Stat 2911 Lecture Notes

Class 27, 2017

Uri Keich

Quantile-Quantile and probability plots, Extrema and Order Statistics (Rice 9.8, 10.2.3, 3.7)

Visual comparison of distributions

The Q-Q (quantile-quantile) plot of F_X and F_Y is defined as the graph

$$\{(F_X^{-1}(p), F_Y^{-1}(p)) : p \in (0,1)\}$$

2) $Y = a\underline{X} + b$ (scale: a, location: b) $\quad a>0,\ b\in\mathbb{R}$

Let $p\in(0,1)$ and denote $y_p = F_Y^{-1}(p)$. Since

$$F_Y(y) = P(a\underline{X}+b \le y)$$
$$= P(\underline{X} \le (y-b)/a)$$
$$= F_{\underline{X}}\left(\frac{y-b}{a}\right),$$

$$x_p := F_{\underline{X}}^{-1}(p)$$
$$= \frac{y_p - b}{a}$$

$$\Rightarrow y_p = a x_p + b .$$

$y_p = \varphi(x_p)$ where $Y = \varphi(\underline{X})$ this applies for any strictly $\uparrow \varphi$.

Therefore the Q-Q plot in this case is a section of the line $y = ax + b$.

Moreover, the y-intercept of the line is b = location, and the slope is a = scale.

In particular, recall $Y \sim N(\mu, \sigma^2)$ can be represented as $Y = \sigma\underline{X} + \mu$ where $\underline{X} \sim N(0,1)$.

$\Rightarrow$ The Q-Q plot of $\underline{X}$ against Y will be the line with slope σ and y-intercept of μ.

What if $F_{\underline{X}}^{-1}$ or F_Y^{-1} do not exist?

More generally the Q-Q plot is defined as

$$\{(Q_{F_X}(p), Q_{F_Y}(p)) : p \in (0,1)\},$$

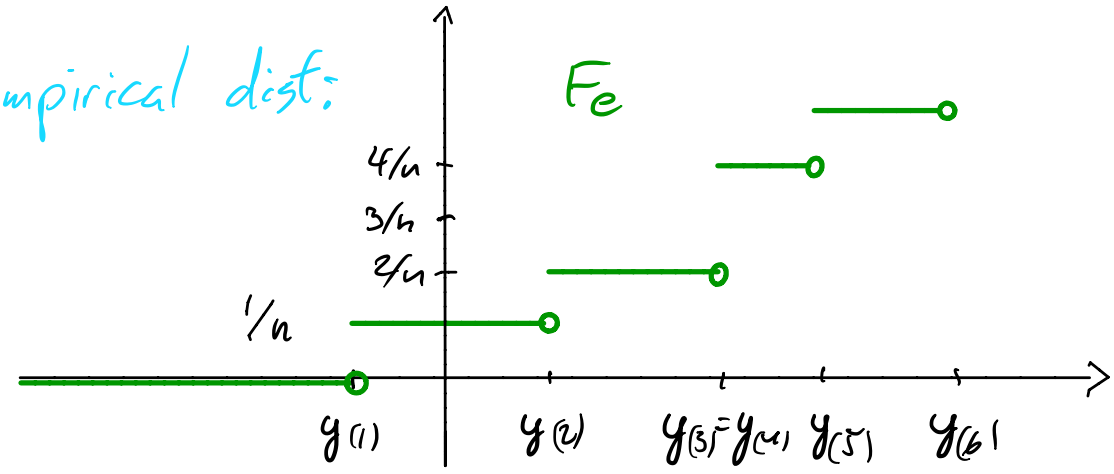
where $Q_F(p) = \min\{x : F(x) \geq p\}$ $(= F^{-1}(p)$ if F^{-1} exists$)$

Typically we wouldn't compare CDFs.

Instead, given a sample $y_1, \dots, y_n$ from an unknown dist. F_Y, we want to visually inspect the fit between the sample and a suspected dist. F_X.

We can do this by comparing F_X with F_e, the empirical (discrete) dist. of the sample.

An empirical dist: F_e



$y_{(i)}$ are the ordered sample points (order statistics)

so $\{y_{(i)}\}_1^n = \{y_i\}_1^n$ (including multiplicities!) and

$y_{(1)} \leq y_{(2)} \leq y_{(3)} \dots \leq y_{(n)}$. Formally,

$$F_e(y) = F_e^{(n)}(y) = \frac{|\{y_i : y_i \leq y\}|}{n}.$$

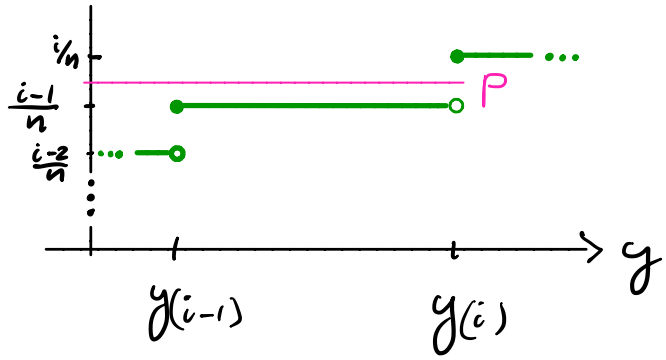
What happens as $n\to\infty$? What is $\lim_{n\to\infty} F_e^n(y)$?

For a fixed $y\in\mathbb{R}$ and $n\in\mathbb{N}$, if $Y_1,\dots,Y_n$ are iid F_Y-distributed RVs, $|\{Y_i : Y_i\le y\}|\sim$, where $p=F_Y(y)$. Therefore,

$$F_e^n(y)=\frac{|\{Y_i : Y_i\le y\}|}{n}\xrightarrow[n]{}$$ by LLN.

In fact $F_e^n\xrightarrow[n]{} F_Y$ uniformly on $\mathbb{R}$ with prob. 1 (a.s.) (Glivenko-Cantelli)

How does Q_{F_e} look like?



Clearly, for any $p\in\left(\frac{i-1}{n},\frac{i}{n}\right]$, $Q_{F_e}(p)=y_{(i)}$ so we can choose which p in $\left(\frac{i-1}{n},\frac{i}{n}\right)$ we select and accordingly pair $Q_{F_Z}(p)$ with $Q_{F_e}(p)=y_{(i)}$.

The Q-Q plot is commonly defined using $p_i=\frac{i}{n+1}$:

$$\frac{i-1}{n}<p_i=\frac{i}{n+1}<\frac{i}{n}\qquad i=1,2,\dots,n$$

so the plot consists of the n-points graph:

$$\left\{\left(Q_{F_Z}\left(\tfrac{i}{n+1}\right),\, y_{(i)}\right)\ :\ i=1,\dots,n\right\}.$$

If F_X^{-1} exists then the Q-Q plot is the graph:

$$\{(F_X^{-1}(\tfrac{i}{n+1}), y_{(i)}) : i=1,\dots,n\}.$$

The R function qqnorm which draws the Q-Q plot of a sample $\{y_i\}$ against $F_X = \Phi(N(0,1))$ uses $p_i = \frac{i-1/2}{n}$ for $n \geq 10$.
We will return to the selection of p_i later.

As mentioned, $F_{\hat{F}}^n(y) \xrightarrow[n]{} F_y(y)$ and it follows that with $i_n \xrightarrow[n]{} \infty$ s.t. $p_{i_n} = \frac{i_n}{n+1} \xrightarrow[n]{} p$, with prob. 1

$$\lim_{n\to\infty} y^n_{(i_n)} = \lim_{n\to\infty} Q^n_{\hat{F}}(p_{i_n}) = F_y^{-1}(p),$$

provided F_y^{-1} exists and is cont. at p (e.g. F_y is cont. and strictly $\uparrow$).

It follows that as $n\to\infty$ the (empirical) Q-Q plot will converge to the distributional Q-Q plot:

$$\{(F_X^{-1}(p), F_y^{-1}(p)) : p \in (0,1)\}$$

In particular, if $y = aX+b$, we expect the empirical Q-Q plot to roughly follow a straight line. Useful for gauging if your sample came from a normal dist.

Q-Q plots compare quantiles or F^{-1}.
Probability (P-P) plots compare the CDFs (F):
If F and G are CDFs then the probability plot comparing F and G is defined by the graph

$$\{(G(x), F(x)) : x \in \mathbb{R}\}$$

Advantage over Q-Q plots: the scale is constant.
If G^{-1} exists then this graph is essentially the same as

$$\{(p, F(G^{-1}(p))) : p \in (0,1)\},$$

which suggests the following alternative definition of a probability plot:

$$\{(p, F(Q_G(p))) : p \in (0,1)\}.$$

And for comparing a sample $y_1, \dots, y_n$ with F

$$\{(i/(n+1), F(y_{(i)})) : i = 1, \dots, n\}$$

In summary:
Given a sample $y_1, \dots, y_n$ we can visually gauge how well it fits the CDF F by plotting

Q-Q: $F^{-1}\left(\frac{i}{n+1}\right)$ vs. $y_{(i)}$

P-P: $i/(n+1)$ vs. $F(y_{(i)})$

Theoretical PP & QQ against normal

User: Anon

May 15, 2017

The following three distributions are visually compared with the standard normal distribution using QQ plots and PP.

- $N(10, \sigma^2 = 25)$
- A gamma distribution with shape $a = 4$ and rate $\lambda = 0.4$ which has the same mean (location) and standard deviation (scale) as the normal $N(10, \sigma^2 = 25)$ distribution
- Cauchy distribution with location $\mu = 10$ and scale $\sigma = 5$ again:

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sigma}.$$

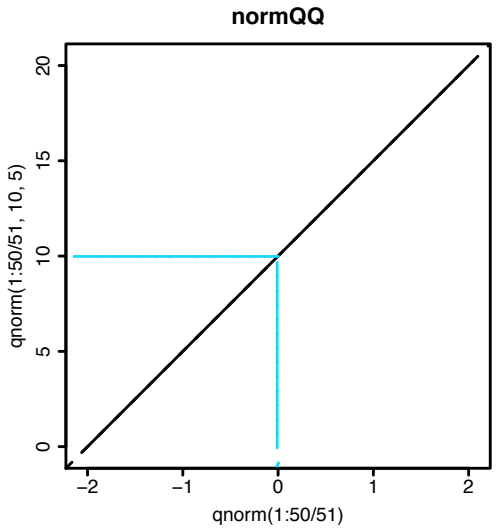
We first compare all three theoretical distributions with the $N(0, 1)$ distribution.

```
> par(mfrow=c(3,2), pty="s", ps=20, cex=0.25)
> plot(qnorm(1:50/51),qnorm(1:50/51,10,5),type="l",main="normQQ")
> abline(c(10,5), lty="dashed")
> plot(1:50/51,pnorm(qnorm(1:50/51,10,5)),type="l",main="normPP")
> plot(qnorm(1:50/51),qgamma(1:50/51,4,.4),type="l",main="gammaQQ")
> abline(c(10,5), lty="dashed")
> plot(1:50/51,pnorm(qgamma(1:50/51,4,.4)),type="l",main="gammaPP")
> abline(c(10,5), lty="dashed")
> plot(qnorm(1:50/51),qcauchy(1:50/51, 10, 5),type="l",main="CauchyQQ")
> abline(c(10,5), lty="dashed")
> plot(1:50/51,pnorm(qcauchy(1:50/51, 10, 5)),type="l",main="CauchyPP")
> abline(c(10,5), lty="dashed")
>
```

$\{(F^{-1}(p), G^{-1}(p)) : p = i/51\}$

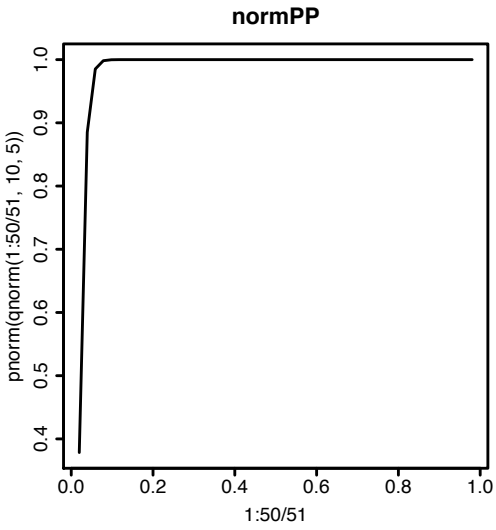
$\{(p, FG^{-1}(p)) : p = i/51\}$

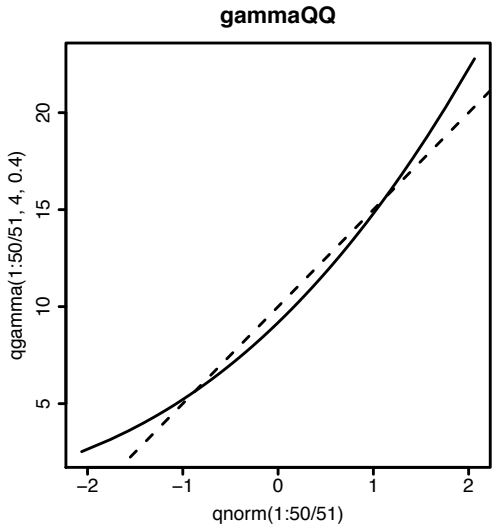
$$F(3) = \Phi(3) = \int_{-\infty}^{3} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt$$

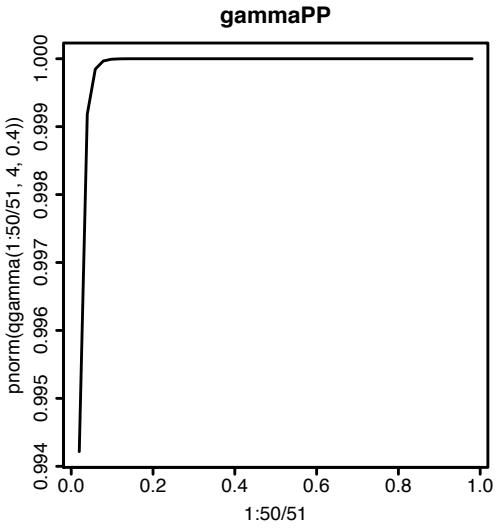


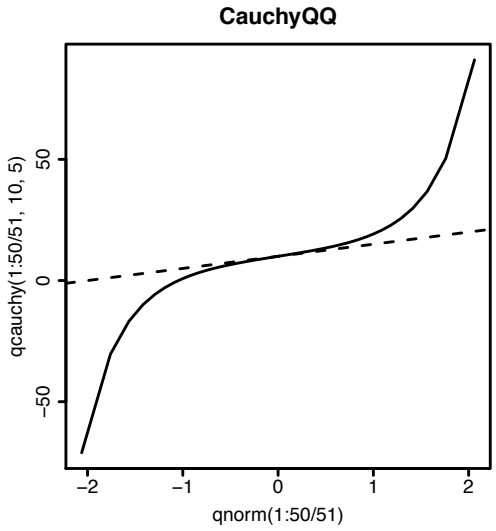
$X \sim N(0,1)$

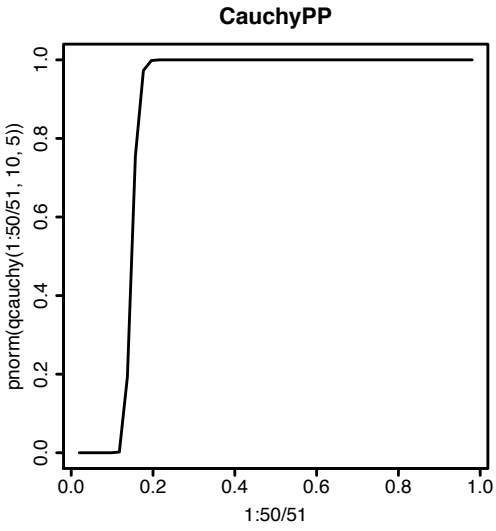
$Y = 5 \cdot X + 10$











P-P and Q-Q plots: sample size $n = 50, 1000$

User: Anon

May 15, 2017

Samples of size $n = 50$ are taken from the same 3 distributions as before:

- $N(10, \sigma^2 = 25)$
- A gamma distribution with shape $a = 4$ and rate $\lambda = 0.4$ which has the same mean (location) and standard deviation (scale) as the normal $N(10, \sigma^2 = 25)$ distribution
- Cauchy distribution with location $\mu = 10$ and scale $\sigma = 5$ again:

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2} \cdot \frac{1}{\sigma}.$$

All samples are then visually compared against the standard normal distribution and the normal sample is also compared against the correct $N(10, \sigma^2 = 25)$ distribution. These plots are then followed by those generated from samples of size $n = 1000$.

```
> n = 50
> x = rnorm(n, 10, 5)
> d = sqrt(12*25)/2
> z = rgamma(n,4,.4)
> y = rcauchy(n, 10, 5)
> par(mfrow=c(3,2), pty="s", ps=20, cex=0.25)
> plot(1:n/(n+1), pnorm(sort(x),10,5), main="PPnormal (correct mean/sd)")
> abline(c(0,1), lty="dashed")
> plot(1:n/(n+1), pnorm(sort(x),0,1), main="PPnormal (incorrect mean/sd)")
> abline(c(0,1), lty="dashed")
> plot(qnorm(1:n/(n+1),10,5), sort(x), main="QQnormal (correct mean/sd)")
> abline(c(0,1), lty="dashed")
> plot(qnorm(1:n/(n+1)), sort(x), main="QQnormal (incorrect mean/sd)")
> abline(c(mean(x), sd(x)), lty="dashed")
> plot(qnorm(1:n/(n+1)), sort(z), main="QQgamma")
> abline(c(mean(z), sd(z)), lty="dashed")
> plot(qnorm(1:n/(n+1)), sort(y), main="QQCauchy")
> abline(c(10, 5), lty="dashed")
>
```

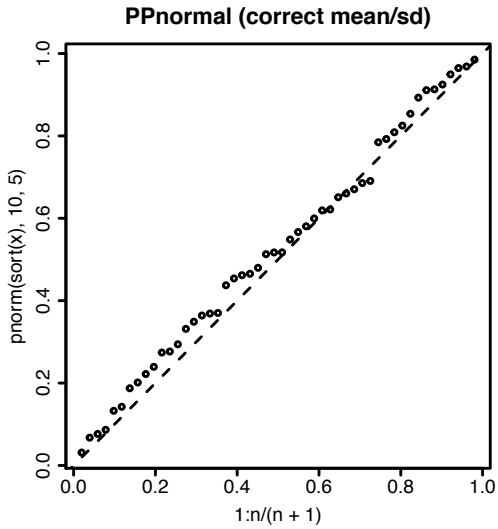
$\left\{\left(F^{-1}\left(\frac{i}{n+1}\right), y_{(i)}\right) : i = 1, \ldots, n\right\}$

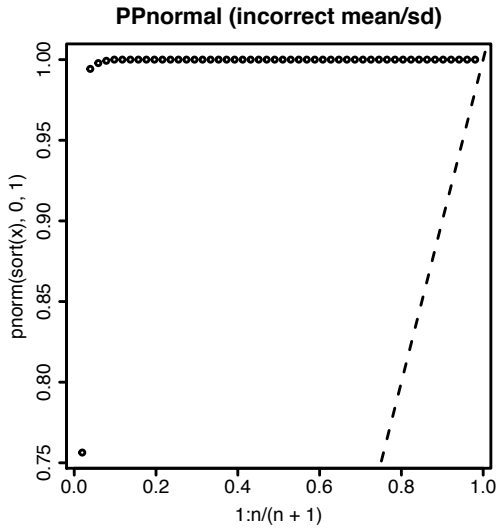
$\left\{\left(\frac{i}{n+1}, F(y_{(i)})\right) : i = 1, \ldots, n\right\}$

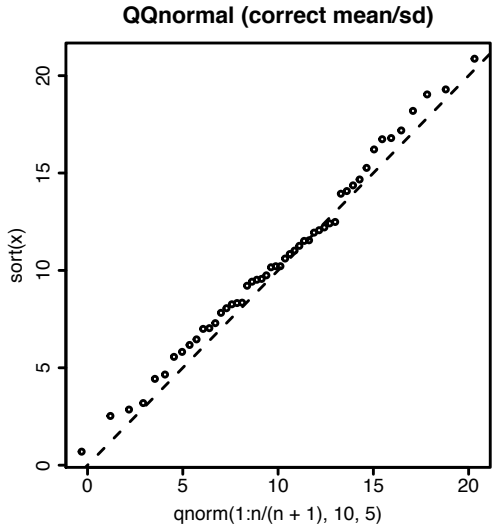
$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt$$

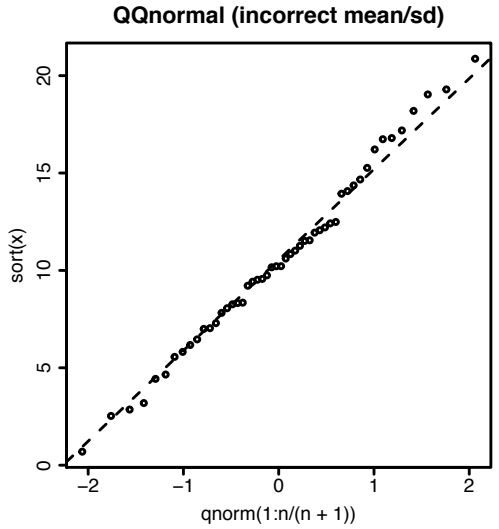
n=50

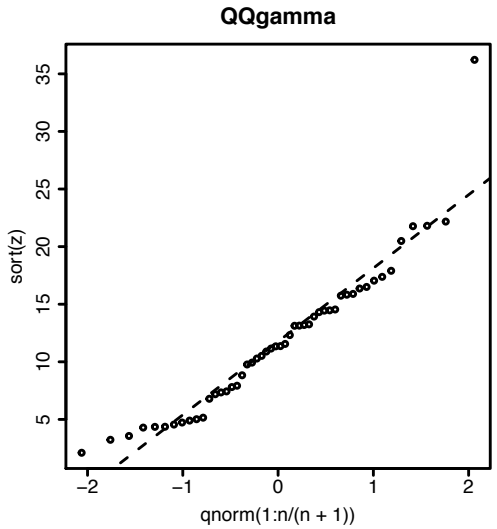
27.10

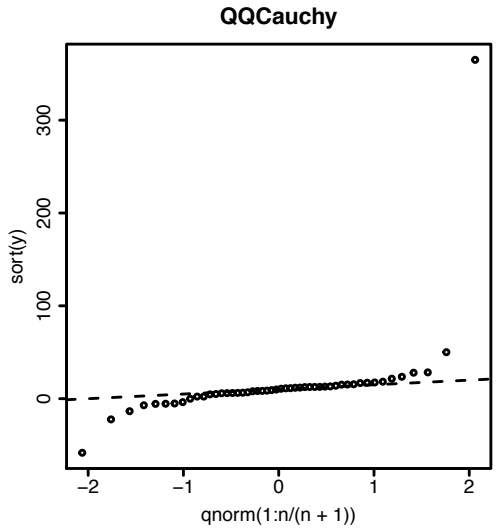

PPnormal (correct mean/sd)
pnorm(sort(x), 10, 5)
1:n/(n + 1)
0.0
0.2
0.4
0.6
0.8
1.0


PPnormal (incorrect mean/sd)
pnorm(sort(x), 0, 1)
1:n/(n + 1)
0.75
0.80
0.85
0.90
0.95
1.00

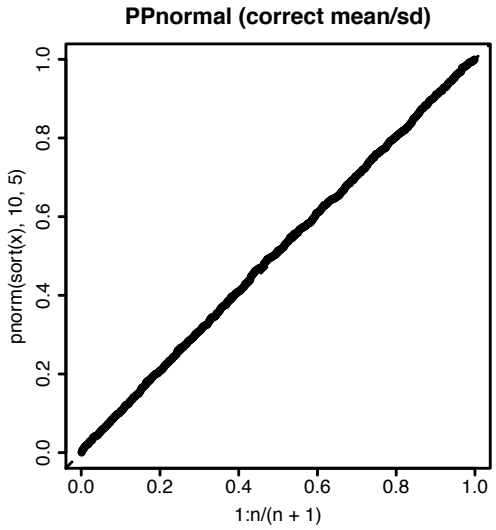

QQnormal (correct mean/sd)
sort(x)
qnorm(1:n/(n + 1), 10, 5)
0
5
10
15
20

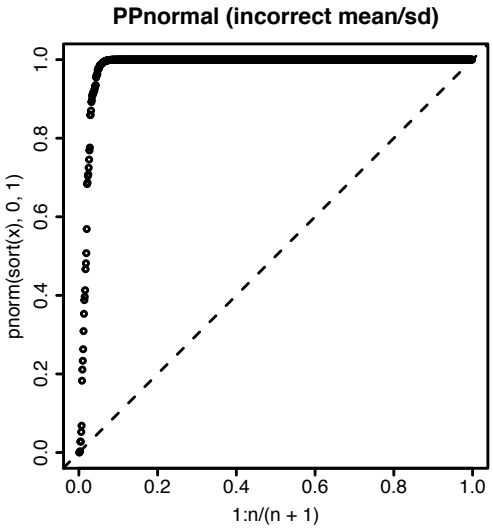

QQnormal (incorrect mean/sd)
sort(x)
qnorm(1:n/(n + 1))
−2
−1
0
1
2

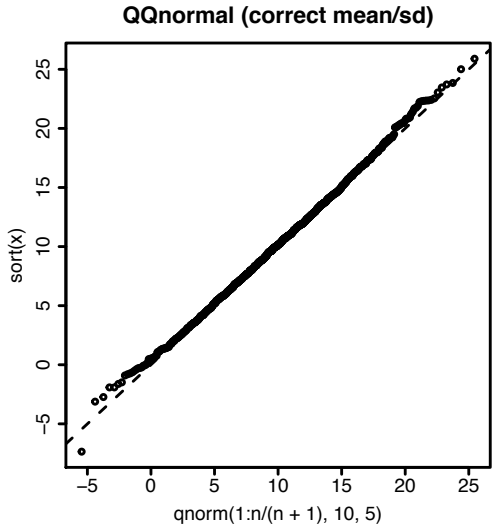

QQgamma
sort(z)
qnorm(1:n/(n + 1))
5
10
15
20
25
30
35

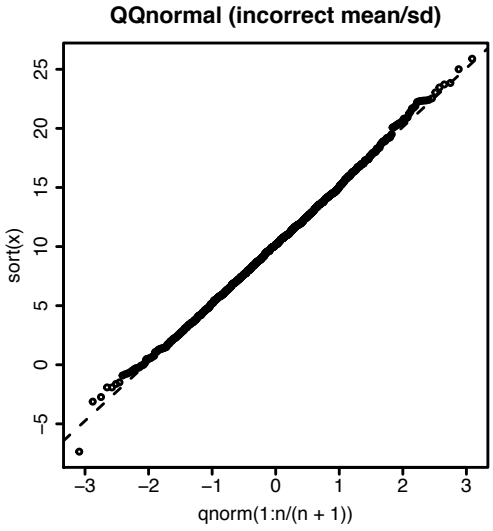

QQCauchy
sort(y)
qnorm(1:n/(n + 1))
0
100
200
300

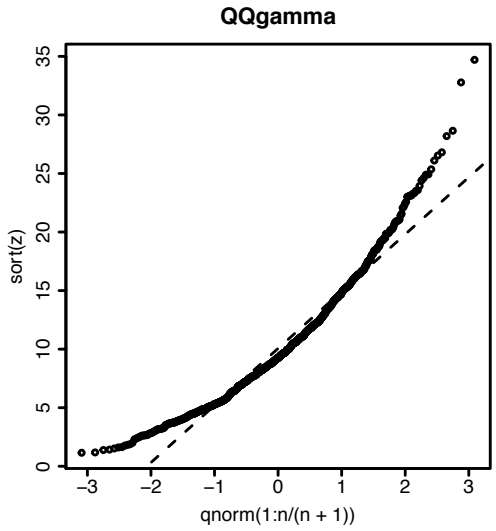
(omitted repeated code) $n = 1000$

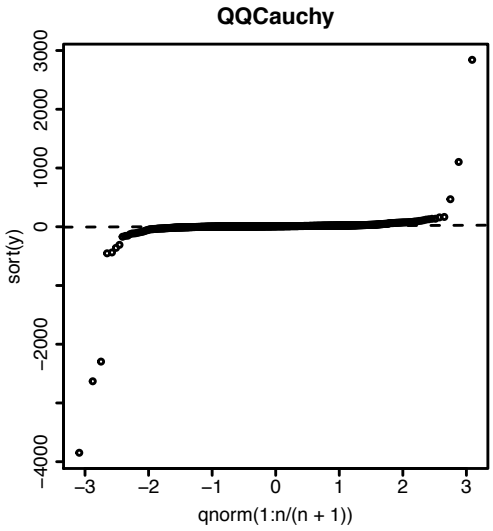












Extrema and Order Statistics

Let $X_1, \ldots, X_n$ be iid RVs with CDF F.

Let $U = X_{(n)} = \max\{X_1, \ldots, X_n\}$

$V = X_{(1)} = \min\{X_1, \ldots, X_n\}$

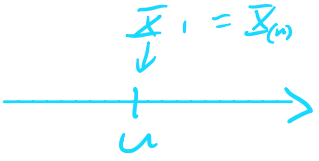
What are the dists of U and V?

$$\begin{aligned} F_U(u) &= P(U \le u) \\ &= P(X_1 \le u, \ldots, X_n \le u) \\ &= \prod_1^n P(X_i \le u) \\ &= [F(u)]^n \end{aligned}$$

If F has a density f, then by chain rule so does F_U and

$$f_U(u) = n[F(u)]^{n-1} f(u)$$

Intuitively: "$f_U(u)\,du$" = "$f(u)\,du$"



$\times [F(u)]^{n-1}$

$\times n$ (which $i = (n)$)

Similarly,

$$\begin{aligned} F_V(v) &= 1 - P(V > v) \\ &= 1 - P(X_1 > v, \ldots, X_n > v) \\ &= 1 - \prod_1^n P(X_i > v) \\ &= 1 - [1 - F(v)]^n \end{aligned}$$

If F has a density f, then by chain rule

$$f_V(v) = n\,[1 - F(v)]^{n-1} f(v) \quad .$$

More generally, let $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ be the sorted values of $X_1, \dots, X_n$.

The $X_{(i)}$ are called the order statistics, so $X_{(1)}$ is the first, $X_{(i)}$ is the i^{th}, and $X_{(n)}$ is the last order statistic.

Claim. If F is a cont. function (so $\forall x \; P(X=x)=0$)
then with prob. 1 $X_{(1)} < X_{(2)} < \dots < X_{(n)}$.

Proof. exercise (easy if F has a density)

What is the dist. of $X_{(k)}$?