

The central limit theorem

Sum of IID binary random variable

$$X_i = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p \end{cases}$$

$$S_n = \sum_{i=1}^n X_i$$

The distribution of S_n

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

Mean: $\mu(S_n) = np$

Standard Deviation: $\sigma(S_n) = \sqrt{np(1 - p)}$

Normalizing mean and variance

$$Z_n = \frac{S_n - \mu(S_n)}{\sigma(S_n)} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

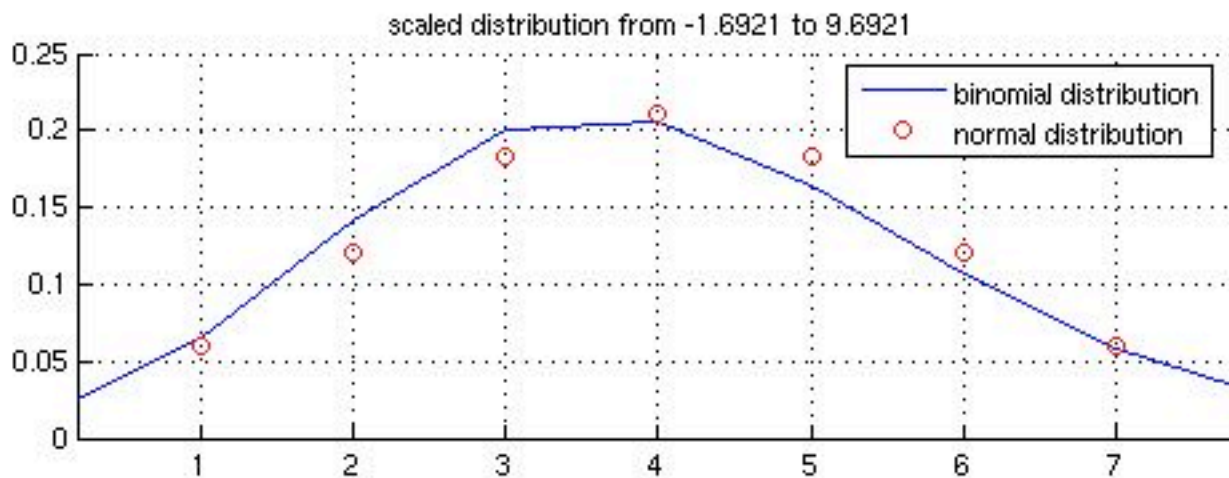
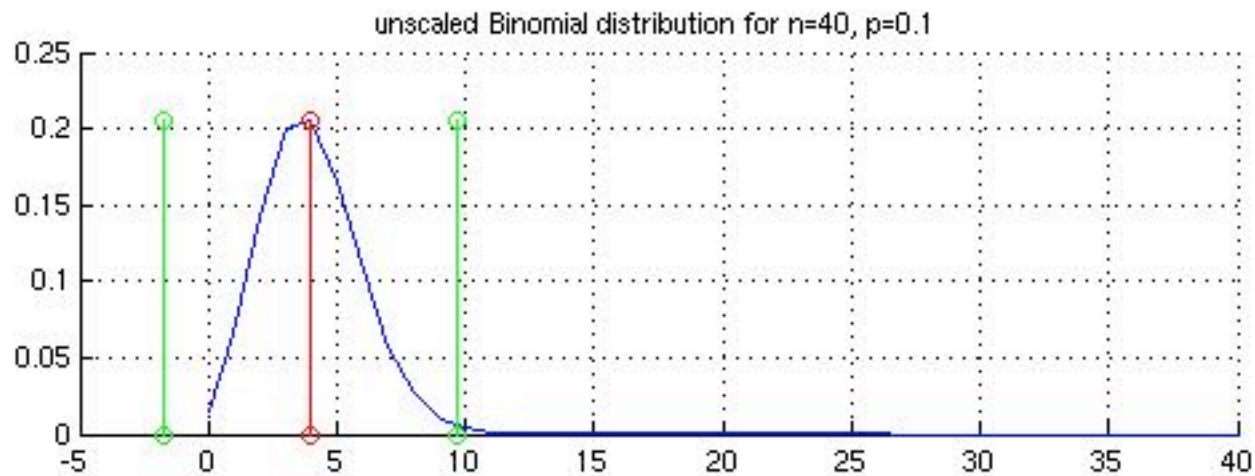
Shift

Scale

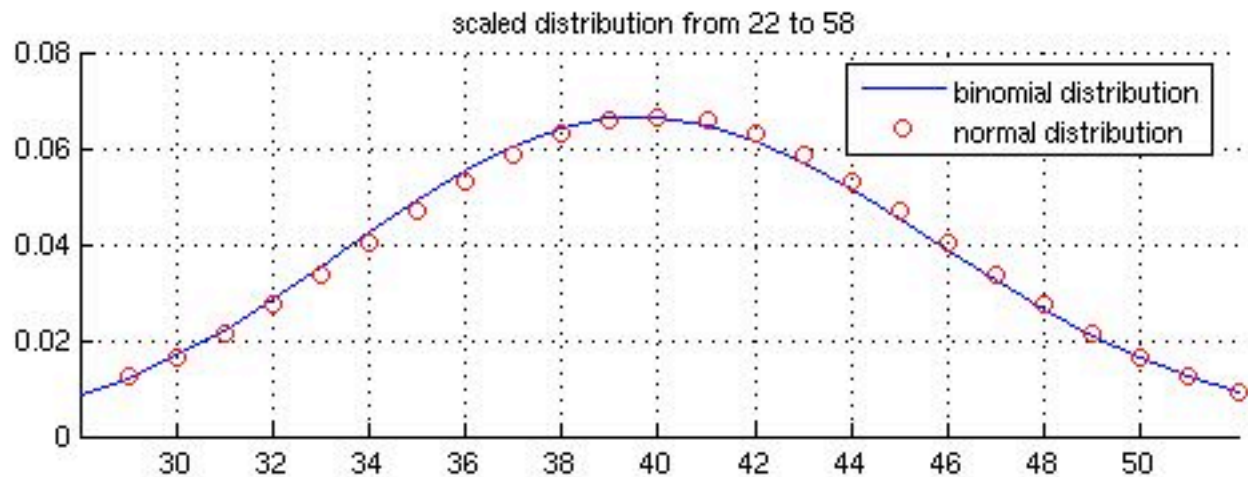
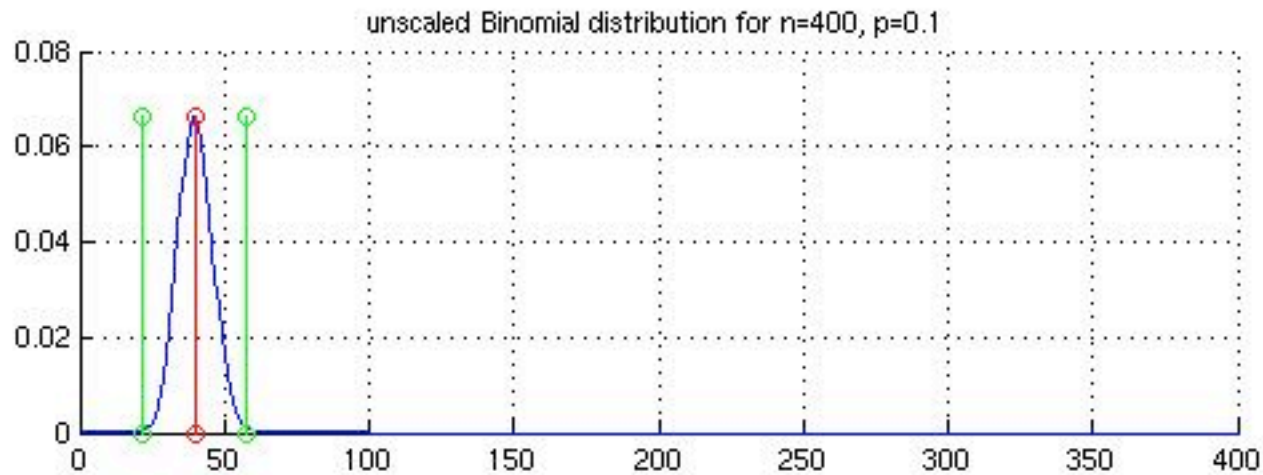
New mean = 0

New variance = 1

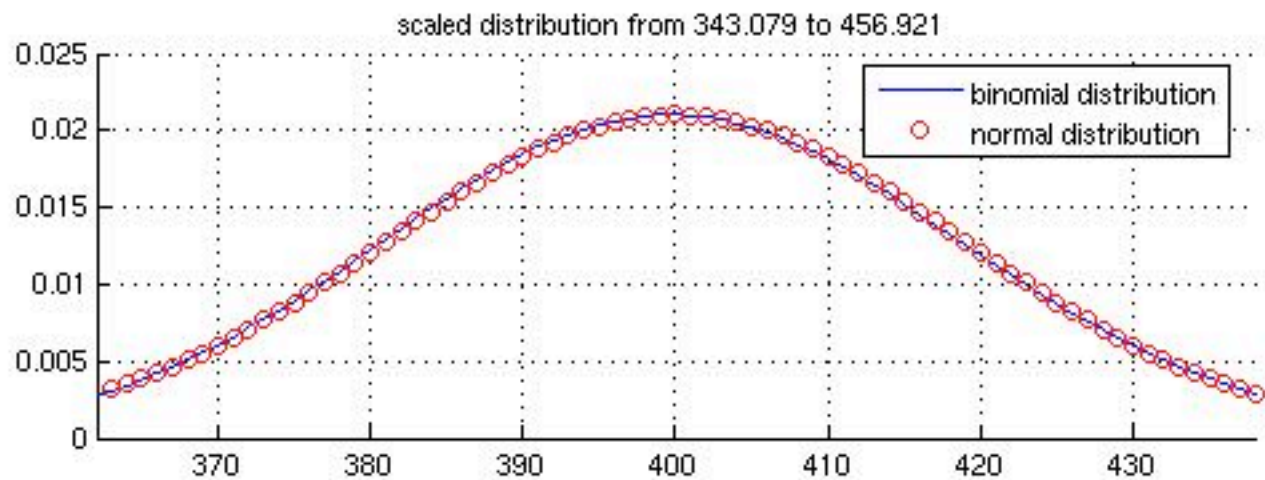
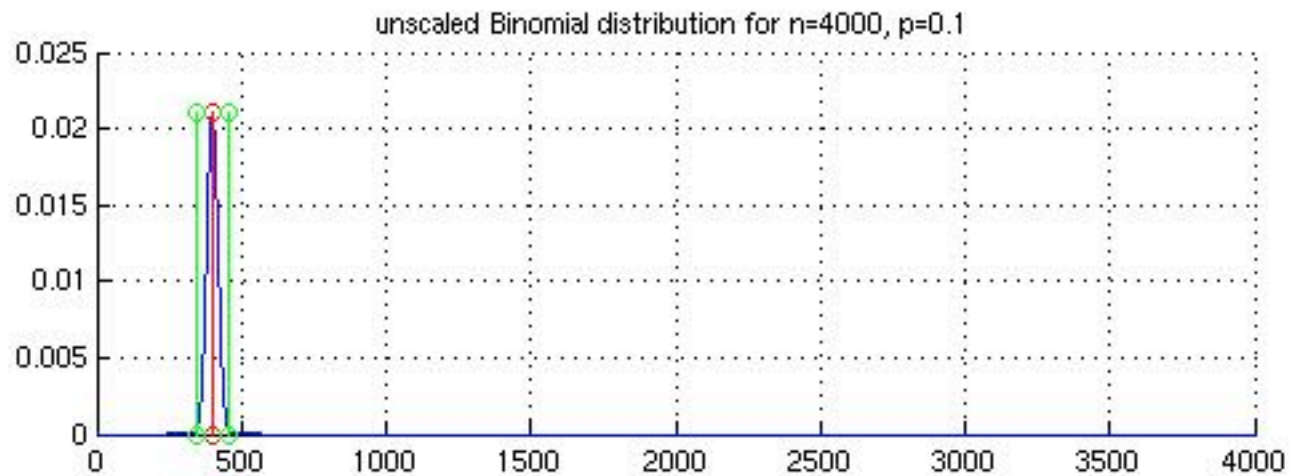
$$p=0.1, n=40$$



$p=0.1, n=400$



$$p=0.1, \quad n=4000$$



Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be IID Random variables with common mean μ and variance σ^2

$$\text{Define: } Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then the CDF of Z_n converges to the standard normal CDF:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx,$$

In the sense that

$$\forall z, \quad \lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z)$$

Probability vs. Statistics

- **Probability Theory:** model → data
mechanism is given
infer
probability of events (=data)
- **Statistics:** data → model
data is given
infer
properties of mechanism

From binomial distribution to inferring probabilities

Given p, n we know that the distribution of $\frac{S_n}{n}$ is

$$\mathcal{N}(\mu, \sigma^2) = \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Given that **m** out of **n** coin tosses came up head, what is the probability of head?

$$\frac{m}{n} \pm \frac{c}{\sqrt{n}}$$

Counting butterflies

- About 100,000 monarch butterflies hanging on 5-10 trees.
- How do you count them?
- Catch 1,000, tag them and free them
- wait for them to fly around and mix
- Catch another 1,000 and count how many of those are tagged.

Selecting sample points in the Falluja, Iraq



Interview people in 30 houses closest to each **X**



Image © 2006 DigitalGlobe

©2006 Google

enter 33° 21'04.94"N 43° 47'01.69"E

Streaming ||||| 100%

Eye alt 27959

Statistical tests

- The more specific the question - the more accurate the answer.
- Rather than asking for an **estimate** of p , we want to have some confidence that $p > q$
- q is the null hypothesis

Example - proving efficacy of drug

- Company X wants to prove to the FDA that their drug is effective against relapse of disease Y.
- Suppose we know that the general probability of a relapse within a year is q
- Out of n randomly selected patients, m have a relapse within a year.

Significance test

- Skeptical opinion: no effect = probability of relapse when using drug remains q .
- q is the null hypothesis.
- To reject the null hypothesis we need to show that m/n is significantly smaller than q

p-values

- What threshold should we use to reject the null hypothesis.
- **p-value** for threshold **T**: the probability of observing a sample with **$D \geq T$** under the null hypothesis

The χ^2 test

$$D = \frac{(m - np)^2}{np} + \frac{(n - m - (1 - p)n)^2}{n(1 - p)} = \frac{(m - np)^2}{np(1 - p)} = \left(\frac{m - \mu(m)}{\sigma(m)} \right)^2$$

- $p=0.1, n=100, m=5$ $D=2.8$
- $p=0.1, n=1000, m=50$ $D=27.8$
- Increasing n , keeping n/m fixed increases significance
- Significance increases with D .

The χ^2 Distribution

If X_i are k independent, normally distributed random variables with means μ_i and variances σ_i^2 , then the random variable

$$Q = \sum_{i=1}^k \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

is distributed according to the chi-square distribution. This is usually written

$$Q \sim \chi_k^2.$$

The χ^2 test for multinomials

t possible outcomes.

null hypothesis: $p_1, \dots, p_t$ $\sum_{i=1}^t p_i = 1$

we take n independent measurements and
get counts

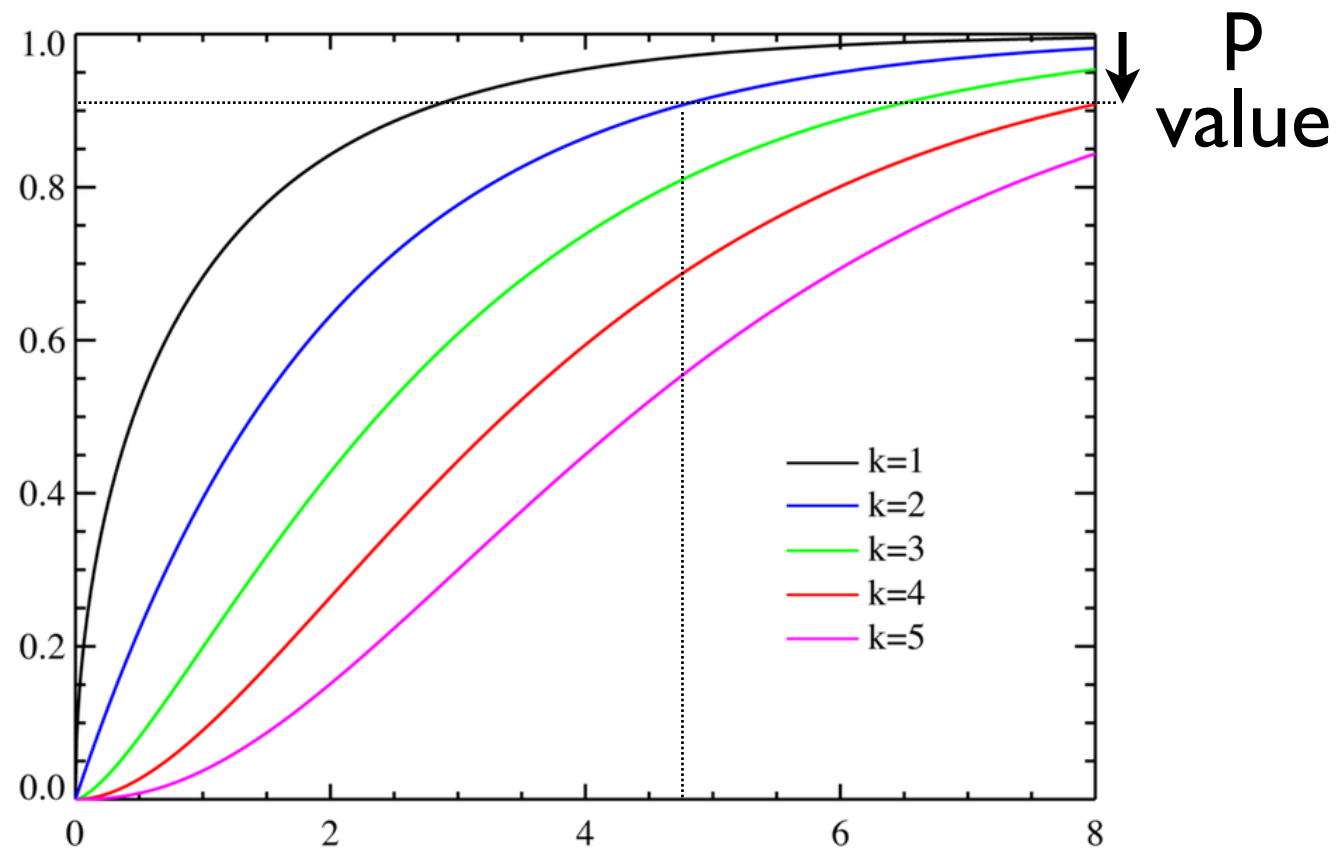
$$m_1, \dots, m_t \quad \sum_{i=1}^t m_i = n$$

The Pearson χ^2 statistic with t-1 degrees
of freedom:

$$D = \sum_{i=1}^t \frac{(m_i - np_i)^2}{np_i}$$

Significance of the χ^2 test

- Assume that n is large enough that the multinomial distribution is approximately normal (central limit theorem).
- D has a χ^2 distribution with $k=t-1$ degrees of freedom.



Using the χ^2 test to compare histograms

- If we partition the reals into t predefined bins, we get a multinomial distribution.
- We can use the chi-square test to compare histograms.
- Is n large enough?
- If t too large - we cannot assume normality
- If t too small - we miss details
- **Recommendation:** use the KS test to compare the PDFs

Summary

- **Central limit theorem:** summing many independent random variables with finite mean and variance creates a distribution close to the normal distribution.
- Sample means are approximately normal.
- Careful experimental design is needed to insure that observations are independent.
- The chi-square test measures the deviation of a sample from a normal null distribution.