Sampling Distribution and Central Limit Theorem

**Sampling Distribution and CLT pt.1 & 2**

* Distribution of the sample (sample distri.): display of the distribution of actual data values collected in one sample
* Sampling distribution: display of summary statistics for many different samples
* The statistic can be mean, proportion, sum, variance, etc.
* Sampling distribution models how a statistic from a sample varies from sample to sample
* Randomness: outcome is random if we don’t know which particular value it takes; however, we may know the possible values it can have
* Pattern
* The more dice you toss: the sample average goes closer to 3.5 and the shape is approaching a bell-shaped curve
* The larger the sample size in each sampling: the sample mean is approaching to the population mean and the sampling distribution is approaching the normal distribution
* Central Limit Theorem (CLT): the sampling distribution of any mean approaches to normal as the sample size gets larger
* CLT conditions:
* Independence assumption or randomization condition
* Sample size condition: if the popul. Is unimodal and symmetric, a small sample is okay and if the popul. Is skewed, a much larger sample size may be needed
* Regardless of the popul. Distribution, the sampling distrib. Of mean approaches to an approximately normal distribution (z distribution) when sample size is large enough
* The clt can be applied to other statistics (proportion, sum, variance)
* Which normal?
* Mean centered at population mean u
* SD = o/sqrt(n)
* Mean varies less than individual observations
* Having one student with GPA 4.0
* Having the mean of all the students with GPA 4.0
* Equations & standard error
* Mean: the mean of mean approaches population mean
* Standard error (stand. Dev of mean) or SE: se = o/sqrt(n)
* Estimated SE using the standard deviation of a sample
* Why CLT is important
* As the sampling distribution of the mean is approximately a normal distribution;

We can construct a confidence interval easily

Construct hypothesis testing easily

When sample size is large enough, one sample can be good to estimate the population mean

* Histogram shows the distribution of all ISAT 251 students’ height is a distribution of a sample
* Based on the CLT, the sampling distribution will approximate the normal distribution when sample size is large enough, regardless of the variable’s distribution in the population
* Histogram shows the distribution of average birth weight of babies is an example of sampling distribution of mean
* CLT is important because we can model the sampling distribution of some statistic of a variable simply with the normal distribution even if that variable is not distributed normally

Friday 10.05 video lectures

**Lab Exercise 4 – Hand Span & Height Normal Distribution**

**Stats w/R: 3.6: Sampling Distribution and the CLT**

* Sampling error
* Sample statistics can provide an approximation to what the true population parameters will be
* Distribution of a sample: how frequently different outcomes have been observed within a small portion of the population
* Distribution of a population: how frequently different outcomes can be observed in a larger population (theoretical)
* Sampling distribution: mean or proportion describes what you see after repeatedly sampling a population an infinite number of times
* Sampling distribution model of a mean tells that the mean of the sampling distribution is the mean of the population
* The bigger the n, the thinner the sampling distribution will be meaning the standard error of your estimate will be smaller
* CLT: averages or sum of each repeated samples from a population with the collection
* Bigger sample size is usually better, because this will force the varability of the sampling distribution to become tinier

Confidence Intervals: One Mean

**Confidence Intervals of One Mean, t Distribution pt.1 & 2**

Confidence Intervals

* Can we be sure that a sample mean be the truth (population mean)?
* Random sample
* Random errors
* A confidence interval (CI) is an interval containing the most believable values for a parameter
* A range of values that’s likely to include the true value (the population parameter)
* Point estimate ± margin of error
* A CI is an interval (%) containing the most believable values for a parameter
* For the sampling distribution of the mean with large sample size
* Equations (pg.17)
* 90% CI = 1.645
* 95% CI = 1.960
* 99% CI = 2.576

Margin of Error

* In a confidence interval, the extent of the interval on either side of the observed statistic value (e.g. sample mean) is called the margin of error
* Smaller margin of error
* Narrower or wider CI?
* More or less precise?
* For a sampling distribution of the mean with the large enough sample size

ME = z\* x se = z\* x o/sqrt(n) = z\* x s/sqrt(n)

* Z\* (z star) is the critical value & describe how wide the CI to be
* Se is standard error

How to Interpret CI

* True mean (or population mean) of (your research project) is between lower CI bound and upper CI bound

When the sample size is small or not large enough

* A family of distributions similar to the normal model (z distributions) with larger variation is needed
* Variation depends on the sample size
* T distributions (student’s t)

T Distribution (Student’s t)

* A bell-shaped distribution has slightly thicker tails (i.e., more variability) than a normal distribution
* The distribution of the t-models depends on a parameter called degrees of freedoms

Degrees of Freedom

* A distribution representing the number of independent quantities that are left after we’ve estimated the parameters
* The number of data values minus the number of estimated parameters
* Df = n-1, where n is the sample size
* The larger the degrees of freedom, the (larger or smaller) the se in your t distribution?
* Df (up) = n (up) – 1 vs. se (down) = s/sqrt(n) (up)

Margin of error with t distribution

* For the sampling distribution of the mean with unknown and a smaller sample size
* ME = t\* x se = t\* x s/sqrt(n)
* T\* is the critical value & describe how wide the CI to be
* For the same % CI, t\* is larger than z\*

Video lectures 10.12

* When we have a smaller sample size, t distribution (students’ t) should be used to construct a CI for mean
* 99% CI IS widest
* 90% has highest precision
* Confidence interval help us to figure out the range of likely value for a population parameter based on statistics from a sample

**3.7: P-Values, Confidence Intervals, and Controversy**

* null hypothesis & alternative hypothesis
* p-value: probability that sample actually came from the population that the null hypothesis describes
* portion of the area underlying a probability distribution function
* CI, power analysis, effect sizes, accuracy in parameter estimation (aipe), Meta-analysis, Not statistically significant, Data open and accessible

**4.1 One mean**

* Qt(ci, df=29)
* Arguments to the ci.mean (245)

**Lab exercise 5 – rolling dice**

Confidence Intervals for Two means and Paired Means

*Confidence intervals for two means*

CIs with two means

* Comparing the means of two groups
* The differences between the averages of the 2 groups
* True difference?
* Two groups statistically different?
* The sampling distribution for the difference between two means
* Normal distribution for the sampling distribution of x1 – x2

Mean = u1 – u2

CIs for the Difference between the Two Means

* Confidence intervals
* The population standard deviation is known is unknown
* Critical value & standard error
* The variances of the two groups are the equal or not equal: standard error & degree of freedom
* Use the proper standard error or standard deviation for se
* Use the proper degree of freedom for t\*

Assumptions & Conditions for CIs for two means

* Independence
* Independent observations within each group
* Independent between the two groups
* Sample size
* Small enough (each group)
* Large enough (each group)

When both groups have large sample size, CLT starts to work

* Nearly normal condition
* When sample sizes are small
* Check it for each group
* Check whether there are outliers
* Equal variance assumption for pooled t-interval

*Confidence intervals for paired means*

* What if the data in the 2 groups are not independent; moreover, they are paired

Cis with Paired means

* Paired data: an observation of a quantitative variable in one group is paired with an observation in the second group
* Dependent observations within each pair
* The sampling distribution for the mean of the difference between the observations in each pair
* D = mean, where di = xi – yi and I from 1 to n
* N is the number of pairs
* Normal distribution of d: mean and standard dev

Cis for the Difference between Paired Means

* Confidence intervals
* CI = point estimate ± margin of error

Where critical value is t\*, n is the sample size or the number of pairs, and se(d) is standard error of the mean of the difference between each pair

* The popul. Stand dev is known or unknown

Critical value & standard error

Interpretation of CI for paired mean

* I am (90,95,99) % confident that the true (what the difference is referring to) of (your research object) is between (lower bound of the CI) and (upper bound of the CI).

Assumptions & conditions of CI for paired means

* Paired data condition
* Understand how the data were collected
* Independence
* The differences must be independent

Pairs are random samples

Independent individual observations

* Sample size
* Small enough
* Large enough

When the number of pairs is large enough, CLT starts to work

* Near normal condition
* When sample is smaller check the distribution of the differences
* Target value or true value a CI of two means estimates = differences between two means
* A CI of paired means estimates = mean of differences
* Two groups of data that have equal variance can pool data together using pooled standard deviation, during the process of calculating CIs of two menas

Video lectures 10.19

**4.2: Two Means**

* Point estimate (y1 – y2)
* Real difference between population means (u1 – u2)
* Sampling distribution and standard error of the difference between means
* Form of the confidence interval for two means
* Choose the right expression for standard error
* Equal and unequal variance
* Calculate pooled standard dev
* Choose the right expression for degrees of freedom
* Arguments to the ci.twomeans function

**4.3: Paired Means**

* Dependent observations
* Sampling distribution and standard error of the difference between means
* Standard error is a characteristic of a random sampling process, not of the population
* To ci.paired

**Lab Exercise 6 – m&ms bag weight**

CIs: Proportion(s)

*CIs for One Proportion & Two Proportions*

Basic ideas & Notations about Proportions

* Binary responses
* # of observations or subjects: n
* # of “successes”: the number of observations or subjects given you’re interested in
* # of “failures”: n - # of successes
* Proportion of successes (“sample proportion”): # of successes/n
* Proportion of failures: # of failures/n = 1 – p
* The population proportion (of successes) (truth or true value): p
* Q = 1-p
* Always z-distribution

Common Approach to CIs for One Proportion

* the Wald Confidence Interval
* equation in book

Other Options for CIs for One Proportion

* the Wald Confidence Interval
* General & common
* Inaccurate & not reliable: coverage probability is lower than it should cover
* The agrestic-Coull correction confidence interval
* Simple adjustment to wald
* Easy to calculate

The Agresti-Coulli Correction

- add 2 to the # of successes (and add 2 to the # of failures) and add 4 to the total observation

> new n = original n + 4

New p =

How would you interpret the CI you get in your example?

I am (90,95,99) % confident that the true proportion at (your research object) is between (lower bound CI) and (upper bound CI).

* Use percentages rather than fractions or decimals

Assumptions & Conditions for One proportion

* Independence assumption or randomization condition
* Sample size is small enough
* 10% condition (less than 10% of the population)
* Sample size is large enough (to use the normal distribution)
* success/failure condition
* at least 10 successes & at least 10 failures

CIs for Two proportions

* if we would to like to compare of proportions from 2 groups..
* wald confidence intervals

Assumptions & conditions for two proportions

* independence assumption or randomization condition
* independent groups
* sample size is small enough (10% condition for each group population)
* sample size is large enough (success/failure condition with at least 10 successes and failures for each group)

How to interpret the CI?

I am (90,95,99) % confident that the true difference in proportion of (rp) is between lower bound CI and upper bound CI

**4.4: One proportion**

* many variations on equations for the CI for one proportion
* The Wilson score interval is the one you should use if you’re relying on the R software to compute your intervals
* The Agresti-Coull correction is the one you should use if you want a reasonably accurate confidence interval that you would also like an easy way to compute analytically
* Coverage Probabilities

**4.5: Two proportions**

* Agresti-Caffo (“Adjusted Wald”) CI is the one you should use if you want a reasonably accurate confidence interval that you would also like an easy way to compute
* Adding 1 success to the observations in each group and adding 2 total observations to each group
* Newcombe Hybrid Score is the one to use if you are OK with relying on the R software to compute and very accurate

Sampling Survey, Observation Studies, and Experiments

Experiment vs Observation Studies

* Experiment
* Controlled or manipulated environment
* Control group and treatment group
* Effect and Causality
* Medical or clinical test, placebo (control group), manipulated environment, try to find effect or causality, blind study, mostly quantitative, a/b test
* Observation Studies
* Prospective or retrospective observation
* Sample survey
* Relationship and association
* Survey, mostly find association, mostly qualitative, prospective or retrospective, Up series

Sample Survey

* Randomness is important
* Representativeness is important as well
* Sample should reflect the population
* Biased sample: under- or over-emphasized some characteristics
* Common bias
* Voluntary response bias
* Nonresponse bias
* Response bias
* Undercoverage

Common Sampling Strategies

* Simple Random Sample
* Cluster Sampling
* Stratified Sampling
* Systematic Sampling

**Stats w/R: 3.1**

* What is randomness?
* Probability is the long run relative frequency of an event
* The law of large numbers say that when you look at all of your outcomes from a multitude of trials, the results will converge on some expected value – even when the individual outcomes are random
* Common sampling strategies (167)
* Simple random sample
* Systematic sampling
* Stratified sampling
* Cluster sampling
* Convenience sampling
* Voluntary response sampling
* Preventing bias: undercoverage and non-response bias
* Both forms of selection bias
* Voluntary response bias
* Biased nonresponse
* Response bias
* Social desirability bias

**3.2: Experiments vs. Observational Studies**

* Experiments
* Experimental units
* Factor has some factor levels
* Control group
* Treatments
* Blocking
* Aspects of experiment (175)
* Experimental unit, control group, randomization, factor, factor level, treatment, blocking, blind, double blind, placebo, replication, and hawthorne effect
* Observational studies
* Prospective (longitudinally) or retrospective