The mean variance relationship

Analysis of variance

- Variance: (Fisher, 1918)
- ANOVA: (Fisher, 1921)

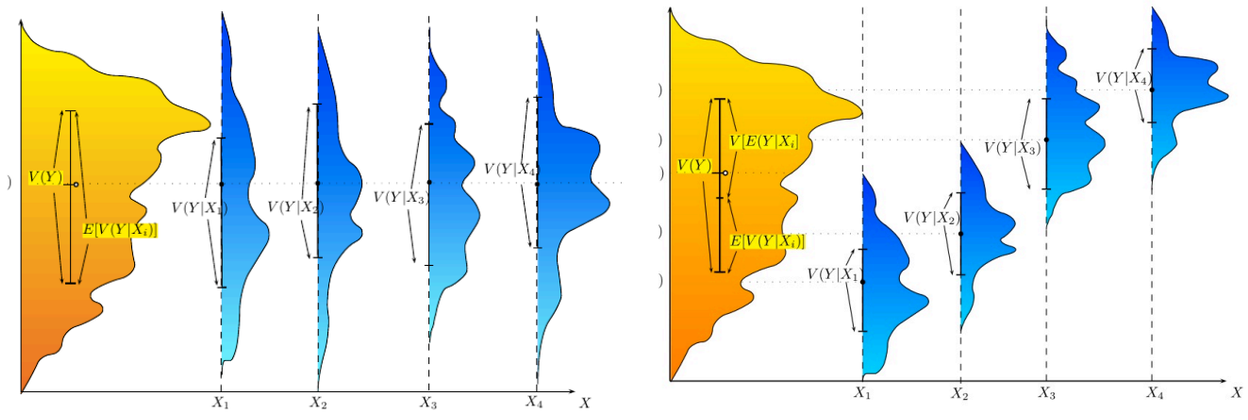


Figure 1: ANOVA : Fair fit

Same mean **Different means**

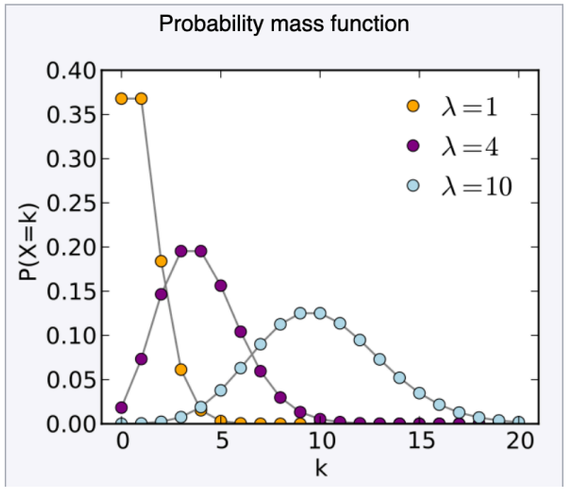
- Likelihood: (Fisher, 1921)
- Maximum likelihood: (Fisher, 1922)

$$P(X \mid \theta)$$

https://en.wikipedia.org/wiki/Analysis_of_variance

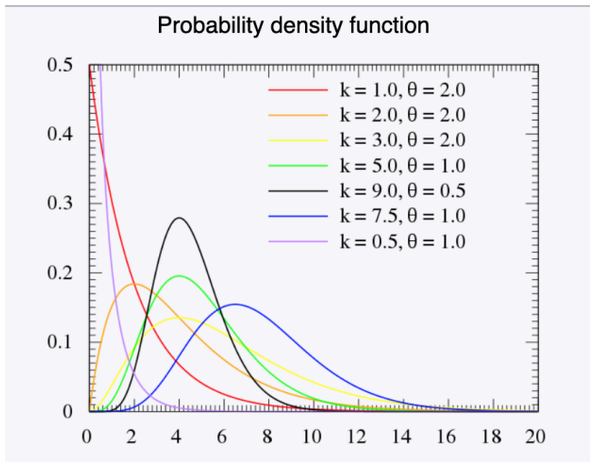
Mean vs variance

Poisson Distribution



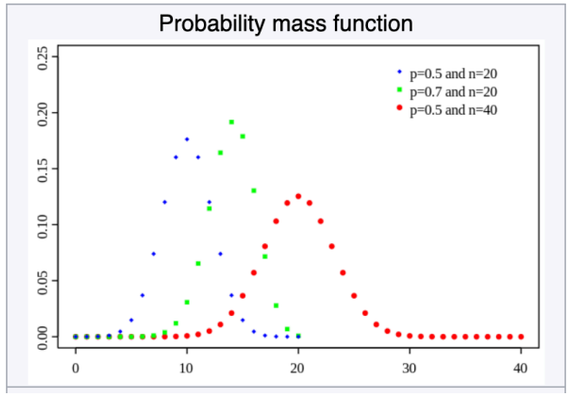
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ

Gamma



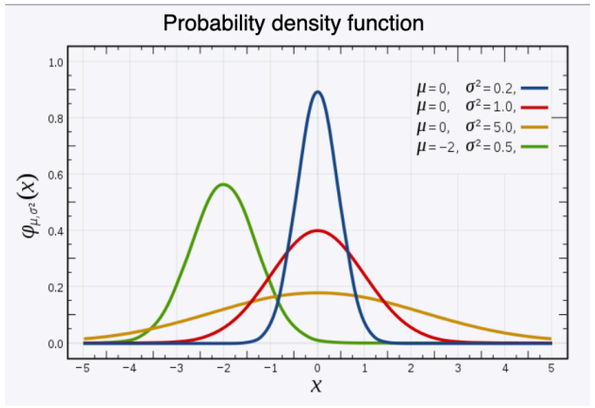
Mean	$\mathrm{E}[X] = k\theta$	$\mathrm{E}[X] = \frac{\alpha}{\beta}$
Median	No simple closed form	No simple closed form
Mode	$(k-1)\theta$ for $k \geq 1$	$\frac{\alpha - 1}{\beta}$ for $\alpha \geq 1$
Variance	$\mathrm{Var}(X) = k\theta^2$	$\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$

Binomial distribution



Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$

Normal distribution



Mean	μ
Median	μ
Mode	μ
Variance	σ^2

Bartlett 1936

Square root transform for analysis of variance

THE SQUARE ROOT TRANSFORMATION IN ANALYSIS OF VARIANCE.

By M. S. BARTLETT.

1. *Introduction.*

THE analysis of variance has by now been used in such a vast number of problems that, in spite of its wide range of applicability, it would be surprising if examples had not occurred where its direct use was of doubtful value. Certain of these cases can, however, sometimes be more legitimately solved by a suitable transformation of our variate. The common occurrence of a type of experiment for which the square-root transformation has often been found useful, justifies a closer inspection of this particular transformation. Some illustrations of the type of experiment referred to are included in the paper.

2. *Theoretical Discussion.*

Just as, in order to stabilize the variance, the logarithmic transformation suggests itself when the standard error of a variate is proportional to its mean value, so when the variance is proportional to the mean, the square root may be considered. For the mean m large, and actually equal to the variance, we have

$$\sigma^2(\sqrt{x}) = \sigma^2(x)\left(\frac{\partial\sqrt{x}}{\partial x}\right)^2 = \tfrac{1}{4} \quad . \quad . \quad . \quad . \quad (1)$$

approximately; or more generally $\frac{1}{4}\lambda$, if $\sigma^2(x) = \lambda m$.

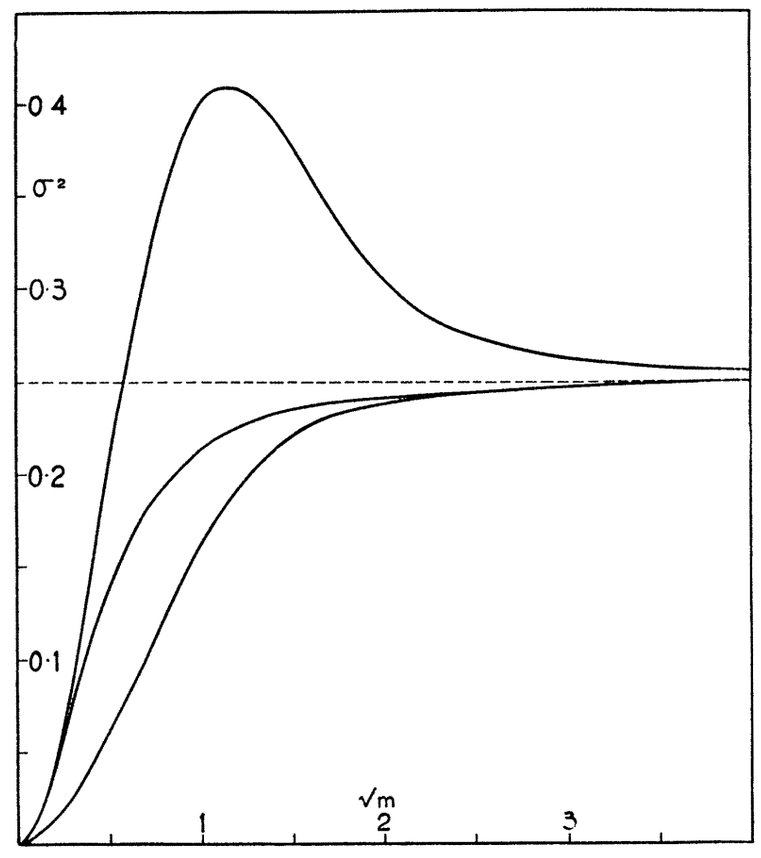


FIG. 1.—The three curves (the largest variance first) are for
1. Poisson, $\sqrt{x}$,
2. Type III, $\sqrt{x}$,
3. Poisson, $\sqrt{x+\frac{1}{2}}$.

Bartlett 1936

Square root transform for analysis of variance

- When μ > 10 -> sqrt(x)
- When 10 > μ > 2 -> sqrt(x+1/2)
- When μ < 2 -> Discontinuous nature of Poisson makes it hard to say anything.

> Thus above a mean of 10, say, $\sqrt{x}$ may be considered, from 10 to about 2 or 3, $\sqrt{x+\frac{1}{2}}$ is preferable. Below a mean of about 2 or 3 the discontinuous nature of the Poisson distribution has, of course, become so violent that any variate is of little use quite apart from questions of variance, unless a large number of replications is available.

Anscombe 1948

Transformations for Poisson, Binomial, NB

THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA

By F. J. ANSCOMBE, *Rothamsted Experimental Station*

1. Introduction

Bartlett (1936) showed that if **r** is a Poisson variable with mean m and **y** is a random variable whose values y are derived by the transformation

$$y = \sqrt{r} \tag{1·1}$$

from the values r of **r**, then **y** is distributed rather more nearly normally than **r** with variance approximately $\frac{1}{4}$ if m is large, and the technique of analysis of variance may be applied to **y**.* He also showed that

$$y = \sqrt{(r+\tfrac{1}{2})} \tag{1·2}$$

is a better transformation, if slightly less convenient to use, as **y** then has more nearly a constant variance of $\frac{1}{4}$, even when m is not large. Similar transformations were proposed for a binomial variable, and Beall (1942) gave the transformation analogous to (1·1) appropriate to a negative binomial variable.

I begin by considering the transformation

$$y = \sqrt{(r+c)} \tag{1·3}$$

of a Poisson variable **r**, and show that for large m **y** has a most nearly constant variance (namely, $\frac{1}{4}$) when $c = \frac{3}{8}$; a result due to A. H. L. Johnson.

The similar transformation for a binomial variable **r**, with mean m and total number n, is

$$y = \sin^{-1}\sqrt{\left(\frac{r+c}{n+2c}\right)}. \tag{1·4}$$

The optimum value of c is $\frac{3}{8}$ if m and $n-m$ are large. The variance is approximately $\frac{1}{4}(n+\frac{1}{2})^{-1}$.

For a negative binomial variable **r**, with mean m and exponent k, the latter being constant and known, the corresponding transformation is

$$y = \sinh^{-1}\sqrt{\left(\frac{r+c}{k-2c}\right)}. \tag{1·5}$$

The optimum value of c is roughly $\frac{3}{8}$ if m is large and $k > 2$, and the variance is approximately $\frac{1}{4}\psi'(k)$, where $\psi'(t)$ denotes the second derivative of $\ln\Gamma(t)$ with respect to t. A simpler transformation, known to have an optimum property (i.e. to be the best of that degree of complexity) for m large and $k \geqslant 1$, is

$$y = \ln(r+\tfrac{1}{2}k); \tag{1·6}$$

the variance is approximately $\psi'(k)$. This is equivalent to setting $c = \frac{1}{2}k$ in (1·5). If k is large, $\psi'(k) = 1/(k-\frac{1}{2})$ approximately.

Nelder and Wedderburn 1972

GLMs: Non-normal likelihoods

Two extreme models are conceivable for any set of data, the *minimal model* which contains the smallest set of terms that the problem allows, and the *complete model* in which all the Ys are different and match the data completely so that $\hat{\mu} = z$. An extreme case of the minimal model is the null model, which is equivalent to fitting the grand mean only and effectively consigns all the variation in the data to the random component of the model, while the complete model fits exactly and so consigns all the variation in the data to the systematic part. The model-fitting process with an ordered model thus consists of proceeding a suitable distance from the minimal model towards the complete model. At each stage we trade increasing goodness-of-fit to the current set of data against increasing complexity of the model. The fitting of the parameters at each stage is done by maximizing the likelihood for the current model and the matching of the model to the data will be measured quantitatively by the quantity $-2L_{\max}$ which we propose to call the *deviance*. For the four special distributions the deviance takes the form:

Normal	$\sum(z-\hat{\mu})^2/\sigma^2,$
Poisson	$2\{\sum z \ln(z/\hat{\mu}) - \sum(z-\hat{\mu})\},$
Binomial	$2[\sum z \ln(z/\hat{\mu}) + \sum(n-z)\ln\{(n-z)/(n-\hat{\mu})\}],$
Gamma	$2p\{-\sum \ln(z/\hat{\mu}) + \sum(z-\hat{\mu})/\hat{\mu}\}.$

Nelder and Wedderburn 1972

GLMs: Non-normal likelihoods

Two extreme models are conceivable for any set of data, the *minimal model* which contains the smallest set of terms that the problem allows, and the *complete model* in which all the Ys are different and match the data completely so that $\hat{\mu} = z$. An extreme case of the minimal model is the null model, which is equivalent to fitting the grand mean only and effectively consigns all the variation in the data to the random component of the model, while the complete model fits exactly and so consigns all the variation in the data to the systematic part. The model-fitting process with an ordered model thus consists of proceeding a suitable distance from the minimal model towards the complete model. At each stage we trade increasing goodness-of-fit to the current set of data against increasing complexity of the model. The fitting of the parameters at each stage is done by maximizing the likelihood for the current model and the matching of the model to the data will be measured quantitatively by the quantity $-2L_{\max}$ which we propose to call the *deviance*. For the four special distributions the deviance takes the form:

Normal	$\sum(z-\hat{\mu})^2/\sigma^2$,
Poisson	$2\{\sum z \ln(z/\hat{\mu}) - \sum(z-\hat{\mu})\}$,
Binomial	$2[\sum z \ln(z/\hat{\mu}) + \sum(n-z)\ln\{(n-z)/(n-\hat{\mu})\}]$,
Gamma	$2p\{-\sum \ln(z/\hat{\mu}) + \sum(z-\hat{\mu})/\hat{\mu}\}$.

1.3. *The Generalized Linear Model*

We now combine the systematic and random components in our model to produce the generalized linear model. This is characterized by

(i) A dependent variable z whose distribution with parameter θ is one of the class in Section 1.1.

(ii) A set of independent variables $x_1, \ldots, x_m$ and predicted $Y = \sum \beta_i x_i$ as in Section 1.2.

(iii) A linking function $\theta = f(Y)$ connecting the parameter θ of the distribution of z with the Y's of the linear model.

Generative models

$$\mu = X\beta$$

$$Y \sim \text{Poisson}(s \cdot \exp(\mu))$$

- How to know this does the right thing?

Assessing generative models

- Held out log likelihood

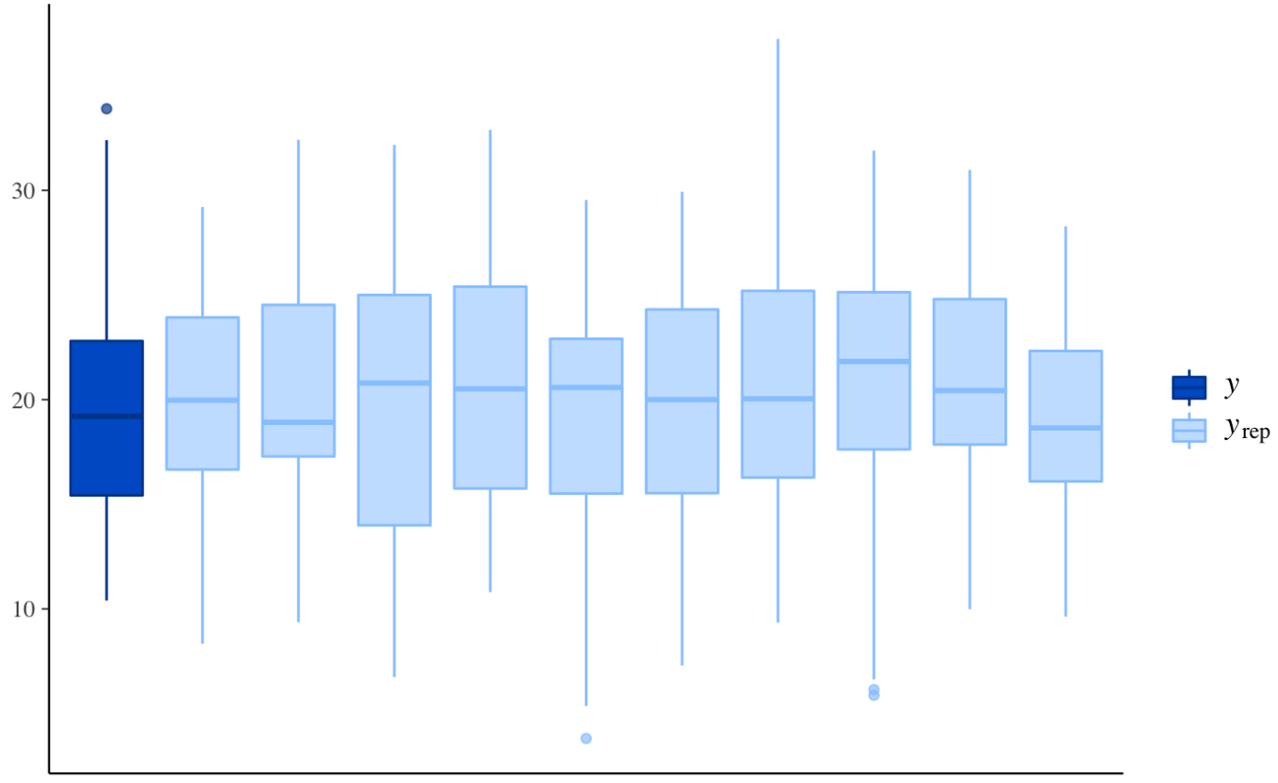
- Posterior predictive checks

Graphical Posterior predictive checks

- Fit model
- Sample data
- Does summary statistics look similar between real and sampled data?

Simple regression example

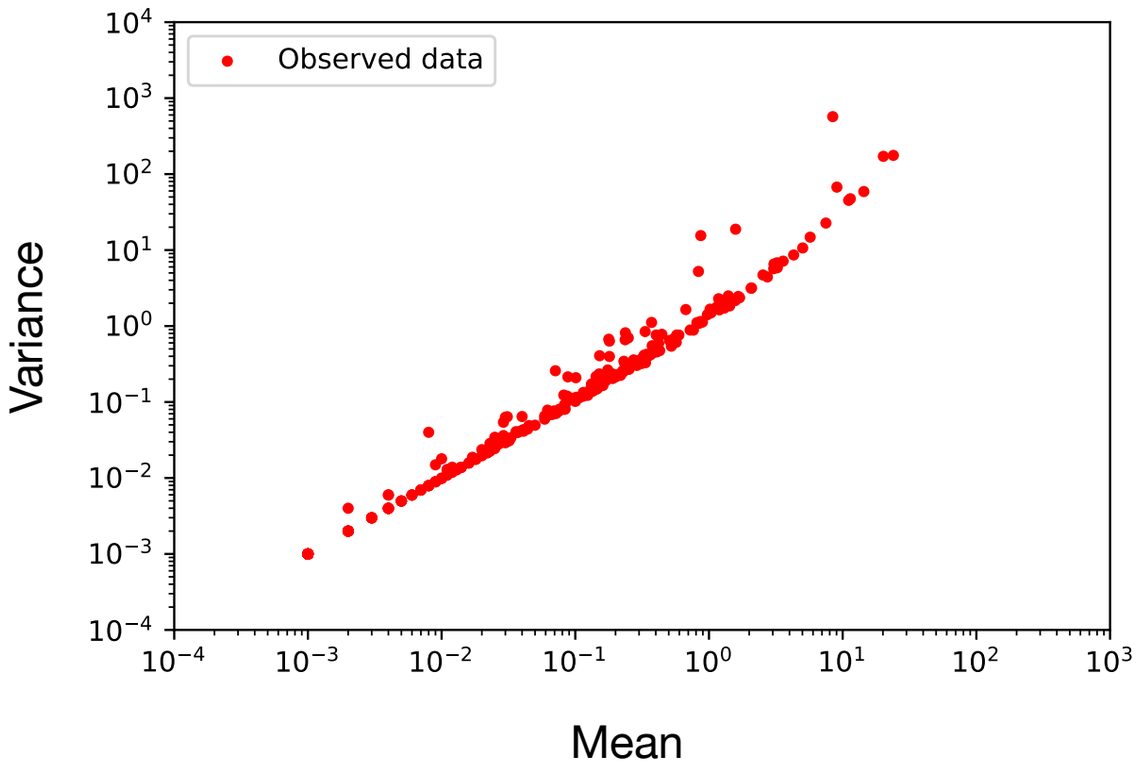
```
fit <- stan_glmer(
  mpg ~ wt + am + (1|cyl),
  data = mtcars,
  iter = 400, # iter and chains small just to keep example quick
  chains = 2,
  refresh = 0
)
```



http://mc-stan.org/rstanarm/reference/pp_check.stanreg.html

Using the mean-variance relation

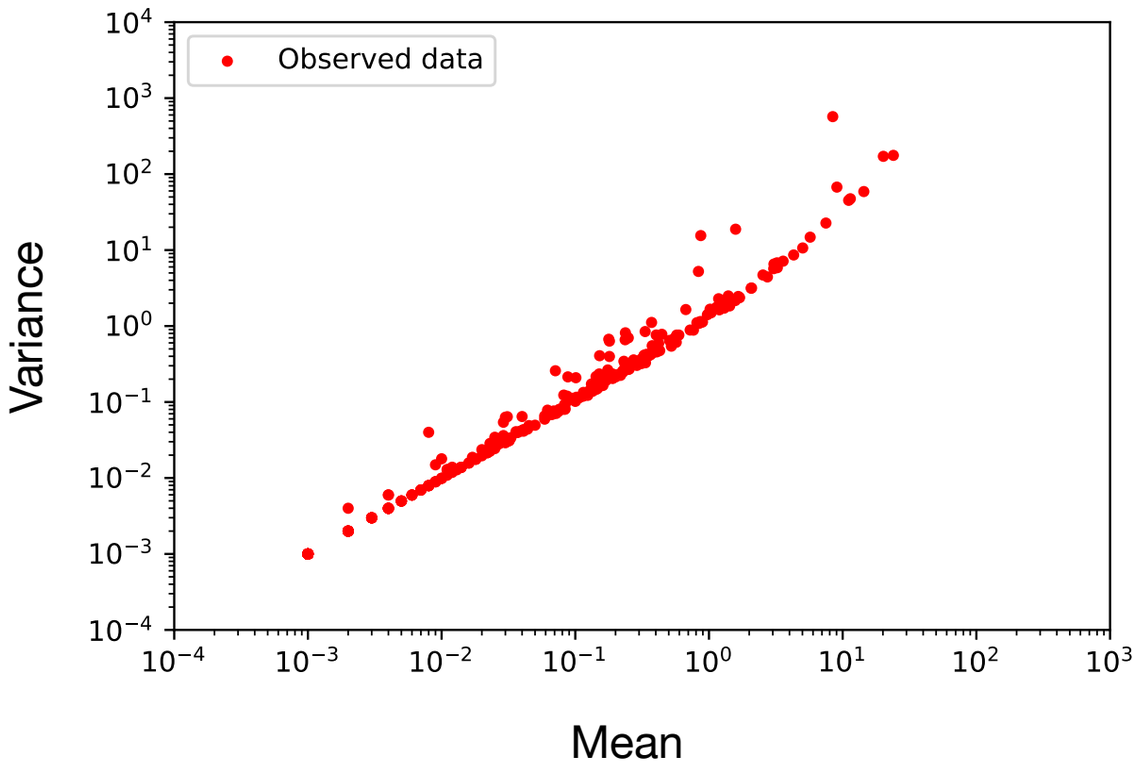
500 genes, 1 000 cells



Data from *Pijuan-Sala et al* 2019

Using the mean-variance relation

500 genes, 1 000 cells



500 genes, 10 000 cells

Fit scVI model

Apply to 1 000 held out cells

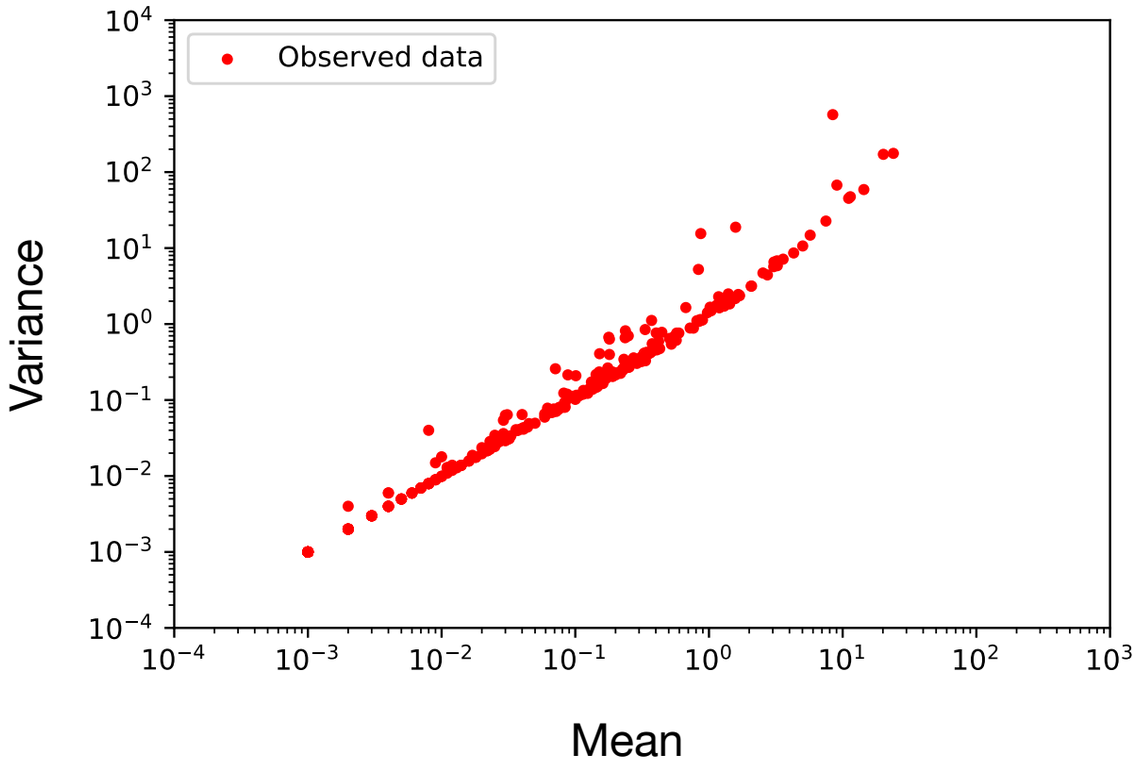
Sample count matrix for those

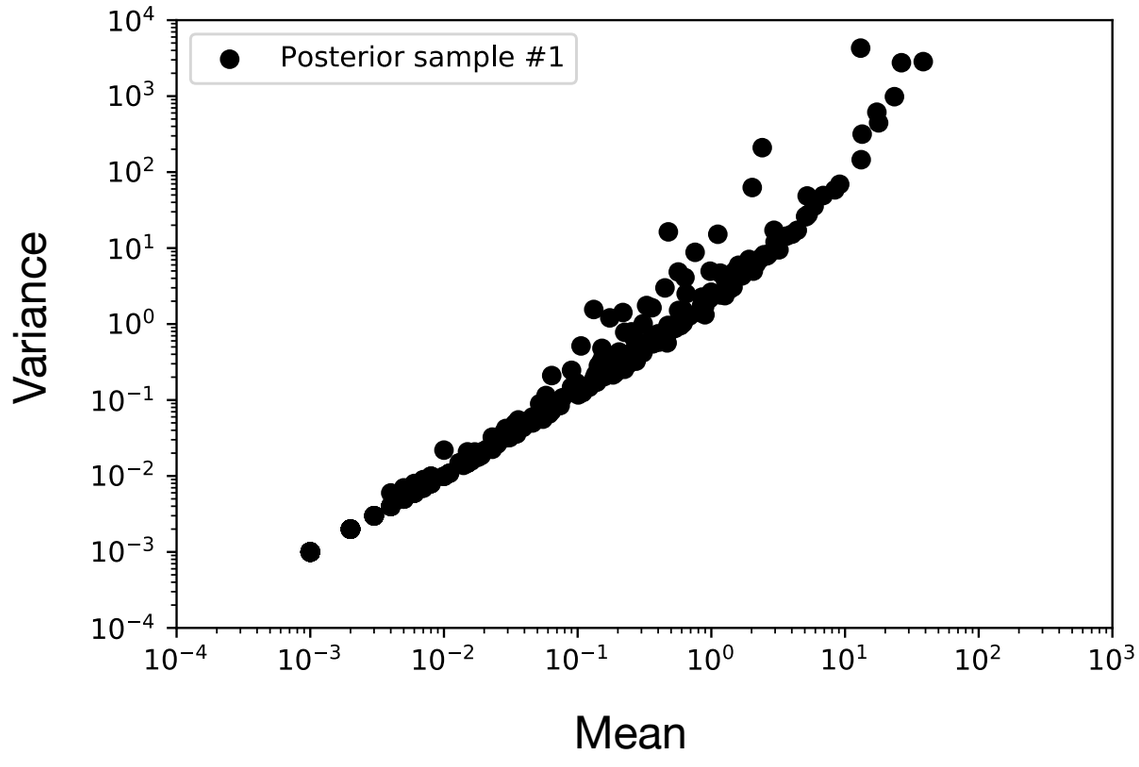
Calculate mean and variance from sampled matrix

Plot

Data from *Pijuan-Sala et al* 2019

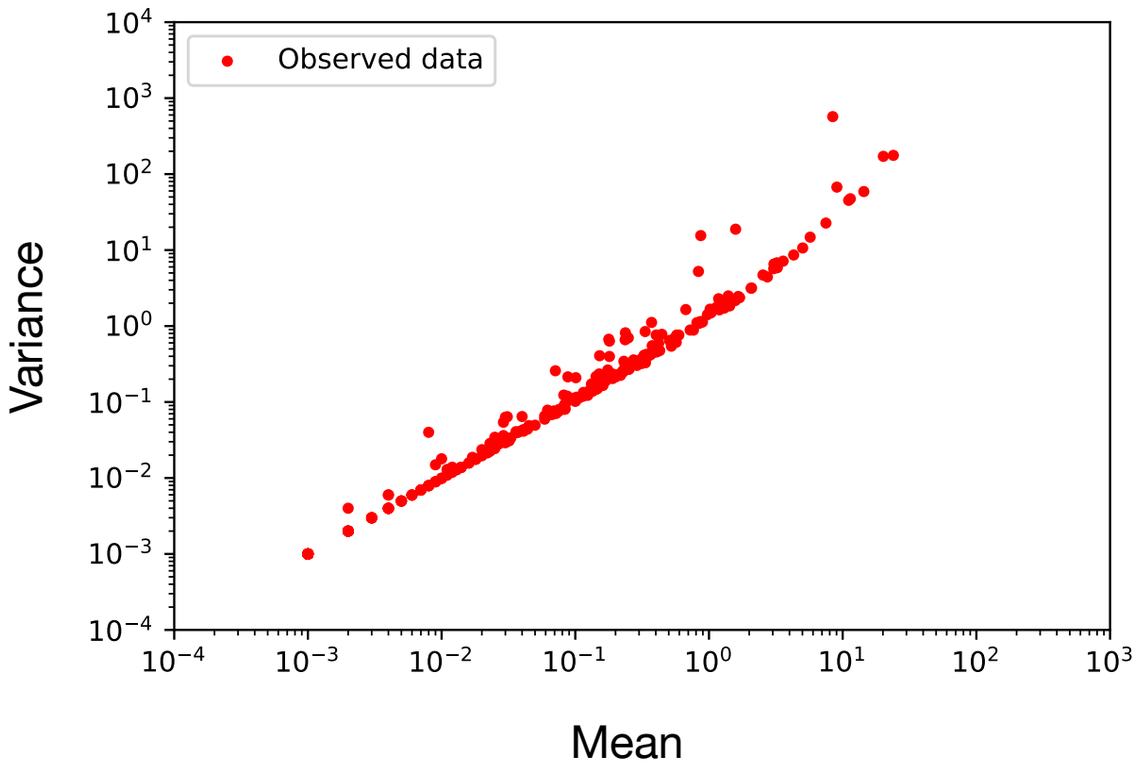
Using the mean-variance relation

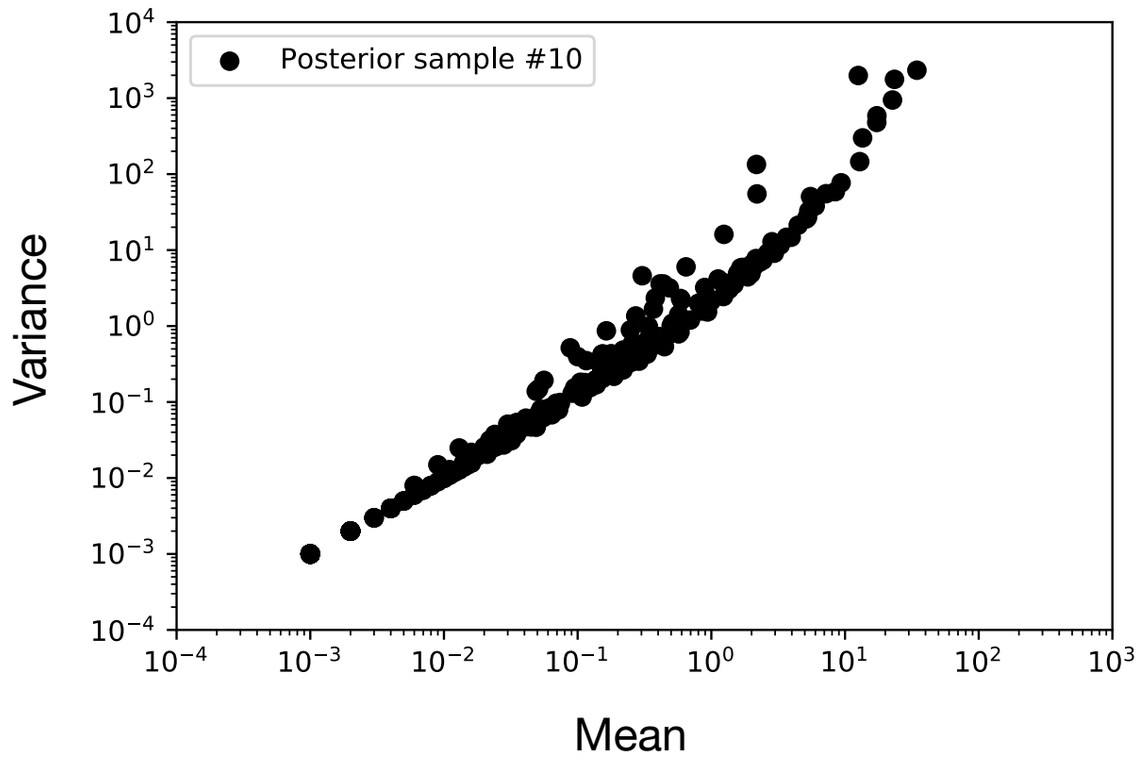




Data from *Pijuan-Sala et al* 2019

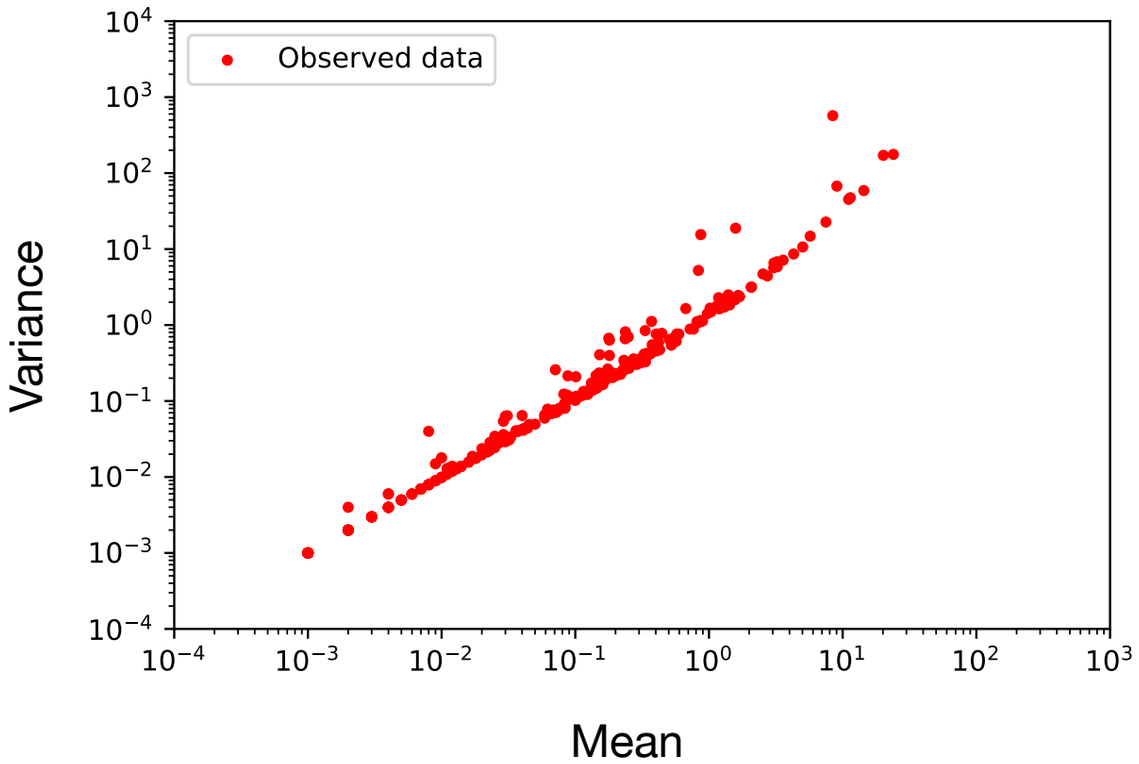
Using the mean-variance relation

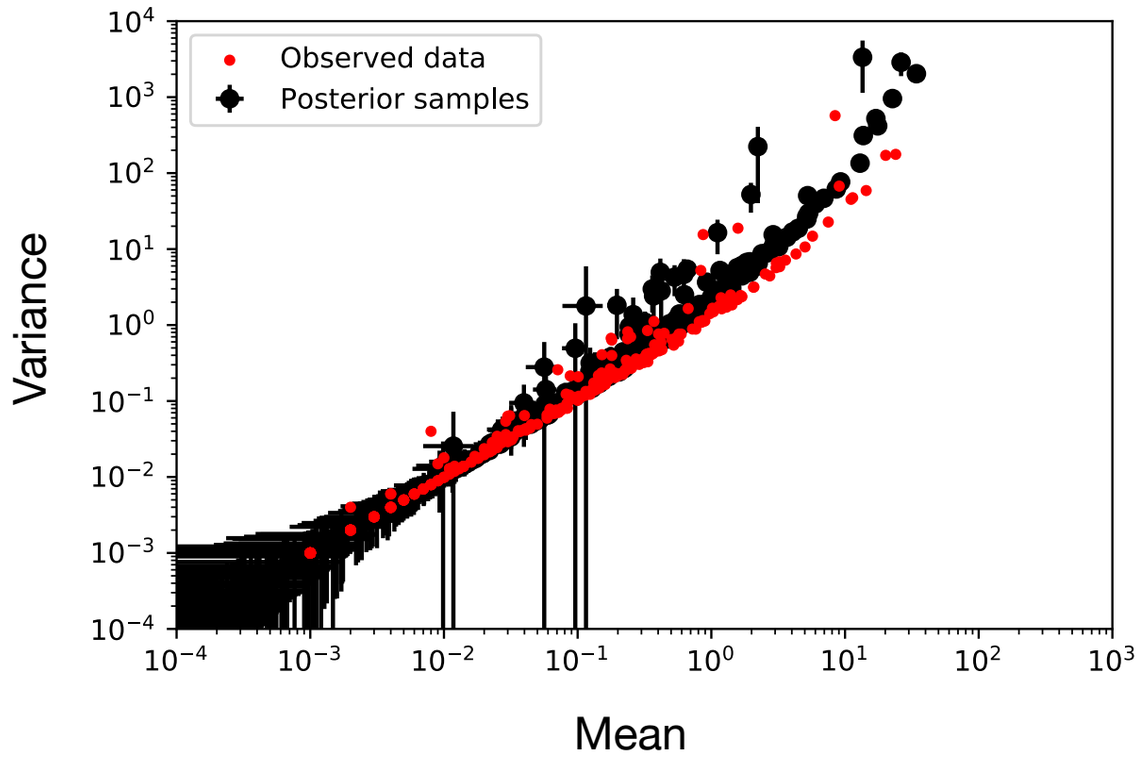




Data from *Pijuan-Sala et al* 2019

Using the mean-variance relation





Data from *Pijuan-Sala et al* 2019

References

- *Fisher* 1918 **The Correlation between Relatives on the Supposition of Mendelian Inheritance**
- *Fisher* 1921 **On the “probable error” of a coefficient of correlation deduced from a small sample**
- *Fisher* 1922 **On the mathematical foundations of theoretical statistics**
- *Bartlett* 1936 **The Square Root Transformation in Analysis of Variance**
- *Anscombe* 1948 **The transformation of Poisson, Binomial and Negative-Binomial data**
- *Nelder & Wedderburn* 1972 **Generalized Linear Models**
- *Stigler* 2008 **The Epic Story of Maximum Likelihood**
- *St-Pierre, Shikon, & Schneider* 2017 **Count data in biology - Data transformation or model reformation?**
- *Lopez et al* 2018 **Deep generative modeling for single-cell transcriptomics**
- *Pijuan-Sala et al* 2019 **A single-cell molecular map of mouse gastrulation and early organogenesis**