

Statistical Variation

Roberto Puch-Solis

FORSTAT Workshop
Edinburgh, Scotland, June 13th - 14th 2008

A sexual assault case

- ◆ A semen stain was found in victim's underwear
- ◆ A Y-STR profile was obtained from the stain and matches the Y-STR profile of the suspect
- ◆ Evidential weight is calculated as a proportion of people having the profile in a population

A sexual assault case

The statistician working for the defence wrote:

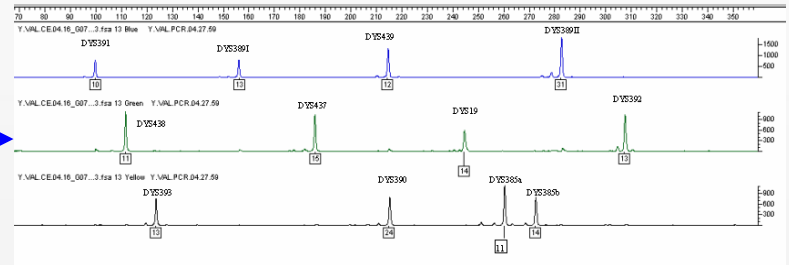
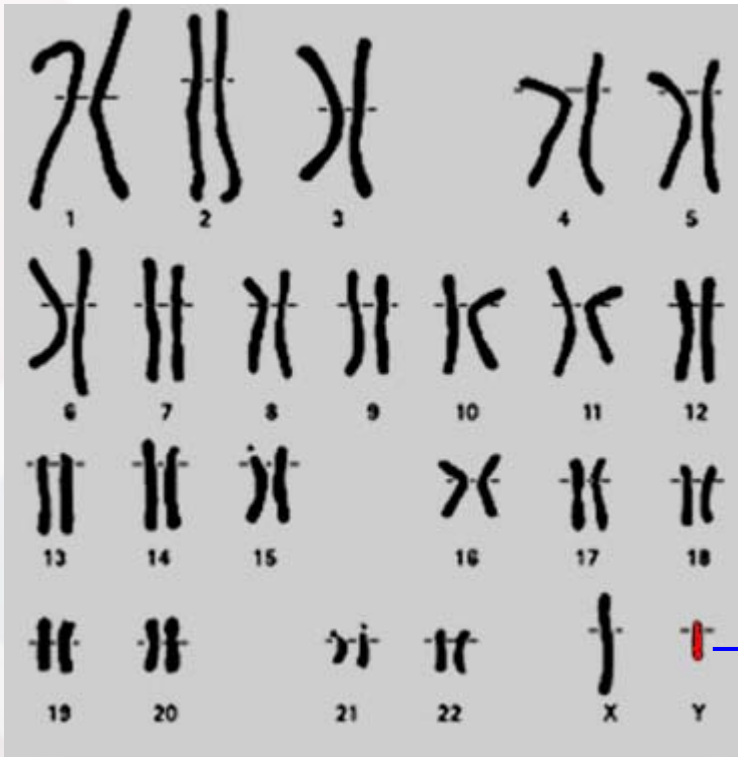
“ ...Mr Smith’s Y STR profile has also been observed in 40 profiles from a European database of 5,000 Euro-Asian profiles. Taking statistical variation into account, the latter figure is consistent with a Euro-Asian frequency of the profile of about 1 man in every 100 (\approx upper 97.5% confidence limit)”

The question that the practitioner asked was:

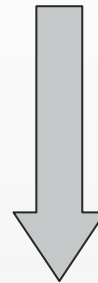
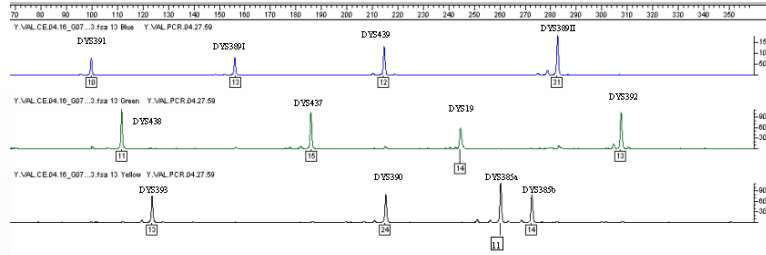
What does “*taking statistical variation into account*” mean?

In this talk, the statistical tools for answering this question will be presented

Y-STR Profile



Y-STR Profile



Y-STR Site	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385a	DYS385b	DYS437	DYS438	DYS439
Crimestain	14	13	31	24	10	13	13	11	14	15	11	12

Y-STR Profile

- ◆ From DNA located on the male Y chromosome
- ◆ Inherited solely from the father
- ◆ All males who are related by a paternal link will have the same Y-STR profile, up to mutations
- ◆ More males chosen at random from the population will have the same Y-STR profile, in comparison to autosomal DNA profiles

Calculating a proportion

The proportion of men having this Y-STR profile in question is:

$$\frac{40}{5,000} = 0.008$$

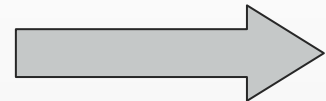
The proportion of men not having this Y-STR profile is:

$$\frac{4,960}{5,000} = 0.992$$

Notation; shorthand



Y-STR Site	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385a	DYS385b	DYS437	DYS438	DYS439
Crimestain	14	13	31	24	10	13	13	11	14	15	11	12



γ



A probability distribution

The probability of a man having Y-STR profile γ is:

$$\Pr(\Gamma = \gamma) = 0.008$$

The probability of a man not having Y-STR profile γ is:

$$\Pr(\Gamma = \text{not } \gamma) = 0.992$$

A probability distribution

A probability distribution for Γ :

$$\Gamma = \begin{cases} \gamma & \text{with probability } 0.008 \\ \text{not } \gamma & \text{with probability } 0.992 \end{cases}$$

A discrete probability distribution

Γ is an example of discrete random variable:

- ◆ Variable
 - ◆ Takes several values: γ and not γ
- ◆ Random
 - ◆ It takes values according to probabilities: 0.008 and 0.992
- ◆ Discrete
 - ◆ The values it takes are either finite or countably infinite
 - ◆ Finite set $\{1,2,3\}$, $\{a,b,c\}$.
 - ◆ Countably infinite: $\{1,2,3,5,6,\dots\}$

A discrete probability distribution

An example of a discrete probability distribution

$$\Gamma = \begin{cases} \gamma & \text{with probability } 0.008 \\ \text{not } \gamma & \text{with probability } 0.992 \end{cases}$$

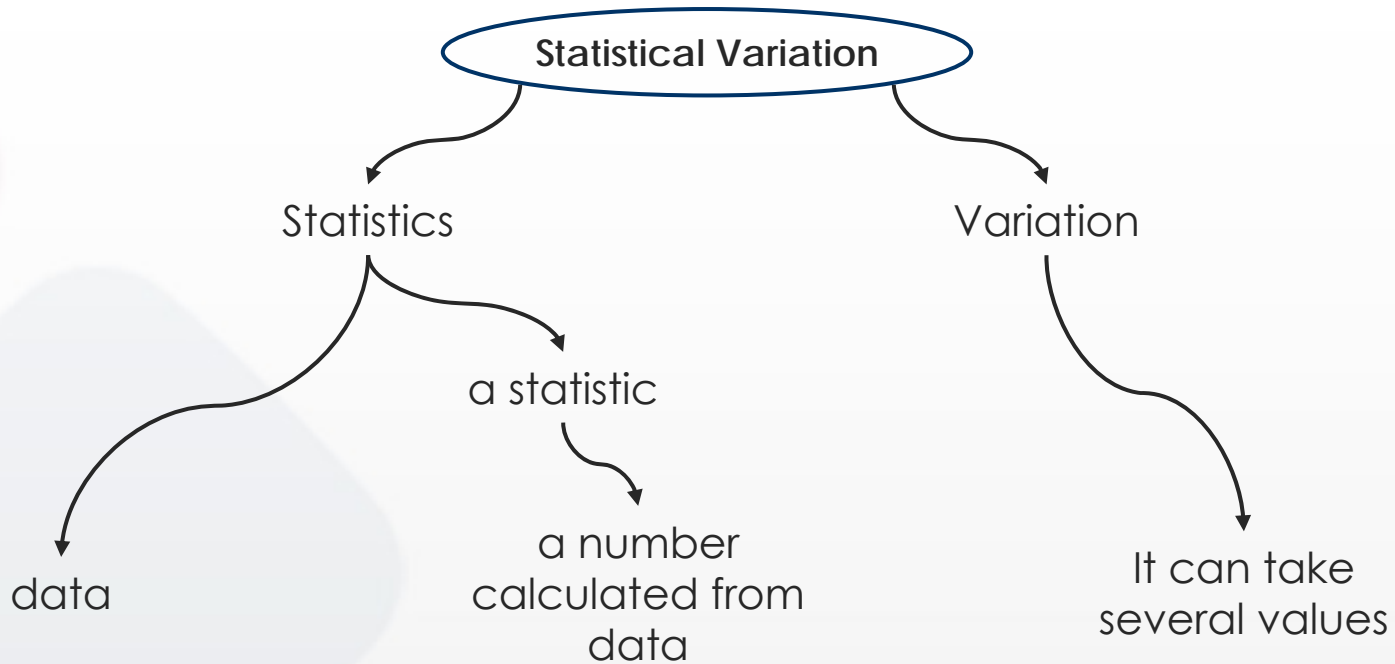
A discrete probability distribution

Important remarks:

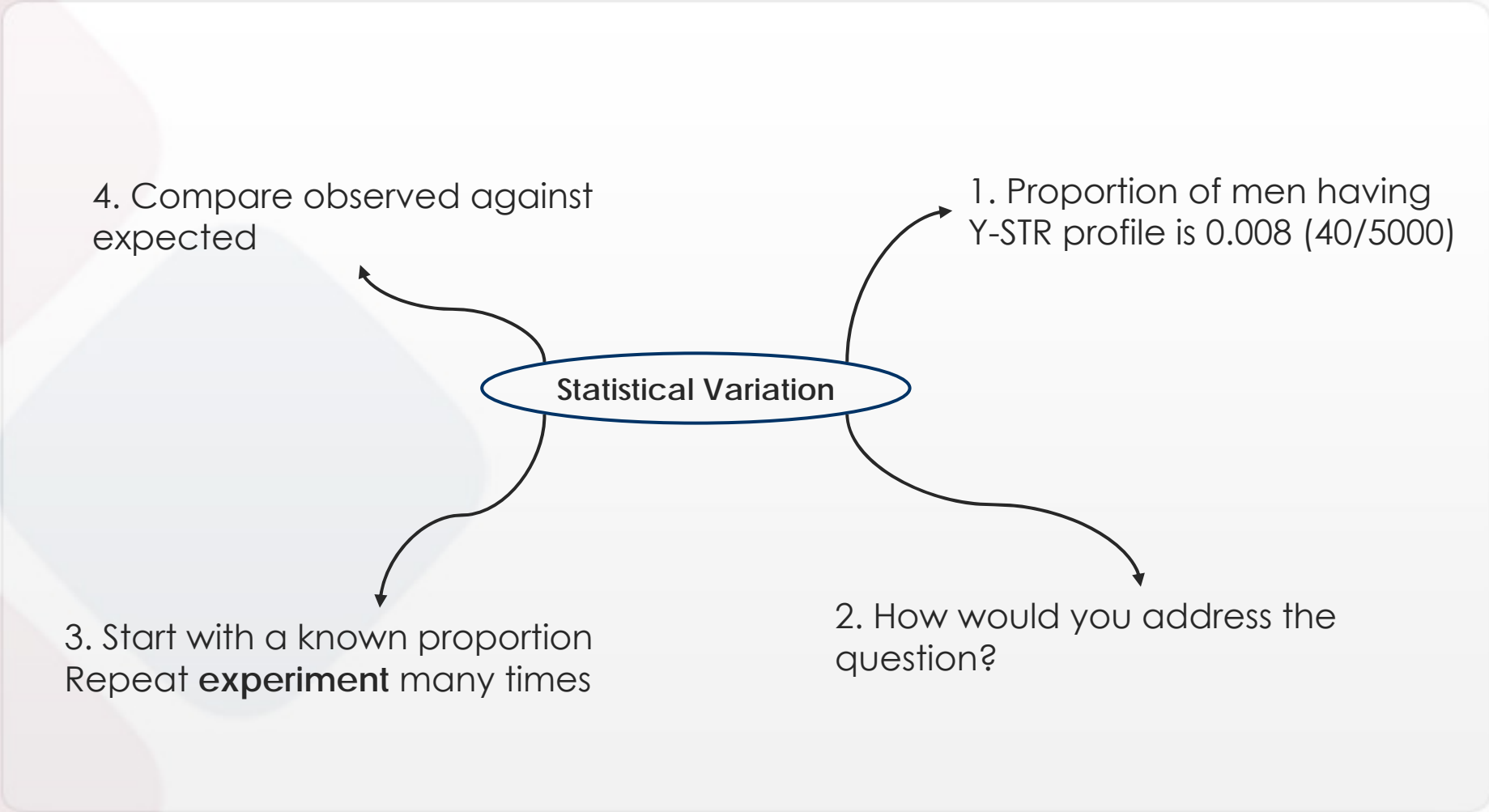
- ◆ The set of values is exhaustive: all values are considered
- ◆ Probabilities add to one
- ◆ A discrete probability distribution is also called probability mass function, abbreviated to pmf

$$\Pr(\Gamma = \gamma) = 0.008$$

$$\Pr(\Gamma = \text{not } \gamma) = 0.992$$



- It is about values that a statistic can take when computed from different data sets
- The more data the more precise the statistic



4. In statistical terms:
computer simulation or simply, **simulation**

1. Toss a fair coin 100 times;
Known proportion: 0.5

Computer experiment

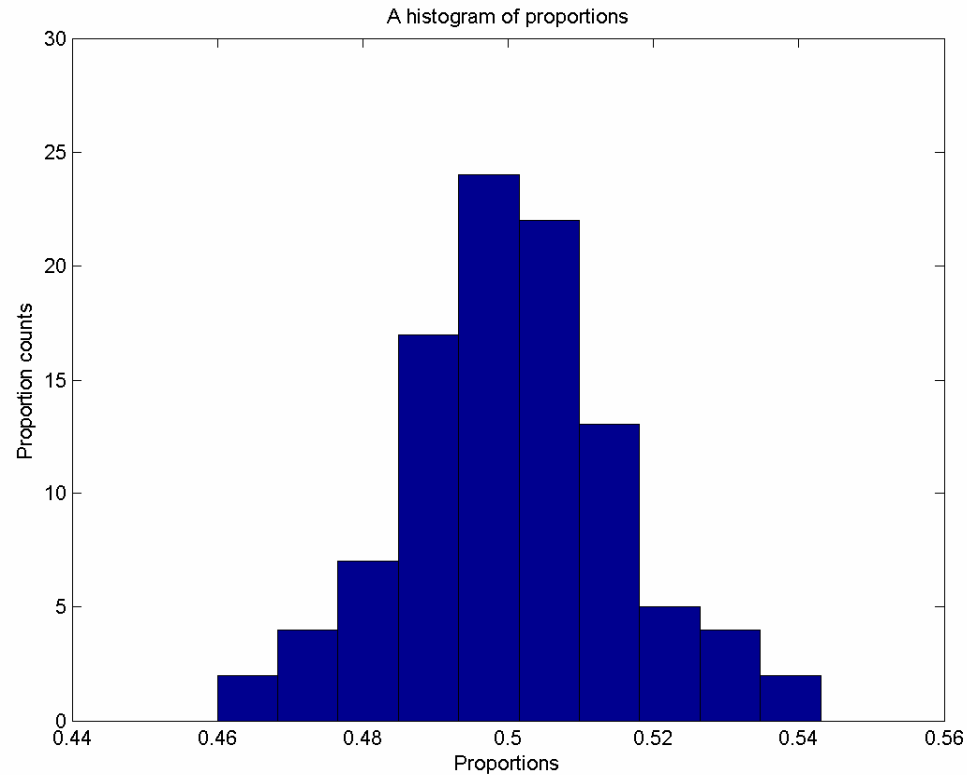
3. Do it a again 1,000 times

2. Compute the proportion of heads

Statistical variation

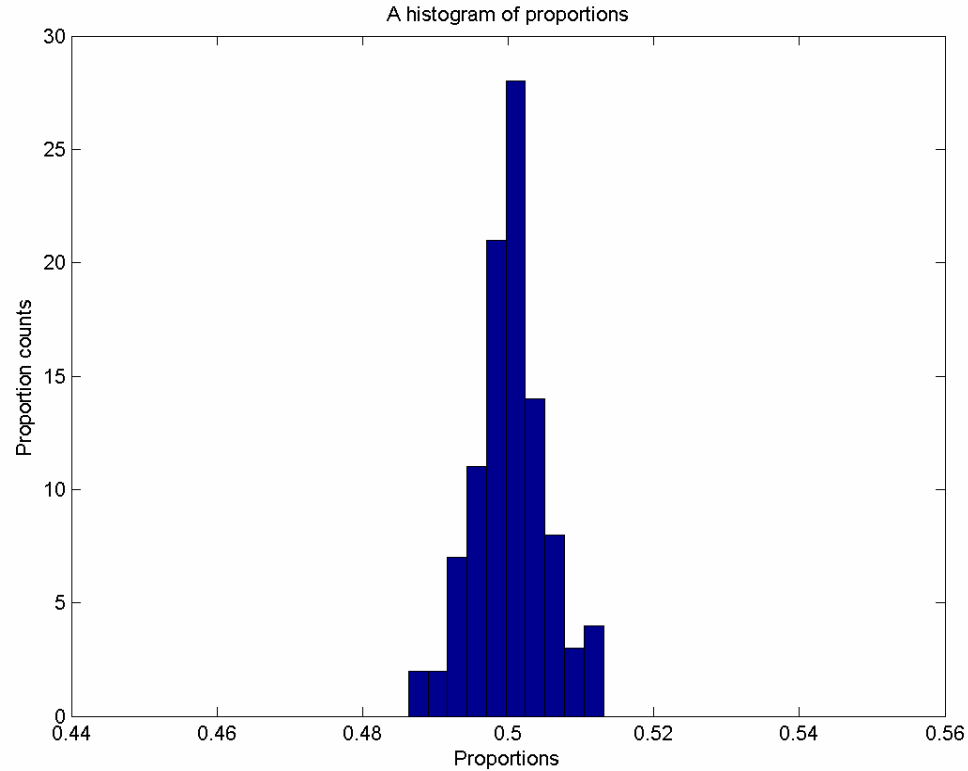
Prop	Order Prop	Rep Prop	Count
0.519	<u>0.460</u>	0.460	1
0.494	0.467		
0.512	0.470		
0.489	0.472	0.470	5
0.501	0.473		
0.482	<u>0.473</u>		
0.503	0.477		
0.500	0.479		
0.515	0.482		
0.472	0.482	0.480	7
0.484	0.482		
0.496	0.483		
0.521	<u>0.484</u>		
0.506	0.486		
0.500	0.486		
0.503	0.488		
0.477	0.488		
0.509	0.488		
0.508	0.488	0.490	12
M	M	M	

Statistical variation



A histogram of 1,000 proportions where each proportion was calculated with 100 observations

Statistical variation



A histogram of 1,000 proportions where each proportion was calculated with 1,000 observations

Brief historical note

Normal distribution; Gaussian distribution

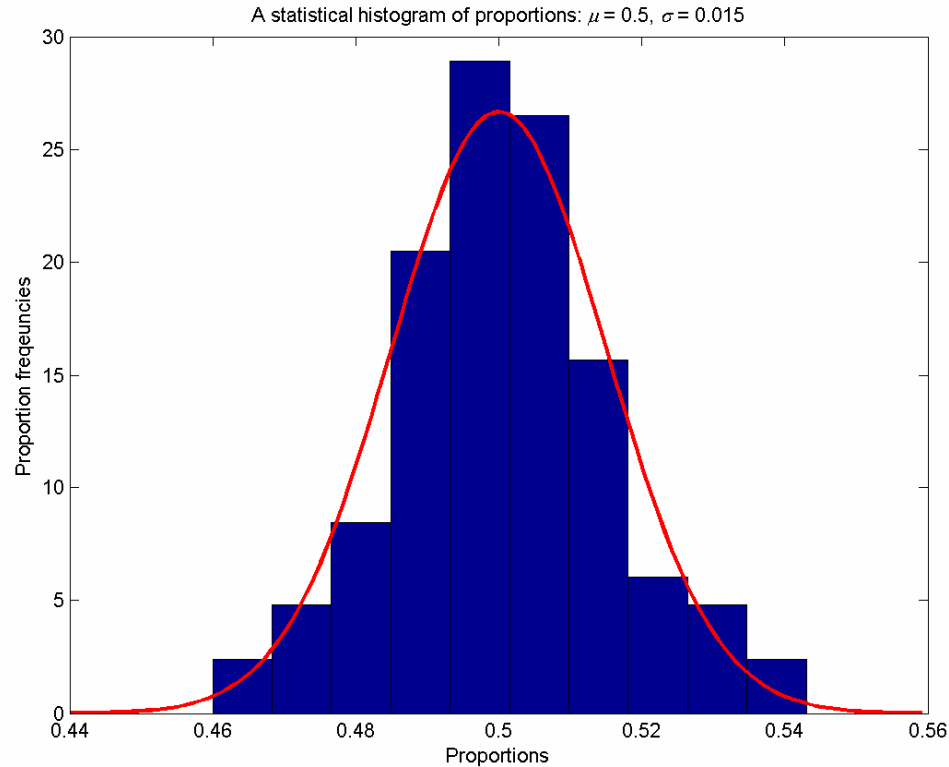
Abraham de Moivre, France

Bell-shaped distribution

Johann Carl Friedrich Gauss, Germany

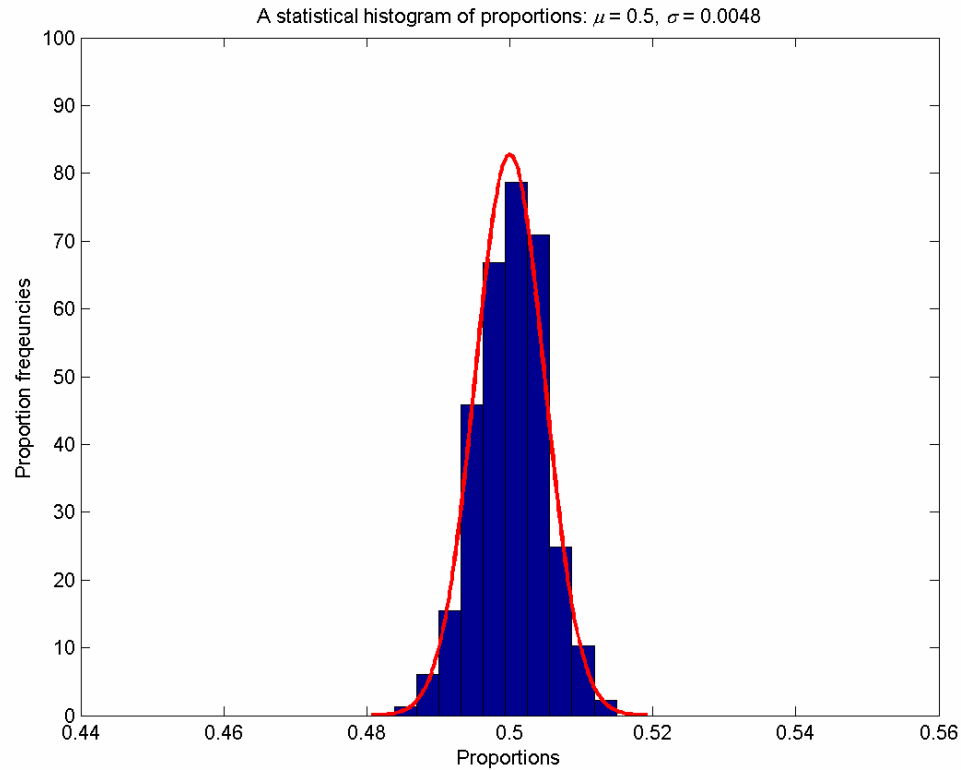
Francis Galton, UK

Quantification of statistical variation



- Statistical histogram of 1,000 proportions
- Each proportion was calculated with 100 observations.
- A Normal distribution with $\mu=0.5$ and $\sigma=0.015$

Quantification of statistical variation



- Statistical histogram of 1,000 proportions
- Each proportion was calculated with 1,000 observations.
- A Normal distribution with $\mu=0.5$ and $\sigma=0.048$

$$X \sim N(\mu, \sigma^2)$$

X is normally distributed
with mean μ and variance σ^2 (or standard deviation σ)

Normal distribution

$$f(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

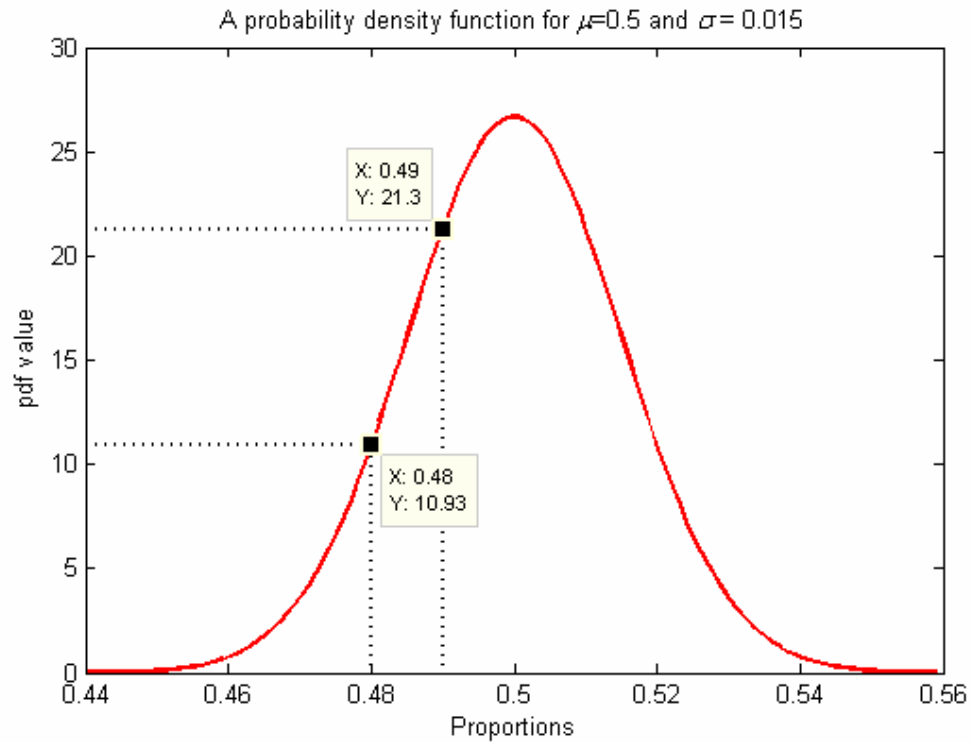
constant $e \approx 2.72$

constant $\pi \approx 3.14$

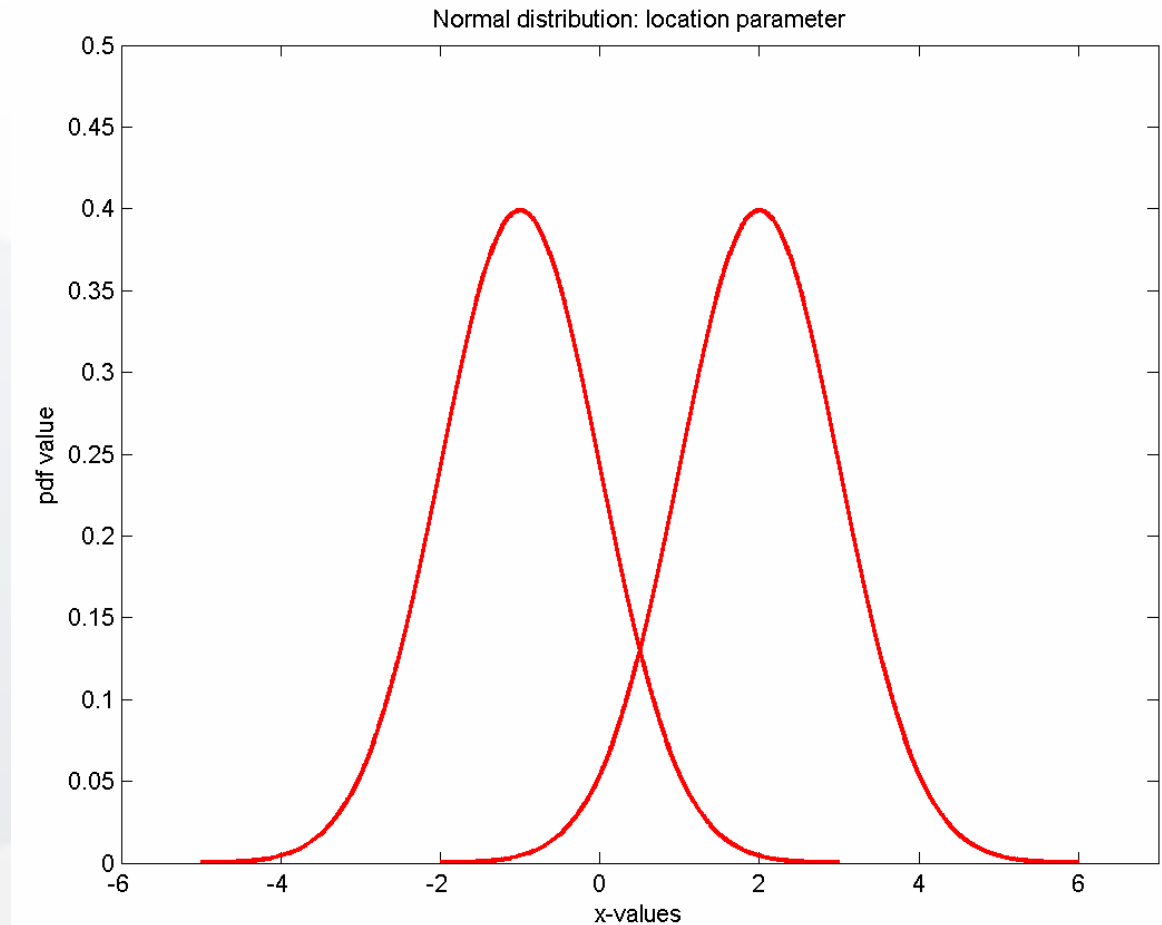
Location
parameter μ

Scale
parameter σ

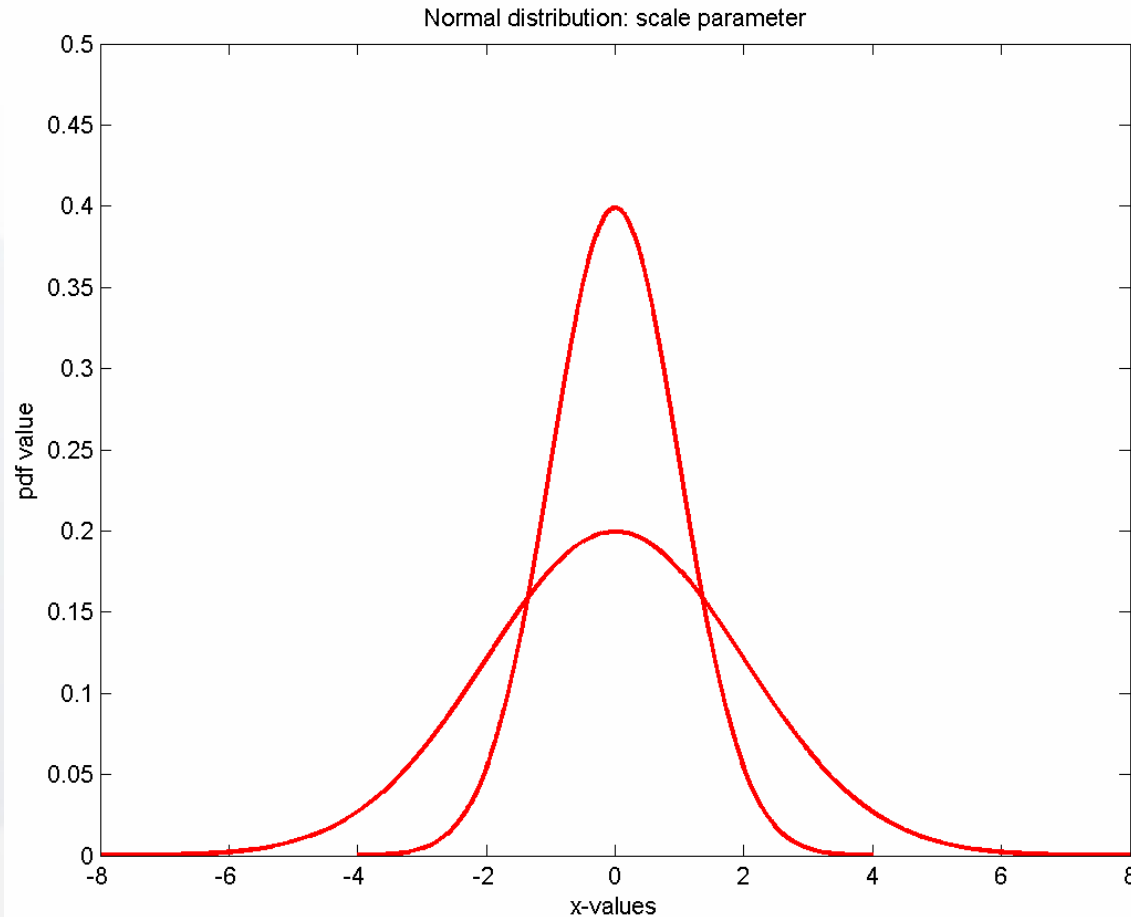
Normal distribution



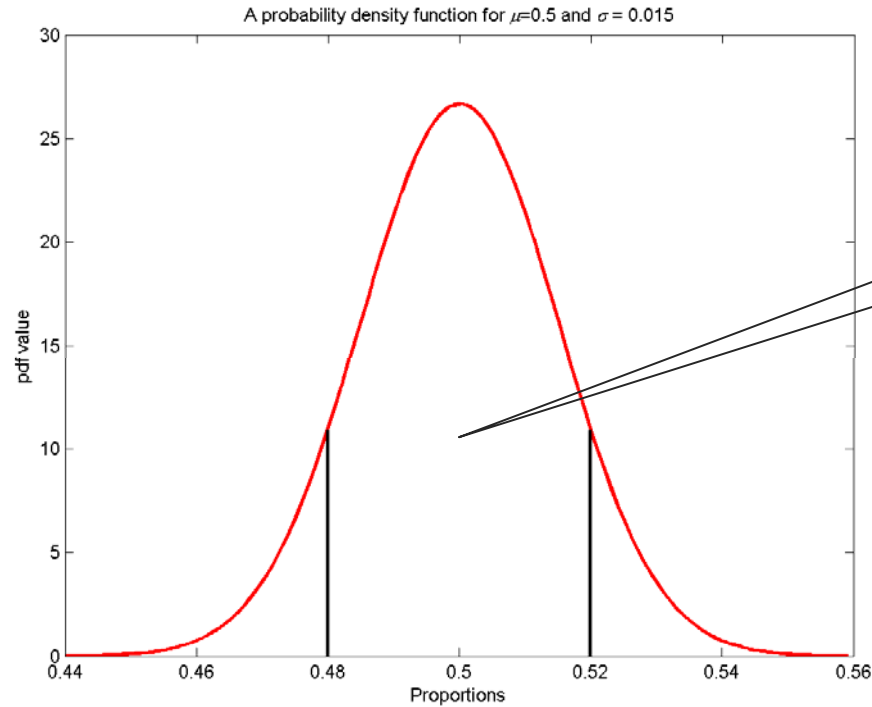
Normal distribution: location parameter



Normal distribution: scale parameter

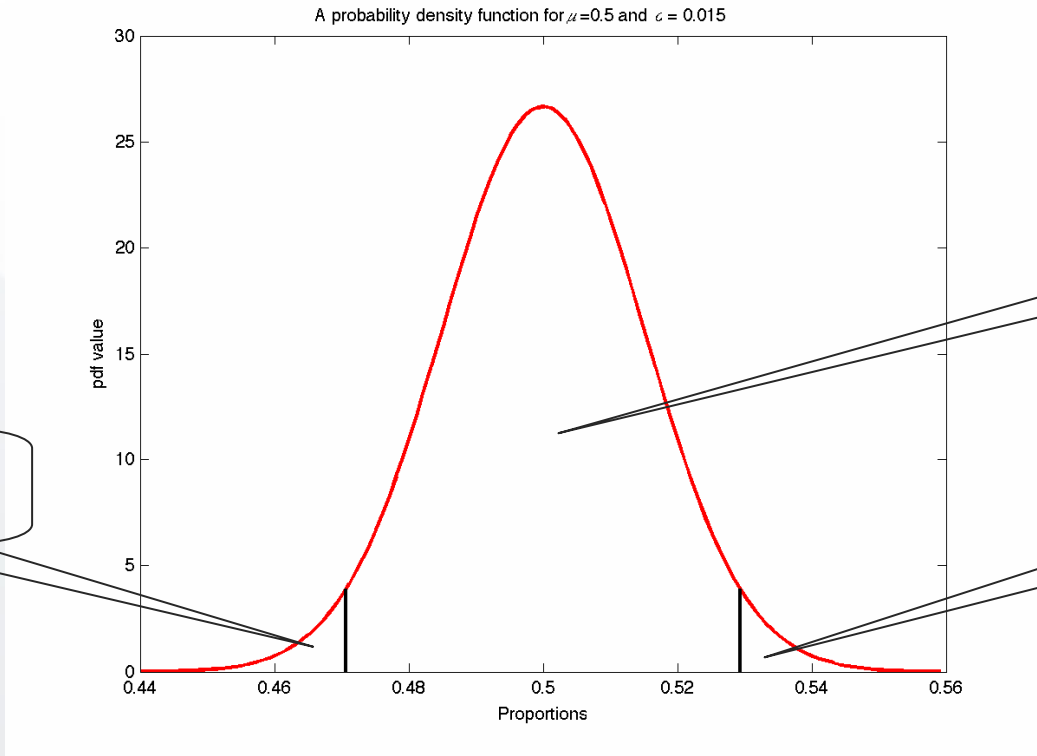


Normal distribution



$$\Pr(0.48 \leq p \leq 0.52) \approx 0.82$$

Normal distribution



$$\Pr(x_1 \leq p \leq x_2) = 0.95$$

$$X \sim N(\mu, \sigma) \Leftrightarrow \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$X \sim N(0, 1)$$

Standard Normal distribution

$\mu = 0$ and $\sigma = 1$

X follows a standard normal distribution

Use the best quantity we have: observed proportion; how?

4. But, what can we do when the proportion is not known?

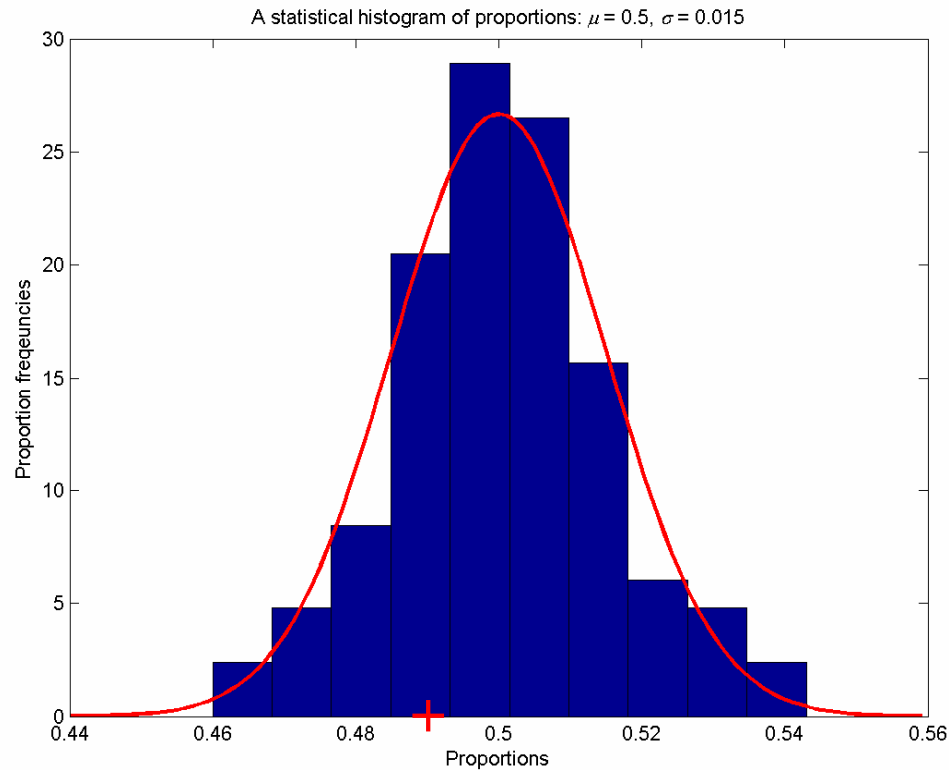
1. We ran an experiment

Where are we?

3. We assumed a known proportion

2. Discovered the normal distribution

Confidence Interval for a proportion



Observed proportion is 0.49

Random variable Y has mean μ and variance σ^2

Central limit theorem

The sample mean (average) \bar{y} , based on n observations,
has an approximate normal distribution with
mean μ and variance σ^2/n , for large n ,

$$\bar{Y} \sim N\left(\mu, \sigma^2/n\right)$$

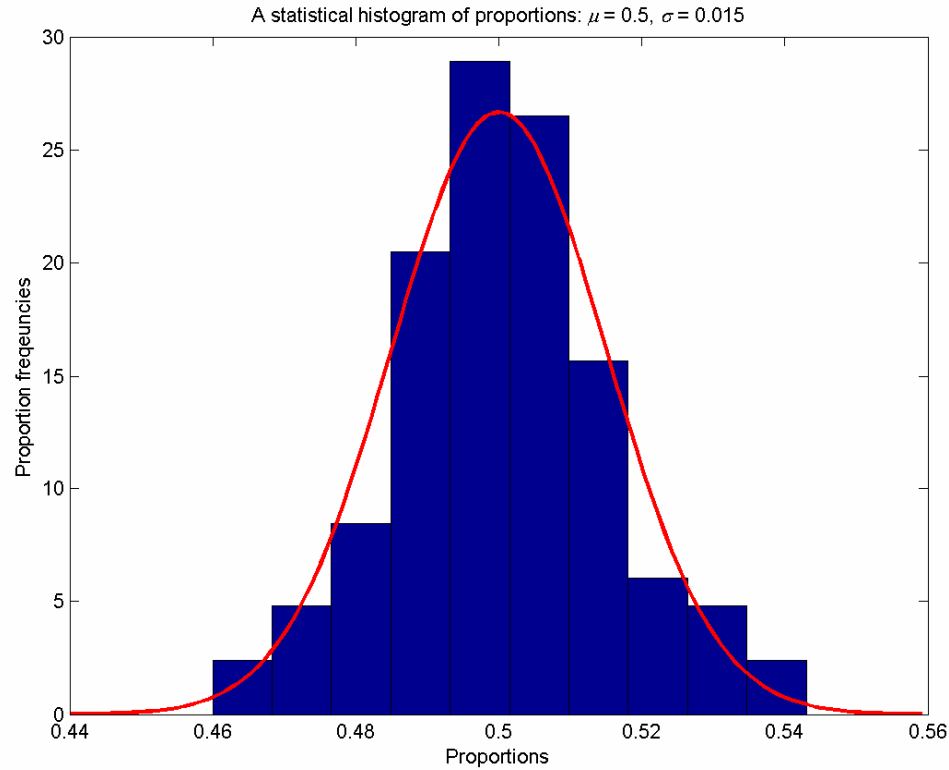
Confidence interval for a proportion

$$Y = \begin{cases} 0 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5 \end{cases}$$

Ten observations: 0,1,0,0,1,1,0,0,1,0

$$\text{The mean } \bar{y} = \frac{0+1+0+0+1+1+0+0+1+0}{10} = \frac{4}{10} = 0.4 = \hat{p}$$

Confidence Interval for a proportion



Therefore the bell-shaped distribution

Confidence interval for a proportion

$$Y = \begin{cases} 0 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5 \end{cases}$$
$$\begin{aligned} \mu &= E[Y] \\ &= 0 \times 0.5 + 1 \times 0.5 \\ &= 0.5 \end{aligned}$$
$$\begin{aligned} \sigma^2 &= E[Y^2] - E[Y]^2 \\ &= (0 \times 0.5 + 1^2 \times 0.5) - 0.5^2 \\ &= 0.025 \end{aligned}$$
$$\bar{Y} \sim N\left(0.5, \frac{0.025}{n}\right)$$

Confidence interval for a proportion

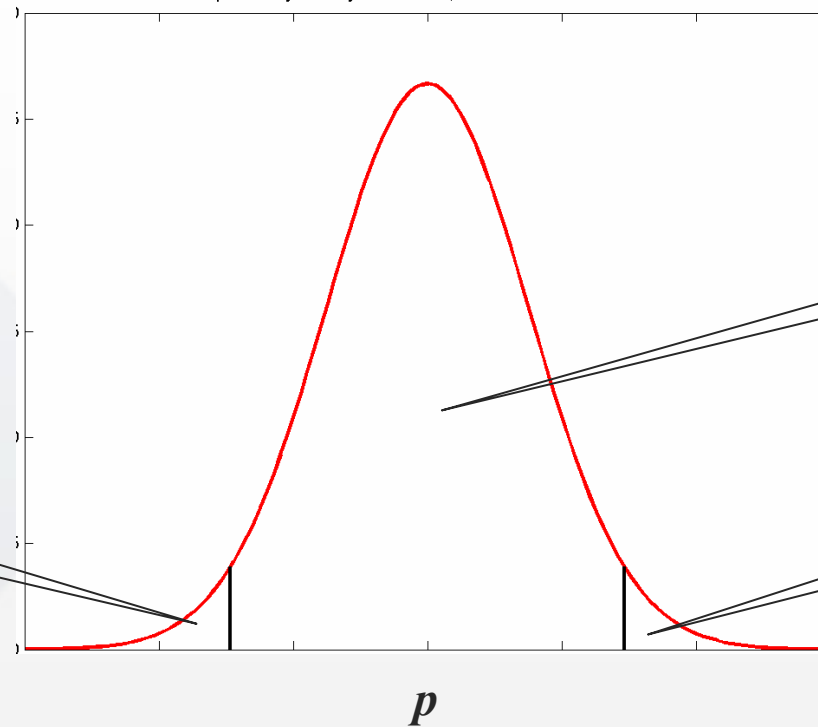
$$Y = \begin{cases} 0 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

$$\begin{aligned} \mu &= E[Y] \\ &= 0 \times (1-p) + 1 \times p \\ &= p \end{aligned}$$

$$\begin{aligned} \sigma^2 &= E[Y^2] - E[Y]^2 \\ &= (0 \times (1-p) + 1^2 \times p) - p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

$$\bar{Y} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

A Normal distribution for \bar{Y}



$$\Pr(x_1 \leq p \leq x_2) = 0.95$$

Confidence Interval for a proportion

α is the confidence level

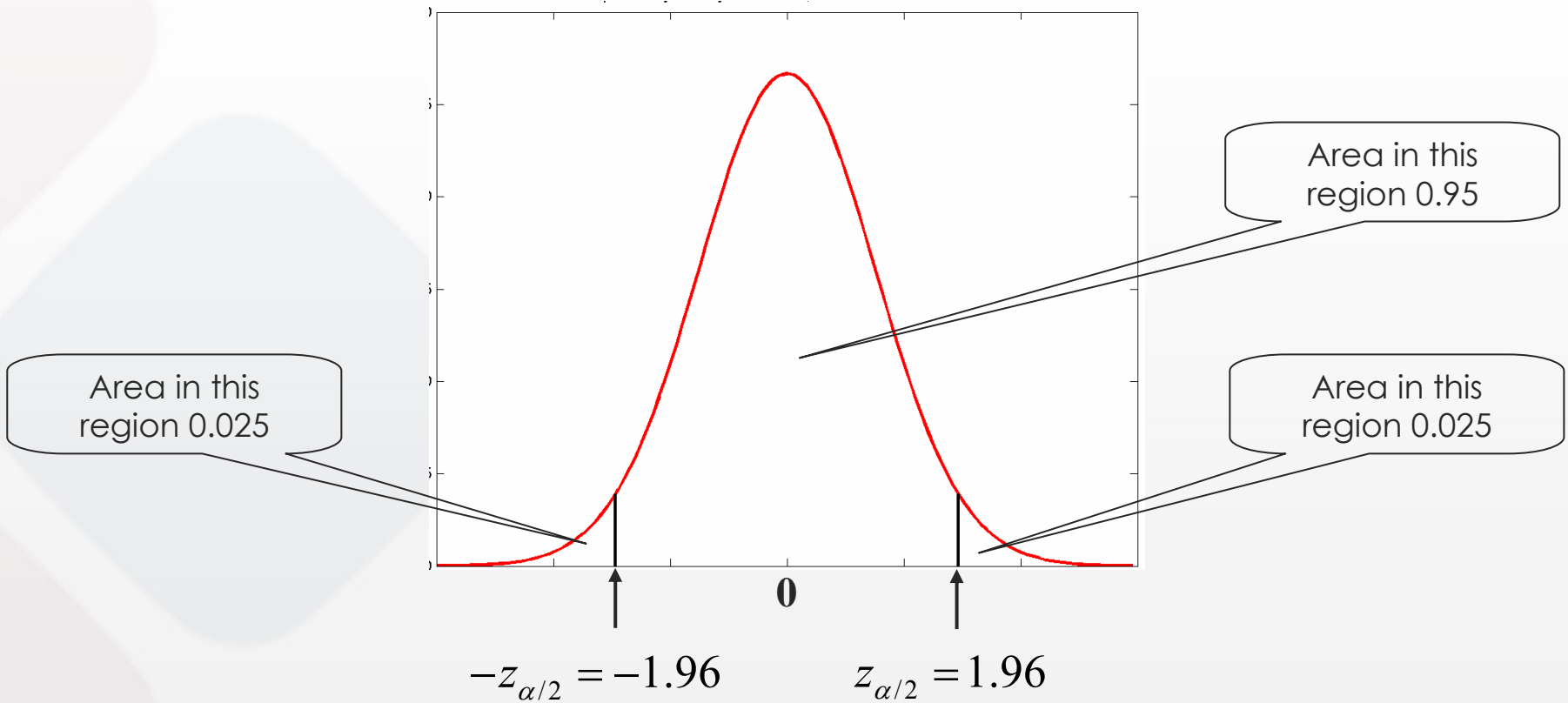
\hat{p} is the observed proportion

$$\left[\hat{p} - z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

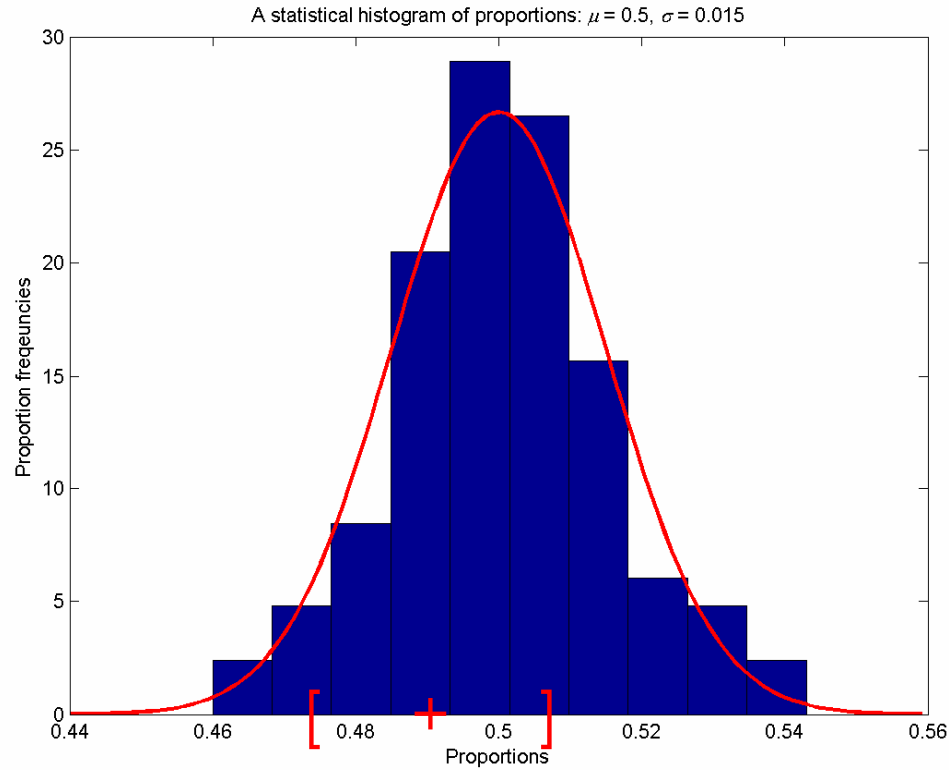
$z_{\alpha/2}$ is obtained from a standard normal distribution

n is the number of observations

The standard normal distribution

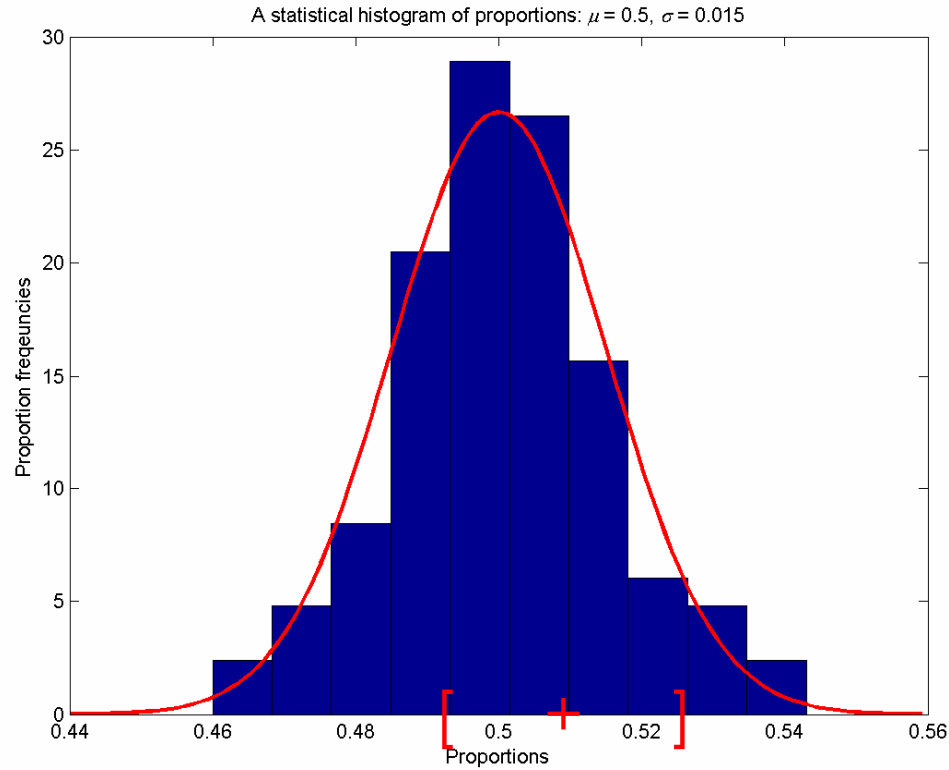


Confidence Interval for a proportion



Observed proportion is 0.49

Confidence Interval for a proportion



Observed proportion is 0.51

“ ...Mr Smith’s Y STR profile has also been observed in 40 profiles from a European database of 5,000 Euro-Asian profiles. Taking statistical variation into account, the latter figure is consistent with a Euro-Asian frequency of the profile of about 1 man in every 100 (\approx upper 97.5% confidence limit)”

Back to the case

- $\hat{p} = \frac{40}{5,000} = 0.008$
- $n = 5,000$
- $\hat{p}(1 - \hat{p}) = 0.008 \times 0.992 = 0.0079$
- $z_{\alpha/2} = 1.96$
- $z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 1.96 \times \sqrt{\frac{0.0079}{5,000}} = 0.0025$

Back to the case

$$\left[\hat{p} - z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

$$[0.008 - 0.0025, 0.008 + 0.0025]$$

$$[0.0055, 0.0105]$$

Lower 2.5% confidence limit is 0.0055

Upper 97.5% confidence limit is 0.0105

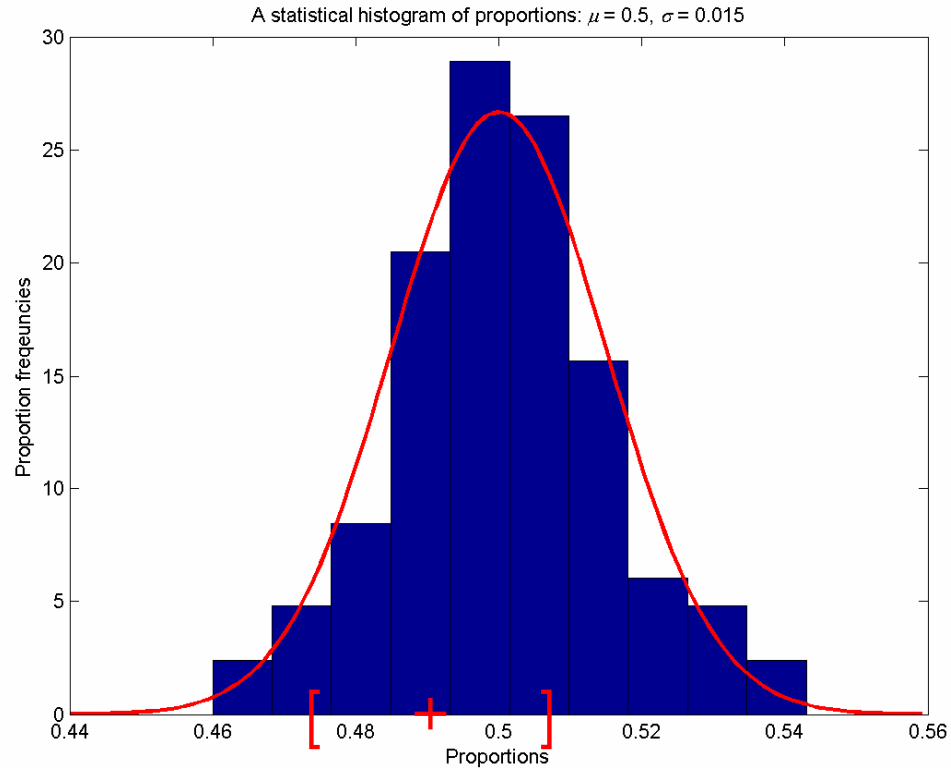
$$0.0105 \approx 0.01 = \frac{1}{100}$$

One man in every
hundred

“ ...Mr Smith’s Y STR profile has also been observed in 40 profiles from a European database of 5,000 Euro-Asian profiles. Taking statistical variation into account, the latter figure is consistent with a Euro-Asian frequency of the profile of about 1 man in every 100 (\approx upper 97.5% confidence limit)”

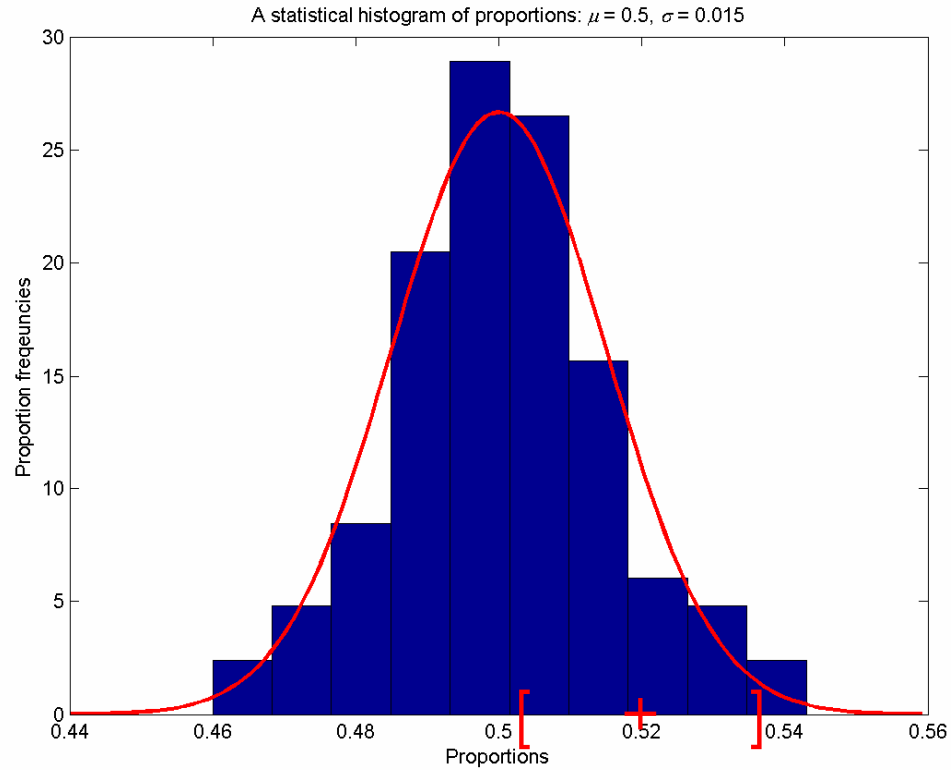
Why the upper 97.5% confidence limit? Because it is favourable to the defendant.

Confidence Interval for a proportion

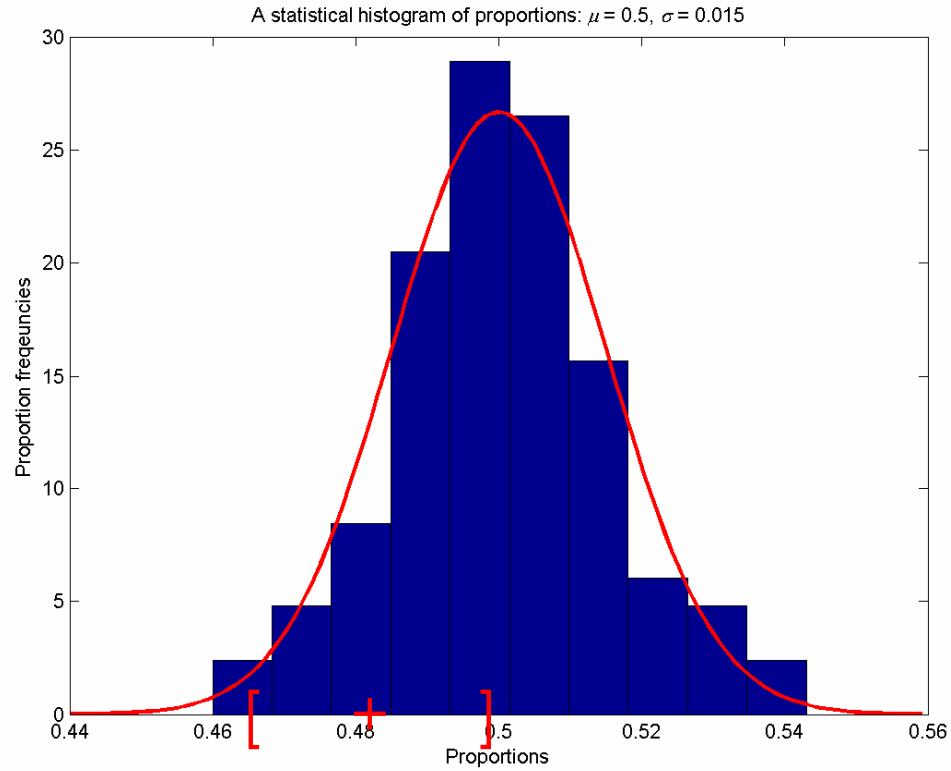


Observed proportion is 0.49

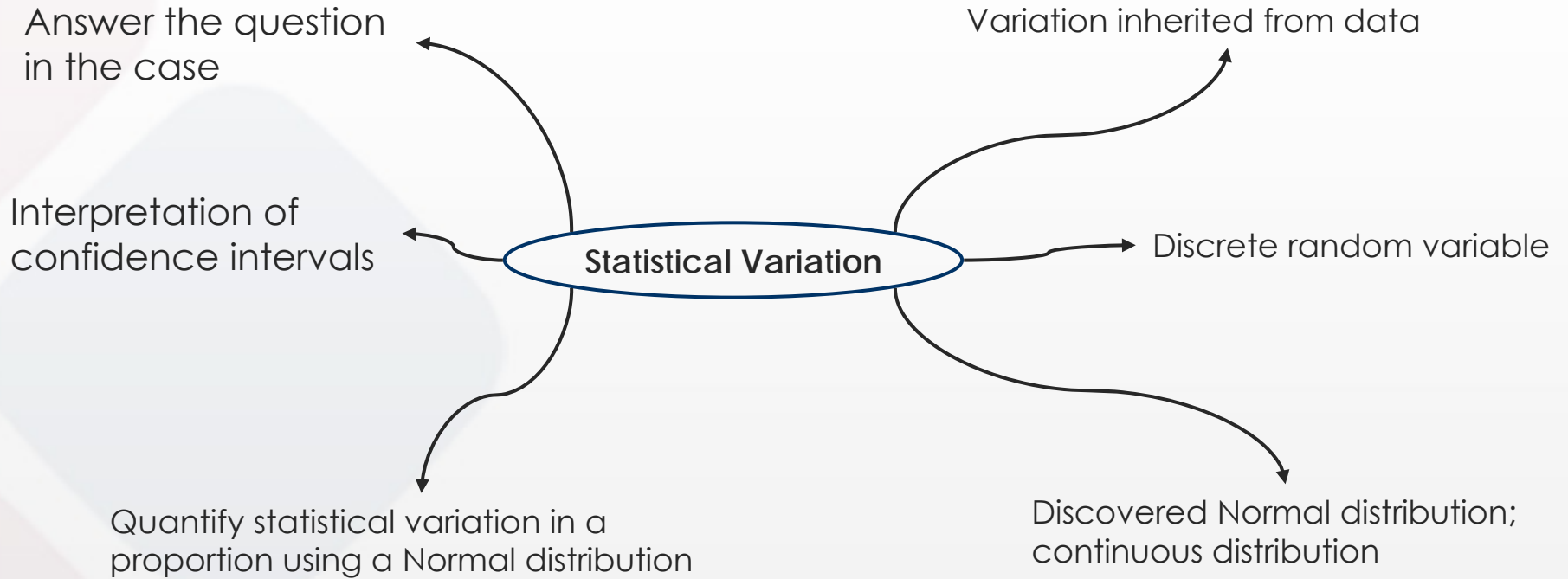
Confidence Interval for a proportion



Confidence Interval for a proportion

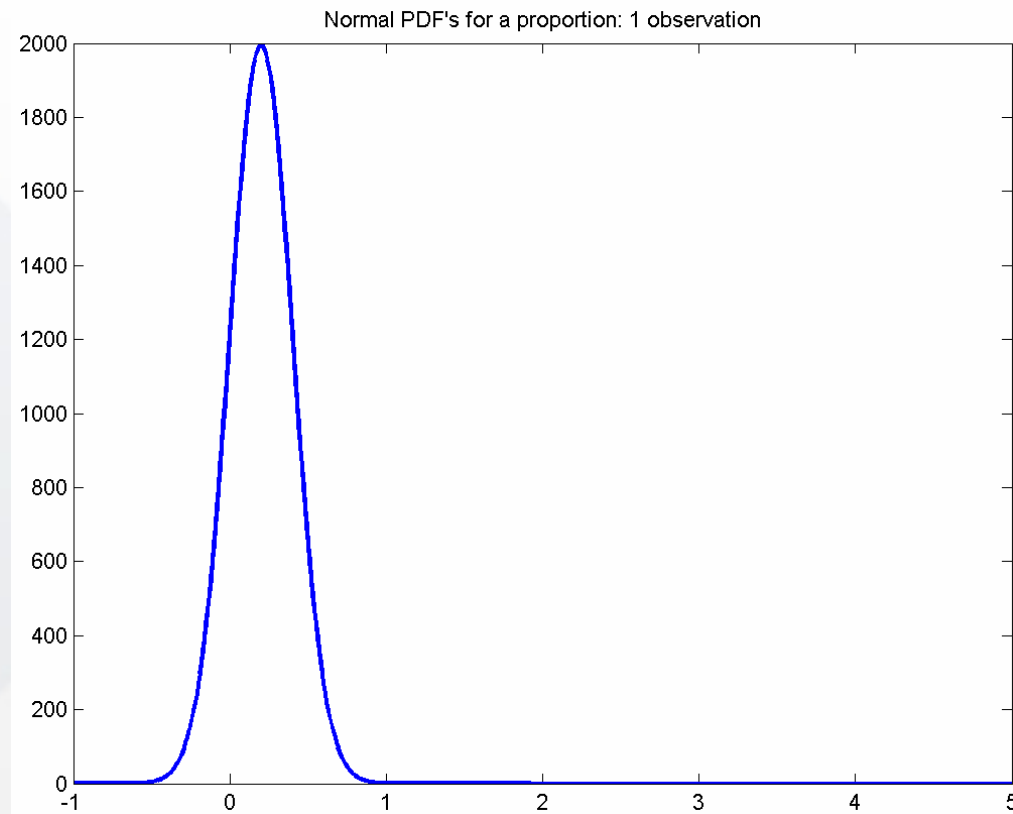


Summary

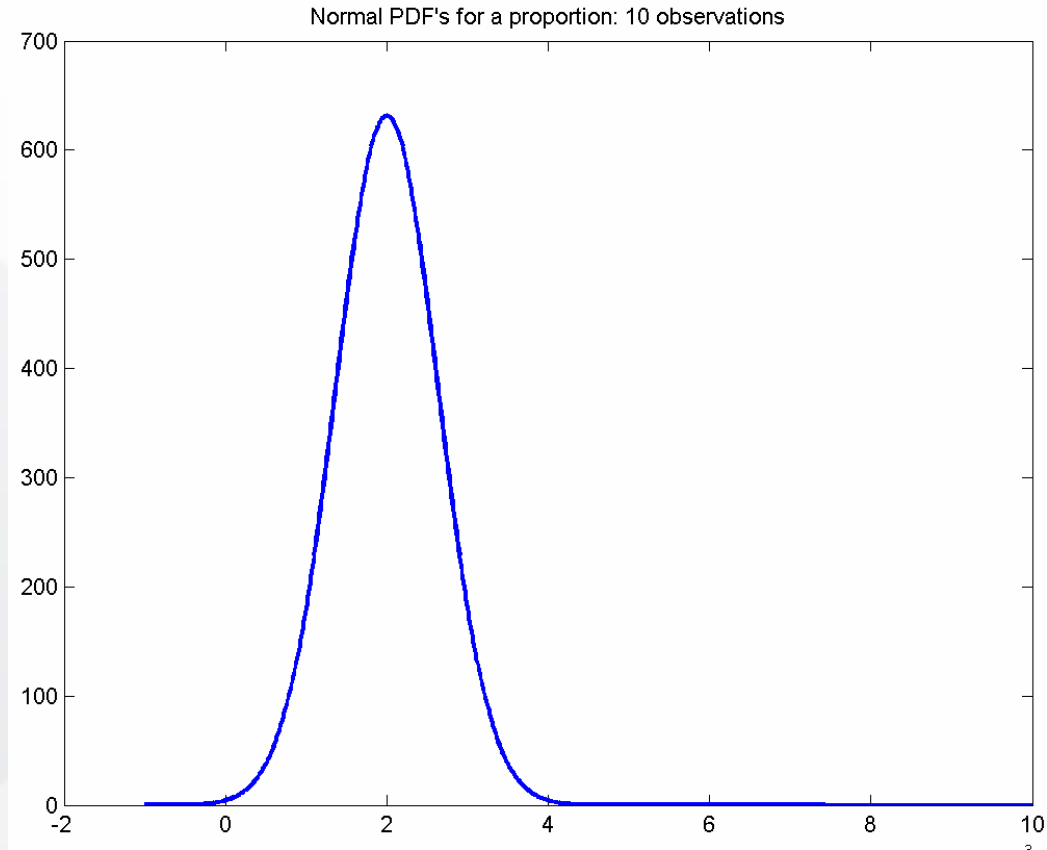


Normal distribution for a proportion

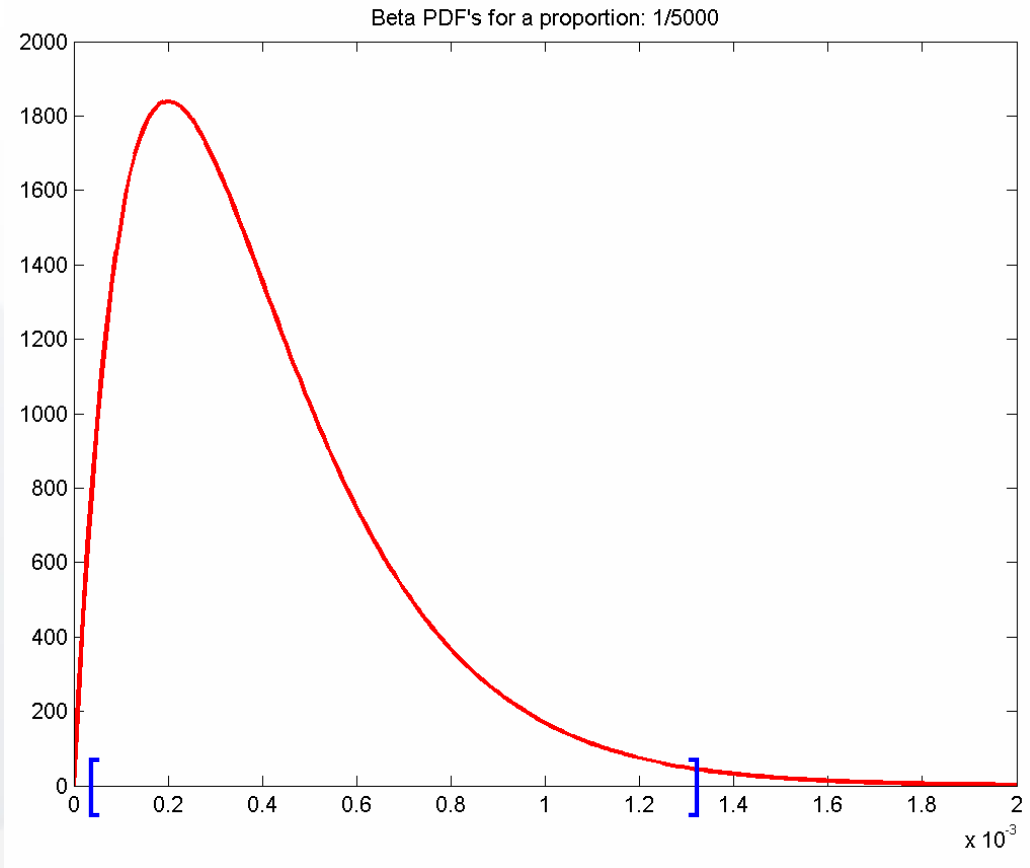
A confidence interval is reliable if the proportion is not close to 0 or 1



Normal distribution for a proportion



Preferably: use a Beta distribution



95% Probability interval