• From last time
• Big idea for the day: all statements we make from evals are probabilistic
• Quick refresher - sample vs. population
  – Parameters of prob distribution vs. statistics of the sample
• The value of hypothesis testing
  – State your goal, test whether or not you achieved it
  – eg. a good thesis statement is a testable hypothesis
  – X is faster than Y
  – Z has negligible overhead
• We measure a sample mean, but it is really just an estimate of population mean
  – What is the full population here, and what is the sample?
    * Full populations: All possible executions of our experiment
    * Sample: The ones we actually run
  – We can get a confidence interval that helps us understand what we would get if we ran the experiment more times.
  – Book explanation of way to get confidence level
    * Get multiple samples (multiple trials per sample), compute stats on the means, treat that as a sample set and take confidence intervals
  – Again, iid comes up, and this is why you need to be careful in experiment design
    * When might you not meet identically distributed criteria? Caching, warmup, different conditions over time of day, etc.
  – Standard error — not to be confused with standard deviation or STDERR
    * Different samples drawn from the same population would have different means
    * The standard error of the sample mean is how close to that real mean you can expect to get
    * We are viewing the set of sample means as a distribution and basically looking at its variance
    * …using the normal distribution, thanks to cool properties of that distribution WRT iid variables
  – Don’t confuse with credible intervals
    * Probability of the true values
    * Requires a priori knowledge/estimation of distribution
• Confidence interval for sample mean
  - Lower: \( \bar{x} - \frac{z_{1-\alpha/2}s}{\sqrt{n}} \)
  - Upper: \( \bar{x} + \frac{z_{1-\alpha/2}s}{\sqrt{n}} \)
  - Why is this symmetric?
    * Because we’re modeling the sample means using the normal distribution
  - \( \bar{x} \) is sample mean
  - \( s \) is sample stddev
  - \( z_{1-\alpha/2} \) is \((1 - \alpha/2)\) quantile of unit normal dist \((\mu = 0 \text{ and } \sigma = 1)\) - note, you are picking \( \alpha \)
  - \( n \) is the sample size
  - So, what does this tell us?
    * We are \( x\% \) certain that the population mean is between \( x \) and \( y \)
  - What do we need to apply this result?
    * iid sample
    * Large samples (30 or greater)
    * Or sample itself is normally distributed
  - When is it not worth computing this?
    * When the means are extremely far apart
  - When is it important?
    * Close enough that it’s possible that means lie within each others’ confidence intervals
    - Testing for mean of particular value - does it lie within the CI?
    - When might you want your mean to be the same as another mean?
      * Showing insignificant overhead
  - Showing significance: Paired samples (eg. same benchmarks)
    - Take samples for two systems under the same workload
    - Compute statistics of the difference
    - Compute CI of mean of the difference
    - If CI contains zero, not statistically different: The hypothesis “the two systems are the same” is supported by the data
  - Showing significance: visual check
    - Draw both confidence intervals and means
    - If CIs don’t overlap, one is clearly better
    - If CIs do overlap, both means fall inside CI of the other: effectively the same
    - If the mean of one is in the CI of the other, but this is not true for both, t-test required
  - Showing significance: t-test (eg. truly random samples)
    - Best to leave the implementation of this up to someone else
    - Degrees of freedom: number of independent sources of data that go into the model: number of samples minus steps that go into the estimation
    - eg. R includes this as a module
- Fun fact: t-test invented as a way of measuring the quality of beer (Guinness Stout)

- Picking CIs
  - As discussed before, degree of confidence has to do with the gain/loss of being outside the range
  - Reiterate plane example, you don’t want to fly on a plane built with only 99% confidence intervals

- Picking a sample size
  - What should our goal be when picking a sample size?
    * Give us a high degree that our sample mean is close to the population mean
    * While taking a reasonable amount of time to run experiments
  - Data dependent - on variance, which is intuitive
  - Getting a good mean
    * Pick the confidence level we want (say, 95%)
    * Pick the accuracy level (how far on either side)
    * The accuracy equal to the confidence bounds, solve for $n$
      * $n = \left( \frac{100 \times z}{r \times x} \right)^2$
      * $s$ is sample standard deviation
      * $\pi$ is sample mean
      * $z$ is as above ($z_{1-\frac{r}{2}}$, quantile of unit normal distribution (note, contains the confidence level))
      * $r$ - accuracy in plus or minus $r\%$
  - Comparing two systems
    * Goal is to end up with non-overlapping intervals at some confidence level
    * For comparing two, upper edge of lower must be below lower edge of upper
    * Run some sample experiments to get an initial mean and stddev for each
    * Fill in all numbers for both confidence intervals except $n$
    * Set the upper bound of one to be lower than the lower bound of the other
    * Solve for $n$

- For next time
  - Bring your laptop
  - Sign up for GENI account
  - Read GENI paper posted on Canvas