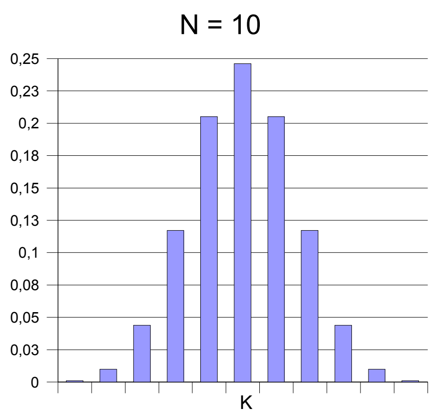
Normal approximation to Binomial

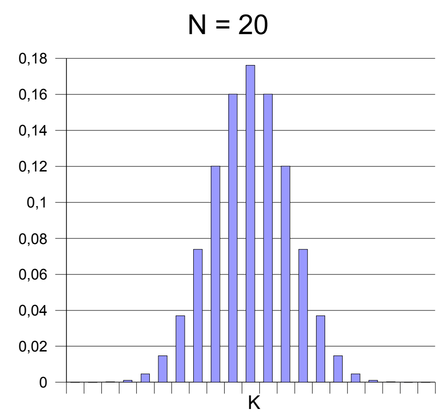
24.10.2005
GE02: day 2 part 4

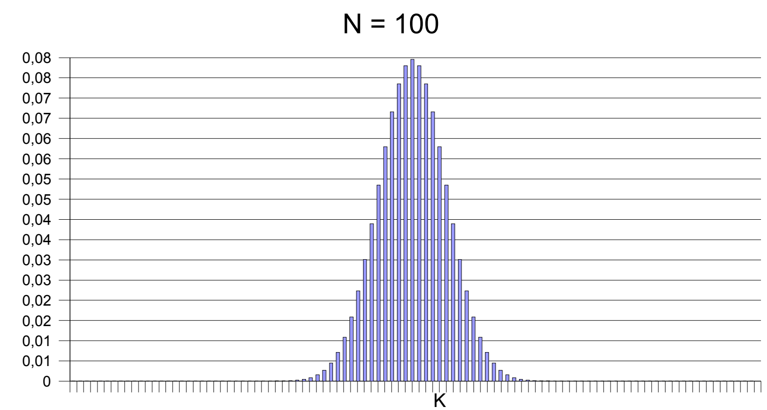
Yurii Aulchenko
Erasmus MC Rotterdam

Binomial distribution at different N

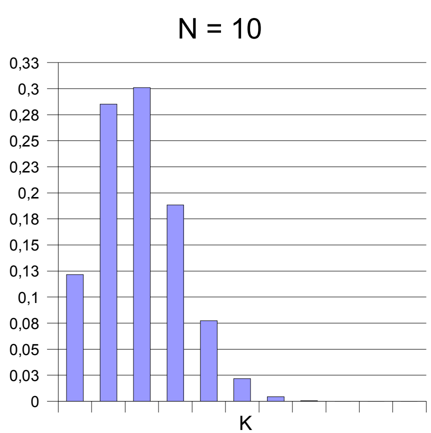
$p = 0.5$

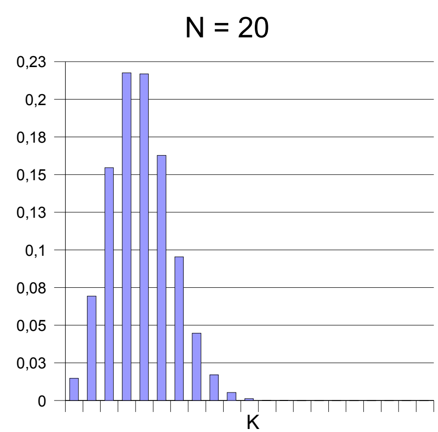


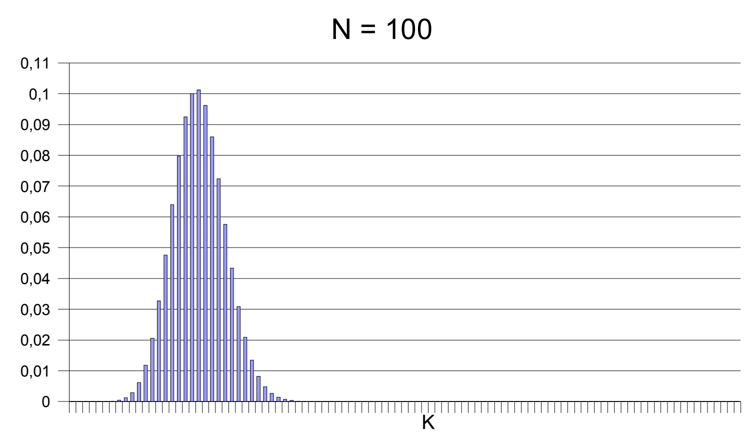




$p = 0.19$







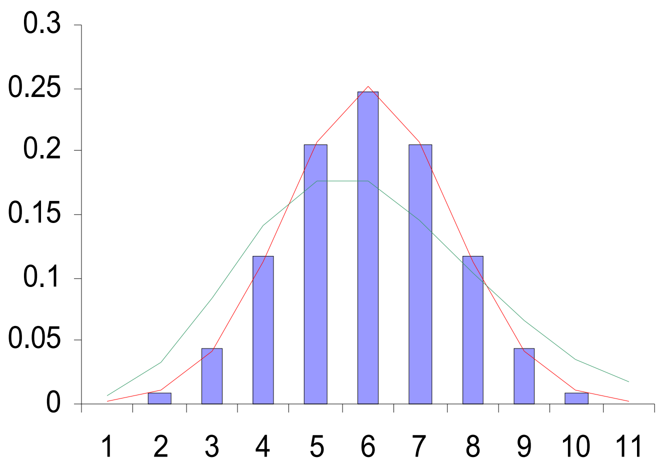
Normal approximation of Binomial

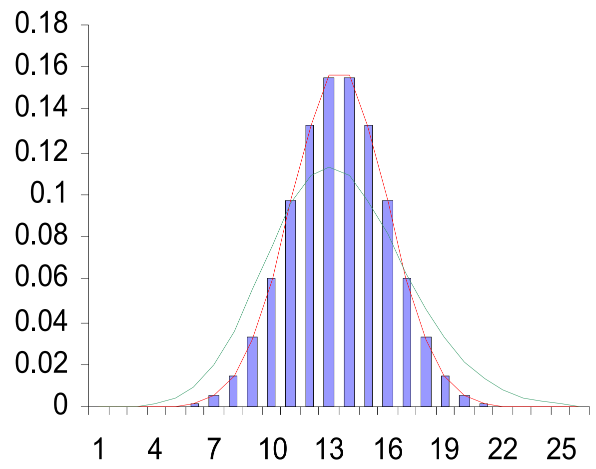
- *n* must be large, say >100
- If *np* > 5, use Normal approximation
  - $P_{\mu,\sigma,}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - where mean $\mu = np$ and variance $\sigma^2 = np(1-p)$

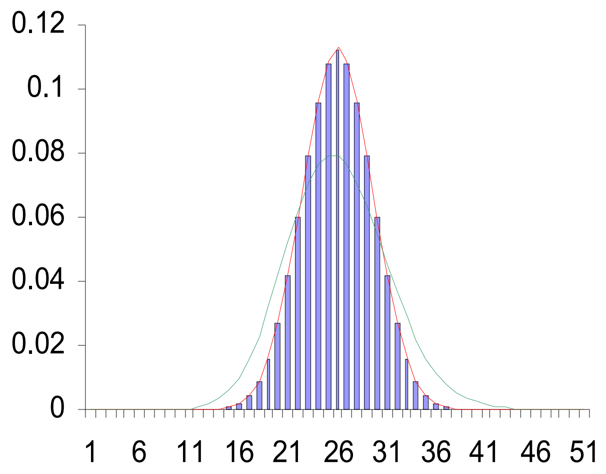
Poisson approximation of Binomial

- If *np* is about 1-4, use Poisson approximation
  - $P_\lambda(x) = \dfrac{e^{-\lambda}\lambda^k}{k!}$
  - where $\lambda = np$

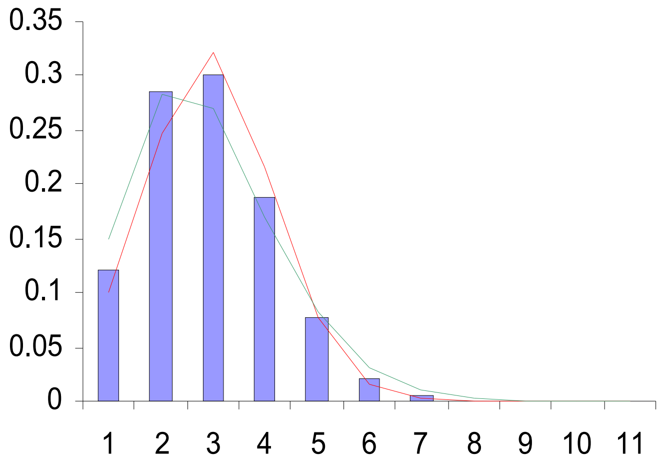
Approximating Binomial
p=0.5, *n*=10, 25, 50

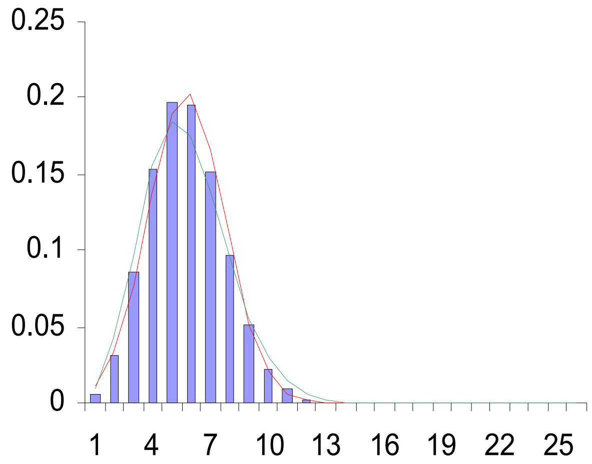


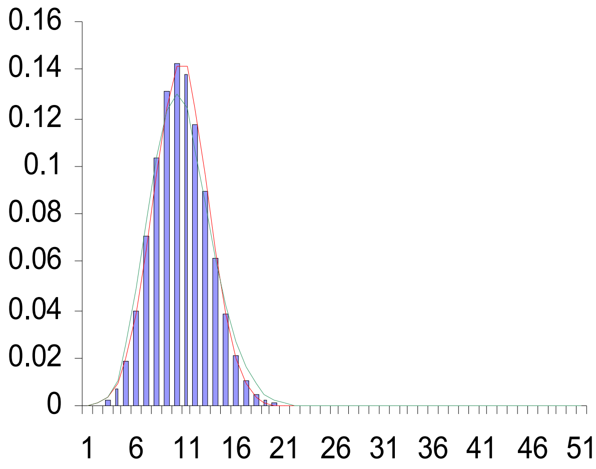




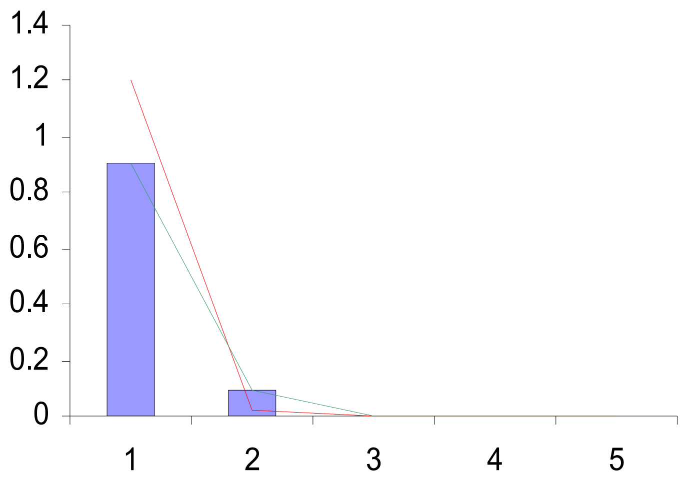
Approximating Binomial
p=0.19, *n* =10, 25, 50

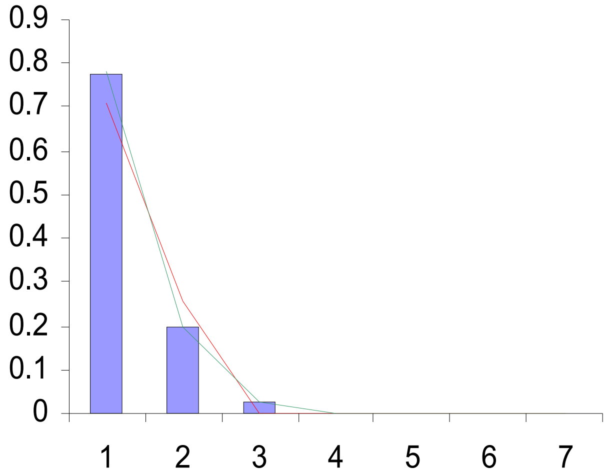


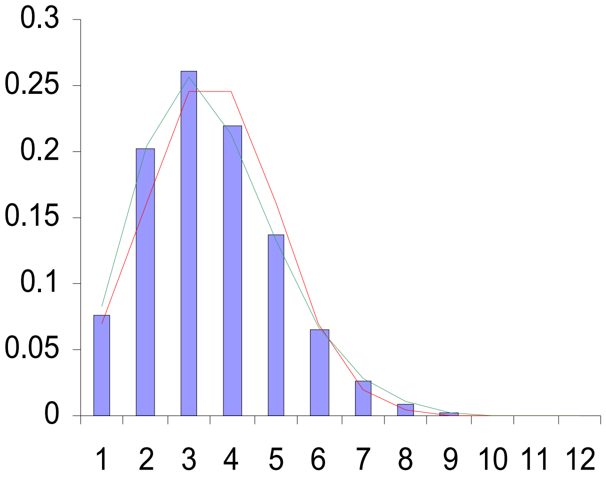




Approximating Binomial
p=0.01, *n* =10, 25, 50







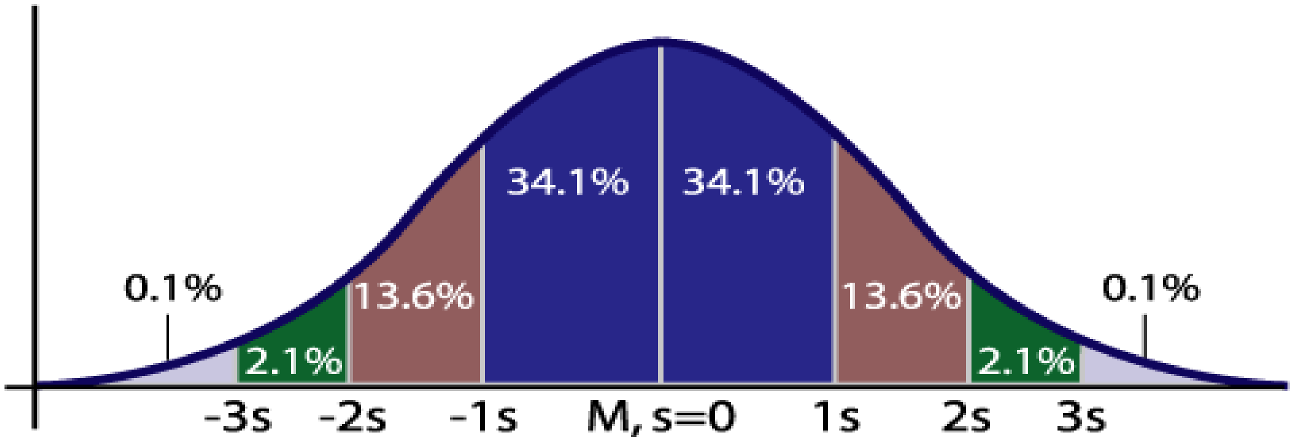
Standard Normal distribution

- Normal distribution with mean 0 and variance 1

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

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

Facts about Standard Normal



- $P(x \leq 1) = 0.84$ $P(-1 \leq x \leq 1) = 2\ 0.84 - 1 = 0.68$ (68%)
- $P(x \leq 2) = 0.977$ $P(-2 \leq x \leq 2) = 2\ 0.977 - 1 = 0.045$ (4.5%)
- $P(x \leq 3) = 0.999$ $P(-3 \leq x \leq 3) = 2\ 0.999 - 1 = 0.003$ (99.7%)
- $P(x \leq 1.64) = 0.95$ $P(-1.64 \leq x \leq 1.64) = 0.1$
- $P(x \leq 1.96) = 0.975$ $P(-1.96 \leq x \leq 1.96) = 0.05$
- $P(x \leq 2.57) = 0.995$ $P(-2.32 \leq x \leq 2.57) = 0.01$

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

http://www.math.unb.ca/~knight/utility/NormTble.htm

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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974

If k is Binomial with mean μ and variance σ^2

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

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

Problem

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

- Solution
  - Do it with Binomial
  - Use Normal approximation

Answer

- Coin is tossed 200 times. Estimate the chances that number of heads will deviate from 100 by 5 at most
- The parameters of the Binomial are μ=*np*=100 and variance $\sigma^2 = np(1-p)$=50 (then σ is 7.07)

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

$$P(105 \le k \le 95) = \Phi\left(\frac{105 - 100 + \frac{1}{2}}{7.07}\right) - \Phi\left(\frac{95 - 100 - \frac{1}{2}}{7.07}\right) = \Phi\left(\frac{5.5}{7.07}\right) - \Phi\left(\frac{-5.5}{7.07}\right) =$$

$\Phi(0.79) - \Phi(-0.79) = 1 - 2\,\Phi(0.79) = ???$

Number!

$$\Phi(0.778) – \Phi(-0.778) = 1 – 2\ \Phi(0.778) =$$

$$= 2\ 0.785 – 1 = 0.57$$

We leave exact computations and comparison for the exercises session

Task

- 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

Answer

- 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)

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

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

$1 - 0.93 = 0.07$

Task

- The frequency of some 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) ≥ 0.95
- P(at least one carrier) = 1 – P(no carriers) =

  $1 - 0.99^n$
- $1 - 0.99^n = 0.95$
- $0.99^n = 0.05$
- $n = \ln(0.05)/\ln(0.99) = 298.07$

Problem

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

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

$P(\geq 3 \text{ carriers}) = 1 - P(0 \text{ carriers}) - P(1 \text{ carrier}) - P(2 \text{ carriers}) =$

$1 - 0.99^{n} - n\ 0.01\ 0.99^{n-1} - \frac{1}{2}\, n\ (n-1)\ 0.01^{2}\ 0.99^{n-2}$

$1 - 0.99^{n} - n\ 0.01\ 0.99^{n-1} - \frac{1}{2}\, n\ (n-1)\ 0.01^{2}\ 0.99^{n-2} = 0.95$

$0.99^{n} - n\ 0.01\ 0.99^{n-1} - \frac{1}{2}\, n\ (n-1)\ 0.01^{2}\ 0.99^{n-2} = 0.05$

- ???

Use Normal approximation

- P(≥3 carriers) = P(>2 carriers)
- P(>2 carriers) ≥ 0.95 $$P(k > \alpha) = 1 - \Phi\left(\frac{\alpha - \mu + \frac{1}{2}}{\sigma}\right)$$
- $\mu = 0.01\, n$; $\sigma^2 = n\ 0.01\ 0.99$
- $1 - \Phi([2 - 0.01n + ½\ ]/[0.099n]) = 0.95$
- $\Phi([2 - 0.01n + ½\ ]/[0.099n]) = 0.05$
- $[2 - 0.01n + ½\ ]/[0.099\sqrt{n}] = -\ 1.64$
- $1.64\ 0.099\ \sqrt{n} - 0.01\ n = -\ 2\ ½$

Solving quadratic equation

- If there is equation of the form

  $A \sqrt{n} - B\, n = - C$

- The solution is

  $n = [A^2 + 2\, B\, C + A \sqrt{(A^2 + 4\, B\, C)}\,] / 2\, B^2$

Answer is…

- $1.64\ 0.099 \sqrt{n} - 0.01\ n = -2\ ½$
- $n = [0.026 + 0.05 + 0.058]/[0.0002] = 670.4$ (671)