Recap: Sampling distributions of the sample mean

Sampling distribution of the sample mean

Sample mean:

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

What is the expectation of the sample mean? μ

What is the variance of the sample mean? $\frac{\sigma^2}{n}$

What is the distribution of the sample mean?

- Converges in probability to μ (WLLN)
- Normal as $n \to \infty$ (CLT)

CLT Application

CLT: Asymptotic but considered valid for most distributions when n>30

Assume Applied Statistics grades are normally distributed with mean 5.5 and standard deviation 1.8.

a. If I select one student at random, what is the probablity that their grade is greater than 8?

b.If I select 16 students at random, what is the probability that they will have a mean grade greater than 8?

c. Is our inference for (b) valid even though n<30? Why?

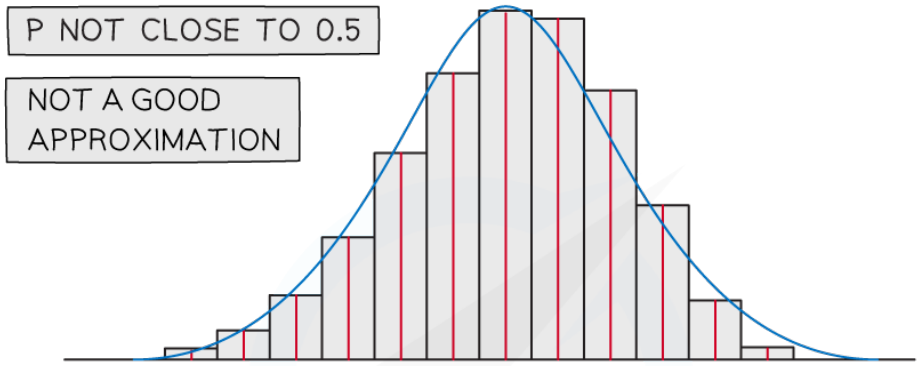
CLT Application: Binomial

Janice and Barbara decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the names. They carpool to school for 36 days. Calculate the following probabilities.

- Find the probability that Janice is the driver at most 21 days.
- Find the probability that Barbara is the driver for exactly 19 days.

CLT Application: Binomial

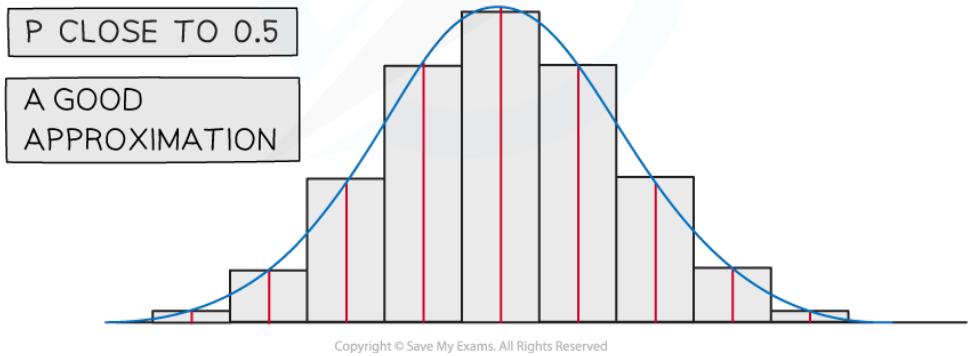
- X_i: Bernoulli with parameter p
- $S_n: X_1 + \cdots + X_n \sim Bin(n, p)$
- CLT: $\frac{S_n - np}{\sqrt{n\,p(1-p)}} \sim N(0,1)$



A binomial distribution $X \sim \mathrm{B}(n, p)$ can be approximated by a normal distribution $X_N \sim \mathrm{N}(\mu, \sigma^2)$ provided

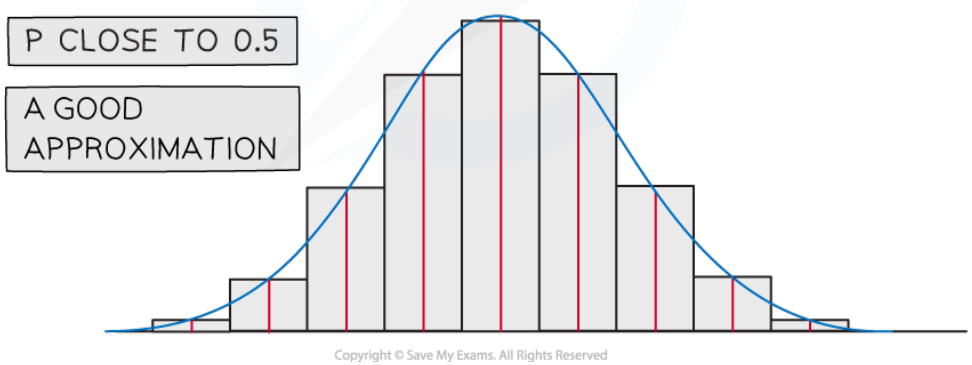
n is large or

p is close to 0.5



Correction for Continuity

- The **binomial distribution is discrete** and the **normal distribution is continuous**
- A **continuity correction** takes this into account when using a normal approximation
- The probability being found will need to be changed from a discrete variable, X, to a continuous variable, X_N
  - For example, $X = 4$ for binomial can be thought of as for normal as every number within this interval rounds to 4
  - Remember that for a normal distribution the probability of a single value is zero



CLT Application: Binomial

Janice and Barbara decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the names. They carpool to school for 36 days. Calculate the following probabilities.

- Find the probability that Janice is the driver at most 21 days.
- Find the probability that Barbara is the driver for exactly 19 days. (De Moivre-Laplace CLT to the binomial)

Confidence Intervals for a Proportion

Facebook’s categorization of user interests

Most commercial websites (e.g. social media platforms, news out- lets, online retailers) collect a data about their users’ behaviors and use these data to deliver targeted content, recommendations, and ads.

Facebook's categorization of user interests

Most commercial websites (e.g. social media platforms, news out- lets, online retailers) collect a data about their users' behaviors and use these data to deliver targeted content, recommendations, and ads.

Pew Research asked a representative sample of 850 American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook's categorization of user interests

Most commercial websites (e.g. social media platforms, news out- lets, online retailers) collect a data about their users' behaviors and use these data to deliver targeted content, recommendations, and ads.

Pew Research asked a representative sample of 850 American Facebook users how accurately they feel the list of categories Facebook has listed for them on the page of their supposed interests actually represents them and their interests.

Q: Are the categories listed by Facebook accurate?
A: 578 "yes", 272 "no"
Estimate the true proportion of American Facebook users who think the Facebook categorizes their interests accurately.

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

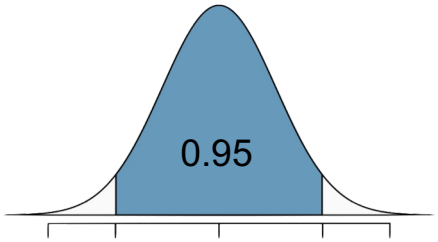
Facebook's categorization of user interests

$$\hat{p} = 0.68, \quad n = 850$$

Facebook's categorization of user interests

$$\hat{p} = 0.68, \quad n = 850$$

Distribution of $\hat{p}$: $N(p, \sqrt{\frac{p(1-p)}{n}})$



We do not know $\sqrt{\frac{p(1-p)}{n}}$

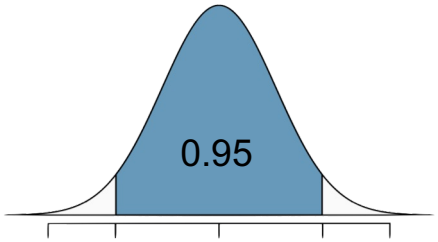
Substitute with the $\text{SE} = \frac{sample\ std}{\sqrt{n}} = 0.016$

$$P\left(-1.96 < \frac{\hat{p} - p}{SE} < 1.96\right) \geq 0.95$$

Facebook's categorization of user interests

$$\hat{p} = 0.68, \quad n = 850$$

Distribution of $\hat{p}$: $N(p, \sqrt{\frac{p(1-p)}{n}})$



We do not know $\sqrt{\frac{p(1-p)}{n}}$

Substitute with the $\mathrm{SE} = \frac{sample\ std}{\sqrt{n}} = 0.016$

$$P\left(-1.96 < \frac{\hat{p} - p}{SE} < 1.96\right) \geq 0.95$$

$$P(\hat{p} - 1.96 \times SE < p < \hat{p} + 1.96 \times SE) \geq 0.95$$

approximate 95% confidence interval

Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence intervals

$$\hat{p} \pm 1.96 \times SE$$

$$CI = [\hat{p} - 1.96 \times SE,\ \hat{p} + 1.96 \times SE]$$

For $\hat{p} = 0.68,\ SE = 0.074$

$$CI = [\hat{p} - 1.96 \times 0.074,\ \hat{p} + 1.96 \times 0.074]$$
$$= [0.666, 0.694]$$

What does 95% confident mean?

If I repeat the poll again and again, and compute 95% confidence intervals [A, B] as follows:

$$\text{point estimate} \pm 1.96 \times \text{SE}$$

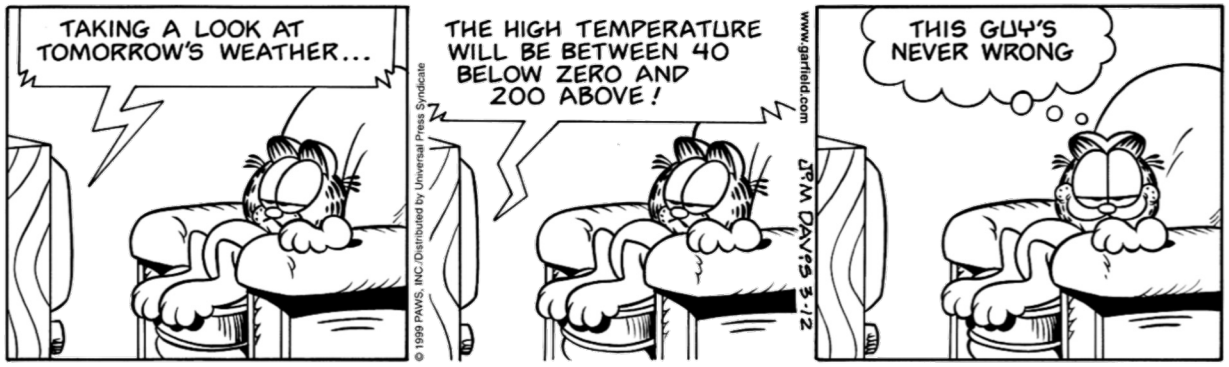
then 95% of these confidence intervals will include the true value p.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Changing the confidence level

$$\text{point estimate} \pm z^\star \times SE$$

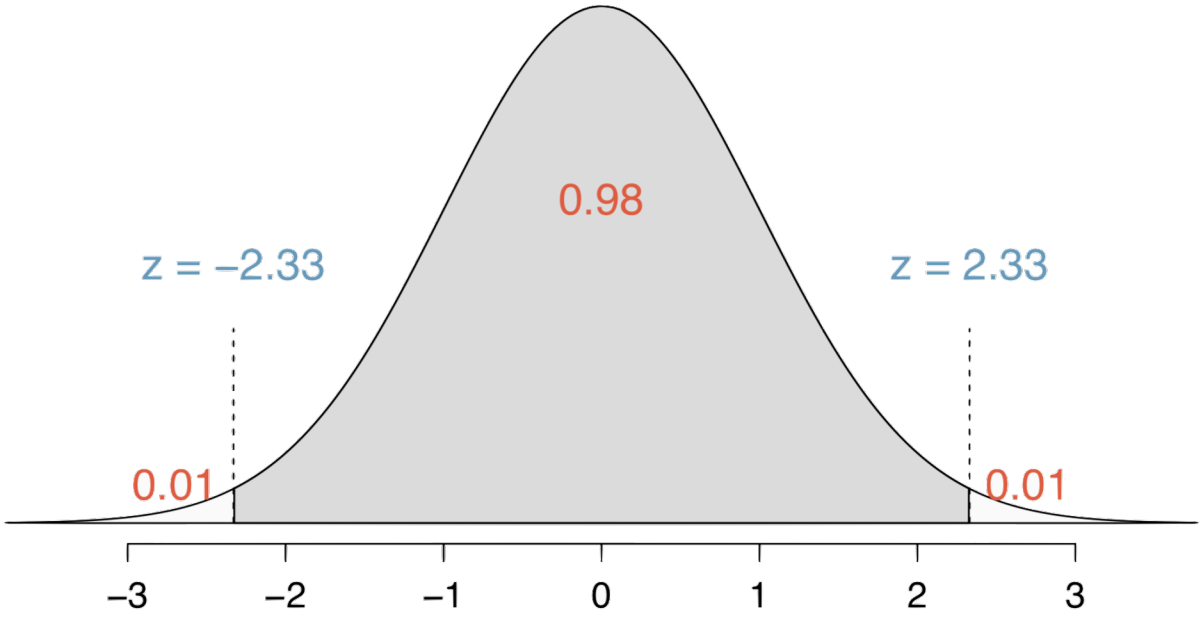
- In a confidence interval, $z^\star \times SE$ is called the **margin of error**, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) Z = 2.05
(b) Z = 1.96
(c) Z = 2.33
(d) Z = −2.33
(e) Z = −1.65

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) Z = 2.05
(b) Z = 1.96
(c) Z = 2.33
(d) Z = −2.33
(e) Z = −1.65



Interpreting confidence intervals

Confidence intervals are ...

- always about the population
- are not probability statements
- only about population parameters, not individual observations
- only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter

Practice: Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

A more accurate interval

Confidence interval, a general formula

$$\textit{point estimate} \pm z^* \times SE$$

Confidence interval

Confidence interval, a general formula

$$\textit{point estimate} \pm z^* \times SE$$

Conditions when the point estimate = $\bar{x}$

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
2. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

Confidence interval

Confidence interval, a general formula

$$\textit{point estimate} \pm z^* \times SE$$

Conditions when the point estimate = $\bar{x}$

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
2. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

Note: We will discuss working with samples where $n < 30$ in the next chapter.

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* ± *1.96 x SE*.
- Then about 95% of those intervals would contain the true population mean (μ).
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.