Sampling from a population of “0”s and “1”s

Consider a survey in which you want to ask one “yes” or “no” question. Examples: “Do you own your own home?”, “Are you going to vote for Al Gore for president?”, etc. You are interested in the proportion of people responding “yes”.

- Envision the population as a box of tickets, each marked with either “0” or “1”
- A member of the population who can answer “yes” to the question of interest is like a “1” ticket
- Taking a sample of size n from this population is like taking n tickets from the box
- To find the sample proportion, $\hat{p}$, we divide the number who answered “yes” by the sample size, n
- Finding the number who answered “yes” is the same as adding up the values on the n tickets

Connecting the binomial to the survey problem

- If we take a sample of size n from the box of tickets, the sum of the sequence of “0”s and “1”s gives the number of successes
- The n sampled tickets can be viewed as n trials (practically independent if n is much smaller than the total population N)
- The percentage of “1” tickets in the box is the probability of success for the trials
- Binomial probabilty function can be used to calculate the probability of observing a given number of “yes” answers in a sample from the population
- In real life, though, the sample size is often large enough that you wouldn’t want to calculate $\binom{n}{x}p^x(1-p)^{n-x}$ on your calculator

Can we use the CLT?

- The sample proportion is actually the sample mean (remember: "yes" answers are "1"s, "no" answers are "0"s)
- If the number of success in the sample is denoted as Y, $\hat{p} = \frac{Y}{n}$
- This means that the CLT could come into play if the sample size is large enough
- General rule for determining how large n should be depends on p; both endpoints of $p \pm 2\sqrt{\frac{pq}{n}}$ should be between 0 and 1
- In order to use the CLT, we need to know what the expected value and variance for $\hat{p}$, the sample proportion, are

Mean and variance of sample proportion

What are the mean and variance for $\hat{p} = \frac{Y}{n}$?

Visual: sampling distribution for sample proportion

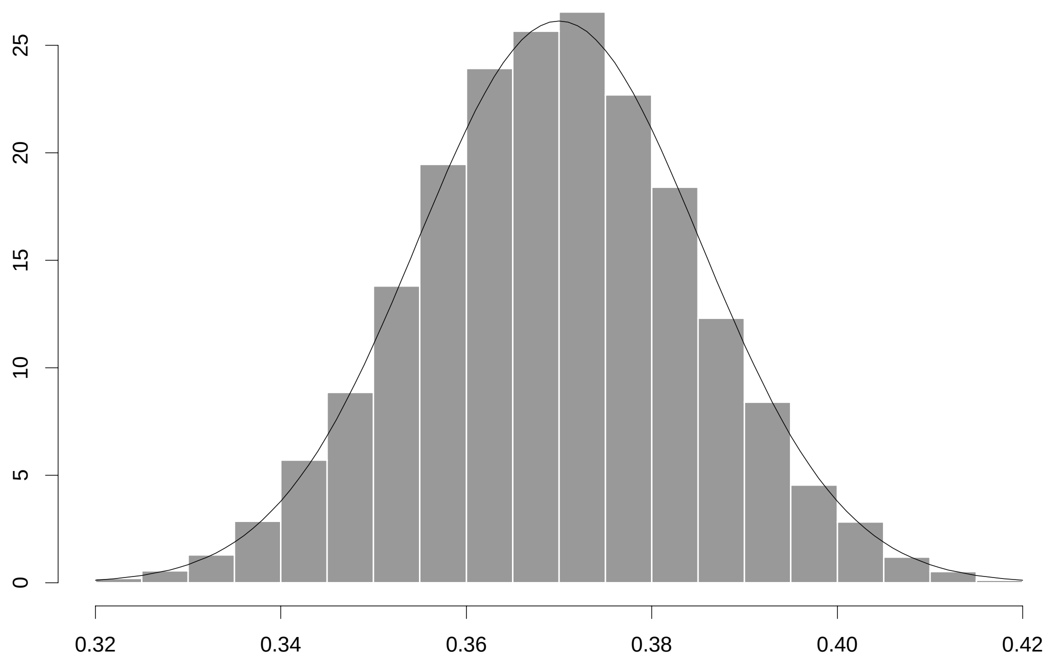


Figure 1: Histogram of 10000 sample proportions

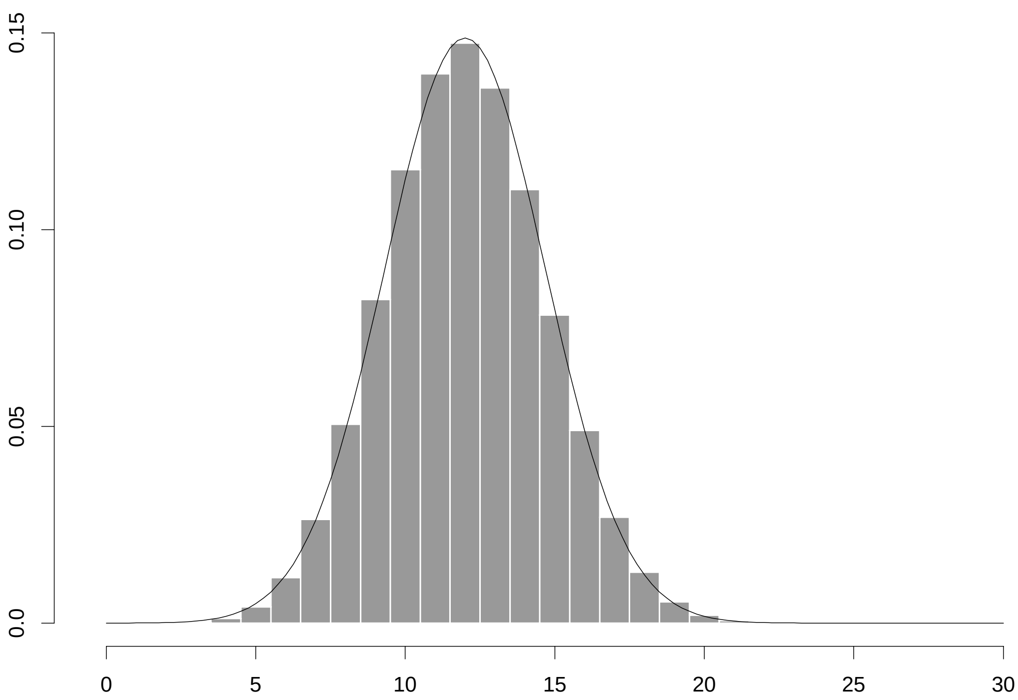
Basics of using CLT for proportions

Assume women constitute 37% of all union members, so that $p = 0.37$. A simple random sample of 1000 union members is selected.

What is the probability that the sample proportion of women will be within ± 0.03 of the population proportion?

Normal approx. to binomial

What's the probability that a sample of 30 voters contains 15 or more who voted Democratic, given that 40% of the population voted Democratic? With patience, could solve this by calculator. Can also approximate using the normal distribution.



Why is the number of successes Y normally distributed? Y is a linear function of a normally distributed random variable $\hat{p} = \frac{Y}{n}$, so Y must be normally distributed.

The continuity correction

In the last diagram, you can see that the normal curve leaves out some area of the bars that should be included. The approximation can be improved by proper choice of boundary points; the key to this *continuity correction* is to reword the question in your mind.

- In the discrete case, phrases like "less than 21" are equivalent to their counterparts of the form "20 or less"
- Rephrase the question. Another example is "15 or more" re-worded as "more than 14".
- Then look for $P(Y > 14.5)$ in the latter case, $P(Y < 20.5)$ in the former
- Phrases like "exactly 25" can be rewritten as $P(24.5 < Y < 25.5)$ to make the transition to the continuous case

Ex: using the continuity correction

Of your first 30 great-great-grandchildren, what is the chance there will be more than 22 boys? Find the exact answer, then make an approximation using the normal. (Assume male and female children are equally likely, and the genders are independent.)

Ex: using the continuity correction

What's the probability that a sample of 30 voters contains 15 or more who voted Democratic, given that 40% of the population voted Democratic?

Summary of sampling distributions so far (part 1)

Let's summarize the various sampling distributions that we talked about in chapter 7.

For a normal population with mean μ:

- If the standard deviation σ is known, $\bar{Y}$ is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- If the standard deviation σ is unknown, $\frac{\bar{Y}-\mu}{\frac{s}{\sqrt{n}}}$ has a t distribution with $n-1$ degrees of freedom.

Summary of sampling distributions so far (part 2)

If the population is non-normal with mean μ and known standard deviation σ, using CLT:

- If $n \geq 30$, $\bar{Y}$ is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- When you have no information about the shape of the population distribution, you shouldn't assume you can use the CLT unless $n \geq 30$ (even though I used it for an example with $n = 25$ in class on Tuesday just so I could use small enough numbers to do the math in my head).

Summary of sampling distributions so far (part 3)

If the population consists of all "yes"s and "no"s ("1"s and "0"s) and n is large enough so that both endpoints of $p \pm \sqrt{\frac{p(1-p)}{n}}$ fall in the interval (0,1) (using CLT):

- The sample proportion $\hat{p}$ is normally distributed with mean p (population proportion) and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.
- The number of "yes"s, Y, is normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$ **and** we should use the continuity correction (helps when approximating the discrete distribution of Y using the continuous normal distribution).