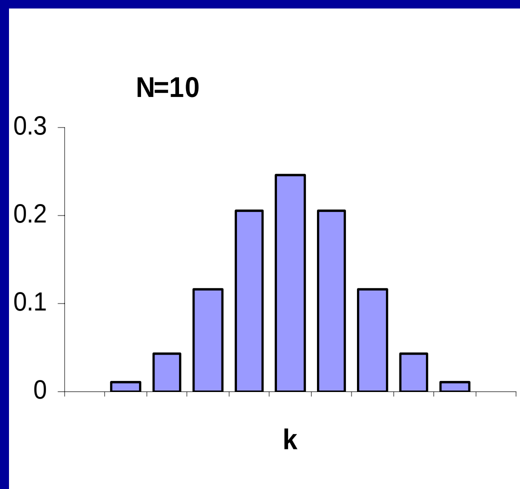
Normal approximation to Binomial

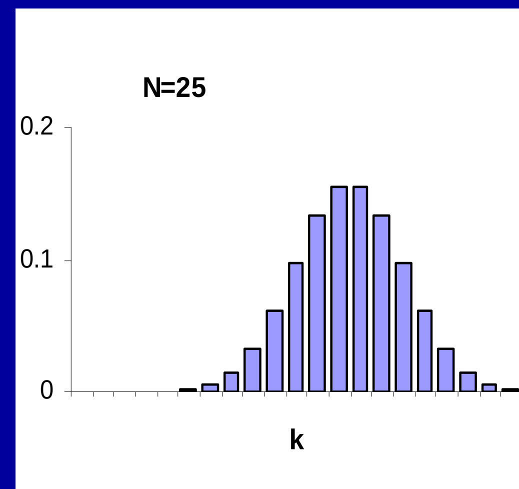
24.10.2007
GE02: day 3 part 3

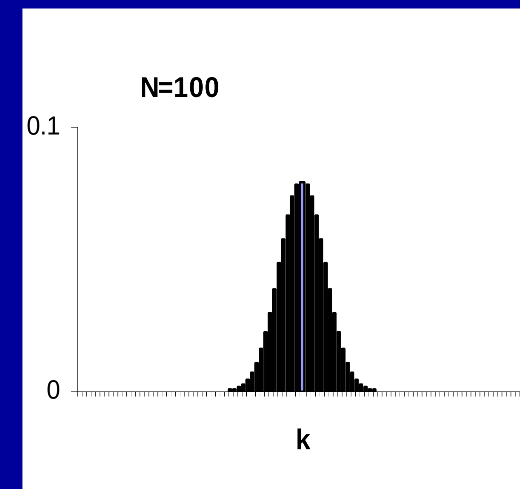
Yurii Aulchenko
Erasmus MC Rotterdam

Binomial distribution at different *n* and p

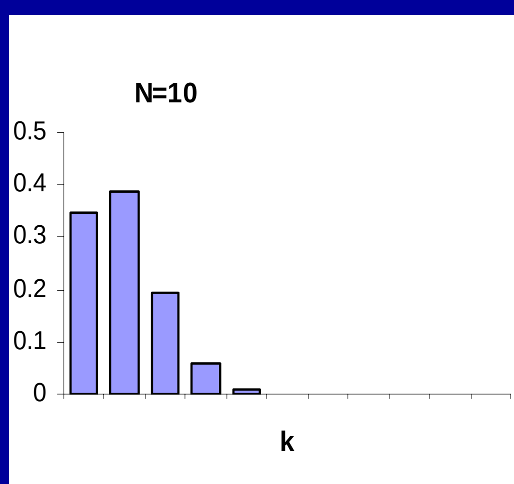
- P=0.5

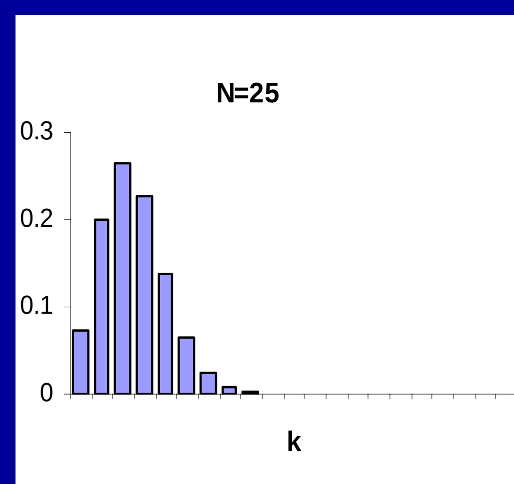


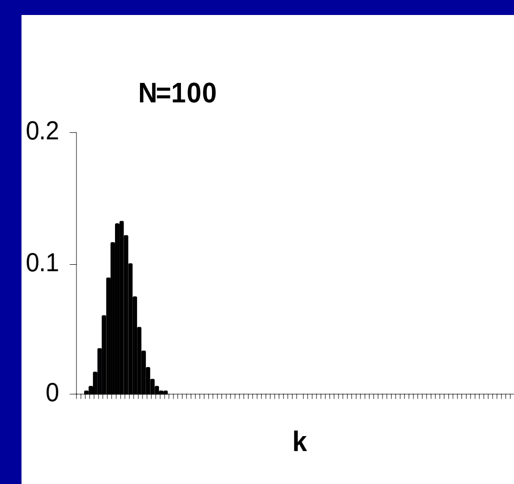




- P=0.1







Carl Friedrich Gauss (1777-1855)

- Developed Normal (Gaussian) distribution to describe measurement error

Normal approximation

- *n* must be large, say >100
- If *np* > 5, use Normal approximation

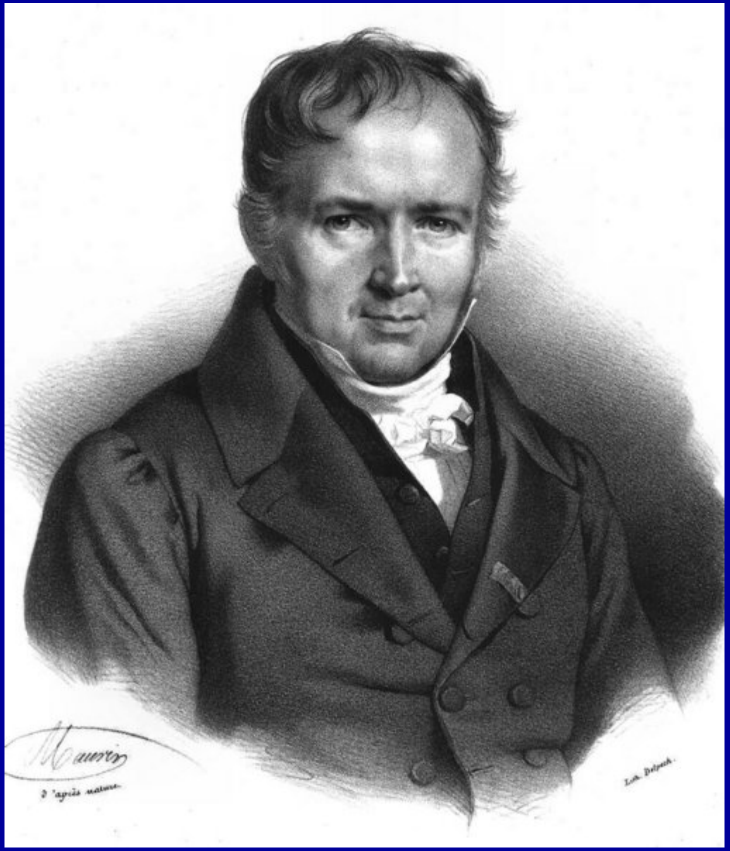
$\text{Binomial}_{n,p}(x) \propto P_{\mu,\sigma}(x) =$

  - where mean $\mu = np$ and variance $\sigma^2 = np(1-p)$

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Siméon D. Poisson (1781-1840)

- Book on "Research on the Probability of Judgments in Criminal and Civil Matters"
- His distribution described time till some rare event happens



Poisson approximation

- If *np* is about 1-4, use Poisson approximation

$$\text{Binomial}_{n,p}(x) \propto P_\lambda(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

  - where $\lambda = np$

Problem

- What approximation would you use under these scenarios?

	n		
p	100	250	1000
0.5	?	?	?
0.01	?	?	?
0.001	?	?	?

Solution

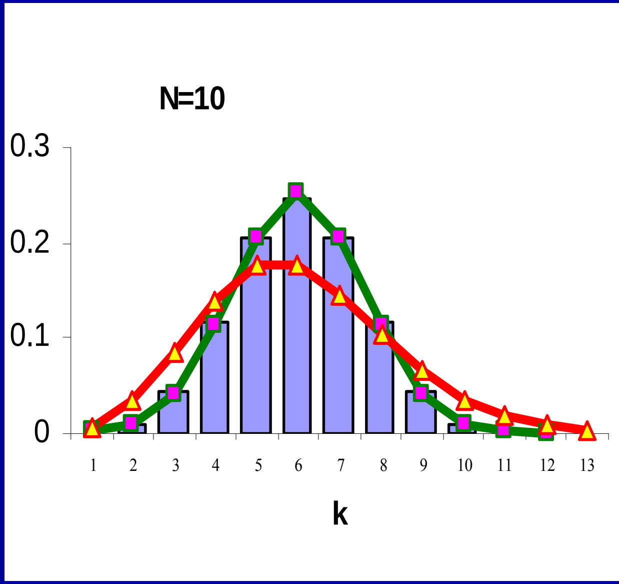
- Table of np's:

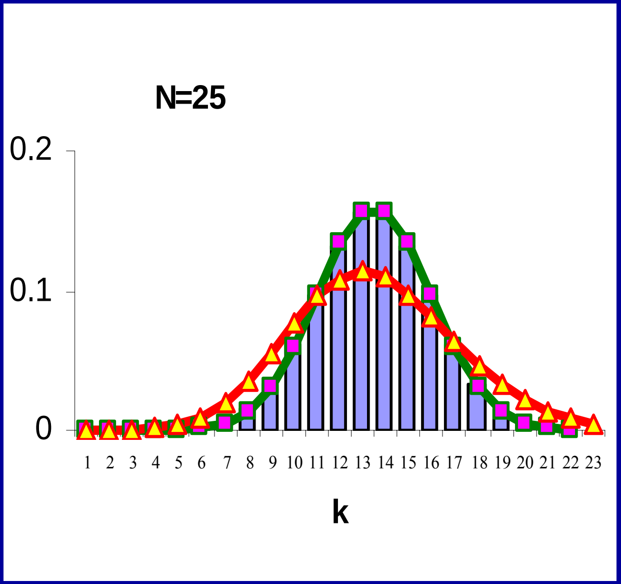
p	n: 100	250	1000
0.5	50	125	500
0.01	1	2.5	10
0.001	0.1	0.25	1

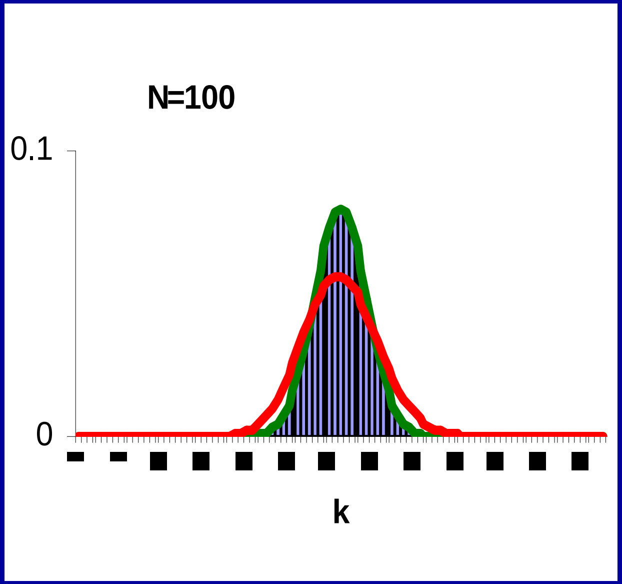
- Approximation:

p	n: 10	25	100
0.5	N	N	N
0.01	P	P	N
0.001	P	P	P

Approximating Binomial at p=0.5

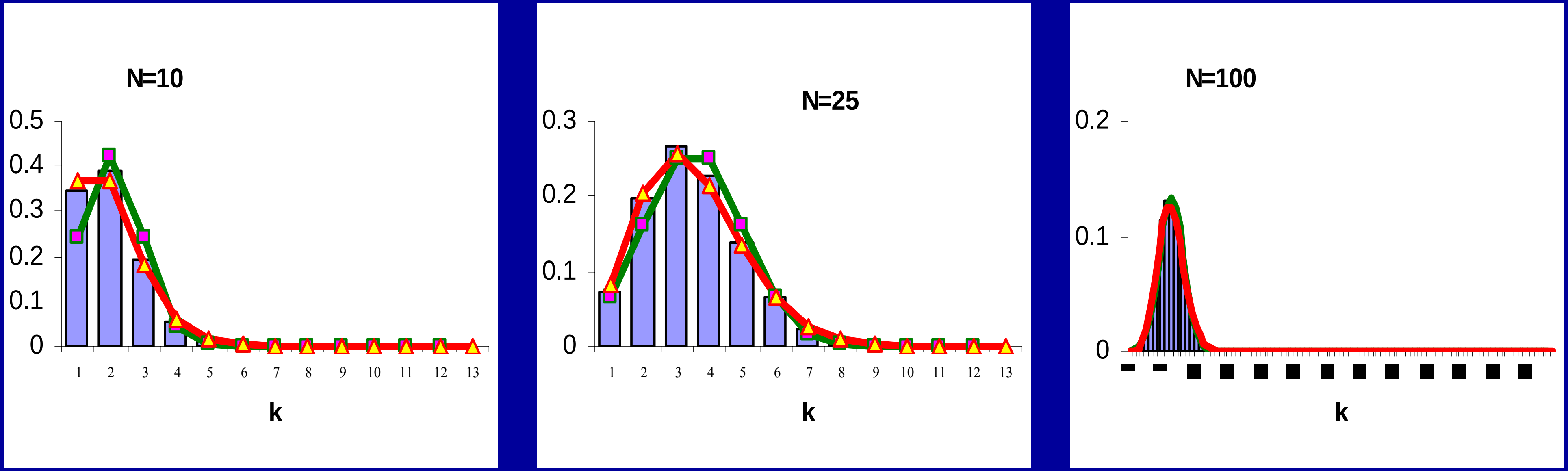






Green:	Normal approximation
Red:	Poisson approximation

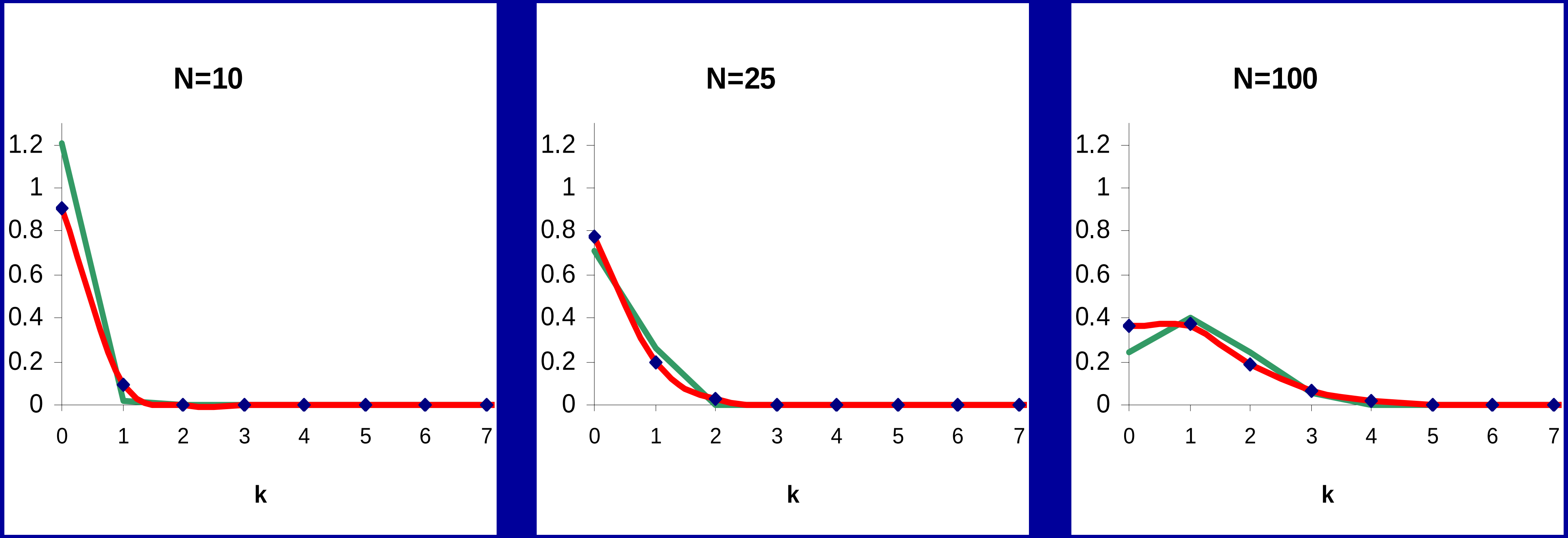
Approximating Binomial at p=0.1



Green: Normal approximation

Red: Poisson approximation

Approximating Binomial at p=0.01



Green: Normal approximation

Red: Poisson approximation

Approximating Binomial(k≤x)

- Binomial$_{n,p}$(k) ∝ P$_\lambda$(k) = $\dfrac{e^{-\lambda}\lambda^{k}}{k!}$
  - where λ = *np*

- Binomial$_{n,p}$(k≤x) ∝ P$_\lambda$(k≤x) = $\displaystyle\sum_{-\infty}^{x}\frac{e^{-\lambda}\cdot\lambda^{x}}{x!}$

$P_{\lambda}(k \le x)$

Table IV. Cumulative Poisson Distribution

	Lambda									
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.905	0.819	0.741	0.670	0.607	0.549	0.497	0.449	0.407	
1	0.995	0.982	0.963	0.938	0.910	0.878	0.844	0.809	0.772	0.736
2	1.000	0.999	0.996	0.992	0.986	0.977	0.966	0.953	0.937	0.920
3	1.000	1.000	1.000	0.999	0.998	0.997	0.994	0.991	0.987	0.981
4				1.000	1.000	1.000	0.999	0.999	0.998	0.996
5										0.999
6										1.000

Probability to sample 0 alleles with frequency of 0.01 among 100 chromosomes:
n=100, p=0.01, k=0, λ=np=1

Problem

- A mutation of microsatellite marker occurs in rate of 10^{-3} per meiosis
- In a complex pedigree, including 1500 meioses, 2 Mendelian errors were observed

- **P1:** What is the chance to have 2 or more errors under assumption that all errors represent new mutations?
- **P2:** What would be the number of errors, after which you would conclude that there is genotyping error (at α=0.05)?

Solution P1: Binomial

- $P(\mathbf{k} \geq 2) = 1 - P(\mathbf{k} \leq 1) =$

  $1 - P(\mathbf{k}=1) - P(\mathbf{k}=0) =$

  $1 - 1500 \cdot 0.001 \cdot 0.999^{1499} - 0.999^{1500} =$

  0.442

Solution P2: Binomial

- Idea: compute
  - P($\mathbf{k}\geq 2$)
  - P($\mathbf{k}\geq 3$)
  - P($\mathbf{k}\geq 4$)
  - P($\mathbf{k}\geq 5$)
  - P($\mathbf{k}\geq 6$)
  - …
- And check when it becomes $\leq$ 0.05

Solution P1: Poisson

- $P(\mathbf{k} \geq 2) = 1 - P(\mathbf{k} \leq 1)$

- $\lambda = \mathbf{np} = 1500 \cdot 10^{-3} = 1.5$

- Using the table

$$P_{\lambda=1.5}(\mathbf{k} \geq 2) = 1 - P_{\lambda=1.5}(\mathbf{k} \leq 1) =$$
$$1 - 0.558 = 0.442$$

Solution P2: Poisson

- P(**k**≥X) = 1 – P(**k**≤[X–1]) ≤ 0.05
- P(**k**≤[X-1]) ≥ 0.95

- Idea: check the column with λ = 1.5 and see at what **k** it becomes more then 0.95, then add 1 to this number

- Answer: 5 errors

	Lambda					
k	1.1	1.2	1.3	1.4	1.5	1.6
0	0.333	0.301	0.273	0.247	0.223	0.202
1	0.699	0.663	0.627	0.592	0.558	0.525
2	0.900	0.879	0.857	0.833	0.809	0.783
3	0.974	0.966	0.957	0.946	0.934	0.921
4	0.995	0.992	0.989	0.986	[illegible]	0.976
5	0.999	0.998	0.998	0.997	0.996	0.994
6	1.000	1.000	1.000	0.999	0.999	0.999
7				1.000	1.000	1.000

Standard Normal

- Normal density function with mean 0 and variance 1:

$$P(k=x)=\phi(x)=\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{x^2}{2}\right)$$

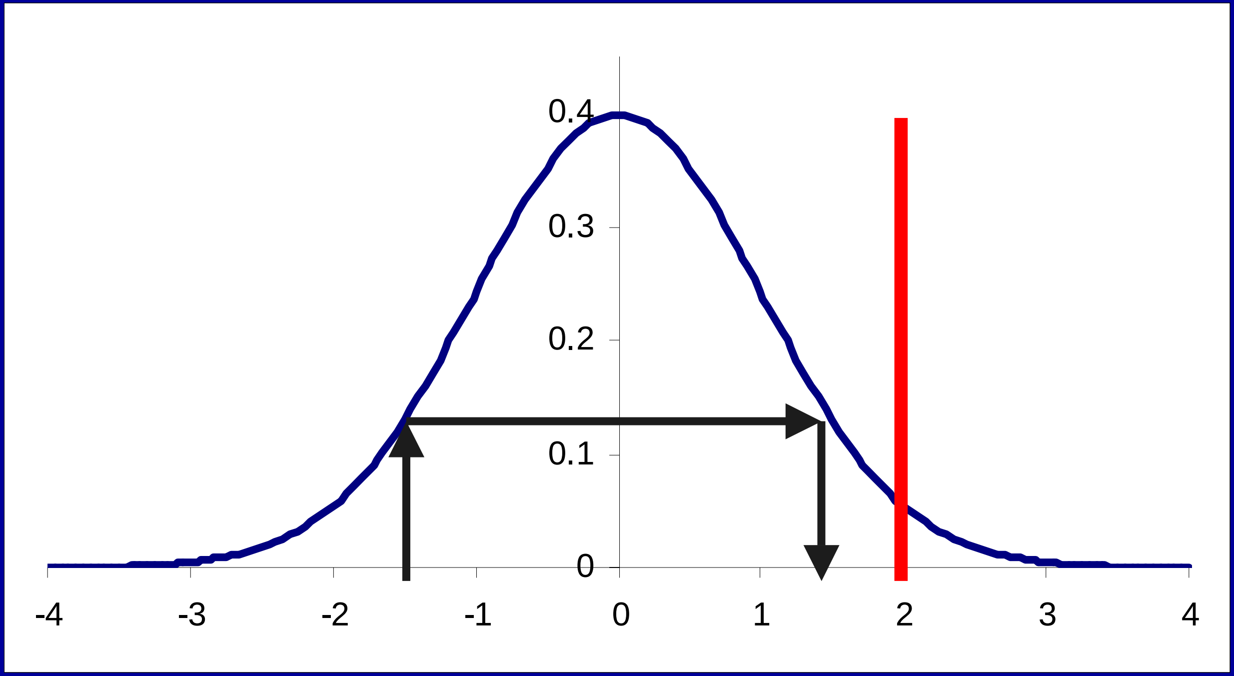
- Its integral is termed Normal distribution:

$$P(k\leq x)=\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{x^2}{2}\right)dx$$

Standard Normal

- We know a lot about this function and many statistical techniques are based on that

Facts about Standard Normal



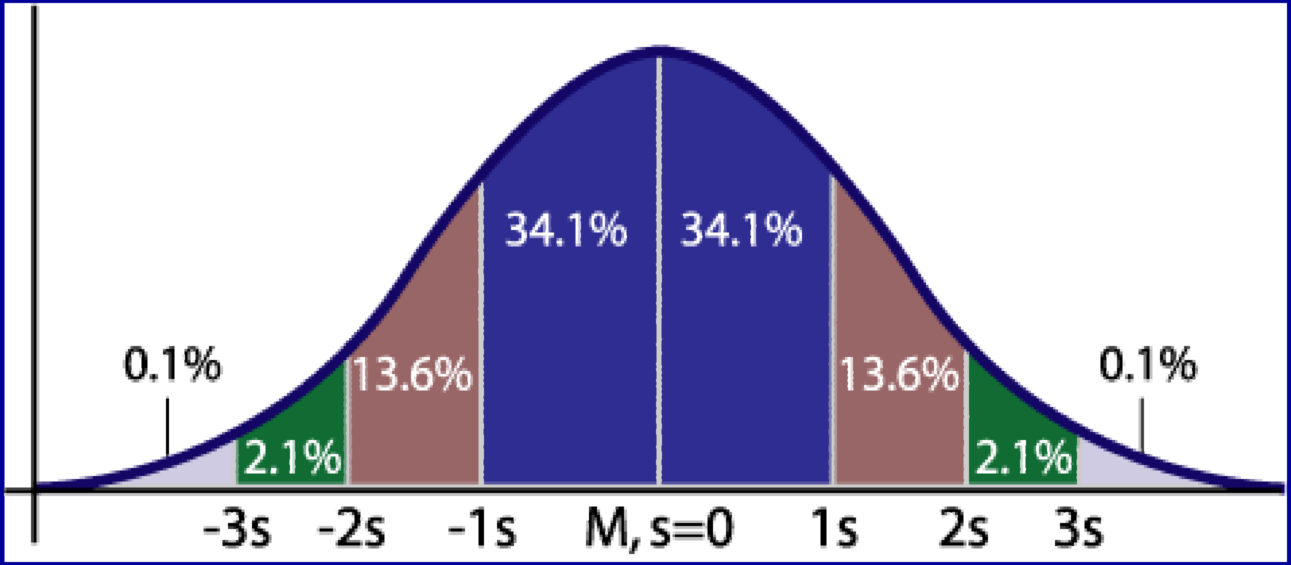
It is symmetric distribution, therefore

$$\phi(-x) = \phi(x)$$

It has area of 1, therefore

$$\mathbf{\Phi(x) = 1 - \Phi(-x)}$$

Facts about Standard Normal



$P(x \le 1) = 0.84$	$P(-1 \le x \le 1) = 2\ (1 - 0.84) = 0.32$
$P(x \le 2) = 0.977$	$P(-2 \le x \le 2) = 2\ (1 - 0.977) = 0.955$
$P(x \le 3) = 0.999$	$P(-3 \le x \le 3) = 2\ (1 - 0.999) = 0.997$

$P(x \le 1.64) = 0.95$	$P(-1.64 \le x \le 1.64) = 0.90$
$P(x \le 1.96) = 0.975$	$P(-1.96 \le x \le 1.96) = 0.95$
$P(x \le 2.57) = 0.995$	$P(-2.32 \le x \le 2.57) = 0.99$

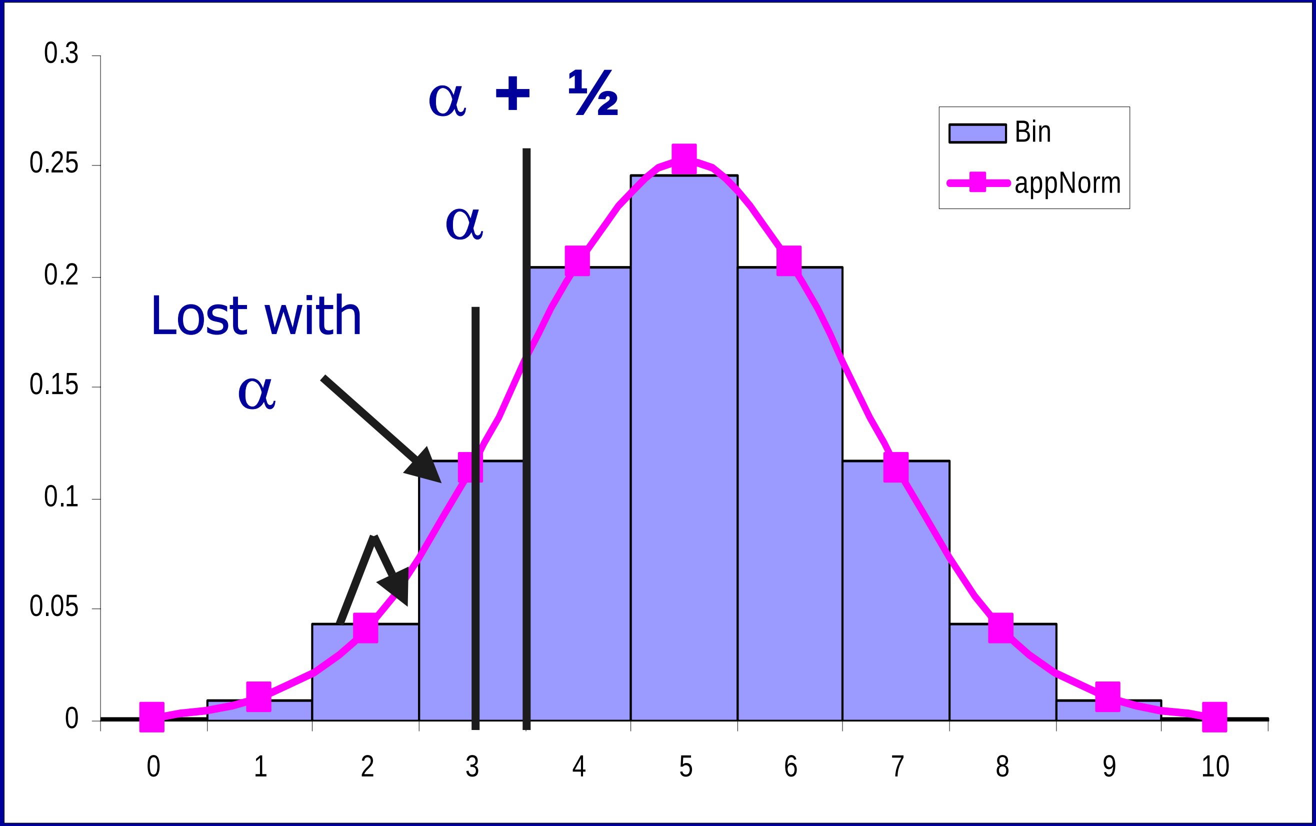
$P(x \leq Z) = \Phi(Z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Approximating Binomial with Normal

mean $\mu = np$ and variance $\sigma^2 = np(1-p)$

$$P(k \le \alpha) \approx \Phi_{\mu,\sigma}(\alpha + 0.5)$$



Using Standard Normal

$$P(k \le \alpha) = \Phi\left(\frac{(\alpha + 0.5) - \mu}{\sigma}\right)$$

$$P(k > \alpha) = 1 - \Phi\left(\frac{(\alpha + 0.5) - \mu}{\sigma}\right)$$

$$P(k \ge \alpha) = 1 - \Phi\left(\frac{(\alpha - 0.5) - \mu}{\sigma}\right)$$

$$P(\alpha \le k \le \beta) = \Phi\left(\frac{\beta - \mu + 0.5}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu - 0.5}{\sigma}\right)$$

Problem

- Coin is tossed 200 times.
- Estimate probability that the number of heads is between 95 and 105, included – that is to say that it deviates from 100 by 5 at most

- Suggestion
  - $n>100$, $np=100 \Rightarrow$ use Normal approximation

Solution

The parameters of the Binomial are μ=*np*=100 and variance σ^2 = *np(1-p)*=50 (then σ is 7.07)

$$P(\alpha \le k \le \beta) = \Phi\left(\frac{\beta - \mu + 0.5}{\sigma}\right) - \Phi\left(\frac{\alpha - \mu - 0.5}{\sigma}\right)$$

$$P(95 \le k \le 105) = \Phi\left(\frac{105.5 - 100}{7.07}\right) - \Phi\left(\frac{94.5 - 100}{7.07}\right) = \Phi\left(\frac{5.5}{7.07}\right) - \Phi\left(\frac{-5.5}{7.07}\right) =$$

Φ(0.78) – Φ(-0.78) = Φ(0.78) – (1 – Φ(0.78)) =

2 Φ(0.78) – 1 = ???

Number!

$$\Phi(0.778) - \Phi(-0.778) =$$
$$1 - 2\mathrm{x}\Phi(0.778) =$$
$$2\mathrm{x}0.782 - 1 =$$
$$= 0.56$$

Z	0.00	0.01	0.07	.08	0.09
0.0	0.5000	0.5040	0.5279	.5319	0.5359
0.1	0.5398	0.5438	0.5675	.5714	0.5753
0.2	0.5793	0.5832	0.6064	.6103	0.6141
0.3	0.6179	0.6217	0.6443	.6480	0.6517
0.4	0.6554	0.6591	0.6808	.6844	0.6879
0.5	0.6915	0.6950	0.7157	.7190	0.7224
0.6	0.7257	0.7291	0.7486	7517	0.7549
0.7	0.7580	0.7611	0.779	[illegible]	0.7852
0.8	0.7881	0.7910	0.8078	0.8106	0.8133

We leave exact (Binomial) computations and comparison for the exercises session

Problem

- Frequency of a disease allele is 0.03
- In a sample of 100 people
  - What number of carrier is expected?
  - What is the chance to have 10 or more carriers?

- Assume HWE

Solution

- Carrier frequency is roughly 6%
- The parameters of the Binomial are $\mu = np = 6$ and variance $\sigma^2 = np(1-p) = 5.64$ (then σ is 2.37)
- Use Normal (np>5)

$$P(k > \alpha) = 1 - P(k \le \alpha) = 1 - \Phi\left(\frac{\alpha - \mu + \frac{1}{2}}{\sigma}\right)$$

$$P(k > 9) = 1 - \Phi\left(\frac{9 - 6 + \frac{1}{2}}{2.37}\right) = 1 - \Phi\left(\frac{3.5}{2.37}\right) = 1 - \Phi(1.48)$$

= 1 – 0.93 = 0.07

Problem

- The frequency of a genetic variant is 0.01
- How many people you need to sample to have 95% probability that at least one is carrier?

Solution

- P(at least one carrier) $\geq$ 0.95
- P(at least one carrier) =

  1 – P(no carriers) = 1 – 0.98^n
- 1 – 0.98^n = 0.95
- $0.98^n = 0.05$
- n = ln(0.05)/ln(0.98) = 148.28

Problem

- The frequency of a genetic variant is 0.01
- How many people you need to sample to have 95% probability to have at least THREE carriers?

Straight solution

- $P(\geq 3 \text{ carriers}) \geq 0.95$

$1 - P(0 \text{ carriers}) - P(1 \text{ carrier}) - P(2 \text{ carriers}) \geq 0.95$

$1 - [0.98^{n}] - [n \cdot 0.02 \cdot 0.98^{n-1}] - [\frac{1}{2} \cdot n \cdot (n-1) \cdot 0.02^{2} \cdot 0.98^{n-2}] \geq 0.95$

$[0.98^{n}] + [n \cdot 0.02 \cdot 0.98^{n-1}] + [\frac{1}{2} \cdot n \cdot (n-1) \cdot 0.02^{2} \cdot 0.98^{n-2}] \leq 0.05$

- ?!!? Solution ?!!?

Using Poisson

- As **p** is low (0.02), Poisson approximation may work well

Idea of solution

- Event of interest is $\mathbf{k} \geq 3$
- $P(\mathbf{k} \geq 3) = 1 - P(\mathbf{k} \leq 2) \geq 0.95$
- $P(\mathbf{k} \leq 2) \leq 0.05$

- In Poisson, $\lambda = \mathbf{np}$
- If you find out what λ gives $P(\mathbf{k} \geq 3) = 0.95$, then n is $\lambda / \mathbf{p}$

Solution

- At **k**=2
  - λ**=**6.2 gives P(**k**≤2) = 0.054
  - λ**=**6.4 gives P(**k**≤2) = 0.046

- Therefore sample size should be between
  - 6.2/0.02 = 310 and
  - 6.4/0.02 = 320

Solution with Normal 1

- P(k≥3 carriers) = P(>2 carriers)
- P(k>2 carriers) ≥ 0.95

$$P(k > \alpha) = 1 - \Phi\left(\frac{\alpha - \mu + \frac{1}{2}}{\sigma}\right)$$

- $\mu = 0.02\, n$; $\sigma^2 = n\, 0.02\; 0.98$

$$P(k > 2) = 1 - \Phi\left(\frac{2 - 0.02 \cdot n + \frac{1}{2}}{\sqrt{0.02 \cdot 0.98 \cdot n}}\right) \geq 0.95$$

Solution with Normal 2

$$\Phi\left(\frac{2-0.02\cdot n+\dfrac{1}{2}}{\sqrt{0.02\cdot 0.98\cdot n}}\right)<0.05$$

- Use table:

$$\frac{2-0.02\cdot n+\dfrac{1}{2}}{\sqrt{0.02\cdot 0.98\cdot n}}<-1.64$$

$$1.64\cdot\sqrt{0.0196}\cdot\sqrt{n}-0.02\cdot n<-2\frac{1}{2}$$

Solving quadratic equation

- If there is equation of the form
  A $\sqrt{n}$ – B n = – C

- Solution is

$$\frac{A^2+2\cdot B\cdot C+A\cdot\sqrt{A^2+4\cdot B\cdot C}}{2\cdot B^2}$$

Answer is…

$$A = 1.64\sqrt{0.0196} \approx 0.23; \quad B = 0.02; \quad C = 2\frac{1}{2}$$

$$\frac{A^2 + 2 \cdot B \cdot C + A \cdot \sqrt{A^2 + 4 \cdot B \cdot C}}{2 \cdot B^2} =$$

$$\frac{0.23^2 + 2 \cdot 0.02 \cdot 2.5 + 0.23 \cdot \sqrt{0.23^2 + 4 \cdot 0.02 \cdot 2.5}}{2 \cdot 0.02^2} =$$

$$\frac{0.053 + 0.1 + 0.23 \cdot \sqrt{0.053 + 0.2}}{0.0008} = \frac{0.153 + 0.23 \cdot 0.503}{0.0008} = 335$$

Binomial computations: exercises session

Useful excel functions

- Cumulative binomial $P_{n,p}(x \leq k)$

  =binomdist(k,n,p,1)

- Cumulative standard normal $\Phi(x \leq k)$

  =normdist(k,0,1,1)

- Poisson $P_{\lambda}(x \leq k)$

  =posisson(k,λ,1)

- Chi-squared with m d.f., $\chi_m^2(x \geq k)$

  =chidist(k,m)