# Some of the common univariate probability distributions

| Continuous | Discrete |
|---|---|
| Normal | Bernoulli |
| Student's t | Binomial |
| F | Geometric |
| Chi-squared | Hypergeometric |
| | Negative Binomial |
| | Poisson |

## Normal distribution N(μ, σ)

- Also called Gaussian,
- Defined by 2 parameters - mean and standard deviation: N(μ, σ)
- Standard normal distribution N(0,1), area under the curve = 1
- Z-score of an observation -> how many standard deviations it falls above or below the mean (if the observation is 1 SD above the mean, it's z-score is 1, if it is 1.5 SD below the mean, it's z-score is -1.5)

$$Z = \frac{x - \mu}{\sigma}$$

We can use normal curve to find percentiles. Here is an applet that you can use to enter z-score for an observation, and it will show you area under the curve for that interval.

Approximately 68%, 95%, and 99.7% of observations fall within 1,2, and 3 standard deviations from the mean in the normal distribution

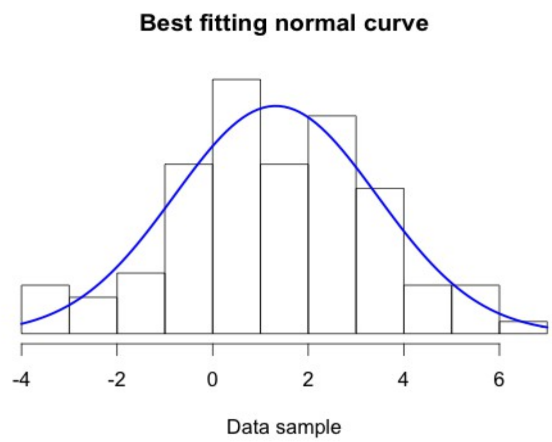
### *Testing the appropriateness of the normal assumption*

1. Histogram with the best fitting normal curve overlaid on the plot
   - Create histogram with the best fitting normal curve overlaid on the plot (sample mean and standard deviation are used as the parameters of the best fitting normal curve).
   - The closer the curve fits the histogram, more likely the normal distribution

In R:

- generate normally distributed sample
- draw histogram of the data and overlay best fitting normal curve

```
sample <- rnorm(100, mean=1, sd=2)
hist(sample, prob=TRUE, yaxt='n', main="Best fitting normal curve", ylab="", xlab="Data sample")
curve(dnorm(x, mean=mean(sample), sd=sd(sample)), add=TRUE, col="blue", lwd=2.5)
```

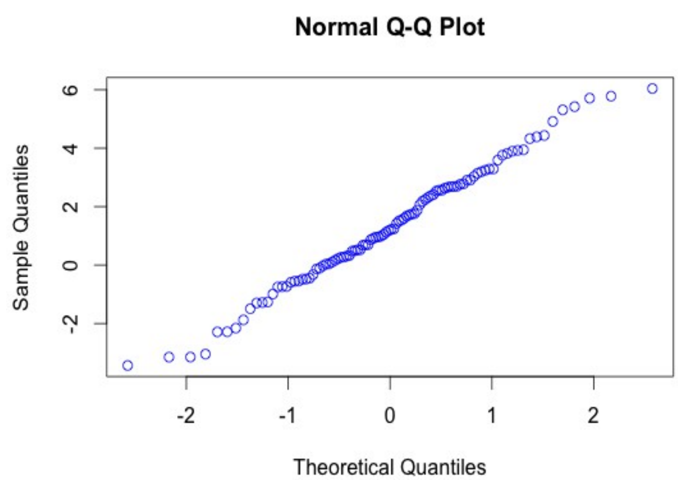



2. Examine normal probability plot (also called quantile-quantile plot in the case of normal distribution)
   - The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality
   - closer the points to a perfect straight line, more likely the normal distribution.

In R:

1. create normally distributed data sample
2. create normal probability plot (qq plot)

```
sample <- rnorm(100, mean=1, sd=2)
qqnorm(sample, col="blue")
```



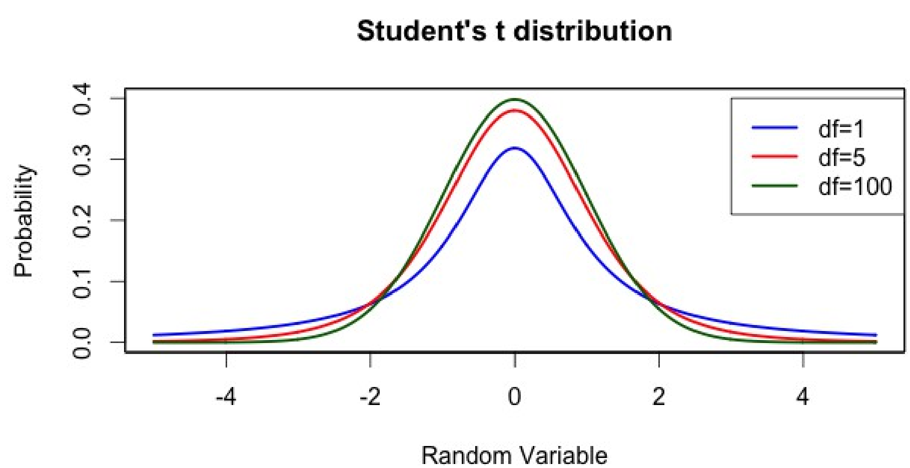
# Student's t – distribution

- Parameter degrees of freedom (df) – describes the shape of the distribution
- Used instead of normal distribution when the sample size is small and population standard deviation is unknown.
- For df ≥ 30, the *t* distribution is nearly indistinguishable from the normal distribution.

In R:

- Create Student's *t* – distribution with 1, 5 and 100 degree of freedom.

```
x <- seq(-5,5, by=0.01)
probt <- dt(x, df=1)
plot(x, probt, type='l', col='blue', main = "Student's t distribution",
     xlab = "Random Variable",xlim = c(-5,5), ylim=c(0, 0.4), lwd = 2.5, ylab = "Probability")
par(new=TRUE)
probt <- dt(x, df=5)
plot(x, probt, type='l', col='red', xlab = "", ylab='',xaxt = "n", yaxt="n",
     lwd = 2.5, xlim = c(-5,5), ylim=c(0, 0.4))
par(new=TRUE)
probt <- dt(x, df=100)
plot(x, probt, type='l', col='dark green', xlab = "", ylab='', xaxt = "n", yaxt="n",lwd = 2.5,
       xlim = c(-5,5), ylim=c(0, 0.4))
legend(3, 0.4, c("df=1", "df=5", "df=100"), lty=1, lwd= 2.5,col=c("blue","red", "dark green"))
```



# Chi-squared Distribution $\chi^2(k)$

If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom:

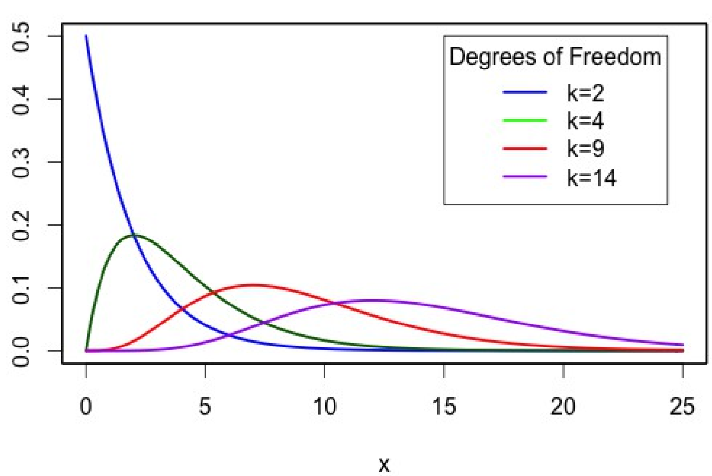
$$Q \sim \chi^2(k)$$

It has just one parameter (degrees of freedom -k). As the degrees of freedom increases distribution becomes more symmetric, center moves to the right, and variability increases.

In R:

- Chi-square distribution with 2,4,9 and 14 degrees of freedom:

```
curve(dchisq(x, 2), xlim = c(0,25), ylim=c(0, 0.5), col='blue',
      lwd = 2.5, ylab='')
par(new=TRUE)
curve(dchisq(x, 4), xlim = c(0,25), ylim=c(0, 0.5),col='dark green',
      lwd = 2.5, xaxt = "n", yaxt="n", xlab = "", ylab='')
par(new=TRUE)
curve(dchisq(x, 9), xlim = c(0,25), ylim=c(0, 0.5),col='red',
      lwd = 2.5, xaxt = "n", yaxt="n", xlab = "", ylab='')
par(new=TRUE)
curve(dchisq(x, 14), xlim = c(0,25), ylim=c(0, 0.5),col='purple',
      lwd = 2.5, xaxt = "n", yaxt="n", xlab = "", ylab='')
legend(15, 0.5, title = "Degrees of Freedom",
       c("k=2", "k=4", "k=9", "k=14"),
       lty=1, lwd= 2.5,col=c("blue","green", "red", "purple"))
```



# Bernoulli Distribution Bern(p)

- Has exactly two outcomes (a success - often denoted by 1, and failure - denoted by 0), and 1 trial
- If x is a random variable that takes value 1, with probability of success p, and 0 with probability 1-p, then x is a Bernoulli random variable, with mean and standard deviation

$$\mu = p \qquad \sigma = \sqrt{p(1-p)}$$

# Geometric Distribution

Describes the waiting time until a success for repeated Bernoulli trials (probability distribution of the number of *X* Bernoulli trials needed to get one success)
Trials are independent
If p denotes probability of success, then the probability of finding the first success on $n^{th}$ trial:

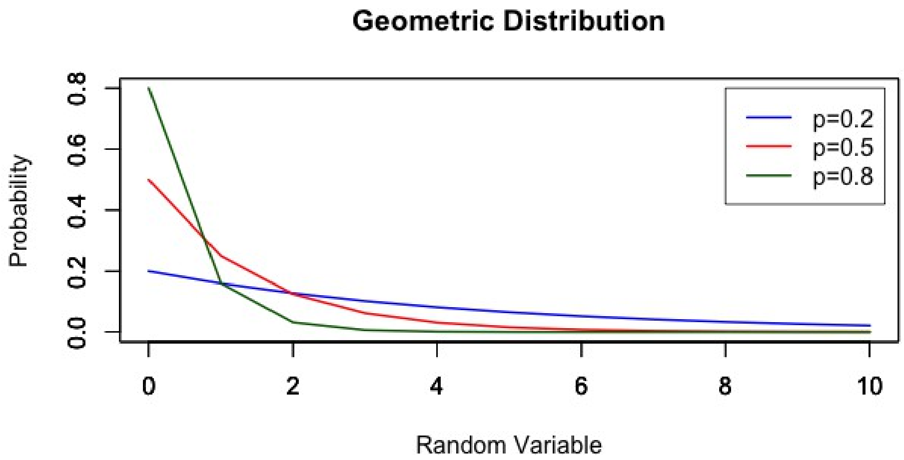
$$(1-p)^{n-1} p$$

Mean and standard deviation:

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

In R:

```
x <- seq(0,10, by=1)
probg <- dgeom(x, 0.2)
plot(x, probg, type='l', col='blue', lwd = 2, main = "Geometric Distribution",
       xlab = "Random Variable", ylab="Probability", ylim=c(0, 0.8),)
par(new=TRUE)
probg<- dgeom(x, 0.5)
plot(x, probg, type='l', col='red', lwd = 2, xlab="",ylab ="", ylim=c(0, 0.8))
par(new=TRUE)
probg <- dgeom(x, 0.8)
plot(x, probg, type='l', col='dark green', lwd =2, xlab="",ylab="", ylim=c(0, 0.8))
legend(8, 0.8, c("p=0.2","p=0.5","p=0.8"), lty=1, lwd= 2,col=c("blue","red", "dark green"))
```




# Binomial Distribution B(n, p)

- Parameters:
  n – number of trails
  p – success probability in each trial

- Binomial distribution describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p.
- So, conditions for using Binomial distribution:
  1. there is a fixed number, *n*, of identical trials
  2. for each trial there are two possible outcomes (success/failure)
  3. probability of success, *p*, remains constant for each trial
  4. trails are independent
  5. k = number of successes observed for n trials

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

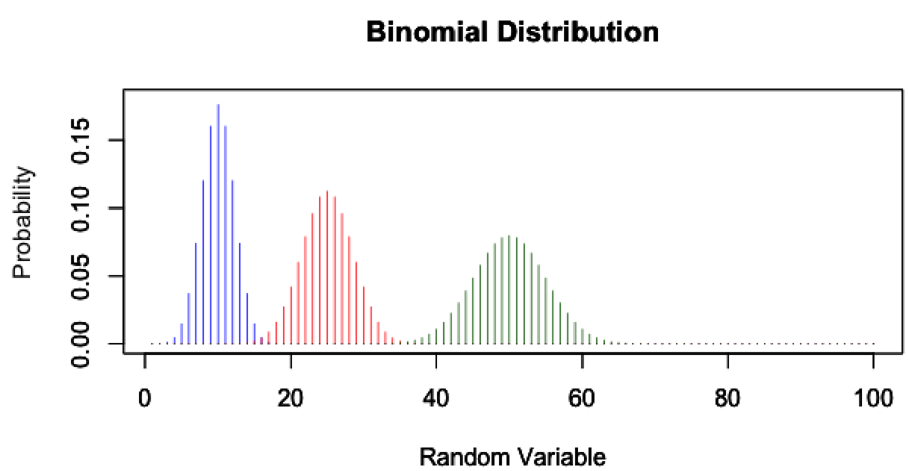
Mean and standard deviation of number of observed successes:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

In R:

- Generate binomial distribution and draw histograms for 20, 50 and 100 trials and probability of 0.5

```
x <- seq(1,100, by=1)
probx <- dbinom(x, 20, 0.5)
plot(x, probx, type='h', col='blue', main = "Binomial Distribution", xlab = "Random Variable",
       ylab="", ylim=c(0, 0.18))
par(new=TRUE)
probx <- dbinom(x, 50, 0.5)
plot(x, probx, type='h', col='red', main = "Binomial Distribution", xlab = "Random Variable",
       ylab="", ylim=c(0, 0.18))
par(new=TRUE)
probx <- dbinom(x, 100, 0.5)
plot(x, probx, type='h', col='dark green', main = "Binomial Distribution", xlab = "Random
       Variable", ylab="", ylim=c(0, 0.18))
```




## Negative Binomial Distribution NB(r, p)

- Parameters:
  r > 0 — number of failures until the experiment is stopped
  p - probability that individual trial is a success (trials are assumed to be independent)
- Describes the probability of observing the k$^{th}$ success on the n$^{th}$ trial:

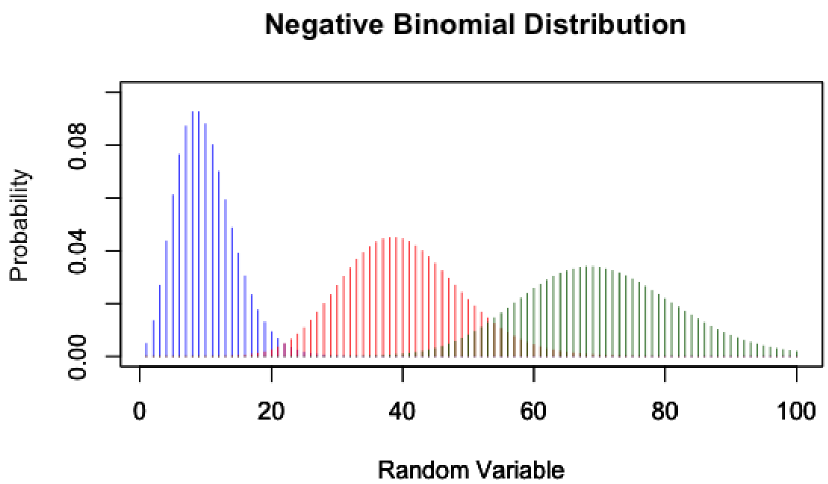
$$\binom{n-k}{k-1} p^k (1-p)^{n-k}$$

k – number of successes (fixed number)
n – number of trials (unlike binomial distribution, it's not fixed, it's a random variable )

In R:

- Create negative binomial distribution with 10, 40 and 70 successful trials:

```
x <- seq(1,100, by=1)
probx <- dnbinom(x, 10, 0.5)
plot(x, probx, type='h', col='blue', main = " Negative Binomial Distribution", xlab = "Random
Variable", ylab="", ylim=c(0, 0.1))
par(new=TRUE)
probx <- dnbinom(x, 40, 0.5)
plot(x, probx, type='h', col='red', xlab = "Random Variable", ylab="", ylim=c(0, 0.1))
par(new=TRUE)
probx <- dnbinom(x, 70, 0.5)
plot(x, probx, type='h', col='dark green', yaxt='n', xlab = "Random Variable",
ylab="Probability", ylim=c(0, 0.1))
```




# Hypergeometric Distribution

- arises when sampling is performed from a finite population without replacement (thus making trials dependent on each other)
- describes the probability of g successes in n draws without replacement from a finite population of size N containing a maximum of G successes:

$$\frac{\binom{G}{g}\binom{N-G}{n-g}}{\binom{N}{n}}$$

# Poisson Distribution

Often useful for estimating the number of <u>rare events</u> in a <u>large population</u> over a unit of time.
Assumes that events are <u>independent</u>

$$P(observe\, k\, rare\, events) = \frac{\lambda^k e^{-\lambda}}{k!};$$

$\lambda$ – how many rare events do we expect to observe

- Mean and standard deviation

$$\mu = \lambda \qquad \sigma = \sqrt{\lambda}$$

In R:

- Create Poisson Distributions with means = 1, 5 and 10

```
x <- seq(0,100, by=1)
probp <- dpois(x, 1)
plot(x, probp, type='l', col='blue', main = "Poisson Distribution",xlab = "Number of Occurances",
     xlim = c(0,20), ylim=c(0, 0.4), lwd = 2.5, ylab = "Probability")
par(new=TRUE)
probp <- dpois(x, 5)
plot(x, probp, type='l', col='red', xlab = "", ylab='',xaxt = "n", yaxt="n",
     xlim = c(0,20), ylim=c(0, 0.4), lwd = 2.5)
par(new=TRUE)
probp <- dpois(x, 10)
plot(x, probp, type='l', col='dark green', xlab = "", ylab='', xaxt = "n", yaxt="n",
     xlim = c(0,20), ylim=c(0, 0.4),lwd = 2.5)
legend(15, 0.4, c(expression(lambda~"=1"),expression(lambda~"=5"),expression(lambda~"=10")),
       lty=1, lwd= 2.5,col=c("blue","red", "dark green"))
```