Statistical Analysis of Data

Prof. Americo Cunha Jr

americocunha.org

@AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr

Estimators for statistical moments

$X_1, \cdots, X_n$ are independent observations of X

- Sample mean

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Sample skewness

$$\widehat{\gamma_1} = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu})^3}{\left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2 \right)^{3/2}}$$

- Sample variance

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2$$

- Sample kurtosis

$$\widehat{\beta_2} = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu})^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2 \right)^2}$$

main_data_analysis1.m (1/2)

```
clc; clear all; close all;
a = 2; b=5; Ns = 1000;

X      = betarnd(a,b,Ns,1);
mu     = mean(X)
sigma2 = var(X)
sigma  = std(X)
gamma1 = skewness(X)
beta2  = kurtosis(X)

figure(1)
plot(1:Ns,X,'o')
ylim([0 1]);

figure(2)
hist(X)
xlim([0 1]);
```

Statistical analysis in Matlab/Octave

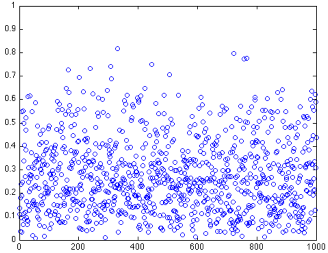
mu = 0.2849

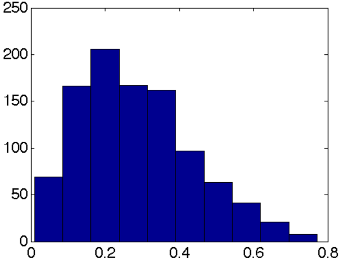
sigma2 = 0.0234

sigma = 0.1528

gamma1 = 0.5900

beta2 = 2.8555





main_data_analysis2.m (2/2)

```
clc; clear all; close all;
a = 2; b=5; Ns = 1000;

rng_stream = RandStream('mt19937ar','Seed',30081984);
RandStream.setGlobalStream(rng_stream); % Matlab 2013

X      = betarnd(a,b,Ns,1);
mu     = mean(X)
sigma2 = var(X)
sigma  = std(X)
gamma1 = skewness(X)
beta2  = kurtosis(X)

figure(1)
plot(1:Ns,X,'o')
ylim([0 1]);
figure(2)
hist(X)
xlim([0 1]);
```

Estimators for PDF and CDF

- Histogram

$$\widehat{p}_n(x) = \sum_{m=-\infty}^{+\infty} \frac{\nu_m}{n\,h_m} \, \mathbb{1}_{\mathcal{B}_m}(x)$$

- Kernel Density Estimator

$$\widehat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \, K\left(\frac{x - X_i}{h}\right)$$

- Empirical CDF

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(X_i \leq x\right),$$

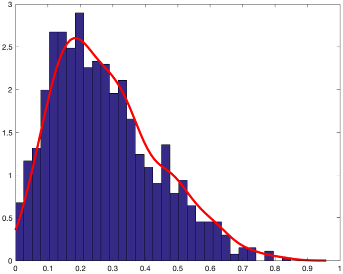
randvar_pdf.m

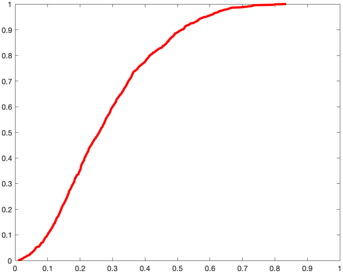
```
function [bins,freq,area] = randvar_pdf(data,numbins)

   Ns = length(data);

   data_max = max(data);
   data_min = min(data);
   binwidth = (data_max-data_min)/(numbins-1);

   bins     = (data_min:binwidth:data_max);
   freq     = histc(data,bins);
   freq     = freq/(Ns*binwidth);
   area     = binwidth*sum(freq);

end
```

main_histogram_ecdf.m

```
clc; clear; close all;

a = 2; b = 5; Ns = 1000;

X     = betarnd(a,b,Ns,1);
Nbins = round(sqrt(Ns));

[X_bins,X_freq,X_area] = randvar_pdf(X,Nbins);
[X_ksd ,X_supp1      ] = ksdensity(X);
[X_cdf ,X_supp2      ] = ecdf(X);

figure(1)
bar(X_bins,X_freq,1.0);
hold on
plot(X_supp1,X_ksd,'r','linewidth',3)
xlim([0 1]);
hold off

figure(2)
plot(X_supp2,X_cdf,'r','linewidth',3)
xlim([0 1]); ylim([0 1]);
```

PDF and CDF estimation in Matlab





Construction of a confidence interval/envelope

- p-th quantile of distribution F_X

$$Q(p) = \inf\left\{x \in \mathbb{R} : p \leq F_X(x)\right\}, \quad 0 < p < 1$$

- envelope with probability P_c

$$\mathcal{P}\left\{r^- < X \leq r^+\right\} = P_c$$

$$r^+ = Q\left((1+P_c)/2\right) \qquad r^- = Q\left((1-P_c)/2\right)$$

- estimation via percentiles
  $X_1 < X_2 < \cdots < X_n$ are independent observations of X

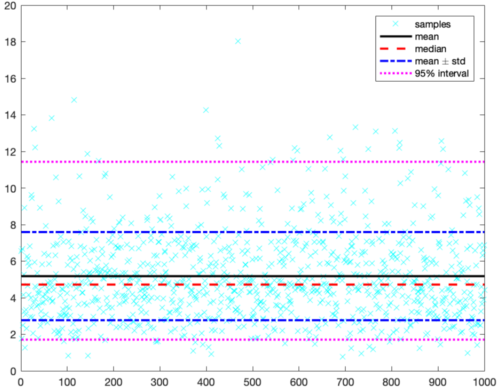
$$\widehat{r}^{\,+} = X_{n^+} \qquad n^+ = \texttt{floor}\left(n\,(1+P_c)/2\right)$$

$$\widehat{r}^{\,-} = X_{n^-} \qquad n^- = \texttt{floor}\left(n\,(1-P_c)/2\right)$$

main_conf_interval.m

```
clc; clear; close all;

a = 5.0; b = 1.0; Ns = 1000; Pc = 95;

X = gamrnd(a,b,Ns,1);
mu = mean(X); sigma = std(X); mu50 = median(X);
r_plus = 0.5*(100 + Pc); r_minus = 0.5*(100 - Pc);
X_upp = prctile(X,r_plus); X_low = prctile(X,r_minus);

figure(1)
plot(X,'xc');
hold on
line([1 Ns],[mu mu             ],'Color','k','LineStyle','- ','linewidth',2);
line([1 Ns],[mu50 mu50         ],'Color','r','LineStyle','--','linewidth',2);
line([1 Ns],[mu-sigma mu-sigma],'Color','b','LineStyle','-.','linewidth',2);
line([1 Ns],[X_low X_low       ],'Color','m','LineStyle',': ','linewidth',2);
line([1 Ns],[mu+sigma mu+sigma],'Color','b','LineStyle','-.','linewidth',2);
line([1 Ns],[X_upp X_upp       ],'Color','m','LineStyle',': ','linewidth',2);
legend('samples','mean','median','mean \pm std','95% interval')
hold off
```

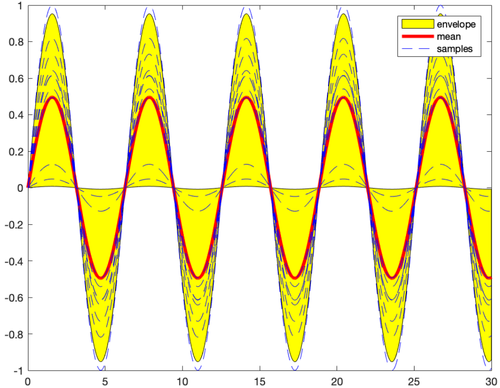
Confidence interval for a random variable



main_curve_envelope.m

```
clc; clear; close all;

Ns = 50; Pc = 95;

A = rand(Ns,1); x = 0:0.01:30; Y = A*sin(x);
r_plus  = 0.5*(100 + Pc); r_minus = 0.5*(100 - Pc);
Y_upp = prctile(Y,r_plus); Y_low = prctile(Y,r_minus);

figure(1)
fh1 = plot(x,mean(Y),'r','linewidth',3);
hold on
fh2 = plot(x,Y(1:10,:),'--b','linewidth',0.5);
fh3 = fill([x fliplr(x)],[Y_upp fliplr(Y_low)],'y');
uistack(fh3,'top');
uistack(fh1,'top');
uistack(fh2,'top');
legend('envelope','mean','samples')
hold off
```

Confidence envelope for a random curve



References

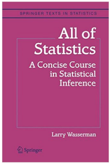
V. Stodden, *Reproducing Statistical Results*, **Annual Review of Statistics and Its Application**, 2:1–19, 2015.

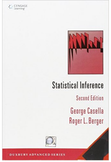
L. Wasserman, **All of Statistics: A Concise Course in Statistical Inference**, Springer, 2004.

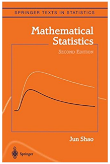
G. Casella, **Statistical Inference**, Thomson Press (India) Ltd, 2008.

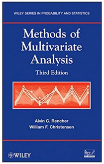
J. Shao, **Mathematical Statistics**, Springer, 2nd Edition, 2007

A. C. Rencher, **Methods of Multivariate Analysis**, John Wiley & Sons, 3rd edition, 2012









How to cite this material?

A. Cunha Jr, *Statistical Analysis of Data*, 2021.

@AmericoCunhaJr

@AmericoCunhaJr @AmericoCunhaJr