>Controls</td>
<td>43</td>
<td>307</td>
<td>350</td>
<td>0.1229</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>528</td>
<td>700</td>
<td>0.2457</td>
</tr>
</tbody>
</table>

Pt. Est. [95% Conf. Interval]

Odds ratio: 4.1674  [2.7931, 6.2772] (exact)

Incorrect, but calculate RR of exposure = 0.3686 / 0.1229 = 2.99

When to use OR vs RR?

- Depends on the study design
  - Cross-sectional
    - OR of being exposed
    - OR of having the outcome
  - Case-control
    - OR of having been exposed
  - Cohort or RCT
    - RR of developing outcome (cumulative risk)
    - HR of instantaneously, conditionally getting the outcome
### What is the appropriate denominator?

- **Per number of population** → understand it as a public health problem
- **Per number of vehicles** → useful for insurance companies
- **Per number of km of road network**
- **Per density**
  - Population per km of road network
  - Vehicles per km of road network
- **Per exposure**
  - Amount of fuel/petrol consumed
  - Amount of vehicle-km traversed
Different interpretations of indicators

![Graph showing different interpretations of indicators.]

Bangderwa et al. (1985) Accident Analysis & Prevention

Fatalities by country per capita income

![Graph showing fatalities by country per capita income.]

Mohan et al. (2017) Journal of Safety Research
Not all indicators are based on ‘binary events’

- Continuous
  - Injury severity score (ISS) – lesion specific
  - Abbreviated injury severity (AIS) – person specific
  - Cost
    - Emergency care, Hospitalization
    - Loss of earning income

- Categorical ordinal
  - Severity on a VAS

- Categorical nominal
  - Anatomical place of injury
  - Type of injury – fracture, burn, laceration,

Statistics used to estimate a parameter

- How large a sample should I have?
  - Enough to account for variability

For estimation

The necessary sample size $n$ when estimating the effect of an intervention in a given group is

- higher if want higher level of confidence
- higher if want higher precision = lower half-width of confidence interval
- higher if there is more variability

Conceptually

$$n \propto \frac{(\text{confidence}) \cdot (\text{variability})}{(\text{width})^2}$$

Class exercise:
Is a sample of size 3 better or not than a sample of size 10?
Confidence intervals

Limits on precision:

- Specify a lower bound and an upper bound within which we are highly confident the true (unknown) value of the parameter lies
- The interval is centered around our point estimate
- The width of the confidence interval
  - Increases with our desire for bigger confidence
  - Is larger if there is larger variability (‘noise’ in the system)
  - Is smaller if based on larger sample sizes

Example

Are 16-year-old drivers with teenager passengers at higher risk of a crash than without other passenger?

<table>
<thead>
<tr>
<th>Driver age</th>
<th>Relative risk (all)</th>
<th>95% confidence intervals</th>
<th>Driver alone</th>
<th>95% confidence intervals</th>
<th>With passengers</th>
<th>95% confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>3.28</td>
<td>3.05-3.51</td>
<td>2.28</td>
<td>1.85-2.85</td>
<td>4.72</td>
<td>4.12-5.31</td>
</tr>
<tr>
<td>17</td>
<td>2.84</td>
<td>2.53-3.20</td>
<td>1.77</td>
<td>1.35-2.22</td>
<td>3.32</td>
<td>2.85-3.80</td>
</tr>
<tr>
<td>18</td>
<td>2.47</td>
<td>2.34-2.59</td>
<td>1.77</td>
<td>1.30-2.23</td>
<td>2.93</td>
<td>2.48-3.35</td>
</tr>
<tr>
<td>19</td>
<td>2.19</td>
<td>2.06-2.33</td>
<td>1.61</td>
<td>1.20-2.07</td>
<td>2.83</td>
<td>2.47-2.64</td>
</tr>
<tr>
<td>20-24</td>
<td>1.90</td>
<td>1.82-2.00</td>
<td>1.50</td>
<td>1.24-1.72</td>
<td>2.15</td>
<td>1.84-2.47</td>
</tr>
<tr>
<td>25-29</td>
<td>1.43</td>
<td>1.35-1.53</td>
<td>1.28</td>
<td>1.04-1.55</td>
<td>1.92</td>
<td>1.60-2.28</td>
</tr>
<tr>
<td>30-54</td>
<td>1.16</td>
<td>1.06-1.27</td>
<td>1.13</td>
<td>0.90-1.39</td>
<td>1.81</td>
<td>1.50-2.18</td>
</tr>
<tr>
<td>55+</td>
<td>1.03</td>
<td>1.00-1.07</td>
<td>1.00</td>
<td>0.87-1.18</td>
<td>1.81</td>
<td>1.57-2.03</td>
</tr>
</tbody>
</table>

Table 3. Relative risk of fatal crash involvement by driver age and passenger presence (FARS, 1990–1995)


The 30-59 age group is the reference group for relative risk calculations.
Statistics used to test a claim

- Suppose $\delta$ is a difference that would be considered clinically/contextually important (on a population level or individual level)

The Significance of a statistic is not related to the Importance of the statistic

References


- Bangdiwala SI, Anzola-Perez E, Glizer IM. Statistical considerations for the interpretation of commonly utilized road traffic accident indicators: Implications for developing countries. Accident Analysis & Prevention, 1985, 17(6):419-427.
Exercise

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

- How do we answer this question?
  - What type of study?
  - How do 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?