Statistical Methods for Insurance: Statistical distributions

Di Cook & Souhaib Ben Taieb, Econometrics and Business Statistics, Monash University
W3.C2

Overview of this class

- Random variables
- Central limit theorem
- Estimation
- Quantiles
- Goodness of fit
- READING: CT6, Section 1.3-1.9

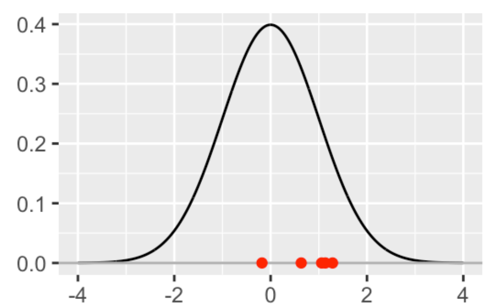
Random variables vs random samples

- Conceptually we think about a random variable (X) having a distribution

$$X \sim N(\mu, \sigma)$$

- Once we collect data, or use simulation we will have a realisation from the distribution, a random sample, observed values:

```
#> [1]  1.06 -0.18  0.63  1.14  1.29
```

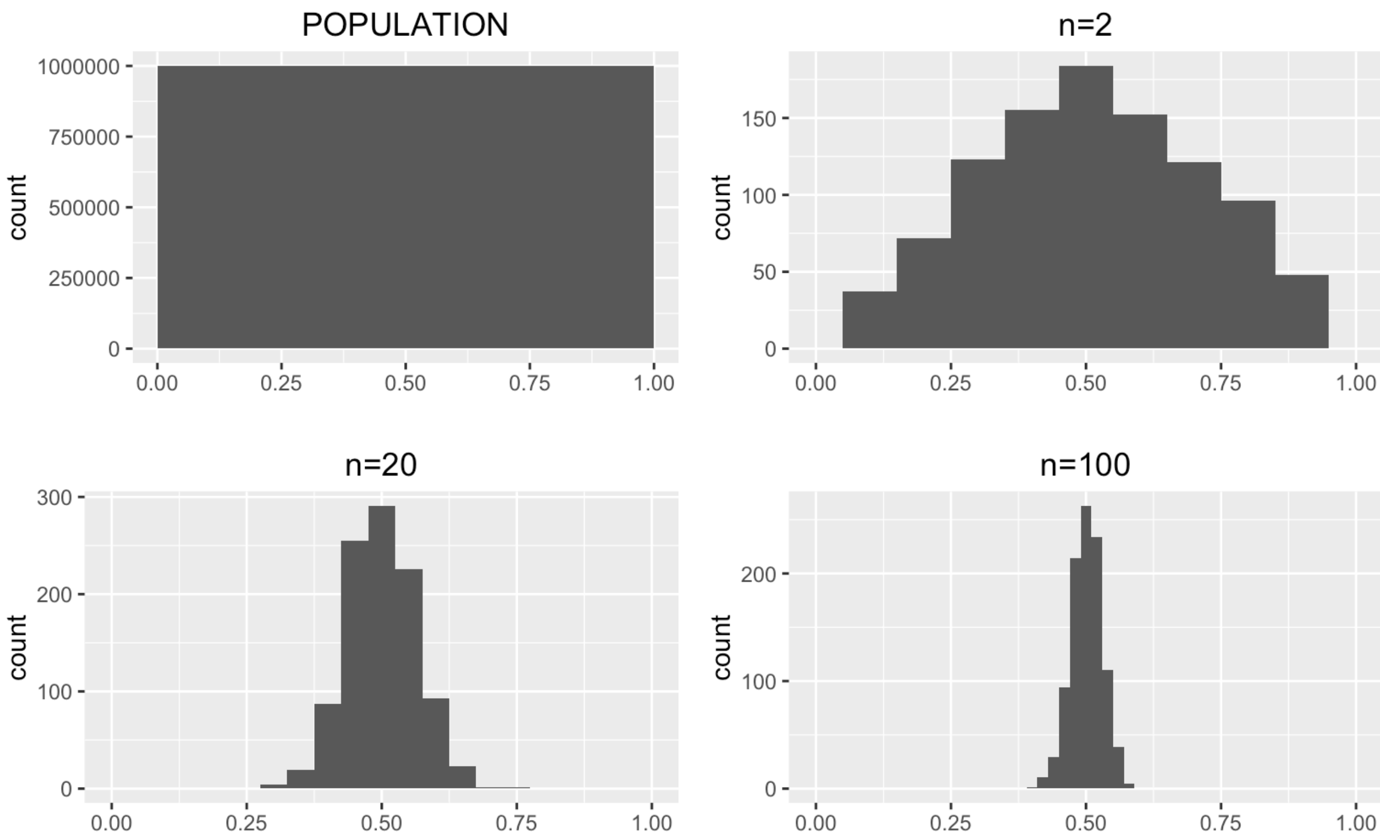


Central limit theorem

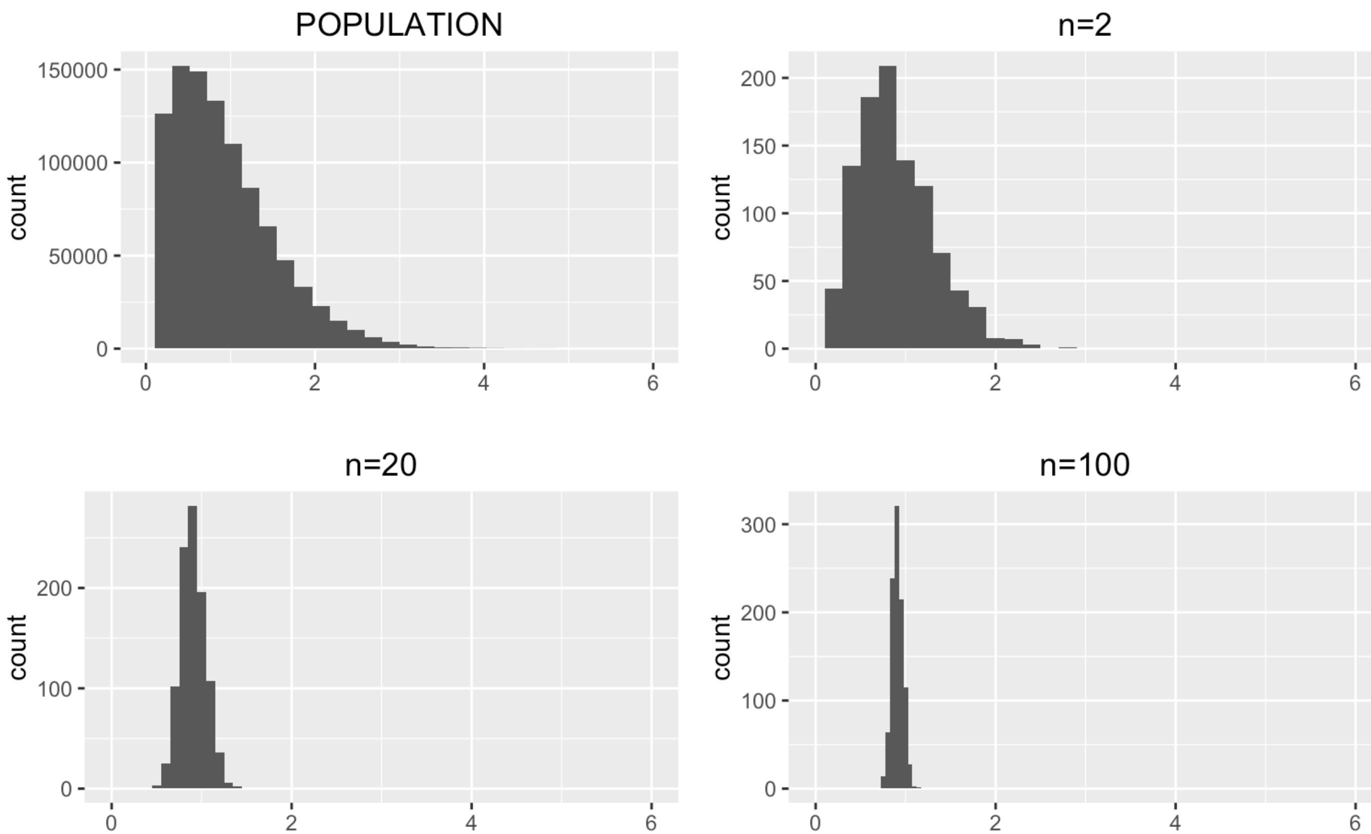
- Why the normal model is so fundamental
- Regardless what distribution X has, the mean of a sample will have a normal distribution, if the sample size is large:

"Let $\{X_1, \ldots, X_n\}$ be a random sample of size n — that is, a sequence of independent and identically distributed random variables drawn from a distribution mean given by μ and finite variance given by σ^2. The sample average is defined $\bar{X} = \sum_{i=1}^{n} X_i/n$, then as n gets large the distribution of $\bar{X}$ approximates $N(\mu, \sigma/\sqrt{n})$."

Example: Uniform parent



Example: Weibull parent



6/23

Estimation

- Estimate parameters of a distribution from the sample data
- Common approach is maximum likelihood estimation
- Requires assuming we know the basic functional form

Maximum likelihood estimate (MLE)

- Estimate the unknown parameter θ using the value that maximises the probability (i.e. likelihood) of getting the observed sample
- Likelihood function

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta) \\ &= f(x_1 | \theta) f(x_2 | \theta) \cdots f(x_n | \theta) \\ &= \prod_{i=1}^{n} f(x_i; \theta) \end{aligned}$$

- This is now a function of θ.
- Use function maximisation to solve.

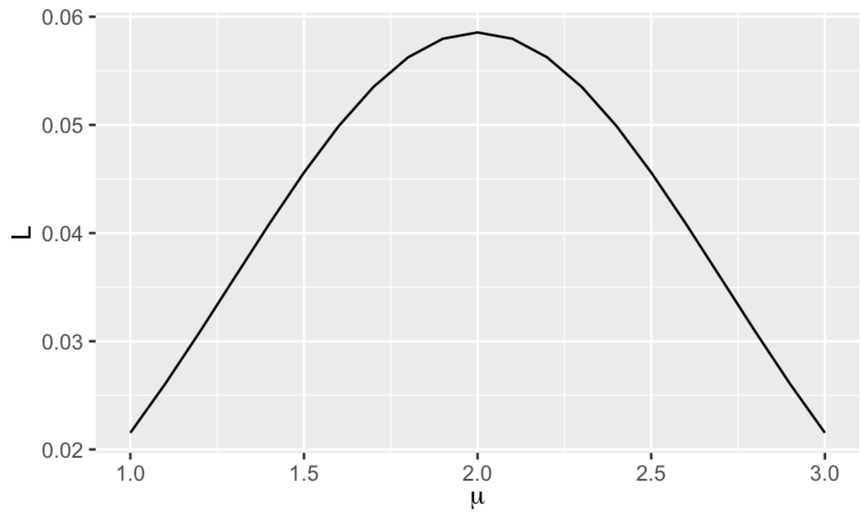
Example - Mean of normal distribution, assume variance is 1

- MLE estimate of the population mean for a normal model is the sample mean
- Run this numerically
- Suppose we have a sample of two: $x_1 = 1.0, x_2 = 3.0$
- Likelihood

$$
\begin{aligned}
L(\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(1.0-\mu)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(3.0-\mu)^2}{2}} \\
&= \frac{1}{2\pi} e^{-\frac{(1-\mu)^2+(3-\mu)^2}{2}}
\end{aligned}
$$

9/23

Plot it



- The maximum is at 2.0. This is the sample mean, which we can prove algebraically is the MLE.

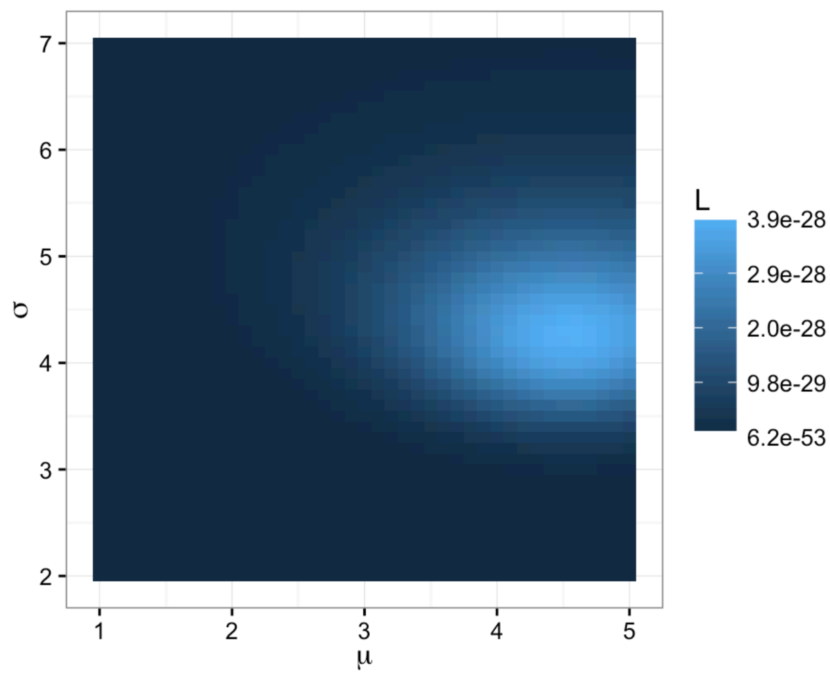
Estimate mean and variance

Sample

```
#>  [1]  7.31  3.96  2.34  0.55  5.12 10.33  3.74 -6.30 -1.14  6.55  9.13
#> [12] 10.34  5.93  1.72  3.71  3.68 -1.36 11.71  1.99  5.13  8.35  7.50
```

We know it comes from a normal distribution. What are the best guesses for the μ, σ?

Compute the likelihood for a range of values of both parameters.



Quantiles

- `quantiles` are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities
- 2-quantile is the median (divides the population into two equal halves)
- 4-quantile are quartiles, Q_1, Q_2, Q_3, dividing the population into four equal chunks
- quantiles are values of the random variable X
- useful for comparing distributions

Example:

- 12-quantiles for a $N(0, 1)$

```
qnorm(seq(1/12,11/12,1/12))
#>  [1] -1.4e+00 -9.7e-01 -6.7e-01 -4.3e-01 -2.1e-01 -1.4e-16  2.1e-01
#>  [8]  4.3e-01  6.7e-01  9.7e-01  1.4e+00
```

- 23-quantiles from a $Gamma(2, 1)$

```
qgamma(seq(1/23,22/23,1/23), 2)
#>  [1] 0.33 0.49 0.63 0.75 0.87 0.99 1.11 1.23 1.35 1.48 1.61 1.75 1.90 2.06
#> [15] 2.23 2.42 2.63 2.88 3.18 3.55 4.06 4.91
```

Percentiles

- indicate the value of X below which a given percentage of observations fall, e.g. 20th percentile is the value that has 20% of values below it
- 17th percentile from $N(0, 1)$

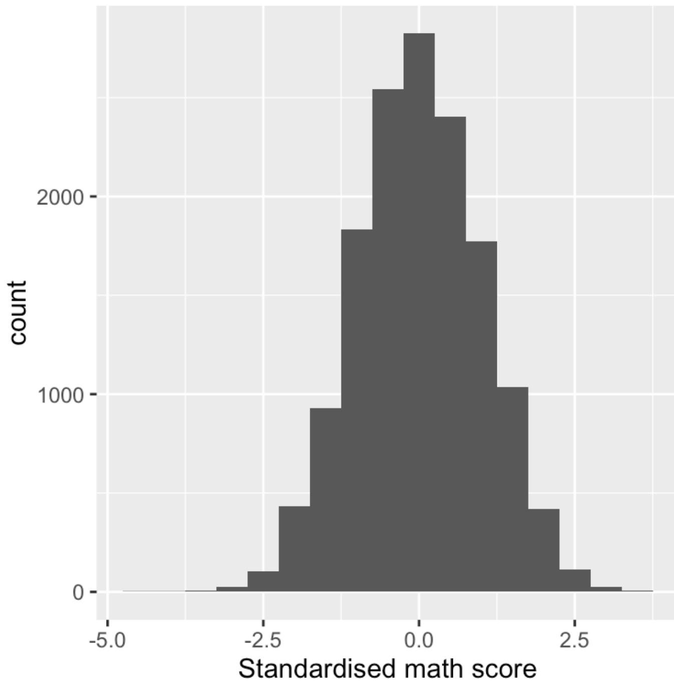
```
qnorm(0.17)
#> [1] -0.95
```

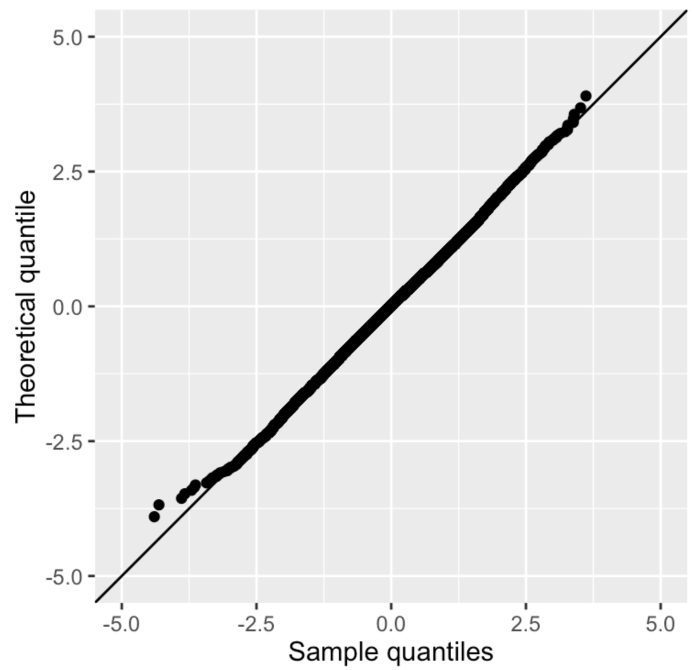
- 78th percentile from $Gamma(2, 1)$

```
qgamma(0.78, 2)
#> [1] 2.9
```

Goodness of fit

- Quantile-quantile plot (QQplot) plots theoretical vs sample quantiles
- Lets check the distribution of PISA math scores





16/23

QQ-Plot computation

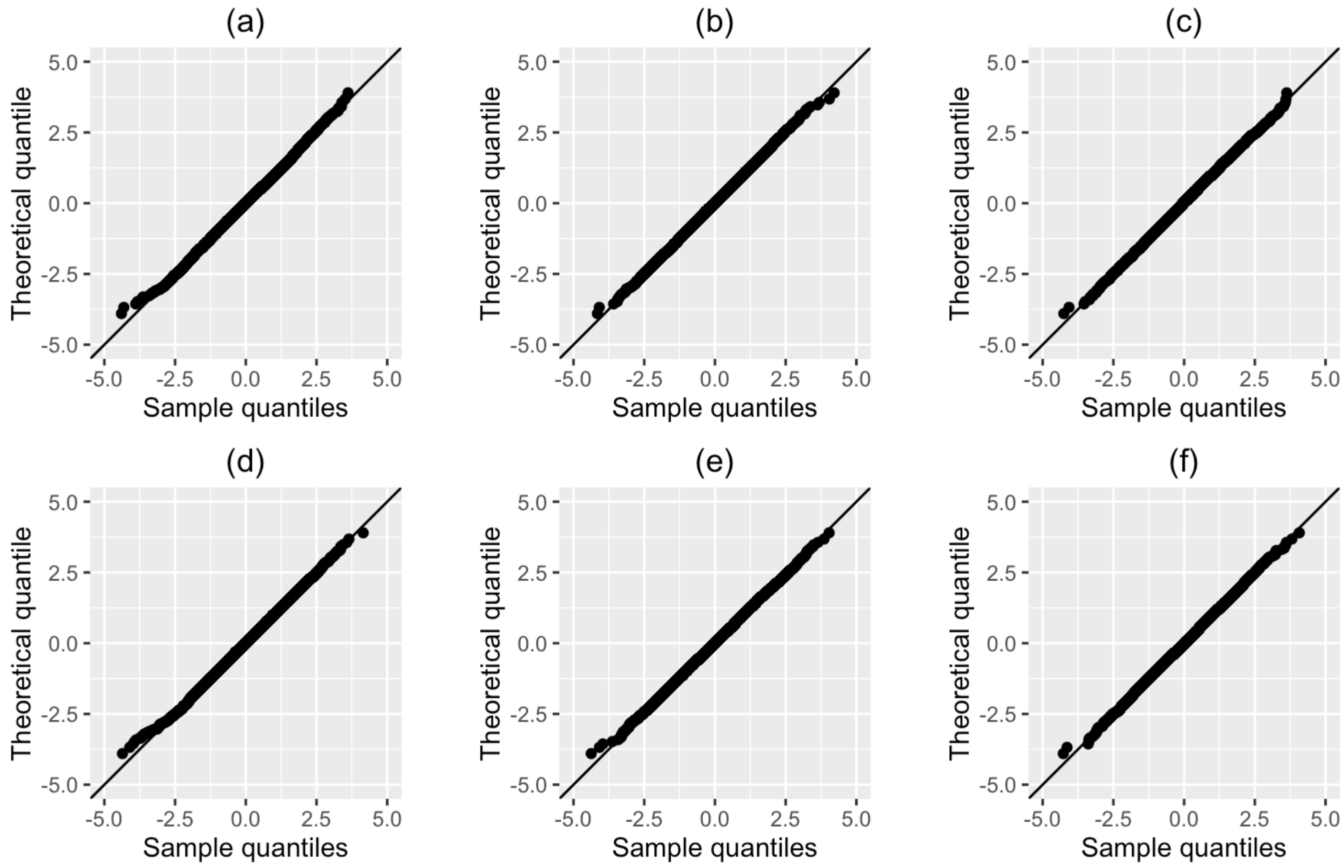
1. Sort and standardise the sample values from low to high
2. Theoretical quantiles, $n =$ sample size

$$\begin{array}{cc} 1 - 0.5^{(1/n)} & i = 1 \\ \dfrac{i - 0.3175}{n + 0.365} & i = 2, \ldots, n-1 \\ 0.5^{(1/n)} & i = n \end{array}$$

3. Plot the theoretical vs sample quantiles

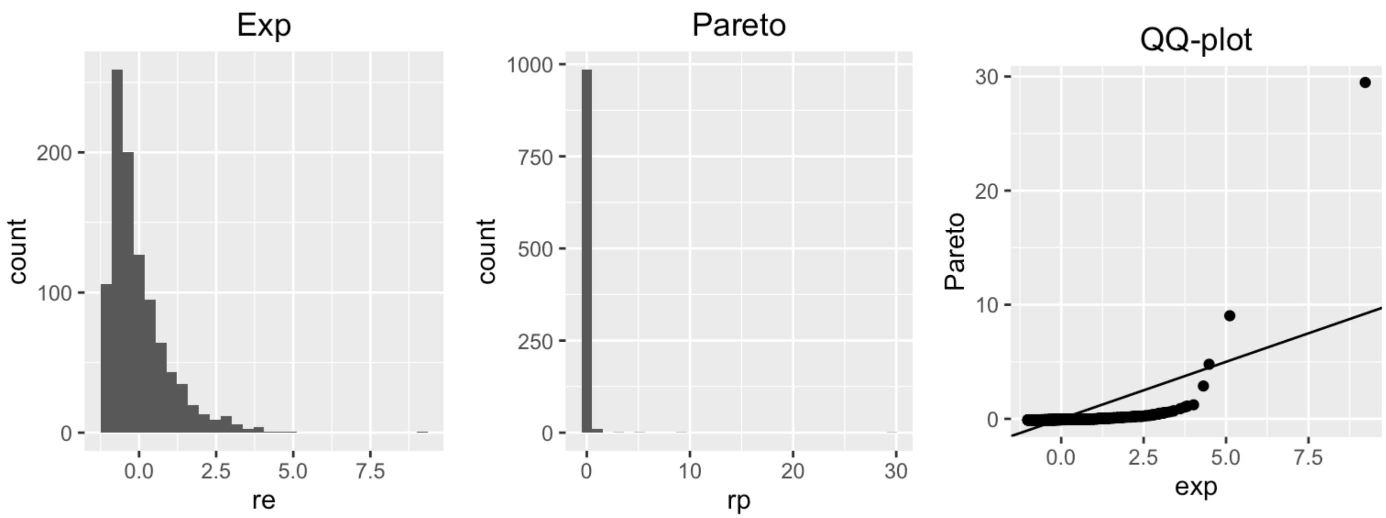
Reading QQ-plots

- The points should lie along the $X = Y$ line, for the sample to be consistent with the distribution.
- How close is good enough?
- It depends on the sample size.
- Simulate some samples of the same size from the target distribution, and make QQ-plots of these, to compare with the actual data


(a)
(b)
(c)
(d)
(e)
(f)
Theoretical quantile
Sample quantiles
5.0
2.5
0.0
-2.5
-5.0

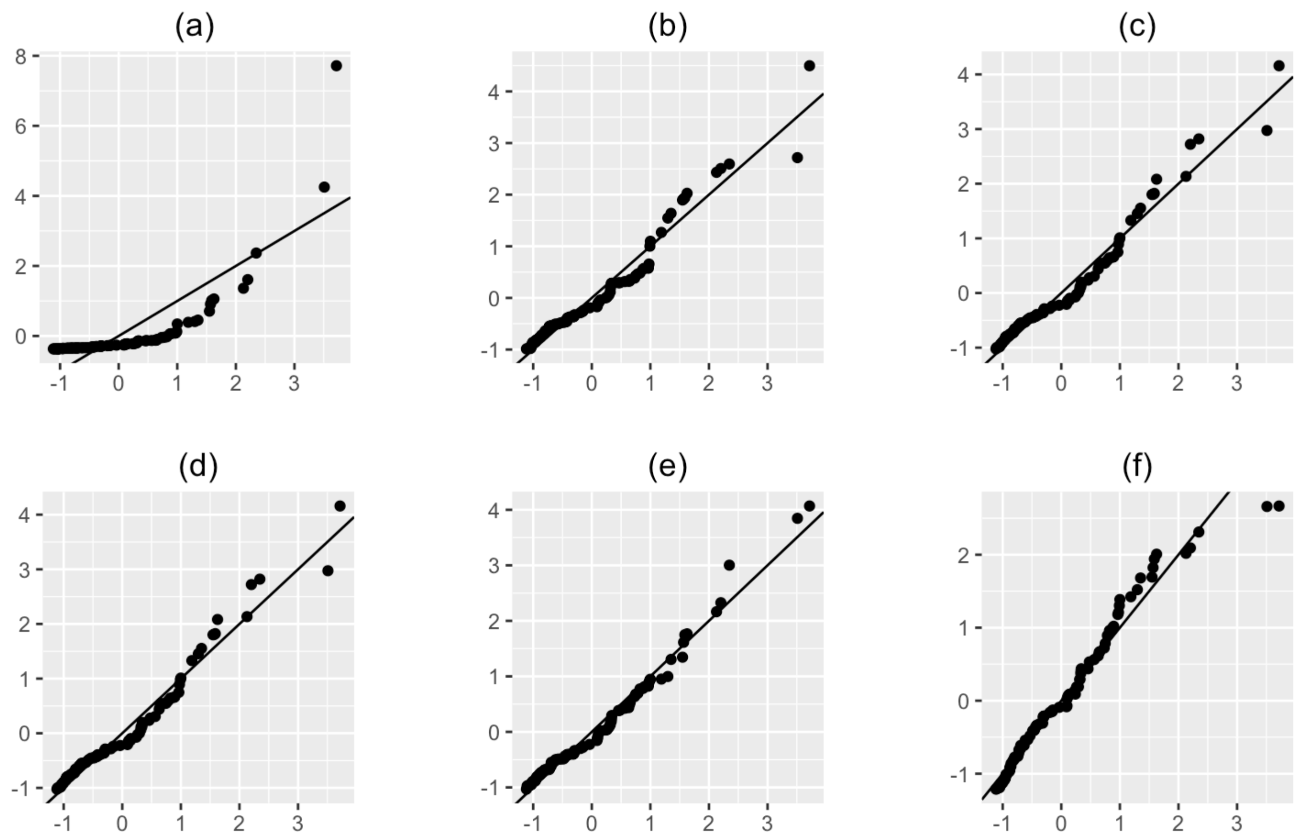
19/23

How different is exponential from Pareto?



20/23

How different can exponentials be?



Resources

- wikipedia
- PSU 414

Share and share alike