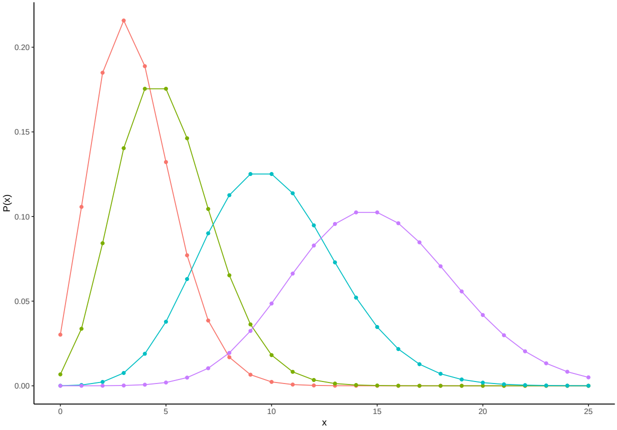
# *Negative binomial regression*

Mark Andrews
Psychology Department, Nottingham Trent University

mark.andrews@ntu.ac.uk

# *The Poisson Distribution*

- The mean of a Poisson distribution is equal to its rate parameter λ.
- Its variance is also equal to λ.




As λ increases, so too does the variance.

# *Means and variances in a Poisson distribution:*

- In a Poisson distribution, the variance of a sample should be approximately the same as the mean of a sample.
- Example 1:

```
x <- rpois(25, lambda = 5)
c(mean(x), var(x), var(x)/mean(x))
```

```
## [1] 5.4400000 5.4233333 0.9969363
```

- Example 2:

```
x <- rpois(25, lambda = 5)
c(mean(x), var(x), var(x)/mean(x))
```

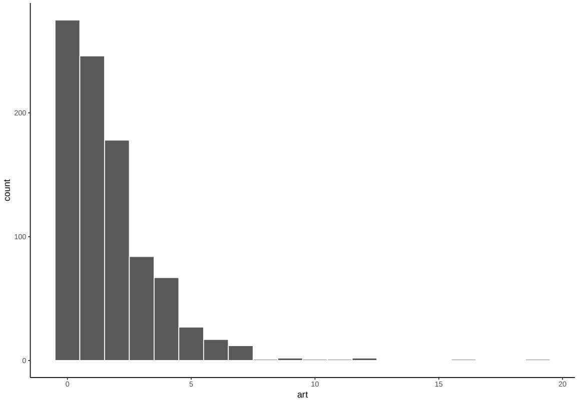
```
## [1] 5.400000 6.416667 1.188272
```

# *Overdispersion*

- If the variance of a sample is greater than would be expected according to a given theoretical model, then we say the data is *overdispersed*.
- In count data, if the variance of a sample is much greater than its mean, we say it is overdispersed.
- Using a Poisson distribution in this situation, this is an example of model mis-specification.
- It will also usually underestimate the standard errors in the Poisson model.

# *Overdispersion*

- In the `bioChemists` data set, we have counts of the number of articles published by PhD students in the last three years (`publications`):




```
var(publications)/mean(publications)
```

```
## [1] 2.191358
```

# *Overdispersion*

- This leads standard errors to be *under*estimated if we use a Poisson model:

```
M <- glm(publications ~ 1, family=poisson)
summary(M)$coefficients
```

```
##              Estimate Std. Error  z value     Pr(>|z|)
## (Intercept) 0.5264408 0.02540804 20.71945 2.312911e-95
```

# *Fixing overdispersion using a Quasi-poisson model*

- A *quasi* Poisson model allows us to correct over-dispersion

```r
M <- glm(publications ~ 1, family=quasipoisson)
summary(M)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.5264408 0.03761239 13.99647 1.791686e-40
```

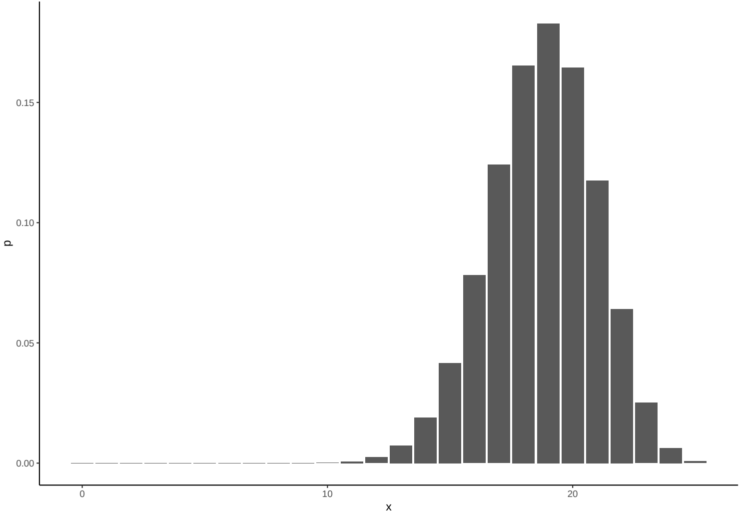
- It does so by calculating an overdispersion parameter (roughly, the ratio of the variance to the mean) and multiplying the standard error by its square root.
- In this example, the overdispersion parameter is 2.1913892 and so its square root is 1.4803341.
- Alternatively, a *negative binomial regression* is an alternative to Poisson regression that can be used with overdispersed count data.

# *Negative binomial distribution*

- A negative binomial distribution is a distribution over non-negative integers.
- To understand the negative binomial distribution, we start with the binomial distribution:
- If, for example, we have a coin whose probability of coming up heads is $\theta$, then the number of Heads in a sequence of $n$ flips will follow a binomial distribution.
- In this example, an outcome of Heads can be termed a *success*.

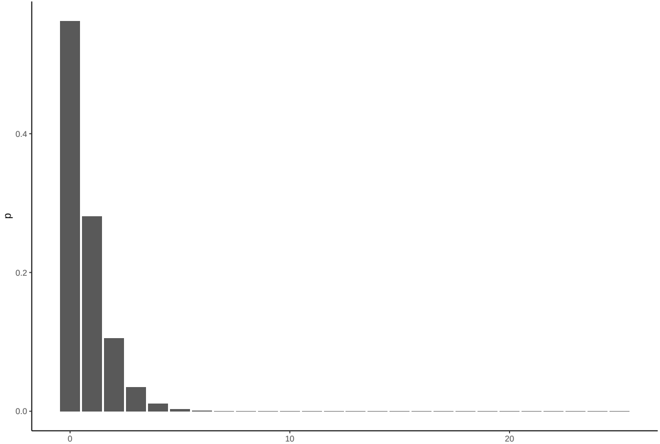
# *Negative binomial distribution*

- Here is a binomial distribution where $n = 25$ and $\theta = 0.75$.



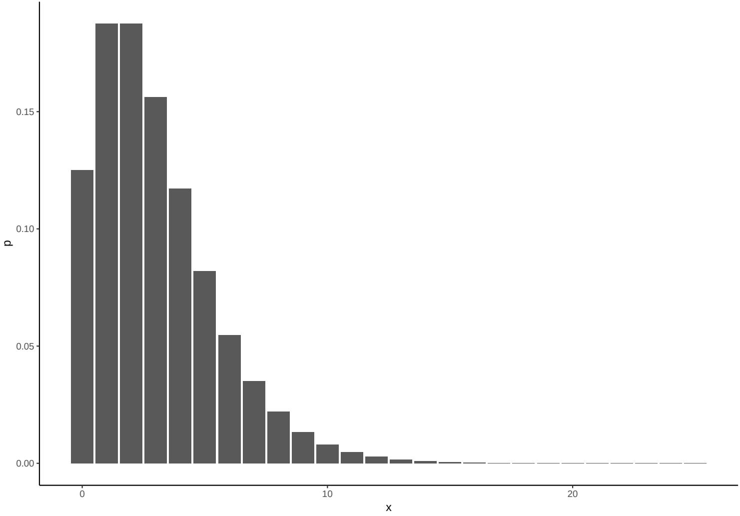
# *Negative binomial distribution*

- A *negative* binomial distribution gives the probability distribution over the number of *successes* (e.g. Heads) before $r$ *failures* (e.g. $r$ Tails).
- For example, here we have the number of Heads (*successes*) that occur before we observe $r = 2$ Tails (*failures*), when the probability of Heads is $\theta = 0.75$:



# Negative binomial distribution

- Here, we have the number of Heads (*successes*) that occur before we observe $r = 3$ Tails (*failures*), when the probability of Heads is $\theta = 0.5$:



# *Negative binomial distribution*

- The probability mass function for the negative binomial distribution is:

$$P(x = k|r, \theta) = \binom{r+k-1}{k} \theta^r (1-\theta)^k$$

or more generally

$$P(x = k|r, \theta) = \frac{\Gamma(r+k)}{\Gamma(r)k!} \theta^r (1-\theta)^k,$$

where $\Gamma()$ is a Gamma function ($\Gamma(n) = (n-1)!$).

- In R, for any $k$, $r$, and $\theta$, we can calculate $P(x = k|r, \theta)$ using `dnbinom`, e.g. $P(x = k = 2|r = 3, \theta = 0.75)$ is

```
dnbinom(2, 3, 0.75)
```

```
## [1] 0.1582031
```

# *Negative binomial distribution*

- In the negative binomial distribution, the mean is

$$\mu = \frac{\theta}{1-\theta} \times r,$$

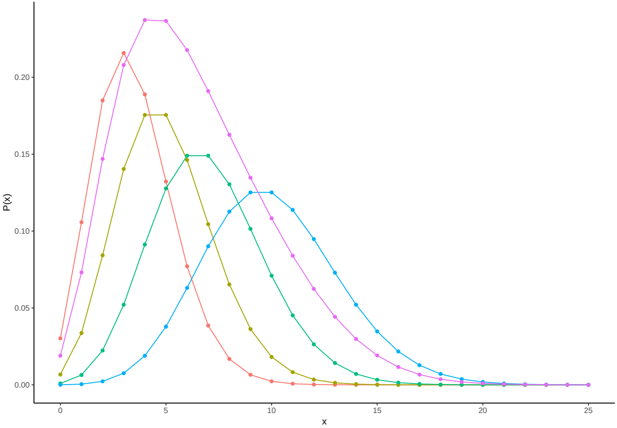
and so

$$p = \frac{r}{r+\mu},$$

and we can generally parameterize the distribution by $\mu$ and $r$.

# *Why use negative binomial distribution?*

- A negative binomial distribution is equivalent as weighted sum of Poissons.




- So it is appropriate to use when the data can be seen as arising from a mixture of Poisson distributions, each with different means.

# *Negative binomial regression*

- In negative binomial regression, we have observed counts $y_1, y_2 \dots y_n$, and some predictor variables $x_1, x_2 \dots x_n$, and we assume that

$$y_i \sim \text{NegBinomial}(\mu_i, r),$$

where $\text{NegBinomial}(\mu_i, r)$ is a negative binomial with mean $\mu_i$ and a dispersion parameter $r$, and then

$$\log(\mu_i) = \beta_0 + \beta x_i.$$