Approximation and Approximjation Regions

Laurie Davies

Fakultät Mathematik, Universität Duisburg-Essen

Bristol, Wednesday 12th October 2011

Approximation

Role of truth in statistics:

Frequentist approach

Behave as if the model were true and then optimize, usually in an asymptotic sense
– optimal rates of convergence; a seal of quality?
Under which circumstances is this reasonable?
Other examples?

Approximation

Bayesian approach

- Additivity of priors.
- Other examples?

Approximation

Approximation as closeness to the truth (Henry Wynn)

Maximum likelihood minimizes the Kullback-Leibler divergence from the 'truth'.

Require a concept of approximation not based on truth

Approximatio sine veritate: D. W. Müller, Heidelberg

Close to the data rather than the truth:

Approximating Data.

Approximation

Data

$$\boldsymbol{x}_n = (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Model P probability measure over $\mathbb{R}^n$.

Give sense to the statement that P is (or is not) an adequate approximation to $\boldsymbol{x}_n$?

The idea applies to the given data; approximations are sought for the data at hand.

Approximation

P is an adequate approximation if 'typical' samples generated under P

$$\boldsymbol{X}_n(P) = (X_1(P), \ldots, X_n(P))$$

'look like' $\boldsymbol{x}_n$.

'typical' is a number α, $0 < \alpha < 1$, $\alpha = 0.95$.

'looks like' is a region $E_n(\alpha, P) \subset \mathbb{R}^n$

$$\boldsymbol{P}(\boldsymbol{X}_n(P) \in E_n(\alpha, P)) = \alpha.$$

Typical samples $\boldsymbol{X}_n(P)$ lie in $E_n(\alpha, P)$. The approximation is adequate if $\boldsymbol{x}_n \in E_n(\alpha, P)$.

Approximation

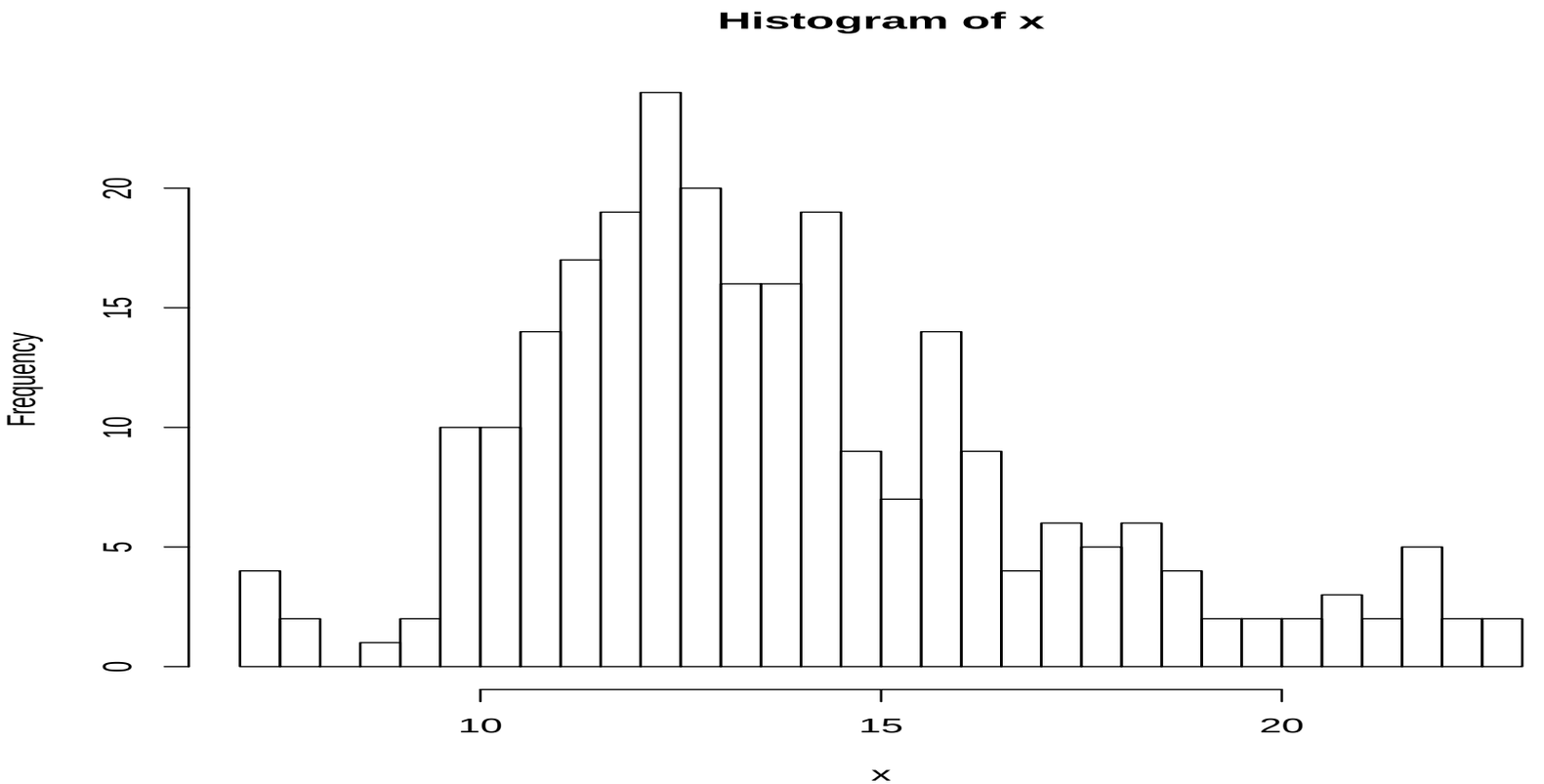
Generate 999 samples of size n under the model and insert the real sample.

The statistician now chooses 950 'typical' samples ($\alpha = 0.95$) or alternatively 50 'atypical' samples.

If the real sample is amongst the typical samples then the model is an adequate approximation.

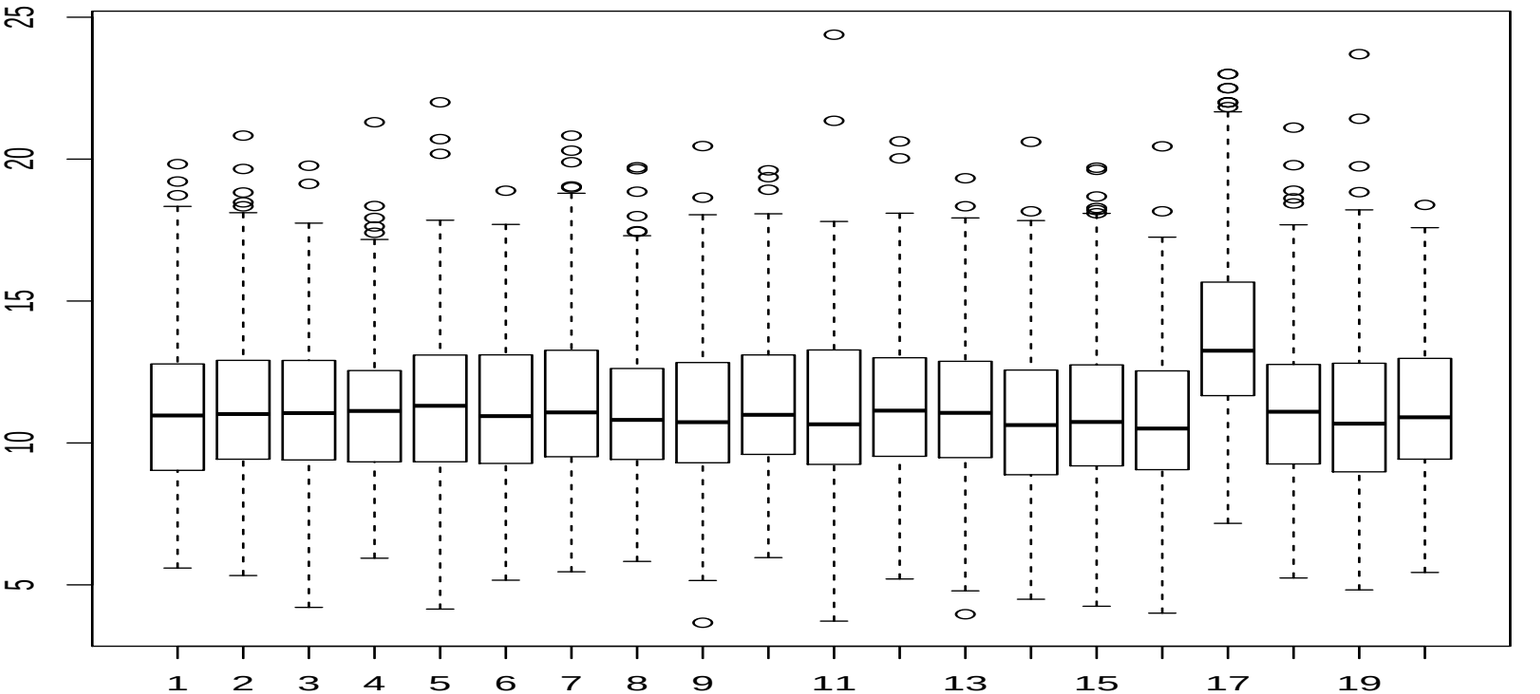
Approximation

Lengths of study in semesters (one semester = six months) of $n = 258$ German students



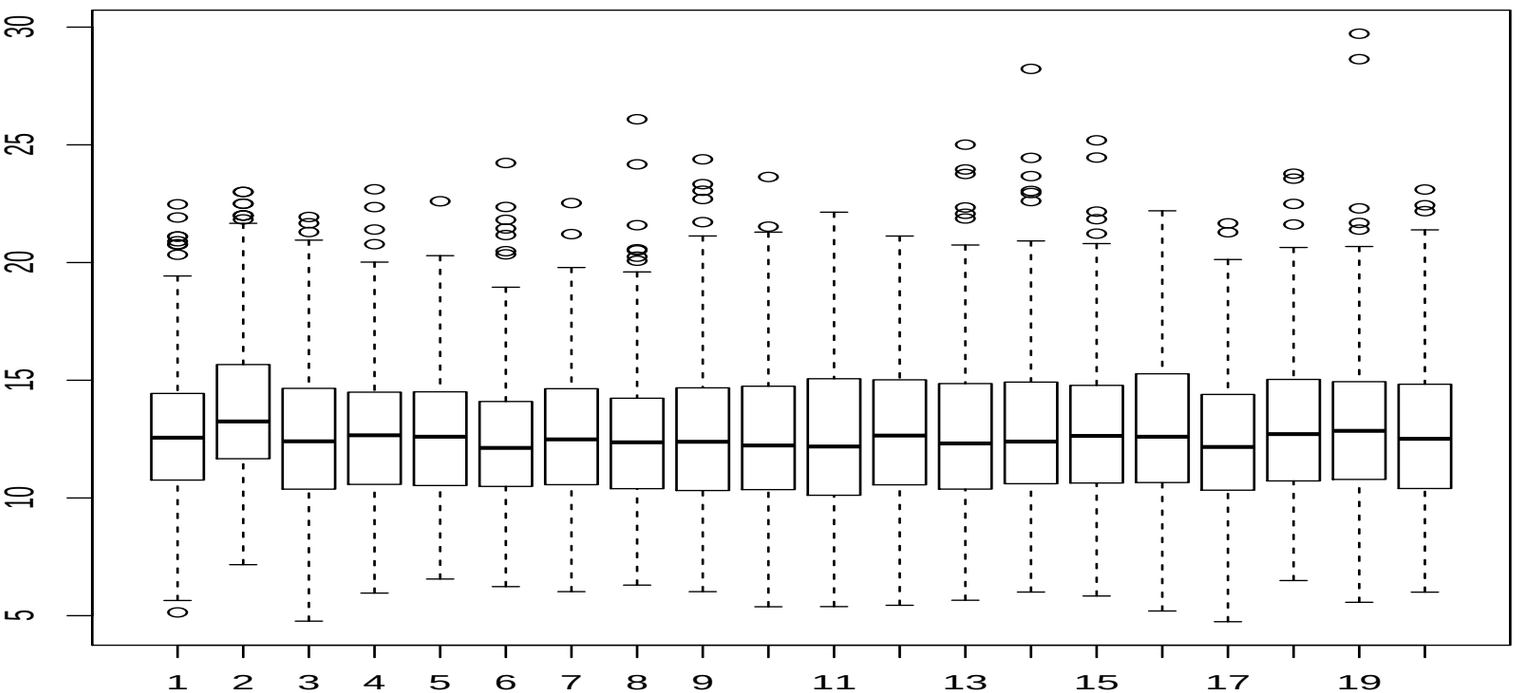
Approximation

19 samples of size $n = 258$ of the $\Gamma(16, 0.7)$ model. Boxplots



Approximation

19 samples of size $n = 258$ of the $\Gamma(16, 0.8)$ model. Boxplots



Approximation

Neyman, Scott and Shane (1953) *On the spatial distribution of galaxies*

in

Synthetic plots: history and examples

Talk by David Brillinger (2005)?

Approximation

$E_n(\alpha, P)$ specifies 'looks like', $\alpha = 0.95$.

Data $\boldsymbol{x}_n$, model $N(\mu_0, 1)$ for some $\mu_0 \in \mathbb{R}$.

Base decision on the mean. Large or small values atypical.

$$E_n(0.95, N(\mu_0, 1)) = \left\{ \boldsymbol{y}_n : \mu_0 - \frac{1.96}{\sqrt{n}} \leq \bar{\boldsymbol{y}}_n \leq \mu_0 + \frac{1.96}{\sqrt{n}} \right\}$$

Approximation regions

So far only one model P. Consider a family $\mathcal{P}$ of models.

Approximation region

$$\mathcal{A}(\boldsymbol{x}_n, \alpha, \mathcal{P}) = \{P \in \mathcal{P} : \boldsymbol{x}_n \in E_n(\alpha, P)\}$$

Parametric family $\mathcal{P}=\{\mathsf{P}_\theta : \theta \in \Theta\}$.

$$\mathcal{A}(\boldsymbol{x}_n, \alpha, \Theta) = \{\theta : \boldsymbol{x}_n \in E_n(\alpha, P_\theta)\}$$

Functionals. $T : \mathcal{P} \to \mathbb{R}^m$

$$\mathcal{A}(\boldsymbol{x}_n, \alpha, T(\mathcal{P})) = \{T(P) : \boldsymbol{x}_n \in E_n(\alpha, P)\}$$

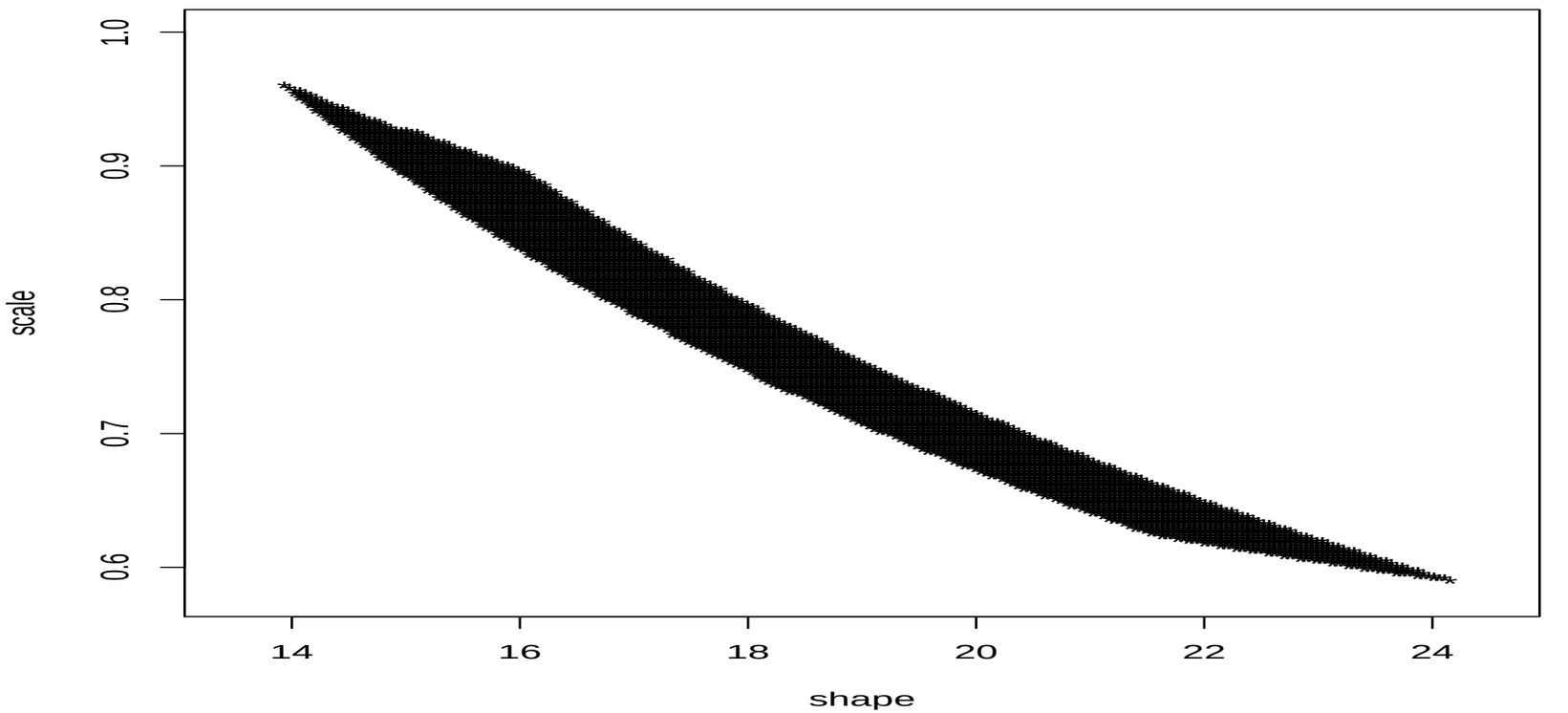
Approximation regions

Take last example but let μ vary: $\mathcal{P} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$.

$$\mathcal{A}(\boldsymbol{x}_n, 0.95, \mathcal{P}) = \left[\bar{\boldsymbol{x}}_n - \frac{1.96}{\sqrt{n}}\,,\, \bar{\boldsymbol{x}}_n + \frac{1.96}{\sqrt{n}}\right]$$

Approximation regions

Student data. Approximation region for the scale and shape parameters of the gamma distribution based on the mean and standard deviation.



Approximation regions

Student data. Approximation region for the scale and shape parameters of the gamma distribution based on total variation deviation.

$$d_{tv}(\mathbb{P}_n, P) = \sum_i |\mathbb{P}_n(\{i\}) - P(\{i\})|$$

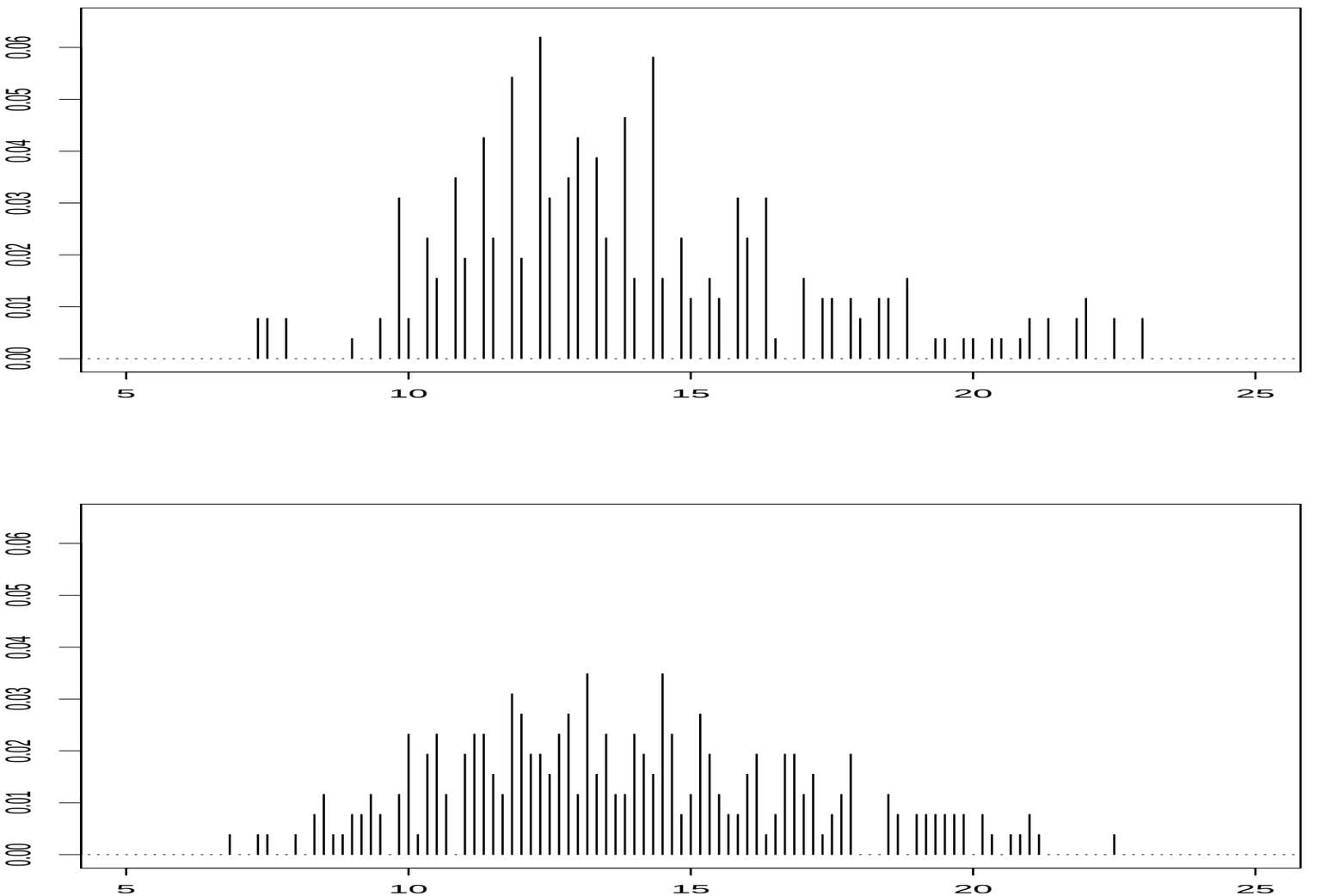
Take P to be a rounded $\Gamma(\gamma, \lambda)$.
Define $q(n, \alpha, P) = q(n, \alpha, \gamma, \lambda)$ by

$$\mathbb{P}(d_{tv}(\mathbb{P}_n(P), P) \leq q(n, \alpha, P)) \geq \alpha$$

$$\mathcal{A}(\boldsymbol{x}_n, 0.95, \mathcal{P}) = \{P : d_{tv}(\mathbb{P}_n, P) \leq q(n, 0.95, P)\} = \emptyset$$

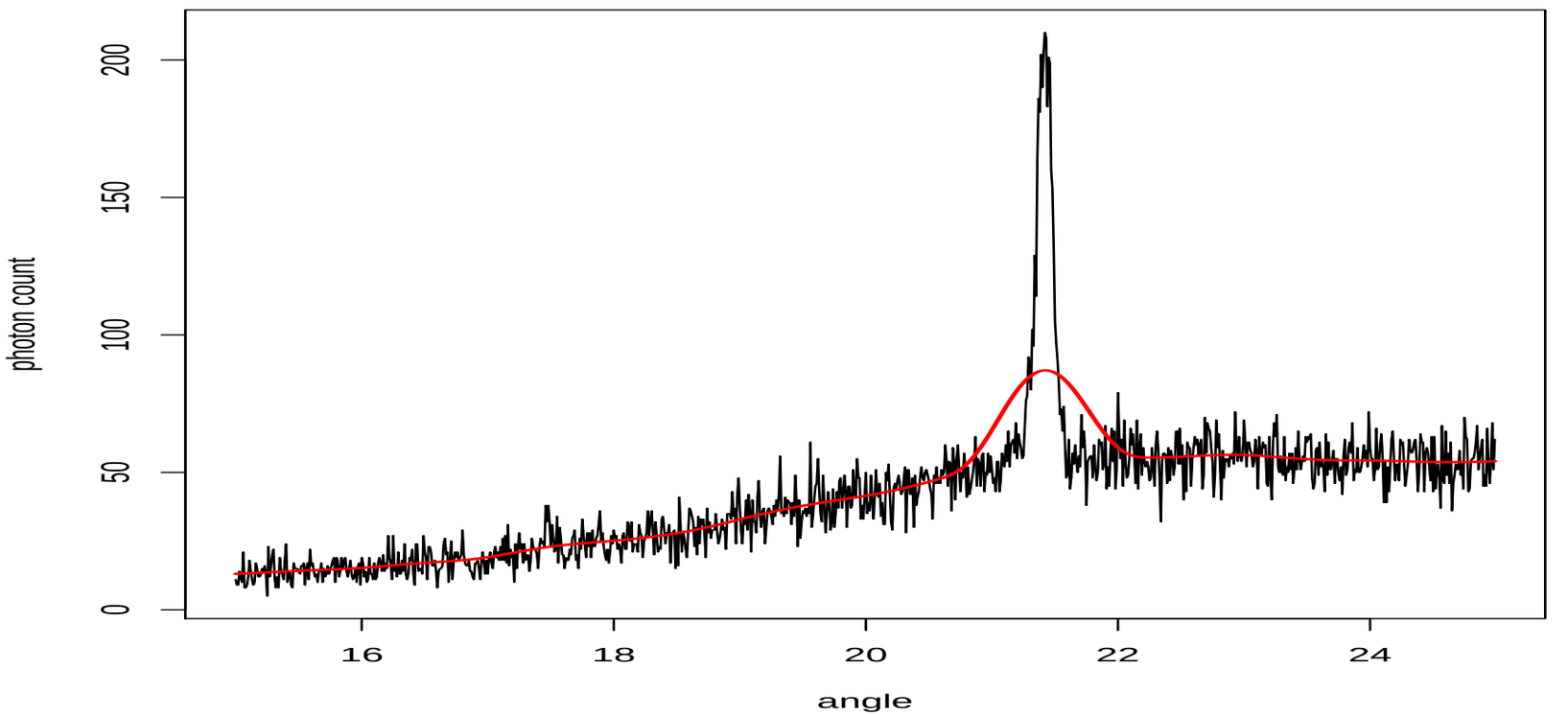
Approximation regions

The histograms look different.



Approximation regions

Non-parametric regression.
Number of photons as a function of the angle of diffraction.
Thin film physics.



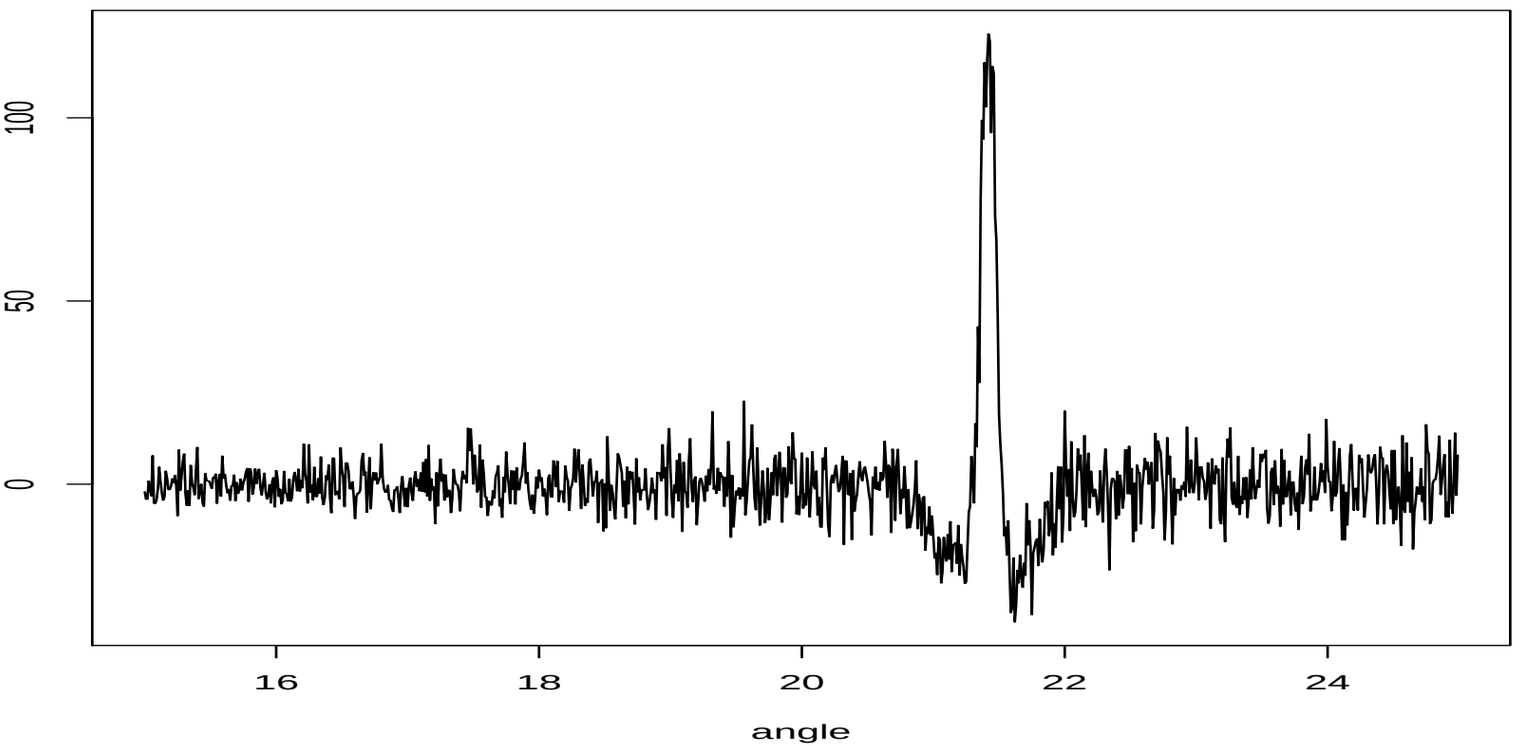
Kernel estimator with bandwidth $h = 0.8$.

Approximation regions

Model

$$X(t) = f(t) + \sigma Z(t), 15 \leq t \leq 85.$$

Assume σ specified, eg $\sigma = 7.27$. Look at residuals.



Approximation regions

Under the model the residuals $X(t) - f(t)$ are i.i.d. $N(0, \sigma^2)$.

For any interval $I \subset [15, 85]$ the $X(t) - f(t), t \in I$ are i.i.d. $N(0, \sigma^2)$
This implies

$$w_n(\boldsymbol{X}_n(f), f, I) = \frac{1}{\sqrt{|I|}} \sum_{t_j \in I} (X(t_j) - f(t_j)) \stackrel{D}{=} N(0, \sigma^2)$$

where $|I|$ = number of observations in interval I. It holds that

$$\boldsymbol{P}\left(\max_I |w_n(\boldsymbol{X}_n(f), f, I)| \leq \sigma \sqrt{\tau_n(\alpha) \log n} \right) = \alpha$$

for some $\tau_n(\alpha)$.

Approximation regions

$\tau_{1000}(0.9) = 2.77,\ \tau_{1000}(0.95) = 3.91$

$$\lim_{n\to\infty} \tau_n(\alpha) = 2 \quad \text{for all} \quad \alpha$$

Approximation region

$$\mathcal{A}(\boldsymbol{x}_n, \alpha, \mathcal{P}) = \left\{ g : \max_I |w_n(\boldsymbol{x}_n, g, I)| \leq \sigma \sqrt{\tau_n(\alpha) \log n} \right\}$$

For the thin film data with $\alpha = 0.95$

$$\sigma \sqrt{\tau_n(\alpha) \log n} = 33.15$$

Approximation regions

For $g = \hat{f}_{0.8}$ and $I = [21.36, 21.48]$

$$w_n(\boldsymbol{x}_n, \hat{f}_{0.8}, I) = 363.47$$

The kernel estimator with band width $h = 0.8$ does not give an adequate approximation to the data.

Reduce h until a satisfactory approximation is obtained. Largest such h is $h = 0.062$.

Approximation regions