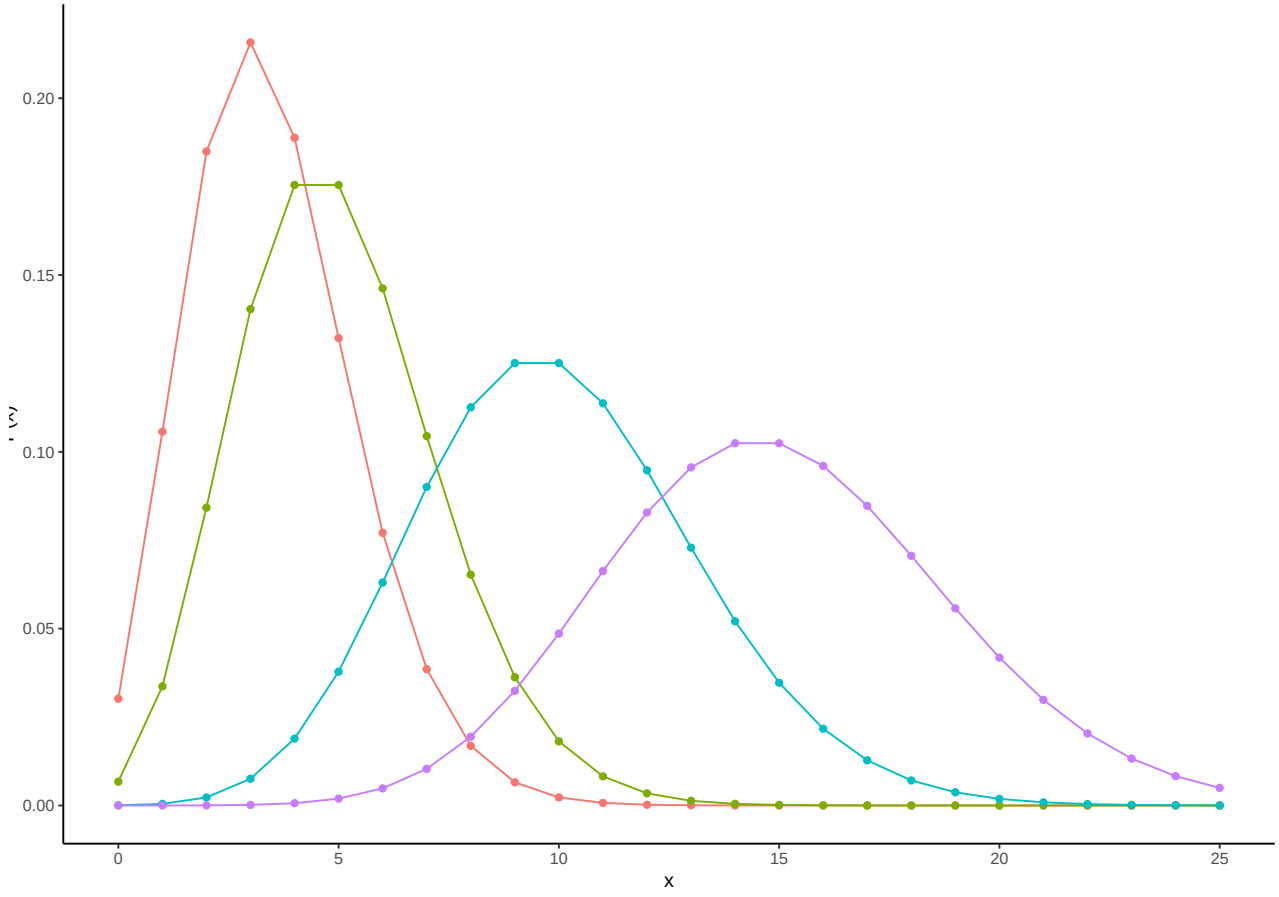
Negative binomial regression

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(100, lambda = 5)
c(mean(x), var(x), var(x)/mean(x))
```

```
## [1] 5.140000 4.869091 0.947294
```

- Example 2:

```
x <- rpois(100, lambda = 5)
c(mean(x), var(x), var(x)/mean(x))
```

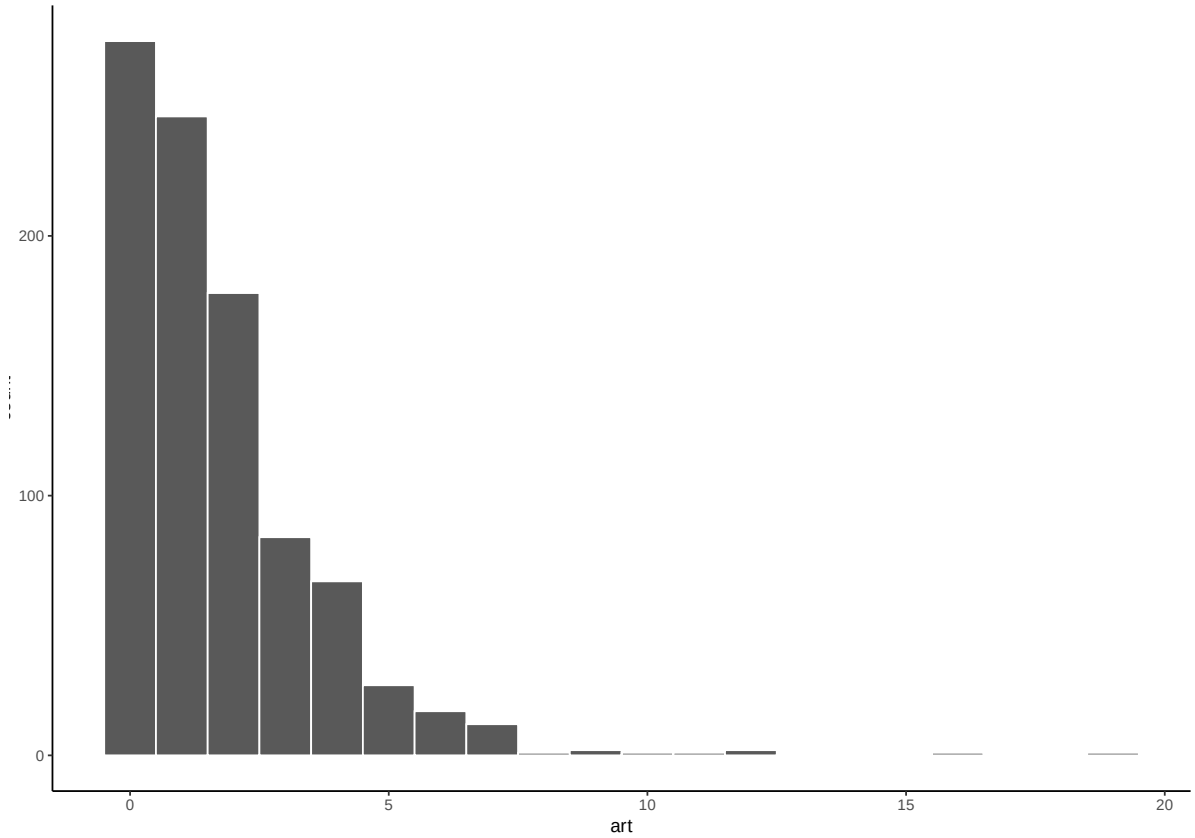
```
## [1] 4.6400000 3.8084848 0.8207941
```

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

```
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
```

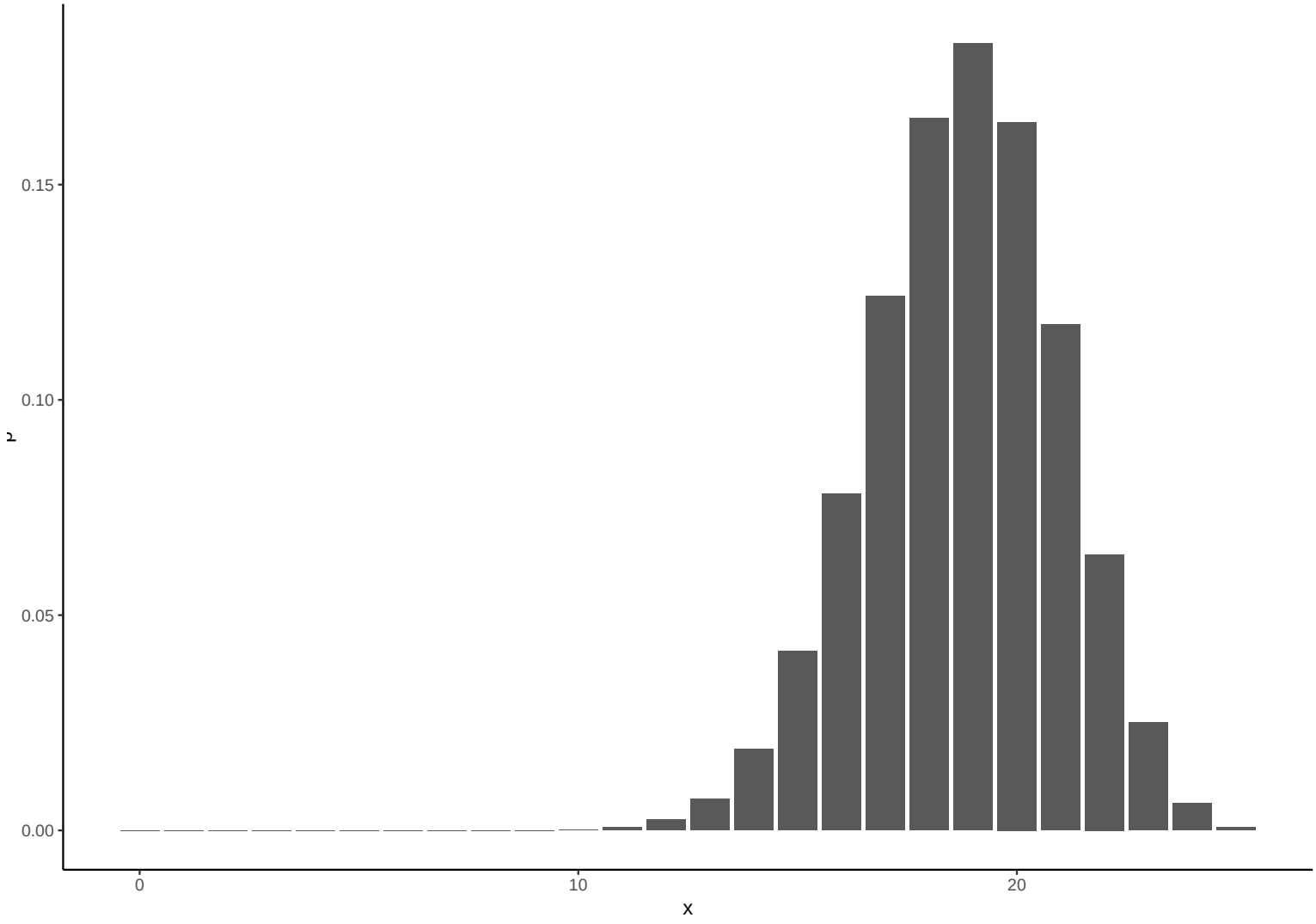
- It does so by calculating an overdispersion parameter (roughly, the ratio of the variance to the mean) and multiplying the standard error by its square root.
- In this example, the overdispersion parameter is 2.1913892 and so its square root is 1.4803341.
- Alternatively, a *negative binomial regression* 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 θ, 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*.

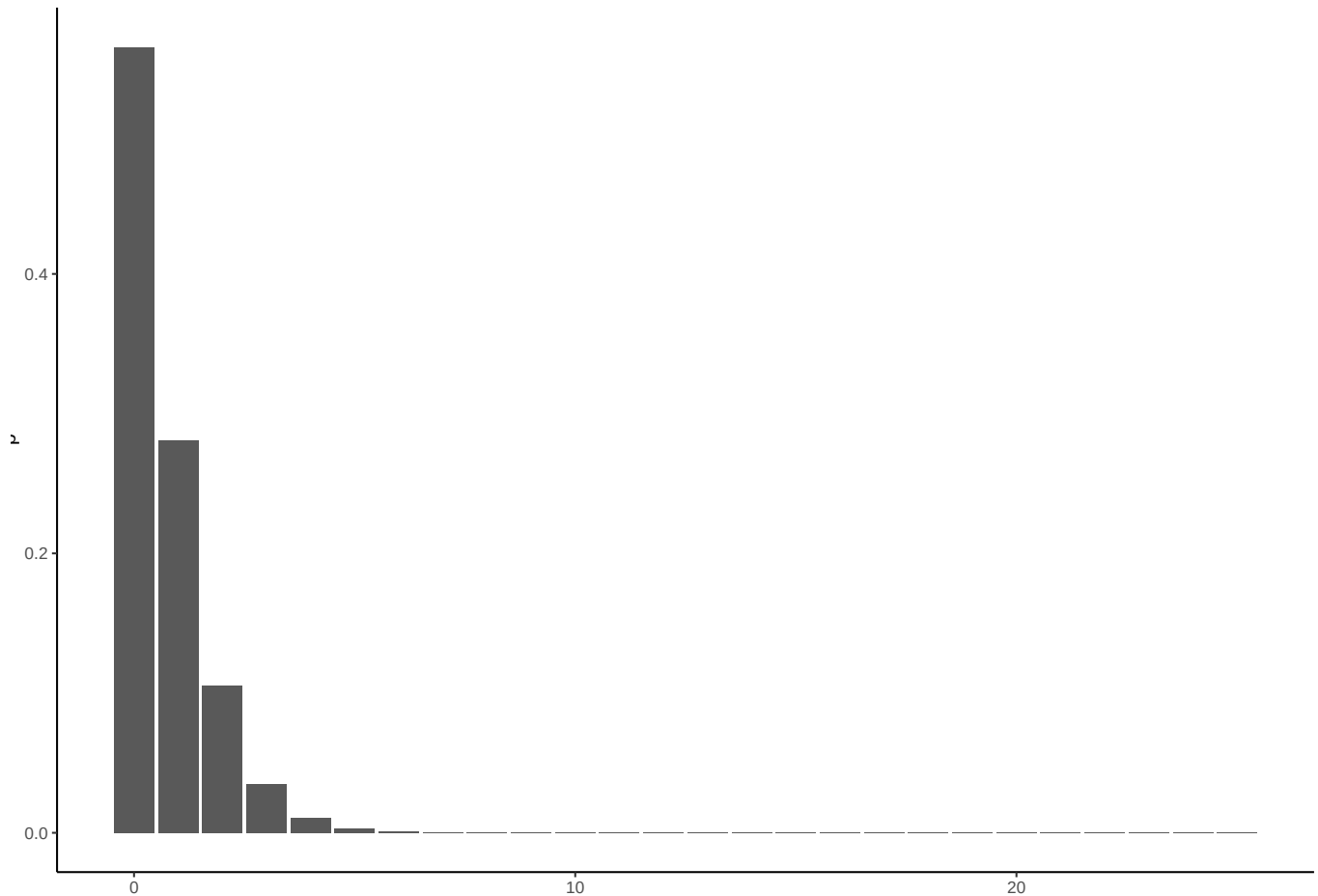
Negative binomial distribution

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



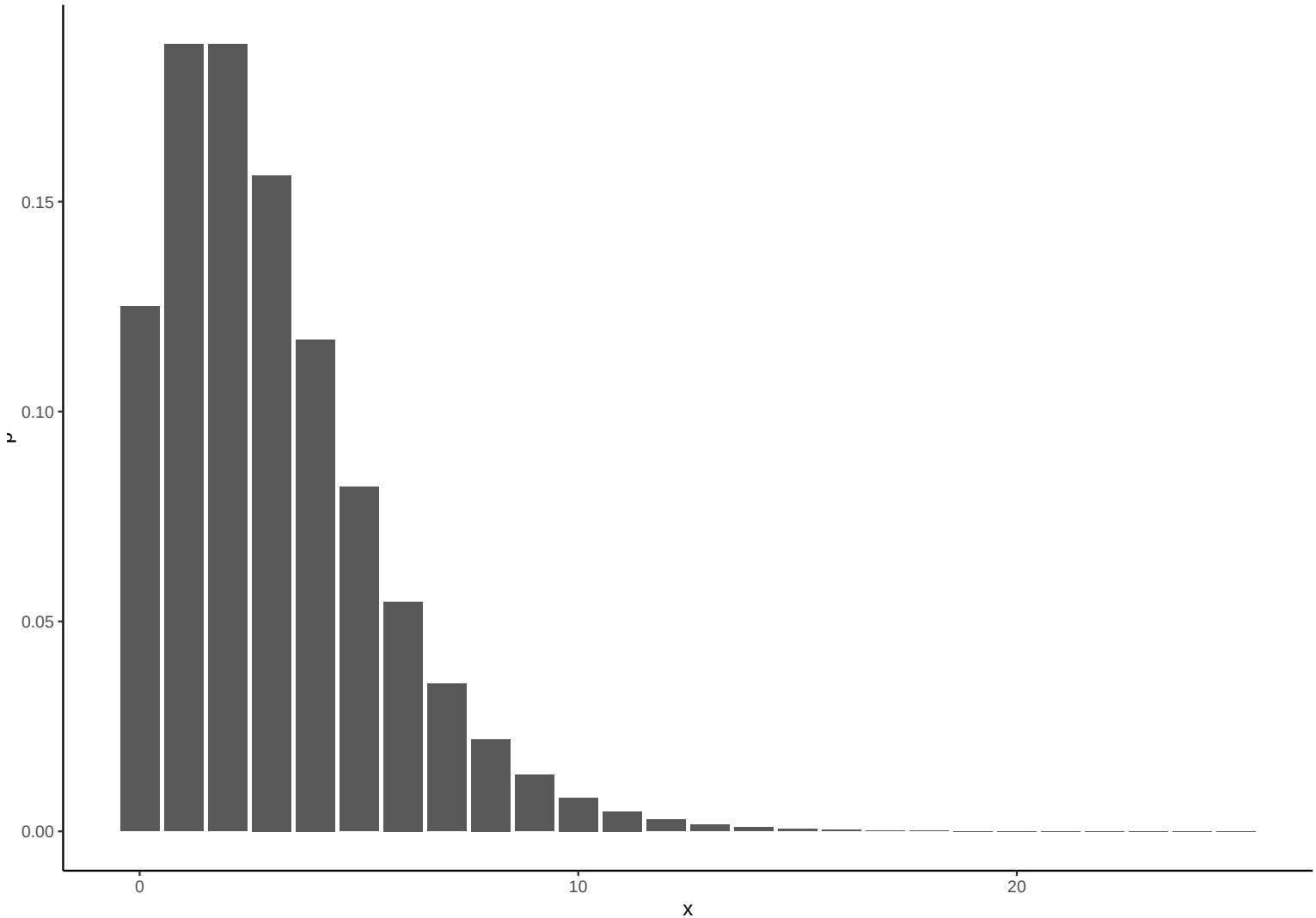
Negative binomial distribution

- 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 θ, 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

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

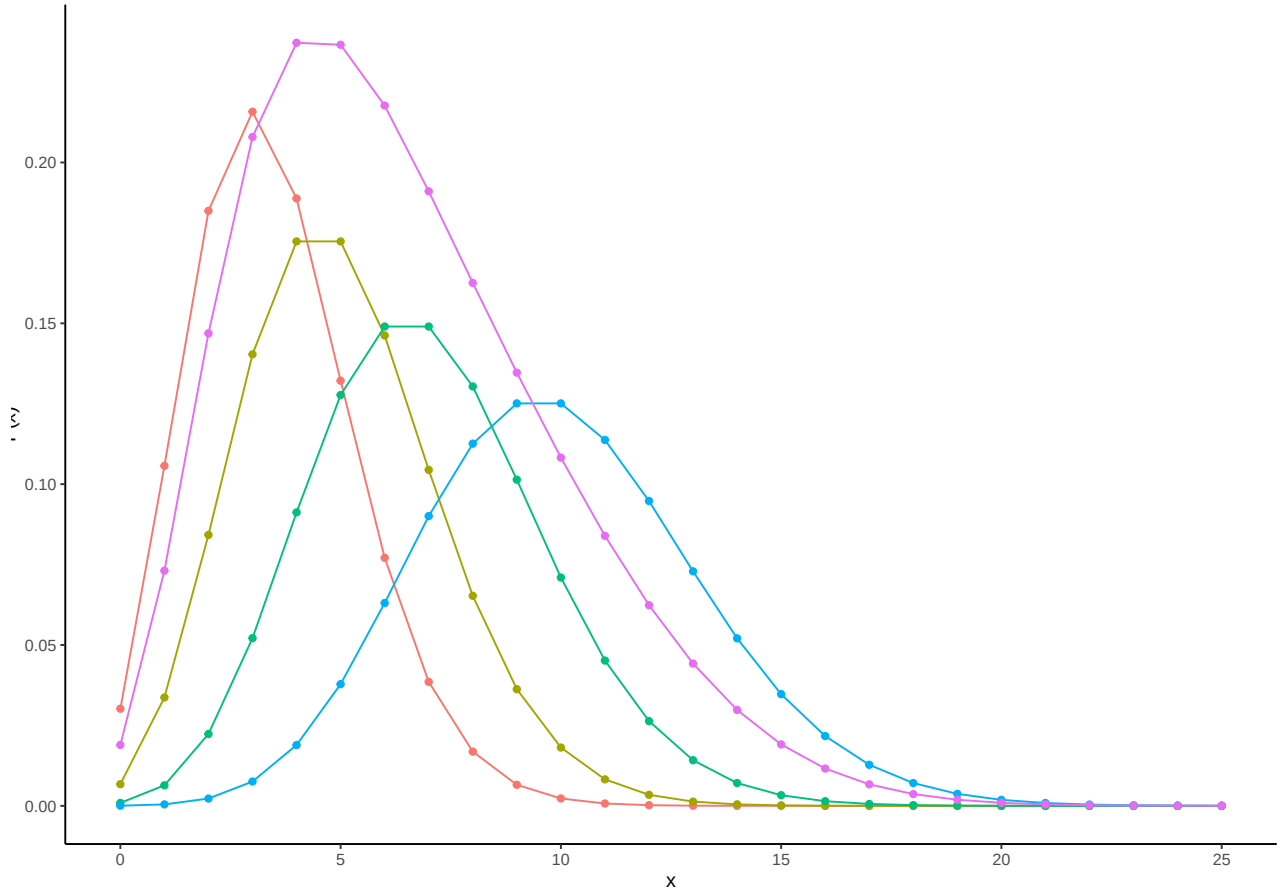
and so

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

and we can generally parameterize the distribution by μ 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 μ_i and a dispersion parameter r, and then

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

Negative binomial regression

- Because it has an overdispersion parameter, we do not have model mispecification or under estimation of standard errors:

```
M <- glm.nb(publications ~ 1)
summary(M)$coefficients
```

```
##             Estimate Std. Error  z value     Pr(>|z|)
## (Intercept) 0.5264408 0.03586252 14.67942 8.734017e-49
```

Prediction

- Here, we predict publications by gender:

```
M <- glm.nb(publications ~ gender, data=biochemists_Df)
summary(M)$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)  0.6326491 0.04716825 13.412604 5.101555e-41
## genderWomen -0.2471766 0.07203652 -3.431268 6.007661e-04
```

- Prediction works exactly like in Poisson regression.
- The predicted log of the mean of number of publictions for men is 0.63265 and for women it is $0.63265 - 0.24718$, and the predicted means are $e^0.63265 = 1.8825928$ and $e^{0.63265-0.24718} = 1.4703052$

Coefficients

- In negative binomial regression,

$$e^{\beta},$$

  where β is a coefficient, is the factor by which the mean of the outcome variable changes for a unit change in the predictor.
- For example, the coefficient for gender above is -0.24718, and males are coded by zero in the dummy variable for sex, and so $e^{-0.24718} = 0.7810001$ is the factor by which the mean publication rate changes as we go from males to females.

Model Fit with Deviance

- As in the case of logistic and Poisson regression, we estimate the parameters, e.g. α and β, using maximum likelihood estimation.
- Once we have the maximum likelihood estimate for the parameters, we can calculate *goodness of fit* using deviance.
- As before, the *deviance* of a model is defined

$$-2 \log L(\hat{\alpha}, \hat{\beta} | \mathcal{D}),$$

where $\hat{\alpha}$, $\hat{\beta}$ are the mle estimates.

Model Fit with Deviance: Model testing

- The difference in the deviance of a null model minus the deviance of the full model is

$$\Delta_D = D_0 - D_1.$$

- Under the null hypothesis, Δ_D is distributed as χ^2 with k degrees of freedom, where k is the difference in the number of parameters in the two models.