ST 518: Data Analytics II

Categorical & Count Data

Reflections on Data Analytics I

What's Different with Categorical and Count Data?

- Categorical Data are Distinctly Non-Normal
- Complicated Variance Properties
- Different Questions involving Categorical/Count Data

Recalling Data Analytics I

In the Data Analytics I course you covered methods for modeling continuous response variables:

- Simple and multiple linear regression
- Analysis of variance (ANOVA) models
- Linear mixed effects models
- Issues for large datasets

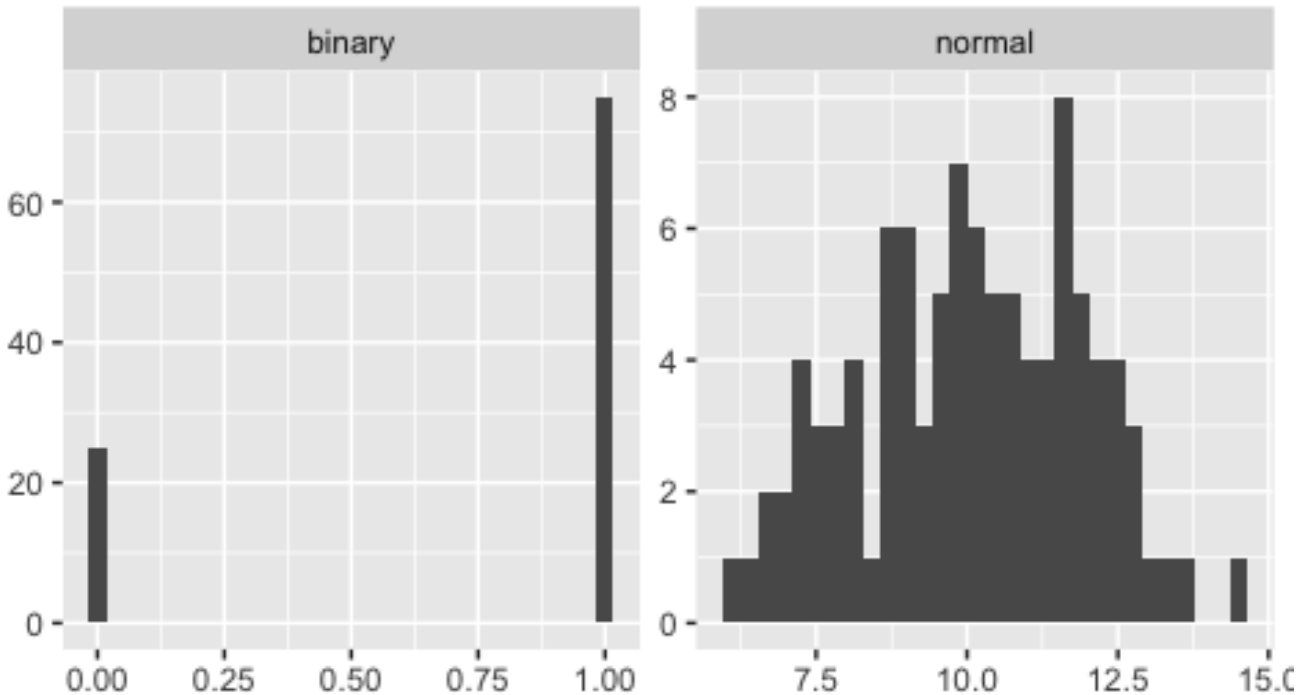
And, you saw in that course that those methods are fairly robust to departures from the Normality assumption that underlies most of those models.

What's the Big Deal?

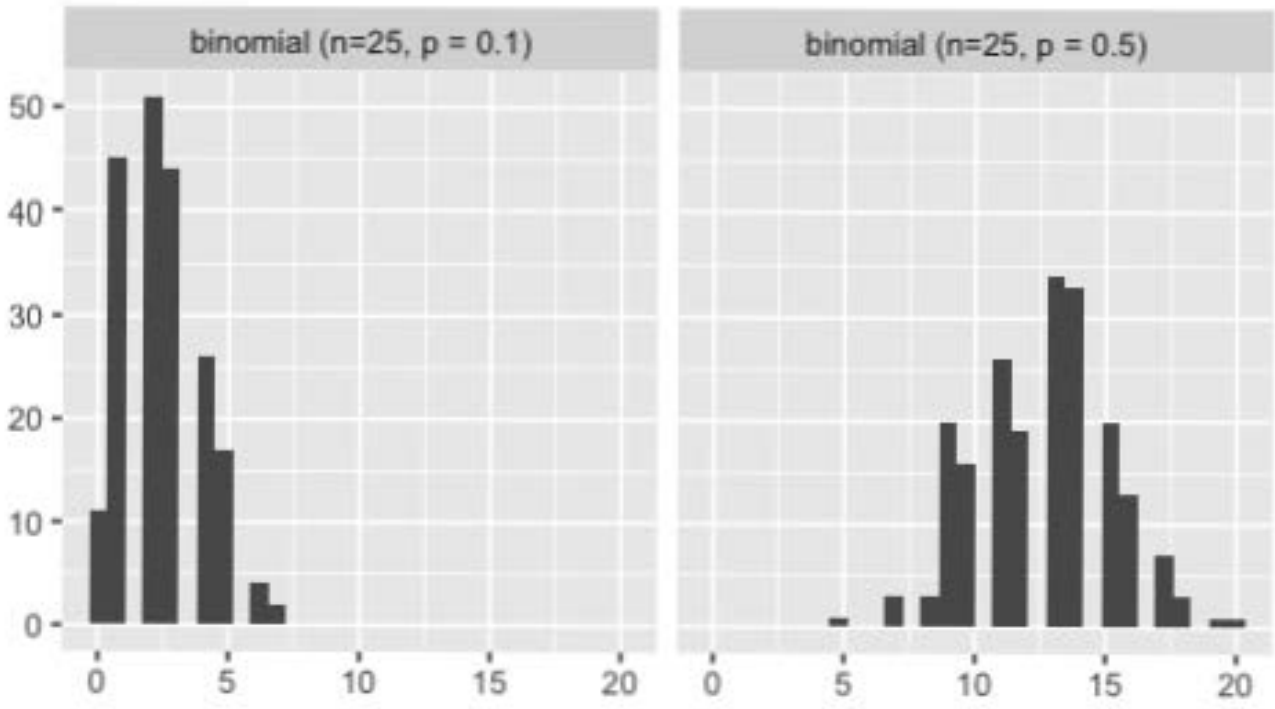
Why do we have to deal with categorial and count data separately?!

Distinctly Non-Normal

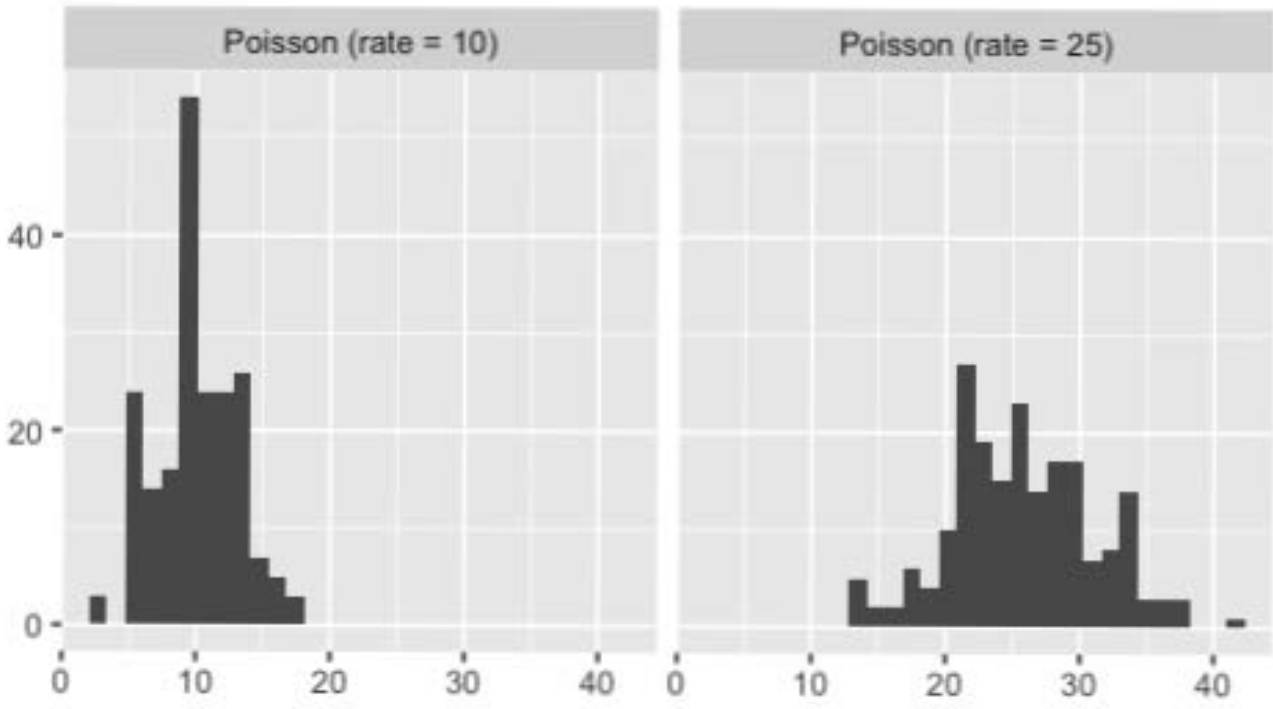
Consider binary data—where a response can take only one of two values (e.g., true/false, yes/no, 0/1).



Variance Properties (Binomial Counts)



Variance Properties (Poisson Counts)



Complicated Variance Properties

A key distinction between a continuous random variable and categorical or count random variabe is that **with a categorical or count random variable, it's fairly typical that changes in the mean correspond to changes in the variance.**

- Recall from Data Analytics I that the ANOVA and multiple regression models are NOT very robust to changes in the constant variance assumption.
- This means that for categorical and count data, we have to consider a different suite of statistical methods that accommodate the variance features of these types of data.

Questions about Means

In addition to these complications introduced by the variance properties of categorical and count data, we are often interested in distinctly different types of questions with these data than we are with continuous response data.

- When our response variable is continuous, we typically address questions about the mean of that response variable; for example:

    1. How does the mean change with respect to a change in an explanatory variable.
    2. Are the means in two or more populations different from each other, and if so, in what ways?

Questions about Proportions, Risk and Odds

With categorical data (and some count data), we typically address questions about **proportions** (or probabilities), **risk** and **odds**.

- We'll define and discuss these quantities in more detail in the next module of the course.

With some count data, we still may be interested in questions about the mean, but we nevertheless have to address the issue raised by the histograms on the earlier slides—with count data, changes in the mean correspond to changes in the variance as well.