03 - Density Estimation

Data Mining

SYS 6018 | Fall 2019

03-density.pdf

Contents

1 Density Estimation Intro **2**

1.1 Required R Packages . . . 2
1.2 Distributions . . . 2
1.3 Example: Default Classification . . . 4
1.4 Example: Association Analysis . . . 4
1.5 Example: Disease Outbreak Detection . . . 4
1.6 Estimation . . . 5

2 Parametric Density Estimation **6**

2.1 Method of Moments Estimation (MOM) . . . 6
2.2 Maximum Likelihood Estimation (MLE) . . . 7
2.3 Bayesian Estimation . . . 14

1 Density Estimation Intro

1.1 Required R Packages

We will be using the R packages of:

- `tidyverse` for data manipulation and visualization
- `fitdistrplus` for parametric estimation

```
library(tidyverse)     # install.packages("tidyverse")
library(fitdistrplus)  # install.packages("fitdistrplus")
```

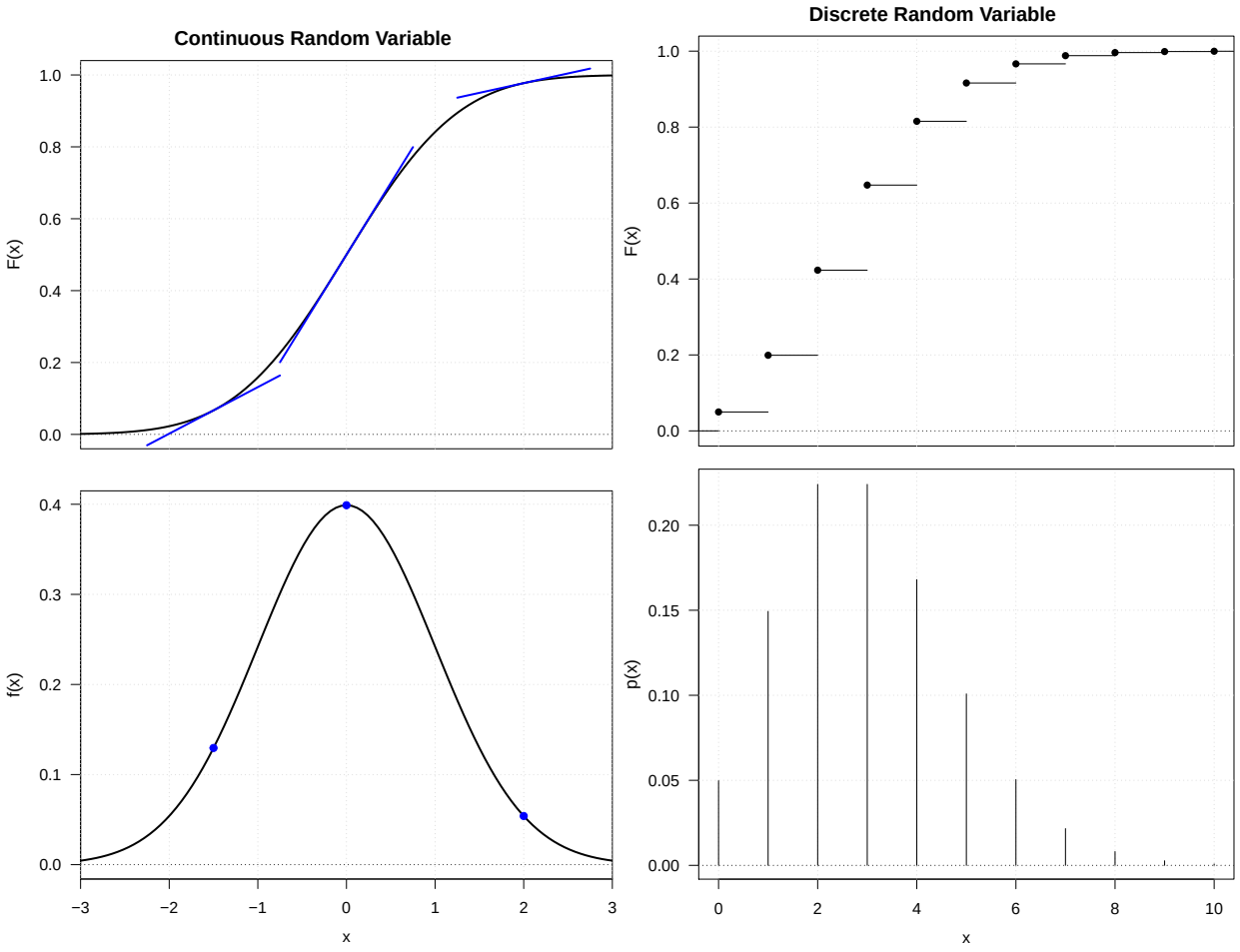
1.2 Distributions

- For many problems, an optimal decision can be formulated if we know the **distribution** of the relevant random variable(s) .
  - The random variable(s) are the unknown or unobserved values.
- Often, only certain properties of the distribution (expected value, variance, quantiles) are needed to make decisions.
- Much of statistics is involved with estimation of the distributions or their properties.

1.2.1 Random Variables

Let X be a **random variable** of interest.

- The **cumulative distribution function (cdf)** is $F(x) = \Pr(X \leq x)$.
  - $F(x)$ is the probability that the random variable X ("big X") will take a value less than or equal to x ("little x").
- For *discrete* random variables, the **probability mass function (pmf)** is $f(k) = \Pr(X = k)$.
  - $f(k) \geq 0,\ \sum_k f(k) = 1$
  - $f(k) = F(k) - F(k-1)$
- For *continuous* random variables, the **probability density function (pdf)** is $f(x) = \frac{d}{dx}F(x)$.
  - $f(x) \geq 0,\ \int_{\infty}^{\infty} f(x) = 1$



1.2.2 Parametric Distributions

A **parametric** distribution, $f(x; \boldsymbol{\theta})$ is one that is fully characterized by a set of parameters, $\boldsymbol{\theta}$. Examples include:

- Normal/Gaussian
  - parameters: mean μ, standard deviation σ
- Poisson
  - parameter: rate λ
- Binomial
  - parameters: size n, probability p
- There are also multivariate versions: Gaussian $N(\mu, \Sigma)$.

If we can model (assume) the random variable follows a specific parametric distribution, then we only need to estimate the parameter(s) to have the entire distribution characterized. The parameters are often of direct interest themselves (mean, standard deviation).

More details about some common parametric distributions can be found in the Distribution Reference Sheet

1.2.3 Non-Parametric Distributions

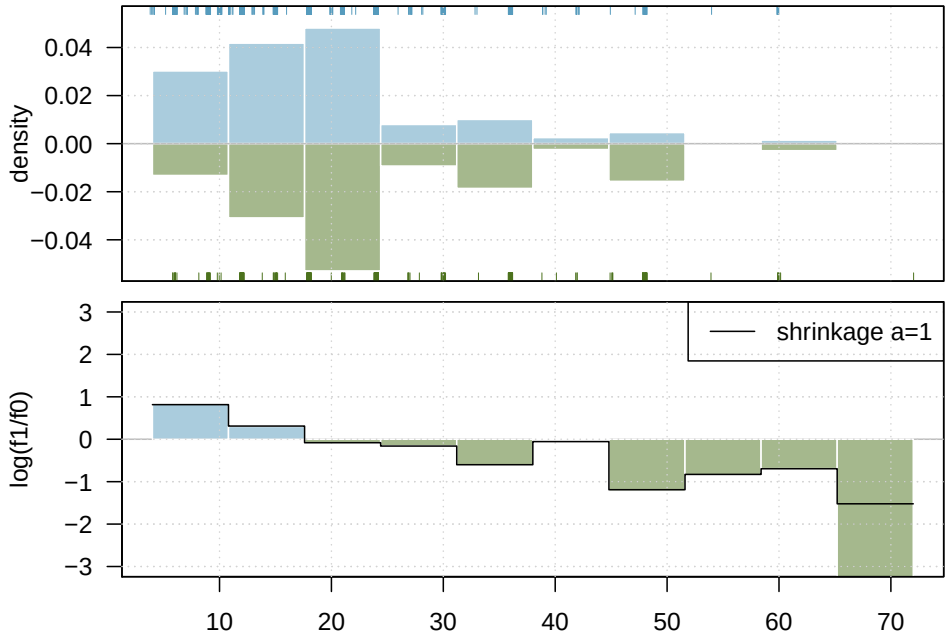
A distribution can also be estimated using **non-parametric** methods (e.g., histograms, kernel methods, splines). These approaches do not enforce a parametric family (which is essentially a type of prior knowledge), but let the data *fully* determine the shape of the density/pmf/cdf. As you might imagine more data is required for these methods to work well. Non-parametric approaches are excellent for exploratory data analysis, but can also be very useful for other types of modeling (e.g., classification, anomaly detection).

1.3 Example: Default Classification

Density estimation can be useful in *classification problems*, where the goal is to determine which class a new observation belongs to.

Below are two *histogram* density estimates; one for customers of a German bank that have good credit (blue) and the other for customers who defaulted (green). If a new customer is observed to have $X = 5$, then the evidence favors them having good credit because $X = 5$ is more likely under customers with good credit.

The bottom plot shows the corresponding log density ratio, which can help the bank make a decision on the customer's credit-worthiness.



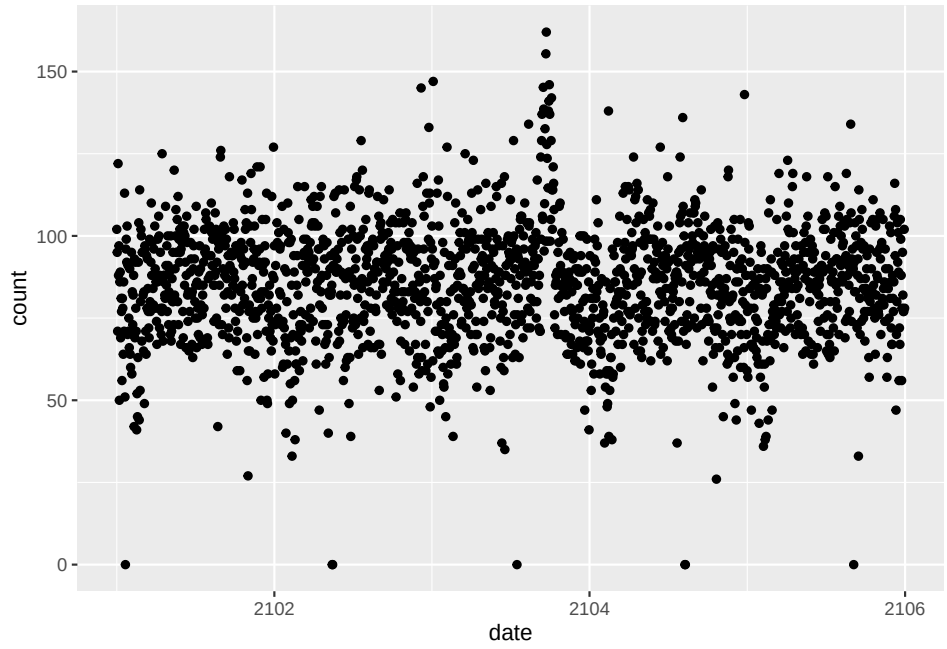
1.4 Example: Association Analysis

Density estimation can be useful in *association analysis*, where the goal is to find the regions with unusually high density (bump-hunting).

1.5 Example: Disease Outbreak Detection

Density estimation can be useful in *anomaly detection systems*, where the goal is to (often quickly) determine the time point when observations starting coming from a new or different distribution.

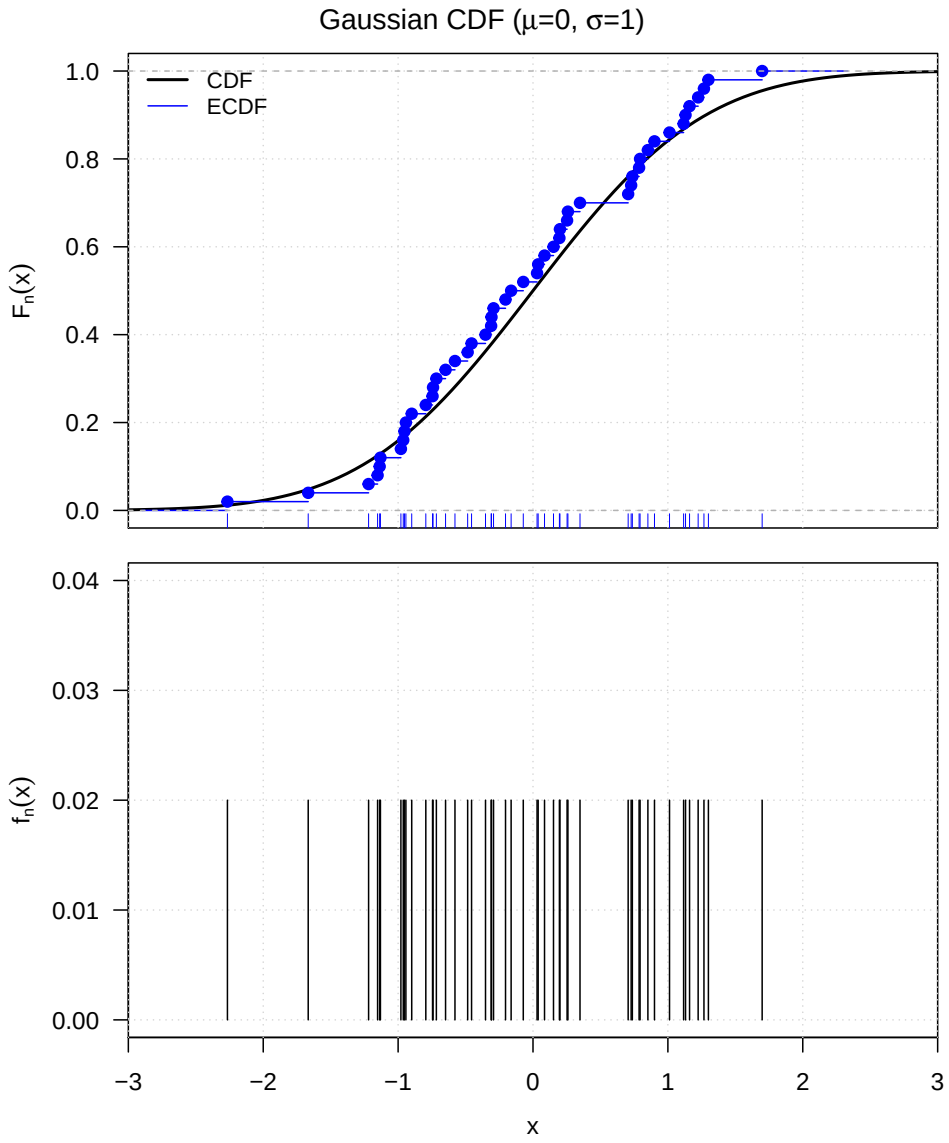
Below is simulated disease outbreak data representing the number of cases of some reported symptoms in an Emergency Department. If we can estimate the distribution of the baseline, or normal counts, on each day, then we will be able to flag an anomaly whenever the observations become *unlikely*.



1.6 Estimation

- These problems would be relatively easy to solve if we knew the exact distributions of the random variables of interest.
- Unfortunately, this is usually never the case (but of course: flipping coins, drawing cards, and playing with urns is different).
- We must use data to estimate the aspects/parameters of a distribution necessary to make good decisions.
  - And it is important to be mindful of the resulting uncertainty (bias, variance) in our estimation.

1.6.1 Empirical CDF and PDF



Data: $n = 50$ from $X \sim N(0, 1)$

2 Parametric Density Estimation

2.1 Method of Moments Estimation (MOM)

- Let X be a random variable with *pdf/pmf* $f(x; \theta)$ parameterized by $\theta \in \Theta$.
- Let $D = \{X_1, X_2, \ldots, X_n\}$ be the observed data.
- *Method of Moments (MOM)* estimators match the sample moments to the theoretical moments
  - This works when the parameter(s) can be written as functions of the moments.
  - To estimate p parameters, use p moments.

- 1st moments
  - The 1st *theoretical* moment $E[X]$ is the mean.
  - The 1st *sample* moment is the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - If $g_1(\theta) = E[X]$, then set $g_1(\theta_{\text{MM}}) = \bar{x}$ and solve for θ_{MM}.
- 2nd (central) moments
  - The 2nd *theoretical* central moment $V(X) = E[(X-\mu)^2]$ is the variance.
  - The 2nd *sample* central moment is the sample variance $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
  - If $g_2(\theta) = V(X)$, then set $g_2(\theta_{\text{MM}}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and solve for θ_{MM}.
- Maximum Likelihood estimation usually produces a *better* estimate, so we will focus on the MLE approach.

2.2 Maximum Likelihood Estimation (MLE)

- Let X be a random variable with *pdf/pmf* $f(x;\theta)$ parameterized by $\theta \in \Theta$.
- Let $D = \{X_1, X_2, \ldots, X_n\}$ be the observed data.
- *Maximum Likelihood Estimation (MLE)* uses the value of θ that maximizes the *likelihood*:

$$L(\theta) = P(X_1, X_2, \ldots, X_n; \theta)$$

- The likelihood is written as a function of θ and treating the observed data as known
- When the observations are *independent*, this becomes:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

- The log-likelihood (under independence) becomes:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$$

- And the MLE is:

$$\begin{aligned}\theta_{\text{MLE}} &= \arg\max_{\theta\in\Theta} L(\theta) \\ &= \arg\max_{\theta\in\Theta} \log L(\theta)\end{aligned}$$

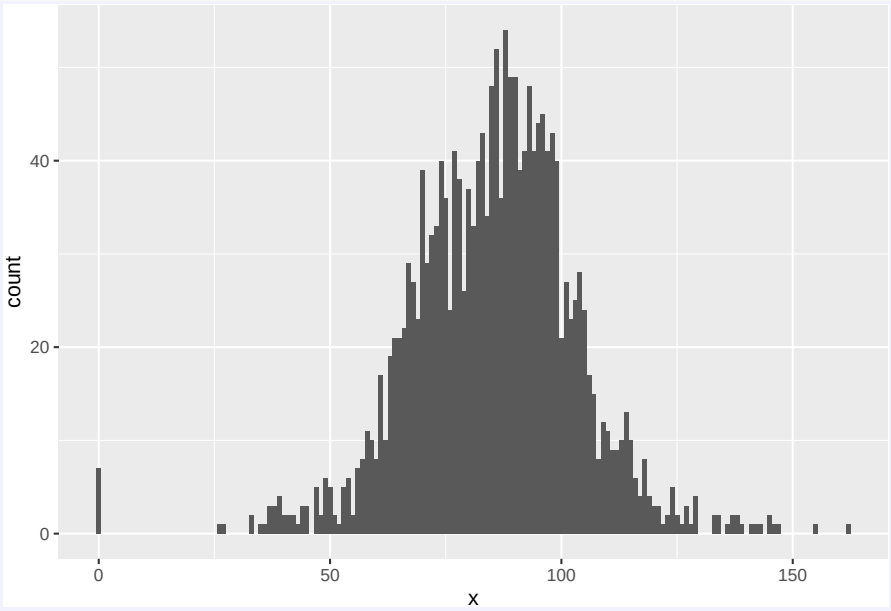
Your Turn #1

In the disease outbreak example, we needed to estimate the distribution of reported symptoms of some disease *on a normal day*. Then *unusual or rare* counts could be considered anomalous and a potential indication of a disease outbreak or bio-attack.

Estimate the baseline density of ED counts.

```r
#-- Load Data
url = 'https://raw.githubusercontent.com/mdporter/SYS6018/master/data/ED-counts.csv'
x = readr::read_csv(url)$count

#-- empirical pmf
ggplot() + geom_bar(aes(x=x))
```

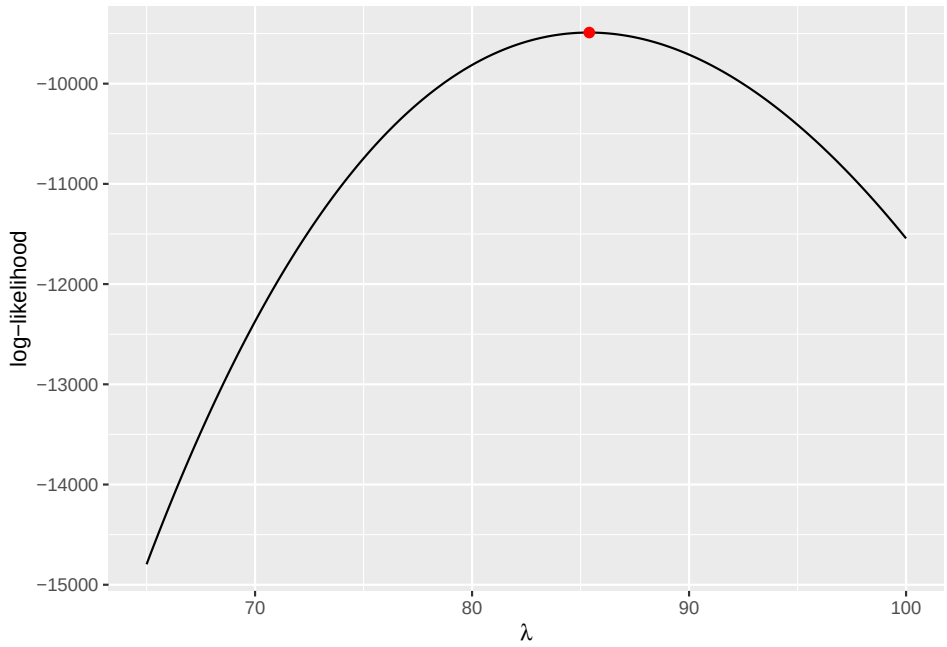


1. Use a Poisson model.
2. Use a Negative Binomial model.
3. Use a Gaussian model.
4. What is the probability that we would get more than > 150 or < 50 counts on a *regular* day?

Note: Distribution Reference Sheet

2.2.1 Poisson MLE: Grid Search

- Notation:
  - $X \sim Pois(\lambda)$
  - $\Pr(X = x) = f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$ where $f(x)$ is a pmf and $x = \{0, 1, \ldots\}$.
  - $E[X] = V[X] = \lambda$
- Calculate the log-likelihood over a range of λ values and choose the one that gives the maximum.



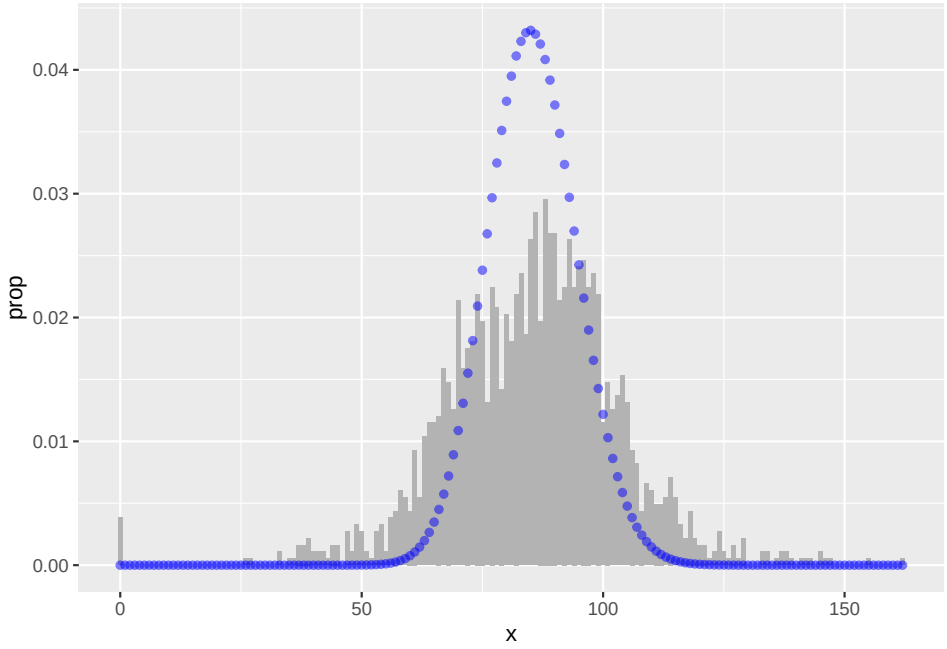
- The grid search gives $\hat{\lambda} = 85.402$
  - Searched 200 values between 65 and 100.
- See density.R for the R code.

2.2.2 Poisson MLE: Calculus

Your Turn #2

Derive the MLE using Poisson model using calculus.

Estimated pmf using Poisson MLE



- Does the Poisson model look like a good one?
- Where do you see lack of fit?

2.2.3 Negative Binomial: MLE

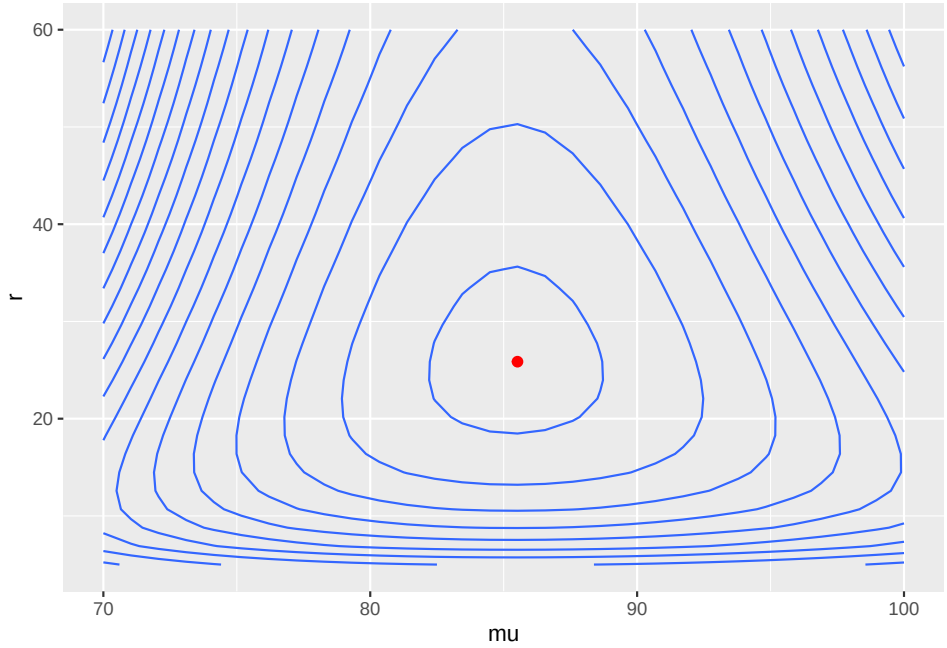
- The Negative Binomial distribution can help for modeling data with overdispersion
- Notation:
  - $X \sim NBin(\mu, r)$ (using *mean* parameterization)
  - $\mu > 0, r > 0$
  - $E[X] = \mu, V[X] = \mu + \mu^2/r$
  - Mean representation: https://en.wikipedia.org/wiki/Negative_binomial_distribution

$$\Pr(X = x; r, \mu) = \frac{\Gamma(r+x)}{x!\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^x$$

  - $n! = \Gamma(n-1)$
  - $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$

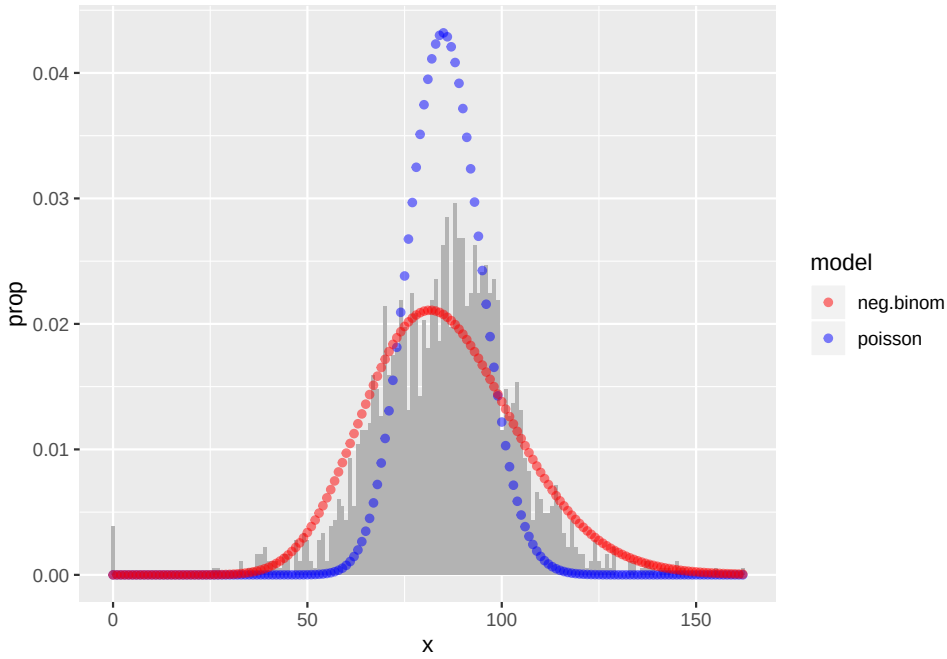
Your Turn #3

Use maximum likelihood to estimate $\theta = (r, \mu)$.



- Use R `fitdistrextra` package.

```
library(fitdistrplus)
opt = fitdist(data=x, distr="nbinom", method="mle")
nb.pars = opt$estimate
```



- Does the Negative Binomial model look better than Poisson?
- Are there any remaining concerns?

2.2.4 Example: Gaussian/Normal

- Data are non-negative integers, not continuous, so Gaussian is clearly "wrong". But as the famous saying goes:

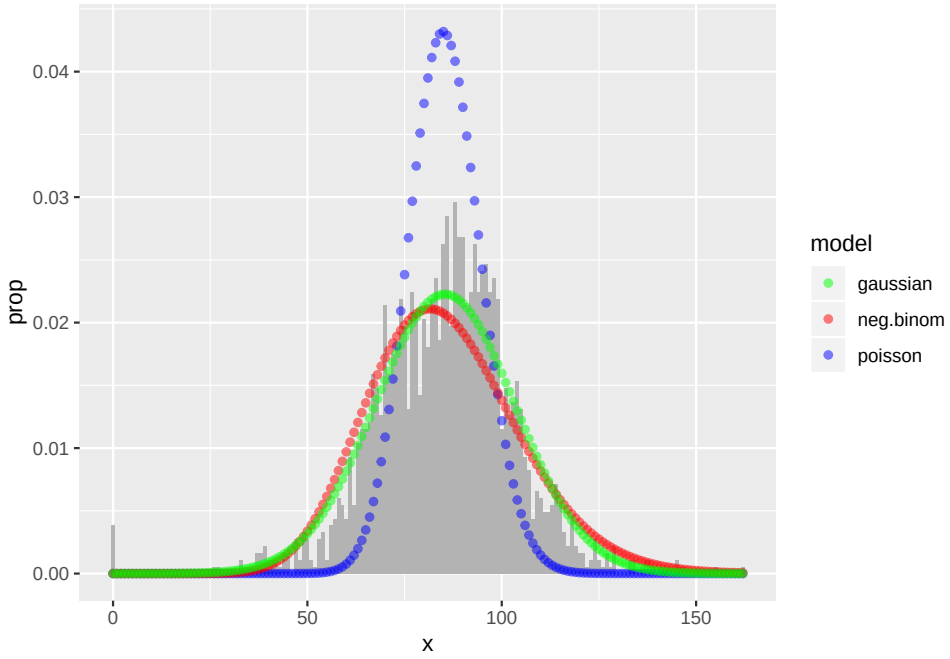
  "All models are wrong, but some are useful". - George E. P. Box

- Notation:
  - $X \sim N(\mu, \sigma)$
  - $\mu \in \mathbf{R}, \sigma > 0$
  - $E[X] = \mu, V[X] = \sigma^2$

Your Turn #4

Find the MLE for (μ, σ).

2.2.5 Comparison of Models



- Models:
  1. Poisson: $\lambda = 85.3817$
  2. Neg.Binom: $r = 25.6266, \mu = 85.3857$
  3. Gaussian: $\mu = 85.3817, \sigma = 17.919$

Your Turn #5

Which model do you choose? Why?

2.3 Bayesian Estimation

In Bayesian analysis, the parameter(s) are *random variables*.

- In MLE, the parameters are assumed fixed, but unknown.

Prior knowledge, any information known about the parameter(s) *before the data are seen*, is captured in the

prior distribution.

- Let $g(\theta)$ be the (possibly multivariate) prior pmf/pdf

Bayes theory gives us the *posterior distribution*,

$$f(\theta|D) = \frac{P(D|\theta)g(\theta)}{\int_{\theta\in\Theta} P(D|\theta)g(\theta)\,\mathrm{d}\theta}$$

- $P(D|\theta) = P(X_1, X_2, \ldots, X_n) =$ *likelihood*
- $\int_{\theta\in\Theta} P(D|\theta)g(\theta)\,\mathrm{d}\theta = P(D)$ is the *normalizing constant* (not function of θ).
- $P(\theta|D)$ is the *posterior distribution*, which contains the updated knowledge about the parameter(s).

2.3.1 Bayesian Point Estimation

1. Posterior Mean

$$\hat{\theta}_{\mathrm{PM}} = E[\theta|D] = \int_{\theta\in\Theta} \theta f(\theta|D)\,\mathrm{d}\theta$$

2. MAP (Maximum a posteriori)

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta\in\Theta} f(\theta|D) \\ &= \arg\max_{\theta\in\Theta} P(D|\theta)g(\theta) \\ &= \arg\max_{\theta\in\Theta} \left(\log P(D|\theta) + \log g(\theta)\right)\end{aligned}$$