topic 13: CLT + Confidence Intervals

stuff to learn today:

1. first statistical tests!
   a. setting up hypotheses
   b. one sample z-test
2. Central Limit Theorem
3. sampling (standard error)
4. confidence intervals
5. the t-distribution

Our First Statistical Test

A data scientist wants to examine if there is an effect on IQ scores when using tutors. To analyze this, she conducts **IQ tests on a sample of 40 students** and wants to compare her students' IQ to the general population IQ. The way an IQ score is structured, we know that **a standardized IQ test has a mean of 100 and a standard deviation of 16**. When she tests her group of students, however, **she gets an average IQ of 103**. Based on this finding, does tutoring have an effect?

Statistical Testing Process

1. Set up hypotheses
2. Pick the statistical test based on your experiment
3. Pick your alpha (level of significance)
4. Calculate your test statistic
5. Find your p-value
6. Interpret

1. Setting up hypotheses

- every experiment starts with a **null** and an **alternative** hypothesis
- typically, the null hypothesis represents no effect
- is the test one-tailed (left or right) or two-tailed?
- for tutoring to have an effect, we want to **reject the null hypothesis** (vs **fail to reject**)
- the data we have:
  - population: $\mu = 100$, $\sigma = 16$; sample: $x_{bar} = 103$, $n = 40$
- H_0: $x_{bar} = \mu$ H_A: $x_{bar} \geq \mu$

2/3. Picking the Test & Alpha

- our first test is the **one sample z test**
- conditions:
  - you know the **population** mean and standard deviation

- alpha (α) is your cut-off value; the threshold for rejecting your null hypothesis - typically set at 0.05
- confidence level = 1 - α

- you'd interpret your results as: "with a confidence level of 95%, we can (fail to) reject the null hypothesis that..."

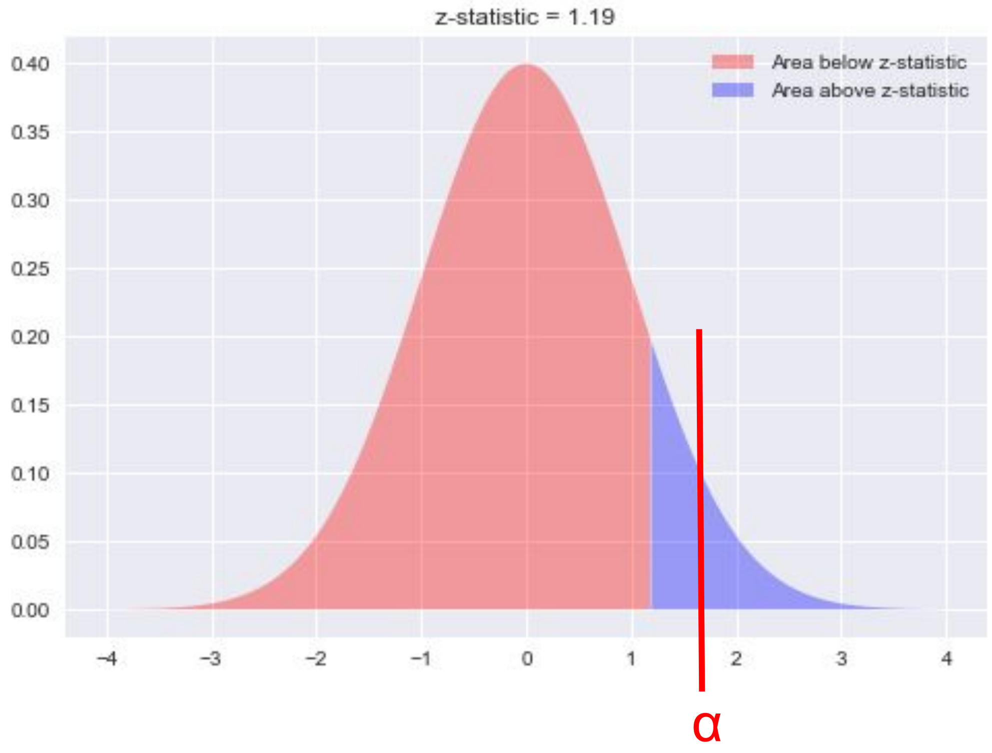
4/5. Calculating the test statistic & p-value

- for a z-test, the z-score is our test statistic
- the data we have:
  - population: $\mu = 100$, $\sigma = 16$; sample: $x_{bar} = 103$, $n = 40$

- $z = (x_{bar}-\mu)/(\sigma/\sqrt{n}) = (103-100)/(16/\sqrt{40}) = 1.19$
- look up the associated probability from the z-table -- 0.883
- average IQ of tutored students is greater than 88% of population

- for a right-tailed test, the p-value is 1 - 0.883 = 0.12, which is greater than our alpha of 0.05

6. Interpretation



central limit theorem

- the CLT states that **independent random variables summed together will converge to a normal distribution as the number of variables increases**
- the distribution of sample means of **any population**, as the sample size increases, will converge to a normal distribution

- this allows us to use sample statistics to make inferences or estimates on the population

sampling statistics: standard error

- *standard error (SE)* is very similar to standard deviation. the higher the number, the more spread out your data is
- while the standard error uses statistics (sample data), standard deviations use parameters (population data)
- the calculation for the standard error of the sample mean is:
- $\sigma_{xbar} = \sigma/\sqrt{n} \approx s/\sqrt{n}$

- σ is the population standard deviation, which we *approximate* with the sample standard deviation (s)

confidence intervals

- a **confidence interval** is a range of values above and below the point estimate that **captures the true population parameter** at some predetermined confidence level
  - a 95% of confidence intervals constructed in this manner contain the true population mean
  - a 95% confidence interval does NOT contain 95% of the values

- calculated by taking: **mean ± margin of error**
- the margin of error depends on the underlying distribution

- if you know the population standard deviation, margin of error = $z * (\sigma / \sqrt{n})$
  - z is the z-score depending on your confidence level (like a two-tailed test)
  - can be looked up on the z-table

confidence intervals

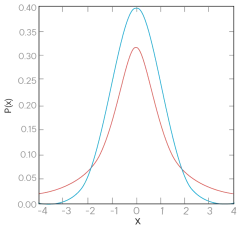
- a **confidence interval** is a range of values above and below the point estimate that **captures the true population parameter** at some predetermined confidence level
  - a 95% of confidence intervals constructed in this manner contain the true population mean
  - a 95% confidence interval does NOT contain 95% of the values

- calculated by taking: **mean ± margin of error**
- the margin of error depends on the underlying distribution

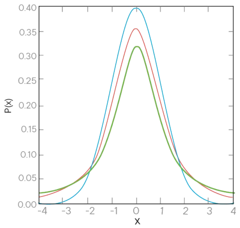
- if you know the population standard deviation, margin of error = $z * (\sigma / \sqrt{n})$
  - z is the z-score depending on your confidence level (like a two-tailed test)
  - can be looked up on the [z-table]

confidence intervals - t-distribution

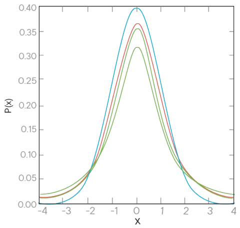
- a primer to our second kind of test (a t-test!)
- when we did a z-test, we **knew** the population standard deviation
- t-distributions are used when we don't know the population standard deviation
    - new parameter: degrees of freedom - the more dof, the more "normal"
    - degrees of freedom = n - 1
- t-table



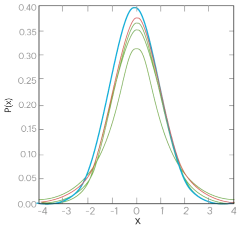
1 degree of freedom



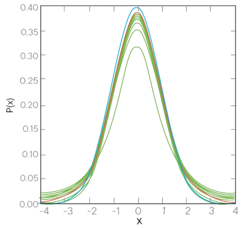
2 degree of freedom



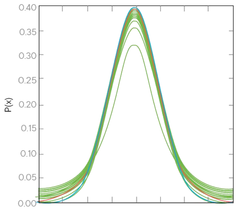
3 degree of freedom



5 degree of freedom



10 degree of freedom



30 degree of freedom

confidence intervals - t-distribution

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$

To review some vocabulary and terms:

(1) $\bar{x}$ is a "point estimate" of μ

(2) $\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$ is an "interval estimate" of μ

(3) $\frac{S}{\sqrt{n}}$ is the "standard error of the mean"

(4) $t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$ is the "margin of error"