Lecture 2

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 8, 2020

Table of contents

Outline

- Define probability calculus
- Basic axioms of probability
- Define random variables
- Define density and mass functions
- Define cumulative distribution functions and survivor functions
- Define quantiles, percentiles, medians

Probability measures

A **probability measure**, P, is a real valued function from the collection of possible events so that the following hold

1. For an event $E \subset \Omega$, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1$
3. If E_i, $i = 1, \ldots, \infty$ are mutually exclusive events ($E_i \cap E_j = \emptyset$ for every $i \neq j$)

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

This is called **countable additivity**

Additivity

Part 3 of the definition implies **finite additivity**

$$P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$$

where $\{E_i\}$, $i = 1, \ldots, n$ are mutually exclusive

For $n = 2$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

where $E_1 \cap E_2 = \emptyset$

Note

- P is defined on $\mathcal{F}$ a collection of subsets of Ω
- Example $\Omega = \{1, 2, 3\}$ then

$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

- The number of possible events, $|\mathcal{F}| = 2^3$ when $|\Omega| = 3$
- When Ω is a continuous set, the definition gets much trickier. In this case we assume that $\mathcal{F}$ is sufficiently rich so that any set that we're interested in will be in it

Consequences

You should be able to prove all of the following:

- $P(\emptyset) = 0$
- $P(E) = 1 - P(E^c)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- if $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = 1 - P(A^c \cap B^c)$
- $P(A \cap B^c) = P(A) - P(A \cap B)$
- $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$
- $P(\cup_{i=1}^n E_i) \geq \max_i P(E_i)$

Example

Proof that $P(E) = 1 - P(E^c)$

$$\begin{aligned} 1 &= P(\Omega) \\ &= P(E \cup E^c) \\ &= P(E) + P(E^c) \end{aligned}$$

■

Example

Proof that $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$

$$\begin{aligned} P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &\leq P(E_1) + P(E_2) \end{aligned}$$

Assume the statement is true for $n-1$ and consider n

$$\begin{aligned} P(\cup_{i=1}^n E_i) &\leq P(E_n) + P(\cup_{i=1}^{n-1} E_i) \\ &\leq P(E_n) + \sum_{i=1}^{n-1} P(E_i) \\ &= \sum_{i=1}^{n} P(E_i) \end{aligned}$$

■

Example

The National Sleep Foundation (`www.sleepfoundation.org`) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Similarly, they report that 58% of adults in the US experience insomnia. Does this imply that 71% of people will have at least one sleep problems of these sorts?

Example continued

Answer: No, the events are not mutually exclusive. To elaborate let:

$$\begin{aligned} A_1 &= \{\text{Person has sleep apnea}\} \\ A_2 &= \{\text{Person has RLS}\} \\ A_3 &= \{\text{Person has insomnia}\} \end{aligned}$$

Then (work out the details for yourself)

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &+ P(A_1 \cap A_2 \cap A_3) \\ &= .71 + \text{Other stuff} \end{aligned}$$

where the "Other stuff" has to be less than 0

Example: LA Times from Rice page 26

The LA Times reported that the risk of HIV infection was $\approx 1/500$ for a single act of intercourse. They conclude that 500 acts of intercourse yields a 100% probability of infection. **Implication:** I have an oddly shaped coin whose probability of a head is $1/500$. According to LA times, if I flip the coin twice, there is a $2/500$ chance of getting a head on at least one of those tosses.

$$\begin{aligned} A_1 &= \{\text{Head on toss } 1\} \\ A_2 &= \{\text{Head on toss } 2\} \\ B &= \{\text{Head on toss 1 and } 2\} = A_1 \cup A_2 \end{aligned}$$

$$P(B) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = \frac{2}{500} - P(A_1 \cap A_2)$$

Example: Birthday problem

In a given room what is the probability of at least two people having the same birthday? Assume that birthdays occur randomly throughout the year.

Calculation: $P(\text{at least two}) = 1 - P(\text{none})$

2 people

$$P(\text{at least two}) = 1 - \frac{365}{365} \times \frac{364}{365} = 0.0027$$

3 people

$$P(\text{at least two}) = 1 - \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} = 0.008$$

n people $(n < 365)$

$$P(\text{at least two}) = 1 - \frac{365 \times 364 \times \ldots (365 - n + 1)}{365^n}$$

Example: Birthday problem

R code:

```
n=1:364
pn=n
for (i in 1:364)
   {pn[i]<- 1-prod(365:(365-i+1))/365^i}
```

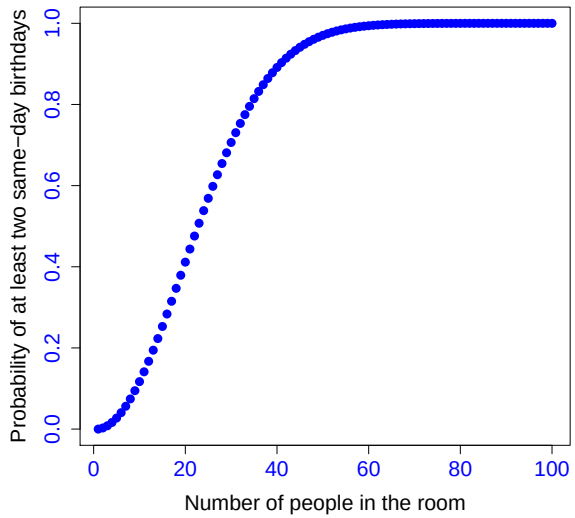
Some results

```
pn[23]
[1] 0.5072972

pn[57]
[1] 0.9901225
```

Example: R plotting

```
plot(n[1:100],pn[1:100],type="p",pch=19,
   col="blue",lwd=3,
   xlab="Number of people in the room",
   ylab="Probability of at least two same-day
        birthdays",
   cex.lab=1.5,cex.axis=1.5,col.axis="blue")
```



Example: The Monte Hall problem

There are three doors with a prize behind one and worthless joke prizes behind the other two. The contestant selects a door. Then Monte Hall shows the contents of one of the remaining two (Monte Hall never opens the door to actual prize.) After showing the contents, the contestant is asked if they would like to switch. Should they?

Answer: YES

Suppose you choose door number 1

	Scenario 1	Scenario 2	Scenario 3
Door 1	Prize	Goat	Goat
Door 2	Goat	Prize	Goat
Door 3	Goat	Goat	Prize
Switch	Loose	Win	Win

Using R to sample

```
# a random permutation
   x <- sample ( 1:6)
# sampling with replacement
   x <- sample ( 1:6, 10, replace=T )
# how many are equal to 3?
   sum ( x == 3 )
# nonparametric bootstrap
   x <- sample ( 1:10, 10, replace=T )
# simulating the number of Come out rolls at a game of craps
   wins <- rep ( 0, 1000 )
   for ( i in 1:1000 )
      {d <- sample ( 1:6, 2, replace=T )
      if ( sum(d)== 7 || sum(d) == 11 )
      wins[i] <- 1}
      sum ( wins )
```

Random variables

- A **random variable** is the function that assigns an outcome of an experiment given the design and implementation characteristics of that experiment
- The random variables that we study will come in two varieties, **discrete** or **continuous**
- Discrete random variable are random variables that take on only a countable number of possibilities
  - $P(X = k)$
- Continuous random variable can take any value on the real line or some subset of the real line
  - $P(X \in A)$

Examples of random variables

- The $(0-1)$ outcome of the flip of a coin
- The outcome from the roll of a die
- The BMI of a subject four years later
- The HTN status of a study participant

PMF

A probability mass function evaluated at a value corresponds to the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy

1. $p(x) \geq 0$ for all x
2. $\sum_x p(x) = 1$

The sum is taken over all of the possible values for x.

Example: Bernoulli

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads.

$$p(x) = (1/2)^x(1/2)^{1-x} \quad \text{for} \quad x = 0, 1$$

Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head

$$p(x) = \theta^x(1-\theta)^{1-x} \quad \text{for} \quad x = 0, 1$$

```
rbinom(20,1,0.5)
rbinom(20,1,seq(0,1,length=20))
?rbinom
```

Example: Poisson

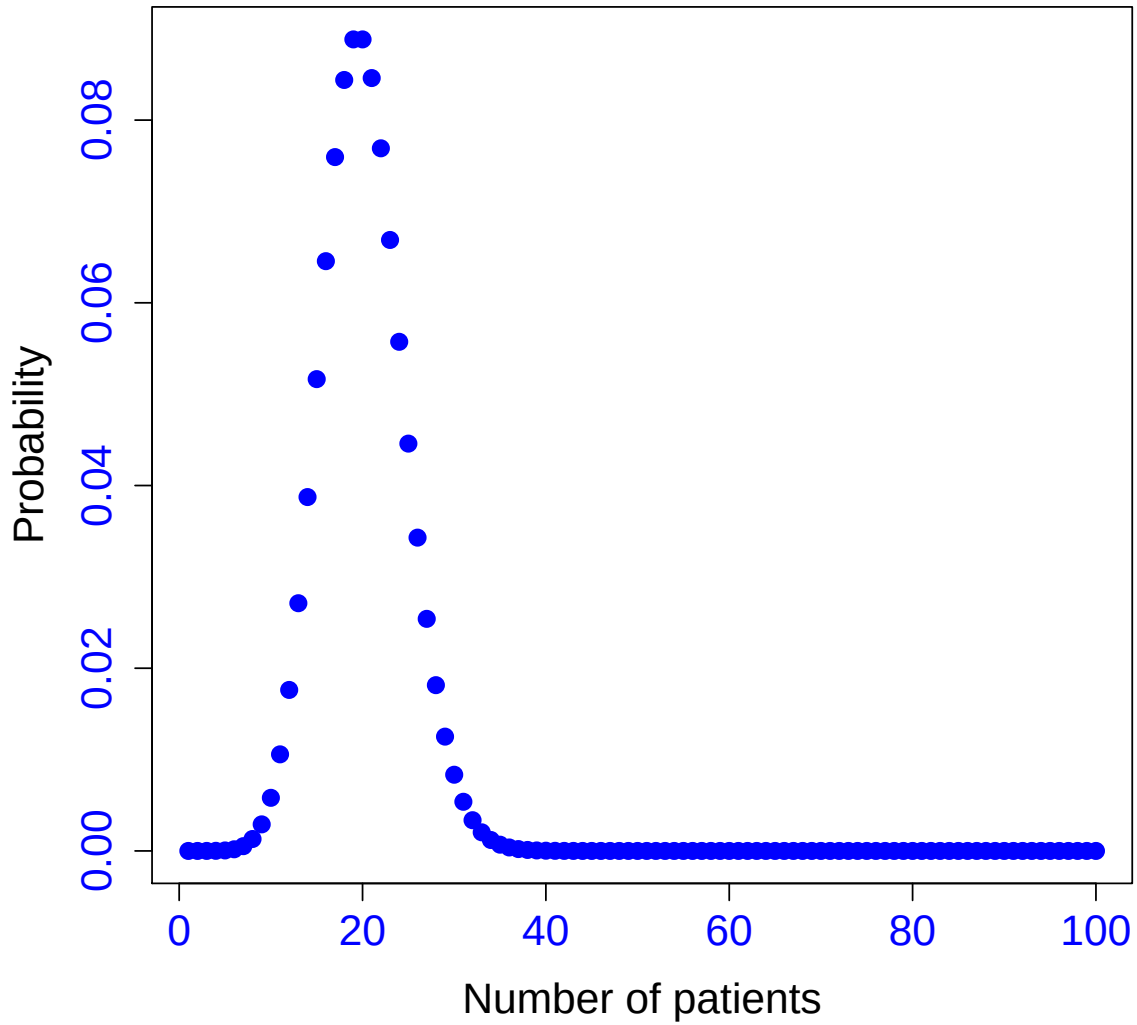
X: number of patients arriving at a clinic on a given day

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for} \quad k = 0, 1, \ldots$$

λ: average number of patients over (∞) number of days
Calculating the PMF and plotting it

```
x=1:100
lambda=20
plot(x,dpois(x,lambda),type="p",pch=19,
col="blue",lwd=3, xlab="Number of patients",
ylab="Probability",cex.lab=1.5,cex.axis=1.5,
col.axis="blue")
```

Poisson PMF


Probability
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
Number of patients

Sampling the Poisson distribution

A sample of 15 independent days with average number of patients $= 20$

```
rpois(15,20)
[1] 25 23 22 24 27 8 19 26 17 14 14 20 16 19 16
```

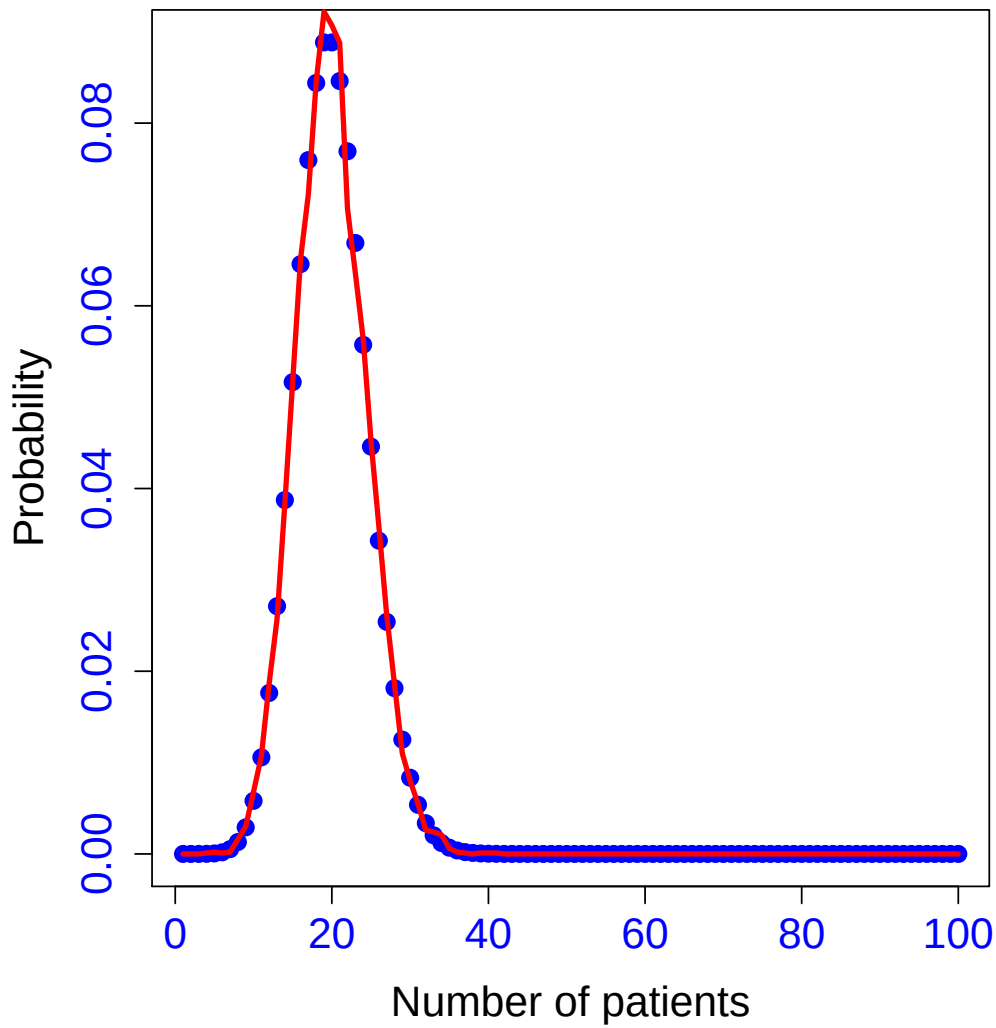
Reconstructing PMF from samples

```
y<-rpois(10000,20)
py=rep(0,100)
for (i in 1:100)
   {py[i]<-mean(y==i)}
```

Add lines to the PMF plot of the Poisson distribution

```
lines(1:100,py,col="red",lwd=3)
```

Reconstructing the Poisson PMF



PDF

A probability density function (pdf), is a function, say $f(\cdot)$, associated with a continuous random variable

> *Areas under pdfs correspond to probabilities for that random variable*

$$P(X \in I) = \int_I f(x)dx$$

To be a valid pdf, the function $f(\cdot)$ must satisfy

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example: Exponential

Assume that the time in years from diagnosis until death of persons with a specific kind of cancer follows a density like

$$f(x) = \begin{cases} \frac{e^{-x/5}}{5} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

More compactly written: $f(x) = \frac{1}{5} e^{-x/5}$ for $x > 0$.
Is this a valid density?

1. e raised to any power is always positive
2. $$\int_0^\infty f(x)dx = \int_0^\infty e^{-x/5}/5 dx = -e^{-x/5}\Big|_0^\infty = 1$$

Example continued

What's the probability that a randomly selected person from this distribution survives more than 6 years?

$$P(X \geq 6) = \int_6^\infty \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_6^\infty = e^{-6/5} \approx .301.$$

Approximation in R

```
pexp(6, 1/5, lower.tail = FALSE)
```

Example continued

Example continued

What's the probability that a randomly selected person from this distribution survives strictly more than 5 and strictly less than 6 years?

$$P(5 < X < 6) = \int_5^6 \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_5^6 = e^{-1} - e^{-6/5}$$

Approximation in R

`pexp(6, 1/5)-pexp(5, 1/5)` $\approx$ `0.067`

This is equal to the probability that the person survives at least 5 years, but no more than 6 years

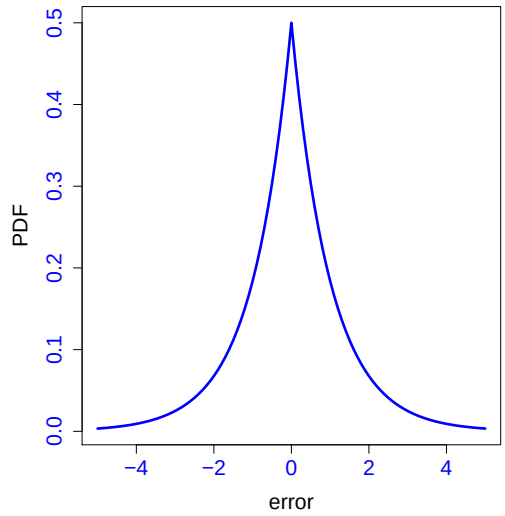
$$P(5 < X < 6) = P(5 \leq X \leq 6)$$

Example: Double Exponential

$$f(x) = \frac{1}{2}\exp(-|x|) \text{ for } -\infty < x < \infty$$

Plotting the pdf of a double exponential

```
x=seq(-5,5,length=101)
fx=exp(-abs(x))/2
plot(x,fx,type="l",col="blue",lwd=3,
xlab="error",ylab="PDF",cex.lab=1.5,
cex.axis=1.5,col.axis="blue")
```



Example continued

Is this a proper pdf?

1. $f(x) = \frac{1}{2}\exp(-|x|) > 0$ for any $x \in (-\infty, \infty)$
2. $$\begin{aligned} \int_{-\infty}^{\infty} \frac{1}{2}\exp(-|x|)dx &= 2\int_{0}^{\infty} \frac{1}{2}\exp(-x)dx \\ &= -\exp(-x)|_0^{\infty} = 1 \end{aligned}$$

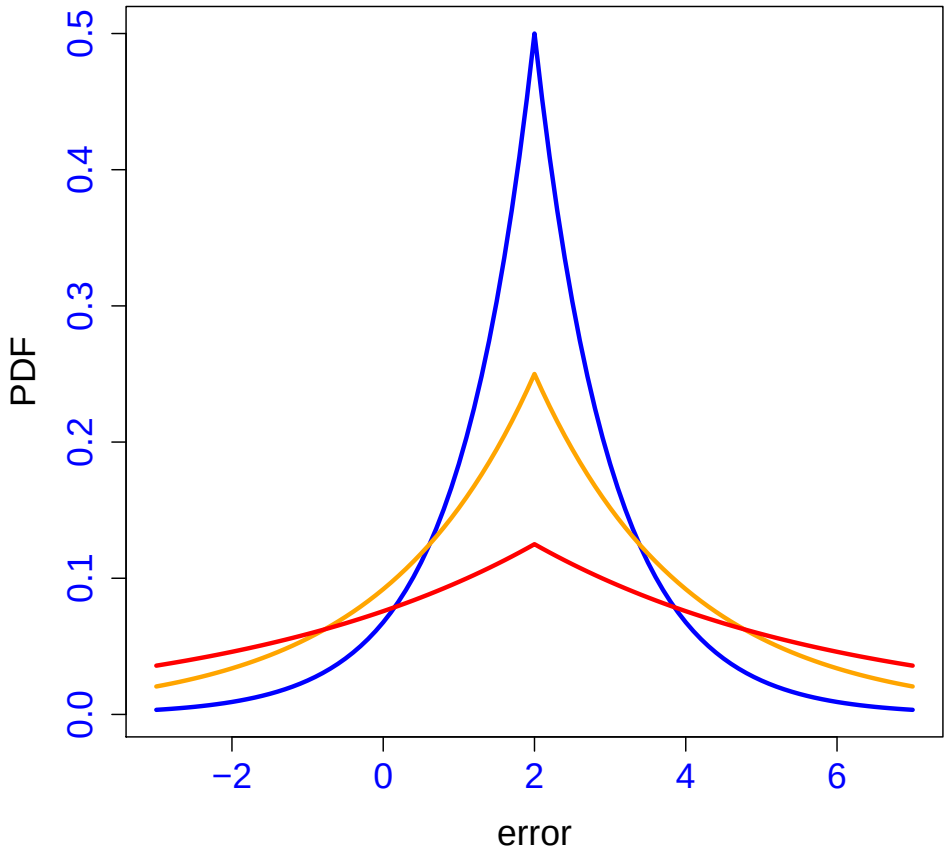
Consider a more general version of $f(\cdot)$

$$f(x) = \frac{1}{2\sigma}\exp\left(-\frac{|x-\mu|}{\sigma}\right) > 0 \text{ for any } x \in (-\infty, \infty)$$

Is this a valid density?

Example: Double Exponential

Plotting double exponentials with mean $\mu = 2$ and $\sigma = 1, 2, 4$



The Double Exponential in R

Install the R package vgam

Construct the empirical pdf from 100 simulated observations

```
mu=2
sigma=2
y=rlaplace(100,mu,sigma)
hist(y,breaks=20,probability=TRUE,xlab="error")
```

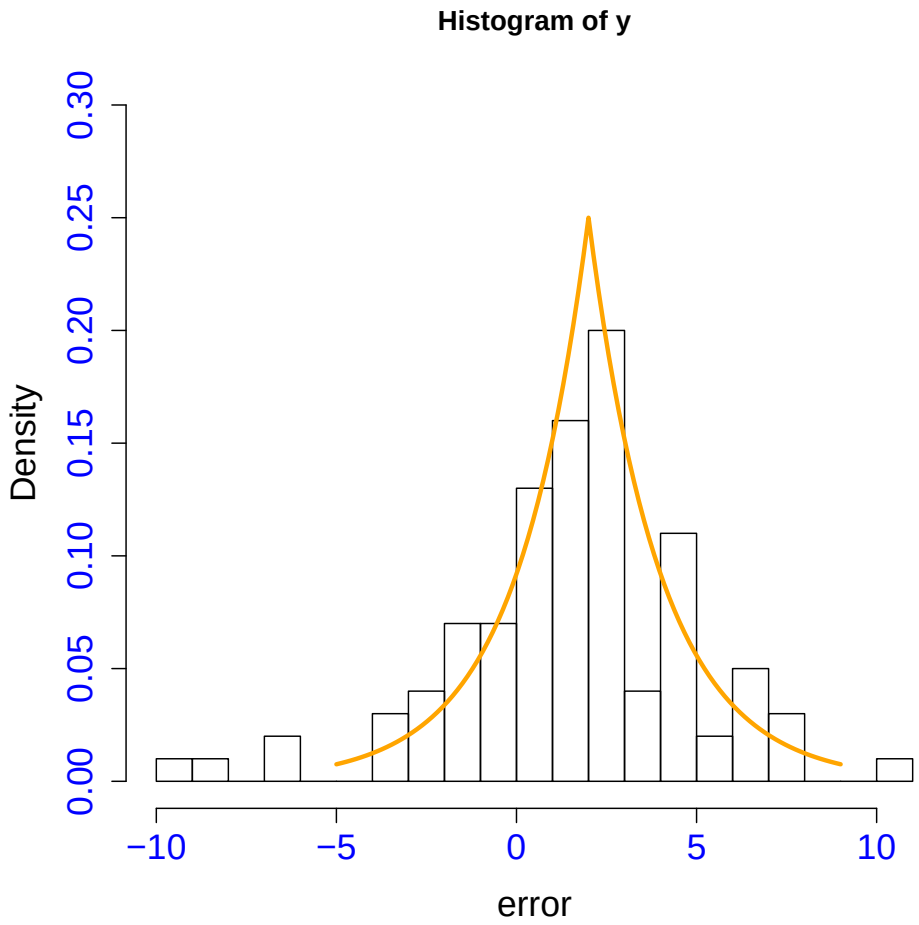
Super impose the theoretical pdf

```
x=seq(mu-7,mu+7,length=101)
fx=exp(-abs(x-mu)/sigma)/(2*sigma)
lines(x,fx,col="orange",lwd=3)
```

Can you do this using only the exponential and the Bernoulli?

Double Exponential PDF

Empirical and theoretical double exponential distributions ($\mu = 2, \sigma = 2$)



CDF and survival function

- The **cumulative distribution function** (CDF) of a random variable X is defined as the function

$$F(x) = P(X \leq x)$$

- This definition applies regardless of whether X is discrete or continuous.
- The **survival function** of a random variable X is

$$S(x) = P(X > x)$$

- Notice that $S(x) = 1 - F(x)$
- For continuous random variables, the PDF is the derivative of the CDF: $f(x) = F'(x)$

Example

What are the survival function and CDF for the exponential density considered before?

$$S(x) = \int_x^\infty \frac{e^{-t/5}}{5} dt = -e^{-t/5}\Big|_x^\infty = e^{-x/5}$$

hence we know that

$$F(x) = 1 - S(x) = 1 - e^{-x/5}$$

Notice that we can recover the PDF by

$$f(x) = F'(x) = \frac{d}{dx}(1 - e^{-x/5}) = e^{-x/5}/5$$

Quantiles

- The α^{th} **quantile** of a distribution with CDF F is the point x_α so that
$$F(x_\alpha) = \alpha$$
- A **percentile** is simply a quantile with α expressed as a percent
- The **median** is the 50^{th} percentile
- You may also hear: tertiles, quartiles, quintiles, deciles

Example

- What is the 25^{th} percentile of the exponential survival distribution considered before?
- We want to solve (for x)

$$\begin{aligned} .25 &= F(x) \\ &= 1 - e^{-x/5} \end{aligned}$$

  resulting in the solution $x = -\log(.75) \times 5 \approx 1.44$
- Therefore, 25% of the subjects from this population live less than 1.44 years
- R can approximate exponential quantiles for you

```
qexp(.25, 1/5)
```

QQ plots

Calculate the quantiles of a exp(5) distribution

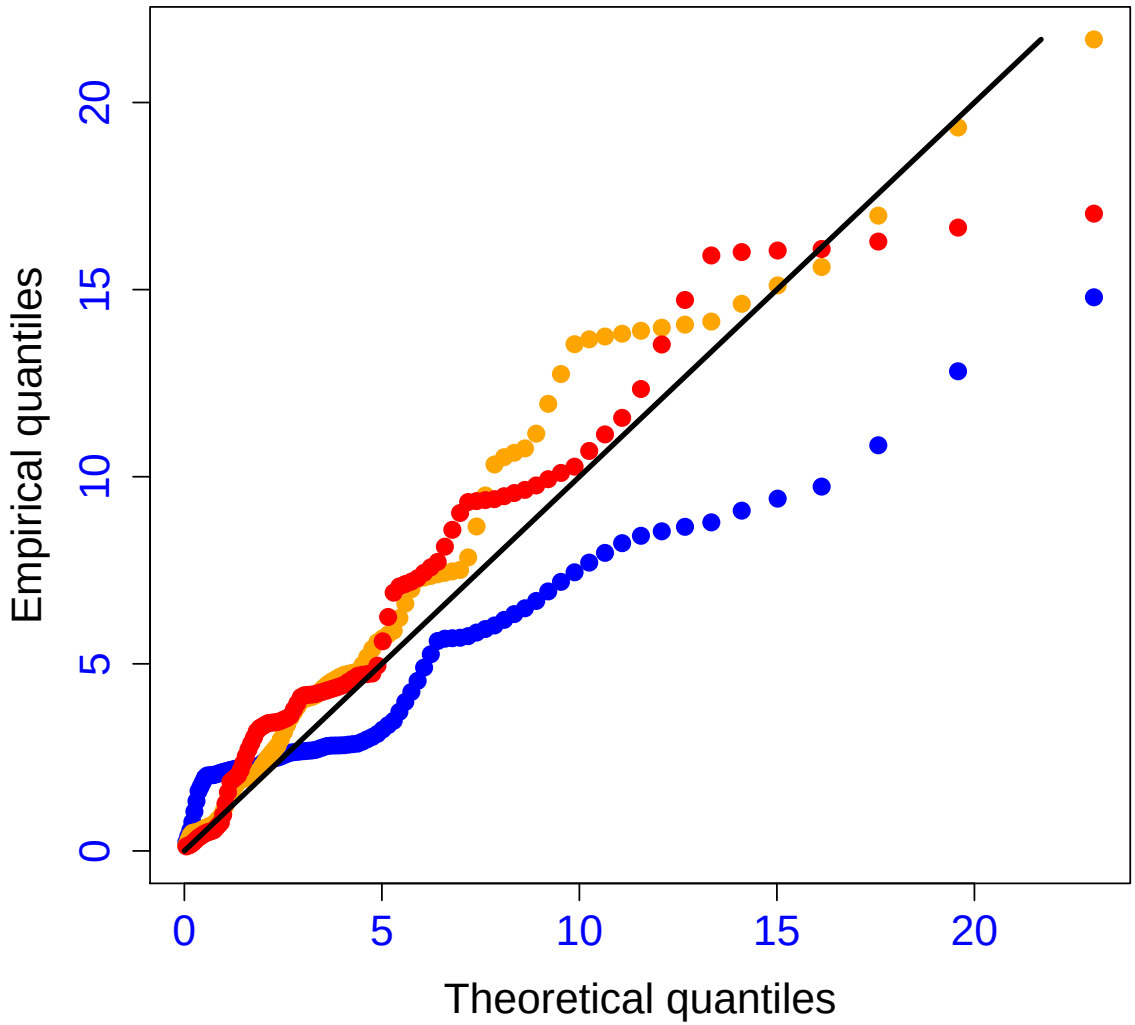
```
tq<-qexp(seq(0.01,0.99,length=100), 1/5)
```

How do they compare to the empirical quantiles?

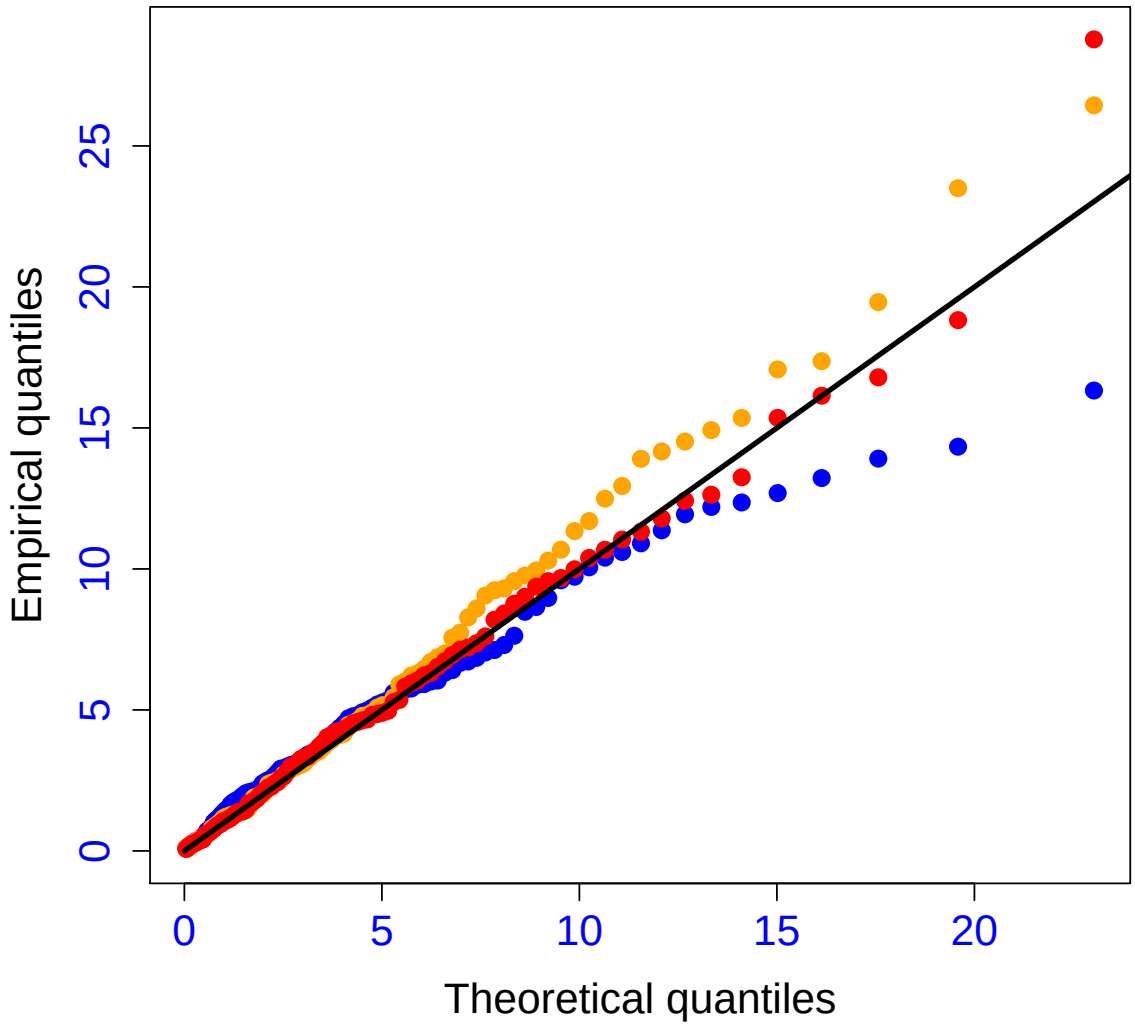
```
x1<-rexp(30, 1/5)
x2<-rexp(30, 1/5)
x3<-rexp(30, 1/5)
eq1<-quantile(x1,seq(0.01,0.99,length=100))
eq2<-quantile(x2,seq(0.01,0.99,length=100))
eq3<-quantile(x3,seq(0.01,0.99,length=100))
```

Plot theoretical versus empirical quantiles (QQ plots)

QQ plots $n = 30$, exp(5)



QQ plots $n = 200$, exp(5)



QQ plots $n = 500$, exp(5)