1. 30 points. White oaks: practice working with likelihoods and distributions.

   Assume that the number of white oak trees in a plot has a **Negative Binomial distribution**.

   (a) Yes, optim() ventures "outside" valid parameter ranges and then recovers. When mu = 16.43, r = -3.95, returns a lnL that is the R missing value (NaN). Then it recovers a few times and produces NaN again.
   The optim() get stuck "outside" the valid parameter ranges. The estimated parameter values are negative. We cannot trust the reported mle's and lnL.

   (b) A reasonable starting value for λ is close to the mean value of the dry woodland counts (3.77). Starting the numerical optimization at $\lambda = 4$, r = 5, gives the following:
   $\hat{\lambda} = 3.773$, $\hat{r} = 3.179$.

   (c) se = 0.6124, 95% CI: (2.572, 4.973).

   (d) When the parameters are log λ and r, the mle's are 1.328 and 3.179. The log likelihood for this fit should be the same as when you use λ as the parameter (-51.552). The only difference is the estimated parameter values for log λ and λ.

   Common mistakes: Using log(λ) and log(r) as the parameters.

   (e) The estimated log scale mean is 1.328 and its standard error is 0.1623. The 95% CI for the log mean is (1.0097, 1.6459). Backtransform the log mean, the 95% CI for the mean is (2.745, 5.186).
   Note: This CI differs from that in question 1c because we are making different assumptions about the distribution of the estimated mean. In 1c, the Wald = asymptotic interval assumes that the mean is normally distributed (at least for sufficiently large sample sizes). In 1e, we assume that the log mean is normally distributed, i.e. that the mean is lognormally distributed.

   (f) C = 2.6528, p = 0.1034.
   Using glm.nb: The null hypothesis model is an intercept only, oak $\sim$ 1; the alternative hypothesis model is different λ for the two woodland types, oak $\sim$ woodland. The log likelhoods for these two models are null = -104.9367, alternative = -103.6103. The test statistics is $C = -2 * (-104.9367 - -103.6103) = 2.6528$. This has an asymptotic Chi-square distribution with 1 degree of freedom (calculated as 2 parameters under the alternative - 1 parameter under the null). The p-value is 0.1034.

   -2 points if you use lnlNB to calculate log-likelihood separately for 2 types, add them up, and then calculate log-likelihood for all data, the difference is not what this question asks for. This is a more general hypothesis than the question asked about (assume that the two woodlands have the same overdispersion parameter).

   Common mistakes: summary(null) gives 2 × log-likelihood. The result times 2 again. The test statistics should be $C = -(-209.873 - -207.221) = 2.652$. Or can use logLik() function to get the log likelihood.

   (g) Estimate = -0.4547, 95% CI = (-0.9927, 0.0833)
   Details: This is much easier done with glm.nb() since the default link function for a Negative Binomial distribution is log. The estimated woodland coefficient in the model, oak $\sim$ woodland, is 0.4547. This is the difference as moist - dry, but we need dry - moist, which = -0.4547. The standard error of either = 0.2745, so the asymptotic CI for log ratio (dry/moist) = (-0.4547 - 1.96*0.2745, -0.4547 + 1.96*0.2745) = (-0.9927, 0.0833).
   Common mistakes: report the coefficient 0.4547 directly.

(h) Estimate: 0.6346, 95% CI: (0.3706, 1.0869)
Details: exponentiate the log scale estimate and confidence endpoints.

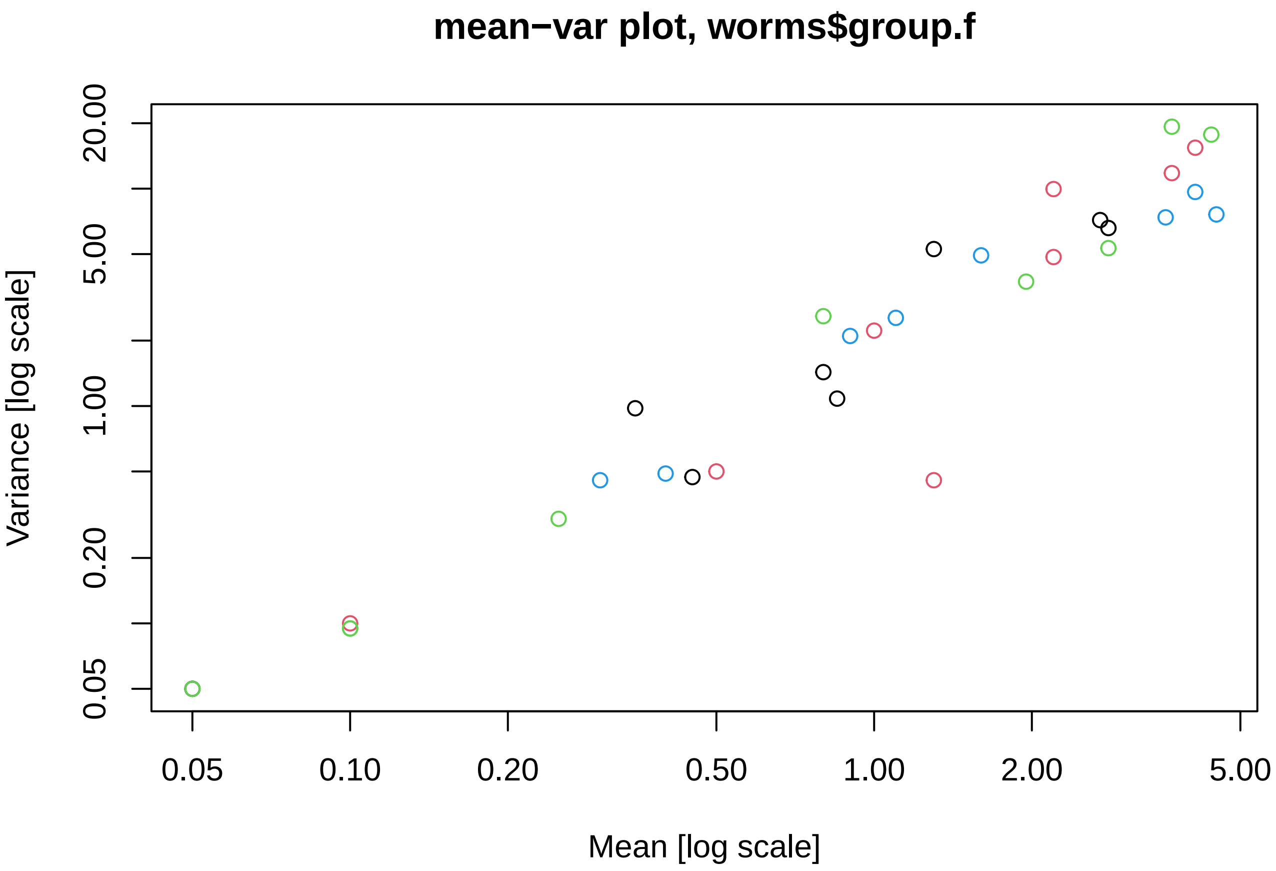
Notes: A function that calculates the Negative Binomial log likelihood of (mean, size) given data is:

```
NBlnl <- function(beta, data) {
  mu <- beta[1]
  r <- beta[2]
  sum(dnbinom(data, size = 1/r, mu=mu, log=T))
  }
```

The 1/r is because dnbinom() expects a "size" parameter for which Var $Y = \mu + size\mu^2$

2. 20 points. ecotoxicology data analysis.

(a) For Poisson distribution, the variance equals to the mean. In this problem, the variance is larger than the mean for most groups with mean ≥ 1 and all groups with mean ≥ 2. So the Negative Binomial distribution is more appropriate.



(b) The sum of the AIC for all species in the Negative Binomial model (1487.723) is smaller than that in the Poisson model (1673.949). So the Negative Binomial distribution is more appropriate.

Common mistakes: compare AIC for each species.

(c) The data does not provide evidence of a time x treatment interaction because the p-value is 0.597.

(d) The time and treatment effects are averages over the omitted variable. So before is the average of before/control and before/impact and after is the average of after/control and after/impact.

I can see how the before vs after comparison described above makes sense as a “how much did things change”, averaging over the two types of site. So yes, the time effect answers the biologically relevant question.

	control	impact
before	-	-
after	-	-

(e) Examining the individual univariate p-values, Species 8(Ali) has a p-value (0.002) smaller than 0.05.

(f) Similarly, control is the average of before/control and after/control. Impact is the average of before/impact and after/impact. Reordering the subtraction, you see that the impact - control comparison is the average of (before/impact - before/control) and (after/impact – after/control). I don’t think this makes much sense, especially as a quantification of impact, because you don’t expect much difference (before/impact - before/control), because that’s before the impact occurred. That gets averaged with after/impact – after/control, where you could expect a difference. The “treatment” effect is likely to be less than the true effect of the treatment.

$$\begin{aligned}
&(before/impact + after/impact) - (before/control + after/control)\\
&= (before/impact - before/control) + (after/impact\text{–}after/control)\\
&= 0 + (after/impact\text{–}after/control)
\end{aligned} \tag{1}$$

Comments: no need to report all digits of the outputs.