PROBABILITY AND DISTRIBUTIONS

Sorana D. Bolboacă

OBJECTIVES

- Bayesian theorem
- Binomial distribution
- Normal distribution
- Chi-square distribution
- Fisher distribution

Bayesian Theorem

- Tests are not events
- Tests detect things that do not exists (false positive) and could miss thinks that exists (false negative)
- Tests give us test probabilities NOT the real probability
- False positive skew results
- People prefer natural numbers as 100 in 10.000 rather than 1%
- Even science is a test

Bayesian Theorem

- Bayes' theorem finds the actual probability of an event from the results of a test
  - Correct the measurement errors
    - If you know the real probability & the chance of false positive and false negative
  - Relate the actual probability to the measured test probability.
    - Given mammogram test results and known error rates, you can predict the actual chance of having cancer

Anatomy of a test

- 2% of women have breast cancer
- 70% of mammograms detect breast cancer when it is present → 30% of mammograms miss the diagnosis
- 10% of mammograms incorrectly detect breast cancer when it is **not** present → 90% of mammograms correctly return a negative result

	Breast cancer + (2%)	Breast cancer - (98%)
Mammo +	70	10
Mammo-	30	90

Anatomy of a test: how to read it!

- 2% of women have breast cancer
- If you already have breast cancer, there is 70% chance that your mammogram will be positive and 30% chance that your test will to be negative
- If you do not have breast cancer, there is 10% chance that your mammogram will be positive and 90% chance that your test will to be negative

	Breast cancer + (2%)	Breast cancer - (98%)
Mammo +	70	10
Mammo-	30	90

Anatomy of a test

- Suppose that you have a patient with a positive result. What are her chances to have breast cancer?

	Breast cancer + (2%)	Breast cancer - (98%)
Mammo +	70	10
Mammo-	30	90

- The probability of a **true positive** = the probability to have breast cancer × probability that the mammogram to be positive = 0.02 × 0.7 = 0.014 → 1.4% chance
- The probability of a **false positive** = the probability not to have breast cancer × probability that the mammogram to be positive = 0. 98 × 0.10 = 0.098 → 9.8% chance

Anatomy of a test

	Breast cancer + (2%)	Breast cancer - (98%)
Mammo +	True positive 0.02*0.70 = 0.014	False positive 0.10*0.98=0.098
Mammo-	False negative 0.02*30 = 0.600	True negative 0.90*0.98 = 0.882

- What is the chance that your patient to really have cancer if she get a positive mammogram.
  - The chance of an event is the number of ways it could happen given all possible outcomes:
  - Probability = (desired event)/ (all possibilities)
  - Probability = 0.014/(0.014+0.098) = 0.125 → the chance that your patient to have breast cancer if the mammogram is positive is 12.5%

Anatomy of a test

- → a positive mammogram only means that the individual chance of breast cancer is 12.5%, rather than expected 70%
- → the mammogram gives a false positive 10% of the time → there will be false positives in any given population
- → the problem can be turned into an equation = Bayes' Theorem

Bayes' Theorem

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times P(A)}{\Pr(B \mid A) \times \Pr(A) + \Pr(B \mid nonA) \times \Pr(nonA)}$$

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \times P(A)}{\Pr(B)}$$

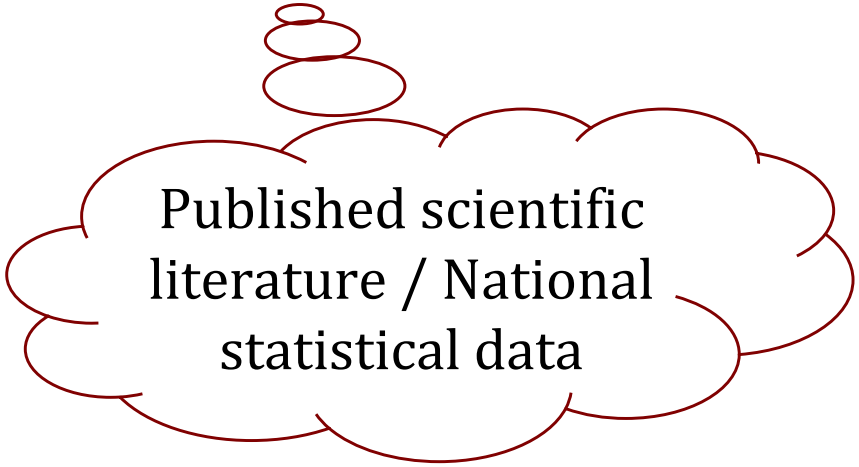
- P(A|B) = probability of having breast cancer (A) given a positive mammogram (B)
- P(B|A) = probability of a positive mammogram (B) given that breast cancer is present (A)
- P(A) = probability of having breast cancer → P(nonA) = probability not to have cancer
- P(B|nonA) = probability of a positive mammogram

Prior vs posterior probability

- Prior probability
  - what you believe before seeing any data
- Posterior probability
  - P(hypothesis|data) = the probability of a hypothesis given the data
  - depends on prior probability & observed data

Bayesian inference

- Prior probability + data → posterior probability



- P(hypothesis is true | observed data)

A good prior is helps, a bad prior hurts, but the prior matter less the more data you have!

Binomial distribution by example

- Describes the probability of having exactly *k* successes in *n* independent Bernouilli trials with probability of success *p*

  P(*k* successes in *n* trials) = COMBIN(n,k)×p^k×$(1-p)^{n-k}$

- How many scenarios yield 1 success in 3 trials?
  - SFF / FSF / FFS
  - n = 3 & k = 1 → =COMBIN(3,1) = 3

- How many scenarios yield 2 successes in 10 trials?
  - n = 10 & k = 2 → =COMBIN(10,2) = 45

Binomial conditions

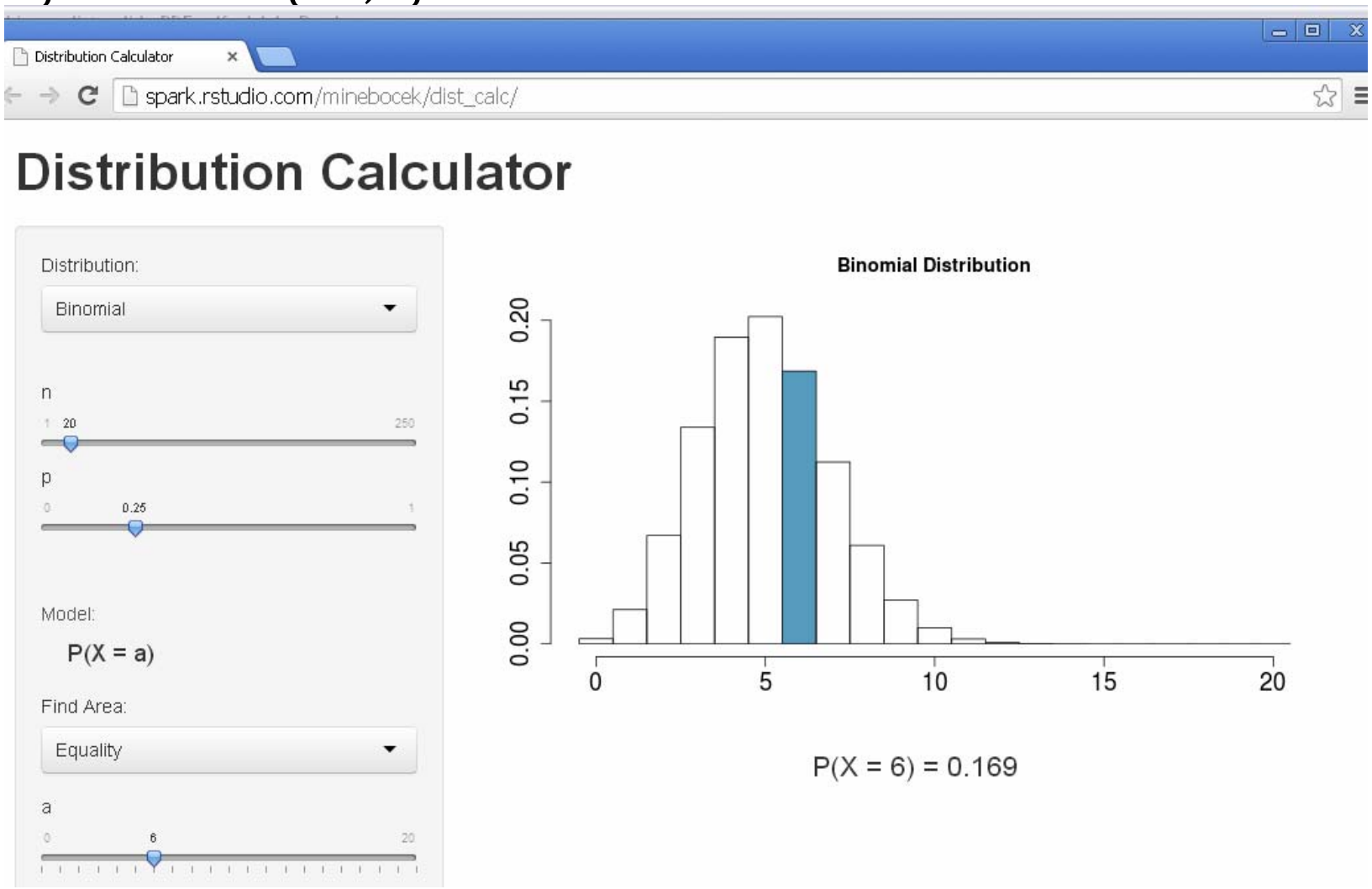
1. The trials are independent
2. The number of n trials must be fixed
3. Each trial outcome is classified as a success (p) or failure (q)
4. The probability of success is the same for each trial

Binomial example

- Generally, only 25% of medical students are engaged at classes (psychologically committed to the classes). Among a random sample of 20 students, what is the probability that 6 students are engaged at classes?
- n=20
- p=0.25 → q = 1-0.25 = 0.75
- k=6
- P(k=6) = combin(20,6)×0.25^{6}×$0.75^{(20-6)}$ = 0.1686

Binomial example

- n=20 | p=0.25 → q = 1-0.25 = 0.75 | k=6
- $P(k=6) = combin(20,6) \times 0.25^{6} \times 0.75^{(20-6)} = 0.1686$



Binomial example

- Among a random sample of 50 students. How many would is expected to be engaged at class? (p = 0.25)
- $\mu = 50 \times 0.25 = 12.5$
- $\Sigma = 50 \times 0.25 \times 0.75 = 9.375$

Expected value (mean) of binomial distribution: $\mu = n \times p$

Standard deviation of binomial distribution: $\sigma = \sqrt{n \times p \times q}$

Normal vs. binomial distribution

- A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution:
  - $n \times p \geq 10$
  - $n \times q \geq 10$
  - $\text{Binomial}(n,p) \sim \text{Normal}(\mu,\sigma)$

Normal vs. binomial distribution

- What is the minimum required n for a binomial distribution with probability of success of 0.25 to closely follow a normal distribution?
- $n \times 0.25 \geq 10 \rightarrow n \geq 10/0.25 \rightarrow n \geq 40$
- $n \times 0.75 \geq 10 \rightarrow n \geq 10/0.75 \rightarrow n \geq 13.33$

Normal distribution

- Unimodal & symmetric (bell shape)
- Many variables are nearly normal not exactly normal
- Normal(μ,σ)

Normal distribution

- A family doctor with ~ 3,000 subjects on the list measure over one year the heart rates (expected to be normal distributed). Three statistics were reported: mean = 75, minimum = 45, and maximum = 150. Which of the following is most likely to be the standard deviation of the distribution?

A. $2 \rightarrow 75 \pm 3 \times 2 = (69; 81)$

B. $5 \rightarrow 75 \pm 3 \times 5 = (60; 90)$

C. $10 \rightarrow 75 \pm 3 \times 10 = (45; 105)$

D. $12 \rightarrow 75 \pm 3 \times 12 = (39; 111)$

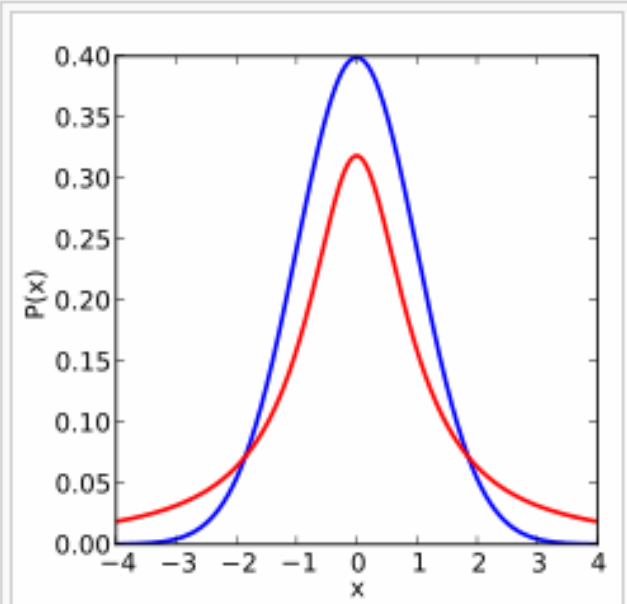
E. $15 \rightarrow 75 \pm 3 \times 15 = (30; 120)$

t-distribution

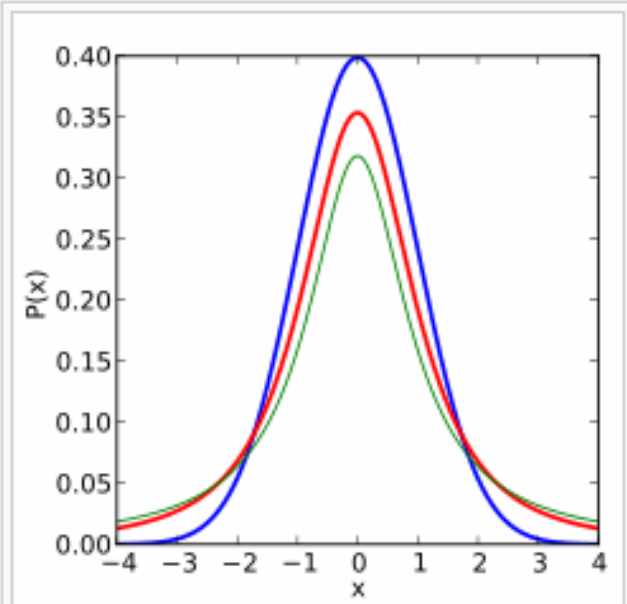
- Student's t-distribution (or simply the t-distribution) arise when the estimation of the mean of a normally distributed population is done on a small sample and population standard deviation is unknown.

- normal distribution describes a full population
- t-distributions describe samples drawn from a full population
  - the t-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution

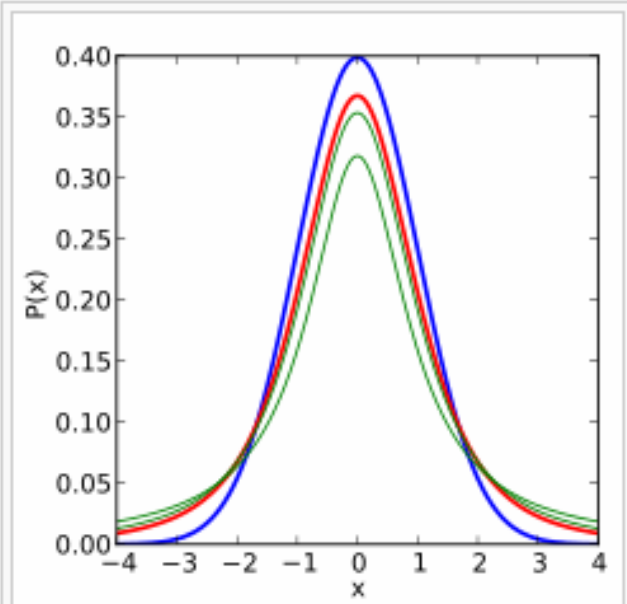
Density of the *t*-distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).
Previous plots shown in green.



1 degree of freedom



2 degrees of freedom



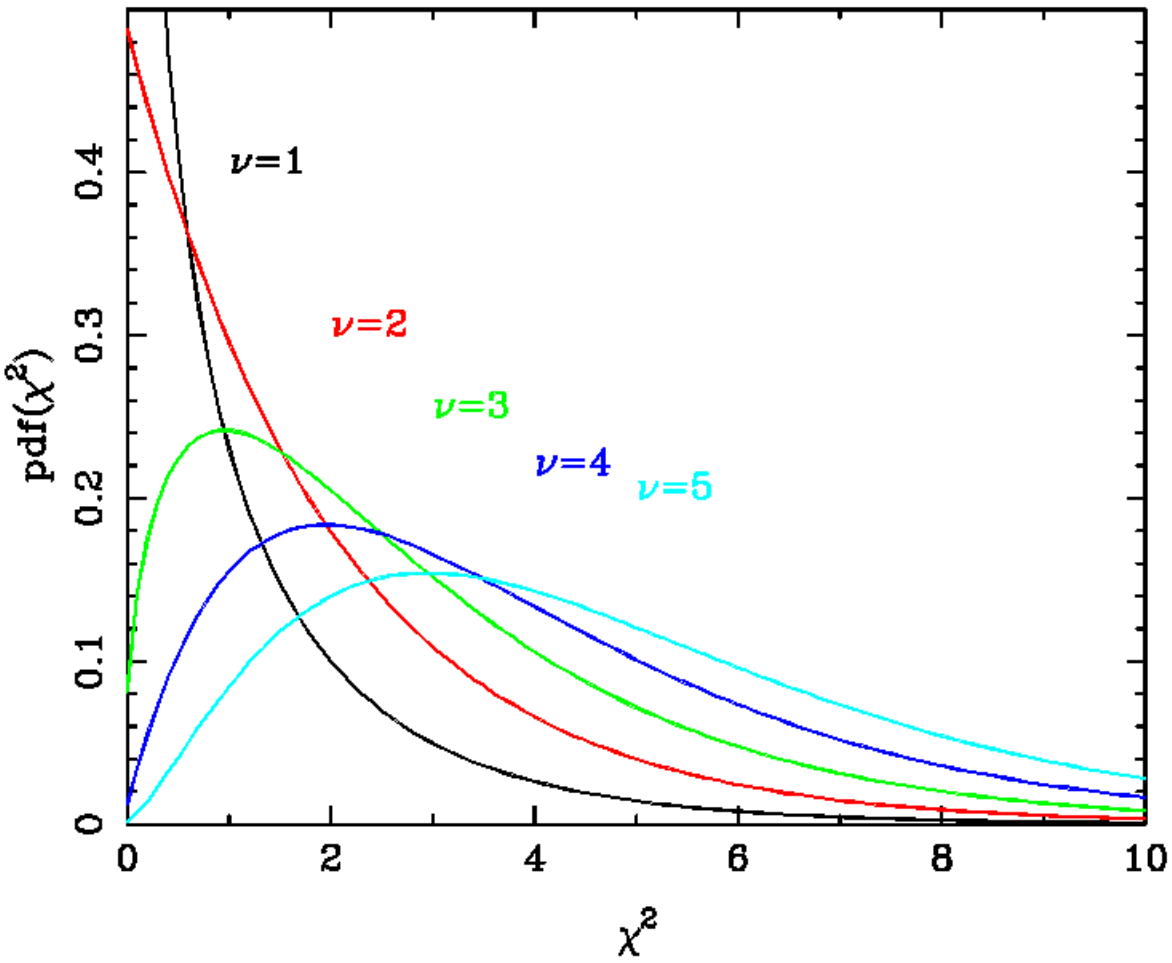
3 degrees of freedom

Chi-square distribution

- Related to the sampling distribution of variances (along with F-distribution)
- A random variable *X* has a Chi-square distribution if it can be written as a sum of squares: $X = Y_1^2 + Y_2^2 + \ldots + Y_n^2$, where Y_1, Y_2, … Y_n are mutually independent standard normal random variables
- Minimum value = 0
- Maximum value = infinite
- Most values are between 0 and 1
- Is the distribution of a sum of squares
- The shape depends on the number of added squared deviates

Chi-square distribution

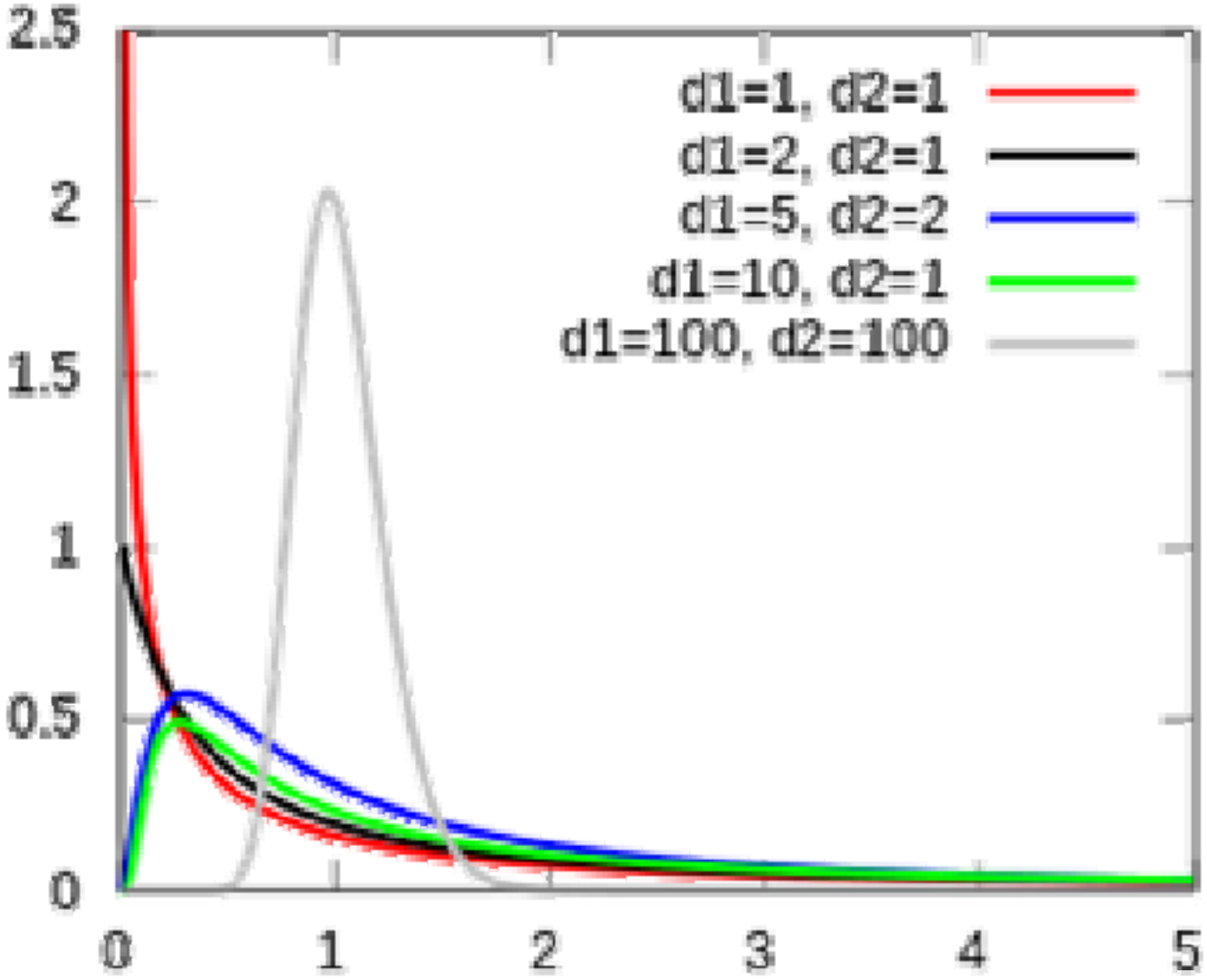
- The expected values = df
- The expected variance of the distribution is 2×df



Fisher distribution

- D-distribution is a ratio (of two variance estimates / two chi-square defined by its degrees of freedom)
- Depended of two parameters (two degrees of freedom)
- Take values between 0 and infinte
- It is used in several statistical tests: equality of variances, ANOVA, regression

Fisher distribution



X	X Measures	f(x)	Values of X	E(x)	V(x)
Continuous uniform	Outcomes with equal density (continuous)	$\frac{1}{b-a}$	$a \le x \le b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Exponential	Time between events; time until an event	$\lambda e^{-\lambda x}$	$x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal	Values with a bell-shaped distribution (continuous)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$-\infty < x < \infty$	μ	σ
Standard normal (Z)	Standard scores	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	$Z = \frac{x-\mu}{\sigma}$	0	1
Binomial approximation	Number of successes in large number of trials	Approx. normal if $np \ge 5$ and $n(1-p) \ge 5$ by CLT	$Z = \frac{x-np}{\sqrt{np(1-p)}}$	np	$np(1-p)$
Poisson approximation	Number of occurrences in a fixed time period (large average)	Approx. normal if $\lambda > 30$	$z = \frac{x-\lambda}{\sqrt{\lambda}}$	λ	λ
$\bar{X}$	Average of $x_1, x_2, \ldots, x_n$	Exactly normal if x is normal. Approx. normal if $n \ge 30$ by CLT	$Z = \frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}}$	μ_x	$\frac{\sigma_x^2}{n}$
$\hat{p}$	Proportion or percentage of successes in binomial with $np \ge 5$, $n(1-p) \ge 5$	Approx. normal if $np \ge 5$ and $n(1-p) \ge 5$ by CLT	$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	p	$\frac{p(1-p)}{n}$

Problems

- A cohort study has been conducted and a probability of developing a disease of 0.05 in the exposed group was observed.

1. How many people would be expected to develop the disease on a randomly sample of exposed subject? Provide a margin of error (± 1 standard deviation) for the estimate.
2. What is the probability that at most 10 exposed people to develop the disease?

Problems

- A cohort study has been conducted and a probability of developing a disease of 0.025 in the exposed group was observed.

1. How many people would be expected to develop the disease on a randomly sample of exposed subject? Provide a margin of error (± 1 standard deviation) for the estimate.

- $X \sim Binomial(500, 0.025)$
- $M(X) = 500 \times 0.025 = 12.5$
- $V(X) = 500 \times 0.025 \times 0.975 = 12.19$
- $\sqrt{12.19} = 3.49$

Answer: 12.5±3.49

Problems

- A cohort study has been conducted and a probability of developing a disease of 0.025 in the exposed group was observed.

2. What is the probability that at most 10 exposed people to develop the disease?
   - Is asking for a cumulative probability:
   - $P(X \leq 10) = P(X=0) + P(X=1) + \ldots + P(X=10)$
   - $P(X \leq 10) = combin(500,0) \times 0.025^{0} \times 0.975^{500} + \ldots + combin(500,10) \times 0.025^{10} \times 0.975^{490}$

Problems

- A case-control study was conducted to investigate if smoking is a risk factor for lung cancer. If the probability of being a smoker among subjects with lung cancer is 0.70. What is the chance that in a group of 8 subjects to have:

1. Less than 2 smokers?
2. More than 5 smokers?
3. What is the expected value and variance of the number of smokers?

Problems

A case-control study was conducted to investigate if smoking is a risk factor for lung cancer. If the probability of being a smoker among subjects with lung cancer is 0.70. What is the chance that in a group of 8 subjects to have:

Less than 2 smokers? P(X<2) = P(X=0) + P(X=1) = 0.000066 + 0.001225

P(X<2) = 0.001290

X	P(X)
0	=Combin(8,0)*(0.7^0)*(0.3^8) = 0.000066
1	=Combin(8,1)*(0.7^1)*(0.3^7) = 0.001225
2	=Combin(8,2)*(0.7^2)*(0.3^6) = 0.010002
3	=Combin(8,3)*(0.7^3)*(0.3^5) = 0.046675
4	=Combin(8,4)*(0.7^4)*(0.3^4) = 0.136137
5	=Combin(8,5)*(0.7^5)*(0.3^3) = 0.254122
6	=Combin(8,6)*(0.7^6)*(0.3^2) = 0.296475
7	=Combin(8,7)*(0.7^7)*(0.3^1) = 0.197650
8	=Combin(8,8)*(0.7^8)*(0.3^0) = 0.057648

Problems

- A case-control study was conducted to investigate if smoking is a risk factor for lung cancer. If the probability of being a smoker among subjects with lung cancer is 0.70. What is the chance that in a group of 8 subjects to have:

2. More than 5 smokers?

X	P(X)
0	=Combin(8,0)*(0.7^0)*(0.3^8) = 0.000066
1	=Combin(8,1)*(0.7^1)*(0.3^7) = 0.001225
2	=Combin(8,2)*(0.7^2)*(0.3^6) = 0.010002
3	=Combin(8,3)*(0.7^3)*(0.3^5) = 0.046675
4	=Combin(8,4)*(0.7^4)*(0.3^4) = 0.136137
5	=Combin(8,5)*(0.7^5)*(0.3^3) = 0.254122
6	=Combin(8,6)*(0.7^6)*(0.3^2) = 0.296475
7	=Combin(8,7)*(0.7^7)*(0.3^1) = 0.197650
8	=Combin(8,8)*(0.7^8)*(0.3^0) = 0.057648

P(X>5) =

P(X=6)+P(X=7)+P(X=8)=

0.296475 + 0.197650 + 0.057648=

0.551774

P(X>5) = 0.551774

Problems

- A case-control study was conducted to investigate if smoking is a risk factor for lung cancer. If the probability of being a smoker among subjects with lung cancer is 0.70. What is the chance that in a group of 8 subjects to have:

3. What is the expected value and variance of the number of smokers?

- $M(X) = 8*0.7 = 5.6$
- $V(X) = 8*0.7*0.3 = 1.7$
- $Stdev(X) = \sqrt{1.7} = 1.30$