Test 2 Checklist

1. probability of an event "E"
   (a) $0 \leq P(E) \leq 1$
   (b) the complement of E ("not E") is denoted $\overline{E}$

$$P(\overline{E}) = 1 - P(E)$$

2. types of probability:
   (a) empirical: relative frequency of historical results, experiments, simulations
   (b) subjective: educated guess based on the knowledge/belief of the observer
   (c) theoretical: mathematical model, deductive reasoning
3. probability distribution for a random variable x (x=L1, $P(x)$=L2)
   (a) sample space Ω is the set of possible outcomes
   (b) for empirical distributions, the sum of frequency is the sample size n
   (c) PDF = probability = proportion = relative frequency
   (d) use CDF to get percentiles
   (e) $\sum P(x) = 1$
   (f) "expected value" is the mean
4. randomness, simulations, law of large numbers
5. equally likely outcomes $P(E) = \frac{\#\text{ E}}{\#\ \Omega}$
6. joint probability
   (a) "not" - do the complement
   $P(\text{not } A) = 1 - P(A)$
   (b) "and" - multiply probabilities of each step, given the previous ones
   $P(A \text{ and } B) = P(A)P(B|A)$
   (c) A and B are independent if $P(A \text{ and } B) = P(A)P(B)$ (the events don't inform each other)
   (d) with vs without replacement
   (e) "or" - use addition, but don't double-count if the events are not disjoint (Venn diagram)
   $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
   (f) if an event could happen different ways, add the probabilities
   e.g. probability of 2 heads in three tosses = P(HHT) + P(HTH) + P(THH)
   (g) "at least one" easier to work with if you translate to "not zero"
7. binomial distribution $X \sim BI(n, p)$
   (a) x is the sum of repeated, independent, and identical binary trials
   (b) theoretical mean $\mu = np$, st dev $\sigma = \sqrt{np(1-p)}$
   (c) binompdf(n,p,x) gives the probability of exactly x
   (d) binomcdf(n,p,x) gives the probability of less than or equal to x
   (e) be able to recognize binomial distribution in applications

8. normal distribution $X \sim N(\mu, \sigma)$
   (a) bell shaped with thin tails
   (b) 68, 95, 99.7 "empirical rule", and connection with z-scores
   (c) draw a picture when you work these problems
   (d) normcdf(a, b, μ, σ) gives the probability that x is between a and b
   (e) invnorm(A, μ, σ) gives the x value with area A to the left of it
9. sampling distribution
   (a) empirical/sample statistics $(\overline{x}, s)$ versus theoretical/population statistics (μ, σ).
   (b) $\overline{x}$ (the group average) is itself a random variable, and
   $$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$
   this is called the "standard error" of $\overline{x}$
   (c) Central Limit Theorem: if the individuals are independent, then as n gets larger,
   $$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
   (d) use the CLT when you are interested in the sample/group mean or total
10. Linear regression and correlation
   (a) scatterplots
   (b) use LinRegTtest to find and work with the regression line $y = a + bx$ that best fits the data.
   (c) the slope of the line is the coefficient of the x.
   (d) You can use the regression line to estimate y for a given x:
      - interpolation if x is within known points
      - extrapolation if it lies outside known points
   (e) the sample correlation r is between -1 and 1
   (f) $|r|$ measures how tightly the cloud of data points cluster around the regression line.
   (g) r has the same sign as the slope
   (h) Be able to explain positive/negative and weak/strong correlation intuitively.
   (i) Know the difference between correlation and causation.
11. models
   (a) be able to select an appropriate model for the given problem or data
   (b) discrete (e.g. binomial) vs continuous (e.g. normal) distributions
   (c) area under the graph represents probability
   (d) for a continuous random variable and a number c, $P(x = c) = 0$; because a line contains zero area
   (e) translate words into equations or inequalities, and then find probability using PDF or CDF
   (f) be able to work two step problems
   (g) be able to solve for parameters (e.g. n, p, μ, σ) to satisfy a given requirement, possibly using guess-check
   (h) remember that many models assume that observations are independent