Announcements

- The third hour-exam will be held on Friday, July 26 and covers Chapters 4-6.
- The use of cell phones and other wireless devices is not permitted on the exam. You will need to bring a separate calculator for the exam.
- Sharing of calculators is not permitted during the exam.

Example

Consider the following 19 scores from a 100 point exam:

43, 57, 59, 65, 67, 67, 68, 70, 72, 73,

74, 76, 79, 80, 83, 85, 87, 90, 98

Find the upper quartile for these scores.

Median = 73

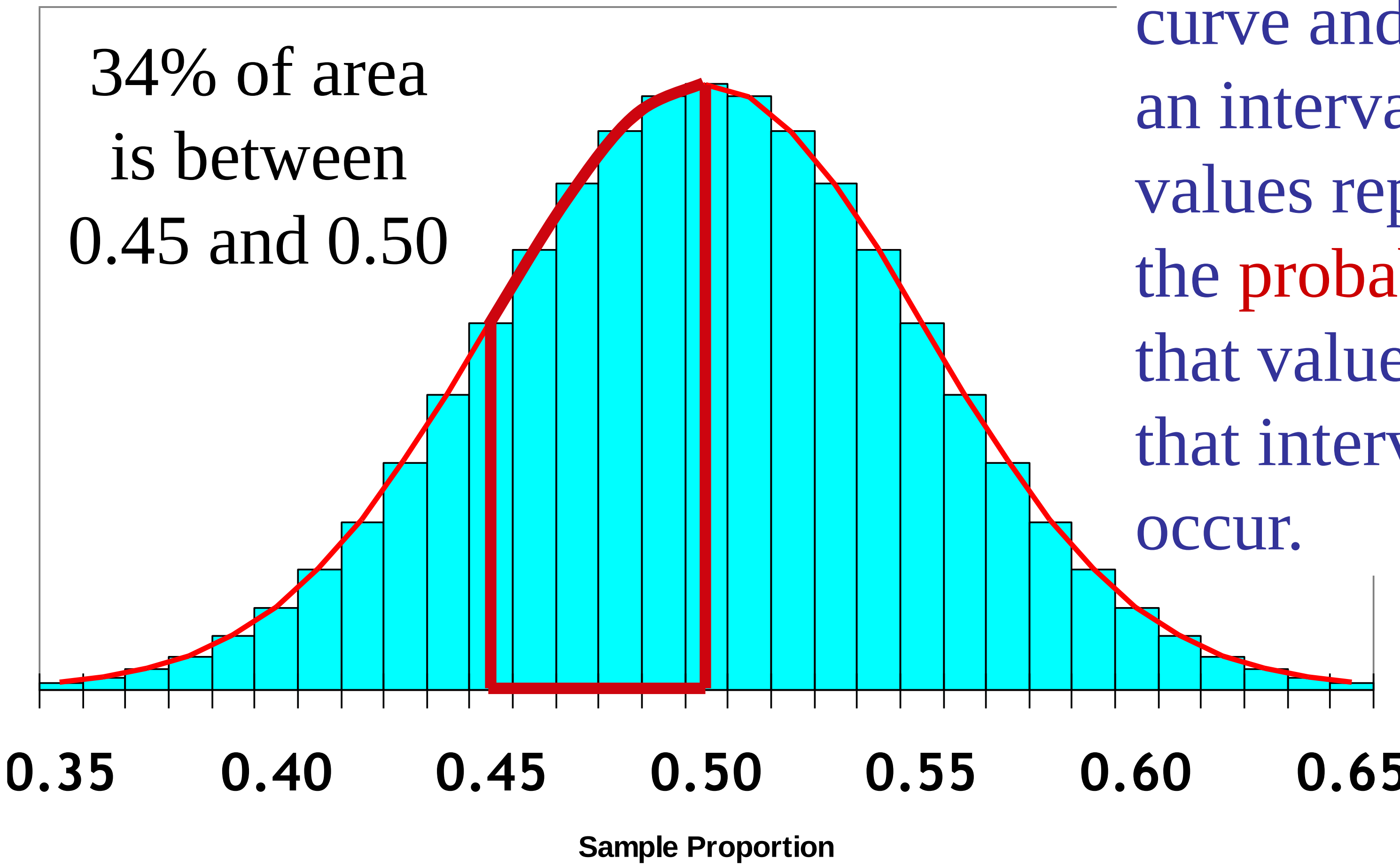
$Q_3 = 83$

The Normal Distribution (Section 6-C)

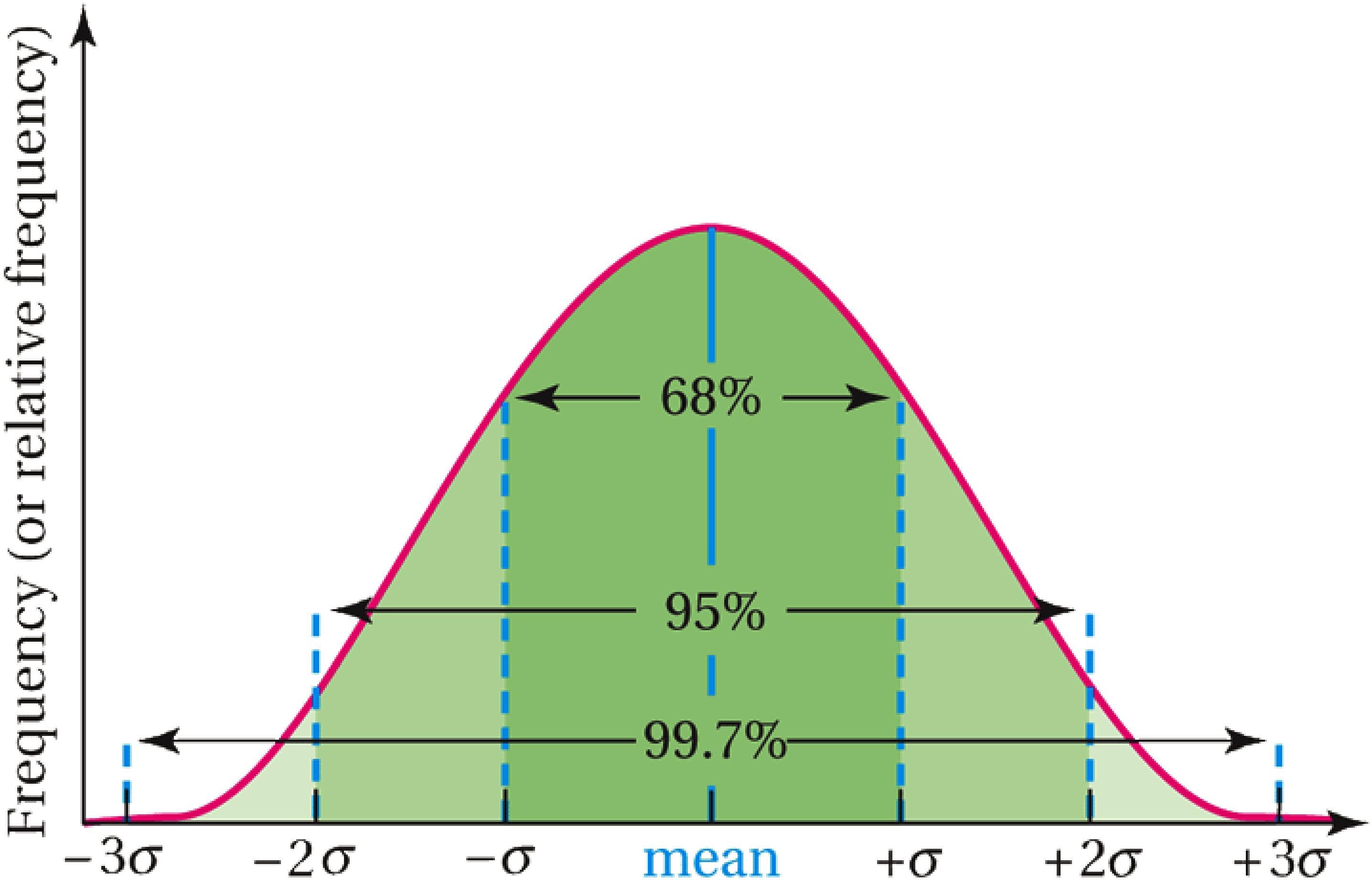
The **normal distribution** is a symmetric, bell-shaped distribution with a single peak. Its peak corresponds to the mean, median, and mode of the distribution.

Approximately normal distribution
(proportion of heads in 100 coin tosses)

The area under the curve and above an interval of values represents the probability that values within that interval will occur.



The 68-95-99.7 Rule for a Normal Distribution



Standard Scores

The number of standard deviations that a data value lies above or below the mean is called its **standard score** (or z-score), defined by

$$z = \text{standard score} = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$$

Data Value		**Standard Score**
above the mean	→	positive
below the mean	→	negative

Standard Scores

Example: If the mean were 21 with a standard deviation of 4.7 for scores on a nationwide test, find the *z*-score for a 30. What does this mean?

$$z = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$$

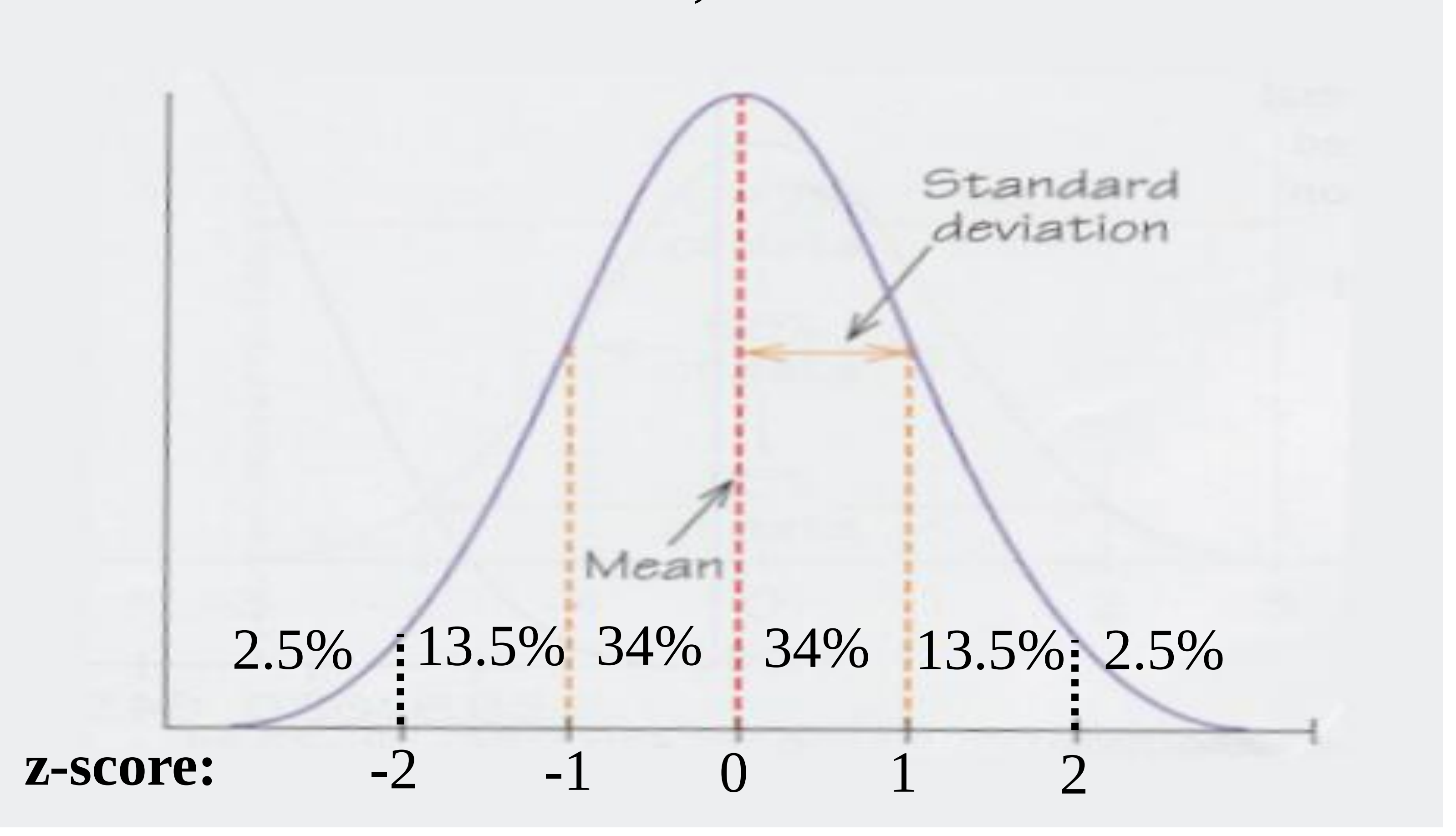
$$= \frac{30 - 21}{4.7} \approx 1.91$$

This means that a test score of 30 would be about 1.91 standard deviations above the mean of 21.

Areas Under Normal Curve

$$68 \div 2 = 34, \quad 95 \div 2 = 47.5$$

$$47.5 - 34 = 13.5, \quad 50 - 47.5 = 2.5$$

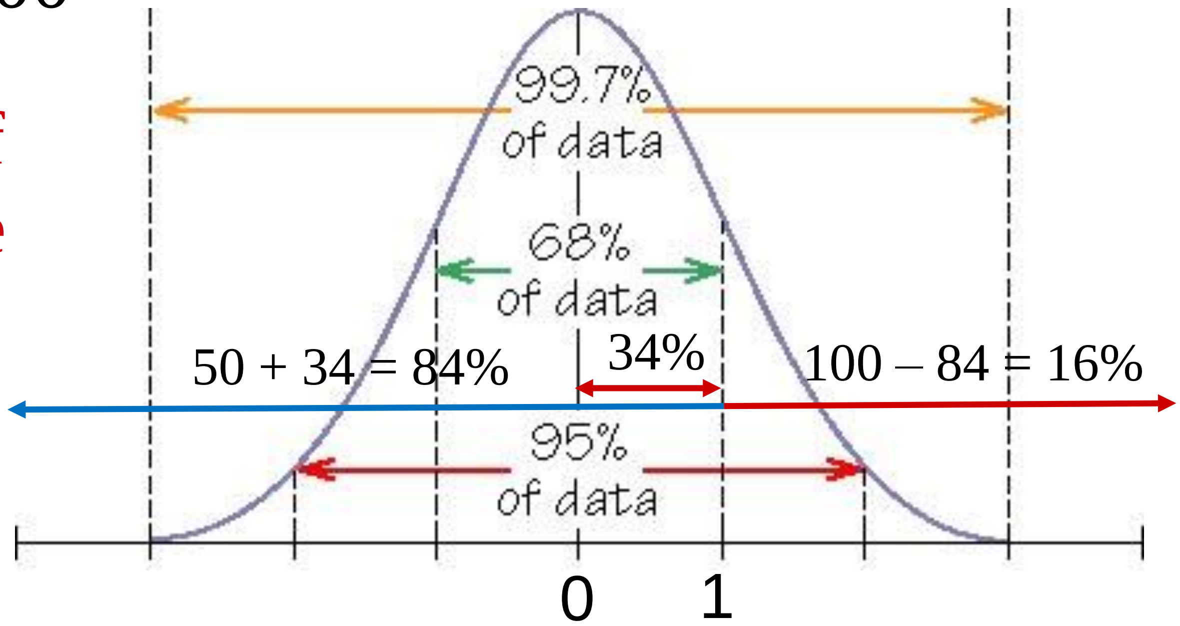


Example

The annual income of residents in a certain county is normally distributed, with a mean of \$42,000 and a standard deviation of \$10,000. What percentage of residents have income over \$52,000?

$$z = \frac{52{,}000 - 42{,}000}{10{,}000} = 1 \quad \text{(84th percentile)}$$

About 16% of residents have income over \$52,000.



Unit 6D

Statistical Inference

Statistical Significance

A set of measurements or observations in a statistical study is said to be **statistically significant** if it is unlikely to have occurred by chance.

Common levels of significance:

- *At the 0.05 level* – The probability of an observed difference occurring by chance is 1 in 20 or less.
- *At the 0.01 level* – The probability of an observed difference occurring by chance is 1 in 100 or less.

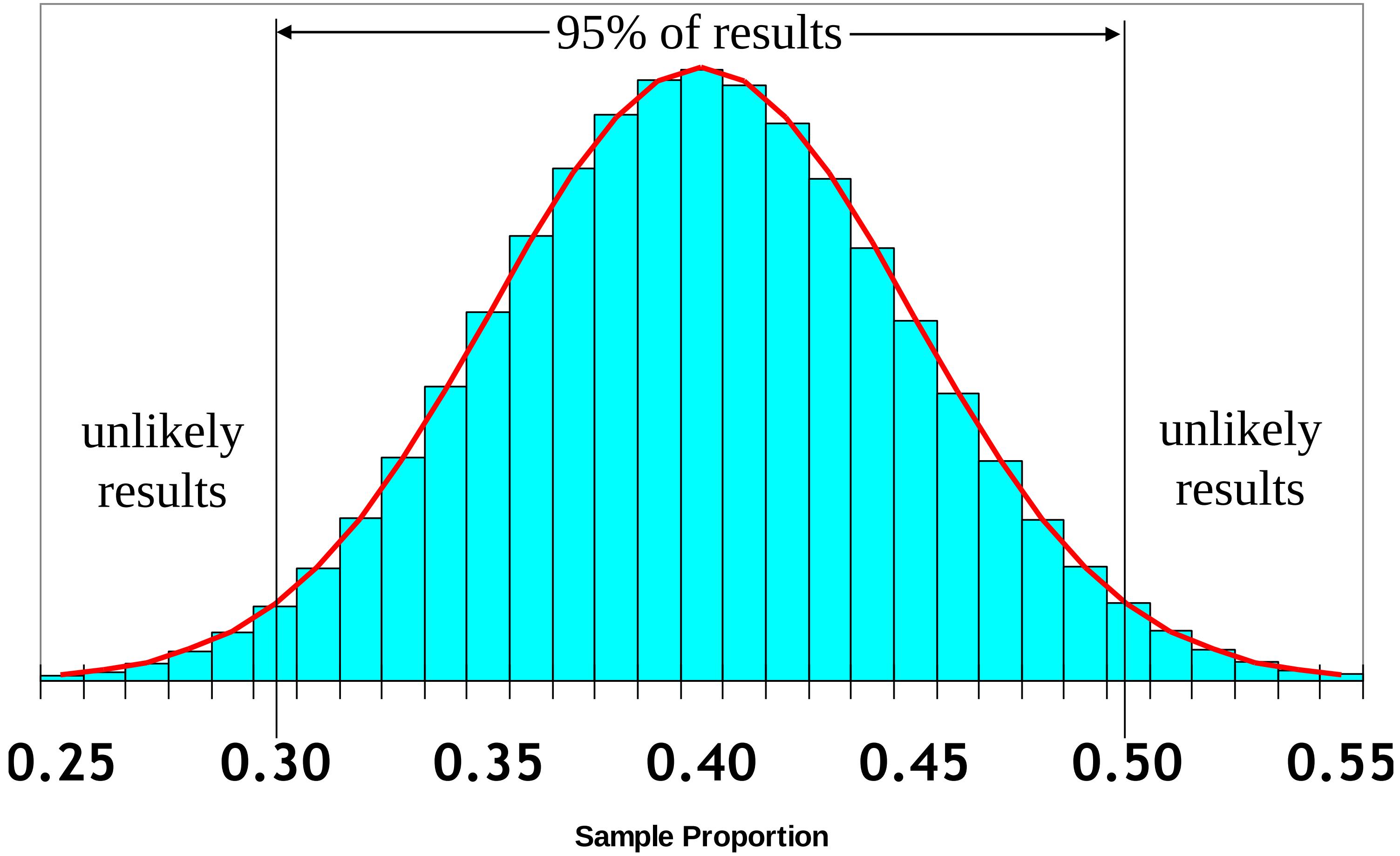
Example: proportion of dog owners

In a random sample of 100 U.S. adults, 34% were dog owners.

- If 34% of the entire population were actually dog owners, finding that 34% of our random sample were dog owners would be a reasonable (not unlikely) result.
- What if 40% of the population were actually dog owners ($p = 0.40$), would a sample with 34% dog owners be likely?
- What if 50% of the population were dog owners?

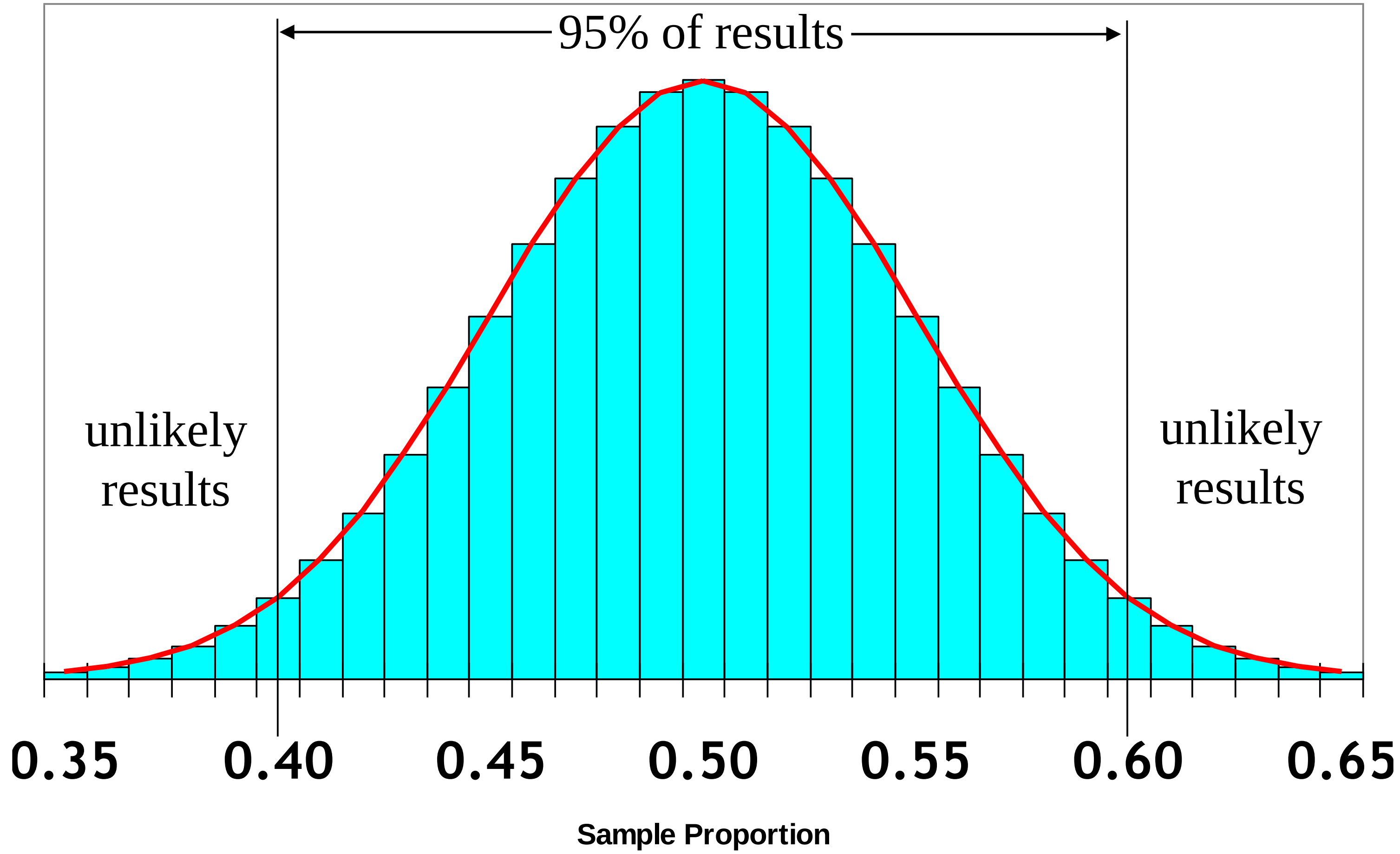
To answer these questions consider the ***distribution of the sample proportion***.

Distribution when $p = 0.40$ and $n = 100$



It would NOT be unlikely to observe 34% dog owners.

Distribution when p = 0.50 and n = 100



It would be unlikely to observe 34% dog owners.

Sampling Distribution of a Sample Proportion

Choose an SRS of size *n* from a population that contains population proportion *p* of successes. The **sample proportion** is

$$\hat{p} = \frac{\text{count of successes in the sample}}{n}$$

Success refers to whatever response we are counting.

- **Shape**: for large n, the sampling distribution is **approximately normal**.
- **Center**: the **mean** of the sampling distribution is ***p***.
- **Spread**: the **standard deviation** of the sampling distribution is

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

Formula for 95% Confidence Interval

Choose a simple random sample of size n from a large population that contains an unknown proportion p of successes. Calculate $\hat{p}$, the proportion of successes in the sample.

The **margin of error** = $2\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

A **95% confidence interval for *p*** is

$$\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ to } \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This interval contains *p* about 95% of the time.

Quick Estimate of Margin of Error

- Easier, but cruder estimate of *p*.
- Produces larger margin of error.
- Is most inaccurate when *p* is close to 0 or 1.

$\sqrt{p(1-p)}$ is largest when $p = 0.5$

Use 0.5 to estimate *p* when calculating margin of error.

$$\text{Margin of error} = 2\sqrt{\frac{p(1-p)}{n}} \approx 2\sqrt{\frac{0.5 \times 0.5}{n}} = \frac{1}{\sqrt{n}}.$$

Margin of Error and Confidence Intervals

Example: A survey of 1200 people finds that 47% plan to vote for Smith for governor.

Find the *margin of error*.

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1200}} \approx 0.029$$

Find the *95% confidence interval* for the survey.

47% – 2.9% = 44.1%

47% + 2.9% = 49.9%

We can be 95% confident that the true proportion of people who plan to vote for Smith is between 44.1% and 49.9%.

Examples

- If n = 1100 and p = .52, margin of error = $2 \times \sqrt{(0.52 \times 0.48 / 1100)} = 0.030$. Quick method gives $1/\sqrt{1100} = 0.030$.
- If n = 400 and p = .90, margin of error = $2 \times \sqrt{(0.90 \times 0.10 / 400)} = 0.030$. Quick method gives $1/\sqrt{400} = 0.050$.

Quick method is not very accurate when p is close to 0 or 1.