MM104/MM106/BM110
Statistics and Data Presentation

Lecture 5:
Sampling distributions
Binomial Distribution
Distribution of the sample proportion

Chris Robertson

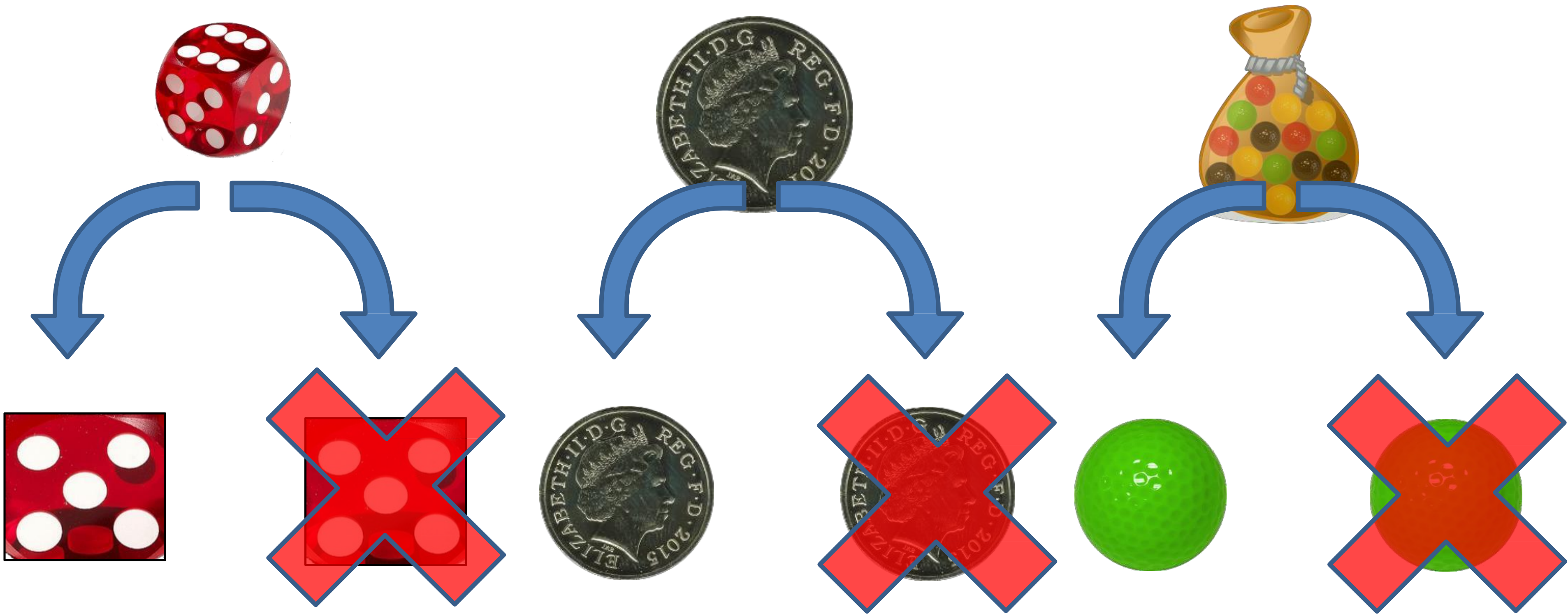
DEPARTMENT OF MATHEMATICS & STATISTICS
University of Strathclyde Science

Binomial distribution

Binomial Random Variables

It describes cases when each attempt (*trial*) can have only *two* outcomes: success or failure.

Discrete Random Variable

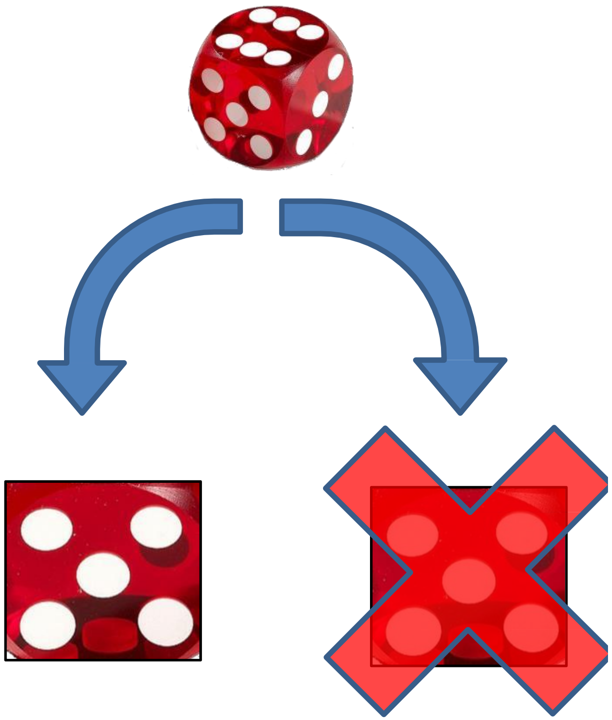


Number of successes → discrete random variable!!

Binomial Random Variables

Rolling a die and counting if you roll a 5 or not

5 – success; 1,2,3,4,6 - fail

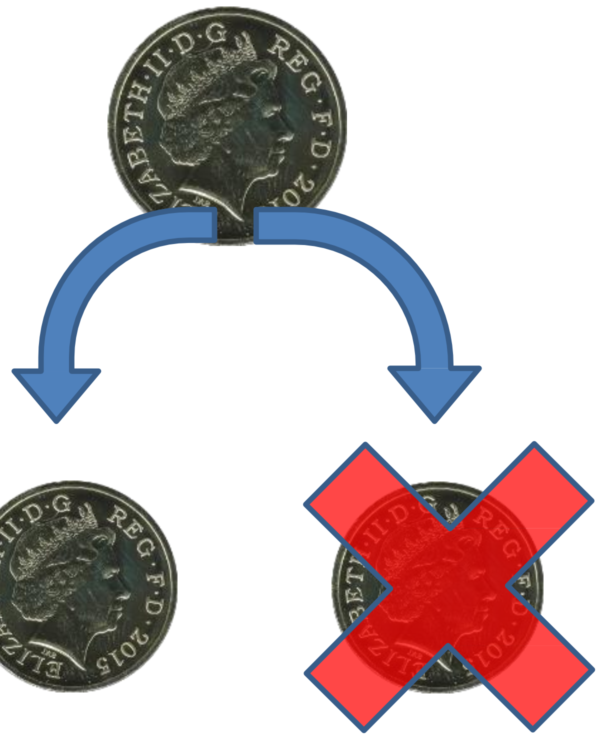


Number of successes ⟹ discrete random variable!!

Binomial Random Variables

Tossing a coin

Head – success; Tail - fail

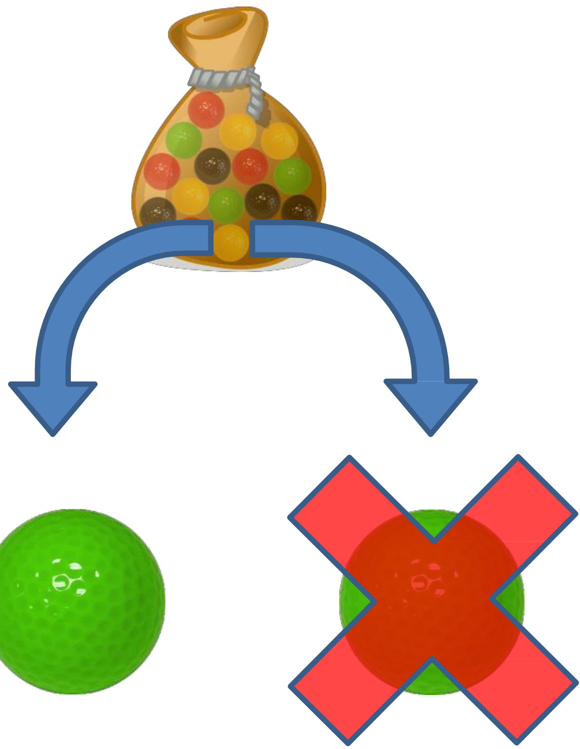


Number of successes → discrete random variable!!

Binomial Random Variables

Drawing a coloured ball from a bag

Green Ball – success; any other colour - fail



Number of successes $\longrightarrow$ discrete random variable!!

Binomial Random Variables

- Distribution associated with X, the number of "successes" in *n* independent trials, each having the same probability *p* of success.
- Many important examples,
  - elections (a trial is a vote, a success is "voting for candidate X").
  - Testing for a disease –
    - success have disease; failure no disease
  - Gambling – success win ; fail lose stake

- Mean of the distribution is *n*p and standard deviation $\sqrt{np(1-p)}$.

You'll learn the mathematical details next year!

Binomial Random Variables

Mean of the distribution is np and standard deviation $\sqrt{np(1-p)}$.

If you toss a coin 10 times and the probability of a tail is 0.5;

$$\mu = 10 \times 0.5 = 5$$
$$\sigma = \sqrt{10 \times 0.5 \times 0.5} = \sqrt{2.5} = 1.5811$$

If you play roulette 100 times and bet £1 on a single number each time the probability of winning is 1/37 (36 numbers plus 0 where the bank wins;

$$\mu = 100 \times \left({}^{1}/_{37}\right) = 2.7$$
$$\sigma = \sqrt{100 \times \left({}^{1}/_{37}\right) \times \left({}^{36}/_{37}\right)} = \sqrt{2.6297} = 1.6216$$

Binomial Random Variables

- If the number of trials (n) is large then the distribution of the number of successes is symmetric and bell-shaped

- Binomial distribution converges to normal distribution for large n.

- Large means $np > 5$; $n(1-p) > 5$

- So we can use the normal distribution to find probabilities for proportions in large samples

The exact distribution is known as the binomial distribution – next year if studying Maths/Statistics

Binomial Random Variables

- Number of times, x, picking red ball out of $n=25$ trials (15 balls, out of which 6 are red).

— Success: picking a red ball.

— Success probability: $p = \frac{6}{15} = 0.4$

— Binomial, mean: $np = 25 * 0.4 = 10$

(Number of times I expect, on average, to get a red ball out of the 25 trials)

— Standard deviation:

$$\sqrt{np(1-p)} = 2.449$$

Example

— In a large company, the number of hours that full-time employees work in any week is known to be Normally distributed with $\mu = 30$h and standard deviation $\sigma = 3.9$h. If there are 276 employees in the company, what is the mean number of **employees** being paid over-time each week (assuming payment of overtime is based on the rule outlined before i.e more that 31.5 hours per week)?

Success: finding one employee working overtime.

Probability of success: $p = P(X > 31.5) = 0.35026$

If the number of employees is n=276, the average number of times the "work overtime" event (i.e. a "success") happens is:

$np = 276 * P(x > 31.5) \qquad = 276 * 0.35026 = 93.672$

Distribution of the sample proportion

Sampling distribution of sample proportion

If x is the number of individuals with a characteristic of interest within a sample of size n from the population, then

sample proportion $\hat{p} = \frac{x}{n}$

is a random variable – i.e. the proportion of the sample with the characteristic.

p is the population proportion – proportion in the population with the characteristic, which is estimated by $\hat{p}$, the sample proportion.

Sampling distribution of sample proportion

Example: Voters in a poll.

- From a population of n voters, number of voters for option 1, x, is like number of successes for option 1.

Random variable!!

- The *proportion* of voters that voted for option 1 (i.e. proportion of successes), $\hat{p} = x/N$, is also a random variable.

Sampling distribution of sample proportion

- If you have binomial data with mean np and standard deviation $\sqrt{np(1-p)}$, where n is the sample size and p is the success probability in the population.
  - Then $\hat{p}$, the sample proportion, is a random variable, which is the proportion of successes.
  - If np>5 **and** $n(1-p)$>5 the sampling distribution for $\hat{p}$ can be approximated by a normal distribution with mean p and standard error $\sqrt{p(1-p)/n}$.

i.e. the proportion of successes can be treated as a normal random variable!

Example

- Number of times, x, picking red ball out of $n=25$ trials (15 balls, out of which 6 are red, i.e. $\hat{p}=6/15=0.4$).

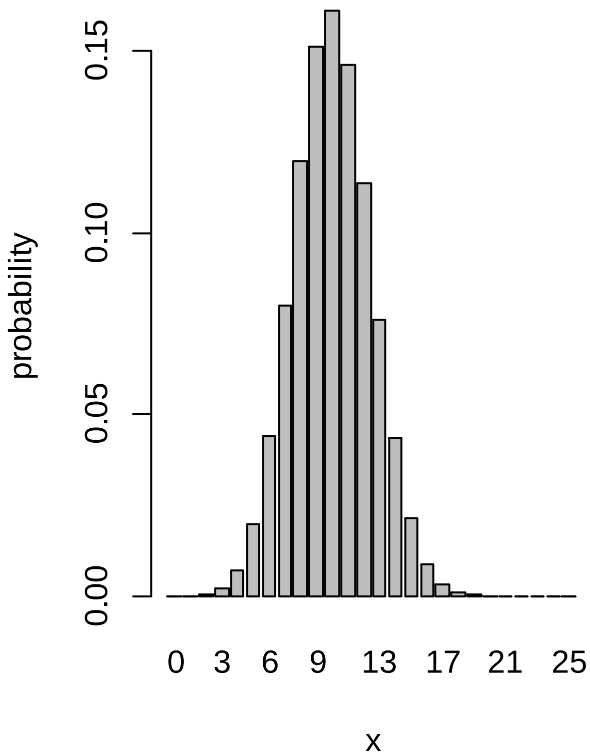
$$n\hat{p} = 10, n(1 - \hat{p}) = 15$$

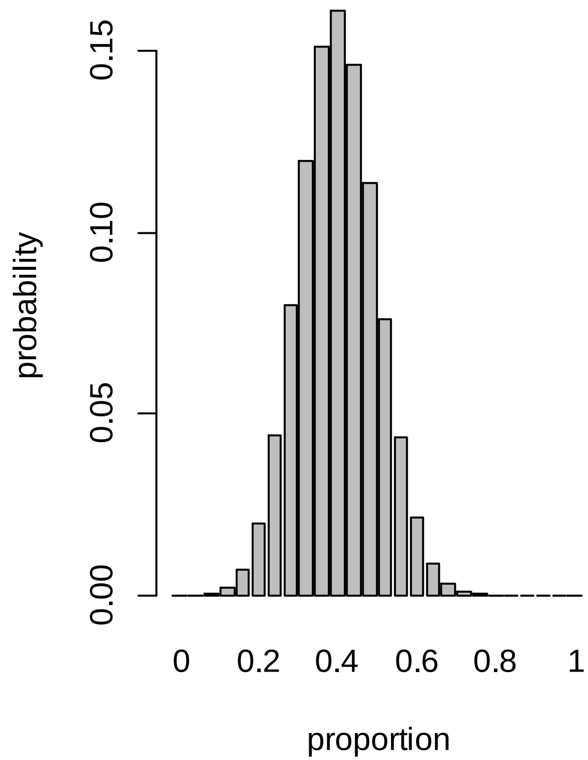
$\hat{p}$
APPROX.
NORMAL!!

— Sample proportion, $\hat{p}$, approx. normal with mean $\hat{p} = 0.4$ and standard error $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.4(1-0.4)/25} = 0.098$

Example Distributions

- Essentially the same distribution, but with horizontal axis rescaled:





Example – probability calculations

— In the Brexit vote, 48.1 % of the electorate voted for “Remain”. On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for “Remain”?

(b) Fewer than 50% of a sample of 300 voters will vote for “Remain”?

Example – probability calculations

— In the Brexit vote, 48.1 % of the electorate voted for “Remain”. On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for “Remain”?

(b) Fewer than 50% of a sample of 300 voters will vote for “Remain”?

This bit is just asking you for the population proportion from the question. It is the population as the vote was the whole population of voters.

This bit of the question is asking you to calculate the probability about a proportion in the sample – even though it is expressed in terms of percentage

Percentage 0-100

Proportion 0-1

Percentage = 100*Proportion

Example – probability calculations

— In the Brexit vote, 48.1 % of the electorate voted for "Remain". On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for "Remain"?

Success prob. $p = 0.481$

$\boldsymbol{n}$

(b) Fewer than 50% of a sample of 300 voters will vote for "Remain"?

$np > 5$ and $n(1 - \mathrm{p}) > 5$ (so can use the normal distribution to get the probability)

i.e. $P(\hat{p} < 0.50)$

$\hat{p}$ variable: fraction votes for "Remain"

$$\widehat{\boldsymbol{p}} = \mathbf{0.481},$$

Mean (μ)

$$\boldsymbol{\sigma} = \sqrt{\frac{\widehat{\boldsymbol{p}}(\mathbf{1} - \widehat{\boldsymbol{p}})}{\boldsymbol{n}}} = \mathbf{0.0288}$$

Standard deviation

Example – probability calculations

— In the Brexit vote, 48.1 % of the electorate voted for "Remain". On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for "Remain"?

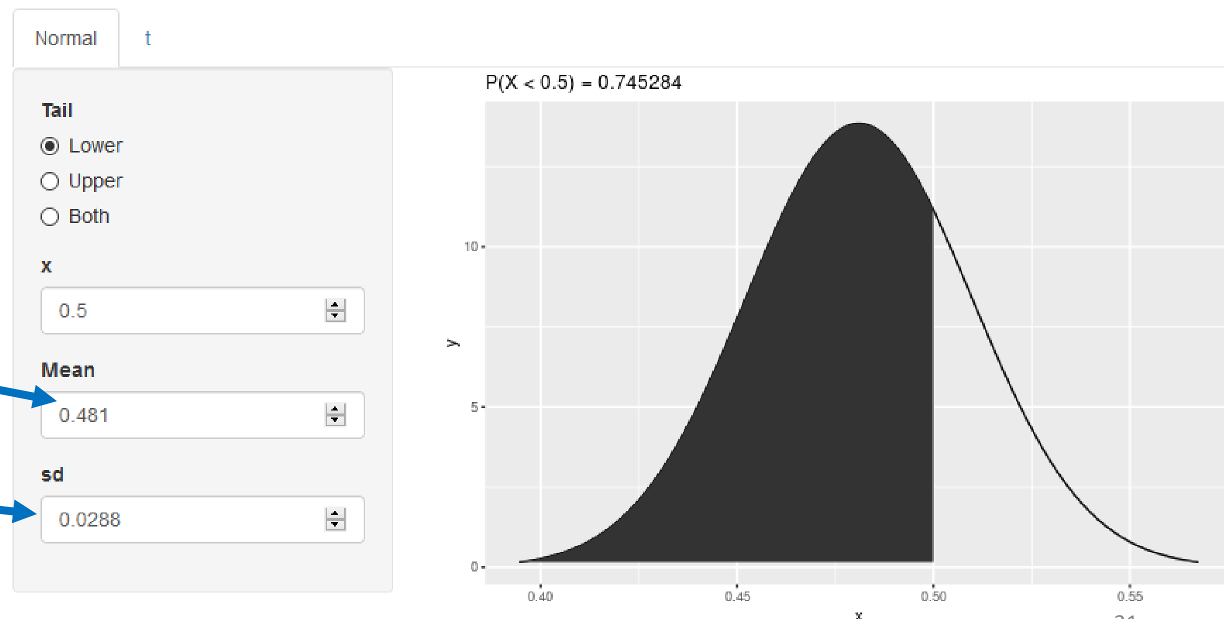
n

(b) Fewer than 50% of a sample of 300 voters will vote for "Remain"?

i.e. $P(\hat{p} < 0.50)$

$$\boldsymbol{\hat{p} = 0.481,}$$

$$\boldsymbol{\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0288}$$



Example – probability calculations

— In the Brexit vote, 48.1 % of the electorate voted for “Remain”. On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for “Remain”?

n

(b) Fewer than 50% of a sample of 300 voters will vote for “Remain”?

Answer

i.e. $P(\hat{p} < 0.50)$

$$\hat{p} = 0.481,$$

$$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0288$$

Example – Change the sample size

— In the Brexit vote, 48.1 % of the electorate voted for “Remain”. On the assumption that allegiances have remained unchanged, what is the probability that in a repeated vote

(a) a randomly chosen elector will vote for “Remain”?

n

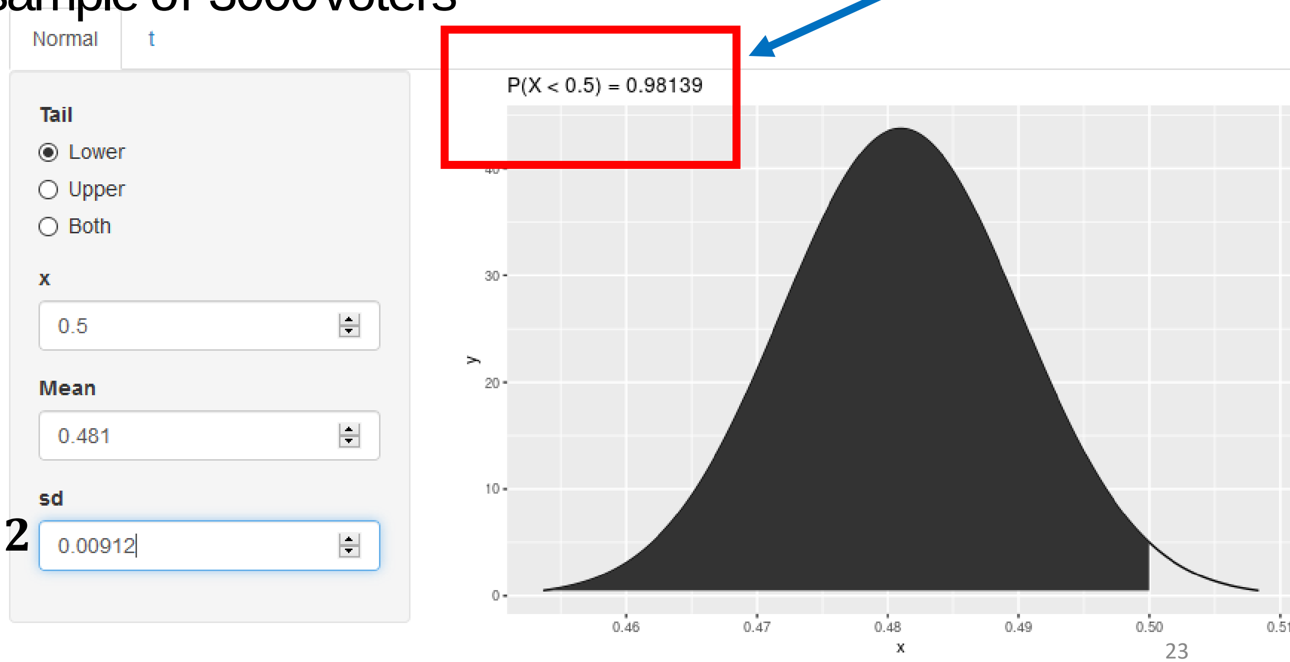
(b) Fewer than 50% of a sample of 3000 voters will vote for “Remain”?

Answer

$$\text{i.e. } P(\hat{p} < 0.50)$$

$$\hat{p} = 0.481,$$

$$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.00912$$

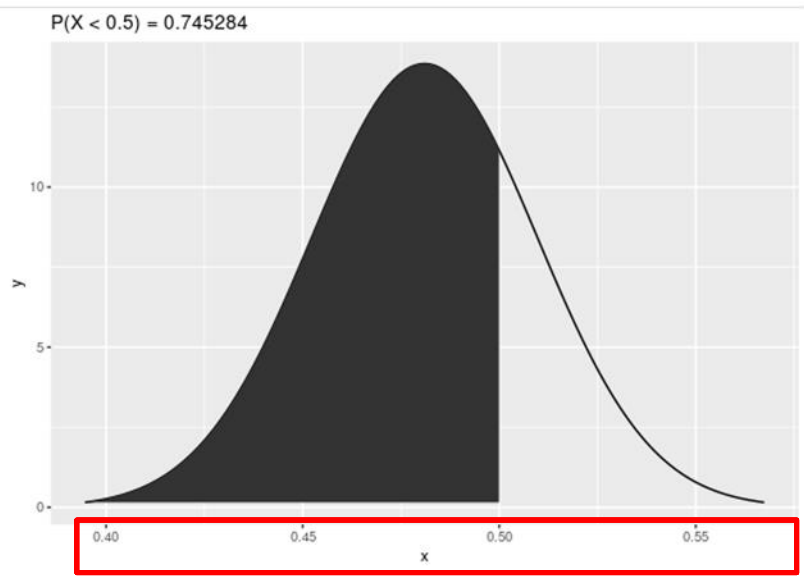


Example – Increasing Sample Size

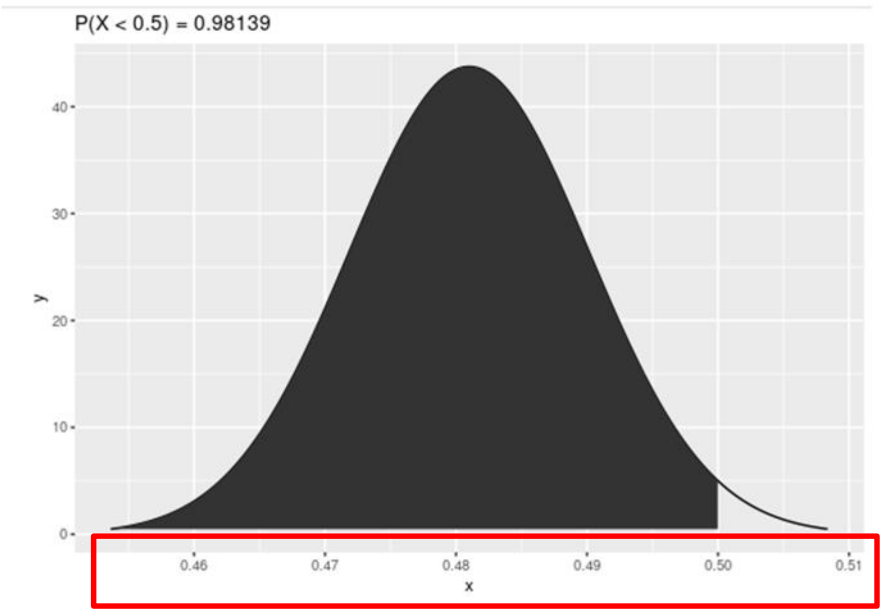
STANDARD ERROR DECREASES WITH INCREASING SAMPLE SIZE

$$\hat{p} = 0.481,\quad \sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0288$$

$$\hat{p} = 0.481,\quad \sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.00912$$



$n = 300$



$n = 3000$

Note the x axis scales on the graphs – there is much less of a range on the right.

Probability calculation

- **Example:**
  - A factory production line has been producing components with 14% of them failing the acceptance test, and getting scrapped. A random sample of 700 components from the line is taken for test. If the defectives rate is *still* 14%, what is the probability that 16.4% or more of components in the sample are defective?

"Success" prob. $\hat{p} = 0.14$

Sample size $n = 700$

$n\hat{p} > 5$ and

$n(1-\hat{p}) > 5$

(Approx. Normal)

(new variable=fraction defectives in the sample)

$$\hat{p} = 0.14,$$

$$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.01311$$

Calculate

$P\,(X > 0.164)$?

Probability calculation

- # Example:

  — A factory production line has been producing components with 14% of them failing the acceptance test, and getting scrapped. A random sample of 700 components from the line is taken for test. If the defectives rate is *still* 14%, what is the probability that 16.4% or more of components in the sample are defective?

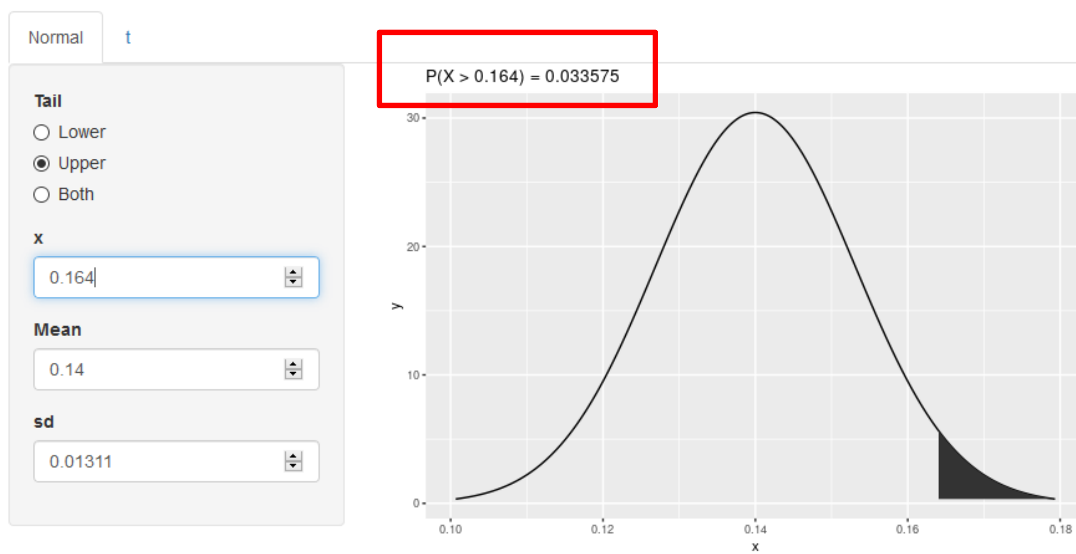
$$P\ (X > 0.164)\ ?$$

"Success" prob. $\hat{p} = 0.14$

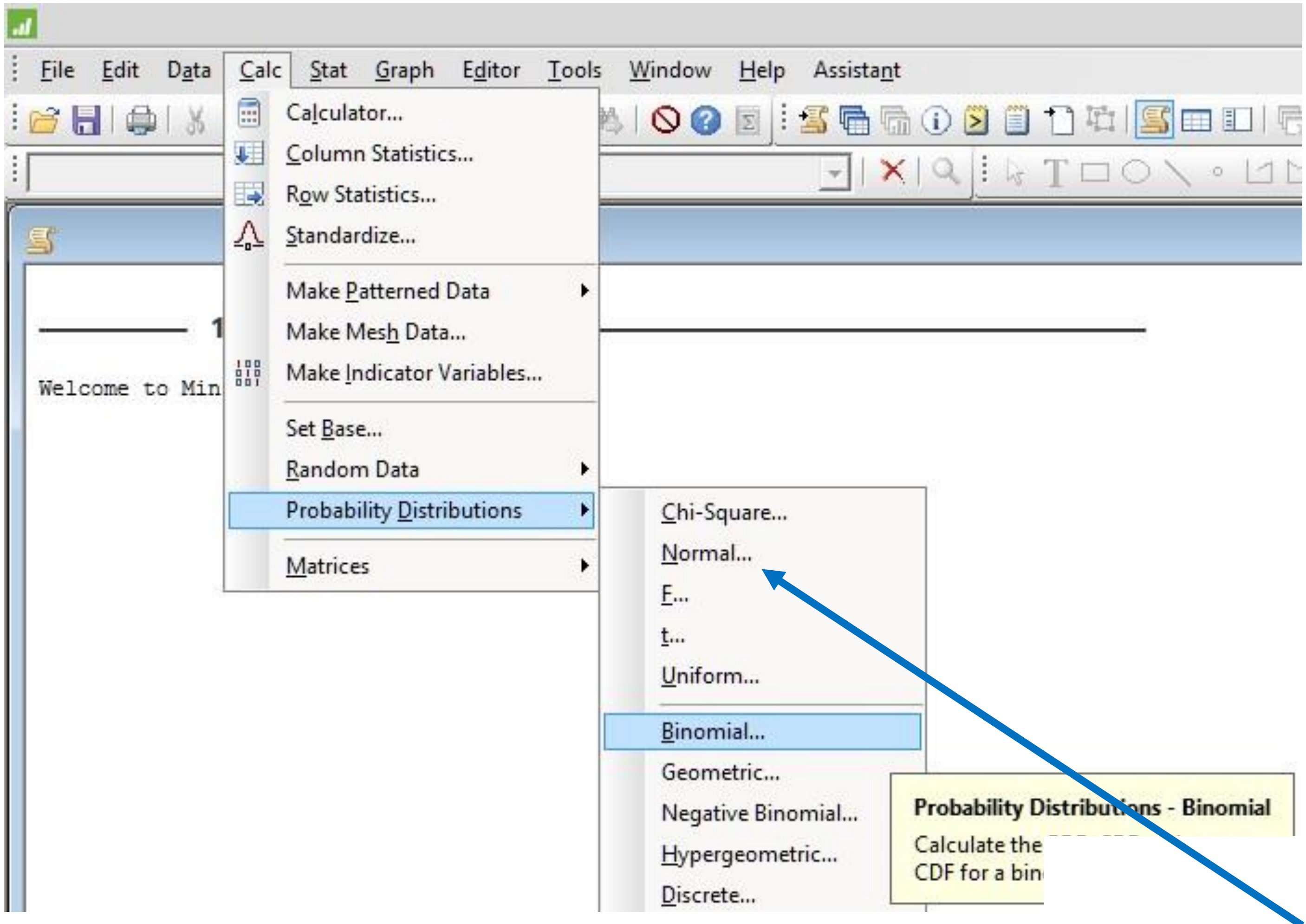
Sample size $n = 700$

$$\hat{p} = 0.14,$$

$$\sigma = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.01311$$



Probability calculations in Minitab



You can use Minitab to make probability distribution calculations if you wish – you do not have to use the tables in myplace

There are also many online calculators.

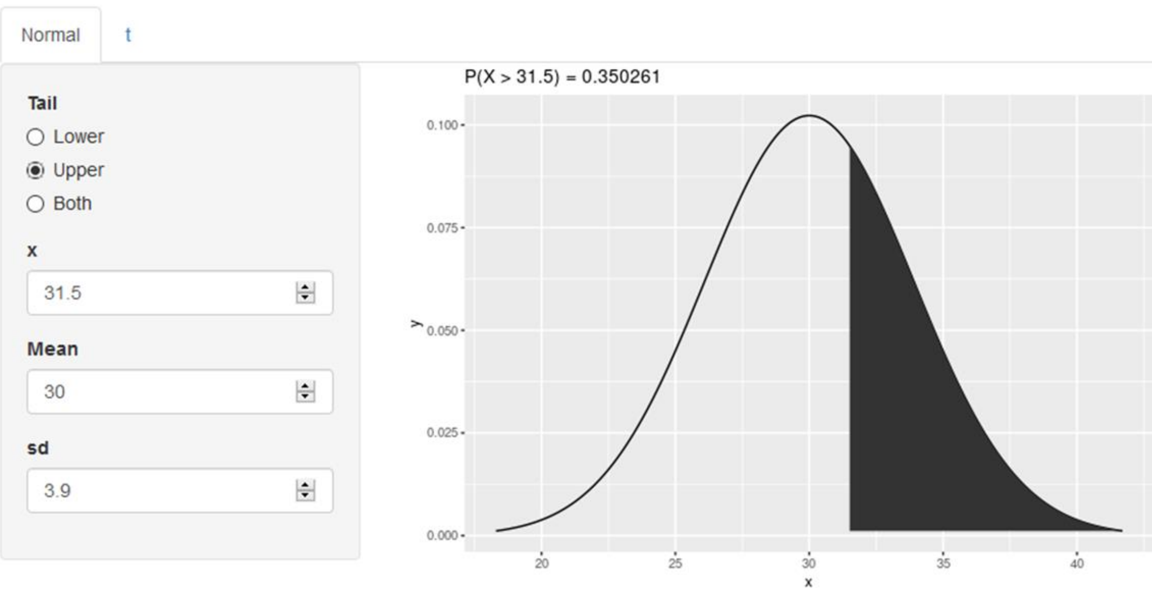
Pick Normal

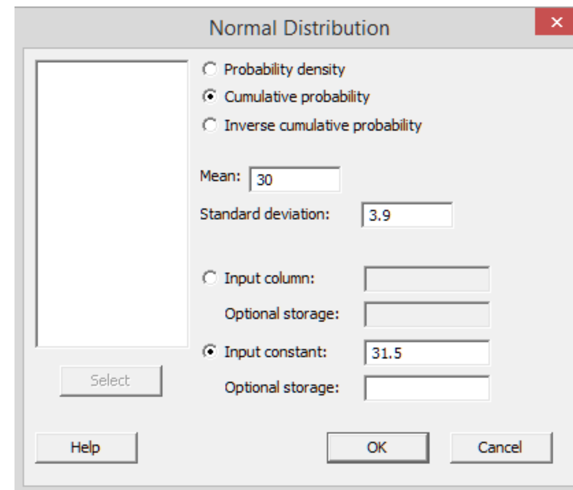
Probability calculations in Minitab

- Examples:

Normal, $\mu = 30$ and $\sigma = 3.9$

$P\ (X\ < 31.5)\ ?$





Cumulative Distribution Function

Normal with mean = 30 and standard deviation = 3.9

x	P(X ≤ x)
31.5	0.649739

$P(X\ < 31.5)\ = 1\text{-}\ P(X\ > 31.5)$
$= 1\text{-}0.350261 = 0.649739$

Key Points

- The sampling distribution of the sample proportion follows a normal distribution in large samples
- The mean of the sampling distribution of the sample proportion is the population proportion

$$\mu_{\hat{p}} = p$$

- The standard deviation of the sampling distribution is known as the standard error of the sample mean

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Large means $np > 5;\ n(1-p) > 5$