NITheCS
National Institute for
Theoretical and Computational Sciences

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

# CHPC & NITHECS CODING SUMMER SCHOOL

# Probability & Statistics

## Categorical Data & Probability

**paul j. van Staden (PhD)**
Department of Statistics
University of Pretoria

# PROBABILITY: Contingency Tables

## EXAMPLE: Jerusalema

- A researcher was interested in whether animals could be trained to line dance.
- She took 80 cats and tried to train them to line dance by giving them either food or love as a reward.
- At the end of the week, she counted how many cats could line dance and how many could not.

EXAMPLE adapted from:

Field, A. (2009). *Discovering Statistics using SPSS (and sex and drugs and rock 'n' roll)*, SAGE Publications Ltd, London, UK.

| | | Line dance | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Reward | Food | 28 | 12 | 40 |
| | Love | 16 | 24 | 40 |
| | Total | 44 | 36 | 80 |

**2-by-2 contingency table showing how many cats could line dance after being trained with different rewards**

- One of the cats is randomly selected.
- Consider the following two events:

  $A$: The cat could line dance.

  $B$: The cat was given food as reward.

- What is the probability that the cat could line dance?

$$P(A) = \frac{44}{80} = 0.55$$

- What is the probability that the cat could not line dance?

$$P(\overline{A}) = 1 - P(A) = 1 - \frac{44}{80} = \frac{36}{80} = 0.45$$

- What is the probability that the cat's reward was food?

$$P(B) = \frac{40}{80} = 0.5$$

- **What is the probability that the cat could line dance and that the cat's reward was food?**

$$P(A \cap B) = \frac{28}{80} = 0.35$$

- **What is the probability that the cat could line dance given that the cat's reward was food?**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{28/80}{40/80} = \frac{28}{40} = 0.7$$

- **Does learning to line dance depend on the type of reward?**

$$P(A) = \frac{44}{80} = 0.55$$

$$P(A|B) = \frac{28}{40} = 0.7$$

$$P(A) \neq P(A|B)$$

**Therefore $A$ and $B$ are not independent events.**

**In effect, learning to line dance depends on the type of award.**

# CATEGORICAL DATA: Testing for independence

## EXAMPLE: Jerusalema continues…

| | | Line dance | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| Reward | Food | 28 (22) | 12 (18) | 40 |
| | Love | 16 (22) | 24 (18) | 40 |
| | Total | 44 | 36 | 80 |

**2-by-2 contingency table showing how many cats could line dance after being trained with different rewards**

- **Use Pearson's chi-square test to test whether the type of reward and learning to line dance are independent.**

  $H_0$: Type of reward and learning to line dance are ***independent***.
  $H_A$: Type of reward and learning to line dance are ***dependent***.

  **Compare the observed counts with the expected counts.**

  **Observed counts:** $O_{i,j}$

  **Expected counts:** $E_{i,j} = \frac{i^{\text{th}} \text{ row total} \times j^{\text{th}} \text{ column total}}{n}$

  **Test statistic:** $\chi^2 = \sum_i \sum_j \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$

## Python code:

```
# Import packages and functions
import numpy as np
from scipy.stats import chi2_contingency
# Give the contingency table with observed counts
ct = np.array([[28, 12], [16, 24]])
# Do the test for independence
chi2_contingency(ct)
```

## Output:

```
Chi2ContingencyResult(statistic=6.1111111111111107,
                      pvalue=0.01343346520577156,
                      dof=1,
                      expected_freq=array([[ 22.,  18.],
                                           [ 22.,  18.]]))
```

## Since $p$-value < 0.05, $H_0$ is rejected at a 5% significance level.

## Type of reward and learning to line dance are **dependent**.

# EXERCISE: Titanic

- The following contingency table summarizes the survival of 2201 passengers on the Titanic according to their economic status (class on ship).

| | | Survived | |
|---|---|---|---|
| | | No | Yes |
| Class | 1st | 122 | 203 |
| | 2nd | 167 | 118 |
| | 3rd | 528 | 178 |
| | Crew | 673 | 212 |

**4-by-2 contingency table summarizing the survival of passengers on the Titanic according to economic status**

- Test whether the survival of the passengers on the Titanic was independent from their economic status (class).

# PROBABILITY: Tree Diagrams & Bayes' Rule

## EXAMPLE: Vital Burning

- **In the forensic assessment of burned bodies, it is of crucial importance to determine whether the victim was exposed to the fire before or after death.**
- **The presence of soot in the respiratory tract and stomach is considered a good indicator of vital burning.**
- **Bernitz *et al*. (2014) proposed that tongue protrusion can be used as an additional indicator of vital burning.**

Bernitz, H., van Staden, P.J., Cronjé, C.M. and Sutherland, R. (2014). Tongue protrusion as an indicator of vital burning, *International Journal of Legal Medicine*, 128(2), 309–312.

- Suppose it is known that soot is present in 84% of burned victims.
- Of those burned victims with soot present, 75% have protruded tongues, while only 12.5% of the burned victims without soot present have protruded tongues.
- A burned victim is randomly selected.
- Given that this burned victim has a protruded tongue, what is the probability that the victim will have soot in the respiratory tract and stomach?

- **Consider the following two events:**

  $A$: **Soot present.**

  $B$: **Tongue protruded.**


- **We know that:**

  $P(A) = 0.84 \qquad P(B|A) = 0.75 \qquad P(B|\bar{A}) = 0.125$

- **We must calculate:**

  $P(A|B)$

- **To do so, we need to calculate:**

  $P(A \cap B) \qquad P(B)$

- **The probability that a randomly selected burned victim has soot present and a protruded tongue is:**

$$P(A \cap B) = P(A) \times P(B|A) = 0.84 \times 0.75 = 0.63$$

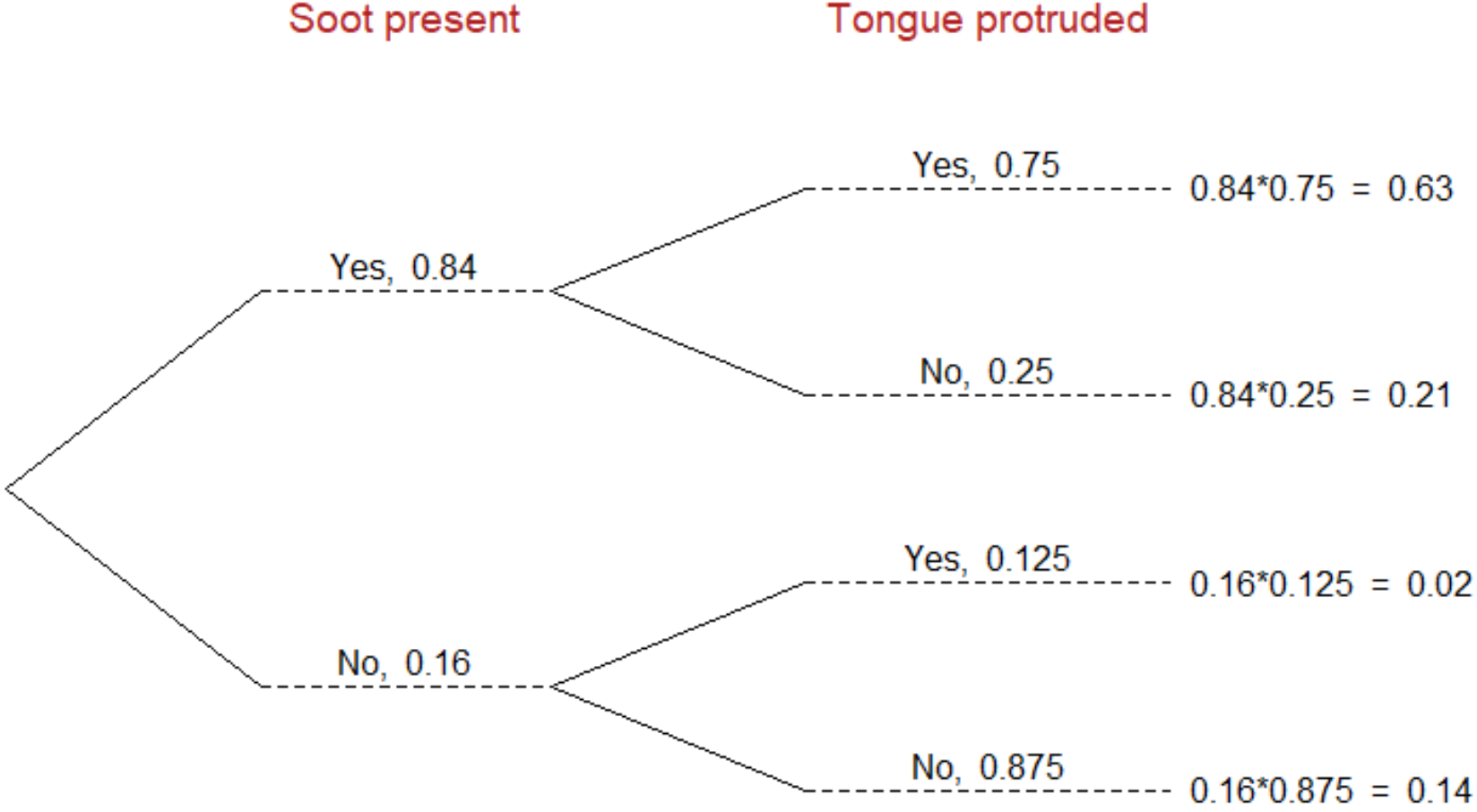
- **Similarly:**

$$P(\overline{A} \cap B) = P(\overline{A}) \times P(B|\overline{A}) = 0.16 \times 0.125 = 0.02$$

$$P(A \cap \overline{B}) = P(A) \times P(\overline{B}|A) = 0.84 \times 0.25 = 0.21$$

$$P(\overline{A} \cap \overline{B}) = P(\overline{A}) \times P(\overline{B}|\overline{A}) = 0.16 \times 0.875 = 0.14$$

# Tree diagram describing soot presence and tongue protrusion



- **The probability that a randomly selected burned victim has a protruded tongue is:**

$$P(B) = P(A \cap B) + P(\overline{A} \cap B) = 0.63 + 0.02 = 0.65$$

- **Finally, given that the randomly selected burned victim has a protruded tongue, the probability that this victim will have soot in the respiratory tract and stomach is:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.63}{0.65} = 0.97$$

# EXERCISE: Vital Burning

- Apart from the presence of soot, a high blood level percentage for carboxyhaemoglobin (COHb), specifically COHb% ≥ 10%, is also indicative of vital burning.
- Suppose again that it is known that soot is present in 84% of burned victims. Assume that of those victims with soot present, 95% have COHb% ≥ 10%, while 65% of the victims without soot present have COHb% ≥ 10%.
- A burned victim is randomly selected.
- Given that this burned victim has COHb% ≥ 10%, what is the probability that the victim will have soot present?