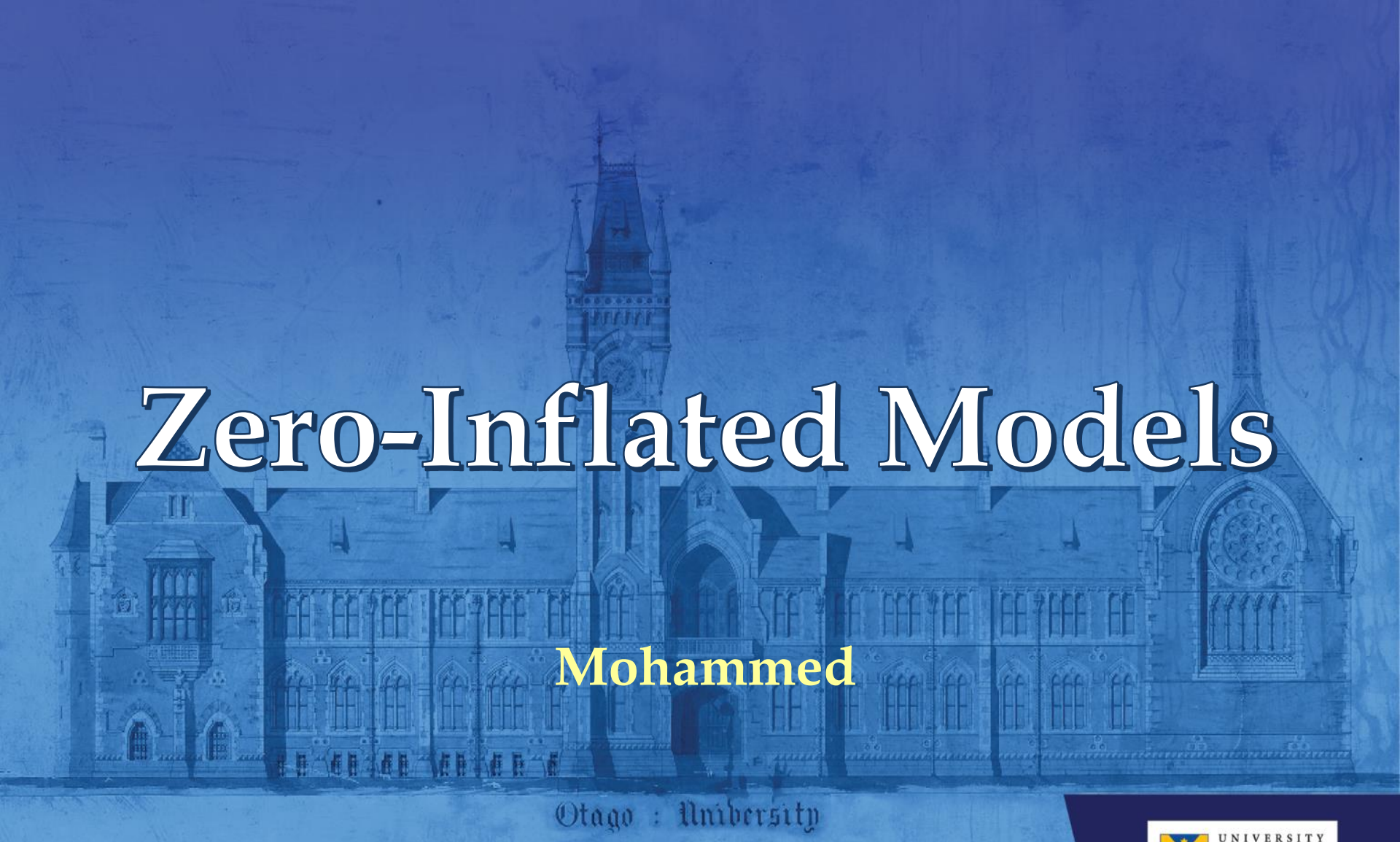
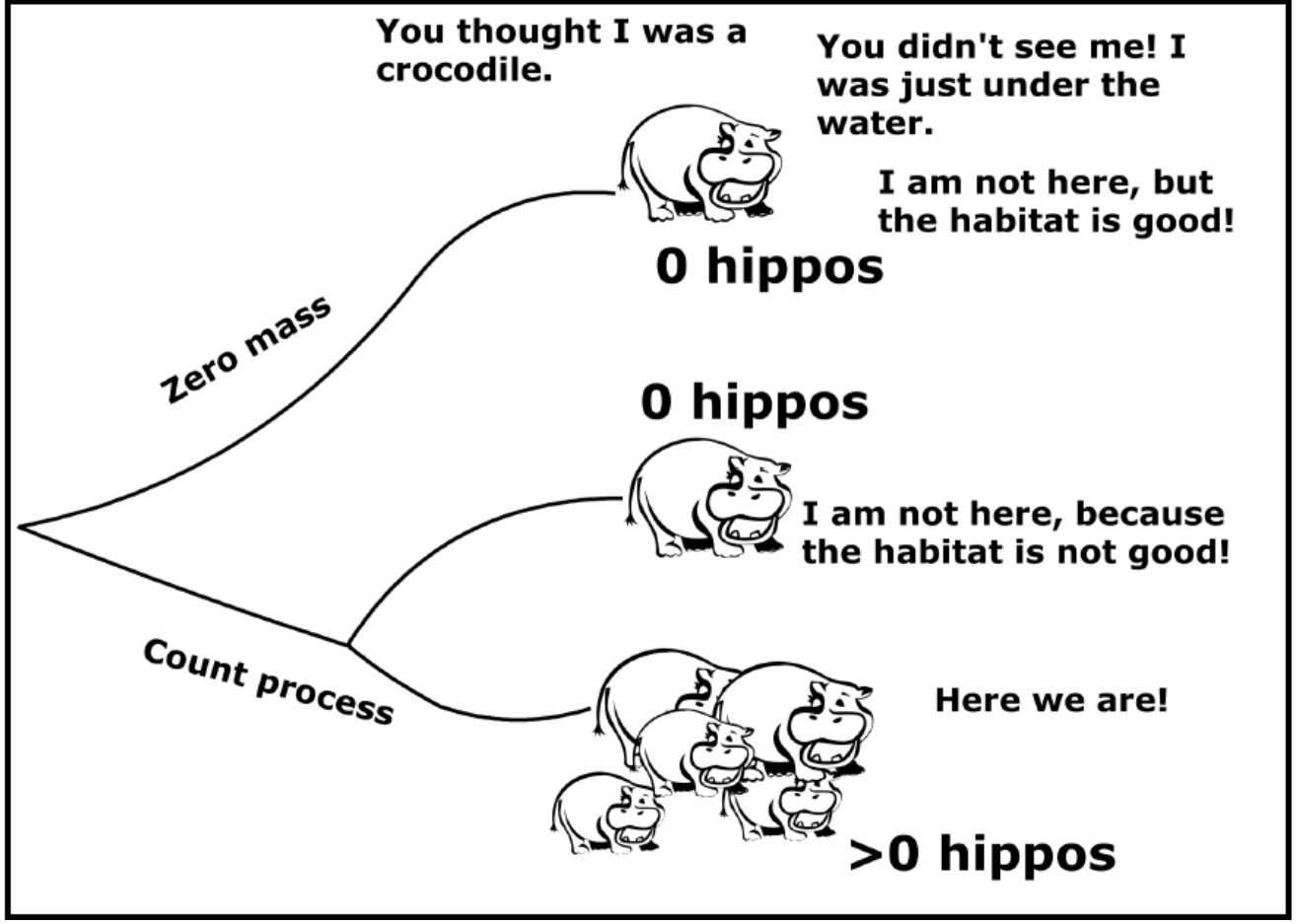
Zero-Inflated Models

Mohammed





What is zero inflation?

Suppose you want to study hippos and the effect of habitat variables on their distribution. When sampling, you may count zero hippos at many sites and as a result standard statistical techniques like regression and GLM are not applicable, therefore zero inflated models should be used

Issues: Excess zeros

- Often, the numbers of zeros in the sample cannot be accommodated properly by a Poisson or Negative Binomial model. Both models would underpredict them.
- There is said to be an "*excess zeros*" problem.
- New models are needed to deal with these type of data.

Issues: Excess zeros

Types:

- Zero-Inflated Poisson (ZIP)
- Zero-Inflated Negative Binomial (ZINB) Models
- Hurdle models
- These models are designed to deal with situations where there is an “excessive” number of individuals with a count of 0.
- Poisson regression models provide a standard framework for the analysis of count data.
- In practice, however, count data are often over-dispersed relative to the Poisson distribution.

Over-dispersion

- Because the Poisson model assumes that the conditional variance of the dependent variable is equal to the conditional mean.
- In most count data sets, the conditional variance is greater than the conditional mean, often much greater, a phenomenon known as *over-dispersion*.

Consequence of over-dispersion

- Standard errors will be underestimated
- Potential for overconfidence in results; rejecting H_0 when you shouldn't!
- Note: over-dispersion doesn't necessarily affect predicted counts (compared to alternative models).

Issues: Excess zeros

- If data consist of non-negative, highly skewed sequence counts with a large proportion of zeros. Zero-inflated models are useful for analysing such data.
- Moreover, the non-zero observations may be over-dispersed in relation to the Poisson distribution, biasing parameter estimates and underestimating standard errors.
- In such a circumstance, a zero-inflated negative binomial (ZINB) model better accounts for these characteristics compared to a zero-inflated Poisson (ZIP).

Zero-Inflated Models

- These models, called *Two-part models*, allow for two different process:
  - one drives whether the value is 0 or positive (participation part), and
  - the other one drives the value of the strictly positive count (amount part).

- Proposed models:
  - Zero inflated models
  - Hurdle models

Zero Inflation – ZIP Models

Structure:

- Zero-inflated Poisson model have two kinds of zeros: “true zeros” and “excess zeros.”
- Two groups of people: Always Zero & Not Always Zero
- Example: Investors (traders) who sometime just did not trade that week versus investors who never ever do.

Zero Inflation – ZIP Models

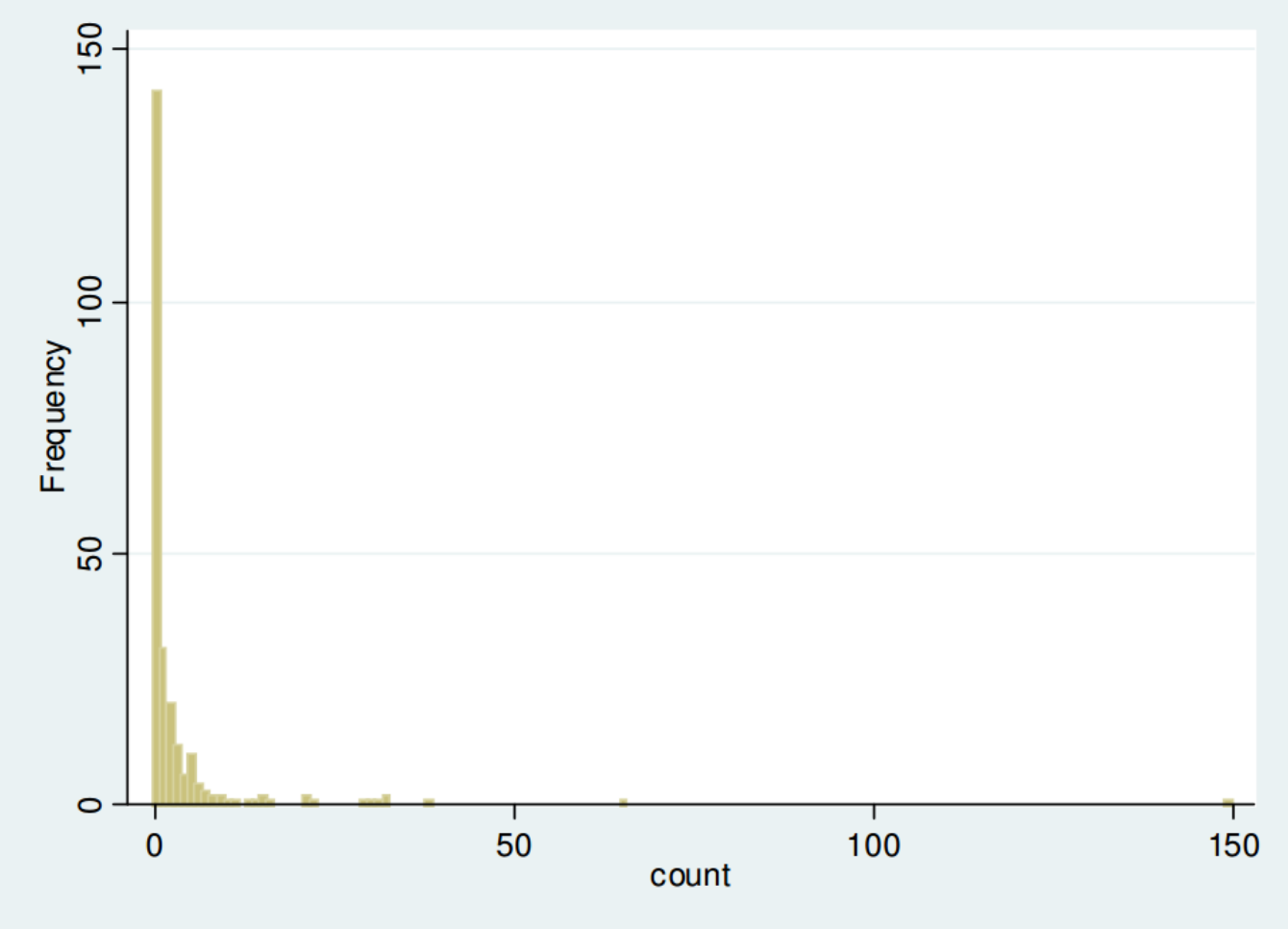
- Two models: (1) for the count and (2) for excess zeros. The key difference is that the count model allows zeros now.
- If we are interested in modelling trading, the zeros from investors who will never trade are not relevant. But, we only observe the zero, not the type of investor. This is the *excess zeros problem*.

Zero-inflated model

Simple definition:

- In statistics, a zero-inflated model is a statistical model based on a zero-inflated probability distribution, *i.e*. a distribution that allows for frequent zero-valued observations.
- Zero-inflated Poisson (ZIP) model is used to model data with excess zeroes

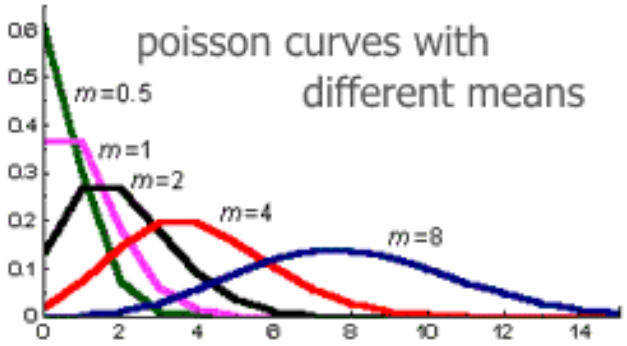
Zero Inflation – ZIP Models



- Note: lots of zeros

Zero-inflated Poisson

- The first zero-inflated model is zero-inflated Poisson model. The zero-inflated Poisson model concerns a random event containing excess zero-count data in unit time
- The Poisson Distribution is a discrete distribution which takes on the values of X = 0, 1, 2, 3,….. It is often used as a model for the number of events in a specific time period..
- Also used to calculate the probability of a number of successes that take place in a certain interval of time or space.



Zero-inflated Poisson

- The zero-inflated Poisson (ZIP) model employs two components that correspond to two zero generating processes.
- The first process is governed by a binary distribution that generates extra zeros. (OR or LOGIT models commonly used)
- The second process is governed by a Poisson distribution that generates count (counting zeroes), some of which may be zero.
- The two model components are described as follows:

Zero-inflated Poisson

$$Pr(y_j = 0) = \pi + (1 + \pi)e^{-\lambda}$$

$$Pr(y_j = h_i) = (1 - \pi)\frac{\lambda^{hi}e^{-\lambda}}{h_i!}, \qquad h_i \geq 1$$

where the outcome variable y_j has any non-negative integer value (h = observed count), λ_i is the expected Poisson count (expected count and variance) for the i^{th} individual (Called mu (μ) in some texts); π is the probability of extra zeros.

The mean is $(1 - \pi)\,\lambda$ and the variance is $\lambda\,(1 - \pi)\,(1 + \lambda\,\pi)$

E.g., the number of insurance claims within a population for a certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim.

Notes on Zero Inflation Models

- Poisson is not nested in ZIP
- Standard tests are not appropriate
- Use Vuong statistic. ZIP model almost always wins.

- Zero Inflation models extend to NB models –ZINB are standard models
  - Creates two sources of over-dispersion
  - Generally difficult to estimate

Zero-Inflated Negative Binomial (ZINB)

- The zero-inflated negative binomial (ZINB) distribution is a mixture of binary distribution that is degenerate at zero and an ordinary count distribution such as negative binomial
- The negative binomial regression can be written as an extension of Poisson regression and it enables the model to have greater flexibility in modelling the relationship between the conditional variance and the conditional mean compared to the Poisson model.
- The binary distribution captures the excess number of zeros, which exceed those predicted by the negative binomial distribution.

Hurdle Models

- A hurdle model is also a modified count model with two parts:

  - one generating the zeros

  - one generating the positive values.

- The models are not constrained to be the same.

- A binomial probability model governs the binary outcome of whether a count variable has a zero or a positive value.

  If $y_i > 0$, the "hurdle is crossed," the conditional distribution of the positive values is governed by a zero-truncated count model.

- Popular models in health economics (use of health care facilities, counselling, drugs, alcohol, etc.).

Take away message

- The zero inflated Poisson (ZIP) model is one way to allow for over-dispersion
- This model assumes that the sample is a “mixture” of two sorts of individuals:
  - one group whose counts are generated by the standard Poisson regression model, and
  - another group (call them the absolute zero group) who have zero probability of a count greater than 0.
- Observed values of 0 could come from either group.
- Although not essential, the model is typically elaborated to include a logistic regression model predicting which group an individual belongs to.

Take away message

But what about the zero-inflated negative binomial (ZINB) model?

- It's certainly possible that a ZINB model could fit better than a conventional negative binomial model regression model.
- But, the latter is a special case of the former, so it's easy to do a likelihood ratio test to compare them (by taking twice the positive difference in the log-likelihoods)
- So next time thinking about fitting a zero-inflated regression model, first consider whether a conventional negative binomial model might be good enough. Having a lot of zeros doesn't necessarily mean that you need a zero-inflated model.

Take away message

- In cases of over-dispersion, the ZIP model typically fits better than a standard Poisson model.
- But there's another model that allows for over-dispersion, and that's the standard negative binomial regression model.
- Experts says; the negative binomial model fits much better than a ZIP model, as evaluated by AIC or BIC statistics and it's a much simpler model to estimate and interpret

Ref: http://statisticalhorizons.com/zero-inflated-models

Thank You