Sampling distributions and the CLT

Objective

(1) Get "big picture" view on drawing inferences from statistical studies.

(2) Understand the concept of sampling distributions & sampling variability.

(3) Learn the Central Limit Theorem & how to use it.

Concept briefs:

* Statistical inference = Rigorous, statistically valid, conclusion drawn about sampled data via probability based analysis.

* Sampling distribution = Hypothetical quantitative variable whose values consist of the same statistic estimated using different samples.

* Central limit theorem for proportions = Sample proportions follow normal model N(p,σ) where p=true population parameter, $\sigma = \sqrt{p(1-p)/n}$.

* Central limit theorem for means = Sample means follow the normal model N(μ,σ) where μ=true mean, σ true SD/$\sqrt{n}$.

* Assumptions for which the above theorems hold are very important.

Statistical inference: What & why

Illustration

A survey of smoking among college students is conducted in 2 different states: IN and CA. The % of smokers in each surveyed sample is

CA: 35% **IN:** 30%

Can we conclude that the % of college students who smoke is larger in CA than in IN? Why, or why not? [Statistically significant difference?]

Assume both surveys used good sampling methods, with the same sample size (e.g, 2000 people) and identical questions.

Illustration 2

A study of dropout rates at 4-year colleges is conducted in the years 2000 and 2004. The mean dropout rates in the samples studied were

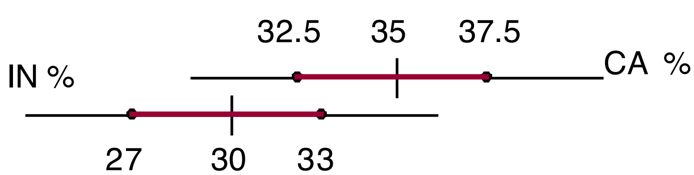
2000: 15% **2004:** 12%

Can we conclude that the % of students who dropped out was lower in 2004 than in 2000? Why, or why not?

Again, assume sound sampling methods, equivalent samples, etc.

Suppose you have more info.

In the smoking study, the margin of error for the IN sample is 3.0%, and for the CA sample 2.5%.



Two key strategies of statistical inference that we will study:

(1) Finding error margins.

[Technical term for this: Confidence intervals]

(2) Determining when sampled differences are statistically significant.

[Technical term for this: Hypothesis test OR Test of significance]

Central to everything we study is the concept of:

* Sampling distributions; and

* Theorems about sampling distributions.

Illustration of sampling distribution concept

* Suppose we want to determine the following two statistics for students currently enrolled at Earlham:

(1) What % of them live off-campus

(2) The mean GPA for all students

* Suppose we randomly sample 100 students to estimate these statistics, and we ask the questions:

(1) Do you live off-campus? (Y or N) {Categorical data}
(2) What is your GPA? {Quantitative data}

and we find:

12.4% students respond "Y" to live off-campus

3.06 is the mean GPA of this sample

Q: How close are these estimates to the true population parameters we want?
Q: If we pick another random sample of 100 students, how much would these same statistics differ?
Q: How can we accommodate sampling differences in our interpretation of statistical data?

Sampling distribution models give a sound theoretical basis for answering these questions.

Concept 1: Distributions that consist of sampled statistics.

* In the above example, imagine calculating the same two population parameters from a 500 different random samples of size 100 each.

* This would give what we call a "Sampling Distribution."

Concept 2: How to describe & analyze sampling distributions.

* We can construct histograms, boxplots, etc.
* We can calculate mean, SD, median, IQR **. . .**

* In the above example, imagine calculating the mean of the % of students who live off-campus, and the mean of the mean GPA!

Concept 3: Normal models for sampling distributions.

* Histograms of sampling distributions are often symmetric, unimodal.
* Normal model applies for analyzing them.

Suppose we have collected data in the above example by asking the following 2 questions (with sample size=10):

(1) Do you live off-campus? (Y or N)
(2) What is your GPA?

Student #	Off campus?	GPA
1	N	2.81
2	N	3.29
3	Y	3.30
4	N	2.72
5	Y	3.75
6	N	1.91
7	N	2.99
8	N	3.80
9	N	3.27
10	N	2.91

Quantitative

Such data yields "mean" type of statistic.

Distinguish between 3 different distributions here: (1) population GPA, (2) sample GPA, and (3) sampling distribution of mean GPA.

Thus, there is a <u>separate value</u> of mean GPA for the population, the sample, and the sampling distribution.

Categorical

Such data yields "proportion" type of statistic.

Think about it:

Suppose you know the shape of the true population distribution for a quantitative variable. [E.g., GPA distribution of students is bimodal with strong left skew.] What shape would you expect the following distributions to have: (1) random sample of size, say, 10% of the population, (2) sampling distribution of mean values within such random samples?

Central limit theorem for "proportion" type statistics

- This refers to % type statistics - typically comes from categorical variables.
- We always convert proportions to fractions (so 100% becomes 1).
- Note that there are only 2 statistics we can have here:

    The proportion & 1 - The proportion.

- Notation: $\hat{p}$ = Proportions estimated from different samples.

    (i.e., $\hat{p}$ denotes the sampling distribution).

    p= True proportion (i.e., population parameter) that we want.

    n = size of the samples.

    $\mu\ (\hat{p})$ = mean of the sampling distribution of $\hat{p}$.

    $\sigma\ (\hat{p})$ = standard deviation of the sampling distribution of $\hat{p}$.

Statement of the central limit theorem:

When the conditions are met, the sampling distribution of a sample proportion follows the normal model $N(p, \sqrt{\frac{p(1-p)}{n}})$, where n = size of the sample, and p= true value of the proportion in the population of interest.

The conditions for this result to hold:

1. The samples must be independent.
   In practice, this holds if sample is random and $n < 10\%$ of population.
2. The samples must be sufficiently large in size.
   In practice, this holds if: $np \geq 10$ and $n(1-p) \geq 10$.

Central limit theorem for "mean" type statistics

- The above result for proportions can be generalized to sample means.
- Notation: $\bar{x}$ = Mean values estimated from different samples.

  (i.e., $\bar{x}$ denotes the sampling distribution).

  μ= True mean in the population of interest.

  σ = True standard deviation in the population of interest.

  n = size of the sample.

Statement of the Central Limit Theorem:

When the conditions are met, the sampling distribution of a sample mean value follows the normal model $N(\mu, \frac{\sigma}{\sqrt{n}})$, where n = size of the sample, μ= true mean value in the population of interest, σ = true standard deviation in the population.

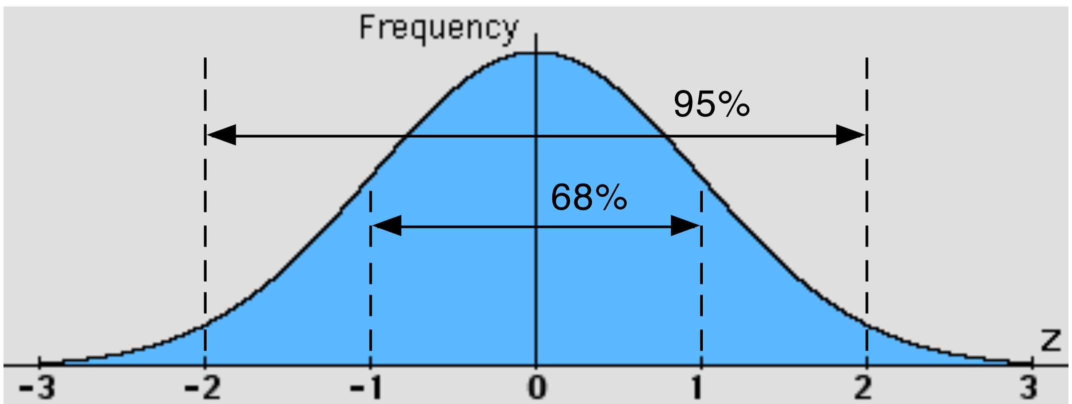
The conditions for this result to hold:

1. The sample must be independent.
   In practice, this holds if sample is random and $n < 10\%$ of population.
2. The sample must be large enough.
   There is no simple rule of thumb to verify this in practice.

The Normal model as a probability model

Recall: Probability = Relative frequency (over the long-term).

The normal model, as we've used it previously, tells us what % of a distribution lies how many SD's from the mean.



Normal model for probabilities:

* Simply convert % to fractions and treat as probability.
* Use z-tables as normal probability distribution tables.
* Note that total probability = total area under normal curve = 1.0

E.g: Suppose the GPA distribution of students is approximately normal, with μ = 3.06; σ = 0.4. What is the probability that a randomly selected student has GPA $\geq$ 3.86?

Ans: 3.86 is exactly 2σ above the mean.

So, P(GPA $\geq$ 3.86) = [1 - 0.95] / 2 = 0.025.

Exercises 14 & 16 **(see copy of exercise on last page)**

Strategy for 14: [part (C) only]

* Assume 50 is large enough to meet the conditions of normal models for means

* Find the mean & SD of normal model (using Central Limit Theorem (CLT) with true mean=$32, true SD=$20, and n=50)

* Find z-score corresponding to $\bar{y}$=$40 (carefully, using the right σ).

* Lookup standard normal table & find area above this z-score.

Strategy for 16 (a):

* Find mean purchase per customer from given total rev. & number of cust.

* Use CLT with μ=$32, σ=$20, and n=312 to construct normal model.

* Find z-score for the mean purchase per customer (watch the σ you use).

* Lookup standard normal table & find area above this z-score.

Strategy for 16 (b):

* Recognize that 10% of worst days correspond to $\bar{y}$ values in the bottom 10% of the normal distribution curve (of the means).

* Lookup 0.10 in the standard normal table & find corresponding z-score.

* Find $\bar{y}$ for this z-score (being careful to use the right σ).

* Total rev. = 312 x $\bar{y}$.

Exercise 28 **(see copy of exercise on last page)**

Solution:

* Q. pertains to sampling distribution of a proportion.

* Can apply CLT if we check conditions & verify they're satisfied:

(1) Independent sample? Check random & $n < 10\%$ of population.

Random: True, since 100 students are randomly picked.

$n < 10\%$: True, if we assume 100 less than 10% of students on campus.

(2) Large enough? Check whether $np > 10$ and $nq > 10$.

$p=0.3$, $q=0.7$. So: $np = 30 > 10$, $nq = 70 > 10$.

* Thus, normal model would be appropriate for the sampling distribution of this proportion. From the CLT, its mean and SD would be:

$$\mu = 0.3, \text{ and } \sigma = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.3)(0.7)}{100}} = 0.0458$$

Model is: N(0.3, 0.0458)

(b) Find z-score for $\hat{p} = 1/3 \Rightarrow z = [1/3 - 0.3] / 0.0458 = 0.7278$.

Lookup z-table & find area for $z > 0.7278$. We get: $1 - 0.7673 = 0.2327$.

Answer: Probability that in this sample more than 1/3 wear contacts is 0.2327.

Exercise 36 **(see copy of exercise on last page)**

Strategy:

* Check conditions for normal model for proportions.
* Identify the "proportion" of interest: % of children with genetic condition.
* Identify p; find mean & SD of normal model; and sketch a rough graph.

* We want to find 20 subjects out of 732 ---> $\hat{p} = 20/732 = .0273$

* Interpret Q. as: "What % of samples have $\hat{p}$ larger than .0273"?

* Find z-score corresponding to $\hat{p} = .0273$

* Lookup standard normal table & find area above this z-score. This is the probability of finding enough subjects for the study.

14. Groceries A grocery store's receipts show that Sunday customer purchases have a skewed distribution with a mean of \$32 and a standard deviation of \$20.

a) Explain why you cannot determine the probability that the next Sunday customer will spend at least \$40.

b) Can you estimate the probability that the next 10 Sunday customers will spend an average of at least \$40? Explain.

c) Is it likely that the next 50 Sunday customers will spend an average of at least \$40? Explain.

16. More groceries Suppose the store in Exercise 14 had 312 customers this Sunday.

a) Estimate the probability that the store's revenues we at least \$10,000.

b) If, on a typical Sunday, the store serves 312 custome how much does the store take in on the worst 10% o such days?

28. Contacts Assume that 30% of students at a university wear contact lenses.

a) We randomly pick 100 students. Let $\hat{p}$ represent the proportion of students in this sample who wear contacts. What's the appropriate model for the distribution of $\hat{p}$? Specify the name of the distribution, the mean, and the standard deviation. Be sure to verify that the conditions are met.

b) What's the approximate probability that more than or third of this sample wear contacts?

36. Genetic defect It's believed that 4% of children have a gene that may be linked to juvenile diabetes. Researchers hoping to track 20 of these children for several years test 732 newborns for the presence of this gene. What's the probability that they find enough subjects for their study?