# Histograms

## non-uniform bin size

Important: Now area becomes important

- if uniform binsize, height of bins usually $\to N_{bin}$, with $\sum N_{bin} = N_{TOT}$

- if non uniform: $N_{bin} = \text{area} \Rightarrow \text{height} = N_{bin}/h_{bin}$

  $\to$ results not easy to interpret

  Thus $\to$ ~~plot~~ calculate/plot pdf

  $f(x)dx = P$ ~~for~~ in $x, x+dx$

  since $P = \frac{N_j}{N_{TOT}}$

  $\Rightarrow f(x) = \frac{N_j}{N_{TOT} \cdot h_j}$

  and $\int f(x)dx \Rightarrow \sum \frac{N_j}{N_{TOT} \cdot h_j} \cdot h_j = \sum \frac{N_j}{N_{TOT}} = 1$ automatically satisfied

$\Rightarrow E(x) = \int x f(x)\, dx \to \sum x_i f_i \cdot h_i = \sum x_i \frac{N_j}{N_{TOT}}$

$Var(x) = \int (x - E(x))^2 f(x)\, dx \to \sum (x_i - E(x))^2 \cdot f_i \cdot h_i$

See: p1_new.pro: bins ordered in such a way that $N_j \approx 10$ everywhere (if possible)

$\to$ ~~hi~~ and/or similar

# How many bins?

1st rule of thumb: at least 5... 10 events/bin

- statistical fluctuations inside bin follow Poisson-statistics

$$\text{error} \approx \sigma = \sqrt{Var} = \sqrt{\underset{\text{expect. value}}{\langle N_j \rangle}} \approx \sqrt{N_j}$$

$$\text{relative error} \approx \frac{\sqrt{N_j}}{N_j} \hat{\approx} \frac{1}{\sqrt{N_j}}$$

for $N_j = 10$, relative error $\approx 33\%$

Note: same argumentation for photon-noise

if specific channel measures $N_j$ photons

$\Rightarrow$ relative error $= \frac{1}{\sqrt{N_j}}$, S/N (inverse ~~revers~~ quantity $\approx \sqrt{N_j}$

(for 200 photons/channel, S/N $\approx 14$, and relative error $= 7\%$

2nd rule of thumb

| total number of events $N_{TOT}$ | # of bins No_bins | average $N_j$ (if uniform) |
|---|---|---|
| $<50$ | 5 to 7 | 7 ... 10 |
| 50 ... 100 | 6 to 10 | 5 ... 10 |
| 100 ... 250 | 7 to 12 | 10 ... 20 |
| $>250$ | 10 to 20 | 12 ... >25 |

Note: various "rules" (based on simplified assumptions)

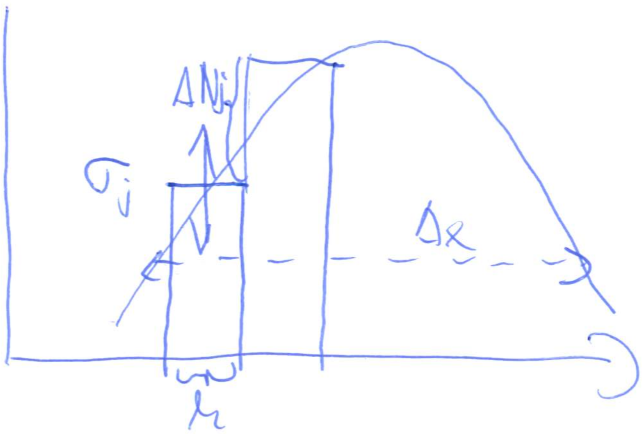
$\Rightarrow h$ propto $\frac{1}{N_{TOT}^{\frac{1}{3}}}$ for larger samples [e.g. David Scott, Biometrika, 1979, 66, 3]

since $No\_bins \approx \frac{\Delta x}{h} \Rightarrow No\_bins \sim (N_{TOT})^{\frac{1}{3}}$

Accounting for "2nd rule" from above, here my personal suggestion (if distribution quite smooth)

| | | | |
|---|---|---|---|
| $N_{TOT}$ | $10 \ldots 30$ | $\Rightarrow$ | $No\_bins \approx N_{TOT}/5$ |
| | $\geq 30$ | | $No\_bins \approx 2 \cdot N_{TOT}^{\frac{1}{3}}$ |

Question: Why $h \sim N_{TOT}^{\frac{1}{3}}$ ?




$h$ chosen in such a way that $\sigma_j \leq \Delta N_j$

$\sigma_j \approx \sqrt{N_j}$ (Poisson) $\qquad N_j \approx N_{TOT} \cdot \frac{h}{\Delta x}$

$\Delta x$ characteristic width of distribution

$$N_{j+1} \approx N_j + \frac{\partial N_j}{\partial x} \cdot h \Rightarrow \Delta N_j \approx N_{j+1} - N_j \approx \frac{\partial N_j}{\partial x} \cdot h \approx \frac{N_j}{\Delta x} \cdot h$$

From $\sigma_j =: \Delta N_j$

$$\sqrt{\frac{N_{TOT} \cdot h}{\Delta x}} = \frac{N_j}{\Delta x} \cdot h = N_{TOT} \left(\frac{h}{\Delta x}\right)^2$$

$$\Rightarrow \left(\frac{h}{\Delta x}\right)^{3/2} = N_{TOT}^{-\frac{1}{2}}$$

$$h = \frac{\Delta x}{N_{TOT}^{\frac{1}{3}}} \quad \text{and} \quad N_{0\_bins} = \frac{\Delta x}{h} \approx N_{TOT}^{\frac{1}{3}}$$

---

Start of first bin

Have seen: start of first bin $\rightarrow$ significant influence on shape/properties of histogramm

subjective choice

more objective choice: "average shifted histogram"
(David Scott, Multivariate Density Estimation, John Wiley, 1992)
calculate histogram for various start values
(e.g., if bin-size = 10, for 0, 2, 4, 6, 8),
and average histograms

$\rightarrow$ New histogramm will have steps at $\Delta$offset
(in the above example, the "new" bin width will be 2)

$\Rightarrow$ To obtain meaningfull results, calculation of pdf
$\left(f(x) = \frac{N_j}{N_{TOT} \cdot h}\right)$ required! e.g., 10, 2 ; 10, 1 ; 20, 4

$\rightarrow$ see average_shift.pro (average_shift, bins, $\Delta$offset)